LINEAR REGRESSION MODEL

Yiqiao YIN
Columbia University

December 11, 2018

Abstract

This is lecture notes for Linear Regression Model offered at Columbia University in 2018 Fall semester. The story of linear regression is well known and I have had the luxury of going over these concepts in details. It comes to my attention that it is worth documenting the experience down in this file for future generations.

Contents

This document is dedicated to Professor Gabriel Young.

1 Statistical Model

Go back to Table of Contents. Please click TOC

1.1 Introduction

A mathematical model is a description of a system using mathematical concepts and language. Consider bacterial growth, we have model $dy/dt = ky$, while $y = A_0 e^{1/t}$. Alternatively, we can use a regression model, $Y = \beta_0 + \beta_1 X_1 + \dots \beta_n X_n$. In this case we are using a functional relation. A functional relation between two variables is expressed by a mathematical formula. If x is the independent var. and y is the dependent variable, then a function relation is of the form

$$y = f(x).$$

The relation is deterministic and not random. On the other hand, a statistical relation, is not a perfect one. In general, the observations for a statistical relation do not fall directly on the curve of relationship. This is commonly expressed as a functional relation coupled with a random error ϵ. If x is the independent variable and Y is the dependent variable, then a statistical relation often takes the form:

$$Y = f(x) + \epsilon$$

while Y and ϵ are random yet $f(x)$ is not random. A statistical relation is also commonly expressed in terms of conditional expectation. That is, for random variables Y and X,

$$Y = E[Y|X = x] + \epsilon$$

which is a function of X variable, and hence, a form of $Y = f(x) + \epsilon$.

The conditional probability mass function of $Y|X = x$ is defined by

$$p(y|X = x) = \frac{P(X = x, Y = y)}{p(X = x)}$$

where $P(X = x, Y = y)$ is the joint distribution of X and Y and $P(X = x)$ is the marginal distribution of X. Note: $P(X = x) > 0$ for all x.

Definition 1.1.1. The conditional expectation of $Y|X = x$ is defined by

$$E[Y|X = x] = \sum_y yp(y|X = x)$$

which is a function of X, taking the form $f(x)$. Note that we can also define conditional variance $\text{Var}[Y|X = x]$.

Remark 1.1.2. Page 5 of slides introduced questions. We answer them here.

1. To estimate $E[Y|X = x]$, we need to choose a method.
2. Some increasing function, maybe on a linear or quadratic.
3. Both X and Y are continuous.
4. Should X and Y be assumed normal? We can't have negative measurement.
5. In a clinical trial, X is typically not random, i.e., dosage level of a drug. In this example,

Definition 1.1.3. Let X and Y be two continuous random variables. The conditional probability density function of $Y|X = x$ is defined by

$$f(y|X = x) = \frac{f(x,0)}{f_X(x)},$$

where $f(x,y)$ is he joint density of X and Y and $f_X(x)$ is the marginal density of X. Note: $f_X(x) > 0$ for all x.

Definition 1.1.4. Let X and Y be two continuous random variables and let $f(y|X = x)$ be the conditional density function of $Y|X = x$. The conditional expectation of $Y|X = x$ is defined by

$$E[Y|X = x] = \int y f(y|X = x) dy.$$

Proposition 1.1.5. *If (X,Y) is a random vector from the bivariate normal distribution, then the conditional expectation and variance of Y given $X = x$ are*

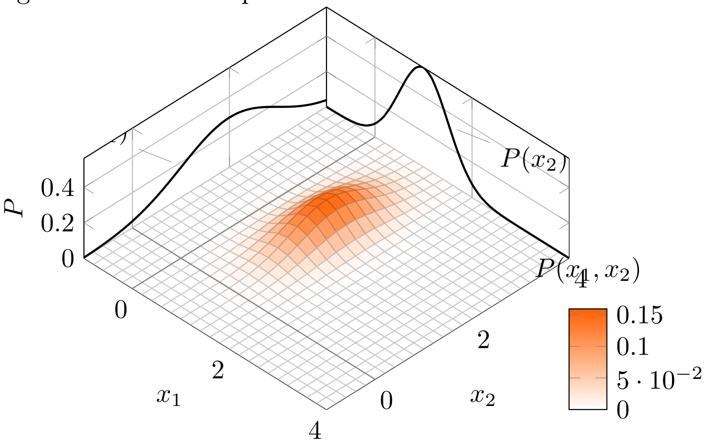
$$E[Y|X = x] = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x - \mu_X) = \beta_0 + \beta_1 x$$

and

$$Var[Y|X = x] = \sigma_Y^2(1 - \rho^2)$$

The three-dimensional plot is presented in the following graph.

Figure 1: This is the picture for bivariate normal distribution.



Remark 1.1.6. We assume linearity for our model. Moreover, we assume that variance of response stays constant.

How do we estimate β_0 and β_1? We can simply guess.

$$\beta_0 = \mu_Y - \beta_1\mu_X, \beta_1 = \rho\frac{\sigma_Y}{\sigma_X}$$

$$\hat{\beta}_0\bar{Y} - \hat{\beta}_1\bar{X}, \hat{\beta}_1 = r\frac{s_Y}{s_X}$$

The above result is the least squares or MLE estimators for β_0 and β_1.

2 Linear Regression Model

Go back to Table of Contents. Please click TOC

2.1 Simple Linear Regression

Let us introduce notation first. Let Y be dependent variable, or response variable. Let x be independent variable, covariate, or predictor. There are n paired observations $(x_1, Y_1), ..., x_n, Y_n)$. The linear regression model states that with parameters β_0, β_1 and σ^2, we have model

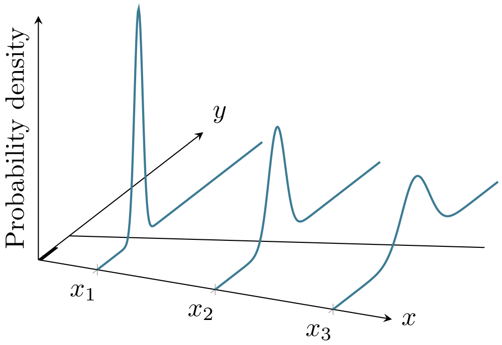
$$Y_i = \beta_0 + \beta_1 x_1 + \epsilon, i = 1, 2, .., n$$

while $\epsilon_i \sim N(0, \sigma^2)$ i.i.d.. In this case, we consider Y_i to be random and x_i to be fixed. We consider ϵ_i to be random error term while β_0 and β_1 are not random. One can think of $\beta_0 + \beta_1 x$ as $E[Y|X = x]$.

To prove the form of expectation and variance of Y, consider the following

$$\begin{aligned}
E[Y_i] &= E[\beta_0 + \beta_i x_i + \epsilon_i] \\
&= E[\beta_0 + \beta_i x_i] + E[\epsilon_i] \\
&= \beta_0 + \beta_i x_i + 0 \\
&= \beta_0 + \beta_i x_i \\
\text{var}[Y_i] &= \text{var}[\beta_0 + \beta_i x_i + \epsilon] \\
&= \sigma^2
\end{aligned}$$

Figure 2: This is the plot for linear regression model. Each x_i there is an estimated response by using $\beta_0 + \beta_i x_i + \epsilon_i$. Note that the normal distribution along the curve should be the same.



2.2 Random Independent Variable

To investigate the question "why isn't the independent variable random?" Consider the following. Set up by assuming independent variable is a random variable, i.e. $X \sim f_X(x)$. Assume normal distribution for error terms. Define response as random variable $Y = \beta_0 + \beta_1 X + \epsilon$, which is a simple linear regression with X random.

We define $Z = X = g_1(x, \epsilon)$, $Y = \beta_0 + \beta_1 x + \epsilon = g_0(x, \epsilon)$, while $X = Z$ and $\epsilon = Y - (\beta_0 + \beta_1 Z)$. Then we need to find the Jacobi

$$|J| = \begin{bmatrix} \frac{\partial x}{\partial z} & \frac{\partial x}{\partial y} \\ \frac{\partial \epsilon}{\partial z} & \frac{\partial \epsilon}{\partial y} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -\beta_1 & 1 \end{bmatrix} = 1$$

while we have

$$\begin{aligned} f_{z,y}(z, y) &= f_{x,\epsilon}(z, y - (\beta_0 + \beta_1 z)) \\ &= f_X(z) f_\epsilon(y - (\beta_0 + \beta_1 z)) \cdot \underbrace{1}_{\text{from Jacobi}} \\ &= f_x(z) \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{1}{2\sigma^2}(y - (\beta_0 + \beta_1 z))\} \\ f_{Y|Z=z}(y) &= \frac{f_{Z,Y}(z, y)}{f_Z(z)} \end{aligned}$$

and then we get $E[Y|Z = z] = \beta_0 + \beta_1 z$ or $E[Y|X = x] = \beta_0 + \beta_1 x$. This bivariate transformation problem will illustrate that the transformation will result in simple linear regression.

2.3 Least Squares Method

Deviations are explained in the following way. $y_i - \bar{y}$ is the deviation of each case y_i from sample mean of response $\bar{y}$. $x_i - \bar{x}$ is the deviation of each case x_i from the sample mean of the predictor variable $\bar{x}$. We also have $(x_i - \bar{x})(y_i - \bar{y})$ to be the product of the deviations.

For sum of squares, we have

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{2} x_i^2 - \frac{1}{n}\Big(\sum_{i=1}^{n} x_i\Big)^2$$

$$S_{yy} = \text{SST} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - \frac{1}{n}\Big(\sum_{i=1}^{n} y_i\Big)^2$$

$$S_{xy}$$

while $\hat{\text{cov}}(x, y) = \frac{s_{xy}}{s_{yy}}$.

Denote some line by $\hat{y} = b_0 + b_1 x$ and let the line at a point (x_i, y_i) be denoted $\hat{y}_i$ which can be expressed by parameters and x_i.

Proposition 2.3.1. *Let*

$$Q(b_0, b_1) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_i))^2$$

Then Q is minimized when

$$\hat{\beta}_1 = b_1 = \frac{S_{xy}}{S_{xx}},$$

and

$$\hat{\beta}_0 = b_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

The proof is trivial. One can simply take partial derivative of Q with respect to b_0 and b_1 to get the estimate.

Proposition 2.3.2. *The line of best fit crosses the point $(\bar{x}, \bar{y})$.*

Proof. Consider

$$\begin{aligned}\hat{y} &= \hat{\beta}_0 + \hat{\beta}_1\bar{x} \\ &= (\bar{y} - \hat{\beta}_1\bar{x}) + \hat{\beta}_1\bar{x} \\ &= \bar{y}\end{aligned}$$

The interpretation of the slope β_1 (or $\hat{\beta}_1$) is for each unit increase of x, the average of the response variable increases (or decreases) by $\hat{\beta}_1$ units. □

Definition 2.3.3. The mean square error denoted by MSE is defined by

$$\text{MSE} = \frac{\text{SSE}}{n-2} = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Example 2.3.4. For one data point, you can draw any line and there is no unit of dispersion. Unit of dispersion starts when you have at least two data points.

Definition 2.3.5. The coefficient of determination denoted r^2 is the proportion of variation in the response variable y explained by the model (or explained by covariate x). The computational formula is given

$$r^2 = 1 - \frac{\text{SSE}}{\text{SST}} = 1 - \frac{\text{SSE}S_{yy}}{=}\frac{\text{SST} - \text{SSE}}{\text{SST}}$$

and note that the r^2 is sample correlation square.

Proposition 2.3.6. *The following properties are of the slope and variance estimators*

1. *$\hat{\beta}_1$ is an unbiased estimator of β_1, $E[\hat{\beta}_1] = \beta_1$*
2. *$\hat{\beta}_0$ is an unbiased estimator of β_0, $E[\hat{\beta}_0] = \beta_0$*
3. *MSE is an unbiased estimator of σ^2, $E[MSE] = \sigma^2$*

Let us note that $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$ is random.

1. Consider the following

$$\begin{aligned}S_{xy} &= \sum(x_i - \bar{X})(Y_i - \bar{Y}) \\ &= \sum(x_i - \bar{X})Y_i - \sum(x_i - \bar{X})\bar{Y} \\ &= \sum(x_i - \bar{X})Y_i - \bar{Y}\sum(x_i - \bar{X}) \\ &= \sum(x_i - \bar{X})Y_i, \text{ note } \sum(x_i - \bar{X}) = 0\end{aligned}$$

2. For the second, consider

$$\sum(x_i - \bar{X}) = \sum x_i - n\bar{X} = \sum x_i - \sum x_i = 0$$

3. For the third, we have

$$\begin{aligned} s_{xx} &= \sum(x_i - \bar{x})^2 \\ &= \sum(x_i - \bar{x})(x_i - \bar{x}) \\ &= \text{expand it} \\ &= \sum(x_i - \bar{x})x_i \end{aligned}$$

Proof. Let us prove the proposition.

1. We want to show $E[\hat{\beta}_1] = \beta_1$. Consider

$$\begin{aligned} E[\hat{\beta}_1] &= E[\frac{s_{xy}}{S_{xx}}] = E[\frac{\sum(x_i - \bar{x})Y_i}{s_{xx}}] \\ &= \underbrace{E[\sum(\frac{s_i - \bar{X}}{s_{xx}})Y_i]}_{\text{linear comb. of } Y_i} \\ &= \frac{1}{s_{xx}}\sum(x_i - \bar{x})E[Y_i] \\ &= \frac{1}{s_{xx}}\sum(x_i - \bar{X})\underbrace{(\beta_0 + \beta_1 X_i)}_{\text{from } E[Y_i]} \\ &= \frac{\beta_0}{s_{xx}}\underbrace{\sum(x_i - \bar{x})}_{=1} + \frac{\beta_1}{s_{xx}}\sum(x_i - \bar{x})x_i \\ &= \beta_1 \end{aligned}$$

2. Rest of the proof is in text [1].
3. Note $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = \bar{y} - \hat{\beta}_1\bar{x} + \hat{\beta}_1 x_i = \bar{y} + \hat{\beta}_1(x_i - \bar{x})$, which is a common way to express least square line of fit. Then we have

$$\begin{aligned} \sum e_i &= \sum(y_i - \hat{y}_i) \\ &= \sum(y_i - (\bar{y} + \hat{\beta}_1(x_i - \bar{X}))) \\ &= \sum y_i - n\bar{y} - \hat{\beta}_1\sum(x_i - \bar{X}) \\ &= \sum y_i - \sum y_i - \hat{\beta}_1 0 \\ &= 0 \end{aligned}$$

4. $\sum e_i = 0$
5. Page 19, proof of (iii). Under linear model, we have

$$W = \frac{(n-2)\text{MSE}}{\sigma^2} \sim \chi^2(\text{df} = n - 2)$$

and note that

$$\begin{aligned}\frac{(n-2)\text{MSE}}{\sigma^2} &= \frac{\sum(Y_i - \hat{Y}_i)^2}{\sigma^2} \\ &= \sum(\frac{Y_i - \hat{Y}_i - E[Y_i - \hat{Y}_i]}{\sigma^2})^2\end{aligned}$$

We want to show $E[\text{MSE}] = \sigma^2$. We have

$$\begin{aligned}E[\text{MSE}] &= \frac{\sigma^2}{(n-2)} \cdot \frac{(n-2)}{\sigma^2} E[\text{MSE}] \\ &= \frac{\sigma^2}{(n-2)} E[\frac{(n-2)\text{MSE}}{\sigma^2}] \\ &= \frac{\sigma^2}{(n-2)} E[W] \\ &= \frac{\sigma^2}{(n-2)} (n-2) \\ &= \sigma^2\end{aligned}$$

□

Definition 2.3.7. The residual denoted e_i is the difference between the observed value y_i and its corresponding fitted value $\hat{y}_i$,

$$e_i = y_i - \hat{y}_i$$

The distinction between e_i and ϵ_i is the following. Residuals are from data: $e_i = y_i - \hat{y}_i$ which is not random. Random variable $e_i = Y_i - \hat{Y}_i$ which is random.

2.4 Probability Distributions of Estimators and Residuals

Theorem 2.4.1. *Let $Y_1, Y_2, ..., Y_n$ be an indexed set of independent normal random variables. Then for real numbers $a_1, a_2, ..., a_n$, the random variable $W = a_1Y_1 + a_2Y_2 + \cdots + a_nY_n$ is normally distributed with mean*

$$E[W]$$

and

$$Var[W]$$

We want to express least squares estimators as linear combination of response values

Y_i. Then we have

$$\begin{aligned}
\hat{\beta}_1 &= \frac{1}{s_{xx}} \sum (x_i - \bar{X}) Y_i \\
&= \sum_{i=1}^{n} (\frac{x_i - \bar{X}}{s_{xx}}) Y_i \\
&= \sum_{i=1}^{n} K_i Y_i, K_i \equiv (\frac{x_i - \bar{X}}{s_{xx}}) \\
\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{X} \\
&= \sum_{i=1}^{n} \frac{Y_i}{n} - \bar{X} \sum_{i=1}^{n} K_i Y_i \\
&= \sum_{i=1}^{n} (\frac{1}{n} - \bar{x} K_i) Y_i \\
&= \sum_{i=1}^{n} c_i Y_i, c_i \equiv (\frac{1}{n} - \bar{x} K_i)
\end{aligned}$$

Theorem 2.4.2. *Theorem 2.2 from notes. under the conditions of regression model, least squares estimator $\hat{\beta}_1$ is normally distributed with mean β_1 and variance σ^2/s_{xx}.*

Proof. We have

$$\begin{aligned}
E[\hat{\beta}_1] &= \beta_1 \\
\text{var}[\hat{\beta}_1] &= \text{var}[\sum K_i Y_i] \\
&= \sum K_i^2 \text{var}(Y_i) \\
&= \sum K_i^2 \sigma^2 \\
\text{note: } \sum K_i^2 &= \sum (\frac{x_i - \bar{X}}{s_{xx}}) \\
&= \frac{1}{s_{xx}^2} \sum (x_i - \bar{X})^2 \\
&= \frac{1}{s_{xx}}
\end{aligned}$$

since $\hat{\beta}_1$ is a linear combination of normal random variables. □

Theorem 2.4.3. *(Gauss-Markov Theorem) Under the conditions of regression model (2.1), the least squares estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased and have minimum variance among all other unbiased linear estimators.*

Proof. We will only prove this for $\hat{\beta}_1$. To show $\hat{\beta}_1$ has minimum variance among all other unbiased linear estimators, consider a new estimator $\hat{\beta}_1^*$, where $\hat{\beta}_1^* = \sum_{i=1}^n a_i Y_i$ and $E[\hat{\beta}_1^*] = \beta_1$. Then

$$E[\hat{\beta}_1^*] = \beta_0 \sum_{i=1}^{n} a_i + \beta_1 \sum_{i=1}^{n} x_i a_i = \beta_1$$

which implies $\sum a_i = 0$ and $\sum x_i a_i = 1$. The variance of $\hat{\beta}_1^*$ is

$$\text{var}[\hat{\beta}_1^*] = \sigma^2 \sum_{i=1}^n a_i^2$$

Let us define $a_i \equiv K_i + d_i$, where $K_i = \frac{x_i - \bar{X}}{s_{xx}}$. Then

$$\begin{aligned}
\text{var}(\hat{\beta}_1) &= \sigma^2 \sum_{i=1}^n (K_i + d_i)^2 \\
&= \sigma^2 (\sum_{i=1}^n K_i^2 + \sum_{i=1}^n d_i^2 + 2 \sum_{i=1}^n K_i d_i) \\
\text{note: } \sum_{i=1}^n K_i d_i &= \sum_{i=1}^n K_i (a_i - K_i) \\
&= \sum_{i=1}^n K_i a_i - \sum_{i=1}^n K_i^2 \\
&= \sum_{i=1}^n a_i (\frac{x_i - \bar{X}}{s_{xx}}) - \frac{1}{s_{xx}} \\
&= \quad \underbrace{\frac{\sum a_i x_i - \bar{x} \sum a_i}{s_{xx}}}_{\text{note: } \sum a_i x_i = 1 \text{ and } \bar{X} \sum a_i = 0} \quad - \frac{1}{s_{xx}} \\
\text{var}(\hat{\beta}_1^*) &= \sigma^2 \sum K_i^2 + \sigma^2 \sum d_i^2 \\
&= \text{var}(\hat{\beta}_1) + \sigma^2 \sum d_i^2
\end{aligned}$$

The smallest value of d_i is zero and $\text{var}(\hat{\beta}_1^*)$ is at a minimum when $\sum d_i^2 = 0$, which can only happen if $d_i = 0$ for all i. Hence, $a_i = K_i$ which proves the desired result. □

We want to express fitted values $\hat{Y}_i$ as a linear combination of the response values Y_i. Recall that $c_j \equiv \frac{1}{n} - \bar{x} K_i$ while $K_j \equiv \frac{x_j - \bar{X}}{sxx}$. This case we have

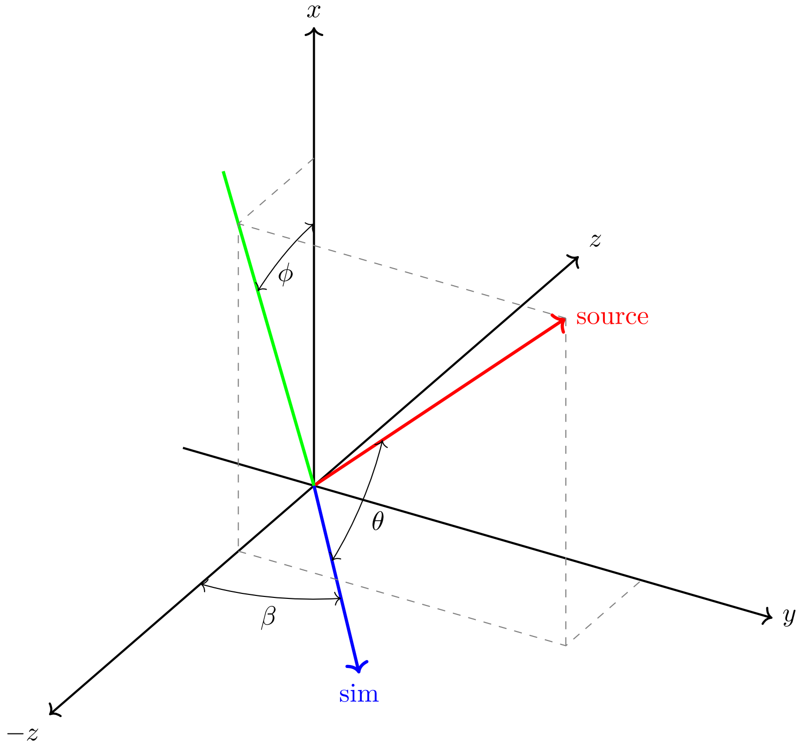
$$\begin{aligned}
\hat{Y}_i &= \hat{\beta}_0 + \hat{\beta}_1 x_i \\
&= \sum_{j=1}^n c_j Y_j + \sum_{j=1}^n K_j Y_j \\
&= \sum_{j=1}^n \left(\frac{1}{n} - \bar{x} \left(\frac{x_j - \bar{x}}{s_{xx}} \right) + \left(\frac{x_j - \bar{x}}{s_{xx}} \right) x_i \right) Y_j \\
&= \sum_{j=1}^n \left(\frac{1}{n} + \frac{(s_j - \bar{x})(x_i - \bar{x})}{s_{xx}} \right) Y_j \\
&= \sum_{j=1}^n h_{ij} Y_j
\end{aligned}$$

Note that h_{ij} is the "hat matrix" which means the $(i,j)^{th}$ element of the hat matrix H.

Theorem 2.4.4. *Let us state some properties of the hat matrix H.*

1. $h_{ij} = h_{ji}$ *symmetric matrix, transpose is itself*
2. $\sum_{j=1}^{n} h_{ij} = 1$ *Consider* $1^T = [1, ..., 1]$
3. $\sum_{j=1}^{n} h_{ij} x_j = x_i$
4. $\sum_{j=1}^{n} h_{ij}^2 = h_{ii}$ *Indempotent matrix*
5. $\sum_{i=1}^{n} h_{ii} = 2$ *number of* β_i *parameters*

Figure 3: This is the projection graph from Theorem 2.4 in class note.



Theorem 2.4.5. *(Theorem 2.5 from lecture)*

$$E[\hat{Y}_i] = \beta_0 + \beta_1 X_i$$

and

$$var[\hat{Y}_i] = \sigma^2 \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{s_{xx}} \right)$$

Proof. Consider

$$\begin{aligned} E[\hat{Y}_i] &= E[\sum_{j=1}^{n} h_{ij} Y_j] \\ &= \beta_0 \sum_{j=1}^{n} h_{ij} + \beta_1 \sum_{j=1}^{n} h_{ij} x_j + 0 \\ &= \beta_0 \cdot 1 + \beta_1 x_i \end{aligned}$$

and we can also consider variance

$$\begin{aligned} \text{var}[\hat{Y}_i] &= \text{var}[\sum_{j=1}^{n} h_{ij} Y_j] \\ &= \sum_{j=1}^{n} h_{ij}^2 \text{var}[Y_j] \\ &= \sigma^2 \sum_{j=1}^{n} h_{ij}^2 \\ &= \sigma^2 h_{ii}, \text{ from indempotent matrix property} \\ &= \sigma^2 \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{s_{xx}} \right) \end{aligned}$$

and $\hat{Y}_i$ is a linear combination of normal random variables. We have distribution

$$\hat{Y}_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2 h_{ii})$$

□

Theorem 2.4.6. *This is Theorem 2.6 from lecture. We have*

$$E[e_i] = 0, var[e_i] = \sigma^2(1 - h_{ii}), h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{s_{xx}}$$

Proof. We prove the following

$$\begin{aligned} e_i &= Y_i - \hat{Y}_i \\ &= Y_i - \sum_{j=1}^{n} h_{ij} Y_j \\ E[e_i] &= E[Y_i - \sum_{j=1}^{n} h_{ij} Y_j] \\ &= \beta_0 + \beta_1 x_i - (\beta_0 + \beta_1 x_i) \\ &= 0 \\ \text{var}[e_i] &= \text{var}[Y_i - \underbrace{\sum_{j=1}^{n} h_{ij} Y_j}_{\text{consists of} Y_i}] \\ &= \text{var}[Y_i - h_{ii} Y_i - \sum_{i \neq j} h_{ij} Y_j] \\ &= \text{var}[(1 - h_{ii}) Y_i - \sum_{i \neq j} h_{ij} Y_j] \\ &= (1 - h_{ii})^2 \sigma^2 + \sum_{i \neq j} h_{ij}^2 \sigma^2 \\ &= (1 - 2h_{ii} + h_{ii})^2 \sigma^2 + \sum_{i \neq j} h_{ij}^2 \sigma^2 \\ &= (1 - 2h_{ii}) \sigma^2 + \sum_{j=1}^{n} h_{ij}^2 \sigma^2 \end{aligned}$$

Then we have e_i's is a linear combination and we have normal distribution $e_i = N(0, \sigma^2(1 - h_{ii}))$. □

Let us introduce relationship between the slope and intercept (from Fall 2017 midterm).

Theorem 2.4.7. *This is Theorem 2.8 from lecture. Let $\hat{\beta}_1$ and $\hat{\beta}_0$ be the least squares estimators of β_1 and β_0. Then*

$$cov(\hat{\beta}_1, \hat{\beta}_0) = -\bar{x} var[\hat{\beta}_1]$$

This is the covariance structure between $\hat{\beta}_0$ and $\hat{\beta}_1$. Note that $\text{var}(\hat{\beta}_1) = \frac{\sigma^2}{s_{xx}}$.

2.5 Maximum Likelihood Estimation

Consider a random sample $X_1, X_2, ..., X_n$ each having common probability density function (or probability mass function) $f(x_i|\theta)$ where θ is a generic parameter of that distribution. θ could also be a vector of parameters. The joing density function (or joint probability mass function) is

$$f(x_1, x_2, ..., x_n|\theta) = f(x_1|\theta) \times f(x_2|\theta) \times \cdots \times f(x_n|\theta)$$

Define the likelihood function as

$$\mathcal{L}(\theta; x_1, x_2, ..., x_n) = f(x_1, x_2, ..., x_n|\theta)$$

which is often convenient working with the log-likelihood function

$$\log(\mathcal{L}(\theta : x_1, ..., x_n)) = \log(f(x_1, ..., x_n|\theta))$$

Example 2.5.1. Let $X_1, ..., X_n$ be a random sample fro man exponential distribution each having common probability density function $f(x_i|\mu) = \frac{1}{\mu}\exp(-\frac{1}{\mu}x_i)$, for $x_i \geq 0$. Find the MLE estimator of μ.

The solution is

$$\begin{aligned}
\mathcal{L}(\mu) &= \prod_{i=1}^{n}(\frac{1}{\mu}\exp(-\frac{1}{\mu}x_i)) \\
&= (\frac{1}{\mu})^n \exp(-\frac{1}{\mu}\sum x_i) \\
l(\mu) &= \log\mathcal{L}(u) \\
&= -n\log(\mu) - \frac{1}{\mu}\sum x_i \\
\frac{dl}{d\mu} &= -\frac{n}{\mu} + \frac{1}{\mu^2}\sum x_i \\
&= 0, \text{ set to} \\
\frac{n}{\mu} &= \frac{1}{\mu^2}\sum x_i
\end{aligned}$$

and thus we have

$$\hat{\mu}_{\text{MLE}} = \frac{1}{\mu}\sum x_i$$

Moreover, we can parametrize $f(x|\lambda) = \lambda e^{-\lambda x}$ which gives us $\hat{\lambda} = \frac{1}{\bar{x}}$.

Let us discuss maximum likelihood estimators for parameters in linear regression. Consider random variable $Y_1, ..., Y_n$ satisfying simple linear regression

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \text{ for } i = 1, ..., n, \text{ and } e_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$$

Recall the probability density function

$$f(y_i|\beta_0, \beta_1, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\Big\{-\frac{1}{2\sigma^2}(y_i - (\beta_0 + \beta_1 x_i))^2\Big\}$$

Consequently, we compute

$$\begin{aligned}
\mathcal{L}(\beta_0, \beta_1, \sigma^2; y_i) &= f(y_1|\beta_0, \beta_1, \sigma) \times \cdots \times f(y_n|\beta_0, \beta_1, \sigma) \\
&= \prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\Big(-\frac{1}{2\sigma^2}(y_i - (\beta_0 + \beta_1 x_i))^2\Big) \\
&= \Big(\frac{1}{\sqrt{2\pi\sigma^2}}\Big)^2 \exp\Big(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_i))^2\Big)
\end{aligned}$$

The log-likelihood function is

$$\log(\mathcal{L}(\beta_0, \beta_1, \sigma^2; y_i)) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_i))^2$$

Taking partial derivative of the equation above with respect to β_1, β_0, and σ^2, we can obtain

$$\hat{\beta}_{1,\text{MLE}} = \hat{\beta}_1 = \frac{s_{xy}}{s_{xx}}$$
$$\hat{\beta}_{0,\text{MLE}} = \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$
$$\hat{\sigma}_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{n-2}{n}\text{MSE}$$

which gives us the maximum likelihood of parameters in the simple linear regression.

Remark 2.5.2. Note the following:

1. The least squares estimates are the same as the maximum likelihood estimates for parameters β_0 and β_1.
2. The maximum likelihhod estimates for σ^2 is biased. This bias becomes negligible for large n. We have E[MSE] $= \sigma^2$ which is unbiased. Then we have $\text{E}[\hat{\sigma}^2_{\text{MSE}}] = \text{E}[\frac{n-2}{n}\text{MSE}] = \frac{n-2}{n}\sigma^2$, which is consistent since $\lim_{n\to\infty}\frac{n-2}{n} = 1$.

2.6 Inferences About Slope Parameter

To assess statistical relationship between response variable Y and covariate x, we want to test the slope parameter. Consider null

$$H_0 : \beta_1 = (\beta_1)_0$$

The most common hypothesized value is zero

$$H_0 : \beta_1 = 0$$

To construct a reasonable test statistic for H_0, we will follow the usual procedure. We want to standardize the slope estimator $\hat{\beta}_1$, i.e.,

$$\text{stat} = \frac{\hat{\beta}_1 - E[\hat{\beta}_1]}{\sigma_{\hat{\beta}_1}}$$

Recall the following. $\text{E}[\hat{\beta}_1] = \beta_1$ unbiased. The variance of estimator $\hat{\beta}_1$ is $\text{var}(\hat{\beta}_1) = \frac{\sigma^2}{s_{xx}}$. The standard error of estimator $\hat{\beta}_1$ is $\frac{\sigma}{\sqrt{s_{xx}}}$. If we standardize $\hat{\beta}_1$, we get

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\sigma^2}{s_{xx}}}}$$

which is the test statistics for testing H_0. In practice, we would use estimate (studentize) so we would have

$$T = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\text{MSE}}{s_{xx}}}}$$

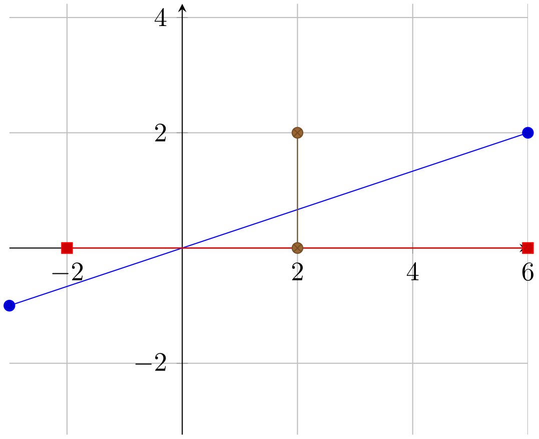
which is what we would use for test statistics as going through simple linear regression. In regression output from a program, we have "estimate" to be $\hat{\beta}_i$, "standard error" to be $\sqrt{\frac{\text{MSE}}{s_{xx}}}$, t-value to be estimate divided by standard error, and "Pr($>|t|$)" would be two-tail p-value. If we are doing one-tail test, we would want to divde this value by 2.

Remark 2.6.1. A statistically significant slope does not always imply a strong correlation. Recall that power is the probability reject null when null is false. Generally, the power of a testing procedure increases as n (sample size) increases. When testing $H_0 : \beta_1 = 0$, we will eventually show significance with large enough n. We can look at R^2 as well.

2.7 Analysis of Variance Approach to Regression Analysis

Consider the following graph

Figure 4: Caption



and the line is $\hat{\beta}_0 + \hat{\beta}_1 x$. At a point x, we have red line to be $\bar{y}$ and we have estimate at x a response $\hat{y}$. This gives us

$$Y_i - \bar{Y} = \hat{Y}_i - \bar{Y} + Y_i + \hat{Y}_i$$

Left hand side is the total deviation of response. The first term of the right hand side is deviation around the fitted regression value around the mean. The second term of the right hand side is deviation around the fitted line.

$$n - 1 = 1 + n - 2$$

Left hand side is total degree of freedom.

Definition 2.7.1. Define the sums of squares regression to be

$$\text{SSR} = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 4$$

Proof. Consider

$$\begin{aligned}\text{SSR} &= \sum (\hat{y}_i - \bar{y})^2 \\ &= \sum (\bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta} x_i - \bar{y})^2 \\ &= \hat{\beta}_1^2 \sum (x_i - \bar{x})^2 \\ &= \hat{\beta}^2 s_{xx}\end{aligned}$$

□

Proposition 2.7.2. *Total variation SST can be partitioned into two sources of variable SSR and SSE. This can be represented with the additive identity*

$$SST = SSR + SSE$$

Proof. Note that $\sum(x_i - \bar{x})e_i = \sum \hat{y}_i e_i - \bar{y}\sum e_i = 0$ and we can compute the following

$$\begin{aligned}
\text{SST} &= \sum(y_i - \bar{y})^2 \\
&= \sum(\hat{y}_i - \bar{y} + y_i - \hat{y}_i)^2 \\
&= \sum(\hat{y}_i - \bar{y})^2 + 2(\hat{y}_i - \bar{y})\underbrace{(y_i - \bar{y}_i)}_{e_i} + (y_i - \hat{y}_i)^2 \\
&= \text{SSR} + 0 + \text{SSE} \\
&= \text{SSR} + \text{SSE}
\end{aligned}$$

□

Proposition 2.7.3. *The expected value of SSR is*

$$E[SSR] = \sigma^2 + \beta_1^2 s_{xx}$$

Proof. For any random variable w, $E[w^2] = \text{var}(w) + (Ew)^2$. Then

$$\begin{aligned}
E[\hat{\beta}^2] &= \text{var}(\hat{\beta}_1) + (E\hat{\beta}^2) \\
&= \frac{\sigma^2}{s_{xx}} + \beta_1^2 \\
\Leftrightarrow s_{xx}E[\hat{\beta}_1^2] &= \sigma^2 + \beta_1^2 s_{xx} \\
\Leftrightarrow E[\text{SSR}] &= \sigma^2 + \beta_1^2 s_{xx}
\end{aligned}$$

□

The motivation of F-statistics: on average

$$F \sim \frac{E[\text{SSR}]}{E[\text{MSE}]} = \frac{\sigma^2 + \beta_1^2 s_{xx}}{\sigma^2} = 1 + \frac{\beta_1^2 s_{xx}}{\sigma^2}$$

If $H_0 : \beta_1 = 0$ is true, then $F \sim 1$ on average. If $H_0 : \beta_1 = 0$ is false, then $F > 1$ (much larger) on average.

Remark 2.7.4. This is a special case of "Cochran's Theorem". If $\beta_1 = 0$ is true, then all Y_i have the same mean $\mu = \beta_0$ and the same variance σ^2. Then $\frac{\text{SSE}}{\sigma^2}$ and $\frac{\text{SSR}}{\sigma^2}$ are independent χ^2 random variables.

Proposition 2.7.5. *Let T be distributed student's t-distribution with degrees of freedom v. Then the random variable T^2 has an F-distribution with degrees of freedom 1 and v. Namely,*

$$T^2 \sim F(1, v)$$

Definition 2.7.6. Consider a realized data set $y_1, ..., y_n$. The likelihood ratio test statistics for testing $H_0 : \theta \in \Theta_0$ versus $H_A : \theta \in \Theta_0^c$ is

$$\lambda(y_1, ..., y_n) = \frac{\max\limits_{\Theta_0} \mathcal{L}(\theta; y_1, ..., y_n)}{\max\limits_{\Theta} \mathcal{L}(\theta; y_1, y_2, ..., y_n)}$$

Remark 2.7.7. You can derive the general linear test through likelihood ratio test.

Remark 2.7.8. Let us note the following.

1. Define the rejection region if $\lambda \leq c$, where $0 \leq c \leq 1$

2. Θ is the full parameter space
3. Θ_0 is the null space, and Θ_0^c is the alternative space
4. $\Theta = \Theta_0 \cup \Theta_0^c$. They are complement of each other within Θ.

Proof. Full Model. $Y = \beta_0 + \beta_1 X + \epsilon$, with $\epsilon \sim N(0, \sigma^2)$. Then

$$\begin{aligned}
\mathcal{L}(\beta_0, \beta_1, \sigma^2) &= \big(\frac{1}{\sqrt{2\pi\sigma^2}}\big)^2 \exp\big\{\frac{-1}{2\sigma^2}\sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2\big\} \\
\Theta \Rightarrow \max_{\Theta} \mathcal{L} \Rightarrow \hat{\beta}_1 &= \frac{s_{xy}}{s_{xx}}, \\
\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\
\hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^n (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2 \\
\mathcal{L}(\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2) &= [2\pi \frac{1}{\pi} \sum (y_i - \hat{y}_i)^2]^{-\frac{n}{2}} \exp\{-\frac{n}{2}\}
\end{aligned}$$

Now let us discuss reduced model. $Y = \beta_0 + \epsilon$, with $\epsilon \sim N(0, \sigma^2)$.

$$\begin{aligned}
\Theta_0 \Leftrightarrow H_0 : \beta_1 = 0 :=& \\
\mathcal{L}(\beta_0, 0, \sigma^2) &= \big(\frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{1}{2\sigma^2} \sum (y_i - \beta_0)^2\}\big)^2 \\
\max_{\Theta_0} \mathcal{L} \Rightarrow \hat{\beta}_0 &= \bar{y} \\
\hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^n (y_i - \bar{y})^2
\end{aligned}$$

Then we have

$$\lambda(y_1, ..., y_n) = \lambda(\underline{y}) = \left[\frac{\sum (y_i - \bar{y})^2}{\sum (y_i - \hat{y}_i)^2}\right]^{-\frac{n}{2}}$$

□

The general linear test gives F-statistics

$$f_{\text{calc}} = [\frac{\text{SSE}_r - \text{SSE}_F}{\text{df}_e - \text{df}_F}] / \frac{\text{SSE}_f}{\text{df}_F} = ?$$

Another way

$$f = \left([[\frac{SST}{SSE}]^{-n/2}]^{-\frac{n}{2}} \right)(n-2)$$

Remark 2.7.9. Note: df_F is the residuals DF from ANOVA table. SSE_F is the residuals sum of square from ANOVA table.

Remark 2.7.10. Consider homework question: $Y = \beta X + \epsilon$. We have $H_0 : \beta = \beta'$ and reduced model to be $Y = \beta' X + \epsilon$. Take $\max \mathcal{L}$ we would have $\hat{\sigma}^2 = \frac{1}{n}\sum (y_i - \beta' x_i)^2$. Then for full model, we have $\max \mathcal{L}$ which will give us $\hat{\beta} = \frac{\sum x_i y_\epsilon}{\sum x_i^2}$ with $\hat{\sigma}^2 = \frac{1}{n}\sum (y_i - \hat{\beta} x_i)^2$. Then we need to fill in λ, which is the test statistic of the likelihood-ratio.

Example 2.7.11. Consider testing whether the intercept β_0 statistically differs from zero. Test $H_0 : \beta_0 = 0$ versus $H_A : \beta_0 \neq 0$. We have full model: $Y = \beta_0 + \beta_1 x + \epsilon$ while reduced model is $Y = \beta_1 x + \epsilon$. Then we have

$$\begin{aligned}
\text{SSE}_F &= 21026 \\
\text{SSE}_R &= \sum (y_i - \hat{\beta}_1 x_i)^2 \text{ while } \hat{\beta}_1 = \frac{\sum x_i y_i}{\sum x_i^2} \\
&= 101678 \\
\text{Note: lm}(y \sim x - 1) & \\
f_{\text{calc}} &= [\frac{101678 - 21026}{(n-1)-(n-2)}]/[\frac{21026}{(n-2)}] \\
&= 19.18
\end{aligned}$$

2.8 Binary Predictor

Consider splitting the response values $y_1, ..., y_n$ into two groups with respective sample sizes n_1 and n_2. Define the dummy variable

$$x_i = \begin{cases} 1 & \text{if group one} \\ 0 & \text{if group two} \end{cases}$$

What will the estimated linear regression model be?

Theorem 2.8.1. *Consider simple linear regression model using independent variable defined as above. Then the least squares estimators are*

$$\hat{\beta}_1 = \bar{y}_1 - \bar{y}_2 \textit{ and } \hat{\beta}_0 = \bar{y}_2$$

where $\bar{y}_1$ and $\bar{y}_2$ are the respective sample means of each group.

What will the test statistic look like when testing β_1? Recall sample T-test. When testing the null hypothesis $H_0 : \mu_1 - \mu_2 = \triangle_0$, the test statistic is

$$t_{\text{calc}} = \frac{\bar{y}_1 - \bar{y}_2 - \triangle_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

where $\bar{y}$, s_1^2 and n_1 are respective sample averages, sample variance and sample size for group one and $\bar{y}_2$, s_2^2 and n_2 are the respective sample average, sample variance and sample size for group two. To compute p-values, we use the students T-distribution with degrees of freedom

$$\text{df} = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n)^2}{n_1-1} + \frac{(s^2/n_2)^2}{n_2-1}}$$

When the population variances are assumed to be equal for the two groups ($\sigma_1^2 = \sigma_2^2 = \sigma^2$), then the pooled test statistic is

$$t_{\text{calc}} = \frac{\bar{y}_1 - \bar{y}_2 - \triangle_0}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n-2}}}$$

where sample pooled variance is

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

To compute p-values, we use the students T-distribution with degrees of freedom

$$\mathrm{df} = n_1 + n_2 - 2$$

This leads us to the following theorem.

Theorem 2.8.2. *We have test statistics*

$$t_{calc} = \frac{\bar{y}_1 - \bar{y}_2 - \triangle_0}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}} = \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{\frac{MSE}{s_{xx}}}}$$

2.9 Prediction

In simple linear regression, there are two fundamental goals:

1. Test if there is a relationship between the response variable Y and covariate x. The first goal is accomplished by testing hypothesis $H_0 : \beta_1 = \beta_0$
2. Predict the response Y given a fixed value of x. This section describes predictions and confidence intervals on predictions.

Definition 2.9.1. Inferences concerning $E[Y_h] = \mu_Y$. The parameter of interest is $\theta = E[Y_h] = \mu_Y$.

Proposition 2.9.2. *Let*

$$\hat{Y}_h = \hat{\beta}_0 + \hat{\beta}_1 x_h$$

where x_h is some fixed value of x. Then

$$E[\hat{Y}_h] = \beta_0 + \beta_1 x_h, \leftarrow \textit{ unbiased}$$

and

$$var[\hat{h}] = \sigma^2 \left[\frac{1}{n} + \frac{(x_h - \bar{x})^2}{s_{xx}} \right] \sim SE(\hat{\theta})^2$$

From the above proposition, the standardized score of $\hat{Y}_h$ is

$$Z = \frac{\hat{\beta}_0 + \hat{\beta}_1 x_h - (\beta_0 + \beta_1 x_h)}{\sqrt{\sigma^2 \left(\frac{1}{n} + \frac{(x_h - \bar{x})^2}{s_{xx}} \right)}} \sim N(0, 1)$$

and since $\hat{Y}_h$ is a linear combination or response variable Y_i, the random variable Z has a standard normal distribution. The studentized of $\hat{Y}_h$ is

$$T = \frac{\hat{\beta}_0 + \hat{\beta}_1 x_h - (\beta_0 + \beta_1 x_h)}{\sqrt{\mathrm{MSE}\left(\frac{1}{n} + \frac{(x_h - \bar{x})^2}{s_{xx}} \right)^2}} \sim \mathrm{t}(\mathrm{df} = n - 2$$

Remark 2.9.3. Derivation of confidence interval for EY_h:

$$T = \frac{\hat{\beta}_0 + \hat{\beta}_1 x_h - (\beta_0 + \beta_1 x_n)}{\sqrt{\mathrm{MSE}(\frac{1}{n} + \frac{(x_h - \bar{x})^2}{s_{xx}}}} = \frac{\hat{\beta}_0 + \hat{\beta}_1 x_h - EY_h}{\hat{\mathrm{SE}}(Y_h)}$$

and then notice

$$P(-t_{\alpha/2, n-2} \leq T \leq t_{\alpha/2, n-2}) = 1 - \alpha$$

rearranage the above inequality and isolate the parameter, we get the confidence interval

$$\hat{\beta}_0 + \hat{\beta}_1 x_h \pm t_{\alpha/2, n-2} \hat{\mathrm{SE}}(\hat{Y}_h)$$

Proposition 2.9.4. *The expected value and variance of the prediction error are respectively given by*

$$E[Y_{h(new)} - (\hat{\beta}_0 + \hat{\beta}_1 x_h)] = 0$$

and

$$Var[Y_{h(new)} - (\hat{\beta}_0 + \hat{\beta}_1 x_h)] = \sigma^2[1 + \frac{1}{n} + \frac{(x_h - \bar{x})^2}{s_{xx}}]$$

Proof.

$$\begin{aligned} E[Y_h - (\hat{\beta}_0 + \hat{\beta}_1 x_h)] &= EY_h - E[\hat{\beta}_0 + \hat{\beta}_1 x_h] \\ &= \beta_0 + \beta_1 x_h - (\beta_0 + \beta_1 x_h) \\ &= 0 \\ \text{Var}[Y_h - (\hat{\beta}_0 + \hat{\beta}_1 x_h)] &= \text{Var}(Y_h) + \text{Var}(\hat{\beta}_0 + \hat{\beta}_1 x_h) \\ &= \sigma^2 + \sigma^2(\frac{1}{n} + \frac{x_h - \bar{x})^2}{s_{xx}}) \\ &= \sigma(1 + \frac{1}{n} + \frac{(x_h - \bar{x})^2}{s_{xx}}) \end{aligned}$$

which give us result. □

In a similar way the confidence interval for $E[Y_h]$, the studentized score of prediction error is given by

$$T = \frac{\text{Y}_{\text{h(new)}} - (\hat{\beta}_0 + \hat{\beta}_1 x_h - 0}{\sqrt{\text{MSE}(1 + \frac{1}{n} + \frac{(x_h - \bar{x})^2}{s_{xx}})}}$$

The $100(1-\alpha)\%$ prediction interval for a single future value of $Y_{\text{h(new)}}$ when $x = x_h$ is

$$(\hat{\beta}_0 + \hat{\beta}_1 x_h) \pm t_{\alpha/2,n-2}\sqrt{\text{MSE}\left(1 + \frac{1}{n} + \frac{(x_h - \bar{x})^2}{s_{xx}}\right)} \sim t(\text{df} = n - 2)$$

The appropriate proposition implies T has a student's t-distribution with $n-2$ degrees of freedom. Consequently, the confidence interval of interest follows.

Let us derive C.I. of for EY_h:

$$T = \frac{\hat{\beta}_0 + \hat{\beta}_1 x_h - \overbrace{(\beta_0 + \beta_1 x_n)}^{EY_h}}{\sqrt{\text{MSE}(\frac{1}{n} + \frac{(x_h - \bar{x})^2}{s_{xx}})}} = \frac{\hat{\beta}_0 + \hat{\beta}_1 x_h - EY_h}{\hat{\text{SE}}(Y_h)}$$

and then we have

$$P(-t_{\alpha/2,n-2} \le T \le t_{\alpha/2,n-2}) = 1 - \alpha$$

Rearrange the above inequality, isolate parameter, we have

$$P(\hat{\beta}_0 + \hat{\beta}_1 x_h - t_{\alpha/2,n-2}\hat{\text{SE}}(\hat{Y}_h) \le EY_h \le \hat{\beta}_0 + \hat{\beta}_1 x_h + t_{\alpha/2,n-2}\hat{\text{SE}}(\hat{Y}_h)$$

and hence we solve for

$$\hat{\beta}_0 + \hat{\beta}_1 x_h \pm t_{\alpha/2,n-2}\hat{\text{SE}}(\hat{Y}_h)$$

2.10 Linear Correlation

Let us introduce a few definition.

Definition 2.10.1. The covariance of random variables X and Y is define by

$$\text{Cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$$

where μ_X and μ_Y are the respective expected values of X and Y.

Note that $\text{Cov}(X, X) = \text{Var}(X)$.

Definition 2.10.2. The correlation of random variables X and Y is defined by

$$\rho = \text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

where σ_X and σ_Y are the respective standard deviations of X and Y. Note:

$$\rho = E[(\frac{x - \mu_x}{\sigma_x}\frac{y - \mu_y}{\sigma_y})] = E[Z_x Z_y]$$

Proposition 2.10.3. *This proposition states the following.*

1. *For $a, c > 0$ or $a, c < 0$, $Corr(aX + b, cY + d) = Corr(X, Y)$*
2. *$-1 \leq corr(X, Y) \leq 1$*
3. *If X and Y are independent, then $\rho = 0$.*
4. *$\rho = 1$ or $\rho = -1$ if and only if $Y = aX + b$ for some real numbers a, b with $a \neq 0$.*

Suppose for $(X_1, Y_1), ..., (X_n, Y_n)$ are random ordered pairs each coming from a bivariate normal distribution. Consequently, the conditional expectation and variance of Y given $X = x$ are

$$E[Y|X = x] = \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x - \mu_1) = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x - \mu_x)$$

and

$$\text{Var}[Y|X = x] = \sigma_2^2(1 - \rho^2) = \sigma_Y^2(1 - \rho^2)$$

and note

$$\text{Var}[Y|X = x] = \sigma_2^2(1 - \rho^2) = \sigma_Y^2$$

and one can derive the following

$$E[Y|X = x] = \mu_Y - \rho\frac{\sigma_Y}{\sigma_X}\mu_x + \rho\frac{\sigma_Y}{\sigma_X}x$$

Notice for the simple linear regression model,

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = r\frac{S_x}{S_x} \text{ and } \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$

where s_x and s_y are the sample standard deviations and r is the sample correlation between variables x and y.

Proof. Note

$$\begin{aligned} r\frac{S_y}{S_x} &= \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}\frac{\sqrt{\frac{S_{yy}}{n-1}}}{\sqrt{\frac{s_{xx}}{n-1}}} \\ &= \frac{s_{xy}}{s_{xx}} \end{aligned}$$

□

Assume the pairs (X_1, Y_1), ..., (X_n, Y_n) are random, the correlation coefficient as an estimator is given by

$$R = \frac{\sum_{i=1}^n X_i Y_i - \frac{1}{n}\bigg(\sum_{i=1}^n X_i\bigg)\bigg(\sum_{i=1}^n Y_i\bigg)}{\sqrt{(\sum X_i^2 - \frac{1}{n}(\sum X_i)^2)(\sum Y_i^2 - \frac{1}{n}(\sum Y_i)^2)}}$$

Note that if $H_0 \rho = 0$ is true, we have

$$T = \frac{R\sqrt{n-2}}{\sqrt{1-R^2}} \sim (\text{df} = n-2)$$

Consider testing hypothesis

$$H_0 : \rho = 0$$

under the null, the test statistic is

$$t_{\text{calc}} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

Note that $\beta_1 = \rho \frac{\sigma_y}{\sigma_x}$

2.11 Simultaneous Inferences

Let us motivate this subsection with the following.

1. Consider making inference with confidence level 95% of both true slope and the true intercept.
2. The difficulty is that these would not provide 95% confidence that the conclusions of both β_1 and β_0 and the true intercept β_0.
3. If the inferences were independent, the probability of both being correct would be $(0.95)^2 = 0.9025$.
4. The inferences are not independent.

Recall, in any hypothesis testing procedure,

$$P(\text{Type I error}) = \alpha$$

The family-wise error rate is defined as

$$P(\text{At least one type I error})$$

To compute family-wise error rate, consider running a pairwise procedure on β_1 and β_0. Then we have

$$\begin{aligned} P(\text{At least one type I error in 2 trials}) &= 1 - P(\text{no type I error in 2 trials}) \\ &= 1 - (1-\alpha)^2 \end{aligned}$$

and can be generated to $1-(1-\alpha)^K$. if there are K trials. Showing false significance in a testing procedure is a bad thing. Ideally, researchers want to control for making too many Type I errors. There have been many different procedures developed to control for the family-wise error rate.

Then

$$P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2)$$

By De Morgan's law

$$P(A_1^c \cap A_2^c) = P((A_1 \cup A_2)^c)$$

and the probability that both intervals are correct is

$$\begin{aligned} P(A_1^c \cap A_2^c) &= 1 - P(A_1 \cup A_2) \\ &= 1 - (P(A_1) + P(A_2) - P(A_1 \cap A_2)) \end{aligned}$$

Using the fact that $P(A_1 \cap A_2) \geq 0$, we obtain the Bonferroni inequality:

$$P(A_1^c \cap A_2^c) \geq 1 - P(A_1) - P(A_2)$$

for which we have setting

$$P(A_1^c \cap A_2^c) \geq 1 - (\alpha + \alpha) = 1 - 2\alpha$$

We can easily use Bonferroni inequality to obtain a family confidence coefficient of at least $1 - \alpha$ for estimating β_0 and β_1. We do this by estimating β_0 and β_1 separately with confidence levels of $1 - \alpha/2$. namely,

$$1 - \alpha/2 - \alpha/2 = 1 - \alpha$$

To find the critical value in two-tailed tests (or centered confidence intervals, we divide the significance by 2.

The $1 - \alpha$ family intervals for estimating β_0 and β_1 are

$$\hat{\beta}_0 \pm t\hat{\sigma}_{\hat{\beta}_0}$$

Example 2.11.1. Compute

$$\begin{aligned} t_{\frac{\alpha}{4},n-2} &= t_{0.05/4,5} = qt(1 - 0.0125, 5) = 3.16 \\ \beta_0 &:= 607 \pm 3.16(138.76) \\ &= (168.74, 1046.67) \\ \beta_1 &:= 23.01 \pm 3.16(2.19) \\ &= (18.09, 31.94) \end{aligned}$$

which we can use "confint(lm(y x), level = 1-0.05/2)" to construct Bonferroni confidence interval as well.

Extensions of the Bonferroni procedure

1. The Bonferroni procedure can also be applied for prediction
2. The critical value can be generalized

   $$t_{\alpha/2K,\text{df}}$$

   where K is the number of predictions (or intervals) and df is the degrees of freedom of the linear model. Use R command "predict(method, newdata=x.data, interval="confidence", level = 1 - 0.05/2)"

The critical value is larger than for a regular confidence interval for $\hat{y}$. Note the line is a t-value. The working-hotelling $100(1-\alpha)\%$ confidence band for the simple linear regression model has the following boundary values at any level x_h:

$$(\hat{\beta}_0 + \hat{\beta}_1 x_h) \pm W\sqrt{\text{MSE}\left(\frac{1}{n} + \frac{(x_h - \bar{x})^2}{S_{xx}}\right)}$$

Note that $t_v^2 = f_{1,v}$.

Remark 2.11.2. Scheife method. For predicting k new observations, $\hat{Y}_h \pm W \cdot \text{SE}$ where $W = \sqrt{k \cdot f_{1-\alpha,k,n-2}}$ and $\text{SE} = \sqrt{\text{MSE}\big(\frac{1}{n} + \frac{(x_h - \bar{x})^2}{s_{xx}}\big)}$

3 Multiple Regression I

Go back to Table of Contents. Please click TOC

Consider

$$Y = \beta_0 + \beta_1 x + \epsilon, x = \begin{cases} 1 & \text{control group} \\ 0 & \text{drug group} \end{cases}, \epsilon \overset{\text{i.i.d}}{\sim} N(0, \sigma^2)$$

What other variables should we include in the model?

1. Athletes tend to have a lower resting heart rate. Maybe we should include X_2 to be initial resting heart rate.
2. Age also influences resting heart rate. Introduce X_3 to be age.
3. Other variables... etc.

We can also extend model, meaning adding more x_i's variables.

3.1 Matrix Algebra

Note $\sum y_i = 1^T \underline{y}$ and $\bar{y} = \frac{1}{n} 1^T \underline{y}$

Definition 3.1.1. For a square matrix A, the inverse denoted A^{-1}, is a matrix that satisfies

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \text{ and } A^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

Please refer to linear algebra course for the rest of matrix algebra.

3.2 Random Vector and Matrix

Let $Y = (Y_1, ..., Y_n)^T$ be a random vector. Then we have expected value and covariance matrix of Y, respectively, $E(Y)$ and $\text{Var}(Y)$. The covariance matrix is also defined by

$$\text{Var}(Y) = E[[Y - E(Y)][Y - E(Y)]^T]$$

Answer. If $Y_1, ..., Y_n$ iid with variance σ^2, then we have

$$\begin{aligned} \text{Var}(Y) &= \text{Var} \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \\ &= \sigma^2 I \end{aligned}$$

□

3.3 Matrix Form of Multiple Linear Regression Model

Definition 3.3.1. Consider a data consisting of $p - 1$ covariates $X_1, ..., X_{p-1}$. Then the design matrix is defined by

$$X = (1_n, X_1, ..., X_{p-1}) = \begin{pmatrix} 1 & x_{11} & \dots & x_{1,p-1} \\ 1 & x_{12} & \dots & x_{2,p-1} \\ 1 & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \dots & x_{n,p-1} \end{pmatrix}$$

Answer. Regression Model. (Scaler Form)

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{p-1} x_{i,p-1} + \epsilon, i = 1, 2, ..., n, \epsilon_i \sim_{\text{iid}} N(0, \sigma^2) \\ \Rightarrow Y_1 &= \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_{p-1} x_{1,p-1} + \epsilon_1 \\ \ldots &= \ldots \\ Y_n &= \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_{p-1} x_{n,p-1} + \epsilon_n \end{aligned}$$

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \ldots & x_{1,p-1} \\ 1 & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \vdots & x_{n,p-1} \end{bmatrix} \cdot \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_n \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

which can be simplified as

$$\underline{Y} = \underline{X}\underline{\beta} + \underline{\epsilon}, \Sigma = \text{Var}(\underline{\epsilon}) = \sigma^2 \underline{I}, \underline{\epsilon} \sim \text{MN}(\underline{0}, \sigma^2 \underline{I})$$

□

3.4 Estimation of the Multiple Linear Regression Model

Recall simple linear regression, the least squares estimators are derived by minimizing

$$Q(b_0, b_1) = \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_i))^2$$

with respect to b_0 and b_1. We need an analogous criterion using matrix for the multiple regression model. First, define

$$Q(b_0, b_1, ..., b_{p-1}) = \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_{i1} + \cdots + b_{p-1} x_{i,p-1}))^2$$

and $b = (b_0, b_1, ..., b_{p-1})^T$. Then Q can be expressed

$$Q(b_0, ..., b_{p-1}) = Q(b) = (Y - Xb)^T (Y - Xb)$$

Proposition 3.4.1. *Let A be defined above then A is minimized when*

$$b = (X^T X)^{-1} X^T Y$$

Denote the minimum $\hat{\beta}$. Hence,

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

Further, the minimum value of Q is

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

Answer. We have

$$\begin{aligned}
Q(\underline{b}) &= (\underline{Y} - \underline{X}\underline{b})^T(\underline{Y} - \underline{X}\underline{b}) \\
&= Y^TY - Y^Xb - (Xb)^tY + (Xb)^T(Xb), \text{ ignore underline} \\
\frac{\partial Y^TXb}{\partial b} &= Y^TX \\
\frac{\partial b^TX^TY}{\partial b} &= (X^TY)^T \\
\frac{\partial b^T(X^TX)b}{\partial b} &= b^T(X^TX) + b^T(X^TX)^T = 2b^T(X^TX) \\
\frac{\partial Qb}{\partial b} &= 0 - 2Y^TX + 2b^T(X^TX) \overset{\text{Set}}{=} 0 \\
b^T(X^TX)^T &= Y^TX \\
(X^TX)b &= X^TY \\
b &= (X^TX)^{-1}X^TY
\end{aligned}$$

□

Remark 3.4.2. Recall $\frac{d}{db}YXb = YX$

Remark 3.4.3. Note

$$\begin{aligned}
\frac{\partial^2 Q(b)}{\partial b} &= \frac{\partial}{\partial b}\big[-2Y^TX + 2b^T(X^TX)\big] \\
&= 0 + 2\frac{\partial}{\partial b}b^T(X^TX) \\
&= 2(X^TX)^T \\
&= 2(X^TX)
\end{aligned}$$

Remark 3.4.4. Note that (X^TX) is positive definite (all eigen-values are greater than zero). Hence, $Q(b)$ achieves the minimum at $\hat{\beta} = (X^TX)^{-1}X^TY$

Remark 3.4.5. Let $a \in \mathbb{R}^{p-1}$ then $d = xa_{p-1}$. Then $d^Td = (Xa)^T(Xa)$

Example 3.4.6. THe State of Vermont is divided into 10 districts – they correspond roughly to counties. The following data represent the percentage of live births of babies weighing under 2500 grams (y), the fertility rate for females younger than 19 or older than 34 years of age (x_1), total high-risk fertility rate for females younger than 17 or older than 35 years of age (x_2), percentage of mothers with fewer than 12 years of education (x_3), percentage of births to unmarried mothers (x_4), and percentage of mothers not seeking medical care until the third trimester (x_5).

Answer.

$$\begin{aligned}
\frac{\partial^2 Q(\underline{b})}{\partial \underline{b}} &= \frac{\partial}{\partial \underline{b}}\big[-2\underline{Y}^T\underline{X} + 2\underline{b}^T(\underline{X}^T\underline{X})\big] \\
&= 0 + 2\frac{\partial}{\partial \underline{b}}[\underline{b}^T(\underline{X}^T\underline{X})] \\
&= 2(\underline{X}^T\underline{X})^T
\end{aligned}$$

The matrix is positive definite. Hence, $Q(\underline{b})$ achieves its minimum of the point at $\hat{\beta} = (\underline{X}^T\underline{X})^T\underline{X}^T\underline{Y}$ □

Remark 3.4.7. Why is $(\underline{X}^T\underline{X})^T$ positive definite?

Answer. Let $a \in \mathbb{R}^{p-1}$ for $a \neq 0$ and set $\underline{d} = \underline{Xa}$. Then $\underline{d}^T\underline{d} = (\underline{Xa})^T(\underline{Xa}) > 0$ □

Definition 3.4.8. A set of vectors $\{v_i\}$, each in $\mathbb{R}^n$, is linearly independent if

$$c_1v_1 + \cdots + c_pv_p = 0$$

has only the trivial solution, i.e. $c_1 = \cdots = c_p = 0$.

Example 3.4.9. Consider

$$v_1 = \begin{bmatrix} 3 \\ -2 \\ 7 \end{bmatrix}, v_2 = \begin{bmatrix} 8 \\ -16 \\ 3 \end{bmatrix}$$

and we have only one solution for $c_1v_1 + c_2v_2 = 0$ which is $c_1 = c_2 = 0$

Definition 3.4.10. A set of vectors $\{v_i\}$, each in $\mathbb{R}^n$, is linearly independent if there exists scalars c_i, not all zero, such that

$$c_1v_1 + \dots c_pv_p = 0$$

Example 3.4.11. Consider

$$v_1 = \begin{bmatrix} 3 \\ -2 \\ 7 \end{bmatrix}, v_2 = \begin{bmatrix} -6 \\ 4 \\ -14 \end{bmatrix}$$

Definition 3.4.12. The span of a set of vectors $\{v_i\}$, each in $\mathbb{R}^n$, is the collection of all vectors that can be written in the form

$$c_1v_1 + \dots c_pv_p$$

Note

1. span$\{v_1, ..., v_p\}$ is the set of all linear combinations of $v_1, ..., v_p$
2. say span of a subspace

Example 3.4.13. Consider

$$v_1 = \begin{bmatrix} 3 \\ -2 \\ 7 \end{bmatrix}, v_2 = \begin{bmatrix} -2 \\ 12 \\ 9 \end{bmatrix}, b = \begin{bmatrix} 8 \\ -16 \\ 5 \end{bmatrix}$$

and we have $2v_1 - v_2 = b$

Definition 3.4.14. The column space of a matrix A is the set $\mathcal{C}(A)$ of all linear combinations of the columns of A. If $A = [v_1, ..., v_n]$, then $\mathcal{C}(A) = \text{span}\{v_1, ..., v_n\}$. Then $\text{col}(A) = \mathcal{C}(A)$.

Example 3.4.15. Consider

$$A = \begin{bmatrix} 3 & -2 \\ -12 & 12 \\ 7 & 9 \end{bmatrix}, b = \begin{bmatrix} 8 \\ -16 \\ 3 \end{bmatrix}$$

and you can compute $b \in \mathcal{C}(A)$ or $b \in \text{col}(A)$.

Definition 3.4.16. The rank of a matrix A, denoted rank(A), is the number of linearly independent columns of A.

Example 3.4.17. A 3 by 3 matrix A can have rank of 2, because v_1 and v_2 are linearly independent, but $2v_1 - v_2 - v_3 = 0$.

Theorem 3.4.18. *The following statements are equivalent if A is a p by p matrix*

1. *A is invertible*
2. *$\mathcal{C}(A) = \mathbb{R}^p$*
3. *rank$(A) = p$*

Theorem 3.4.19. *If X is a n by p matrix with rank of p, we cannot proceed.*

Example 3.4.20. Consider a design matrix X given by

$$\vec{X} = \begin{bmatrix} 1 & 3.1 & 1 & 0 \\ 1 & 4.2 & 1 & 0 \\ 1 & 7.3 & 1 & 0 \\ 1 & 10.1 & 1 & 1 \\ 1 & 11.6 & 1 & 1 \\ 1 & 13.8 & 1 & 1 \end{bmatrix}$$

which means rank$(X) = 3$ since $\text{col}_1 = \text{col}_3 + \text{col}_4$. We can also check rank$(X^TX) = 3$.

Remark 3.4.21. There are ways around this. Generalized inverse $A^\dagger$, then we have $AA^\dagger A = A$.

Example 3.4.22. Single factor ANOVA, we have $Y_{ij} = \mu + \alpha_j + \epsilon_{ij}$ with $j = 1, 2, 3$.

3.5 Fitted Values and Residuals

Let vector of fitted values $\hat{Y}_i$ be denoted by $\hat{Y}$,

$$\hat{Y} = (\hat{Y}_1, \dots, \hat{Y}_n)^T$$

and in matrix form

$$\hat{Y} = X\hat{\beta} = X((X^TX)^{-1}X^TY) = (X(X^TX)^{-1}X^T)Y = HY$$

and we simply have $\hat{Y} = X\hat{\beta}$ in matrix form.

Definition 3.5.1. The hat matrix denoted by H is defined by

$$H = X(X^TX)^{-1}X^T$$

Note p is often used.

The hat matrix H shows that the fitted values $\hat{Y}_i$ are a linear combination of the response values Y_i.

$$\hat{Y} = HY \Rightarrow \hat{Y}_i = \sum_{i=1}^n h_{ij}Y_j$$

The hat matrix H plays an important role in regression diagnostics. Recall the studentized residuals

$$t_i = \frac{e_i}{\sqrt{\text{MSE}(1 - h_{ii})}}, \text{ where } h_{ii} = \text{inverse}$$

Note that the hat matrix H is symmetric

$$H^T = [X(X^TX)^{-1}X^T]^T = (X^T)^T((X^TX)^{-1})^TX^T = X(X^TX)^{-1}X^T = H$$

remember to show $(A^{-1})^T = (A^T)^{-1}$, $(AB)^T = B^TA^T$.

The hat matrix H is indempotent

$$\begin{aligned} H^2 &= HH \\ &= (X(X^TX)^{-1}X^T)(X(X^TX)^{-1}X^T) \\ &= H \end{aligned}$$

Recall the hat values for simple linear regression

$$H = X(X^TX)^{-1}X^T, X = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

and now we have

$$h_{ij} = \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x}}{s_{xx}}$$

Now we discuss residuals vector using hat matrix.

$$e = (e_1, ..., e_n)^T$$

and then

$$e = Y - \hat{Y} = Y - HY = (I_n - H)Y$$

Definition 3.5.2. The mean square error denoted MSE is defined by

$$\text{MSE} = \frac{\text{SSE}}{n-p} = \frac{\sum(y_i - \hat{y}_i)^2}{n-p}$$

Theorem 3.5.3. *Let Y be random variable vector with mean $E[Y] = \mu$ and covariance V or $[Y] = \Sigma$ and let A be a matrix of scalars. Then the random vector $W = AY$ has mean vector and covariance matrix*

$$E[W] = E[AY] = AE[Y] = A\mu$$

Theorem 3.5.4. *Let A be symmetric, then quadratic from A^TYA has expectation and variance*

$$E[Y^TAY] = 2\mathit{tr}(A\Sigma) + \mu^TA\mu$$

and

$$\mathit{var}[Y^TAY] = 2\mathit{tr}(A\Sigma A\Sigma) + 4\mu^TA\Sigma A\mu$$

where $\text{tr}(B)$ is the trace of the matrix B.

Theorem 3.5.5. *The least squares estimator of $\hat{\beta}$ is an unbiased estimator of parameter vector β.*

Answer. First note that $E[Y] = E[X\beta + \epsilon] = X\beta$. Then we derive

$$\begin{aligned} E[\hat{\beta}] &= E[(X^TX)^{-1}X^TY] \\ &= (X^TX)^{-1}X^TE[Y] \\ &= (X^TX)^{-1}X^TX\beta \\ &= I\beta = \beta \end{aligned}$$

□

Theorem 3.5.6. *The covariance matrix of least squares estimator $\hat{\beta}$ is*

$$Var[\hat{\beta}] = \sigma^2 (X^T X)^{-1}$$

Answer. First note $\text{Var}[Y] = \Sigma = \sigma^2 I$. Thus,

$$\begin{aligned}
\text{Var}[\hat{\beta}] &= \text{Var}[(X^T X)^{-1} X^T Y] \\
&= (X^T X)^{-1} X^T \text{Var}[Y]((X^T X)^{-1} X^T)^T \\
&= (X^T X)^{-1} X^T \sigma^2 I X
\end{aligned}$$

□

Theorem 3.5.7. *The mean square error MSE is an unbiased estimator of parameter σ^2*

Answer.

$$\begin{aligned}
E[\text{SSE}] &= E[Y^T (I - H) Y] \\
&= \text{trace}((I - H\sigma^2 I) + \underbrace{(X\beta)^T (I - H)(X\beta)}_{=0} \\
&= \sigma^2 \text{trace}(I - H) + 0 \\
&= \sigma^2 (n - p)
\end{aligned}$$

and note that

$$E[\frac{\text{SSE}}{n-p}] = \frac{1}{n-p} E[\text{SSE}] = \frac{\sigma^2(n-p)}{(n-p)} = \sigma^2$$

□

We conclude the following table:

	Estimate	Standard Error
Simple Linear Regression	$\hat{\beta}_1 = S_{xy}/S_{xx} = (S_{xx})^{-1} S_{xy}$	$\text{Var}(\hat{\beta}_1 = \sigma^2 (S_{xx})^{-1}$
Regression through Origin	$\hat{\beta} = \sum x_i y_i / \sum x_i^2 = (\sum x_i)^{-1} \sum x_i y_i$	$\text{var}(\hat{\beta}) = \sigma^2 (\sum x_i^2)^{-1}$
Multiple Regression	$\hat{\beta} = (X^T X)^{-1} X^T Y$	$\text{var}(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$

3.6 Non-linear Response Surfaces

Consider multiple linear regression model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{p-1} x_{i,p-1} + \epsilon_i$$

that is linear in the parameters.

Definition 3.6.1. The mean square error denoted MSE is defined by

$$\text{MSE} = \frac{\text{SSE}}{n-p} = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-p}$$

Note that

$$\hat{\sigma}^2_{\text{MSE}} = \frac{1}{n} \sum (y_i - \hat{y}_i)^2$$

3.7 Analysis of Variance for Multiple Linear Regression

Suppose we are interested in overall relationship between response variable and all covariates $x_1, ..., x_{p-1}$. To assess the overall relationship, we test the null alternative pair:

$$H_0 : \beta_1 = \cdots = \beta_{p-1} = 0, H_A : \text{at least one } \beta_0 \neq 0$$

The F-stat as a random variable is F = MSR/MSE. Under null, the appropriate proposition implies F has a F-distribution with respective degrees of freedom $df_1 = p-1$ and $df_2 = n-p$. The corresponding test statistic is

$$f_{\text{calc}} = \frac{\text{MSR}}{\text{MSE}}$$

F is a random variable and f_{calc} is a single realization of F based on the data set.

The analysis of variance table for linear regression, ANOVA, is

	Df	Sum Sq	Mean Sq	F-value	Pr(> F)
Regression	$p-1$	SSR	MSR = SSR/$(p-1)$	f_{calc} = MSR/MSE	P-value
Residuals	$n-p$	SSE	MSE = SSE/$(n-p)$		

where

$$\text{SSR} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2, \text{SSE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2, \text{SST} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

with important identities

1. $(p-1) + (n-p) = n-1$
2. SSR + SSE = SST

Let us develop overall F-test. Consider full model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1} x_{i,p-1} + \epsilon_i$$

with degrees of freedom $n-p$ and reduced model to be, under null, e.g. $H_0 : \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0$, to be

$$Y_i = \beta_0 + \epsilon_i$$

with degrees of freedom $n-1$. Then we can compute F-statistic

$$\begin{aligned} F_{\text{calc}} &= \frac{\text{SSE(E)} - \text{SSE(F)}}{n-1-(n-p)} \Big/ \frac{\text{SSE(F)}}{n-p} \\ &= \frac{\sum (Y_i - \bar{y})^2 - \sum (Y_i - \hat{y}_i)^2}{p-1} \Big/ \frac{\sum (Y_i - \hat{y}_i)^2}{n-p} \\ &= \frac{\text{SST} - \text{SSE}}{p-1} \Big/ \frac{\text{SSE}}{n-p} \\ &= \frac{\text{MSR}}{\text{MSE}} \end{aligned}$$

3.8 Coefficient of Multiple Determination

Let us introduce the definition.

Definition 3.8.1. The coefficient of multiple determination, denoted R^2 is defined by

$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}$$

and the interpretation is that we say there is R^2 percent of the variation in the response Y explained by the covariates $x_1, ..., x_{p-1}$.

1. Note R^2 is always between 0 and 1.
2. For simple linear regression, $r^2 = R^2$.
3. There is not a correlation coefficient r for multiple linear regression.
4. Every time a new variable is added to the model, the coefficient of multiple determination R^2 increases. It never decreases.
5. Every time a new variable is added to the model, SSE decreases. It never increases.
6. To adjust for R^2 always increasing, we can divide the sums of squares SSE and SST by their respective degrees of freedom. This leads to the adjusted coefficient of multiple determination.

Definition 3.8.2. The adjusted coefficient of multiple determination, denoted R_a^2 is defined

$$R_a^2 = 1 - \frac{\text{SSE}/(n-p)}{\text{SST}/(n-1)} = 1 - \left(\frac{n-1}{n-p}\frac{\text{SSE}}{\text{SST}}\right) = 1 - \left(\frac{n-1}{n-p}\right)(1-R^2)$$

Note that we have $\lim_{n\to\infty} R_a^2 = 1 - (1-R^2) = R^2$. If n is large relative to p, then we have $R_a^2 \sim R^2$.

3.9 Inference on the Slope Parameters

Ceteris parabius is a latin phrase meaning "with other things being equal or held constant". The above notion is key when interpreting and testing slope parameters in a multiple linear regression model. To further understand this, recall

$$E[Y_i] = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{p-1} x_{i,p-1}$$

then we have $\frac{\partial E[Y_i]}{\partial x_1} = \beta_1$ and $\frac{\partial E[Y_i]}{\partial x_j} = \beta_j$. For simple linear regression, we can compute t-test

$$\frac{\beta_1 \hat{-} 0}{\sqrt{\frac{\text{MSE}}{s_{xx}}}}$$

Wew ant to compute analogous results for multiple linear regression. Before continuing, consider some more results from probabiity theory,

$$\Sigma_{\hat{\beta}} = \text{Var}[\hat{\beta}] = \begin{pmatrix} \sigma^2_{\hat{\beta}_0} & \text{Cov}(\hat{\beta}_0, \hat{\beta}_1) & \dots & \text{Cov}(\hat{\beta}_0, \hat{\beta}_{p-1}) \\ \text{Cov}(\hat{\beta}_1, \hat{\beta}_0) & \sigma^2_{\hat{\beta}_1} & \dots & \text{Cov}(\hat{\beta}_1, \hat{\beta}_{p-1}) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(\hat{\beta}_{p-1}, \hat{\beta}_0) & \text{Cov}(\hat{\beta_{p-1}}, \hat{\beta}_1) & \dots & \sigma^2_{\hat{\beta}_{p-1}} \end{pmatrix}$$

We can estimate covariance matrix

$$\Sigma\hat{\beta} = \text{MSE}$$

Let us discuss linear transformation of β. Discuss motivation upfront. Consider full model $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$. Suppose we want to test (1) $H_0 : \beta_1 = 0$, (2) $H_0 : \beta_1 = \beta_2$, (3) $H_0 : \beta_1 = \beta_2 = 0$.

We can use, respectively for each case, (1) f-test or t-test, (2) f-test or t-test, and (3) f-test. Let us write all of them out in matrices.

Matrix form:

- $H_0 : c^T\beta{=}0$ using $\beta = [\beta_0, ..., \beta_{p-1}]^T$ where $c^T = [0, 1, 0]^T$.

- $H_0 : c^T\beta = 0$ where $c^T = [0, 1, -1]^T$ while the result is scalar; and $\sum c_i = 0$, e.g. a contrast.
- $H_0 : c^T\beta = 0$ where $c^T = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$ and in this case we are dealing with a vector.

Let us only look at cases 1 and 2. Define $\Psi = c^T\beta$ the parameter and $\hat{\Psi} = c^T\hat{\beta}$ the estimator. Then we have c^T to be a vector of known constants. Now we want to find expectation and variance so that we can find standard statistics and studentized statistics.

$$\begin{aligned} E[\hat{\Psi}] &= E[c^T\hat{\beta}] = c^T E[\hat{\beta}] \\ &= c^T\beta \\ &= \Psi \end{aligned}$$

and

$$\begin{aligned} \text{Var}[\hat{\Psi}] &= \text{Var}[c^T\hat{\beta}] = c^T\text{Var}(\hat{\beta})[c^T]^T \\ &= c^T\text{Var}(\hat{\beta})c^T \\ &= \sigma^2 c^T(X^TX)^{-1}c \end{aligned}$$

and thus

$$t_{\text{calc}} = \frac{\hat{\Psi} - \Psi_0}{\sqrt{c^T\hat{\text{Var}}\hat{Psi}}}$$

this will give us a $T \sim t(\text{df} = n - p)$.

For case 1, recall $H_0 : \beta_1 = 0$ with $c^T = [0, 1, 0]$. Then we have (i)

$$\begin{aligned} E[\hat{\Psi}] &= [0, 1, 0]\begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} = \beta_1 \\ &= [0, 1, 0]\begin{pmatrix} \text{var}(\hat{\beta}_0) & \text{Cov}(\hat{\beta}_0, \hat{\beta}_1) & \text{Cov}(\hat{\beta}_0, \hat{\beta}_1) \\ \text{cov}(\hat{\beta}_1, \hat{\beta}_0) & \text{Var}(\hat{\beta}_1) & \text{Cov}(\hat{\beta}_1, \hat{\beta}_2) \\ \text{Cov}(\hat{\beta}_2, \hat{\beta}_0) & \text{Cov}(\hat{\beta}_2, \hat{\beta}_1) & \text{Var}(\hat{\beta}_2) \end{pmatrix}\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \\ &= [0, 1, 0]\begin{bmatrix} \text{Cov}(\hat{\beta}_0, \hat{\beta}_1) \\ \text{Var}(\hat{\beta}_1) \\ \text{Cov}(\hat{\beta}_2, \hat{\beta}_1) \end{bmatrix} = \text{var}(\hat{\beta}_1) \end{aligned}$$

and we have $t_{\text{calc}} = \frac{\hat{\beta}_1}{\sqrt{\text{var}(\hat{\beta}_1}}$ when testing $H_0 : \beta_1 = 0$ (or any $\beta_j = 0$), we extract the corresponding event of $\hat{\sigma}^2(X^TX)^{-1}$ (switch $\hat{\sigma}$ for MSE). Note this is $T \sim t(\text{df} = n - p)$ so df is $n - 3$.

For case 2, we have $H_0 : \beta_1 = \beta_2$ or we write $\beta_1 - \beta_2 = 0$ while $c^T = [0, 1, -1]$. We have $E[\hat{\Psi}] = c^T[\beta_0, \beta_1, \beta_2]^T = \beta_1 - \beta_2$ for expectation and for variance, we have

$$\begin{aligned}
\text{Var}(\hat{\Psi}) &= c^T \text{Var}(\hat{\beta}) c \\
&= [0, 1, -1] \begin{bmatrix} dd \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ -1 \end{bmatrix} \\
&= [0, 1, -1] \begin{bmatrix} \text{Cov}(\hat{\beta}_0, \hat{\beta}_1) - \text{Cov}(\hat{\beta}_0, \hat{\beta}_2) \\ \text{Var}(\hat{\beta}_1) - \text{Cov}(\hat{\beta}_1, \hat{\beta}_2) \\ \text{Cov}(\hat{\beta}_2, \hat{\beta}_1) - \text{Var}(\hat{\beta}_2) \end{bmatrix} \\
&= 0 + 1(\text{Var}(\hat{\beta}_1) - \text{Cov}(\hat{\beta}_1, \hat{\beta}_2)) \\
&- 1(\text{Cov}(\hat{\beta}_2, \hat{\beta}_1) - \text{Var}(\hat{\beta}_2)) \\
&= \text{Var}(\hat{\beta}_1) + \text{var}(\hat{\beta}_2) - 2\text{Cov}(\hat{\beta}_1, \hat{\beta}_2)
\end{aligned}$$

and $t_{\text{calc}} = \frac{\hat{\beta}_1 - \hat{\beta}_2}{\sqrt{\text{Var}(\hat{\beta}_1 - \hat{\beta}_2)}}$ with $T \sim t(\text{df} = n - p = n - 3)$

For case 1, we have full model (write everything out) and with df = n-3. we have reduced model $Y = \beta_0 + \beta_2 X_2 + \epsilon$ with df = n-2. Then

$$\begin{aligned}
f_{\text{calc}} &= \frac{\text{SSE}_R - \text{SSE}_F}{(n-2) - (n-3)} / \frac{\text{SSE}_F}{n-3} \\
&= \frac{\text{SSE}_R - \text{SSE}_F}{1} / \frac{\text{SSE}_F}{n-3}
\end{aligned}$$

For case 2, we have null (..). Write full model (everything out..). Write reduced model with design matrix

$$\begin{bmatrix}
\beta_0 & X_1 & X_2 \\
1 & 1 & 0 \\
1 & 1 & 0 \\
\vdots & \vdots & \vdots \\
1 & 1 & 0 \\
1 & 0 & 1 \\
\vdots & \vdots & \vdots \\
1 & 0 & 1 \\
1 & 0 & 1
\end{bmatrix}$$

reduced model is $Y = \beta_0 + \beta_1(X_1 + X_2) + \epsilon$. We have f-test

$$f_{\text{calc}} = \frac{\text{SSE}_R - \text{SSE}_F}{(n-2) - (n-3)} / \frac{\text{SsE}_F}{n-3}$$

In R, we do $m.full <- lm(Y \sim X_1 + X_2)$ and $m.reduced <- lm(Y \sim I(X_1 + X_2))$ with $anova()$

Recall connection with simple linear regression estimators

$$\Sigma_{\hat{\beta}} = \text{MSE}(X^T X)^{-1} \begin{bmatrix} \frac{1}{n} + \frac{\bar{x}^2}{s_{xx}} & -\frac{\bar{x}}{s_{xx}} \\ -\frac{\bar{x}}{s_{xx}} & \frac{1}{s_{xx}} \end{bmatrix}$$

and moreover we have

$$\text{cov}(\hat{\beta}_0, \hat{\beta}_1) = -\bar{x}\text{var}(\hat{\beta}_1)$$

Suppose we are interested in the marginal relationship between response variable Y and the jth covariate x_j. The slope parameter of interest is β_j. To see if x_j is marginally significant, we test the null hypothesis $H_0 : \beta_j = \beta_{j0}$. Note the most common hypothesized value is zero. Also, the studentized score is

$$T = \frac{\hat{\beta}_j - \beta_{j0}}{\hat{\sigma}_{\hat{\beta}_j}}$$

where $\hat{\sigma}_{\hat{\beta}_j}$ is the estimated standard error of $\hat{\beta}_j$. The appropriate proposition implies T has a student's t-distribution with $n - p$ degrees of freedom.

The interpretation if β_j test statistically significant: the covariate x_j is statistically related to the response variable Y when holding all other covariates constant.

Hypothesis: $H_0 : \beta_2 = 0$ which is equivalent as $H_o : \psi = c^T\beta = 0$ and we have $c^T = [0, 0, 1, 0, 0, 0]$. We compute

$$t_{\text{calc}} = \frac{\hat{\beta}_2}{\sqrt{\hat{\sigma}^2_{\hat{\beta}_2}}} = \frac{\hat{\psi}}{\sqrt{\sigma^2\psi}}$$

which would be the t-statistic result in R output.

For marginal F-test for β_j, we have full model

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_j x_j + \cdots + \beta_{-1} x_{p-1} + \epsilon$$

and we have $H_0 : \beta_j = 0$ with degrees of freedom for (full) model to be $n - p$. Then we have reduced model

$$Y = \beta_0 + \beta_1 x_1 + \cdots + 0 + \cdots + \beta_{p-1} x_{p-1} + \epsilon$$

with degrees of freedom for (reduced) model to be $n - (p - 1)$. Then we have f-test

$$f_{\text{calc}} = \frac{\text{SSE}_R - \text{SSE}_F}{(n - (p - 1)) - (n - p)} / \frac{\text{SSE}_F}{n - p} = t^2_{\text{calc}}$$

In this case, we have $f \sim F(df_1 = 1, df_2 = n - p)$.

4 Diagnostics and Remedial Measures

Go back to Table of Contents. Please click TOC

Consider

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1} x_{i,p-1} + \epsilon_i$$

for $i = 1, 2, ..., n$ with $\epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$. What major assumptions are we having? The response function $E[Y]$ is linear. The errors ϵ are normally distributed. The errors have constant variance (homoscedasticity). The error ϵ are independent and identically distributed.

Definition 4.0.1. The ith residual is defined by

$$e_i = Y_i - \hat{Y}_i, i = 1, 2, ..., n$$

Note the summation of residuals not necessarily go to zero, e.g. $\sum \epsilon_i \neq 0$. Any analyzing the residuals provides insight on whether or not the regression assumptions are satisfied.

The sample mean and sample variance of the residuals are

$$e = \frac{1}{n}\sum_{i=1}^n e_i = 0, s_e^2 = \frac{1}{n-p}\sum_{i=1}^n e_i^2 = \text{MSE}$$

Although the errors ϵ_i are independent random variables, the residuals are not independent random variables. This can be seen by the following two properties:

$$\sum_{i=1}^n e_i = 0 \text{ and } \sum_{i=1}^n x_{ik} e_i = 0, k = 1, 2, ..., p-1$$

Note that we have $HX = X$, and $(I - H)X = X - X = 0$.

4.1 Residual Diagnostics

We want to standardize the residuals. With that said, let us introduce the following definition.

Definition 4.1.1. Let e_i be the residual defined above and let MSE be the mean square error defined in 3.15 from notes. Then the ith semistudentized residual is defined by

$$e_i^* = \frac{e_i - \bar{e}}{\sqrt{\text{MSE}}} = \frac{e_i}{\text{MSE}}, \text{ for } i = 1, 2, ..., n$$

Recall that the residual vector can be expressed as

$$e = (I_n - H)Y$$

Proposition 4.1.2. *The mean and variance of the residual vector e are respectively,*

$$E[e] = 0$$
$$Var[e] = \sigma^2 (I_n - H)$$

Proof. Note that

$$\begin{aligned} E[e] &= E[(I - h)Y] = (I - H)E[Y] \\ &= (I - H)X\beta = X\beta - HX\beta \\ &= X\beta - X\beta \\ &= 0 \end{aligned}$$

Next, we solve

$$\begin{aligned}\text{Var}[(I-H)Y] &= (I-H)\underbrace{\text{Var}[Y]}_{\sigma^2 I}(I-H)^T \\ &= \sigma^2(I-H)^2 \\ &= \sigma^2(I-H)\end{aligned}$$

□

Consequently, the Ith studentized residual is

$$t_i = \frac{e_i}{\sqrt{\text{MSE}(1-h_{ii})}}$$

where h_{ii} is the ith diagonal element of the hat matrix H.

Remark 4.1.3. Note that when $h_{ii} \equiv 1$, then t_i is large (because bottom of the fraction is getting smaller).

A useful refinement to make residuals more effective for detecting outlying Y observations is to measure the ith residual e_i when the fitted regression is based on all of the cases except the ith one. Denote $\hat{Y}_{(i)}$ the fitted regression equation based on all cases except the ith one. Denote $\hat{Y}_{i(i)}$ the fitted response value based on predicted model $\hat{Y}_{i(i)}$.

Consequently, the deleted residual denoted d_i is defined by

$$d_i = Y_i - \hat{Y}_{i(i)}$$

and note

$$\text{PRESS} = \sum_{i=1}^n (Y_i - \hat{Y}_{i(i)})^2 = \sum_{i=1}^n d_i^2$$

We want to studentize the residuals, i.e. we want to find an expression for

$$t_i = \frac{d_i}{\hat{\sigma}_{d_i}}$$

An algebraically equivalent expression for d_i, that does not require a recomputation of the fitted regression function omitting the ith case is

$$d_i = \frac{e_i}{1-h_{ii}}$$

Note that to compute the deleted residuals d_i, for each case, we do not need to fit a regression.

Define MSE as the mean square error based on all cases except the ith one. The following equation analogous to above relates MSE with the regular MSE.

$$(n-p)\text{MSE} = (n-p-1)\text{MSE} + \frac{e_i}{1-h_{ii}}$$

Using the above relation, the studentized deleted residuals can be expressed

$$t_i = \frac{d_i}{\hat{\sigma}_{d_i}} = e_i\sqrt{\frac{n-p-1}{\text{SSE}-(1-h_{ii})-e_i^2}}$$

Using the deleted studentized residuals in diagnostic plots is a common technique of validating the regression assumptions. The deleted studentized residuals are particularly useful in identifying outlying Y values.

Use the studentized or deleted studentized residuals to construct residual plots. Some recommendations follow:

1. Scatter plot matrix of all variables. Linearity, Constant Variance, General exploratory analysis
2. Plot of the studentized residuals against all (or some) of the predictor variables. Linearity, Constant Variance, Normality, Independence
3. Plot of the studentized residuals against fitted values. Same as the previous plot
4. Plot of the studentized residuals against time or other sequence. Independence, Normality
5. Plots of the studentized residuals against omitted predictor variables. Model validation.
6. Box plot (or histogram) of the studentized residuals. Normality
7. Normal probability plot (QQ Plot) of studentized residuals. Normality, Linearity

Example 4.1.4. Heavy tail: H; Short tail: S; Left skewed: convex; Right skewed: concave.

4.2 F Test for Lack of Fit

Although visually inspecting the residual plots gives insight on whether the regression assumptions have been satisfied, there are also formal testing procedures to check these claims. This section introduces an important testing procedure for determining whether a specific type of regression function adequately fits the data. For simple linear regression, the lack-of-fit hypothesis test procedure tests the question

$$\text{Is this linear function } E[Y] = \beta_0 + \beta_1 \text{ appropriate for this data}$$

Equivalently we want to test hypothesis

$$H_0 : E[Y] = \beta_0 + \beta_1 x \text{ or } H_A : E[Y] \neq \beta_0 + \beta_1 x$$

To construct a reasonable test statistic, we will use the general linear F-statistic. note the lack-of-fit test requires repeat observations of one or more x levels. Let n_j be the number of experiment units of the jth group.

$$\sum_{j=1}^{c} n_j = n$$

In order to construct the test statistics, we need to consider the following full and reduced models. We have full model

$$Y_{ij} = \mu_j + \epsilon_{ij}$$

with

$$\epsilon_{ij} \overset{\text{iid}}{\sim} N(0, \sigma^2)$$

Note: above, we are including more parameters than the simple linear regression model.

$$\mu_j, j = 1, 2, ..., e$$

Next, we have reduced model

$$Y_{ij} = \beta_0 + \beta_1 x_j + \epsilon_{ij}$$

with $\epsilon_{ij} \overset{\text{iid}}{\sim} N(0, \sigma^2)$. The least squares estimators of the full and reduced models are respectively. We realize that full model has $\hat{\mu}_j = \bar{y}_j$ and reduced model has $\hat{\beta}_0, \hat{\beta}_1$ same as bar. The residuals of the full and reduced models are respectively

$$e_{ij} = y_{ij} - \bar{y}_j$$

and

$$e_{ij} = y_{ij} - (\hat{\beta}_0 - \hat{\beta}_1 x_{ij})$$

The sum of squared residuals and degrees of freedom for the full model are respectively

$$\text{SSE}_F = \sum_j \sum_i (y_{ij} - \bar{y}_j)^2 = \text{SSPE}$$

and $df_F = \sum_{j=1}^{c}(n_j - 1) = n - c$. The sum of squared residuals and degrees of freedom for the reduced model are respectively

$$\text{SSE}_R = \sum_j \sum_i (y_{ij} - \hat{y}_j)^2 = \text{SSE}$$

and $df_R = n - 2$.

Applying the general lienar F-statistic, we get

$$\begin{aligned} f_{\text{calc}} &= \frac{\text{SSE}_R - \text{SSE}_F}{\text{df}_R - \text{df}_F} / \frac{\text{SSE}_F}{df_F} \\ &= \frac{\text{SSE} - \text{SSPE}}{(c-2)} / \frac{\text{SSPE}}{n-c} \\ F &\sim F(\text{df}_1 = c - 2, \text{df}_2 = n - c) \end{aligned}$$

Under null hypothesis, $H_0 : E[Y] = \beta_0 + \beta_1 x$, the F lack-of-fit test statistic is

$$F^* = \frac{(\text{SSE} - \text{SSPE})/(c-2)}{\text{SSPE}/(n-2)} = \frac{\text{MSLF}}{\text{MSPE}}$$

4.3 Remedial Measures

Fixing heteroscedasticity: transforming the response variable Y may remedy heteroscedasticity. If the error variance is not constant but changes in a systematic fashion, "weighted least squares" is an appropriate technique for modeling the data set.

Remark 4.3.1. Non-constant variance and non-normality often go hand and hand.

Fixing outliers: When outlying observations are present, use of the least squares estimators for the simple linear regression model may lead to serious distortions in the estimated regression function. When the outlying observations do not represent recording errors and should not be discarded, it may be desirable to use a procedure that places less emphasis on such outlying observations.

Estimate parameters using a robust loss function, e.e. minimize $Q(b) = \sum |y_i - \hat{y}_i|$. The loss function can be $f(x) = |x|$.

4.4 Robustness of the T-test

Definition 4.4.1. An inference procedure is robust if probability calculations (p-values or confidence intervals, e.g. standard errors) remain fairly accurate even when a condition is violated.

- The t-procedure is sensitive to outliers. Hence the t-procedure is not robust in the presence of outliers.
- The t-procedure is robust under violations of normality when there are no outliers.
- The t-procedure is not robust when the sample size is small.
- The t-procedure is robust when the sample size is large and there are no outliers.

Box-Cox transformation: it is often difficult to determine from residual diagnostic which transformation of Y is most appropriate for correcting violations of the regression model. The Box-Cox procedure automatically identifies a transformation from the family of power transformations on Y. Consider

$$Y_i^\lambda = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1}x_{i,p-1} + \epsilon_i, i = 1, 2, ..., n, \epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$$

Example 4.4.2. Consider

$$Y = \beta_0 + \beta_1 x + \epsilon$$

and we have $\mathbb{E}Y = \beta_0 + \beta_1 x$. We solve

$$\log(\mathbb{E}Y) = \beta_0^* + \beta_1^* x$$
$$\mathbb{E}Y = e^{\beta_0^* + \beta_1^* x}$$

For every single unit increase in x, $\mathbb{E}Y$ is multiplied by $e^{\beta_1^*}$, e.g.

$$e^{\beta_0^* + \beta_1^*(x+1)} = e^{\beta_0^* + \beta_1^* x} e^{\beta_1^*} = e^{\beta_1^*}\mathbb{E}Y$$

4.5 General Least Squares (Weighted Least Squares)

Often transforming the response variable Y will help in reducing or eliminating unequal variances of the error terms. Transforming Y may create an inappropriate regression relationship. Weighted least squares is a technique for modeling a data set when the error variance is not constant but changes in a systematic fashion. This maintains the original shape of the response function.

The generalized linear regression model is

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1}x_{i,p-1} + \epsilon_i$$

with

$$\epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma_i^2), i = 1, ..., n.$$

Consider using method of maximum likelihood estimation. The likelihood function follows

$$\mathcal{L}(\beta) = \prod_{i=1}^n \frac{1}{(2\pi\sigma_i^2)^{1/2}} \exp\left\{ - \frac{1}{2\sigma_i^2}(y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_{p-1}x_{i,p-1})^2 \right\}$$

Defining the reciprocal of the variance σ_i^2 as the weight w_i, $w_i = \frac{1}{\sigma_i^2}$. Notice that

$$\mathcal{L}(\beta) = \prod_{i=1}^n \left(\frac{w_i}{(2\pi\sigma_i^2)}\right)^{1/2} \exp\left\{ - \frac{1}{2}\sum_{i=1}^n w_i(y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_{p-1}x_{i,p-1})^2 \right\}$$

Thus we can estimate the weighted least squares model by maximizing $\mathcal{L}(\beta)$ or equivalently by minimize the objective function

$$Q_w(\beta) = \sum_{i=1}^n w_i(y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_{p-1}x_{i,p-1})^2$$

Let

$$W = \begin{pmatrix} w_1 & 0 \dots & 0 \\ 0 & w_2 & \dots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ 0 & \dots 0 & w_n \end{pmatrix}$$

Minimizing W_q with respect to b yields normal equations and least squares estimator

$$(X^TWX)\hat{\beta}_w = X^TWY$$

and

$$\hat{\beta}_w = (X^TWX)^{-1}X^TWY$$

The variance-covariance matrix of $\hat{\beta}_w$ is

$$\text{Var}(\hat{\beta}_w) = \sum_{\hat{\beta}_w}(X^TWX)^{-1}$$

In practice, the magnitudes of σ_i^2 often vary in a regular fashion with one or several predictor variables X_k. Examples include the megaphone shape when inspecting the residual plots. Notice that e_i^2 is an estimator of σ_i^2 when using unweighted least squares and $|e_i|$ is an estimator of $|\sqrt{\sigma_i^2}|$.

For iterative least squares, we fit the regression model using unweighted least squares. Regress the squared residuals e_i^2 against appropriate predictors. Or regress the absolute residuals $|e_i|$ against appropriate predictors. Use the estimated model computed from $e_i^2 \sim X_k$ as the variance function $\hat{v}_i$. Or use the estimated model computed from $|e_i| \sim X_k$ as the standard deviation function $\hat{s}_i$. The weights are then computed using

$$w_i = \frac{1}{\hat{v}_i} \text{ or } w_i = \frac{1}{\hat{s}_i^2}$$

5 Multiple Regression II

Go back to Table of Contents. Please click TOC

5.1 Extra Sums of Squares

An extra sum of squares measures the marginal reduction in the error sum of squares when one of several predictor variables are added to the regression model, given that other predictor variables are already in the model. Extra sum of squares measures the marginal increase in the regression sum of squares when one or several predictor variables are added to the regression model.

Consider the regression model $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$. Let x_1 be the extra variable when x_2 is already in the model

$$\text{SSR}(x_1|x_2) = \text{SSE}(x_2) - \text{SSE}(x_1, x_2)$$
$$\text{SSR}(x_1|x_2) = \text{SSR}(x_1, x_2) - \text{SSR}(x_2)$$

Note that let x_2 be the extra variable when x_1 is already in the model, we have

$$\text{SSR}(x_2|x_1) = \text{SSE}(x_1) - \text{SSE}(x_1, x_2)$$
$$\text{SSR}(x_2|x_1) = \text{SSR}(x_1, x_2) - \text{SSR}(x_1)$$

Let x_3 be the extra variable when x_1, x_2 are already in the model

$$\text{SSR}(x_3|x_1, x_2) = \text{SSE}(x_1, x_2) - \text{SSE}(x_1, x_2, x_3)$$
$$\text{SSR}(x_3|x_1, x_2) = \text{SSR}(x_1, x_2, x_3) - \text{SSR}(x_1, x_2)$$

A variety of decompositions exist.

$$\text{SSR}(x_1, x_2, x_3) = \text{SSR}(x_1) + \text{SSR}(x_2|x_1) + \text{SSR}(x_3|x_1, x_2)$$
$$\text{SSR}(x_1, x_2, x_3) = \text{SSR}(x_2) + \text{SSR}(x_3|x_2) + \text{SSR}(x_1|x_2, x_3)$$
$$\text{SSR}(x_1, x_2, x_3) = \text{SSR}(x_1) + \text{SSR}(x_2, x_3|x_1)$$

There are three types of sums of squares: Type I, Type II, and Type III. Let us introduce the following definition.

Definition 5.1.1. Type I sums of squares decomposes SSR by

$$\text{SSR}(x_1, x_2, ..., x_{p-1}) = \text{SSR}(x_1) + \text{SSR}(x_2|x_1) + \text{SSR}(x_3|x_1, x_2) + \cdots + \text{SSR}(x_{p-1}|x_1, ..., x_{p-2})$$

Note that Type I sums of squares is also called sequential sums of squares. The function anova() uses type I sums of squares. The type I sums of squares ANOVA table collapses to the standard table.

Definition 5.1.2. Type II sum of squares is relevant for factorial designs.

Definition 5.1.3. Type III sums of squares are used to test a single covariate after controlling for all other covariates.

5.2 Uses of Extra Sums of Squares in Tests for Regression Coefficients

Test whether a single $\beta_i = 0$. We assume hypothesis $H_0 : \beta_k = 0$ versus $H_A : \beta_k \neq 0$. We use

$$t_{\text{calc}} = \frac{\hat{\beta}_k}{\hat{\sigma}_{\hat{\beta}_k}}$$

for $t^2 = f$. In this case we have

$$\begin{aligned} f_{\text{calc}} &= \left(\frac{\text{SSE}_R - \text{SSE}_F}{\text{df}_R - \text{df}_F}\right) \div \frac{\text{SSE}_F}{\text{df}_F} \\ &= \frac{(\text{SSR}(X_k | X_1, \dots X_{k-1}, X_{K+1}, \dots, X_{p-1})/1}{(\text{SSE}(X_1, ..., X_{p-1})/(n-p)} \end{aligned}$$

Let us start by discussing the following scenarios of hypothesis testing.

Test whether some $\beta_k = 0$, we assume hypothesis,

$$\begin{aligned} H_0 : &\beta_q = \beta_{q+1} = \dots \beta_{p-1} = 0 \\ H_A : &\text{At least one } \beta_j \neq 0 \end{aligned}$$

and we can compute

$$f_{\text{calc}} = \frac{\text{SSR}(X_q, X_{q-1}, \dots, X_{p-1} | X_1, ..., X_{q-1})/(p-1)}{\text{SSE}(X_1, ..., X_{p-1})/(n-p)}$$

Test whether all $\beta_k = 0$, we write hypothesis

$$\begin{aligned} H_0 : &\beta_1 = \beta_2 = \dots = \beta_{p-1} = 0 \\ H_A : &\text{At least one } \beta_j \neq 0 \end{aligned}$$

$$f_{\text{calc}} = \frac{\text{SSR}(X_1, \dots, X_{p-1})/(p-1)}{\text{SSE}(X_1, ..., X_{p-1})/(n-p)}$$

Consider two predictors x_1, x_2. We have the following definition

Definition 5.2.1. The relative marginal reduction in the variation in Y associated with x_1 when x_2 is already in the model is

$$R^2_{Y_{1|2}} = \frac{\text{SSE}(x_2) - \text{SSE}(x_1, x_2)}{\text{SSE}(x_2)} = \frac{\text{SSR}(x_1 | x_2)}{\text{SSE}(x_2)}$$

The quantity is known as the coefficient of partial determination. The definition can be extended to more general cases.

$$\begin{aligned} R^2_{Y_{1|2,3}} &= \frac{\text{SSR}(x_1 | x_2, x_3)}{\text{SSE}(x_2, x_3)} \\ R^2_{Y_{2|1,3}} &= \frac{\text{SSR}(x_2 | x_1, x_3)}{\text{SSE}(x_1, x_3)} \\ R^2_{Y_{4|1,2,3}} &= \frac{\text{SSR}(x_4 | x_1, x_2, x_3)}{\text{SSE}(x_1, x_2, x_3)} \end{aligned}$$

Whether testing whether some $\beta_k = 0$, the general F statistic can be stated equivalently in terms of the coefficients of multiple determination for the full and reduced models. The formula follows

$$F = \frac{R^2_{Y|1 \dots p-1} - R^2_{Y|1 \dots q-1}}{p-q} \div \frac{1 - R^2_{Y|1 \dots p-1}}{n-p}$$

where $R^2_{Y|1\dots p-1}$ denotes the coefficient of multiple determination when Y is regressed on all x variables, and $R^2_{Y|1\dots q-1}$ denotes the coefficient when Y is regressed on $x_1, ..., x_{q-1}$ only.

Definition 5.2.2. The square root of a coefficient of partial determination is called a coefficient of partial correlation.

The coefficient of partial correlation is given the same sign as that of the corresponding regression coefficient in the fitted regression function.

5.3 Multicollinearity

Let us begin with a definition.

Definition 5.3.1. In statistics, multicollinearity (or collinearity) is a phenomenon in which two or more predictor variables in a multiple regression model are highly correlated.

Multicollinearity can cause two problems: (1) instability in the slope estimators, e.g. $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_{p-1}$, (2) instability in the standard errors (inflate SE significantly) $\hat{\sigma}_{\hat{\beta}_0}, ..., \hat{\sigma}_{\hat{\beta}_{p-1}}$.

Implications of perfectly correlated predictors. The perfect linear correlation between x_1 and x_2 did not inhibit our ability to obtain a good fit to the data. Since many different response functions $\hat{Y}$ provide the same good fit, we cannot interpret any one set of regression coefficients as reflecting the effects of the different predictor variables.

Proposition 5.3.2. *Another way of stating this problem is that there exist infinite number of models provide a perfected fit.*

Further inspection on uncorrelated predictors. Consider the following model, $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i12} + \epsilon_i$, $\epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$. Denote $\hat{Y}$ sample mean of the response variable Y. Denote s_Y the sample standard deviation of the response variable Y. Denote $\hat{X}_j$ sample mean of the covariate X_j, for $j = 1, 2$. Denote s_j sample standard deviation of the covariate X_j, for $j = 1, 2$. Denote r_{Y_j} sample correlation coefficient between Y and X_j and $r_{12} = r_{21}$ sample correlation coefficient between X_1 and X_2.

Consider the estimated model

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

with

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = (X^T X)^{-1} X^T$$

Equivalently, we can write

$$\begin{aligned} \hat{\beta}_1 &= \left(\frac{s_Y}{s_1}\right)\left(\frac{r_{Y_1} - r_{12} r_{Y_2}}{1 - r_{12}^2}\right) \\ \hat{\beta}_2 &= \left(\frac{s_Y}{s_2}\right)\left(\frac{r_{Y_2} - r_{12} r_{Y_1}}{1 - r_{12}^2}\right) \\ \hat{\beta}_0 &= \bar{Y} - \hat{\beta}_1 \bar{x}_1 - \hat{\beta}_2 \bar{x}_2 \end{aligned}$$

5.4 Higher Order Regression Models

Under the following circumstances shall we consider the use of polynomial models:

- When the true curvilinear response function is indeed a polynomial function

- When the true curvilinear response function is unknown but a polynomial function is an approximation to the true function

Consider one predictor-second order

$$Y_i = \beta_0 + \beta_1 t_i + \beta_2 t_i^2 + \epsilon_i, \epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2), t_i = X_i - \bar{X}$$

Note that the variables x and x^2 are often highly correlated. This induces multicollinearity which can cause computational difficulties when computing $(X^T X)^{-1}$ and instability in the parameter and standard error estimates. Centering a predictors can significantly reduce multicollinearity.

For one predictor-third order, one can consider

$$Y_i = \beta_0 + \beta_1 t_i + \beta_2 t_i^2 + \beta_3 t_i^3 + \epsilon_i, \epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2), t_i = X_i - \bar{X}$$

For one predictor-higher orders

$$Y_i = \sum_{K=0}^{p-1} \beta_K t_i^K + \epsilon_i, \epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$$

Example 5.4.1. How to approach the procedure of fitting such model? One can consider $Y_i = \beta_0 + \beta_1 t_i + \beta_2 t_i^2 + \beta_3 t_i^3 + \epsilon_i$. The Type I sum of squares, the procedure is to fit $EY = \beta_0 + \beta_t$ and compute SSR(t) and do F-test. Then for higher order we simply fit $EY = \beta_0 + \beta_1 t + \beta_2 t^2$ and compute SSR($t^2|t$) and do F-test.

5.5 Qualitative Predictors

Let us introduce the following definition.

Definition 5.5.1. A regression model with $p-1$ predictor variables contains additive effects if the response function can be written in the form

$$E[Y] = f_1(x_1) + f_2(x_2) + \cdots + f_{p-1}(x_{p-1})$$

where $f_1, ..., f_{p-1}$ can be any functions.

Additive example:

$$E[Y] = \underbrace{\beta_0 + \beta_1 X_1 + \beta_2 X_1^2}_{f_1(X_1)} + \underbrace{\beta_3 X_3}_{f_2(X_2)}$$

Non-additive example:

$$E[Y] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$$

Definition 5.5.2. If a regression model is not additive, it is said to contain an interaction effect.

Consider the regression model: $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \epsilon_i$. The response surface is: $E[Y] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$. The change in $E[Y]$ with a unit increase in x_1 when x_2 is held constant is

$$\frac{\partial EY}{\partial X_1} = \beta_1 + \beta_3 X_2 = \text{function of } X_2$$

and the change in $E[Y]$ with a unit increase in x_2 when x_1 is held constant is

$$\frac{\partial EY}{\partial X_2} = \beta_2 + \beta_3 X_1 = \text{function of } X_1$$

To model interactions between qualitative and quantitative predictors, consider the following model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \epsilon_i$$

In this case, we have X_1 to be a binary variable. For $X_1 = 1$, we have $EY = (\beta_0 + \beta_1) + (\beta_2 + \beta_3)X_2$. For $X_1 = 0$, we have $EY = \beta_0 + \beta_2 X_2$.

6 Multiple Regression III

Go back to Table of Contents. Please click TOC

6.1 Overview of the model building process

In this section, we present an overview of the model-building and model-validation process. A detailed description of variable selection for observational studies is presented in next subsection.

For data collection and types of studies, we consider the following. Data collection requirements for building a regression model vary with the nature of the study. This topic deserves more attention but is not a focus of this class. Consider four different types of studies.

- **Controlled experiments:** In a controlled experiment, the experimenter controls the levels of the explanatory variables and assigns treatment, consisting of a combination of levels of the explanatory variable to each experimental unit and observes the response.
- **Controlled experiments with covariates:** Statistical design of experiments uses supplemental information, such as characteristics of the experimental units, in designing the experiment so as to reduce the variance of the experimental error terms in the regression model. Sometimes, however, it is not possible to incorporate this supplemental information into the design of the experiment. Instead, it may be possible for the experimenter to incorporate this information into the regression model and thereby reduce the error variance by including uncontrolled variables or covariates in the model.
- **Confirmatory observational studies:** These studies, based on observational, not experimental, data, are intended to test (i.e. to confirm or not to confirm) hypotheses derived from previous studies or from hunches. For these studies, data are collected for explanatory variables that previous studies have shown to affect the response variable, as well as for the new variable or variables involved in the hypothesis.
- **Exploratory observational studies:** In the social, behavioral, and health sciences, management, and other fields, it is often not possible to conduct controlled experiments.

For an observational study, the key to establishing **causation** is to rule out the possibility of any confounding (or lurking) variables. We must establish that individuals differ only with respect to the explanatory variables. This is often very difficult and most times impossible for observational studies.

Controlled experiments. For single factor ANOVA, $Y_{ij} = \mu + \alpha_j + \epsilon_{ij}$ with $j = 1, ..., K$. Consider $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$. For two-way ANOVA, we have $Y_{ijk} = \mu + \alpha_j + \beta_K + (\alpha\beta)_{jk} + \epsilon_{ijk}$ with $j = 1, ..., J$ and $k = 1, ..., K$. For example, assuming $J = 2$ and $K = 2$, we have $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_2 X_{i1} X_{i2} + \epsilon_i$.

For controlled experiments with covariates, consider single factor ANOVA. Suppose we have $Y_{ij} = \mu + \alpha_j + \gamma c + \epsilon_{ij}$ for $j = 1, ..., K$. Suppose $K = 3$. Then we havea $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i$ as control. We can add interactions with the covariate or two-way ANCOVA.

Model building starts with some preliminary model investigation. First, we identify the functional forms in which the explanatory variables should be entered in the model. Next, identify important interactions that should be included in the model. Note that when incorporating interactions, we typically include both the interaction and the main effects.

In terms of reduction of explanatory variables, it is generally not important for controlled experiments. In studies of controlled experiments with covariates, some reduction of the covariates may take place because investigators often cannot be sure in advance that the selected covariates will be helpful in reducing the error variance. Generally, no reduction of explanatory variables should take place in confirmatory observational studies. The control variables were chosen on the basis of prior knowledge and should be retained for comparison with earlier studies. In exploratory observational studies, the number of explanatory variables that remain after the initial screening typically is still large. Explanatory variable reduction is extremely relevant for this type of study. For exploratory observational studies, we may have many co-linear variables which can cause instability in the slope and standard error estimates. For exploratory observational studies, we may have several good candidate models.

Model refinement and selection: At this stage in the model-building process, the tentative regression model, or the several good regression models in the case of exploratory observational studies, need to checked in detail for curvature and interaction effects. Residual plots are helpful in deciding whether one model is to be preferred over another.

Model validation: Model validity refers to the stability and reasonableness of the regression coefficients, the plausibility and usability of the regression function, and the ability to generalize inferences drawn from the regression analysis.

6.2 Variable Selection

All too often, investigators will screen a set of explanatory variables by fitting the regression model containing the entire set of potential X variables and then simply dropping those for which the t-statistic

$$t_{\text{calc}} = \frac{\hat{\beta}_j}{\hat{\sigma}_{\hat{\beta}_j}}$$

is insignificant. This procedure can lead to the dropping of important inter-correlated explanatory variables. A good search procedure should be able to handle important inter-correlated explanatory variables in such a way that not all of them will be dropped

How many model candidates exist? From any set of $p-1$ predictors, 2^{p-1} potential models can be constructed. Note that $p-1$ predictors implies p beta parameters. For the following criteria, assume $n > p$. Then we have

$$\text{SSE}_p = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

which is not a good measure. The sum of squares error using p beta parameters (or $p-1$ predictors) is denoted SSE_p. We desire a low SSE_p.

Next, we have R-sqare

$$R_p^2 = 1 - \frac{\text{SSE}_p}{\text{SST}}$$

The coefficient of multiple determination using p beta parameters (or $p-1$ predictors) is denoted R_p^2. We desire a high R_p^2. Note that SSE_p is equivalent to R_p^2.

Then we have adjusted R-square

$$R_{a,p}^2 = 1 - (\frac{n-1}{n-p})(1 - R_p^2) = 1 - (\frac{n-1}{n-p})\frac{\text{SSE}_p}{\text{SST}}$$

The adjusted coefficient of multiple determination using p beta parameters (or $p-1$ predictors) is denoted $R_{a,p}^2$. We desire a high $R_{a,p}^2$.

We also have C_p which is defined

$$C_p = \frac{\text{SSE}_p}{\text{MSE}(X_1, X_2, ..., X_{p-1})} - (n - 2p)$$

The Mallows' C_p criterion using p beta parameters (or $p-1$ predictors) is denoted C_p. We have $p-1$ potential predictors in the model. We desire a low C_p.

Moreover, we have AIC which is defined

$$\text{AIC}_p = n\log(\text{SSE}_p) - (n\log(n) - 2p)$$

Akaike's information criterion using p beta parameters (or $p-1$ predictors) is denoted AIC_p. We desire a low AIC_p.

Similar as AIC, we have BIC, which is defined

$$\text{BIC}_p = n\log(\text{SSE}_p) - (n\log(n) - \log(n)p)$$

The Bayesian information criterion using p beta parameters (or $p-1$ predictors) is denoted BIC_p. We desire a low BIC_p.

Last but not least, we want to recall PRESS_p, which is defined

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2,$$

while $\hat{y}_{i(i)}$ is deleted prediction. The prediction sum of squares criterion using p beta parameters (or $p-1$ predictors) is denoted PRESS_p. We desire a low PRESS_p. This is good for model validation.

Let us discuss true mean square error.

Definition 6.2.1. The mean square error of estimator $\hat{\theta}$ is

$$\text{MSE}(\hat{\theta}) = E[(\hat{\theta} - \theta)^2] = E[(\hat{\theta} - E(\hat{\theta}) + E(\hat{\theta}) - \theta)^2]$$

where θ is the parameter of interest. The mean square error can also be expressed as

$$\text{MSE}(\hat{\theta}) = \text{Var}(\hat{\theta}) + (E[\hat{\theta}] - \theta)^2$$

Not to be confused with $\text{MSE} = \frac{1}{n-p}\sum(Y_i - \hat{Y}_i)^2$. Mean square error uses squared loss, i.e. $E[f(\hat{\theta} - \theta)]$ and $f(u) = u^2$.

What is the relationship between mean square error and Mallows' C_p criterion. Let μ_i be the true average of Y_i. Then we have $E[\hat{Y}_i] - \mu_i$ is the bias of $\hat{Y}_i$ against the μ_i only for case i. The mean square error of our model at the ith case.

$$E[(\hat{Y}_i - \mu_i)^2] = \text{var}(\hat{Y}_i) + (E[\hat{Y}_i] - \mu_i)^2$$

The total mean square error of our regression

$$\sum_{i=1}^n \text{var}(\hat{Y}_i) + \sum_{i=1}^n (E[\hat{Y}_i] - \mu_i)^2$$

Define the true Mallows' criterion by

$$\Gamma_p = \frac{1}{\sigma^2}\big(\sum_{i=1}^n \text{var}(\hat{Y}_i) + \sum_{i=1}^n (E[\hat{Y}_i] - \mu_i)^2\big)$$

In this case, Γ_p is a standardized total mean square error. We need to estimate Γ_p, i.e. $\hat{\Gamma}_p = C_p$. Note the following

1. Note that $\sum_{i=1}^n \text{var}(\hat{Y}_i) = \sum_{i=1}^n \sigma^2 h_{ii} = \sigma^2 \sum_{i=1}^n h_{ii} = \sigma^2 p$
2. $E[\text{MSE}(X_1, ..., X_{p-1})] = \sigma^2$
3. $E[\text{SSE}_p] = (n-p)\sigma^2 + \sum(E(\hat{Y}_i) - \mu_i)^2$. Here notice that SSE_p is the quadratic form. Then we have $\text{SSE}_p - (n-p)\text{MSE}$ is an estimator of $\sum(E\hat{Y}_i - \mu_i)^2$ and $\text{MSE}(x_1, ..., x_{p-1})$ is an estimator of σ^2.

Definition 6.2.2. Let us introduce Mallow's C_p:

$$\begin{aligned} \hat{\Gamma}_p &= \frac{1}{\text{MSE}}[\text{SSE}_p - (n-p)\text{MSE} + \text{MSE}p] & (1) \\ &= \frac{\text{SSE}_p}{\text{MSE}} - (n-2p) & (2) \\ & & (3) \end{aligned}$$

Suppose that Y is generated from the true model

$$Y = X\beta + \epsilon, \epsilon \sim N(0, \sigma^2 \mathbb{I})$$

Consider a candidate model using p parameters

$$Y = X_1\beta_1^* + \epsilon, \epsilon \sim N(0, \sigma^2 I)$$

The estimated candidate model is

$$Y^* = H_1 Y = X_1(X_1^T X_1)^{-1} X_1^T Y$$

Note that H_1 is the projection matrix using only p parameters. The sum of squared error of the candidate model is

$$\text{SSE}_p = \text{SSE}_p(x_1, x_2, ..., x_{p-1}) = Y^T(I - H_1)^Y$$

Under the candidate model, the expected SSE_p is

$$\begin{aligned} E[\text{SSE}_p] &= \text{trace}\{(I - H_1)\sigma^2 I\} + (X\beta)^T(I - H_1)X\beta \\ &= \sigma^2(n-p) + (X\beta)^T(I - H_1)X\beta \end{aligned}$$

Notice that under the true model, the second term will vanish, i.e.

$$(X\beta)^T(I - H)X\beta = 0$$

Thus the second term of $E[\text{SSE}_p]$ represents the bias of the candidate model against the true model. Consequently, we can define Mallow's C_p criterion by

$$C_p = \frac{\text{SSE}_p}{\text{MSE}(x_1, x_2, ..., x_{P-1}} - (n-2p)$$

If our estimated candidate model is unbiased, then

$$\begin{aligned} E[C_p] &= \frac{E[\text{SSE}_p]}{E[\text{MSE}[x_1, x_2, ..., x_{P-1}]]} - (n-2p) \\ &= \frac{\sigma^2(n-p)}{\sigma^2} - (n-2p) \\ &= p \end{aligned}$$

Let us introduce AIC.

Suppose that $f(Y)$ is the true density of Y generated by

$$Y = X\beta + \epsilon, \epsilon \sim N(0, \sigma^2 I)$$

Suppose that $g(Y)$ is the density of the candidate model generated by

$$Y = X_1\beta_1^* + \epsilon, \epsilon \sim N(0, \sigma^2 I)$$

Note that $\theta = \begin{pmatrix}\beta_1 & \beta_2 & \dots & \beta_{p-1} & \sigma^2\end{pmatrix}^T$. Define the Kullback-Leibler (KL) divergence, by

$$\text{KL}(f, g) = \int \log\left\{\frac{f(Y)}{g(Y)}\right\} f(y)dy = \int \log\{f(y)\}f(y)dy - \int \log\{g(y)\}f(y)dy$$

Note that the KL divergence is analogous to a distance but is not a true distance because $\text{KL}(f, g) \neq \text{KL}(g, f)$.

The AIC criterion is based on estimating the Kullback-Leibler divergence. Note that

- The first term $\int \log\{f(y)\}f(y)dy$ is unknown but is also constant.
- The second term $\int \log\{g(y)\}f(y)dy$ is the expected negative log-likelihood of the data generated from f under g, i.e.,

  $$E_f[-l(\theta|y)]$$

  where

  $$l(\theta|y) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(Y - X_1\beta_1^*)^T(Y - X_1\beta_1^*)$$

- A natural estimator of the second term second term $\int \log\{g(y)\}f(y)dy$ comes from evaluating the negative log-likelihood at the maximum likelihood estimator, i.e.

  $$-l(\hat{\theta}_{\text{MLE}}|y) = \left(\frac{n}{2}\log(2\pi) + \frac{n}{2}\right) + \frac{n}{2}\log(\text{SSE}_p) - \frac{n}{2}\log(n) = \text{constant} + \frac{n}{2}\log(\text{SSE}_p) - \frac{n}{2}\log(n)$$

  Thus define the AIC criterion as

  $$\text{AIC}_p = 2p - (n\log(\text{SSE}_p) - n\log(n)) = n\log(\text{SSE}_p) - (n\log(n) - 2p)$$

- A smaller AIC value implies the KL divergence between f and g is also small, indicating that g is close to the truth. Typically we choose candidate models with a small AIC_p.

7 Multiple Regression IV

Go back to Table of Contents. Please click TOC

7.1 Further Diagnostics

Identifying outlying Y observations and deleted residuals: The deleted residual denoted d_i is defined by

$$d_i = Y_i - \hat{Y}_{i(i)}$$

The studentized deleted residuals can be expressed as

$$t_i = \frac{d_i}{\hat{\sigma}_{d_i}} = e_i \sqrt{\frac{n-p-1}{\text{SSE}(1-h_{ii}-e_i^2})}$$

Using the deleted studentized residuals in diagnostic plots is a common technique of validating the regression assumptions. The deleted studentized residuals are particularly useful in identifying outlying Y values.

Identifying outlying X observations: The hat matrix, as we saw, plays an important role in determining the magnitude of a studentized deleted residual and therefore in identifying outlying Y observations. The hat matrix is also helpful in directly identifying outlying X observations. In particular, the diagonal elements of the hat matrix are a useful indicator in a multivariate setting of whether or not a case is outlying with respect to its X values.

The diagonal element h_{ii} of the hat matrix have some useful properties. Namely,

$$0 \leq h_{ii} \leq 1 \text{ and } \sum_{i=1}^{n} h_{ii} = p$$

where p is the number of beta parameters.

Definition 7.1.1. The diagonal element h_{ii} of the hat matrix H is called the leverage of the ith case

For simple linear regression, let us derive the following. Consider $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$. Then we have

$$\begin{aligned}
\hat{\beta} &= \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = \begin{bmatrix} \bar{Y} - \frac{s_{xy}}{s_{xx}} \bar{X} \\ \frac{s_{xy}}{s_{xx}} \end{bmatrix} = (X^T X)^{-1} X^T Y \\
\hat{Y}_i &= \sum_{j=1}^{n} h_{ij} Y_j, \text{ while } h_{ij} = \frac{1}{n} + \frac{(x_i - \bar{X})(x_j - \bar{X})}{s_{xx}} \\
\hat{Y} &= HY
\end{aligned}$$

in which case we have leverage $h_{ii} = \frac{1}{n} + \frac{x_i - \bar{X})^2}{\sum_{l=1}^{n}(x_l - \bar{X}))}$.

Remark 7.1.2.

- The fitted value $\hat{Y}_i$ is a linear combination of the observed Y values and h_{ii} is the weight of observation $Y - i$ in determining this fitted value. Thus, the larger h_{ii} is, the more important Y_i is in determining $\hat{Y}_i$.
- h_{ii} is a function only of the X values, so h_{ii} measures the role of the X values in determining how important Y_i is in affecting the fitted value $\hat{Y}_i$.
- The larger h_{ii}, the smaller is the variance of the residual e_i.

Remark 7.1.3. Leverage use:

- A leverage value h_{ii} is usually considered to be large if it is more than twice as large as the mean leverage value, denoted h.
- Equivalently, leverage values greater than $2p/n$ are considered to be outlying with regards to their X value.
- Another guideline suggests that when h_{ii} exceeds 0.5, the leverage is very high, whereas those between 0.2 and 0.5 indicates moderate leverage.

Using the hat matrix to identify hidden extrapolation: To spot hidden extrapolations, we can utilize the direct leverage calculation for the new set of X values for which inferences are to be made:

$$h_{\text{new, new}} = X_{\text{new}}^T (X^T X)^{-1} X_{\text{new}}$$

where X_{new} is the vector containing the X values for which an inference about a mean response or a new observation is to be made, and X is the design matrix. If $h_{\text{new,new}}$ is well within the range of leverage values h_{ii} for the cases in the data set, no extrapolation is involved.

7.2 Identifying Influential Cases

After identifying cases that are outlying with respect to their Y values and or their X values, the next step is to determine whether or not these outlying cases a re influential.

Influence on single fitted value $\hat{Y}_i$: A useful measure of the influence that case i has on the fitted value $\hat{Y}_i$ is given by

$$\text{DFFITS}_i = \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\sqrt{\text{MSE}_{(i)} h_{ii}}}, i = 1, ..., n$$

Inference on all fitted values $\hat{Y}_i$: In contrast to the DFFITS, which considers the influence of the ith case on the fitted value $\hat{Y}_i$, Cook's distance considers the influence of the ith case on all n fitted values.

$$D_i = \frac{\sum_{j=1}^n (\hat{Y}_j - \hat{Y}_{j(i)})^2}{p\text{MSE}}, i = 1, ..., n$$

Influence on the regression coefficients $\hat{\beta}_k$: A useful measure of the influence that case i has on $\hat{\beta}_k$ is given by

$$\text{DFBETAS}_{k(i)} = \frac{\hat{\beta}_k - \hat{\beta}_{k(i)}}{\sqrt{\text{MSE}_{(i)} c_{kk}}}, k = 0, 1, ..., p-1$$

How ith case infer slope $\hat{\beta}_k$? We note that $\text{var}(\hat{\beta} = \sigma^2 (X^T X^{-1}$ and $\text{var}(\hat{Y}) = \sigma^2 H = \sigma^2 X (X^T X)^{-1} X^{-1}$.

7.3 Remedial Measures

Below are some remedial measures for specific violations of the multiple regression model.

Remedy for non-constant variance:

- Try to transform the response variable Y, e.g. $\log(Y)$;
- weighted least squares.

Remedy for multicolinearity:

- Center collinear covariates, $t_{ik} = x_{ik} - \bar{x}$;
- Principle Component Analysis (PCA). Note that you may lose interpretability of the regression coefficients;
- Ridge Regression;
- Other.

Remedy for influential observations:

- Try to transform the response variable Y;
- Fit a smooth function;
- Robust regression: (1) iteratively reweighted least squares (IRLS) robust regression; (2) In general, robust regression requires minimizing the objective function $Q(\beta)$ with respect to β using some robust loss function ψ, i.e. minimize

$$Q(\beta) = \sum_{i=1}^{n} \psi(y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1} x_{i,p-1}))$$

- Least Absolute Deviations (LAD). Here $\psi(u) = |u|$.

Remedy for non-normality of the errors:

- Try to transform the response variable Y;
- Nonparametric techniques: (1) Bootstrap the estimated coefficients $\hat{\beta}_k$; (2) Other;

7.4 Ridge Regression

The main idea is to modify the method of least squares to allow biased estimators of the regression coefficients, which which induces a reduction in the standard errors.

Bias Variance Tradeoff: Recall from inference topics, the mean square error of an estimator $\hat{\theta}$ is defined as

$$\begin{aligned}\text{MSE}(\theta) &= E[(\hat{\theta} - \theta)^2] \\ &= \text{var}[\hat{\theta}] + (E[\hat{\theta}] - \theta)^2 \\ &= \text{var}[\hat{\theta}] + (\text{bias})^2\end{aligned}$$

In multiple regression setting, consider least squares estimator $\hat{\beta}_k$. Then we $\hat{\beta}_k$ is an unbiased estimator of β_k. Then

$$\begin{aligned}\text{MSE}(\hat{\beta}_k) &= \text{var}(\hat{\beta}_k) + (E(\hat{\beta}_k) - \beta_k)^2 \\ &= \text{var}(\hat{\beta}_k) + (\beta_k - \beta_k)^2 \\ &= \text{var}(\hat{\beta}_k)\end{aligned}$$

Key ideas:

- the least squares estimators $\hat{\beta}_k$, $k = 1, 2, ..., p-1$ are the Best Linear Unbiased Estimators (BLUE). This result comes from the Gauss-Markov theorem.
- When predictors are highly correlated, the variance of $\hat{\beta}_k$ becomes inflated. The estimators are still unbiased but they may have a large variance.

- Let $\hat{\beta}_k^R$ be some estimator of coefficient β_k. Then

$$\begin{aligned}\text{MSE}(\hat{\beta}_k^R) &= \text{var}(\hat{\beta}_K^R) + (E(\hat{\beta}_k^R) - \beta_k)^2 \\ &= (E[(\hat{\beta})^2] - (E\hat{\beta}_k^R)^2) + (E(\hat{\beta}_k^R) - \beta_k)^2\end{aligned}$$

  Thus, we can control the variance by constrain the term $E[(\hat{\beta}_k^R)^2]$.

- When an estimator has only a small bias and is substantially more precise than an unbiased estimator, it may well be preferred since it will have a larger probability of being close to the true parameter value.
- Modify the method of least squares to allow biased estimators of the regression coefficients, which induces a reduction in the standard errors.

Penalized least squares: The idea is to implement a penalty (or constraint) on the least squares objective function. This method is motivated by the Lagrange Multiplier optimization technique. In our case we want to minimize

$$Q(\beta) = (Y - X\beta)^T(Y - X\beta) = \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1}x_{i,p-1})^2 \text{ subject to } \sum_{j=1}^{p-1} \beta_j^2 \le c$$

with respect to $\beta = \begin{pmatrix}\beta_0 & \dots \beta_{p-1}\end{pmatrix}^2$. The above optimization problem is equivalent to minimizing

$$Q^R(\beta) = \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1}x_{i,p-1}))^2 + c\sum_{j=1}^{p-1} \beta_j^2$$

with respect to $\beta = \begin{pmatrix}\beta_0 & \beta_1 \dots & \beta_{p-1}\end{pmatrix}^T$.

8 Logistic Regression

Go back to Table of Contents. Please click TOC

Recall the simple linear regression model:

$$Y = \beta_0 + \beta_1 x + \epsilon = \mathbb{E}[Y] + \epsilon$$

with

$$\epsilon \overset{\text{iid}}{\sim} N(0, \sigma^2)$$

and note that in this model the response variable Y is assumed to be a quantitative continuous variable. The covariate x can be quantitative continuous variable. The covariate x can be quantitative or categorical (dummy variable).

The distributional assumptions placed on the errors (ϵ) from the basis of all inferential procedures related to the simple linear regression model. More specifically, normality of the errors allows us to construct confidence intervals for $E[Y_h]$ and run testing procedures on the slope parameter β_1.

An equivalent form of above model is given by

$$Y \overset{\text{iid}}{\sim} N(\beta_0 + \beta_1 x, \sigma^2)$$

From model statements above, notice that the mean of the response variable is equal to a deterministic function. That is,

$$\mathbb{E}[Y] = \beta_0 + \beta_1 x$$

A statistical model can be specified by the response variable's mean and some distributional assumption.

Motivation of the logistic regression model: How do we define a simple (one covariate) regression model that allows for a categorical (binary) response variable? To answer this question, first recall the Bernoulli random variable:

Definition 8.0.1. Any random variable whose possible values are 0 and 1 is called a Bernoulli random variable. That is, $Y \sim \text{Bern}(p)$ or the following form

$$\mathbb{P}(Y = y) = p^y (1-p)^{1-y}, \text{ for } y = 0, 1$$

Also recall that the expected value (or true mean) of a Bernoulli random variable is its success probability. That is, if $Y \sim \text{Bern}(p)$, then $E[Y] = p$.

The answer to the above question: Regress a sigmoidal function $p = f(x)$ on covariate x. Note: A sigmoidal function has an s shape and is bounded between 0 and 1 ($0 < f(x) < 1$). The success probability of a Bernoulli random variable is bounded between 0 and 1 ($0 < p < 1$).

8.1 The Probit Mean Response Function

Consider a health researcher studying the effect of a mother's use of alcohol (X is an index of degree of alcohol use during pregnancy) on the duration of her pregnancy (Y^c). Here we use the superscript c to emphasize that the response variable, pregnancy duration, is a continuous response. This can be represented by a simple linear regression model:

$$Y_i^c = \beta_0^c + \beta_1^c x_i + \epsilon_i^c, \text{ for } \epsilon_i^c \overset{\text{iid}}{\sim} N(0, \sigma^2)$$

and we will assume that ϵ_i^c is normally distributed with mean zero and variance σ^2. If the continuous response variable, pregnancy duration, were available, we might proceed with the usual simple linear regression analysis. However, in this instance, researchers coded each pregnancy duration as preterm or full term using the following rule:

$$Y_i = \begin{cases} 1 & \text{if } Y_i^c \leq 38 \text{ weeks} \\ 0 & \text{if } Y_i^c > 38 \text{ weeks} \end{cases}$$

It follows that

$$\begin{aligned} P(Y_i = 1) = p_i &= P(Y_i^c \leq 38) \\ &= P(p_0^c + \beta_1^c x_i + \epsilon_i^c \leq 38) \\ &= P(\epsilon_i^c \leq 38 - p_0^c - p_1^c x_i) \\ &= P(\frac{\epsilon_i^c - 0}{\sigma} \leq (\frac{38 - \beta_0^c}{\sigma}) + (-\frac{\beta_1^c}{\sigma}) x_i) \\ &= P(z = \frac{\epsilon_i^c}{\sigma} \leq \beta_0 + \beta_1 x_i) \\ &= \Phi(\beta_0^* + \beta_1^* x_i) \end{aligned}$$

The nonlinear regression function known as the probit mean response function is

$$E[Y_i] = p_i = \Phi(\beta_0^* + \beta_1^* x_i),$$

where $\Phi(z)$ is the cdf of the standard normal distribution. The inverse function (Φ^{-1}) of the standard normal cumulative distribution is sometimes called the probit transformation. Hence, the probit response is

$$\Phi^{-1}(p_i) = \beta_0^* + \beta_1^* x_i$$

8.2 The Logistic Mean Response Function

The logistic mean response function is defined by

$$EY_i = p_i = F_L(\beta_0 + \beta_1 x_i) = \frac{e^{(\beta_0 + \beta_1 x)}}{1 + e(\beta_0 + \beta_1 x)} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

The density of a logistic random variable ϵ_L is

$$f_L(\epsilon_L) = \frac{\exp(\epsilon_L)}{(1 + \exp(\epsilon_L))^2}, \epsilon_L \in \mathbb{R}$$

Note that the above density has sd $\pi/sqrt3$. Consider the linear model $Y_i^c = \beta_0^c + \beta_1^c x_i + \epsilon_i^c$, where $\epsilon_i^c \sim \text{logistic}(\text{mean} = 0, \text{sd} = \sigma_c)$. Then we have

$$\begin{aligned} P(Y_i = 1) = p_i &= P(\frac{\epsilon_i^c}{\sigma_c} \leq \beta_0^* + \beta_1^* x_i) \\ &= P(\frac{\pi}{\sqrt{3}} \frac{\epsilon_i^c}{\sigma_c} \leq \frac{\pi}{\sqrt{3}} \beta_0^* + \frac{\pi}{\sqrt{3}} \beta_i^* x_i) \\ &= P(\epsilon_L \leq \beta_0 + \beta_1 x_i) \\ &= F_L(\beta_0 + \beta_1 x_i) \end{aligned}$$

The logistic statistical model: Let $Y_1 Y_2, ..., Y_n$ be independently distributed Bernoulli random variables with respective success probabilities $p_1, p_2, ..., p_n$. Then the logistic regression model is

$$E[Y_i] = p_i = F_L(\beta_0 + \beta_1 x_i) = \frac{e^{(\beta_0 + \beta_1 x_i)}}{1 + e^{(\beta_0 + \beta_1 x_i)}}, i = 1, 2, ..., n$$

The estimated logistic model: we have

$$\hat{p}_i = \frac{e^{(\hat{\beta}_0+\hat{\beta}_1 x_i)}}{1+e^{(\hat{\beta}_0+\hat{\beta}_1 x_i)}}, i = 1, 2, ..., n$$

Note that the quantities $\hat{\beta}_0$ and $\hat{\beta}_1$ are the estimated intercept and slope. Maximum likelihood estimation is a common technique for estimating the logistic model parameters β_1 and β_2, which requires numerical methods. Since the response variable Y is Bernoulli, the Y_i's consist of zeros and ones.

Example 8.2.1. Set up the likelihood function:

$$\begin{aligned}
L(\beta_0, \beta_1|y) &= L(p_1, ..., p_n|y) = \prod_{i=1}^{y} p_i^{y_i}(1-p_i)^{1-y_i} \\
l(\beta_0, \beta_1|y) &= l(p_1, ..., p_n)|y) = \sum_{i=1}^{n} y_i \log p_i + \sum_{i=1}^{n}(1-y_i)\log(1-p_i) \\
&= \sum_{i=1}^{n} y_i \log\big(\frac{p_i}{1-p_i}\big) + \sum_{i=1}^{n}\log(1-p_i) \\
&= \sum_{i=1}^{n} y_i(\beta_0+\beta_1 x_i) + \sum_{i=1}^{n}\log(1+\exp(\beta_0+\beta_1 x_i))
\end{aligned}$$

Odds and Log-Odds: Rearranging equation $E[Y_i]$ gives us

$$\frac{p_i}{1-p_i} = e^{\beta_0+\beta_1 x_i}, i = 1, 2, ..., n$$

The equation above relates the odds of event $\{Y = 1\}$ occurring to a deterministic exponential function. Taking the natural log of both sides gives

$$F_L^{-1}(p_i) = \log\big(\frac{p_i}{1-p_i}\big) = \beta_0 + \beta_1 x_i, i = 1, 2, ..., n$$

The equation above relates the log-odds of event $\{Y = 1\}$ occuring to a deterministic linear function. The inverse function (F_L^{-1}) of the logistic cumulative distribution is called the logit transformation.

Interpretation of the slope parameter β_1: Consider a unit increase in the covariate. The odds of event $\{Y = 1\}$ occurring when the covariate is fixed at x is:

$$\text{odds}_1 = \frac{p_1}{1-p_1} = e^{\beta_0+\beta_1(x)}$$

The odds of event $\{Y = 1\}$ occurring when the covariate is fixed at $x + 1$ is:

$$\text{odds}_2 = \frac{p_2}{1-p_2} = e^{\beta_0+\beta_1(x+1)}$$

Thus,

$$\text{odds ratio} = \frac{\text{odds}_2}{\text{odds}_2} = \frac{e^{\beta_0+\beta_1(x+1)}}{e^{\beta_0+\beta_1 x}} = e^{\beta_1}$$

Equivalently, we have

$$\text{odds}_2 = e^{\beta_1}(\text{odds}_1)$$

Two equivalent interpretations of $\hat{\beta}_1$:

- The odds ratio of event $\{Y = 1\}$ occurring for covariate fixed at $x + 1$ verses covariate fixed at x is equal to $e^{\hat{\beta}_1}$.
- The odds of event $\{Y = 1\}$ occurring are multiplied by $e^{\hat{\beta}_1}$ for every one unit increase in x.

Inference on the regression parameter β_1: Consider testing the null hypothesis: $H_0 : \beta_1 = 0$. The statistic used for this hypothesis test is

$$z_{\text{calc}} = \frac{\hat{\beta}}{\sigma_{\hat{\beta}_1}} = \frac{\hat{\beta}_1}{\hat{\sigma}_{\hat{\beta}_1}}$$

The rejection rules and p-value computations are the same as any z-test. Note that if $H_0 : \beta_1$ is true, then $\Theta = e^{\beta_1} = e^0 = 1$.

Let $Y_1, ..., Y_n$ be independently distributed Bernoulli random variables with respective success probabilities $p_1, ..., p_n$. Also suppose that $x_1, ..., x_n$ is a sequence of 0's and 1's (the covariate is a binary variable). For this situation, the simple logistic regression model is

$$E[Y_i] = p_i = \frac{e^{(\beta_0+\beta_1 x_i)}}{1 + e(\beta_0 + \beta_1 x_i)}, i = 1, 2, ..., n$$

where

$$x_i = \begin{cases} 1 & \text{if the } i^{\text{th}} \text{ case is in the exposed group} \\ 0 & \text{if the } i^{\text{th}} \text{ case is in the unexposed group} \end{cases}$$

Note in this model, both x and y values consist of 0's and 1's. A data set of this nature can be organized with a two-by-two table.

8.3 Multiple Logistic Regression

Let $Y_1, ..., Y_n$ be independently distributed Bernoulli random variables with respective success probabilities $p_1, ..., p_n$. Then the multiple logistic regression model is

$$E[Y_i] = p_i = \frac{\exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1} x_{i,p-1})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1} x_{i,p-1})}$$

The model can be expressed using matrix notation

$$E[Y_i] = p_i = \frac{\exp(X_i^T \beta)}{1 + \exp(X_i^T \beta)}, i = 1, 2, ..., n,$$

where

$$\beta^T = \begin{pmatrix}\beta_0 & \beta_1 \dots \beta_{p-1}\end{pmatrix}, X_i^T = \begin{pmatrix}1 & x_{i1} \dots x_{i,p-1}\end{pmatrix}$$

Equation: the multiple logistic regression model is estimated by the method of maximum likelihood. The log-likelihood function is

$$l(\beta, Y) = \sum_{i=1}^{n} y_i (X_i^T \beta) - \sum_{i=1}^{n} \log[1 + \exp(X_i^T \beta)]$$

Note: Many ideas from regular regression translate to logistic regression.

Inference: We consider the large sample testing procedure. Let G denote the matrix of second-order partial derivatives of the log-likelihood function, the derivatives being taken with respect to the parameters β.

$$G_{p\times p} = [g_{ij}], i, j = 0, 1, ..., p - 1$$
$$g_{00} = \frac{\partial^2 l}{\partial \beta_0^2}, g_{01} = \frac{\partial^2 l}{\partial \beta_0 \partial \beta_1}$$

The matrix G is called the Hessian. Typically it is denoted with **H** but we reserved this letter for the projection matrix. When the second-order partial derivatives in the hessian matrix are evaluated at the maximum likelihood estimates, the estimated approximate variance-covariance matrix of the estimated regression coefficients is

$$\Sigma_{\hat{\beta}} = ([-g_{ij}])^{-1}$$

Thus to test the null hypothesis $H_0 : \beta_k = \beta_{k0}$, we use z-statistic

$$z_{\text{calc}} = \frac{\hat{\beta}_k}{\hat{\sigma}_k}$$

where $\hat{\sigma}_k$ is the square root of the corresponding diagonal element in the estimated variance covariance matrix. This is a z-test not a t-test. The asymptotic approach can also be used for regular linear regression.

Aymptotic approximate distribution of ML estimators: Consider estimating parameter θ using the method of maximum likelihood. Under some standard regularity assumptions, we know the maximum likelihood estimator $\hat{\theta}$ is approximately distributed as

$$\hat{\theta} \sim N(\theta, I^{-1}(\theta))$$

where $\mathcal{I}(\theta)$ is the Fisher information defined by

$$\mathcal{I}(\theta) = -E\bigg[\frac{\partial^2}{\partial\theta^2}\log f(y;\theta)\bigg]$$

Note:

$$\begin{aligned}
\theta^T &= \begin{bmatrix}\theta_1 & \dots & \theta_p\end{bmatrix} \\
\hat{\theta} &= \sim N(\theta, \mathcal{I}^{-1}(\theta)) \\
G = \mathcal{I}(\theta) &= \{-E\big[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log f(y;\theta)\big]\}
\end{aligned}$$

Notice that ML estimator $\hat{\theta}$ has approximated variance $\mathcal{I}^{-1}(\hat{\theta})$.

Not logistic: Consider $Y_i = \beta x_i + \epsilon$, where $\epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$. The goal is to find $\mathcal{I}(\beta)$. Then we compute

$$\begin{aligned}
l(\beta; y) = \log L(\beta; y) &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum(y_i - \beta x_i)^2 \\
\frac{\partial l}{\partial \beta} &= 0 - \frac{2}{2\sigma^2}\sum(y_i - \beta x_i)(-x_i) = \frac{1}{\sigma^2}\sum y_i x_i - \frac{\beta}{\sigma^2}\sum x_i^2 \\
\frac{\partial^2 l}{\partial \beta^2} &= \frac{-\sum x_i^2}{\sigma^2}, -E\big[\frac{\partial^2 l}{\partial \beta^2}\big] = \frac{\sigma x_i^2}{\sigma^2}, \mathcal{I}^{-1}(\beta) = \frac{\sigma^2}{\sum x_i^2}
\end{aligned}$$

Consider testing the null alternative pair

$$H_0 : \beta_q = \beta_{q+1} = \cdots = \beta_{p-1} = 0, H_A : \text{ At least one } \beta_j \neq 0$$

We want to formulate a testing procedure analogous to the general linear F-approach for a logistic setting. The likelihood ratio test statistic denoted G^2 is

$$G^2 = -2\log\big[\frac{L(R)}{L(F)}\big] = -2[\log L(R) - \log L(F)],$$

where $L(R)$ is the likelihood function under the reduced model and $L(F)$ is the likelihood function under the full model. The quantity G^2 is known as the deviance. The large sample theory states G^2 is distributed approximately chi-squared with $p-q$ degrees of freedom,

$$G^2 \sim \chi^(df = p - q)$$

8.4 Inference about Mean Parameter

Denote the vector of the levels of the X variable for which p is to be estimated by X_h:

$$X_h = \begin{pmatrix} 1 & x_{h1} & \dots & x_{h,p-1} \end{pmatrix}$$

and the mean response of interest by p_h:

$$p_h = \frac{\exp(X_h^T \beta)}{1 + \exp(X_h^T \beta)}$$

Denote the point estaimtor of p_h by $\hat{p}_h$ and compute this quantity by

$$\hat{p}_h = \frac{\exp(X_h^T \hat{\beta})}{1 + \exp(X_h^T \hat{\beta})}$$

The estimated approximate variance of $X_h^T \hat{\beta}$ is

$$s_h^2 = X_h^T \hat{\Sigma}_{\hat{\beta}} X_h$$

where $\hat{\Sigma}_{\hat{\beta}}$ is defined as the variance-covariance matrix.

The $(1-\alpha)100\%$ confidence interval for $E[Y_h] = p_h$: The formula for the $(1-\alpha)100\%$ confidence interval for $E[Y_h] = p_h$ is (L^*, U^*), where

$$L^* = \frac{\exp(L)}{1 + \exp(L)}, U^* = \frac{\exp(U)}{1 + \exp(U)}$$

and

$$L = X_h^T \hat{\beta} - z_{\alpha/2} s_h, U = X_h^T \hat{\beta} + z_{\alpha/2} s_h$$

Index

References

[1] Kutner, Michael, "Applied Linear Statistical Models", 4th Edition.