Regression Analysis

Last Updated:

October 13, 2023

Contents

Chapter 1

Introduction

The primary goal in regression is to a devlop a model that relates a set of explanatory variables $X_1, \ldots, X_p$ to a response variable Y, then test the model and use it for inference and predicition.

Given a set of n pairs of data Y_i and X_i, we attempt to fit a straight line to these points, using a simple regression model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

Where ϵ_i represents an unobserved random error term, β_0 is the intercept and β_1 is the slope of the line. β_0 and β_1 are parameters that need to be estimated from observed data. The model can also be expressed in terms of $(X_i - \bar{X})$.

$$Y_i = (\beta_0 + \beta_1 \bar{X}) + \beta_1 (X_i - \bar{X}) + \epsilon_i$$

Where $\bar{X}$ is the sample mean

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

This proposed model is linear in the parameters β_0, β_1, and would still be referred to as linear if we had X_i^2 instead of X_i. This model also makes the assumption that the random error terms ϵ_i are uncorrelated, have mean 0, and variance σ^2. Under these assumptions, we have

$$E(Y_i) = \beta_0 + \beta_1 X_i$$

$$\text{Var}(Y_i) = \sigma^2$$

Chapter 2

Simple Linear Regression

The primary goal in regression is to a devlop a model that relates a set of explanatory variables $X_1, \ldots, X_p$ to a response variable Y, then test the model and use it for inference and predicition.

Given a set of n pairs of data Y_i and X_i, we attempt to fit a straight line to these points, using a simple regression model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

Where ϵ_i represents an unobserved random error term, β_0 is the intercept and β_1 is the slope of the line. β_0 and β_1 are parameters that need to be estimated from observed data. The model can also be expressed in terms of $(X_i - \bar{X})$.

$$Y_i = (\beta_0 + \beta_1 \bar{X}) + \beta_1 (X_i - \bar{X}) + \epsilon_i$$

Where $\bar{X}$ is the sample mean

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

This proposed model is linear in the parameters β_0, β_1, and would still be referred to as linear if we had X_i^2 instead of X_i. This model also makes the assumption that the random error terms ϵ_i are uncorrelated, have mean 0, and variance σ^2. Under these assumptions, we have

$$E(Y_i) = \beta_0 + \beta_1 X_i$$

$$\text{Var}(Y_i) = \text{Var}(\beta_0 + \beta_1 X_i + \epsilon_i) = \sigma^2$$

Thus the mean of Y is a linear function of X however the variance of Y does not depend on a value of X.

The parameters β_0 and β_1 are called the regression coefficients. The slope β_1 is the change in the mean of the distribution of Y produced by a unit change in Y. If the range of data on X includes x = 0, then the intercept β_0 is the mean of the distribution of the response Y when x = 0. If the range of x does not include zero, then β_0 has no practical interpretation.

2.1 Estimating the Parameters with the Method of Least Squares

The parameters β_0, β_1 are unknown and must be estimated from the data.Suppose we have n pairs of data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.

2.1.1 Estimation of β_0 and β_1

The method of least squares is the most popular approach to fitting a regression model. Let Q be the sum of the error terms squared

$$Q = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_i X_i)^2$$

Then we want to minimize Q with respect to the parameters β_1, β_2,

$$\frac{\partial Q}{\partial \beta_0} = -2 \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i) = 0$$

$$\frac{\partial Q}{\partial \beta_1} = -2\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)X_i = 0$$

We can rearrange these equations to get the following equations

$$\begin{aligned}
&-2\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i) = 0 \\
\implies &\sum_{i=1}^{n} Y_i - \sum_{i=1}^{n}\beta_0 - \sum_{i=1}^{n}\beta_1 X_i = 0 \\
\implies &\sum_{i=1}^{n} Y_i = n\beta_0 - \beta_1 \sum_{i=1}^{n} X_i \\
&-2\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)X_i = 0 \\
\implies &\sum_{i=1}^{n} Y_i X_i - \sum_{i=1}^{n}\beta_0 X_i - \sum_{i=1}^{n}\beta_1 X_i^2 = 0 \\
\implies &\sum_{i=1}^{n} Y_i X_i = \beta_0 \sum_{i=1}^{n} X_i - \beta_1 \sum_{i=1}^{n} X_i^2
\end{aligned}$$

These 2 equations are known as the normal equations and the solutions to them, call them b_0, b_1, are

$$b_0 = \bar{Y} - b_1\bar{X}$$

$$b_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})Y_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \sum_{i=1}^{n} k_i Y_i$$

with

$$k_i = \frac{X_i - \bar{X}}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

We also sometimes use a more compact notation, by denoting the corrected sum of squares for X and the sum of cross products of X_i Y_i as

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

$$S_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right) = \sum_{i=1}^{n} y_i(x_i - \bar{x})$$

So, we can write

$$b_1 = \frac{S_{xy}}{S_{xx}}$$

The observed difference between Y_i and the corresponding fitted value $\hat{Y}_i$ is a residual. The ith residual is

$$e_i = y_i - \hat{y}_i = y_i - (b_0 + b_1 x_i)$$

Note that k_i has important properites, such as

$$\sum_{i=1}^{n} k_i = 0,\ \sum_{i=1}^{n} k_i X_i = 1,\ \sum_{i=1}^{n} k_i^2 = \frac{1}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$\begin{aligned}
\sum_{i=1}^{n} k_i &= \frac{\sum_{i=1}^{n}(X_i - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \\
&= \frac{n\bar{X} - n\bar{X}}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = 0
\end{aligned}$$

$$\begin{aligned}\sum_{i=1}^{n} k_i X_i &= \frac{\sum_{i=1}^{n}(X_i - \bar{X})X_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \\ &= \frac{\sum_{i=1}^{n}(X_i^2 - X_i\bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \\ &= \frac{\sum_{i=1}^{n} X_i^2 - \bar{X}\sum_{i=1}^{n} X_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \\ &= \frac{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2}{\sum_{i=1}^{n}(X_i^2 - 2X_i\bar{X} + \bar{X}^2)} \\ &= \frac{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2}{\sum_{i=1}^{n} X_i^2 - 2n\bar{X}^2 + n\bar{X}^2} = 1\end{aligned}$$

$$\begin{aligned}\sum_{i=1}^{n} k_i^2 &= \sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right)^2 \\ &= \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^4} \\ &= \frac{1}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\end{aligned}$$

The equation for the fitted line is then

$$\hat{Y} = b_0 + b_1 X$$

Or alternatively using $X - \bar{X}$,

$$\hat{Y} = (b_0 + b_1\bar{X}) + b_1(X - \bar{X})$$

Theorem 2.1.1 (Gauss Markov Theorem)**.** *The least square estimators b_0, b_1 are unbiased and have minimum variance among all unbiased linear estimators.*

Proof. Consider an unbiased linear estimator

$$\hat{\beta}_1 = \sum_{i=1}^{n} c_i Y_i$$

$\hat{\beta}_1$ must satisfy $E(\hat{\beta}_1) = \beta_1$.

$$\begin{aligned}\beta_1 &= E(\hat{\beta}_1) \\ &= E\left(\sum_{i=1}^{n} c_i Y_i\right) \\ &= \sum_{i=1}^{n} c_i E(Y_i) \\ &= \sum_{i=1}^{n} c_i(\beta_0 + \beta_1 X_i) \\ &= \beta_0 \sum_{i=1}^{n} c_i + \beta_1 \sum_{i=1}^{n} c_i X_i\end{aligned}$$

Therefore, $\sum_{i=1}^{n} c_i = 0$, and $\sum_{i=1}^{n} c_i X_i = 1$. We can also see that the variance is

$$\mathrm{Var}(\hat{\beta}_1) = \sum_{i=1}^{n} c_i^2 \,\mathrm{Var}(Y_i) = \sigma^2 \sum_{i=1}^{n} c_i^2$$

Now, set $c_i = k_i + d_i$ where k_i is as defined previously above and d_i are arbitrary constants. We want to show

that the variance is minimized, so

$$\begin{aligned}\operatorname{Var}(\hat{\beta}_1) &= \sum_{i=1}^{n} c_i^2 \operatorname{Var}(Y_i) \\ &= \sigma^2 \sum_{i=1}^{n} c_i^2 \\ &= \sigma^2 \sum_{i=1}^{n} (k_i + d_i)^2 \\ &= \sigma^2 \left(\sum_{i=1}^{n} k_i^2 + 2 \sum_{i=1}^{n} k_i d_i + \sum_{i=1}^{n} d_i^2 \right)\end{aligned}$$

Note that the variance of b_1 is

$$\operatorname{Var}(b_1) = \operatorname{Var}\left(\sum_{i=1}^{n} k_i Y_i \right) = \sigma^2 \sum_{i=1}^{n} k_i^2 = \sigma^2 \frac{1}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

Now notice that there is a relationship between the variance of $\hat{\beta}_1$ and b_1, namely that the variance of $\hat{\beta}_1$ is the same as b_1 plus an additonal constants but these constants are indeed 0.

$$\begin{aligned}\sum_{i=1}^{n} k_i d_i &= \sum_{i=1}^{n} k_i (c_i - k_i) \\ &= \sum_{i=1}^{n} k_i c_i - \sum_{i=1}^{n} k_i^2 \\ &= \sum_{i=1}^{n} c_i \frac{X_i - \bar{X}}{\sum_{i=1}^{n} (X_i - \bar{X})^2} - \frac{1}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \\ &= \frac{\sum_{i=1}^{n} c_i X_i - \sum_{i=1}^{n} c_i \bar{X}}{\sum_{i=1}^{n} (X_i - \bar{X})^2} - \frac{1}{\sum_{i=1}^{n} (X_i - \bar{X})^2}\end{aligned}$$

We know that $\sum_{i=1}^{n} c_i = 0$ and $\sum_{i=1}^{n} c_i X_i = 1$, so this becomes

$$\sum_{i=1}^{n} k_i d_i = \frac{1 - 0}{\sum_{i=1}^{n} (X_i - \bar{X})^2} - \frac{1}{\sum_{i=1}^{n} (X_i - \bar{X})^2} = 0$$

Therefore,

$$\operatorname{Var}(\hat{\beta}_1) = \sigma^2 \left(\sum_{i=1}^{n} k_i^2 + \sum_{i=1}^{n} d_i^2 \right)$$

Clearly the variance is minimized when $d_i = 0$ for all i, thus

$$\operatorname{Var}(\hat{\beta}_1) = \sigma^2 \sum_{i=1}^{n} k_i^2 = \operatorname{Var}(b_1)$$

Thus the least squares estimator b_1 has minimum variance along all unbiased estimators. □

We may write

$$\hat{Y} = b_0 + b_1 X$$

for the estimated or fitted line, and

$$e_i = Y_i - \hat{Y}_i$$

for the estimated i^{th} residual. The estimate for the variance σ^2 is then

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n - 2}$$

The estimate of the variance σ^2 is also known as the mean square error (MSE).

2.1.2 Properties of Fitted Regression Line

(i) $\sum_{i=1}^{n} e_i = 0$. Recall that $\hat{Y} = b_0 + b_1 X = (b_0 + b_1\bar{X}) + b_1(X - \bar{X})$, and

$$\bar{Y} = b_0 + b_1\bar{X}$$

So $\hat{Y} = \bar{Y} + b_1(X - \bar{X})$, then

$$\begin{aligned}
\sum_{i=1}^{n} e_i &= \sum_{i=1}^{n}(Y_i - \hat{Y}_i) \\
&= \sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} \hat{Y}_i \\
&= \sum_{i=1}^{n} Y_i - \sum_{i=1}^{n}(\bar{Y} + b_1(X_i - \bar{X})) \\
&= n\bar{Y} - n\bar{Y} + b_1 \sum_{i=1}^{n}(X_1 - \bar{X}) \\
&= n\bar{Y} - n\bar{Y} + b_1(n\bar{X} - n\bar{X}) = 0
\end{aligned}$$

(ii) $\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \hat{Y}_i$. This follows from the previous property since

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} \hat{Y}_i = 0 \implies \sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \hat{Y}_i$$

(iii) $\sum_{i=1}^{n} X_i e_i = 0$. This can be shown from the definition

$$\begin{aligned}
\sum_{i=1}^{n} X_i e_i &= \sum_{i=1}^{n} X_i(Y_i - \hat{Y}_i) \\
&= \sum_{i=1}^{n} X_i(Y_i - b_0 - b_1 X_i) \\
&= \sum_{i=1}^{n} X_i Y_i - b_0 \sum_{i=1}^{n} X_i - b_1 \sum_{i=1}^{n} X_i^2 \\
&= b_0 \sum_{i=1}^{n} X_i + b_1 \sum_{i=1}^{n} X_i - b_0 \sum_{i=1}^{n} X_i - b_1 \sum_{i=1}^{n} X_i \\
&= 0
\end{aligned}$$

This is signficant because it tells us that the dot product between the vector of explanatory variables $\vec{X} = (X_1, \ldots, X_i)^T$ is orthogonal to the vector of error terms $\vec{e} = (e_1, \ldots, e_n)^T$, and from the previous property we get that

$$\vec{e} \cdot 1_n = \sum_{i=1}^{n} e_i = 0$$

Hence the vectors $\{1_n, X - \bar{X}1_n\}$ are linearly independent and form a basis of the estimation space.

(iv) By applying the Pythagorean Theorem to the previous property we get

$$\begin{aligned} ||Y||^2 &= ||\hat{Y}||^2 + ||Y - \hat{Y}||^2 \\ \sum_{i=1}^{n} Y_i^2 &= \sum_{i=1}^{n} \hat{Y}_i^2 + \sum_{i=1}^{n} e_i^2 \\ &= \sum_{i=1}^{n} \bar{Y}^2 + b_1^2 \sum_{i=1}^{n} (X_i - \bar{X})^2 + \sum_{i=1}^{n} e_i^2 \\ \implies \sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2 &= b_1^2 \sum_{i=1}^{n} (X_i - \bar{X})^2 + \sum_{i=1}^{n} e_i^2 \\ \sum_{i=1}^{n} (Y_i - \bar{Y})^2 &= b_1^2 \sum_{i=1}^{n} (X_i - \bar{X})^2 + \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 \end{aligned}$$

This shows us the the total sum of squares is equal to the regression sum of squares plus the error sum of squares.

(v) The point $(\bar{X}, \bar{Y})$ is on the fitted line.

(vi) The sum of residuals weighted by their corresponding fitted value is 0, that is

$$\sum_{i=1}^{n} y_i e_i = 0$$

(vii) Under the normality assumption, $e_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$. The method of maximum likelihood leads to the method of least squares.

$$L(\beta_0, \beta_1, \sigma^2) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} \epsilon_i^2\right)$$

So maximizing $L(\beta_0, \beta_1, \sigma^2)$ is equivalent to minimizing $\sum \epsilon_i^2$.

2.1.3 Estimation of σ^2

We need to estimate σ^2 to test hypotheses and construct interval estimates pertinent to the regression model. Ideally we would like this estimate not to depend on the adequacy of the fitted model. This is only possible when there are several observations on y for at least one value of x, or when prior information concerning σ^2 is available. When this approach cannot be used, the estimate of σ^2 is obtained from the residual or error sum of squares.

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

We can substitute $\hat{y}_i$ for $b_0 + b_1 x_i$ and simplify to get

$$SSE = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 - b_1 S_{xy}$$

Morever, the correct sum of squares of the response variable is

$$SST = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Thus,

$$SSE = SST - b_1 S_{xy}$$

The residual sum of squares has $n - 2$ degrees of freedom, because we reserve 2 degrees of freedom for the estimators b_0, b_1. We will later show that the expected value for SSE is

$$E(SSE) = (n - 2)\sigma^2$$

So an unbiased estimator of σ^2 is

$$\hat{\sigma}^2 = \frac{SSE}{n - 2} = MSR$$

The quantity MSR is known as the **residual mean square**. The root of $\hat{\sigma}^2$ is known as the **standard error of regression.**

2.2 Hypothesis Testing on the Slope and Intercept

To preform hypotheses tests and construct confidence intervals, we require that we make the additional assumption that the model errors ϵ_i are normally distributed. Thus, the complete assumptions are that the errors are normally and independently distributed with mean 0 and variance σ^2, written as $\{\epsilon_i\} \overset{\text{iid}}{\sim} N(0, \sigma^2)$. We will discuss how these assumptions can be checked through residual analysis later.

Suppose that we have the model $Y_i = \beta_0 + \beta_1 X_i + \epsilon_1$, where $\{\epsilon_i\} \overset{\text{iid}}{\sim} N(0, \sigma^2)$. Then

(a) $\frac{b_1 - \beta_1}{se(b_1)} \sim t_{n-2}$ where $se^2(b_1) = \frac{MSE}{\sum_{i=1}^n (X_i - \bar{X})^2}$

(b) $\frac{b_0 - \beta_0}{se(b_0)} \sim t_{n-2}$ where

$$se^2(b_0) = MSE\left(\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\right)$$

(c) MSE is an unbiased estimate of σ^2 and is independent of b_0,b_1. Furthermore

$$\frac{(n-2)MSE}{\sigma^2} \sim \chi^2_{n-2}$$

Proof. Proof will be shown when we generalize this using matrices in later sections. □

2.2.1 Using t-tests

Suppose we want to test that the slope is equal to a constant, β, we have the hypotheses

$$H_0 : \beta_1 = \beta,\ H_1 : \beta_1 \neq \beta$$

Since $\{\epsilon_i\} \overset{\text{iid}}{\sim} N(0, \sigma^2)$, the observations y_i are normally distributed with $\beta_0 + \beta_1 x_i$ and variance σ^2. Then, b_1 is a linear combination of the observations, so it is normally distributed with mean β_1 and variance σ^2/S_{xx}. Therefore, our test statistic becomes

$$Z_0 = \frac{b_1 - \beta}{\sqrt{\sigma^2/S_{xx}}}$$

If the null hypothesis is true, then $Z_0 \sim N(0, 1)$. If σ^2 is known then we would use Z_0 to test our hypotheses. However, σ^2 is typically unknown. We've seen that MSE is an unbiased estimator for σ^2, and we've established that $(n-2)MSE/\sigma^2 \sim \chi^2_{n-2}$.

$$t_0 = \frac{b_1 - \beta}{\sqrt{MSE/S_{xx}}}$$

If the null hypothesis is true, $t_0 \sim t_{n-2}$. We compare the observed value t_0 with the upper $\alpha/2$, of the t_{n-2} distribution. So we reject the null hypothesis

$$|t_0| > t_{\alpha/2, n-2}$$

We can also test with the p-value. From the equation for t_0, the denominator is called the **estimated standard error** of the slope.

$$se(b_1) = \sqrt{\frac{MSE}{S_{xx}}}$$

So, we often write t_0 is

$$t_0 = \frac{b_1 - \beta}{se(\beta_1)}$$

We test the intercept in a similar manner,

$$H_0 : \beta_0 = \beta,\ H_1 : \beta_0 \neq \beta$$

We use a similar test statistic,

$$t_0 = \frac{b_0 - \beta}{se(b_0)}$$

and we reject the null hypothesis when $|t_0| > t_{\alpha/2, n-2}$.

2.2.2 Testing Significance

A special case for hypotheses is

$$H_0 : \beta_1 = 0,\ H_1 : \beta_1 \neq 0$$

These hypotheses relate to the **significance of regression**. Failing to reject the null hypothesis means there is no linear relationship between x and y, we would reject the null hypothesis when $|t_0| > t_{\alpha/2,n-2}$.

2.2.3 Analysis of Variance Tables (ANOVA)

Analysis of variance can be used to test significance of regression. The analysis of variance of variance is based on a partitioning of the total variability of the response variable y, given by

$$y_i - \bar{y} = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)$$

Then, taking the sum of the square of both sides

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

Notice that

$$\begin{aligned} 2\sum_{i=1}^{n}(\hat{y}_i - \bar{y})(y_i - \hat{y}_i) &= 2\sum_{i=1}^{n}\hat{y}_i(y_i - \hat{y}_i) - 2\bar{y}\sum_{i=1}^{n}(y_i - \hat{y}_i) \\ &= 2\sum_{i=1}^{n}\hat{y}_i e_i - 2\bar{y}\sum_{i=1}^{n} e_i = 0 \end{aligned}$$

Therefore,

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

The left side is the corrected sum of squares of the observations, which we denote by SST or $SSTO$. Notice that $y_i - \hat{y}_i = e_i$, so that term is the sum of residuals squared SSE. We call $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ the **regression sum of squares**. So we have

$$SST = SSR + SSE$$

The regression sum of squares can also be computed by

$$SSR = b_1^2 S_{xx}$$

The **degrees of freedom** for each sum of squares is as follows.

- The total sum of squares SST has $df_T = n - 1$ since we lose a degree of freedom for the constraint

$$\sum_{i=1}^{n}(y_i - \bar{y}) = 0$$

- The regression sum of squares SSR has $df_R = p - 1$ where p is the number of variables (including y).
- The residual sum of squares SSE has $df_E = n - 2$ degrees of freedom since 2 constraints are placed on $e_i = y_i - \hat{y}_i$ with the estimation for $beta_0$ and β_1.

We create a table to summarize our results from statistical analysis.

Source	SS	DF	MS=SS/df	E(MS)
Regression	$SSR = b_1^2 \sum(X_i - \bar{X})^2$	$p - 1$	MSR	$\sigma^2 + \beta_1^2 \sum(X_i - \bar{X})^2$
Error	$SSE = \sum(Y_i - \hat{Y}_i)^2$	$n - p$	MSE	σ^2
Total	$SSTO = \sum(Y_i - \bar{Y})^2$	$n - 1$		

Each of the sums of squares is a quadratic form where the rank of the corresponding matrix is the degrees of freedom indicated. Chochran's theorem applies and we conclude that the quadratic forms are independent and have chi-sqaured distributions. Note that

$$\frac{SSR}{\sigma^2} = \frac{b_1^2 \sum (X_i - \bar{X})^2}{\sigma^2} \sim \chi^2_{p-1}$$

$$\frac{SSE}{\sigma^2} = \frac{\sum (Y_i - \hat{Y}_i)^2}{\sigma^2} \sim \chi^2_{n-p}$$

Then, the ratio between 2 chi-sqaured distributions divided by their degrees of freedom has a F-distribution with their respective degrees of freedom.

$$F = \frac{SSR/\sigma^2(p-1)}{SSE/\sigma^2(n-p)} = \frac{SSR/(p-1)}{SSE/(n-p)} = \frac{MSR}{MSE} \sim F_{p-1,n-p}$$

The degrees of freedom are determined by the amount of data required to calculate each expression.
To summarize, the ANOVA table indactes how one can test the null hypothesis

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

The null Hypothesis is that the slope of the line is equal to 0. Under the null hypothesis, the expected mean square for regression and the expected mean square error are seperate independent estimates of the variance σ^2.

2.3 Interval Estimation

2.3.1 Confidence Intervals on β_0, β_1, and σ^2.

The width of the confidence intervals are a measure of the quality of the regression line. If the error is normally and independently distributed by our assumption, then $(b_1 - \beta_1)/se(b_1)$ and $(b_0 - \beta_0)/se(b_0)$ follow a t distribution with $n-2$ degrees of freedom. So, a $100(1-\alpha)$ percent. The confidence interval for the slope β_1 is

$$b_1 - t_{\alpha/2,n-2} se(b_1) \leq b_1 \leq b_1 + t_{\alpha/2,n-2} se(b_1)$$

and for the intercept β_0,

$$b_0 - t_{\alpha/2,n-2} se(b_0) \leq b_0 \leq b_0 + t_{\alpha/2,n-2} se(b_0)$$

The interpretation for these intervals is, if we were to take repeated samples of the same size at the same x levels and construct 95% CIs on the slope for each sample, then 95% of those intervals will contain the true value of β_1.

As we've seen earlier, the sampling distribution of $(n-2)MSE/\sigma^2$ follows a chi-sqaure distribution with $n-2$ degrees of freedom, so

$$P\left\{\chi^2_{1-\alpha/2,n-2} \leq \frac{(n-2)MSE}{\sigma^2} \leq \chi^2_{\alpha/2,n-2}\right\}$$

Thus the $100(1-\alpha)$ percent CI on σ^2 is

$$\frac{(n-2)MSE}{\chi^2_{\alpha/2,n-2}} \leq \sigma^2 \leq \frac{(n-2)MSE}{\chi^2_{1-\alpha/2,n-2}}$$

2.3.2 Interval Estimation of the mean Response

Another important part of the regression model is estimating the mean response $E(y)$ for a particular regressor variable x. Assuming that x_0 is any value of the regressor variable within the range of the original data on x that we used to create the moel. Then, an unbiased estimator for $E(y|x_0)$ can be found from the fitting model

$$\widehat{E(y|x_0)} = \hat{\mu}_{y|x_0} = b_0 + b_1 x_0$$

Note that $\hat{\mu}_{y|x_0}$ follows a normal distribution since it is a linear combination of the observations y_i. The variance is

$$\text{Var}(\hat{\mu}_{y|x_0}) = \text{Var}(b_0 + b_1 x_0) = \text{Var}(\bar{y} - b_1(x_0 - \bar{x})) = \frac{\sigma^2}{n} + \frac{\sigma^2 (x_0 - \bar{x})^2}{S_{xx}}$$

The sampling distribution for

$$\frac{\hat{\mu}_{y|x_0} - E(y|x_0)}{\sqrt{MSE(1/n + (x_0 - \bar{x})^2/S_{xx})}}$$

is a t distribution with $n-2$ degrees of freedom. Then the CI is given as

$$\left[\hat{\mu}_{y|x_0} \pm t_{\alpha/2,n-2}\sqrt{MSE\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}\right]$$

2.4 Coefficient of Determination

The quantity

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

is called the **coefficient of determination.** R^2 is also called the proportion of variation explained by the regressor x since SST is a measure of variability in y without considering the effect of x, and SSE is the variability in y after considering x. Since $0 \leq SSE \leq SST$, then $0 \leq R^2 \leq 1$. An R^2 value close to 1 means **most of the variability of y is explained by x.**

2.5 Correlation Coefficient

The pearson correlation coefficient, denoted by ρ, related to b_1 is given as

$$\rho = \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)}\sqrt{\text{Var}(Y)}}$$

This measures the linear correlation between 2 variables. When applied to a sample,

$$r = b_1\left(\frac{S_{xx}}{SST}\right)^{\frac{1}{2}} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2 (Y_i - \bar{Y})^2}} = \frac{S_{xy}}{(S_{xx}SST)^{1/2}}$$

Note that $-1 \leq r \leq 1$. To test hypotheses on ρ, we have 2 cases. The hypotheses for testing if the correlation is 0 is as follows

$$H_0 : \rho = 0,\ H_1 : \rho \neq 0$$

When testing the null hypothesis $\rho = 0$, we use a t statistic given as

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{n-2}$$

When testing

$$H_0 : \rho = \rho_0,\ H_1 : \rho \neq \rho_0$$

We use a Z statistic,

$$Z = \text{arctanh}\, r = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) \sim N\left(\mu_z, \frac{1}{n-3}\right)$$

where

$$\mu_z = \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right)$$

Now we can standardize our statistic to obtain a standard normal test statistic

$$Z_0 = (\text{arctanh}(r) - \text{arctanh}(\rho_0))\sqrt{n-3}$$

We can obtain our confidence interval with

$$\tanh\left(\text{arctanh}(r) - \frac{Z_{\alpha/2}}{\sqrt{n-3}}\right) \leq \rho \leq \tanh\left(\text{arctanh}(r) + \frac{Z_{\alpha/2}}{\sqrt{n-3}}\right)$$

where $\tanh(u) = (e^u - e^{-u})/(e^u + e^{-u})$.

Chapter 3

Multiple Linear Regression

We call a regression model with more than one regressor variable a **multiple regression model.**

3.1 Matrix Approach to Regression

We will first cover simple linear regression in matrix form.

Let $Y = [Y_1, \ldots, Y_n]^T$ be a column data vector, and we'll define the expected value as

$$E(Y) = \begin{bmatrix} E(Y_1) \\ \vdots \\ E(Y_n) \end{bmatrix}$$

Proposition 3.1.1. *If $Z = AY + B$ for a matrix of constants A, and B, then*

$$E(Z) = AE(Y) + B$$

Proof. Simply from the definition of expectations on vectors,

$$E(Z_i) = E\left(\left[\sum_j a_{ij}Y_j\right] + b_i\right) = \sum_j a_{ij}E(Y_j) + b_i$$

So

$$E(Z) = AE(Y) + B$$

□

Definition 3.1.1. *The covariance of a vector of data*

$$Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}$$

is

$$\mathrm{Cov}(Y) = E([Y - E(Y)][Y - E(Y)]^T) = \Sigma$$

Proposition 3.1.2. $\mathrm{Cov}(AY) = A\Sigma A^T$.

Definition 3.1.2. *A random vector Y has a multivariate normal distribution if its density is given by*

$$f(y_1, \ldots, y_n) = \frac{|\Sigma|^{-1/2}}{\exp\left(-\frac{1}{2}(Y-\mu)^T\Sigma^{-1}(Y-\mu)\right)}$$

where

$$Y^T = (y_1, \ldots, y_n),\ \mu^T = (\mu_1, \ldots, \mu_n)$$

we denote this by

$$Y \sim N_n(\mu, \Sigma)$$

Theorem 3.1.1. *Let $Y \sim N_n(\mu, \Sigma)$. Let A be an arbitrary $p \times n$ matrix of constants. Then*

$$Z = AY + B \sim N_p(A\mu + B, A\Sigma A^T)$$

This theorem implies that any linear combination of normal variates has a normal distribution. This theorem won't be proved here.

3.1.1 Derivatives

- $z = a'y \to \frac{\partial z}{\partial y} = a$
- $z = y'y \to \frac{\partial z}{\partial y} = 2y$
- $z = a'Ay \to \frac{\partial z}{\partial y} = A'a$
- $z = y'Ay \to \frac{\partial z}{\partial y} = A'y + Ay$
- If A is symmetric, then $z = y'Ay \to \frac{\partial z}{\partial y} = 2A'y$

3.2 Multiple Regression Models

Suppose we have 2 regressor variables, a multiple regression model that may describe a relationshup with our data is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

The parameter β_1 indicates the expected change in response per unit change in x_1 when x_2 is held constant. Similarly β_2 measures the change in y per unit change in x_2 when x_1 is held constant.

3.2.1 Least Squares Estimation of Regression Coefficients

The method of **least squares** can be used to estimate the regression coefficients. Suppose $n > k$ observations are available, and let y_i denote the ith observed response x_{ij} denote the ith observation or level of regressor. We can write the sample regression model as

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \epsilon_i$$

for $i = 1, 2, \ldots, n$. Then the least squares function is

$$S(\beta_0, \beta_1, \ldots, \beta_k) = \sum_{i=1}^{n} \epsilon_i = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij} \right)^2$$

Similar to the simple linear regression approach with Q, we want to minimize S witih respect to $\beta_0, \beta_1, \ldots, \beta_k$. So the least squares estimators must satisfy

$$\left. \frac{\partial S}{\partial \beta_0} \right|_{\hat{\beta_0}, \ldots, \hat{\beta_k}} = -2 \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \sum_{j=1}^{k} \beta_j x_{ij} \right) = 0$$

and similarly for the rest of the estimators for β_j for $j = 1, \ldots, k$.

$$\left. \frac{\partial S}{\partial \beta_j} \right|_{\hat{\beta_0}, \ldots, \hat{\beta_k}} = -2 \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \sum_{j=1}^{k} \beta_j x_{ij} \right) x_{ij} = 0$$

When simplifying these equations, we get the least squares **normal equations**, which we will put in terms of matrices later on.

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1} + \hat{\beta}_2 \sum_{i=1}^{n} x_{i2} + \cdots + \hat{\beta}_k \sum_{i=1}^{n} x_{ik} = \sum_{i=1}^{n} y_i$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_{i1} + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1}^2 + \hat{\beta}_2 \sum_{i=1}^{n} x_{i1} x_{i2} + \cdots + \hat{\beta}_k \sum_{i=1}^{n} x_{ik} x_{i1} = \sum_{i=1}^{n} x_{i1} y_i$$

$$\vdots$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_{ik} + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1}x_{ik} + \hat{\beta}_2 \sum_{i=1}^{n} x_{ik}x_{i2} + \cdots + \hat{\beta}_k \sum_{i=1}^{n} x_{ik}^2 = \sum_{i=1}^{n} x_{ik}y_i$$

Note that there are $p = k + 1$ normal equations for each of the unknown regression coefficients. The solutions give us the **least squares estimators** $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$.

As mentioned earlier, it is easier to work with matrix notation, so we will express the model using matrices

$$Y = X\beta + \epsilon$$

where

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \; X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots x_{1k} \\ 1 & x_{21} & x_{22} & \cdots x_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots x_{nk} \end{bmatrix}$$

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \; \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

In general, Y is an $n \times 1$ vector of observations, X is an $n \times p$ matrix of the levels of the regressor variables, β is a $p \times 1$ vector of regression coefficients, and ϵ is an $n \times 1$ vector of random variables. Note that from here on out we will denote the transpose of a matrix A^T as A'. We want to find the vector of least sqaures estimators $\hat{\beta}$ that minimizes

$$S(\beta) = \sum_{i=1}^{n} = \epsilon'\epsilon = (Y - X\beta)'(Y - X\beta)$$

This can be expanded as

$$S(\beta) = Y'Y - \beta'X'Y - Y'X\beta + \beta'X'X\beta = Y'Y - 2\beta'X'Y + \beta'X'X\beta$$

This is derived from using the fact that $\beta'X'Y$ is a 1×1 matrix, so it is a scalar and its transpose is the same. Now our equation must satisfy

$$\left.\frac{\partial S}{\partial \beta}\right|_{\hat{\beta}} = -2X'Y + 2X'X\hat{\beta} = 0$$

This gives us

$$X'X\hat{\beta} = X'Y$$

The equations here our are **least-squares** normal equations. They are the same as the previous equations we found earlier not in matrix form.

We can solve the normal equations by multiplying both sides of our equation by $X'X$ to give us

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

provided that the inverse exists.

The fitted regression model corresponding to the levels of regressor vairables $x' = [1, x_1, x_2, \ldots, x_k]$ is then

$$\hat{y} = x'\hat{\beta} = \hat{\beta}_0 + \sum_{j=1}^{k} \hat{\beta}_j x_j$$

So the vector of fitted values $\hat{y}_i$ corresponding to the observed values y_i it

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

The $n \times n$ matrix $H = X(X'X)^{-1}X'$ is known as the **hat matrix**. It maps the vector of observed values to a vector of fitted values (in otherwords, it "puts the hat" on Y).

3.2.2 Properties of the Hat Matrix H

The hat matrix has some useful properties, notably

(a) H is a projection matrix, so it is idempotent and symmetric

$$HH = H$$

$$H' = H$$

(b) The matrix H is orthogonal to the matrix $I - H$, so

$$(I - H)H = H - HH = 0$$

Moreover, $(I - H)$ is idempotent and a project matrix as well.

(c) The vector of residuls, which we will denote as $\vec{e}$, is given as

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

(d) Properties (b) and (c) imply that the observation vector Y is projected onto a space spanned by the columns of H, and the residuals are in an orthogonal space (similar to the case for the simple linear regression model).

$$Y = HY + (I - H)Y$$

(e) Similar to our simple linear model, we can apply the pythagorean theorem (from the fact that the matrices are orthogonal) to obtain

$$||Y||^2 = ||HY||^2 + ||(I - H)||^2$$

We will see later why expressing

$$\vec{e} = Y - X\hat{\beta} = Y - HY = (I - H)Y$$

is useful.

3.2.3 Properties of the Least-Squares Estimators

We can first show that $\hat{\beta}$ is an unbiased estimators,

$$\begin{aligned} E(\hat{\beta}) &= E\left[(X'X)^{-1}X'Y\right] \\ &= E\left[(X'X)^{-1}X'(X\beta + \epsilon)\right] \\ &= E\left[(X'X)^{-1}X'X\beta + X'\epsilon\right] \\ &= E\left[\beta + \epsilon\right] \\ &= E(\beta) + E(\epsilon) \\ &= \beta \end{aligned}$$

This is using the fact that $E(\epsilon) = 0$. So if our model assumptions hold, $\hat{\beta}$ is an unbiased estimator for β. The variance of $\hat{\beta}$ is expressed by the covariance matrix

$$\text{Cov}(\hat{\beta}) = E\left\{[\hat{\beta} - E(\hat{\beta})][\hat{\beta} - E(\hat{\beta})]^T\right\}$$

3.2.4 Estimation of σ^2

We may devlop an estimator for σ^2 using the residual sum of squares,

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2 = e'e$$

Then, we can substitute $e = Y - X\hat{\beta}$,

$$\begin{aligned} SSE &= (Y - X\hat{\beta})'(Y - X\hat{\beta}) \\ &= Y'Y - \hat{\beta}'X'Y - Y'X\hat{\beta} + \hat{\beta}'X'X\hat{\beta} \\ &= Y'Y - 2\hat{\beta}'X'Y + \hat{\beta}'X'X\hat{\beta} \end{aligned}$$

Then since $X'X\hat{\beta} = X'y$, we get

$$SSE = Y'Y - \hat{\beta}'X'Y$$

It can be shown that SSE has $n-p$ degres of freedom since p parameters are estimated in the regression model. This gives us the **residual mean square** or **mean sqaure error**

$$MSE = \frac{SSE}{n-p}$$

It can be shown that MSE is again an unbiased estimator for σ^2, therefore

$$\hat{\sigma}^2 = MSE$$

3.3 Estimation and Hypothesis Testing in Multiple Linear Regression

The true relationship between y and our regressors $x_1, \ldots, x_k$ is unknown, and our multiple linear regression model is used to approximate this. Models sometimes are more complex in structure than the model we've discussed,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon$$

but we can still use a multiple linear regression model. For eaxmple, consider a cubic polynomial model

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \epsilon$$

If we set $x_1 = x$, $x_2 = x^2$, and $x_3 = x^3$, we can rewrite the equation as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

Models that include **interaction effects** may also be analyzed by multiple linear regression methods. For example if we have the model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \epsilon$$

we can set $x_3 = x_1 x_2$, and $\beta_3 = \beta_{12}$, then we get

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

Finally, consider a **second-order model with interaction**

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + \epsilon$$

We can set $x_3 = x_1^3$, $x_4 = x_2^2$, $x_5 = x_1 x2$, $\beta_3 = \beta_{11}$, $\beta_4 = \beta_{22}$, and $\beta_5 = \beta_{12}$, then we can get our linear model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \epsilon$$

Our predictor variables my also be qualitative, for example

$$X = \begin{cases} 0 & \text{if subject is male} \\ 1 & \text{if subject is female} \end{cases}$$

We could also have a transformed response variable

$$\ln Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i$$

Now we want to be able to ask questions about our models adequacy and which regressors are important. To do this, we can conduct various hypothesis tests which also require that our random errors are independent and identically distributed from a normal distribution with mean $E(\epsilon_i) = 0$, and variance $\text{Var}(\epsilon_i) = \sigma^2$.

3.3.1 Testing for Significance of Regression

The test for **significance of regression** is to determine if there is a linear relationship between the repsonse Y and any of the regressors $x_1, \ldots, x_k$. The approriate hypotheses are

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0, \; H_1 : \beta_j \neq 0$$

for *at least one* value of j. Rejection of the null hypothesis implies *at least one* of the regressors contribute signficantly to the model. The test is a generalization of the anaylsis of variance used in simple regression. We define the sum of squares the same way,

$$SST = SSR + SSE$$

If the null hypothesis is true, then $SSR/\sigma^2 \sim \chi^2_k$, where k is the number of regressors in the model $(p-1)$. Similarly, $SSE/\sigma^2 \sim \chi^2_{n-k-1}$ if the null hypothesis holds and it can be shown that SSR/σ^2 is independent of SSE/σ^2. We then use an F statistic

$$F_0 = \frac{SSR/k}{SSE/(n-k-1)} = \frac{MSR}{MSE} \sim F_{k,n-k-1}$$

which follows an F distribution. Then, we reject the null hypothesis $H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$ if

$$F_0 > F_{\alpha,k,n-k-1}$$

We can construct an analysis of variance table to summarize this procedure again.

Source of Variation	Sum of Squares	DF	Mean Square	F-statistic
Regression	SSR	k	MSR	MSR/MSE
Residuals	SSE	$n-k-1$	MSE	
Total	SST	$n-1$		

The **total sum of squares** is

$$SST = \sum_{i=1}^{n} Y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} Y_i\right)^2 = Y'Y - \frac{1}{n}\left(\sum_{i=1}^{n} Y_i\right)$$

The **regression sum of squares** is

$$SSR = \hat{\beta}' X'Y - \frac{1}{n}\left(\sum_{i=1}^{n} Y_i\right)$$

The **residual sum of squares** is

$$SSE = Y'Y - \hat{\beta}' X'Y = Y'(I - H)Y$$

We can show that $Y'J_nY = \left(\sum_{i=1}^n Y_i\right)^2$ where J_n is the matrix with 1 in every entry

$$\begin{aligned}
Y'J_nY &= \begin{bmatrix} Y_1 & \cdots & Y_n \end{bmatrix} \begin{bmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{bmatrix} \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} \\
&= \begin{bmatrix} \sum Y_i & \cdots \sum Y_i \end{bmatrix} \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} \\
&= Y_1 \sum_{i=1}^n Y_i + Y_2 \sum_{i=1}^n Y_i + \cdots + Y_n \sum_{i=1}^n Y_i \\
&= \sum_{i=1}^n Y_i (Y_1 + Y_2 + \cdots + Y_n) \\
&= \left(\sum_{i=1}^n Y_i\right)\left(\sum_{i=1}^n Y_i\right) \\
&= \left(\sum_{i=1}^n Y_i\right)^2
\end{aligned}$$

Thus we can write

$$\begin{aligned}
SST &= \sum_{i=1}^n Y_i^2 - \frac{1}{n}\left(\sum_{i=1}^n Y_i\right)^2 \\
&= Y'Y - \frac{1}{n}Y'J_nY \\
&= Y'\left(Y - \frac{1}{n}J_nY\right) \\
&= Y'\left(I - \frac{1}{n}J_n\right)Y
\end{aligned}$$

It can be shown easily that we can write the regression sum of squares in terms of this matrix J_n

$$SSR = Y'\left(H - \frac{1}{n}J_n\right)Y$$

3.3.2 Tests on Individual Regression Coefficients

Once we have determined that at least one of the regressors is significant, we can investigate which one (or more) is significant. Adding a varialbe to the model always causes the regression sum of squares to increase and the residual sum of squares to decrease. So we must decie whter the increase in SSR is sufficient to warrant using an additional regressor. The added regressor also increases the variance for the fitted value $\hat{y}$, so we must be careful to add only regressors that are signficant. Adding an unimportant regressor can increase the mean square error, which can decrease the usefulness of the model.

To test an individual regression coefficient, say β_j, we use the test

$$H_0 : \beta_j = 0,\ H_1 : \beta_j \neq 0$$

If the null hypothesis is not rejected, then we can remove the corresponding regressor x_j. The **test statistic** for this hypothesis is

$$t_0 = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 C_{jj}}} = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)}$$

Where C_{jj} is the diagonal element of the matrix $(X'X)^{-1}$ corresponding to β_j. We reject the null hypothesis if

$$|t_0| > t_{\alpha/2, n-k-1}$$

3.3.3 Extra Sum of Squares Principle

We may also directly determine the contribution to the regression sum of squares of a regressor, say x_j, by using the **extra sum of squares method.** This procedure can also be used to investigate the contribution of a subset of the regressor variables.

Consider the regression model with k regressors,

$$Y = X\beta + \epsilon$$

where Y is $n \times 1$, X is $n \times p$, β is $p \times 1$, and ϵ is $n \times 1$ with $p = k + 1$. We want to determine if some subset of $r < k$ regressors contribute signficantly to our model. We will partition the regression coefficients into 2 vectors, β_1 is a $(p - r) \times 1$ vector, and β_2 is the $r \times 1$ vector of coefficients we are trying to test. We want to test the following hypotheses

$$H_0 : \beta_2 = 0,\ H_1 : \beta_2 \neq 0$$

So our model will be rewritten as

$$Y = X\beta + \epsilon = X_1\beta_1 + X_2\beta_2 + \epsilon$$

where X_1 is a $n \times (p - r)$ matrix that are the columns of X associated with β_1, and X_2 is an $n \times r$ matrix with the columns of X associated with β_2. This is our **full model**.

For the full model, we have established that

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

and the regression sum of squares is

$$SSR(\beta) = \hat{\beta}'X'Y$$

which has $k = p - 1$ degrees of freedom. We also have the residual mean square

$$MSE = \frac{Y'Y - \hat{\beta}'X'Y}{n - p} = \frac{SSE}{n - p}$$

To find the contribution of the regressors in β_2, we fit the model assuming the null hypothesis $H_0 : \beta_2 = 0$ is true. This gives us the **reduced model**

$$Y = X_1\beta_1 + \epsilon$$

The least squares estimator for β_1 in the reduced model is

$$\hat{\beta}_1 = (X_1'X_1)^{-1}X_1'Y$$

The regression sum of squares is

$$SSR(\beta_1) = \hat{\beta}_1{}'X_1'Y$$

which has $k - r = p - 1 - r$ degrees of freedom. The regression sum of squares due to β_2 given that β_1 is already in the model is

$$SSR(\beta_2|\beta_1) = SSR(\beta) - SSR(\beta_1)$$

with $(p - 1) - (p - 1 - r) = r$ degrees of freedom. This sum of squares is called the **extra sum of squares due to** β_2 because it measures the increase in the regression sum of squares that results from adding $x_{k-r+1}, x_{k-r+2}, \ldots, x_k$ to the model with $x_1, x_2, \ldots, x_{k-r}$. Now we have the extra sum of squares due to β_2 is independent of mean square error, so the null hypothesis $\beta_2 = 0$ can be tested with the following statistic,

$$F_0 = \frac{SSR(\beta_2|\beta_1)/r}{MSE}$$

If $F_0 > F_{\alpha,r,n-p}$, then we reject the null hypothesis and conclude that at least one of the parameters in β_2 is not zero, and at least one of the regressors in X_2 contributes signficantly to the regression model. This is sometimes called a **partial F test** since it measures the contribution of regressors in X_2 given that X_1 is in the model.

If we want to compute the regression squares for multiple variables, say

$$SSR(\beta_2|\beta_1, \beta_0)$$

We compute it in the following way

$$SSR(\beta_2|\beta_1,\beta_0) = SSR(\beta_2,\beta_1,\beta_0) - SSR(\beta_1,\beta_0) = SSR(\beta_1,\beta_2|\beta_0) - SSR(\beta_1|\beta_0)$$

We can partition the regression sum of squares into marginal single degree of freedom components. For example, consider the model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

We can use the following identity,

$$SST = SSR(\beta_1,\beta_2,\beta_3|\beta_0) + SSE$$

We can then decompose the three degree of freedom regression sum of squares as

$$SSR(\beta_1,\beta_2,\beta_3|\beta_0) = SSR(\beta_1,\beta_0) + SSR(\beta_2|\beta_1,\beta_0) + SSR(\beta_3|\beta_2,\beta_1,\beta_0)$$

3.3.4 Testing the General Linear Hypothesis

Suppose that the null hypothesis we want to test is $H_0 : T\beta = 0$ where T is an $r \times p$ matrix such that only r of the p equations in $T\beta = 0$ are independent. That is, the rows are the independent equations and will yield a $r \times 1$ vector of coefficients we'd like to test. Recall the β is a $p \times 1$ vector. The **full model** stays the same, with $Y = X\beta + \epsilon$ and $\hat{\beta} = (X'X)^{-1}X'Y$, and the residual sum of squares is

$$SSE(FM) = Y'Y - \hat{\beta}'X'Y$$

which has $n - p$ degrees of freedom. To obtain the **reduced model**, the r independent equations in $T\beta = 0$ are used to solve for r of the regression coefficients in the full model in terms of the $p - r$ regression coefficient. This gives us the reduced model

$$Y = Z\gamma + \epsilon$$

where Z is an $n \times (p - r)$ matrix, and γ is a $(p - r) \times 1$ vector of the unknown regression coefficients. The estimator for γ is

$$\hat{\gamma} = (Z'Z)^{-1}Z'Y$$

and the residual sum of squares is

$$SSE(RM) = Y'Y - \hat{\gamma}Z'Y$$

which as $n-p+r$ degrees of freedom. The reduced model has less parameters than the full model so $SSE(RM) \geq SSE(FM)$. Now to test the hypothesis $H_0 : T\beta = 0$, we use the difference in residual sum of squares, denoted by SSH,

$$SSH = SSE(RM) - SSE(FM)$$

which has $n-p+r-(n-p) = r$ degrees of freedom. This is known as the sum of squares due to the hypothesis. The test statistic here is then

$$F_0 = \frac{SSH/r}{SSE(FM)/n-p} \sim F_{r,n-p}$$

We reject H_0 if $F_0 > F_{\alpha,r,n-p}$.

3.4 Lack of Fit of the Regression Model

Sometimes, we want to test whether or not our linear model is justified at all. This is different from testing if the slope is 0.

3.4.1 Test for Lack of Fit

Suppose we have n_i observations on the response at the ith level of the regressor x_i for $i = 1, 2, \ldots, m$. Let y_{ij} denote the jth ($j = 1, 2, \ldots, n_i$) observation on the corresponding response variable. In total, there are

$$n = \sum_{i=1}^{m} n_i$$

observations. We partition the residual sum of squares

$$SSE = SS_{PE} + SS_{LOF}$$

Where SS_{PE} is the sum of squares due to **pure error** and SS_{LOF} is the sum of squares due to **lack of fit**. Note that we can parition SSE using the following,

$$y_{ij} - \hat{y}_i = (y_{ij} - \bar{y}_i) + (\bar{y}_i - \hat{y}_i)$$

$\bar{y}_i$ is the average from the n_i observations at x_i, then we can square and sum both sides of the equation to give us

$$\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \hat{y}_i)^2 = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2 + \sum_{i=1}^{m}\sum_{j=1}^{n_i}(\bar{y}_i - \hat{y}_i)^2$$

This gives us that the pure error sum of squares is

$$SS_{PE} = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2$$

Since there are $n_i - 1$ degrees of freedom for pure error at each level x_i that gives us the total degrees of freedom for SS_{PE} as

$$\sum_{i=1}^{m}(n_i - 1) = n - m$$

The sum of squares due to lack of fit is

$$SS_{LOF} = \sum_{i=1}^{m} n_i(\bar{y}_i - \hat{y}_i)^2$$

This is a weighted sum of squared deviations between the mean response $\bar{y}_i$ at each level x_i with its corresponding fitted value $\hat{y}_i$. If the fitted values are close to the average responses, then there is a strong indication the regression function is linear. There are $m - 2$ degrees of freedom for SS_{LOF} since there are m levels and 2 degrees are reserved for estimating the 2 parameters β_0, β_1 to compute $\hat{y}_i$. This gives us our test statistic

$$F^* = \frac{SS_{LOF}/(m-2)}{SS_{PE}/(n-m)} = \frac{MS_{LOF}}{MS_{PE}}$$

The expected value for MS_{PE} is σ^2, and for MS_{LOF} it is

$$E(MS_{LOF}) = \sigma^2 + \frac{\sum_{i=1}^{n} n_i \left[E(y_i) - \beta_0 - \beta_1 x_i\right]^2}{m-2}$$

If the regression function is linear, then $E(y_i) = \beta_0 + \beta_1 x_i$ so $E(MS_{LOF})$ is σ^2. So, we construct our hypotheses as

$$H_0 : E(y_i) = \beta_0 + \beta_1 x_i,\ H_1 : E(y_i) \neq \beta_0 + \beta_1 x_i$$

and we reject the null hypothesis if $F^* > F_{\alpha,m-2,n-m}$. We can summarize these in another analysis of variance table,

Source	Sum of Squares	DF	Mean Square	F
Regression	$SSR = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \hat{y}_i)^2$	1	$SSR/1$	MSR/MSE
Residuals	$SSE(R) = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \hat{y}_i)^2$	$n-2$	$SSE(R)/n-2$	
Lack of Fit	$SS_{LOF} = \sum_{i=1}^{m} n_i(\bar{y}_i - \hat{y}_i)^2$	$m-2$	$SS_{LOF}/m-2$	MS_{LOF}/MS_{PE}
Pure Error	$SS_{PE} = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2$	$n-m$	$SS_{PE}/n-m$	
Total	$\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \bar{y})^2$	$n-1$		

It may be also useful to note that

$$E(SS_{LOF}) = \sigma^2 + \frac{\sum_{i=1}^{m} n_i (E(y_i) - \beta_0 - \beta_1 x_i)^2}{m-2}$$

and $E(SS)_{LOF} = \sigma^2$ when we fail to reject the null hypothesis $H_0 : E(y_i) = \beta_0 + \beta_1 x$, since the second term becomes 0, and

$$E(SS_{PE}) = \sigma^2$$

3.5 Confidence Intervals in Multiple Regression

3.5.1 Confidence Intervals on Regression Coefficients

We want to construct confidence intervals for the regression coefficients β_j. We use the same assumptions that $\epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$. So our observations y_i are normally and independently distributed with mean

$$\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}$$

and variance σ^2. $\hat{\beta}$ is a linear combination of the observations, so it follows a normal distribution with a mean vector β and covariance matrix $\sigma^2(X'X)^{-1}$. This implies that the marginal distribution for any $\hat{\beta}_j$ is normal with mean β_j and variance $\sigma^2 C_{jj}$ where C_{jj} is the jth diagonal entry in the matrix $(X'X)^{-1}$. Thus,

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\sigma}^2 C_{jj}}} \sim t_{n-p}$$

So our $100(1-\alpha)$ percent confidence interval for β_j is

$$\hat{\beta}_j - t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2 C_{jj}} \leq \beta_j \leq \hat{\beta}_j - t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2 C_{jj}}$$

Recall that we refer to

$$se(\hat{\beta}_j) = \sqrt{\hat{\sigma}^2 C_{jj}}$$

as the standard error of $\hat{\beta}_j$.

3.5.2 Confidence Intervals On the Mean Response

We can also construct confidence intervals on a specfic point, such as $x_{01}, x_{02}, \ldots, x_{0k}$. We'll define the vector x_0 as

$$x_0 = \begin{bmatrix} 1 \\ x_{01} \\ x_{02} \\ \vdots \\ x_{0k} \end{bmatrix}$$

The fitted value coressponding to this point is

$$\hat{y}_0 = x_0'\hat{\beta}$$

This is an unbiased estimator for $E(y|x_0)$ since $E(\hat{y}_0) = x_0'\beta = E(y|x_0)$, and the variance is

$$\text{Var}(\hat{y}_0) = \sigma^2 x_0'(X'X)^{-1}x_0$$

Thus our confidence interval for the mean response $E(y|x_0)$ is

$$\hat{y}_0 - t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2 x_0'(X'X)^{-1}x_0} \leq E(y|x_0) \leq \hat{y}_0 + t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2 x_0'(X'X)^{-1}x_0}$$

3.5.3 Simultaneous Confidence Intervals on Regression Coefficients

Theorem 3.5.1 (Bonferroni Inequality)**.** *For two events A_1, A_2, we have that*

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

From DeMorgan's identity, we also have

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

where A_1^c is the complement of A_1.

If we define the events

$$A_1^c : \hat{\beta}_0 \pm t_{1-\alpha/2,n-2} s(\hat{\beta}_0)$$

$$A_2^c : \hat{\beta}_1 \pm t_{1-\alpha/2,n-2} s(\hat{\beta}_1)$$

where $s(\hat{\beta}_0)$,$s(\hat{\beta}_1)$ are the standard deviations of $\hat{\beta}_0$ and $\hat{\beta}_1$. So the event $(A_1^c \cap A_2^c)$ is the event that the intervals simultaneous cover β_0, β_1. From Bonforroni's Inequality, if we have $P(A_1) = P(A_2) = \alpha$, then

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

In general, if we have p parameters and each confidence interval has confidence, $1 - \frac{\alpha}{p}$, then

$$P\left(\bigcap_{i=1}^{p} A_i^c\right) \geq 1 - p\frac{\alpha}{p} = 1 - \alpha$$

Conceptually, say we construct $100(1-\alpha)\%$ confidence intervals on β_1 and β_2 as we have done before with $\alpha = 0.05$. This means, we are 95% confident that the event where β_1 is in its CI occurs, and 95% confident that the event where β_0 is in its interval occurs. So, with the events A_1^c being that β_0 is in its confidence interval, and A_2^c the event β_1 is in its confidence interval, the probability that both β_0 and β_1 are *simultaneously in their respective confidence intervals* is $P(\bigcap A_i^c) \geq 1 - 2\alpha = 0.9$. Thus, if we want to be $100(1-\alpha)\%$ certain that *both β_0 and β_1 simultaneously lie in their respective confidence intervals*, we need to construct our individual confidence intervals with $100(1-\alpha/p)\%$ confidence, where p is the number of coefficients.

Chapter 4

Model Adequacy

We've made various assumptions thus far, such as

1. The relationship between the response y and the regressors is linear
2. The error terms ϵ are normally distributed with mean 0 and variance σ^2.
3. The errors are uncorrelated.

We want to ceheck the validity of these assumptions.

4.1 Residual Analysis

Recall the definition for our observed residuals

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

Their variance is estimated by

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

The residuals are not independent, however, as the n residuals have only $n-p$ degrees of freedom associated with them. This nonindependence of the residuals has little effect on their use for model adequacy checking as long as n is not small relative to the number of parameters p.

4.1.1 Checking for Normality

If the assumption of normality holds, a box plot of the residuals should indicate a symmetric box around the median of 0. A histogram of the residuals can also be used to examine normality. If the residuals follow a similar curve as a normal distribution, then it suggests that the normaility assumption is reasonable. The skewness and kurtosis can also help provide insight in this case, a normal distribution has a skewness of 0 and a kurtosis of 3. Finally, a quantile-quantile plot (also known as a qq-plot) comapres the quantiles of the residual data with the quantiles from a normal distribution. We compute

$$E_k = \sqrt{MSE} \cdot \Phi^{-1}\left(\frac{k-0.375}{n+0.25}\right),\ k = 1, \ldots, n$$

Where Φ is the standard normal, and we plot $e_{(k)}$ vs E_k where $e_{(k)}$ is the residual with rank k. Under normality, we would expect a straight line.

4.1.2 Checking Constant Variance

We use MSE as an estimate for approximating the variance of residual. We can improve the residual scaling by dividing the residuals e_i by the exact standard deviations for the ith residual. Recall that we can use the notion with the hat matrix H to write the vector of residuals as

$$\boldsymbol{e} = (I - H)Y$$

where $H = X(X'X)^{-1}X'$ is the hat matrix. Recall the properties of the hat matrix, namely that it is symmetric and idempotent, and $I - H$ has the same properties. We can write the residuals as

$$\boldsymbol{e} = (I - H)(X\boldsymbol{\beta} + \boldsymbol{\epsilon}) = (I - H)\boldsymbol{\epsilon}$$

The covariance matrix of the residuals is

$$\text{Var}(\boldsymbol{e}) = \text{Var}[(I - H)\boldsymbol{\epsilon}] = (I - H)\,\text{Var}(\boldsymbol{\epsilon})(I - H)' = \sigma^2(I - H)$$

This gives us the variance of the ith residual as

$$\text{Var}(e_i) = \sigma^2(1 - h_{ii})$$

and the covariance between residuals e_i and e_j as

$$\text{Cov}(e_i, e_j) = -\sigma^2 h_{ij}$$

Now, we can studentize the reisduals to obtain

$$r_i = \frac{e_i}{\sqrt{MSE(1 - h_{jj})}}$$

We can plot the studentized residuals vs fitted values to check for non-constancy of variance. The plot should show a random distribution of points. Conversly, non-constance variance would appear as a pattern, an increasing or decreasing collection of points. A scale-location plot can also be used to exam homogeneity, by plotting

$$\sqrt{|r_i|} \text{ vs } \hat{Y}_i$$

If the residuals lie in a narrow band around 0 then there are no evidence to suggest we need corrections. Otherwise, if the residuals show a pattern, either increasing or decreasing, this is a sign that variance is non constant. If a double-bow pattern appears, this is indication that the variance in the middle is larger than the variance at the extremes. If the residuals appear to have a quadratic relationship (i.e a parabola shape), there may be a nonlinear relation that the model has not accounted for.

Chapter 5

Transformations and Weighting to Correct Models

When constructing a regression model, recall that we are making a few assumption

1. The error terms ϵ_i are normally distributed with mean 0 and variance $\sigma^2 N(0, \sigma)^2$, and
2. The error terms are independent and uncorrelated.

In this chapter, the objective is to study methods of building regression models when these assumptions are violated.

5.1 Variance Stabilizing Transformations

The assumption of constant various is one of the requirements for the regression model. This assumption is commonaly violated when the response y has the various that is functionally related to its mean. For common distributions and functional relationships, we can summairze their useful variance-stabilizing relationships.

Relationship of σ^2 to $E(y)$	Transfromation
$\sigma^2 \propto$ constant	$y' = y$ (No transformation)
$\sigma^2 \propto E(y)$	$y' = \sqrt{y}$ (Poisson Data)
$\sigma^2 \propto E(y)[1 - E(y)]$	$y' = \sin^{-1}(\sqrt{y})$ (Binomial Data)
$\sigma^2 \propto E(y)^2$	$y' = y^{-1/2}$
$\sigma^2 \propto E(y)^4$	$y' = y^{-1}$

In the case of the Poisson distribution, the variance is equal to the mean. So it would be useful to transformation the data. If $y \sim \text{Poisson}(\lambda)$, then $\sqrt{y}$ is nearly normally disributed with variance approximately $1/4$ if the mean λ is large.

If we have bionmial variable $y \sim \text{Bin}(n, p)$ with mean $m = np$, then we apply the transformation

$$y' = \sin^{-1}\left(\sqrt{\frac{y + c}{n + 2c}}\right)$$

The optimal value of c is $3/8$ when m and $n - m$ are large. The variance is approximately $\frac{1}{4}\left(n + \frac{1}{2}\right)^{-1}$.

5.2 Transformations to Linearize the Model

Another assumption in our regression model is that the relationship between y and the regressors is linear. Sometimes, prior experience or theoretical considerations may indicate that the relationship between y and the regressors is not linear, but may be able to be linearized by using a suitable transformation. These models are known as **intrinsically** or **transformably lineaer**.

Conisder the exponential function,

$$y = \beta_0 e^{\beta_1 x} \epsilon$$

we can transform this model using logarithms to get

$$y' = \ln y = \ln \beta_0 + \beta_1 X + \ln \epsilon = \beta_0' + \beta_1 x + \epsilon'$$

This transformation required that the new error terms $\epsilon' = \ln \epsilon$ still satisfy our assumptions, namely that they are normally and independently distributed with mean 0 and variance σ^2.
Various types of reciprocal transformations can also be used, for example

$$y = \beta_0 + \beta_1 \frac{1}{x} + \epsilon$$

This can be linearized using a **reciprocal transformation** for $x' = \frac{1}{x}$ to give us

$$y = \beta_0 + \beta_1 x' + \epsilon$$

Other models can be linearized by reciprocal transformations such as

$$\frac{1}{y} = \beta_0 + \beta_1 x + \epsilon$$

using the transformation $y' = \frac{1}{y}$, and

$$y = \frac{x}{\beta_0 + \beta_1 x}$$

can be linearized with 2 reciprocal transsformations. First,

$$y' = \frac{1}{y}$$

then

$$x' = \frac{1}{x}$$

This gives us

$$y' = \beta_0 x' + \beta_1$$

5.2.1 Box-Cox Transformations

Another useful class of transformations when the data appears to be non-normal or non-constant variance is the **power transformation** y^λ, where λ is a parameter to be determined. Box and Cox show how the parameters of the regression model and λ can be estimated simultaneously using the method of maximum likelihood. The appropriate procedure is to use

$$y^{(\lambda)} = \begin{cases} \frac{y^\lambda - 1}{\lambda \dot{y}^{\lambda-1}} & \lambda \neq 0 \\ \dot{y} \ln y & \lambda = 0 \end{cases}$$

Where $\dot{y}$ is the geometric mean of the observations,

$$\dot{y} = \ln^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} \ln y_i\right)$$

Then we fit the model

$$y^{(\lambda)} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

The divisor $\dot{y}^{\lambda-1}$ turns out to be related to the Jacobian of the transformation coverting the response y into $y^{(\lambda)}$. The value of λ is usually determined by trial and error and selecting the value for λ which minimizes the residual sum of squares. We can also construct confidence intervals on λ.

5.3 Generalized and Weighted Least Squares

Linear models that do not satisfy the constant error variance assumption can be fitted by the method of **weighted least squares** to give us constant verious. The idea is to multiply the deviation between the observed and expected value of y_i by a **weight**, w_i, chosen to be inversely proportional to the variance of y_i. In the case of simple linear regression, we'd have

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n} w_i e_i^2 = \sum_{i=1}^{n} w_i (y_i - \beta_0 - \beta_1 x_i)^2$$

The resulting normal equations from the weighted least squares are

$$\hat{\beta}_0 \sum_{i=1}^{n} w_i + \hat{\beta}_i \sum_{i=1}^{n} w_i x_i = \sum_{i=1}^{n} w_i y_i$$

$$\hat{\beta}_0 \sum_{i=1}^{n} w_i x_i + \hat{\beta}_i \sum_{i=1}^{n} w_i x_i^2 = \sum_{i=1}^{n} w_i x_i y_i$$

This model will satisfy that $\text{Var}(\sqrt{w_i}\epsilon_i) = \sigma^2$. We may choose different weights depending on the situation, for example we could choose $w_i = \sqrt{x_i}$, or $w = \sqrt{y}$. Another common approach is to preform the usual regression and to esitmate the variance with the sample variance s_i^2 for each y_i, then

$$w_i = \frac{1}{s_i^2}$$

5.3.1 Weighted Least Squares

We can define the matrix

$$W = \begin{bmatrix} w_1 & 0 & \cdots & 0 \\ 0 & w_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \cdots & w_n \end{bmatrix}$$

Where $\sigma^2 W^{-1}$ is the covariance matrix of ϵ. From the weighted least sqaures normal equations, we get

$$(X'WX)\hat{\boldsymbol{\beta}} = X'WY$$

This is the multiple linear regression version of the same normal equations given in the simple linear regression model, then we can solve it to get

$$\hat{\boldsymbol{\beta}} = (X'WX)^{-1}X'WY$$

$\hat{\boldsymbol{\beta}}$ is the **weighted least-sqaures estimator**. If we multiply each of the observed values for the ith observations (including the intercept) by the sequare root of the weights for the corresponding observations, we transform our data to get

$$X_W = \begin{bmatrix} 1\sqrt{w_1} & x_{11}\sqrt{w_1} & \cdots & x_{1k}\sqrt{w_1} \\ 1\sqrt{w_2} & x_{21}\sqrt{w_2} & \cdots & x_{2k}\sqrt{w_2} \\ \vdots & \vdots & \ddots & \vdots \\ 1\sqrt{w_n} & x_{n1}\sqrt{w_n} & \cdots & x_{nk}\sqrt{w_n} \end{bmatrix}, Y_W = \begin{bmatrix} y_1\sqrt{w_1} \\ y_2\sqrt{w_2} \\ \vdots \\ y_n\sqrt{w_n} \end{bmatrix}$$

The model now becomes

$$Y_W = X_W\boldsymbol{\beta} + \epsilon_W$$

Then the weighted least squares estimate becomes

$$\hat{\boldsymbol{\beta}} = (X'_W X_W)^{-1} X'_W Y_W = (X'WX)^{-1}X'WY$$

We also define the mean square error as

$$MSE_W = \frac{\sum_{i=1}^{n} w_i(y_1 - \hat{y}_i)^2}{n-p} = \frac{\sum_{i=1}^{n} w_i e_i^2}{n-p}$$