MAT 3375 Summary

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.

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

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

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

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

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

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

$$\mathrm{Var}(b_1) = \mathrm{Var}\left(\sum_{i=1}^{n} k_i Y_i\right) = \sigma^2 \sum_{i=1}^{n} k_i^2$$

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,

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

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

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

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) $\sum\limits_{i=1}^{n} e_i = 0$. 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) $\sum\limits_{i=1}^{n} Y_i = \sum\limits_{i=1}^{n}\hat{Y}_i$. 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) $\sum\limits_{i=1}^{n} X_i e_i = 0$. 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

$$H_0 : B_1 = 0,\ H_1 : B_1 \neq 0$$

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

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

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 $\sum\limits_{i=1}^{n}(\hat{y}_i-\bar{y})^2$ the **regression sum of squares**. So we have

$$SST=SSR+SSE$$

The regression sum of squares can also be computed by

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

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.

- $\sum_{i=1}^{n}(Y_i - \bar{Y})^2$ has $n-1$ degrees of freedom since we have 1 constraint on the data that

$$\sum_{i=1}^{n}(Y_i - \bar{Y}) = 0$$

- $b_1^2 \sum_{i=1}^{n}(X_i - \bar{X})^2$ has one degree of freedom because it is a function of b_1
- $\sum_{i=1}^{n}(Y - i - \hat{Y}_i)^2$ has $n-2$ degrees of freedom because it is a function of 2 parameters

2.2.4 Hypothesis Testing

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.

Example.

i	1	2	3	4	5	6	7	8	9	10
Shipment Route (X)	1	0	2	0	3	1	0	1	2	0
Airfreight Breakage (Y)	16	9	17	12	22	13	8	15	19	11

(a) Compute the ANOVA table.

(b) Compute the confidence intervals for the parameters.

(c) Compute a confidence interval for the average response when $X = 1$.

We will first load the data into a data frame in R,

```
X <- c(1,0,2,0,3,1,0,1,2,0)
Y <- c(16,9,17,12,22,13,8,15,19,11)
df <- data.frame(X,Y)
```

Now we can fit a linear model and compute the ANOVA table,

```
model <- lm(formula = df$Y ~ df$X, df)
anova(model)
```

```
## Analysis of Variance Table
##
## Response: df$Y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## df$X       1  160.0   160.0  72.727 2.749e-05 ***
## Residuals  8   17.6     2.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Now we can see our linear model is signficant, and we can compute the confidence intervals for our parameters in R

```
confint(model, level=0.95)
```

```
##                2.5 %    97.5 %
## (Intercept) 8.670370 11.729630
## df$X        2.918388  5.081612
```

And we can compute the confidence interval for our response variable at $X = 1$,

```
model <- lm(formula = df$Y ~ df$X, df)
predict(model, newdata = data.frame(X=1),
        interval = 'confidence', level = 0.95)
```

```
## Warning: 'newdata' had 1 row but variables found have 10 rows
```

```
##     fit      lwr      upr
## 1  14.2 13.11839 15.28161
## 2  10.2  8.67037 11.72963
## 3  18.2 16.67037 19.72963
## 4  10.2  8.67037 11.72963
## 5  22.2 19.78144 24.61856
## 6  14.2 13.11839 15.28161
## 7  10.2  8.67037 11.72963
## 8  14.2 13.11839 15.28161
## 9  18.2 16.67037 19.72963
## 10 10.2  8.67037 11.72963
```