MAT 3375 Summary

Joe Zhang

Fall 2023

1 Introduction

We want to model Y in terms of X. We let $X_1, \ldots, X_p$ be the explanatory variables and Y be the response variable. We want to see how Y changes with $X_1, \ldots, X_p$. The relationship between the explanatory variables and the response variable can also be used for prediction the new value of Y given new value of the explanatory variables. The primary goal in regression is to develop a model that relates the response to the explanatory variables, to test it, and ultimately to use it for inference and prediction.

2 Simple Linear Regression

2.1 The Model

We collect a set of paired data. We plot the n paired data Y_i vs. X_i. If it seems reasonable to fit a straight line to the points, we then postulate the following simple regression model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \tag{1}$$

In the model, ϵ represents an unobserved random error term, β_0 is the intercept, and β_1 is the slope of the line.

Both β_0 and β_1 are labeled parameters. They need to be estimated usually from the observed data.

Alternatively, the model may be expressed in terms of $(X_i - \overline{X})$

$$Y_i = (\beta_0 + \beta_1 \overline{X}) + \beta_1 (X_i - \overline{X}) + \epsilon_i \tag{2}$$

where $\overline{X}$ represents the average of the X_i.

The proposed model is linear in the parameters β_0 and β_1.

The model would still be referred to as linear if instead we had X_i^2 instead of X_i.
(i.e. The model $Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \epsilon_i$ is still linear in the parameters).

2.2 Model Assumptions

We assume the following: The random error terms are uncorrelated, have mean equal to 0, and common variance equal to σ^2. This assumption leads to the following:

- $E[Y_i] = \beta_0 + \beta_1 X_i$

- $Var[Y_i] = \sigma^2$

Caution: A well fitting regression model does not imply causation.

2.3 Least Squares Estimates

We define Q as the sum of square errors

$$\begin{aligned} Q &= \sum_{i=1}^{n} \epsilon_i^2 \\ &= \sum_{i=1}^{n} [Y_i - \beta_0 - \beta_1 X_i]^2 \end{aligned}$$

Then we need to find β_0 and β_1 such that they minimize Q. We do this by differentiating with respect to β_0 and β_1 and then setting the partial derivatives equal to 0. We get that the partial derivatives are:

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

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

By rearranging, we get the following equations:

$$\sum_{i=1}^{n} [Y_i] = n\beta_0 + \beta_1 \sum_{i=1}^{n} X_i$$

$$\sum_{i=1}^{n} [X_i Y_i] = \beta_0 \sum_{i=1}^{n} X_i + \beta_1 \sum_{i=1}^{n} X_i^2$$

Solving the system of linear equations, we let b_0 and b_1 represent the solutions to β_0 and β_1, respectively. We get

$$b_0 = \overline{Y} - b_1 \overline{X} \tag{3}$$

$$b_1 = \frac{\sum (X_i - \overline{X}) Y_i}{\sum (X_i - \overline{X})^2} \tag{4}$$

We can also express the equation of b_1 as

$$b_1 = \sum_{i=1}^{n} k_i Y_i$$

where $k_i = \frac{(X_i - \overline{X})}{\sum (X_i - \overline{X})^2}$

We have the following properties of the k_i:

- $$\sum k_i = 0$$
- $$\sum k_i X_i = 1$$
- $$\sum k_i^2 = \frac{1}{\sum (X_i - \overline{X})^2}$$

To show the properties, we have that

$$\begin{aligned}
\sum k_i &= \frac{\sum (X_i - \overline{X})}{\sum (X_i - \overline{X})^2} \\
&= \frac{(\sum X_i) - n\overline{X}}{\sum (X_i - \overline{X})^2} \\
&= 0
\end{aligned}$$

$$\begin{aligned}
\sum k_i X_i &= \frac{\sum (X_i - \overline{X}) X_i}{\sum (X_i - \overline{X})^2} \\
&= \frac{\sum X_i^2 - \overline{X} \sum X_i}{\sum (X_i - \overline{X})^2} \\
&= \frac{\sum X_i^2 - n\overline{X}}{\sum (X_i - \overline{X})^2} \\
&= 1
\end{aligned}$$

$$\begin{aligned}
\sum k_i^2 &= \frac{\sum (X_i - \overline{X})^2}{(\sum (X_i - \overline{X})^2)^2} \\
&= \frac{1}{\sum (X_i - \overline{X})^2}
\end{aligned}$$

After finding the least squares estimate for β_0 and β_1, which we denote as b_0 and b_1, respectively, the line that fits the data is:

$$\hat{Y} = b_0 + b_1 X \tag{5}$$

Alternatively, we can also have

$$\begin{aligned}
\hat{Y} &= (b_0 + b_1\overline{X}) + b_1(X - \overline{X}) \\
&= \overline{Y} - b_1\overline{X} + b_1\overline{X} + b_1(X - \overline{X}) \\
&= \overline{Y} + b_1(X - \overline{X})
\end{aligned}$$

It is also important to note that the point $(\overline{X}, \overline{Y})$ is on the line.

We can predict Y using X and the line.

2.4 The Gauss-Markov Theorem

The Gauss-Markov Theorem states that the least squares estimators b_0 and b_1 are unbiased and have minimum variance among all unbiased linear estimators.

Recall: An estimator is unbiased if its expected value is the value of its parameter.

To show that b_1 is an unbiased estimator of β_1, we need to show that $E[b_1] = \beta_1$

$$\begin{aligned}
E[b_1] &= \sum k_i E[Y_i] \\
&= \beta_0 \sum k_i + \beta_1 \sum k_i X_i \\
&= \beta_0 \cdot 0 + \beta_1 \cdot 1 \\
&= \beta_1
\end{aligned}$$

To show that b_0 is an unbiased estimator of β_0, we need to show that $E[b_0] = \beta_0$

$$\begin{aligned}
E[b_0] &= E[\overline{Y} - b_1\overline{X}] \\
&= E[\overline{Y}] - E[b_1\overline{X}] \\
&= \frac{1}{n}\sum E[Y_i] - \beta_1\overline{X} \\
&= \frac{1}{n}\sum(\beta_0 + \beta_1 X_i) - \beta_1\overline{X} \\
&= \beta_0 + \beta_1\overline{X} - \beta_1\overline{X} \\
&= \beta_0
\end{aligned}$$

Now, we want to show that b_0 and b_1 have minimum variance among all unbiased linear estimators.

Consider an unbiased estimator for β_1, say, $\hat{\beta_1} = \sum c_i Y_i$, it must satisfy

$$\begin{aligned}
\beta_1 &= E[\hat{\beta_1}] \\
&= \sum c_i E[Y_i] \\
&= \sum c_i[\beta_0 + \beta_1 X_i]
\end{aligned}$$

From this, we must have that $\sum c_i = 0$, $\sum c_i X_i = 1$, and $Var[\hat{\beta_1}] = \sigma^2 \sum c_i^2$.

We set $c_i = k_i + d_i$ for arbitrary d_i. Then we get

$$\begin{aligned}
\sum k_i d_i &= \sum k_i(c_i - k_i) \\
&= [\sum c_i \frac{(X_i - \overline{X})}{\sum(X_i - \overline{X})^2}] - \frac{1}{\sum(X_i - \overline{X})^2} \\
&= [\frac{1}{\sum(X_i - \overline{X})^2} - 0] - \frac{1}{\sum(X_i - \overline{X})^2} \\
&= 0
\end{aligned}$$

If we define the vectors $\mathbf{c}^T = [c_1, c_2, ..., c_n]$, $\mathbf{k}^T = [k_1, k_2, ..., k_n]$, and $\mathbf{d}^T = [d_1, d_2, ..., d_n]$, we get that $\mathbf{k}^T\mathbf{d} = 0$. This shows that $\mathbf{k}$ and $\mathbf{d}$ have inner product 0 and are orthogonal vectors.

Since we have $c_i = k_i + d_i$, we get that $\mathbf{c} = \mathbf{k} + \mathbf{d}$. Since $\mathbf{k}$ and $\mathbf{d}$ are orthogonal, we have that by the Pythagorean theorem, $||\mathbf{c}||^2 = ||\mathbf{k}||^2 + ||\mathbf{d}||^2$. Then, we get that

$$Var[\hat{\beta_1}] = \sigma^2(\sum k_i^2 + \sum d_i^2)$$

The variance is minimized when d_i are all 0. Then $\hat{\beta_1} = b_1$ since $c_i = k_i$.

2.5 Summary of estimates

We may write $\hat{Y} = b_0 + b_1 X$ for the estimated or fitted line, $e_i = Y_i - \hat{Y}_i$ for the estimated ith residual, and we estimate the variance σ^2 by

$$\hat{\sigma^2} = \frac{\sum e_i^2}{n-2}$$

This is also known as the mean square error or MSE.

We have

$$b_1 = \sum k_i Y_i$$

$$\begin{aligned} b_0 &= \overline{Y} - b_1 \overline{X} \\ &= \frac{\sum Y_i}{n} - \overline{X} \sum k_i Y_i \\ &= \sum (\frac{1}{n} - k_i \overline{X}) Y_i \end{aligned}$$

We also have the following propeties of the residuals:

- $\sum e_i = 0$
- $\sum X_i e_i = 0$

To prove the properties, we have:

$$\begin{aligned} \sum e_i &= \sum Y_i - \sum [\overline{Y} + b_1 (X_i - \overline{X})] \\ &= \sum (Y_i - \overline{Y}) \\ &= 0 \end{aligned}$$

$$\begin{aligned} \sum X_i e_i &= \sum X_i Y_i - \overline{Y} \sum X_i - b_1 \sum X_i (X_i - \overline{X}) \\ &= [\sum X_i Y_i - n\overline{Y}\overline{X}] - \frac{\sum X_i Y_i - n\overline{Y}\overline{X}}{\sum (X_i - \overline{X})^2} \sum X_i (X_i - \overline{X}) \\ &= 0 \end{aligned}$$

2.6 The Geometry of Estimation

We let $\boldsymbol{X} = (X_1, ..., X_n)^T$, $\boldsymbol{Y} = (Y_1, ..., Y_n)^T$, $\hat{\boldsymbol{Y}} = (\hat{Y_1}, \hat{Y_2}, ..., \hat{Y_n})^T$

We let $\boldsymbol{e} = (e_1, e_2, ..., e_n)^T$ and $\boldsymbol{1_n} = (1, 1, ..., 1)$. Then we can find that $(\boldsymbol{X} - \overline{X}\boldsymbol{1_n})\boldsymbol{e} = 0$. From this, we know that the vector $\boldsymbol{e}$ is orthogonal to the vectors $\boldsymbol{1_n}$ and $\boldsymbol{X} - \overline{X}\boldsymbol{1_n}$. Since $\hat{\boldsymbol{Y}} = \overline{Y}\boldsymbol{1_n} + b_1(\boldsymbol{X} - \overline{X}\boldsymbol{1_n})$. From this, we get that $\boldsymbol{e}$ is orthogonal to $\hat{\boldsymbol{Y}}$.

Using this, we get the following result:

$$||\boldsymbol{Y}||^2 = ||\hat{\boldsymbol{Y}}||^2 + ||\boldsymbol{e}||^2$$

Since we have that $\hat{\boldsymbol{Y}} = \overline{Y}\boldsymbol{1_n} + b_1(\boldsymbol{X} - \overline{X}\boldsymbol{1_n})$, we get that

$$\begin{aligned} ||\hat{\boldsymbol{Y}}||^2 &= ||\overline{Y}\boldsymbol{1_n}||^2 + ||b_1((X) - \overline{X}\boldsymbol{1_n})||^2 \\ &= \overline{Y}^2\boldsymbol{1_n}^T\boldsymbol{1_n} + b_1^2\sum(X_i - \overline{X})^2 \end{aligned}$$

Then we get that

$$\sum Y_i^2 = n\overline{Y}^2 + b_1^2\sum(X_i - \overline{X})^2 + \sum(Y_i - \hat{Y_i})^2$$

From that, we get

$$\sum(Y_i - \overline{Y})^2 = b_1^2\sum(X_i - \overline{X})^2 + \sum(Y_i - \hat{Y_i})^2 \tag{6}$$

We call $\sum(Y_i - \overline{Y})^2$ the total sum of squares, $b_1^2\sum(X_i - \overline{X})^2$ the regression sum of squares, and $\sum(Y_i - \hat{Y_i})^2$ the error sum of squares. This can be used for inferences in regression, which we will talk about in the next section.

2.7 Inference in regression

Remark: If we assume that the random errors $\epsilon_i \sim N(0, \sigma^2)$, then we get that the likelihood function is

$$L(\beta_0, \beta_1, \sigma^2) = (\frac{1}{\sqrt{2\pi}\sigma})^n e^{\frac{1}{2\sigma^2}\sum\epsilon_i^2}$$

Maximizing this function is equivalent to minimizing $Q = \sum epsilon_i^2$, we get the same results for β_0 and β_1.

We can also obtain an estimate for σ^2. It can be estimated by $MSE = \frac{\sum e_i^2}{n-2}$.

Suppose we have the model $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$, where $\epsilon_i \sim N(0, \sigma^2)$ for $i = 1, ..., n$. Then we have

- $\frac{b_1 - \beta_1}{s(b_1)} \sim t_{n-2}$ where $s^2(b_1) = \frac{MSE}{\sum(X_i - \overline{X})^2}$
- $\frac{b_0 - \beta_0}{s(b_0)} \sim t_{n-2}$ where $s^2(b_0) = MSE(\frac{1}{n} + \frac{\overline{X}^2}{\sum(X_i - \overline{X})^2})$
- MSE is an unbiased estimate of σ^2 and $\frac{(n-2)MSE}{\sigma^2} \sim \chi^2_{n-2}$

We can use the properties above to construct confidence intervals for the parameters and test hypotheses. We get that

- $100(1-\alpha)\%$ CI for $\beta_1 : b_1 \pm t_{n-2}(\frac{\alpha}{2})s(b_1)$
- $100(1-\alpha)\%$ CI for $\beta_0 : b_0 \pm t_{n-2}(\frac{\alpha}{2})s(b_0)$

We can also test hypotheses such as $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$ using the test statistic $T = \frac{b_1}{s(b_1)} \sim t_{n-2}$.

2.8 Example for regression

We consider the following example on grade point averages at the end of the freshman year (Y) as a function of the ACT test scores (X).

- We plot the data
- We obtain the least squares estimates
- We plot the estimated regression function and estimate Y when $X = 30$

The R code below will complete the actions

```
data = read.table("/Users/joezhang/Downloads/Grade point average.txt", header = TRUE, sep = '\t')
names(data)
```

```
## [1] "GPA" "ACT"
```

```
GPA = data$GPA
ACT = data$ACT
fit = lm(GPA~ACT, data = data)
fit
```

```
##
## Call:
## lm(formula = GPA ~ ACT, data = data)
##
## Coefficients:
## (Intercept)          ACT
##     2.14596      0.03735
```

The number under (Intercept) is the least squares estimate for β_0 and the number under ACT is the least squares estimate for β_1.

The code below constructs a 95% confidence interval for both β_0 and β_1.

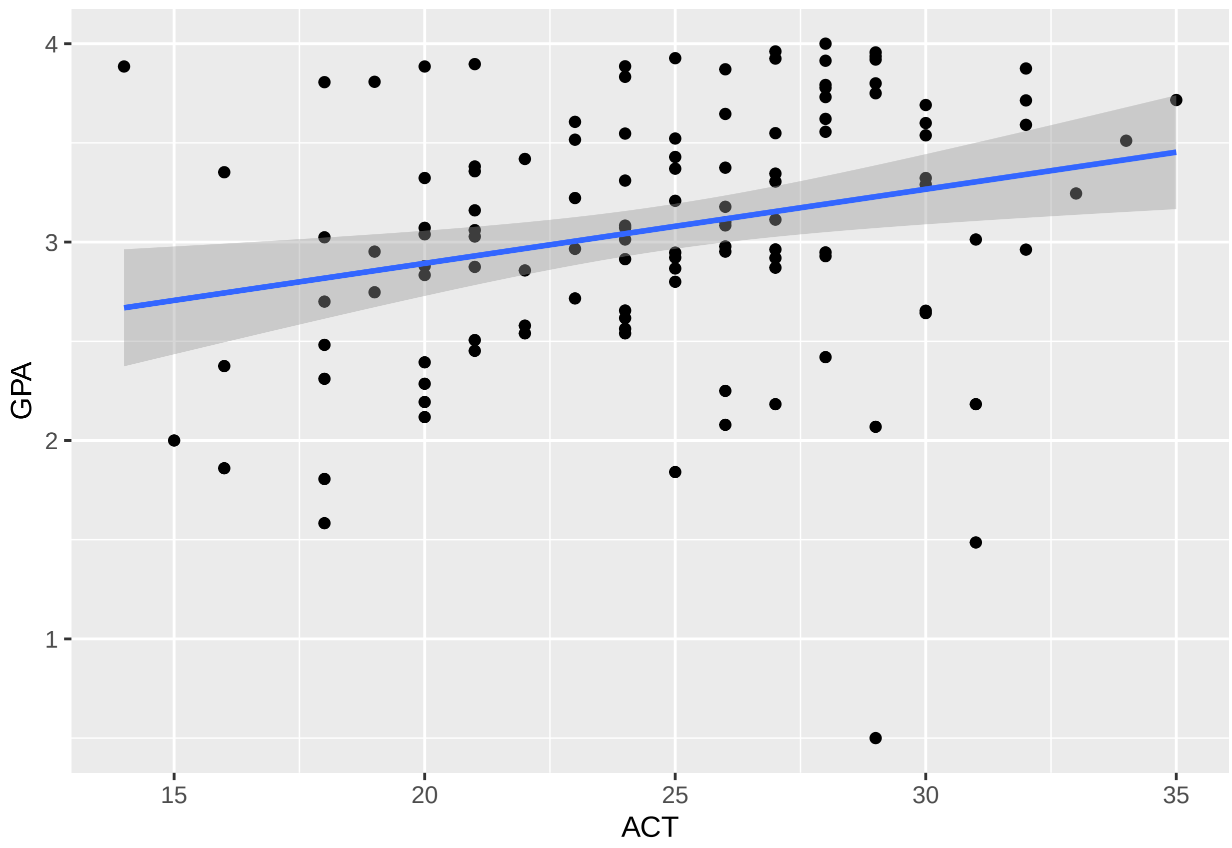
```
confint(fit, level = 0.95)
```

```
##                 2.5 %   97.5 %
## (Intercept) 1.5059161 2.786008
## ACT         0.0118145 0.062880
```

The code below plots the data and also constructs a 95% confidence interval and 95% prediction interval for the average of Y.

```r
library(ggplot2)
ggplot(data, aes(x = ACT, y = GPA)) +
  geom_point()+
  geom_smooth(method = lm, se = TRUE)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

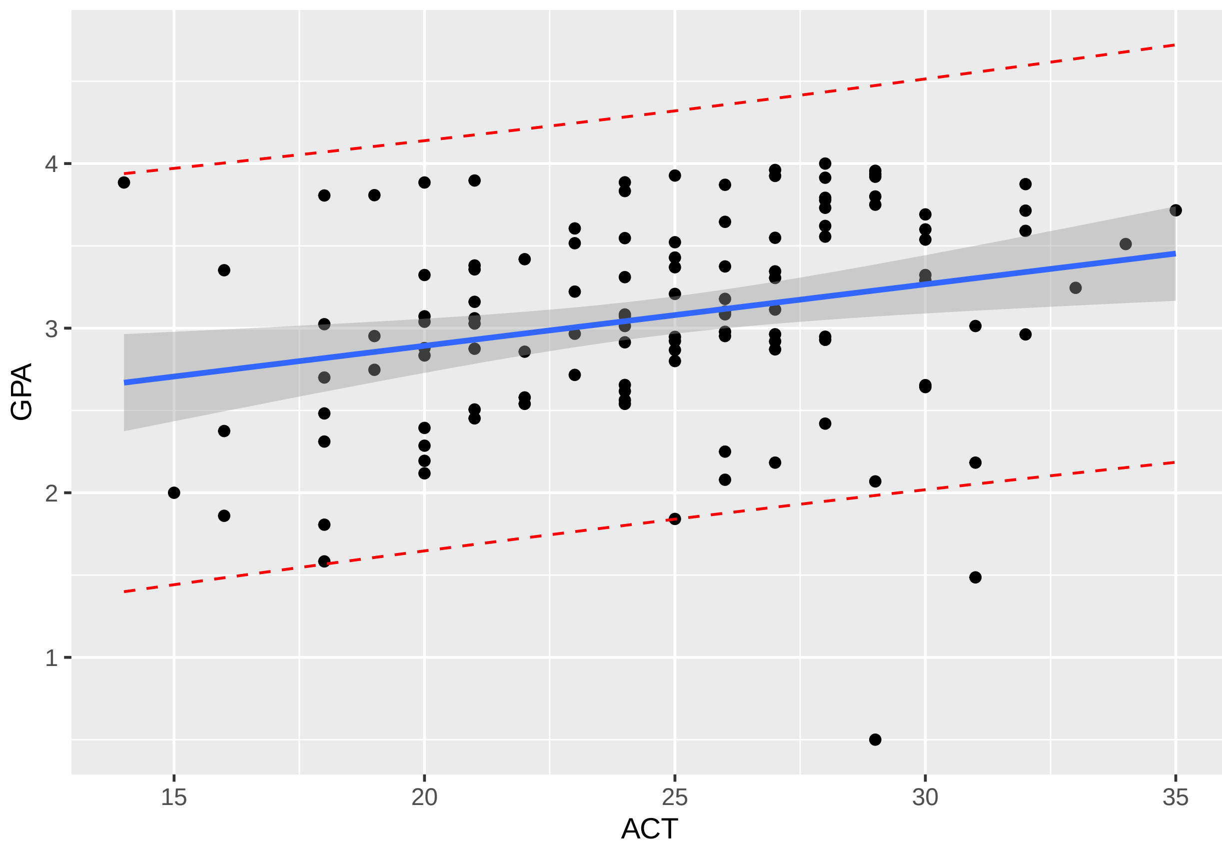



```r
temp_var = predict(fit, interval = 'prediction')
```

```
## Warning in predict.lm(fit, interval = "prediction"): predictions on current data refer to _future_ r
```

```r
new_df = cbind(data, temp_var)
ggplot(new_df, aes(ACT, GPA))+
  geom_point()+
  geom_line(aes(y = lwr), color = 'red', linetype = 'dashed')+
  geom_line(aes(y = upr), color = 'red', linetype = 'dashed')+
  geom_smooth(method = lm, se = TRUE)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



2.9 Analysis of Variance (ANOVA)

Below is the typical format of an analysis of variance (ANOVA) table (for this part, we use $p = 2$):

Table 1: ANOVA Table

Source	Sum of Squares (SS)	df	Mean Square (MS = SS/df)	F statistic	E[MS]
Regression	$SSR = b_1^2 \sum(X_i - \overline{X})^2$	$p-1$	$MSR = \frac{SSR}{p-1}$	$\frac{MSR}{MSE}$	$\sigma^2 + \beta_1^2 \sum(X_i - \overline{X})^2$
Error	$SSE = \sum(Y_i - \hat{Y}_i)^2$	$n-p$	$MSE = \frac{SSE}{n-p}$		σ^2
Total	$SSTO = \sum(Y_i - \overline{Y})^2$	$n-1$			

Each of the sums of squares is a quadratic form where the rank of the corresponding matrix is the degrees of freedom indicated. Cochran's theorem applies and we conclude that the quadratic forms are independent and have Chi-Square distributions. It is well known that the ratio of the two independent Chi-Square divided by their degrees of freedom has a F-distribution (To be seen in section 3 of the notes).

We get that

- $\frac{SSR}{\sigma^2} \sim \chi^2(p-1)$
- $\frac{SSE}{\sigma^2} \sim \chi^2(n-p)$

Then, we get that the F statistic is

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

The degrees of freedom are determined by how much data is required to calculate a particular expression.

$\sum(Y_i - \overline{Y})^2$ has $n-1$ degrees of freedom because of the constraints that $\sum(Y_i - \overline{Y}) = 0$

$b_1^2 \sum(X_i - \overline{X})^2$ has one degree of freedom because it is a function of b_1

$\sum(Y_i - \hat{Y}_i)^2$ has $n-2$ degrees of freedom because it is a function of two parameters.

We'll prove all these using matrices in section 3.

2.10 Testing with ANOVA table

We can use the ANOVA table to test the hypotheses $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$. The null hypothesis states 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 separate independent estimates of the variance σ^2. Hence, if the null hypothesis is true, the F-ratio should be small. On the other hand, if the alternative hypothesis H_1 is true, then the numerator of the F ration will be expected to be large. Consequently, large values of the F statistic are consistent with the alternative. We reject the null hypothesis for large values of F.

In other words, under the null hypothesis, we have that $E[MSR] = \sigma^2$ and $E[MSE] = \sigma^2$. Then the F ratio $F = \frac{MSR}{MSE}$ would be close to 1. Under the alternative hypothesis, $E[MSE] = \sigma^2$. However, $E[MSR] = \sigma^2 + \beta_1^2 \sum(X_i - \overline{X})^2$ since $\beta_1 \neq 0$. Therefore, the F ratio is expected to be large. This is why we reject H_0 for large values of the F ratio.

2.11 Back to GPA data

If we consider the GPA data, we can construct an ANOVA table. We do this using R.

```
anova(fit)
```

```
## Analysis of Variance Table
##
## Response: GPA
##            Df Sum Sq Mean Sq F value Pr(>F)
## ACT         1  3.264  3.2642  8.3917 0.0045 **
## Residuals 117 45.510  0.3890
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This shows that the F value is large and the p-value is small. We can reject H_0 in this case. This means that there is convincing evidence that the slope is not 0 and there is a relationship between the ACT score and GPA.

Now, we want to construct a 95% confidence interval for β_0 and β_1 for the GPA data using the data summary.

```
summary(fit)
```

```
##
## Call:
## lm(formula = GPA ~ ACT, data = data)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7290 -0.3524  0.0407  0.4362  1.2162
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.14596    0.32318   6.640 1.03e-09 ***
## ACT          0.03735    0.01289   2.897   0.0045 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6237 on 117 degrees of freedom
## Multiple R-squared:  0.06692,    Adjusted R-squared:  0.05895
## F-statistic: 8.392 on 1 and 117 DF,  p-value: 0.0045
```

We get that a confidence interval for β_0 can be calculated the following way:

CI for β_0: $b_0 \pm t_{\alpha/2}117 \cdot s(b_0) = 2.14596 \pm 1.98(0.32318) = (1.5059, 2.7860)$

We get that a confidence interval for β_1 can be calculated the following way:

CI for β_1: $b_1 \pm t_{\alpha/2}117 \cdot s(b_1) = 0.03735 \pm 1.98(0.01289) = (0.01181, 0.06288)$

We can do hypothesis testing using t statistics on both β_0 and β_1.

If we test $H_0 : \beta_0 = 0$ versus $H_1 : \beta_0 \neq 0$, we can use the R output and we find that $t = 6.640$, which is significant. We can then reject H_0. Similar with β_1.

However, if we want to test $H_0 : \beta_0 = \beta_{0_1}$ versus $H_1 : \beta_0 \neq \beta_{0_1}$ for some $\beta_{0_1} \neq 0$, then we can't use R. We have to use the test statistic $t = \frac{b_0 - \beta_{0_1}}{s(b_0)} \sim t_{n-2}$ to test and this cannot be computed using R. Similar for β_1.

2.12 Confidence Interval for mean of Y for a given X

We want to construct a confidence interval for the mean of Y^* at a given X^*, or $E[Y^*]$.

To estimate $E[Y^*]$, we know that $E[Y^*] = \beta_0 + \beta_1 X^*$. We can estimate $E[Y^*]$ by

$$\hat{Y}^* = b_0 + b_1 X^* = \sum(\frac{1}{n} + k_i(X^* - \overline{X}))Y_i$$

for a given value of X^*. The estimator is unbiased and has a normal distribution.

We also get that

$$\begin{aligned}
Var[\hat{Y}^*] &= \sigma^2 \sum(\frac{1}{n} + k_i(X^* - \overline{X}))^2 \\
&= \sigma^2 \sum((\frac{1}{n})^2 + k_i^2(X^* - \overline{X}) + 2(\frac{1}{n})k_i(X^* - \overline{X})) \\
&= \sigma^2(\frac{1}{n} + \frac{(X^* - \overline{X})^2}{\sum(X_i - \overline{X})^2})
\end{aligned}$$

The variance of $\hat{Y}^*$ can be estimated by $s^2[\hat{Y}^*] = MSE(\frac{1}{n} + \frac{(X^* - \overline{X})^2}{\sum(X_i - \overline{X})^2})$

We can then use the fact that $\frac{\hat{Y}^* - E[Y^*]}{s[\hat{Y}^*]} \sim t_{n-2}$ to make inference on $E[Y]$. We can then construct a $100(1-\alpha)\%$ confidence interval for $E[Y^*]$ by $\hat{Y}^* \pm t_{\alpha/2, t-2} s[\hat{Y}^*]$.

The width of the confidence interval is different at different values of X^*. In fact, the interval is the narrowest at $X^* = \overline{X}$ and gets wider as it deviates from $\overline{X}$.

2.13 Prediction Interval for Y for a given X

For prediction, we want to find a confidence interval for a new value of Y^* for a given X^*.

Note: Alvo's explanations don't make sense. I used the textbook, internet resources, and Boily's notes to make this section. Please let me know if there's anything I need to correct.

We consider the random variable $Y^* - \hat{Y}^*$ for a given X^*. We can use this to make inferences on the predicted value of Y^*.

We have that $E[Y^* - \hat{Y}^*] = 0$. To show this, we have that

$$\begin{aligned} E[Y^* - \hat{Y}^*] &= E[Y^*] - E[\hat{Y}^*] \\ &= \beta_0 + \beta_1 X^* - E[b_0 + b_1 X^*] \\ &= \beta_0 + \beta_1 X^* - E[b_0] - E[b_1]X^* \\ &= \beta_0 + \beta_1 X^* - \beta_0 - \beta_1 X^* \\ &= 0 \end{aligned}$$

We also have that $Var[Y^* - \hat{Y}^*] = \sigma^2(1 + \frac{1}{n} + \frac{X^* - \overline{X}}{\sum(X_i - \overline{X})^2})$. To show this, we have

$$\begin{aligned} Var[Y^* - \hat{Y}^*] &= Var[Y^*] + Var[\hat{Y}^*] \\ &= \sigma^2 + \sigma^2(\frac{1}{n} + \frac{(X^* - \overline{X})}{\sum(X_i - \overline{X})}) \\ &= \sigma^2(1 + \frac{1}{n} + \frac{(X^* - \overline{X})^2}{\sum(X_i - \overline{X})^2}) \end{aligned}$$

Then we have that $Y^* - \hat{Y}^* \sim N(0, \sigma^2(1 + \frac{1}{n} + \frac{(X^* - \overline{X})}{\sum(X_i - \overline{X})}))$

We estimate the variance of $Y^* - \hat{Y}^*$ by

$$s^2[Y^* - \hat{Y}^*] = MSE(1 + \frac{1}{n} + \frac{(X^* - \overline{X})^2}{\sum(X_i - \overline{X})^2})$$

Then we get that

$$\frac{(Y^* - \hat{Y}^*) - 0}{s[Y^* - \hat{Y}^*]} \sim t_{n-2}$$

Then we can construct a prediction interval for Y^*. The prediction interval is $\hat{Y}^* \pm s(Y^* - \hat{Y}^*)$

2.14 Example: Airfreight Data

```
data = read.table("/System/Volumes/Data/MAT 3375/Summary Sheet/Airfreight Data.txt", header=TRUE, sep =
kable(data)
```

Shipment.Route	Airfreight.breakage
1	16
0	9
2	17
0	12
3	22
1	13
0	8
1	15
2	19
0	11

a. Compute the ANOVA table.
b. Compute confidence intervals for the parameters.
c. Compute a confidence interval for the average response when X = 1.

To compute an ANOVA, table, we simply use the r command

```
x = data$Shipment.Route
y = data$Airfreight.breakage
fit = lm(y~x)
anova(fit)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## 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
```

We conclude that the regression is highly significant since the F statistic has a value of 72.73.

We now want to compute a confidence interval for the coefficients, we do this using the following R command:

```
summary(fit)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##   -2.2   -1.2    0.3    0.8    1.8
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  10.2000     0.6633  15.377 3.18e-07 ***
## x             4.0000     0.4690   8.528 2.75e-05 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.483 on 8 degrees of freedom
## Multiple R-squared:  0.9009, Adjusted R-squared:  0.8885
## F-statistic: 72.73 on 1 and 8 DF,  p-value: 2.749e-05
```

We get that for β_0, a $100(1-\alpha)\%$ confidence interval is $10.2000 \pm t_{\alpha/2,8} \cdot 0.6633$. For β_1, a $100(1-\alpha)\%$ confidence interval is $4.0000 \pm t_{\alpha/2,8} \cdot 0.4690$. In addition, we get that $\hat{\sigma^2} = 2.2$ on 8 degrees of freedom.

To compute a 95% confidence interval for the average response when $X = 1$, we can use the following R commands:

```r
new.dat = data.frame(x=1)
predict(fit, newdata=new.dat, interval="confidence")
```

```
##    fit      lwr      upr
## 1 14.2 13.11839 15.28161
```

To compute a 95% prediction interval for Y at $X = 1$, we can use the following R commands:

```r
new.dat = data.frame(x=1)
predict(fit, newdata=new.dat, interval='prediction')
```

```
##    fit     lwr     upr
## 1 14.2 10.6127 17.7873
```

2.15 Correlation Coefficient

The sample correlation coefficient is defined the following way:

$$r = \frac{\sum(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum(X_i - \overline{X})^2 \sum(Y_i - \overline{Y})}} \tag{7}$$

The correlation coefficient is related to b_1. We can rewrite the equation as

$$r = b_1 (\frac{\sum(X_i - \overline{X})^2}{\sum(Y_i - \overline{Y})^2})^{\frac{1}{2}}$$

The population correlation coefficient is denoted by ρ. It is

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

We use r to estimate ρ.

Under $H_0 : \rho = 0$, we have that

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

We can perform a test for ρ using the R command:

```r
x = p2.10$sysbp
y = p2.10$weight
cor.test(x, y, NULL, method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  x and y
## t = 5.9786, df = 24, p-value = 3.591e-06
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.5513214 0.8932215
## sample estimates:
##       cor
## 0.7734903
```

If we test $H_0 : \rho = \rho_0$, then we use the following fact to make inference:

$$Z = arctanh(r) = \frac{1}{2} \cdot ln(\frac{1+r}{1-r}) \sim N(arctanh(\rho), \frac{1}{n-3})$$

So if we want to test the hypothesis, we use the test statistic: $Z = (arctanh(r) - arctanh(\rho_0))\sqrt{n-3}$.

We reject H_0 for large values of the test statistic.

To compute a confidence interval of ρ, we use the following formula: $[\tanh(arctanh(r) - z_{\alpha/2}(n-3)^{\frac{1}{2}}), tanh(arctanh(r) + z_{\alpha/2}(n-3)^{\frac{1}{2}})]$

3 Matrix Approach to Regression

3.1 Matrix Notations

If we let $\boldsymbol{Y} = [Y_1, ..., Y_n]^T$ be the transpose of the column data vector, then we define the expectation by $\boldsymbol{E}[\boldsymbol{Y}] = [E[Y_1], ..., E[Y_n]]^T$.

Proposition: If $\boldsymbol{Z} = \boldsymbol{A}\boldsymbol{Y} + \boldsymbol{B}$ for some matrix of constants $\boldsymbol{A}$, $\boldsymbol{B}$, then we have $\boldsymbol{E}[\boldsymbol{Z}] = \boldsymbol{A}\boldsymbol{E}[\boldsymbol{Y}] + \boldsymbol{B}$.

To prove this, we let $\boldsymbol{Z} = [Z_1, ..., Z_n]^T$, a_{ij} be the element of the matrix $\boldsymbol{A}$ in the i-th row and j-th column. Let $\boldsymbol{B} = [b_1, ..., b_n]$. Then we get

$$\begin{aligned} E[Z_i] &= E\{[\sum_j a_{ij} Y_j + b_i]\} \\ &= [\sum_j a_{ij} E[Y_j]] + b_i \end{aligned}$$

We define the covariance of $\boldsymbol{Y}$, or the variance-covariance matrix of Y, denoted by $Cov[\boldsymbol{Y}]$, by

$$Cov[\boldsymbol{Y}] = E\{[\boldsymbol{Y} - E[\boldsymbol{Y}]][\boldsymbol{Y} - E[\boldsymbol{Y}]]^T\}$$

. We denote this by $\boldsymbol{\Sigma}$

We have the following property of the variance-covariance matrix:

$$Cov[\boldsymbol{AY}] = \boldsymbol{A\Sigma A^T}$$

where $\boldsymbol{\Sigma}$ is the variance-covariance matrix of $\boldsymbol{Y}$

To prove this, we have that

$$\begin{aligned}
Cov[\boldsymbol{AY}] &= E\{[\boldsymbol{AY} - \boldsymbol{E}[\boldsymbol{AY}]][\boldsymbol{AY} - \boldsymbol{E}[\boldsymbol{AY}]]^T\} \\
&= E\{[\boldsymbol{AY} - \boldsymbol{AE}[\mathbf{Y}]][\boldsymbol{AY} - \boldsymbol{AE}[\mathbf{Y}]]^T\} \\
&= E\{[\boldsymbol{A}[\boldsymbol{Y} - \boldsymbol{E}[\mathbf{Y}]][\boldsymbol{Y} - \boldsymbol{E}[\mathbf{Y}]]^{\boldsymbol{T}}\boldsymbol{A^T}\} \\
&= \boldsymbol{A}E\{[\boldsymbol{Y} - \boldsymbol{E}[\mathbf{Y}]][\boldsymbol{Y} - \boldsymbol{E}[\mathbf{Y}]]^{\boldsymbol{T}}\}\boldsymbol{A^T} \\
&= \boldsymbol{A\Sigma A^T}
\end{aligned}$$

3.2 Multivariate Normal Distribution

A random vector $\boldsymbol{Y}$ has a multivariate normal distribution if its density is given by

$$f(y_1, \dots, y_n) = \frac{|\boldsymbol{\Sigma}|^{\frac{1}{2}}}{(2\pi)^{\frac{n}{2}}} \cdot exp(-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}))$$

where $\boldsymbol{y} = [y_1, ..., y_n]^T$, $\boldsymbol{\mu} = [\mu_1, ..., \mu_n]$, and $\boldsymbol{\Sigma} = Cov[\boldsymbol{Y}]$. We denote this by $Y \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

If we consider the special case where $n = 1$, we have that $\boldsymbol{\Sigma} = \sigma^2$ and $|\boldsymbol{\Sigma}|^{\frac{1}{2}} = \frac{1}{\sigma}$. Then the density function is

$$f(y_1) = \frac{1}{\sigma\sqrt{2\pi}} \cdot exp(-\frac{1}{2}\frac{(y_1 - \mu_1)^2}{\sigma^2})$$

we get back the univariate normal distribution.

Theorem: Let $\boldsymbol{Y} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Let $\boldsymbol{A}$ be an arbitrary $p \times n$ matrix of constants, then we have that

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

Now, if we consider an example where we let $\boldsymbol{Y} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and we let $\boldsymbol{A} = [1, ..., 1]^T$, then we have that

$$\boldsymbol{AY} \sim N_1(\boldsymbol{A\mu}, A\Sigma A^T)$$

where $\boldsymbol{A\mu} = \sum_{i=1}^n \mu_i$, $\boldsymbol{A\Sigma A^T} = \sum \sigma_j^2 + 2\sum_{i \neq j} \sigma_{ij}$.

3.3 Matrix Approach to Linear Regression

If we use the matrix representation in regression, it makes it easy to generalize to fitting several independent variables. This would go beyond 1 independent variable. This approach is also known as Multiple Linear Regression.

We use vectors and matrices to denote the observations of the independent variables, the dependent variable, the coefficients, and the random term.

- We let $\boldsymbol{Y} = \begin{bmatrix} Y_1 & \dots & Y_n \end{bmatrix}^T$ be the transpose of the column vector of obervations of the dependent variable

- We let $\boldsymbol{\beta} = \begin{bmatrix}\beta_1 & \dots & \beta_n\end{bmatrix}^T$ be the transpose of the column vector of coefficients
- We let $\boldsymbol{\epsilon} = \begin{bmatrix}\epsilon_1 & \dots & \epsilon_n\end{bmatrix}^T$ be the transpose of the column vectors of the error terms
- We let $\boldsymbol{X} = \begin{pmatrix} 1 & X_{1,1} & \dots & X_{1,p-1} \\ 1 & X_{2,1} & \dots & X_{2,p-1} \\ \dots & \dots & \dots & \dots \\ 1 & X_{n,1} & \dots & X_{n,p-1} \end{pmatrix}$ be the matrix which incorporates the $p-1$ explanatory variables.

If $\boldsymbol{\epsilon} \sim N_n(0, \sigma^2 \boldsymbol{I_n})$, then the regression model may be expressed as

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon} \sim N_n(\boldsymbol{X\beta}, \sigma^2 \boldsymbol{I_n})$$

where $\boldsymbol{I_n}$ is the $n \times n$ identity matrix and N_n is the multivariate normal distribution.

The above is the same as saying that if $\epsilon_i \sim N(0, \sigma^2)$ for $i = 1, ..., n$, then we have that

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \dots + \beta_{p-1} X_{i,p-1} + \epsilon_i \sim N(\beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \dots + \beta_{p-1} X_{i,p-1}, \sigma^2)$$

for $i = 1, ..., n$.

The matrix approach is much nicer because it is more compact and it's can compute more values easily.

3.4 Least Squares Estimations

We want to find an estimate for the vector $\boldsymbol{\beta}$. To do this, we use the least squares approach. However, we're no longer using just scalars. We're instead dealing with vectors and matrices. We need formulas to take derivatives. Below are some facts for taking derivatives in matrix notation.

- If $z = \boldsymbol{a}^T \boldsymbol{y}$, then we have $\frac{\partial z}{\partial \boldsymbol{y}} = \boldsymbol{a}$
- If $z = \boldsymbol{y}^T \boldsymbol{y}$, then we have $\frac{\partial z}{\partial \boldsymbol{y}} = 2\boldsymbol{y}$
- If $z = \boldsymbol{a}^T \boldsymbol{A} \boldsymbol{y}$, then we have $\frac{\partial z}{\partial y} = \boldsymbol{A}^T \boldsymbol{a}$
- If $z = \boldsymbol{y}^T \boldsymbol{A} \boldsymbol{y}$, then we have $\frac{\partial z}{\partial \boldsymbol{y}} = \boldsymbol{A}^T \boldsymbol{y} + \boldsymbol{A} \boldsymbol{y}$
- If $z = \boldsymbol{y}^T \boldsymbol{A} \boldsymbol{y}$, and $\boldsymbol{A}$ is symmetric, then we have $\frac{\partial z}{\partial \boldsymbol{y}} = 2\boldsymbol{A}^T \boldsymbol{y}$

Using the derivative formulas above, we can derive the least squares estimate of the vector $\boldsymbol{\beta}$. To do this, we need to minimize the function

$$\begin{aligned} Q &= \boldsymbol{\epsilon}^T \boldsymbol{\epsilon} \\ &= \sum_{i=1}^{n} \epsilon_i^2 \\ &= (\boldsymbol{Y} - \boldsymbol{X}\beta)^T (\boldsymbol{Y} - \boldsymbol{X}\beta) \end{aligned}$$

We can differentiate Q and then obtain the estimate for $\boldsymbol{\beta}$. If we differentiate Q, we get

$$\frac{\partial Q}{\partial \boldsymbol{\beta}} = -2\boldsymbol{X}^T (\boldsymbol{Y} - \boldsymbol{X}\beta)$$

. We then set the equation to 0. Then after we solve the equation, we get that a solution for $\boldsymbol{\beta}$ is $\boldsymbol{b} = \boldsymbol{(X^TX)^{-1}X^TY}$.

Therefore, the least squares estimate for $\boldsymbol{\beta}$ is $\boldsymbol{b} = \boldsymbol{(X^TX)^{-1}X^TY}$ if the matrix $(\boldsymbol{X^TX})^{-1}$ exists.

We have that $\boldsymbol{b} = \boldsymbol{(X^TX)^{-1}X^TY}$ is an unbiased estimator of $\boldsymbol{\beta}$. To prove this, we have that

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

This means that the least squares estimates of all the parameters are unbiased estimators of their respective parameters.

Now, we want to find the variance-covariance matrix of $\boldsymbol{b}$.

If we let $\boldsymbol{b} = \boldsymbol{AY}$, where $\boldsymbol{A} = \boldsymbol{(X^TX)^{-1}X^T}$. Then we get

$$\begin{aligned} Cov(\boldsymbol{b}) &= \boldsymbol{A\Sigma A} \\ &= \sigma^2\boldsymbol{AA^T} \\ &= \sigma^2(\boldsymbol{(X^TX)^{-1}(X^TX)(X^TX)^{-1}}) \\ &= \sigma^2(\boldsymbol{X^TX})^{-1} \end{aligned}$$

We get that $Cov(\boldsymbol{b}) = \sigma^2(\boldsymbol{X^TX})^{-1}$. We have therefore computed the variances of all the least-squares estimates of the parameters and the covariances between them. This is the nice thing about matrix notation, we can compute more values in one shot.

Now that we have computed the expectation and variance of $\boldsymbol{b}$, we can now determine its distribution. We get that $\boldsymbol{b} \sim N_p(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X^TX})^{-1})$.

3.5 The Hat Matrix and its Properties

The predicted value of $\boldsymbol{Y}$ is written as

$$\boldsymbol{\hat{Y}} = \boldsymbol{Xb}$$

. We can rewrite the equation as

$$\boldsymbol{\hat{Y}} = \boldsymbol{HY}$$

where $\boldsymbol{H} = \boldsymbol{X(X^TX)^{-1}X}$. We call $\boldsymbol{H}$ the "hat" matrix.

We have that the hat matrix $\boldsymbol{H}$ is a projection matrix onto the estimation space. It projects $\boldsymbol{Y}$ onto the estimation space, leading to $\boldsymbol{\hat{Y}} = \boldsymbol{HY}$. The hat matrix is also idempotent. To show this, we have that

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

The hat matrix is also symmetric, which means that $\boldsymbol{H^T} = \boldsymbol{H}$. To show this, we have

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

We also have that the matrix $(\boldsymbol{I} - \boldsymbol{H})$ is idempotent ($\boldsymbol{I}$ is the identity matrix). To show this, we have

$$
\begin{aligned}
(\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{I} - \boldsymbol{H}) &= \boldsymbol{II} - \boldsymbol{IH} - \boldsymbol{HI} + \boldsymbol{HH} \\
&= \boldsymbol{I} - \boldsymbol{H} - \boldsymbol{H} + \boldsymbol{H} \\
&= \boldsymbol{I} - \boldsymbol{H}
\end{aligned}
$$

We have that the matrix $\boldsymbol{H}$ and the matrix $\boldsymbol{I} - \boldsymbol{H}$ are orthogonal. To show this, we have

$$
\begin{aligned}
\boldsymbol{H}(\boldsymbol{I} - \boldsymbol{H}) &= \boldsymbol{HI} - \boldsymbol{HH} \\
&= \boldsymbol{H} - \boldsymbol{H} \\
&= \boldsymbol{0}
\end{aligned}
$$

We can express the residual vector as $\boldsymbol{e} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$. To show this, we have

$$
\begin{aligned}
\boldsymbol{e} &= \boldsymbol{Y} - \boldsymbol{\hat{Y}} \\
&= \boldsymbol{Y} - \boldsymbol{HY} \\
&= (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}
\end{aligned}
$$

Putting all the properties together, we have that $\boldsymbol{\hat{Y}} = \boldsymbol{HY}$, $\boldsymbol{e} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$, and $\boldsymbol{Y} = \boldsymbol{HY} + (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$. We get that by the Pythagorean theorem, we have

$$||\boldsymbol{Y}||^2 = ||\boldsymbol{HY}||^2 + ||(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}||^2$$

We also get that $Cov[\boldsymbol{e}] = \sigma^2(\boldsymbol{I} - \boldsymbol{H})$, which is estimated by $s^2[\boldsymbol{e}] = MSE(\boldsymbol{I} - \boldsymbol{H})$.

Now, we want to consider the special case where $p = 2$. This is the case with 1 predictor variable, which goes back to simple linear regression. We want to compute the hat matrix for this case.

We let $\boldsymbol{X} = \begin{pmatrix} 1 & (X_1 - \overline{X}) \\ \dots & \dots \\ 1 & (X_n - \overline{X}) \end{pmatrix}$. Then we have that $\boldsymbol{X^T X} = \begin{pmatrix} n & 0 \\ 0 & \sum(X_i - \overline{X})^2 \end{pmatrix}$. Then we get that $(\boldsymbol{X^T X})^{-1} = \begin{pmatrix} \sum(X_i - \overline{X})^2 & 0 \\ 0 & n \end{pmatrix} \frac{1}{n\sum(X_i - \overline{X})}$.

Now, we can compute the hat matrix.

$$
\begin{aligned}
\boldsymbol{H} &= (\boldsymbol{X^T X})^{-1}\boldsymbol{X^T} \\
&= \begin{pmatrix} \sum(X_i - \overline{X})^2 + n(X_1 - \overline{X})^2 & \dots & \sum(X_i - \overline{X})^2 + n(X_1 - \overline{X})^2 + n(X_1 - \overline{X})(X_n - \overline{X}) \\ \dots & \dots & \dots \\ \sum(X_i - \overline{X})^2 + n(X_1 - \overline{X})(X_n - \overline{X}) & \dots & \sum(X_i - \overline{X})^2 + n(X_n - \overline{X})^2 \end{pmatrix} \cdot \frac{1}{n\sum(X_i - \overline{X})^2} \\
&= \begin{pmatrix} \frac{1}{n} & \dots & \frac{1}{n} \\ \dots & \dots & \dots \\ \frac{1}{n} & \dots & \frac{1}{n} \end{pmatrix} + \begin{bmatrix} X_1 - \overline{X} \\ \dots \\ X_n - \overline{X} \end{bmatrix} \begin{bmatrix} X_1 - \overline{X} & \dots & X_n - \overline{X} \end{bmatrix} \frac{1}{\sum(X_i - \overline{X})^2} \\
&= \frac{1}{n}\boldsymbol{J} + \begin{bmatrix} X_1 - \overline{X} \\ \dots \\ X_n - \overline{X} \end{bmatrix} \begin{bmatrix} k_1 & \dots & k_n \end{bmatrix}
\end{aligned}
$$

Note: $\boldsymbol{J}$ is a matrix of 1s.

Now that we have computed the hat matrix for 2 predictor variables, we can compute the least squares regression line in matrix form.

$$\begin{aligned}
\hat{\boldsymbol{Y}} &= \boldsymbol{HY} \\
&= \frac{1}{n}\boldsymbol{JY} + \begin{bmatrix} X_1 - \overline{X} \\ \dots \\ X_n - \overline{X} \end{bmatrix} \begin{bmatrix} k_1 & \dots & k_n \end{bmatrix} \boldsymbol{Y} \\
&= \begin{bmatrix} \overline{Y} \\ \dots \\ \overline{Y} \end{bmatrix} + \begin{bmatrix} X_1 - \overline{X} \\ \dots \\ X_n - \overline{X} \end{bmatrix} b_1 \\
&= \overline{Y}\boldsymbol{1_n} + b_1 \begin{bmatrix} X_1 - \overline{X} \\ \dots \\ X_n - \overline{X} \end{bmatrix}
\end{aligned}$$

Now, we can find the trace and rank of the hat matrix $\boldsymbol{H}$ and we show that it is equal to 2.

$$\begin{aligned}
Rank(\boldsymbol{H}) &= Trace(\boldsymbol{H}) \\
&= \frac{n\sum(X_i - \overline{X})^2 + n\sum(X_i - \overline{X})^2}{n\sum(X_i - \overline{X})^2} \\
&= 2
\end{aligned}$$

3.6 Quadratic Forms

We now want to look at the theory behind the relationship between sums of squares. We first need to look at a fundamental concept.

If we let $Y_1, ..., Y_n$ be a random sample from $N(\mu, \sigma^2)$. A quadratic form in the Y's is defined to be the real quantity $\boldsymbol{Q} = \boldsymbol{Y^T A Y}$, where $\boldsymbol{A}$ is a symmetric positive definite matrix. The singular decomposition of $\boldsymbol{A}$ implies that there exists an orthogonal matrix $\boldsymbol{P}$ such that if $\boldsymbol{\Lambda} = (\lambda_i)$ is the diagonal matrix of eigenvalues of $\boldsymbol{A}$, we have $\boldsymbol{A} = \boldsymbol{P^T \Lambda P}$.

Proportion: $E[\boldsymbol{Y^T A Y}] = Trace[\boldsymbol{A\Sigma}] + E[\boldsymbol{Y}]^T \boldsymbol{A} E[\boldsymbol{Y}]$.

To show this, we have

$$\begin{aligned}
\boldsymbol{Y^T A Y} &= \boldsymbol{Y^T P^T \Lambda P Y} \\
&= \boldsymbol{(PY)^T \Lambda (PY)} \\
&= \sum \lambda_i ||(\boldsymbol{PY})_i||^2
\end{aligned}$$

where $(\boldsymbol{PY})_i$ is the i-th element in the vector PY. The second moment of $(\boldsymbol{PY})_i$ is

$$\begin{aligned}
E[||(\boldsymbol{PY})_i||^2] &= Var[||(\boldsymbol{PY})_i||] + (E[(\boldsymbol{PY})_i])^2 \\
&= (\boldsymbol{P\Sigma P^T})_{ii} + [(\boldsymbol{PE}[\boldsymbol{Y}])_i]^2
\end{aligned}$$

Now, we get

$$
\begin{aligned}
E[\sum \lambda_i ||(PY)_i||^2] &= \sum \lambda_i (\boldsymbol{P\Sigma P^T})_{ii} + \sum \lambda_i [(\boldsymbol{PE}[\boldsymbol{Y}])_i]^2 \\
&= Trace(\Lambda \boldsymbol{P\Sigma P^T}) + \boldsymbol{\mu^T A \mu} \\
&= Trace(\boldsymbol{P^T \Lambda P \Sigma}) + \boldsymbol{\mu^T A \mu} \\
&= Trace(\boldsymbol{A\Sigma}) + \boldsymbol{\mu^T A \mu}
\end{aligned}
$$

Lemma: The mean squared error is an unbiased estimate of σ^2.

To prove this, we have that the residual sum of squares (SSE) is

$$\sum e_i^2 = \sum (Y_i - \hat{Y})^2$$

This can be written in matrix notation as

$$(\boldsymbol{Y} - \boldsymbol{\hat{Y}})^T (\boldsymbol{Y} - \boldsymbol{\hat{Y}})$$

We also know for a fact that $\boldsymbol{Y} - \boldsymbol{\hat{Y}} = \boldsymbol{(I - H)Y}$ and $\boldsymbol{I - H}$ is idempotent. We get that

$$(\boldsymbol{Y} - \boldsymbol{\hat{Y}})^T (\boldsymbol{Y} - \boldsymbol{\hat{Y}}) = \boldsymbol{Y^T (I - H) Y}$$

Then we have that

$$
\begin{aligned}
E[\boldsymbol{Y^T (I - H) Y}] &= Trace(\boldsymbol{(I - H)\Sigma}) + \boldsymbol{\mu^T (I - H) \mu} \\
&= \sigma^2 Trace(\boldsymbol{I - H}) + (\boldsymbol{X\beta})^T (\boldsymbol{I - H})(\boldsymbol{X\beta}) \\
&= \sigma^2 (n - p) + \boldsymbol{\beta^T X^T} (\boldsymbol{I} - \boldsymbol{X(X^T X)^{-1} X^T}) \boldsymbol{X\beta} \\
&= \sigma^2 (n - p) + \boldsymbol{\beta^T (X^T - X^T X (X^T X)^{-1} X^T) X\beta} \\
&= \sigma^2 (n - p) + \boldsymbol{\beta^T (X^T - X^T) X\beta} \\
&= \sigma^2 (n - p) + 0 \\
&= \sigma^2 (n - p)
\end{aligned}
$$

Consequently, we get that

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

3.7 Chi-Squared distribution and F distribution

A random variable U has a chi-squared χ^2_ν distribution with ν degrees of freedom if its density is given by

$$f(u; \nu) = \frac{1}{2^{\frac{\nu}{2}} \Gamma(\nu/2)} u^{(\nu/2)-1} e^{-u/2}$$

for $u > 0, \nu > 0$. The mean of U is ν and the variance of U is 2ν.

A random variable U has a non-central chi-squared distribution $\chi^2_\nu(\lambda)$ with ν degrees of freedom and non-centrality parameter λ if its density is given by

$$f(u;\nu,\lambda) = \sum_{i=0}^{\infty} e^{-\lambda/2} \frac{(\lambda/2)^i}{i!} f(u;\nu+2i)$$

with $u > 0, \nu > 0$. The mean of U is $\nu + \lambda$ and the variance of U is $2\nu + 4\lambda$.

If we let $U_1 \sim \chi^2_{\nu_1}$ and $U_2 \sim \chi^2_{\nu_2}$, then we have that

$$F = \frac{U_1/\nu_1}{U_2/\nu_2} \sim F(\nu_1,\nu_2)$$

If the numerator has a non-central chi-squared distribution, then F has a non-central F distribution.

3.8 Cochran's Theorem

Cochran's Theorem states that if we let Y be a random vector with a multivariate normal distribution $N_n(\boldsymbol{\mu}, \sigma^2 \boldsymbol{I})$ and suppose that we have the decomposition

$$\boldsymbol{Y^T Y} = Q_1 + \cdots + Q_k$$

where $Q_i = \boldsymbol{Y^T A_i Y}$ and $rank(A_i) = n_i$. Then $\frac{Q_i}{\sigma^2}$ are independent and have a non-central chi-squared distribution with n_i degrees of freedom and non-centrality parameter λ_i, where $\lambda_i = \boldsymbol{\mu^T A_i \mu}$.

We have some examples of quadratic forms that are particularly important for analysis.

We let $\boldsymbol{Y} \sim N_n(\boldsymbol{\mu}, \sigma^2 \boldsymbol{I_n})$ be the response vector. We can decompose $\boldsymbol{Y^T Y}$ the following way:

$$\boldsymbol{Y^T Y} = \boldsymbol{Y^T A Y} + \boldsymbol{Y^T} \frac{\boldsymbol{1_n 1_n^T}}{n} \boldsymbol{Y}$$

where $\boldsymbol{A} = \begin{pmatrix} 1-\frac{1}{n} & -\frac{1}{n} & \dots & -\frac{1}{n} \\ -\frac{1}{n} & 1-\frac{1}{n} & \dots & -\frac{1}{n} \\ \dots & \dots & \dots & \dots \\ -\frac{1}{n} & -\frac{1}{n} & \dots & 1-\frac{1}{n} \end{pmatrix}$ (an $n \times n$ matrix with $1 - \frac{1}{n}$ on the diagonals and $-\frac{1}{n}$ on the off-diagonals) and $\boldsymbol{1_n} = \begin{bmatrix} 1 \\ 1 \\ \dots \\ 1 \end{bmatrix}$ (n-dimensional column vector with all 1s).

We can rewrite $\boldsymbol{Y^T A Y}$ the following way:

$$
\begin{aligned}
\boldsymbol{Y^T A Y} &= \begin{bmatrix} Y_1 & \dots & Y_n \end{bmatrix} \begin{pmatrix} 1-\frac{1}{n} & -\frac{1}{n} & \dots & \frac{1}{n} \\ -\frac{1}{n} & 1-\frac{1}{n} & \dots & -\frac{1}{n} \\ \dots & \dots & \dots & \dots \\ -\frac{1}{n} & -\frac{1}{n} & \dots & 1-\frac{1}{n} \end{pmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ \dots \\ Y_n \end{bmatrix} \\
&= \begin{bmatrix} Y_1-\overline{Y} & Y_2-\overline{Y} & \dots & Y_n-\overline{Y} \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ \dots \\ Y_n \end{bmatrix} \\
&= Y_1(Y_1-\overline{Y}) + Y_2(Y_2-\overline{Y}) + \dots + Y_n(Y_n-\overline{Y}) \\
&= \sum Y_i(Y_i-\overline{Y}) \\
&= \sum Y_i^2 - \overline{Y}\sum Y_i \\
&= \sum Y_i^2 - n\overline{Y} \\
&= \sum (Y_i-\overline{Y})^2
\end{aligned}
$$

We can also rewrite $\boldsymbol{Y}^T \frac{\boldsymbol{1_n^T 1_n}}{n}\boldsymbol{Y}$ the following way:

$$
\begin{aligned}
\boldsymbol{Y}^T \frac{\boldsymbol{1_n^T 1_n}}{n}\boldsymbol{Y} &= \boldsymbol{Y}^T \frac{\boldsymbol{J_n}}{n}\boldsymbol{Y} \\
&= \begin{bmatrix} Y_1 & Y_2 & \dots & Y_n \end{bmatrix} \begin{pmatrix} \frac{1}{n} & \frac{1}{n} & \dots & \frac{1}{n} \\ \frac{1}{n} & \frac{1}{n} & \dots & \frac{1}{n} \\ \dots & \dots & \dots & \dots \\ \frac{1}{n} & \frac{1}{n} & \dots & \frac{1}{n} \end{pmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ \dots \\ Y_n \end{bmatrix} \\
&= \begin{bmatrix} \overline{Y} & \overline{Y} & \dots & \overline{Y} \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ \dots \\ Y_n \end{bmatrix} \\
&= n\overline{Y}
\end{aligned}
$$

So now we get that $\boldsymbol{Y^T Y} = \sum(Y_i-\overline{Y})^2 + n\overline{Y}$. We can now look at the degrees of freedom of $\sum(Y_i-\overline{Y})^2$ and $n\overline{Y}$.

We get that $\sum(Y_i-\overline{Y})^2 = \boldsymbol{Y^T A Y}$, where $\boldsymbol{A} = \begin{pmatrix} 1-\frac{1}{n} & -\frac{1}{n} & \dots & -\frac{1}{n} \\ -\frac{1}{n} & 1-\frac{1}{n} & \dots & -\frac{1}{n} \\ \dots & \dots & \dots & \dots \\ -\frac{1}{n} & -\frac{1}{n} & \dots & 1-\frac{1}{n} \end{pmatrix}$. We know that A is idempotent and symmetric. Then we get that $rank(A) = trace(A) = n(1-\frac{1}{n}) = n-1$. This explains why $\frac{\sum(Y_i-\overline{Y})^2}{\sigma^2}$ has a chi-squared distribution with $n-1$ degrees of freedom.

We also know that $\frac{\boldsymbol{1_n 1_n^T}}{n}$ is an $n \times n$ matrix with $\frac{1}{n}$ as all its entries. This makes it an idempotent and symmetric matrix. It has $rank = trace = n(\frac{1}{n}) = 1$.

Therefore, we get that the ranks sum up to n and we have proven that $\frac{\sum(Y_i-\overline{Y})^2}{\sigma^2}$ has a chi-squared distribution with $n-1$ degrees of freedom.

4 Multiple Linear Regression

4.1 Linear models with 2 or more predictors

We usually see models with 2 or more predictor variables rather than 1 in the case of simple linear regression. For instance, with 2 predictors, we have models in the form

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon_i$$

with $\epsilon_i \sim N(0, \sigma^2)$. This model displays a plane in 3 dimensions and β_1 represents the rate of change in a unit increase in X_1 when X_2 is fixed and vice versa for β_2.

In a model with $p-1$ predictors, we have that the model is in the form

$$Y_i = \beta_0 + \sum_{i=1}^{p-1} \beta_k X_{ik} + \epsilon_i$$

where β_k is the rate of change in a unit increase in X_k when all other explanatory variables are held fixed.

4.2 Matrix Approach: Review

To make the equation of the linear model more compact, we use the matrix notation discussed in section 2.

- We let $\boldsymbol{Y} = \begin{bmatrix} Y_1 & \dots & Y_n \end{bmatrix}^T$ be the transpose of the column vector of obervations of the dependent variable
- We let $\boldsymbol{\beta} = \begin{bmatrix} \beta_1 & \dots & \beta_n \end{bmatrix}^T$ be the transpose of the column vector of coefficients
- We let $\boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 & \dots & \epsilon_n \end{bmatrix}^T$ be the transpose of the column vectors of the error terms
- We let $\boldsymbol{X} = \begin{pmatrix} 1 & X_{1,1} & \dots & X_{1,p-1} \\ 1 & X_{2,1} & \dots & X_{2,p-1} \\ \dots & \dots & \dots & \dots \\ 1 & X_{n,1} & \dots & X_{n,p-1} \end{pmatrix}$ be the matrix which incorporates the $p-1$ explanatory variables.

Then we can write the model as

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$$

with $\boldsymbol{\epsilon} \sim N_n(0, \sigma^2 \boldsymbol{I_n})$.

Recall that the least squares estimator for $\boldsymbol{\beta}$ is $\boldsymbol{b} = \boldsymbol{(X^T X)^{-1} X^T Y}$ and the fitted values are $\boldsymbol{\hat{Y}} = \boldsymbol{Xb} = \boldsymbol{HY}$, where $\boldsymbol{H} = \boldsymbol{X(X^T X)^{-1} X^T}$.

We also recall that the variance-covariance matrix of the residuals is $Cov(\boldsymbol{e}) = \sigma^2(\boldsymbol{I - H})$, which can be estimated by $s^2[\boldsymbol{e}] = (MSE)(\boldsymbol{I - H})$. We also have $s^2[\boldsymbol{b}] = (MSE)(\boldsymbol{X^T X})^{-1}$.

We can perform ANOVA on multiple regression models. To do this, it is similar to simple linear regression, except that we need to use matrix notation to write the sums of squares. We have that

- $SSTO = \boldsymbol{Y^T Y} - \frac{1}{n}\boldsymbol{Y^T J Y} = \boldsymbol{Y^T}(\boldsymbol{I} - \frac{1}{n}\boldsymbol{J})\boldsymbol{Y}$
- $SSE = \boldsymbol{e^T e} = \boldsymbol{Y^T (I - H) Y}$
- $SSR = \boldsymbol{b^T X^T Y} - \frac{1}{n}\boldsymbol{Y^T J Y} = \boldsymbol{Y^T}(\boldsymbol{H} - \frac{1}{n}\boldsymbol{J})\boldsymbol{Y}$

These values are all in quadratic form. We get that $SSTO = SSR + SSE$. By Cochran's theorem, we get that SSR, SSE, and SSTO have a chi-squared distribution with degrees of freedom $p-1$, $n-p$, and $n-1$, respectively. This can be shown by computing the ranks of the matrices $\boldsymbol{H} - \frac{\boldsymbol{1}}{\boldsymbol{n}}\boldsymbol{J}$ and $\boldsymbol{I} - \boldsymbol{H}$. The ANOVA table is the same as for simple linear regression except that SSR has degree of freedom $p-1$ and SSE has degree of freedom $n-p$. This is not restricted to $p = 2$. Below is the typical structure of an ANOVA table (this is the same as for simple linear regression).

Table 3: ANOVA Table

Source	Sum of Squares (SS)	df	Mean Square (MS = SS/df)	F statistic	E[MS]
Regression	$SSR = b_1^2 \sum(X_i - \overline{X})^2$	$p-1$	$MSR = \frac{SSR}{p-1}$	$\frac{MSR}{MSE}$	$\sigma^2 + \beta_1^2 \sum(X_i - \overline{X})^2$
Error	$SSE = \sum(Y_i - \hat{Y}_i)^2$	$n-p$	$MSE = \frac{SSE}{n-p}$		σ^2
Total	$SSTO = \sum(Y_i - \overline{Y})^2$	$n-1$			

We can use this to test the null hypothesis $H_0 : \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0$ against H_1 : not all $\beta_k = 0$. We do this by looking at the F statistic $F = \frac{MSR}{MSE}$ and we reject H_0 for large values of F.

We can also do hypothesis tests for individual coefficients. Say we test $H_0 : \beta_k = 0$ against $H_1 : \beta_k \neq 0$, then we can compute the test statistic

$$t = \frac{b_k}{s[b_k]} \sim t_{n-p}$$

and we reject H_0 for large values of t. ($s^2[b_k] = MSE(\boldsymbol{X^T X})^{-1}_{kk}$).

4.3 Extra Sums of Squares Principle

We can use a more general approach to regression to test if we can fit a reduced model rather than a full model to model the data. We first illustrate the case for $p = 2$.

To do this, we let the full model (F) be the model $Y = \beta_0 + \beta_1 X + \epsilon$ and the reduced model (R) be the model $Y = \beta_0 + \epsilon$. We compute the error sum of squares for each model. We get that $SSE(F) = \sum(Y_i - \hat{Y}_i)^2$ for the full model and $SSE(R) = \sum(Y_i - \hat{Y}_i)^2$ for the reduced model. This way we can test $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$ by computing the following statistic:

$$F^* = \frac{\frac{SSE(R) - SSE(F)}{df_R - df_F}}{\frac{SSE(F)}{df_F}}$$

We reject H_0 for large values of F^*, which has an F distribution with degrees of freedom $df_R - df_F$ and df_F, respectively. In other words, $F^* \sim F(df_R - df_F, df_F)$

An immediate application of this approach is to the situation where there are repeat observations at the same values of X (i.e. when there are multiple observed Y values at the same X value). Suppose that the full model is given by

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

where $i = 1, \ldots, n_j$ and $j = 1, \ldots, c$ and $\epsilon_{ij} \sum N(0, \sigma^2)$. The μ_j values are unrestricted parameters when $X = X_j$. To derive their least squares estimates, we want to minimize the following: $Q = \sum_{j=1}^{c} \sum_{i=1}^{n_j} \epsilon_{ij}^2 = \sum_{j=1}^{c} \sum_{i=1}^{n_j} [Y_{ij} - \mu_j]^2$. We take the derivative and we get

$$\begin{aligned}\frac{\partial Q}{\partial \mu_j} &= \frac{\partial \sum_{i=1}^{n_j} (Y_{ij} - \mu_j)^2}{\partial \mu_j} \\ &= -2 \sum_{i=1}^{n_j} (Y_{ij} - \mu_j)\end{aligned}$$

We set the above equal to 0. Then we get that

$$
\begin{aligned}
-2\sum_{i=1}^{n_j}(Y_{ij}-\mu_j) &= 0\\
\sum_{i=1}^{n_j}Y_{ij}-\sum_{i=1}^{n_j}\mu_j &= 0\\
n_j\mu_j &= \sum_{i=1}^{n_j}Y_{ij}\\
\hat{\mu} &= \frac{\sum_{i=1}^{n_j}Y_{ij}}{n_j}
\end{aligned}
$$

So we have that the least squares estimators of μ_j are

$$\overline{Y_j} = \frac{\sum_{i=1}^{n_j}Y_{ij}}{n_j}$$

Therefore, the error sum of squares for this full unrestricted model is

$$SSE(F) = \sum_{ij}(Y_{ij}-\overline{Y_j})^2$$

The corresponding degrees of freedom are

$$df_F = \sum_{j=1}^{c}(n_j-1) = n-c$$

If all $n_j = 1$, then $df_F = 0$ and $SSE(F) = 0$, and the analysis cannot proceed any further.

Now, we have the reduced model

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

which has error sum of squares equal to

$$SSE(R) = \sum_{ij}(Y_{ij}-\hat{Y_{ij}})^2$$

where $\hat{Y_{ij}} = b_0 + b_1X_j$. The degrees of freedom are $df_R = (n-2)$

Now, we can test the hypotheses

$$
\begin{aligned}
H_0 &: E[Y] = \beta_0 + \beta_1 X\\
H_1 &: E[Y] \neq \beta_0 + \beta_1 X
\end{aligned}
$$

by computing the ratio

$$F^* = \frac{\left[\frac{SSE(R)-SSE(F)}{df_R-df_F}\right]}{\left[\frac{SSE(F)}{df_F}\right]}$$

The test is on whether a linear model is justified at all. This is different from just testing that the slope is 0. The main purpose of this test is to see if we can use a linear model instead of a complex model.

We can gain some insight into the components of the F^* ratio. We have that

$$(Y_{ij}-\hat{Y_{ij}}) = (Y_{ij}-\overline{Y_j}) - (\overline{Y_j}-\hat{Y_{ij}})$$

We then get the relationship

$$\sum_{ij}(Y_{ij} - \hat{Y_{ij}})^2 = \sum_{ij}(Y_{ij} - \overline{Y_j})^2 + \sum_{ij}(\overline{Y_j} - \hat{Y_{ij}})^2$$

The components are broken down as follows:

- $SSE(R) = \sum_{ij}(Y_{ij} - \hat{Y_{ij}})^2$ is the error sum of squares for the reduced model
- $SSPE = \sum_{ij}(Y_{ij} - \overline{Y_j})^2$ is the pure error sum of squares
- $SSLF = \sum_{ij}(\overline{Y_j} - \hat{Y_{ij}})^2$ is the error sum of squares due to lack of fit which is independent of i

We also have that the the degrees of freedom of the pure error sum of squares is $df_{PE} = n - c$ and the degrees of freedom of the lack of fit sums of squares is $df_{LF} = c - 2$. An ANOVA table summarizes the analysis:

Table 4: ANOVA Table for Lack of Fit Test

Source	Sum of Squares (SS)	df	Mean Square (MS = SS/df)	F statistic	E[MS]
Regression	$SSR = \sum_{ij}(\hat{Y_{ij}} - \overline{Y})^2$	1	$MSR = SSR$	$\frac{MSR}{MSE}$	$\sigma^2 + \beta_1^2 \sum(X_i - \overline{X})^2$
Error	$SSE(R) = \sum(Y_{ij} - \hat{Y_{ij}})^2$	$n - 2$	$MSE = \frac{SSE(R)}{n-2}$		σ^2
Lack of Fit	$SSLF = \sum_{ij}(\overline{Y_j} - \hat{Y_{ij}})^2$	$c - 2$	$MSLF = \frac{SSLF}{c-2}$	$F^* = \frac{MSLF}{MSPE}$	$\sigma^2 + \frac{\sum n_i(\mu_i - \beta_0 - \beta_1 X_i)^2}{c-2}$
Pure Error	$SSPE = \sum_{ij}(Y_{ij} - \overline{Y_j})^2$	$n - c$	$MSPE = \frac{SSPE}{n-c}$		
Total	$SSTO = \sum(Y_i - \overline{Y})^2$	$n - 1$			

The F^* ratio tests for lack of fit with a simple linear regression model. If there is no lack of fit, the the ratio should be closer to 1 since both the pure error sums of squares and the error sum of squares due to lack of fit are unbiased estimators of σ^2 under H_0. Otherwise, we would expect the ratio to be large.

We can also extend this concept to multiple linear regression. We consider the case where we have 2 predictors.

We define

$$SSR(X_2|X_1) = SSE(X_1) - SSE(X_1, X_2)$$

to be the reduction in the error sum of squares when after X_1 is included, an additional variable X_2 is added to the model. We can rewrite the expression as

$$SSR(X_2|X_1) = SSR(X_1, X_2) - SSR(X_1)$$

Similarly, when we have 3 predictors, we have that

$$SSR(X_1, X_2, X_3) = SSR(X_1) + SSR(X_2|X_1) + SSR(X_3|X_1, X_2)$$

This decomposition enables us to judge the effect an added variable has on the sum of squares due to regression.

We can use this process to test a full model with all predictors and a reduced model with only selected predictors by obtaining the error sum of squares of the full and reduced models.

4.4 Example: Delivery Time Data

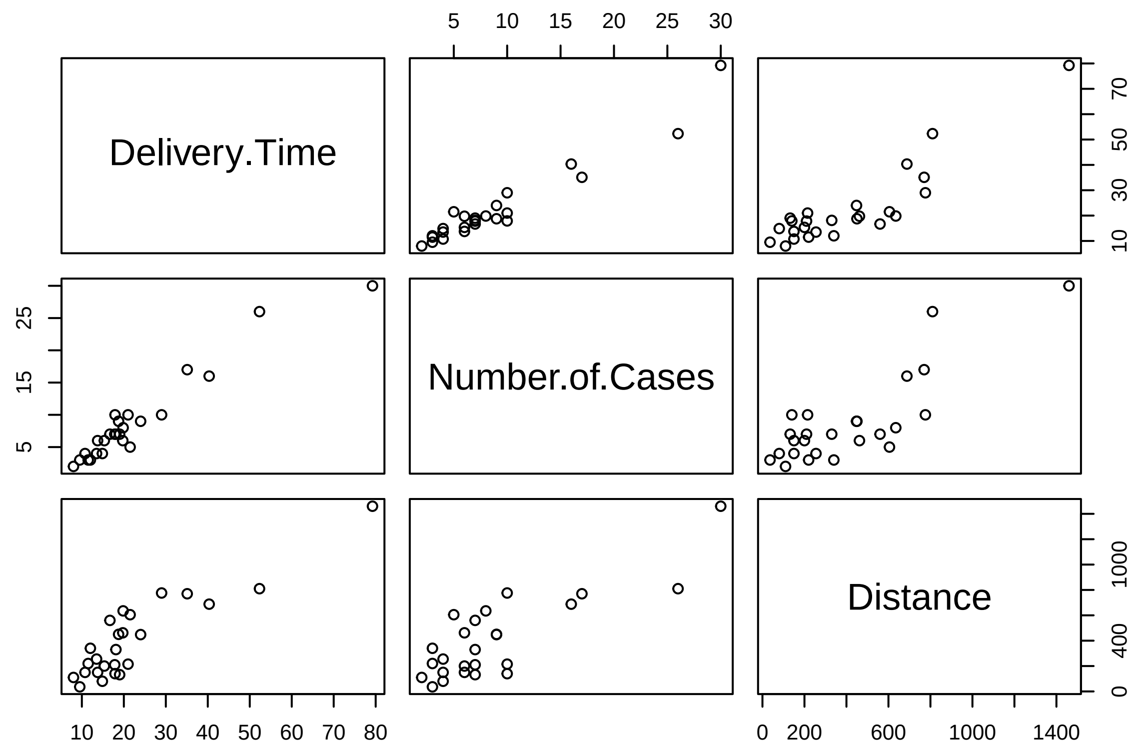
A soft drink bottler is interested in predicting the time required by the route driver to deliver the vending machines in an outlet. We let Y be the delivery time, X_1 be the number of cases of product stocked, and X_2 be the distance walked by the route driver in feet.

We first want to create a scatterplot matrix of the data

```
delivery = read.table("/System/Volumes/Data/MAT 3375/Summary Sheet/Delivery Time.txt", header =TRUE, se
names(delivery)
```

```
## [1] "Delivery.Time"   "Number.of.Cases" "Distance"
```

```
plot(delivery)
```



Now, we want to fit a multiple linear regression model for the data.

```
X_1 = delivery$Number.of.Cases
X_2 = delivery$Distance
Y = delivery$Delivery.Time
model = lm(Y~X_1+X_2, data=delivery)
model
```

```
##
## Call:
## lm(formula = Y ~ X_1 + X_2, data = delivery)
```

```
##
## Coefficients:
## (Intercept)          X_1          X_2
##     2.34123      1.61591      0.01438
```

```r
summary(model)
```

```
##
## Call:
## lm(formula = Y ~ X_1 + X_2, data = delivery)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.7880 -0.6629  0.4364  1.1566  7.4197
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.341231   1.096730   2.135 0.044170 *
## X_1         1.615907   0.170735   9.464 3.25e-09 ***
## X_2         0.014385   0.003613   3.981 0.000631 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.259 on 22 degrees of freedom
## Multiple R-squared:  0.9596, Adjusted R-squared:  0.9559
## F-statistic: 261.2 on 2 and 22 DF,  p-value: 4.687e-16
```

From the summary above, we find that the least squares multiple linear regression function is $\hat{Y} = 2.341231 + 1.615907X_1 + 0.014385X_2$. If we want to test for regression, we see that the F statistic is 261.2 and the p-value is very small. This means that we can reject $H_0 : \beta_1 = \beta_2 = 0$, which means that there is enough evidence to suggest that at least one of X_1 and X_2 have influence on Y. If we look at the t statistics, we have that we can reject the null hypotheses $H_0 : \beta_0 = 0$, $H_0 : \beta_1 = 0$, and $H_0 : \beta_2 = 0$. This is because the t statistics are all statistically significant.

We can also do an ANOVA test on the model

```r
anova(model)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## X_1        1 5382.4  5382.4 506.619 < 2.2e-16 ***
## X_2        1  168.4   168.4  15.851 0.0006312 ***
## Residuals 22  233.7    10.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We have found that the F statistic to test $H_0 : \beta_1 = 0$ is 506.619 and the F statistic to test $H_0 : \beta_2 = 0$ is 15.851, which are both significant. There is convincing evidence that both X_1 and X_2 have influence on Y.

We can also conduct a test using the extra sum of squares principle. We're testing the full model $Y = \beta_0 + \beta_1X_1 + \beta_2X_2$ against the reduced model $Y = \beta_0 + \beta_1X_1$.

```
Full = lm(Y~X_1+X_2, data=delivery)
Reduced = lm(Y~X_1)
anova(Reduced, Full)
```

```
## Analysis of Variance Table
##
## Model 1: Y ~ X_1
## Model 2: Y ~ X_1 + X_2
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     23 402.13
## 2     22 233.73  1     168.4 15.851 0.0006312 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We find that the F statistic is 15.851, which is statistically significant. Therefore, we can reject the null hypothesis that $H_0 : \beta_2 = 0$. This means that we cannot use the reduced model for this data.

4.5 Example: Bank Data

We want to illustrate for testing for lack of fit. For this, we'll use the Bank data. We'll compare the reduced model $Y_{ij} = \beta_0 + \beta_1 X_j + \epsilon_{ij}$ and the full model $Y_{ij} = \mu_j + \epsilon_{ij}$ for a data with repeat observation at the same predictor values.

```
#We first load the data
bank = read.table("/System/Volumes/Data/MAT 3375/Summary Sheet/Bank Data.txt", header=TRUE, sep='\t')
names(bank)
```

```
## [1] "Minimum.Deposit"     "Number.New.accounts"
```

```
#Now, we fit the reduced and full models for the data
x = bank$Minimum.Deposit
y = bank$Number.New.accounts

reduced = lm(y~x)
full = lm(y~0 + as.factor(x))
anova(reduced, full)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ x
## Model 2: y ~ 0 + as.factor(x)
##   Res.Df   RSS Df Sum of Sq      F   Pr(>F)
## 1      9 14742
## 2      5  1148  4     13594 14.801 0.005594 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The F statistic is 14.801, which is statistically significant. Therefore, we reject the null hypothesis that $E[Y] = \beta_0 + \beta_1 X$ and a linear model is not a good fit for the data.

4.6 Simultaneous Confidence Intervals

We have learned to construct a confidence interval for one specific parameter (i.e. confidence intervals for β_0 and β_1 in a simple linear regression model). However, sometimes we want to calculate simultaneous or joint confidence intervals for the entire set of parameters. For example, we may want to construct simultaneous confidence intervals for all the coefficients in a linear regression model (i.e. simultaneous confidence intervals that contain both the intercept and slope in a simple linear regression model). However, the confidence level decreases as we include more parameters to estimate.

For example, we consider 2 parameters: β_0 and β_1 of a simple linear regression model, and we want to construct simultaneous $100(1-\alpha/2)\%$ confidence intervals for the parameters. We can obtain a $100(1-\alpha/2)\%$ for each parameter. However, if we let A_1 be the event that β_0 is in its confidence interval and A_2 be the event that β_1 is in its confidence interval, then if we assume that A_1 and A_2 are independent, then we have that

$$P(A_1 \cap A_2) = P(A_1)P(A_2)$$

Say if we have $P(A_1) = 0.95$ and $P(A_2) = 0.95$, then $P(A \cap B) = (0.95)^2 = 0.9025 < 0.95$. However, the events are never independent, so $P(A_1 \cap A_2)$ is even less.

One strategy is to use the Bonferroni's procedure. Bonferroni's inequality states that for 2 events $\overline{A_1}$, $\overline{A_2}$, we have that

$$P(\overline{A_1} \cap \overline{A_2}) = P(\overline{A_1}) + P(\overline{A_2}) - P(\overline{A_1} \cap \overline{A_2}) \leq P(\overline{A_1}) + P(\overline{A_2})$$

and then we use DeMorgan's identity:

$$P(A_1 \cap A_2) = 1 - P(\overline{A_1} \cup \overline{A_2}) \geq 1 - P(\overline{A_1}) - P(\overline{A_2})$$

We define A_1 as the event that β_0 is contained in its $100(1-\alpha)\%$ confidence interval and A_2 is the event that β_1 is contained in its $100(1-\alpha)\%$ confidence interval. In this case, we have that

$$P(\overline{A_1}) = P(\overline{A_2}) = \alpha$$

and hence, we get that

$$P(A_1 \cap A_2) \geq 1 - P(A_1) - P(A_2) \geq 1 - 2\alpha$$

Now the event $A_1 \cap A_2$ is the event that the intervals $b_0 \pm t(\alpha/2; n-2) \cdot s[b_0]$ and $b_1 \pm t(\alpha/2; n-2) \cdot s[b_1]$ simultaneously cover β_0 and β_1, respectively. The probability of such event is $1 - 2\alpha$. If we have $\alpha = 0.05$, then we get that $1 - 2\alpha = 0.90$. There would be a 0.90 probability that β_0 and β_1 simultaneously fall into the intervals. We would then be 90% confident that the intervals simultaneously cover β_0 and β_1.

On the other hand, if we want to be 95% confident that the intervals simultaneously cover β_0 and β_1, then we need $1-2\alpha-0.95$, which means we need $\alpha = 0.025$. Which means that we need to compute $t(0.025/2; n-2)$. Then we can construct the simultaneous confidence intervals with the t critical value calculated.

In general, when there's p parameters, the probability that they all fall in their respective confidence intervals is $1 - p\alpha$. If we want to be 95% confident that the intervals simultaneously cover the parameters, then we need that $1 - p\alpha = 0.95$ and $\alpha = 0.05/p$. Then we need to compute $t(0.05/p; n-2)$, which may not be possible without a computer.

4.7 Example of Simultaneous Confidence Intervals

We examine the rocket propellant data. We do this using R.

```
rocket = read.table("/System/Volumes/Data/MAT 3375/Summary Sheet/Rocket .txt", header=TRUE, sep='\t')
names(rocket)
```

```
## [1] "Shear.strength"    "Age.of.Propellant"
```

```r
y = rocket$Shear.strength
x = rocket$Age.of.Propellant
fit = lm(y~x)
confint(fit, level = 1-0.05/2)
```

```
##                 1.25 %    98.75 %
## (Intercept) 2519.79245 2735.85227
## x            -44.21747  -30.08971
```

This way, we have computed simultaneous 95% confidence intervals for β_0 and β_1. The intervals are $[2519.79245, 2735.85227]$ and $[-44.21747, -30.08971]$ for β_0 and β_1, respectively.

Alternatively, we can compute the critical value for computing the confidence intervals

```r
qt(0.9875, nrow(rocket) - 2)
```

```
## [1] 2.445006
```

Then we can use this along with the summary data of the model to find the confidence interval.

```r
summary(fit)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -215.98  -50.68   28.74   66.61  106.76
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2627.822     44.184   59.48  < 2e-16 ***
## x            -37.154      2.889  -12.86 1.64e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 96.11 on 18 degrees of freedom
## Multiple R-squared:  0.9018, Adjusted R-squared:  0.8964
## F-statistic: 165.4 on 1 and 18 DF,  p-value: 1.643e-10
```

Then the simultaneous confidence intervals are $2627.822 \pm 2.44(44.184)$ for β_0 and $-37.154 \pm 2.44(2.889)$ for β_1.

5 Model Adequacy Checking

In the previous sections, we talked about simple and multiple linear regression. We recall that the assumptions that we make about the model are

- ϵ_i are normally distributed

- $E[\epsilon_i] = 0$ and $Var[\epsilon_i] = \sigma^2$
- ϵ_i are independent

We now need to check the assumptions. If the assumptions are not met, then the analysis we do with linear models may not closely reflect the actual model. In this section, we talk about how to check for normality of the error terms and the constancy of variance.

The basic tool that we use to check for model adequacy is analyzing the residuals $e_i = Y_i - \hat{Y}_i$.

5.1 Checking for Normality

To check for normality, we can do this using several ways. We do this by examining the shape of the distribution of the data collected. There are 3 ways in which we can do this:

- We can construct boxplots of residuals. Under the normality assumption, the boxplot should show a symmetric box around the median of approximately 0
- We can construct a histogram of the residuals. It provides a graphical check on normality because if it shows a bell shape centered at 0 approximately, then we can assume a normal distribution
- We can construct a quantile-quantile plot. This plot compares the quantiles of the residual data collected (sample quantiles) with the quantiles from a normal distribution. This is the plot of the ranked residuals against the expected value under normality. We let $e_{(k)}$ be the residual with rank k, and E_k be the expected value of the residual with rank k under normality. We have $E_k = \sqrt{MSE}\Phi^{-1}(\frac{k-0.375}{n+0.25})$ for $k = 1, ..., n$. We plot $e_{(k)}$ against E_k. Under normality, one would expect a straight line pattern in the plot.

5.2 Checking for Constancy of Variance

The other assumption we made about our models is that the variance of the random error terms remain constant. We need to use the residuals to verify this assumption.

We note that the variance of the residuals is $Var[e_i] = \sigma^2(1 - h_{ii})$ and the covariance of residuals is $Cov(e_i, e_j) = \sigma^2(1 - h_{ij})$, where h_{ij} is the element of the hat matrix on the ith row and jth column. If we look at the variance of each residual, we notice that the values are different. This means that the residuals have different variances. So, we want to look at the Studentized or standardized residuals

$$e_i^* = \frac{e_i}{s\sqrt{1 - h_{ii}}}$$

where $s^2 = MSE$. The semi-studentized residuals are defined as

$$\frac{e_i}{\sqrt{1 - h_{ii}}}$$

which also has constant variance.

We can check for constancy of variance by making a plot of the Studentized residuals against the fitted values. If the plot shows a random distribution of the points, then we can assume that the variance of the error terms are constant. However, if we see a telescoping increasing or decreasing pattern among the points, then there is evidence of non-constancy of variance. Then the constancy of variance assumption would be violated.

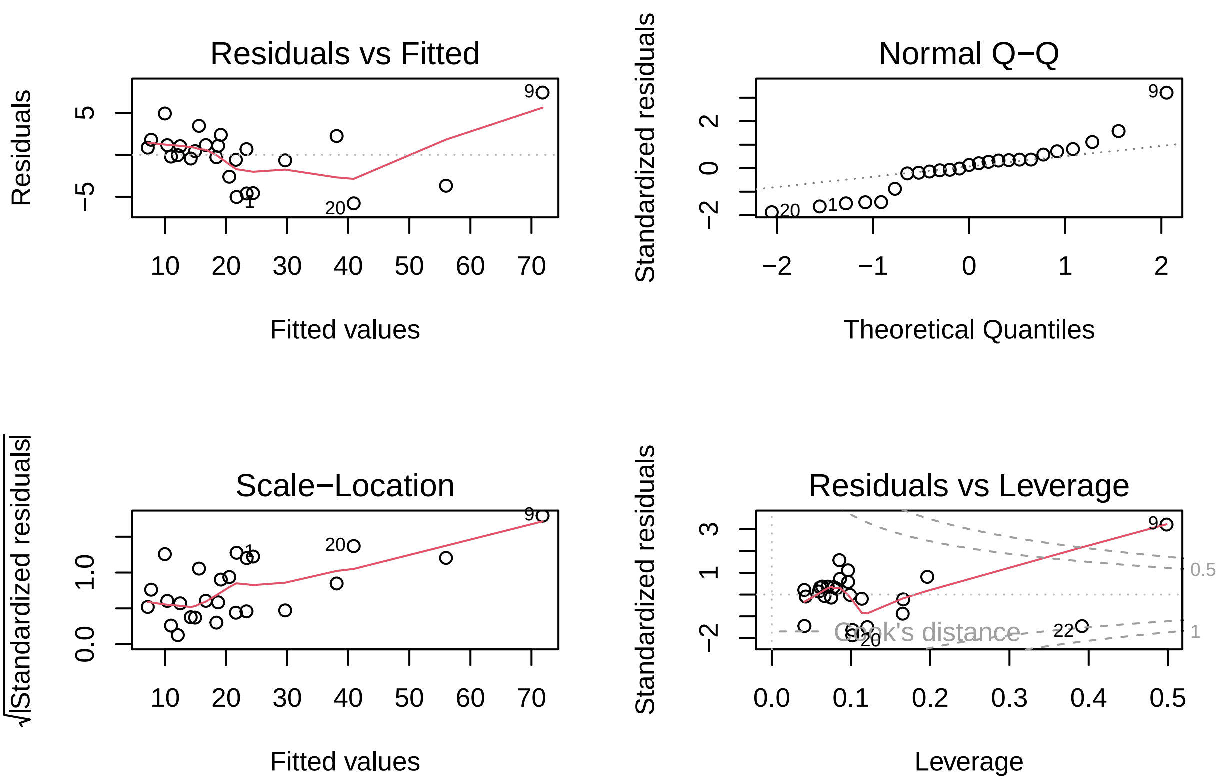
We can also construct a scale-location plot to examine the homogeneity of the variance of the residuals. This plots the square roots of the absolute values of the Studentized residuals vs. the fitted values.

5.3 Example of checking for model adequacy

We look at the delivery time data. We let X_1 represent the delivery number of cases and X_2 represent the delivery distance, and we let Y represent the delivery time.

```r
X1 = delivery$Number.of.Cases
X2 = delivery$Distance
Y = delivery$Delivery.Time

fit = lm(Y~X1+X2, data=delivery)
par(mfrow=c(2,2))
plot(fit)
```



From looking at the plots above, we have that the qq-plot follows a straight line pattern in the middle, but not so much on the sides. Therefore, there may be a problem with the normality assumption for this particular dataset. From looking at the residuals vs. fitted plot, we see that the points are randomly distributed on the plot. This means that we can assume constancy of variance in the dataset.

6 Remedial Measures and Transformations

There are times when the assumptions for fitting and testing a linear model are violated, but we don't want to immediately discard the model. Instead, we can do some transformations to the response or predictor variables. This may help linearize the model and bring it in line with the assumptions made.

6.1 Variance Stabilizing Transformations

When the assumption that the variance of the error terms is constant is not reasonable, we can then transform the data to make the assumption more reasonable. Below are some examples of transformations we apply to response variables from various distributions.

6.1.1 Poisson Distribution

In the case that the response variable Y follows a Poisson distribution, then its variance is equal to its mean. If Y is distributed as a Poisson random variable with mean λ, then $\sqrt{Y}$ is distributed more nearly normally with variance approximately $\frac{1}{4}$ if λ is large. In this case, we can regress $\sqrt{Y}$ against X and fit a linear regression model.

6.1.2 Binomial Distribution

In the case that the response variable Y follows a Binomial distribution $B(n,p)$, we have that its mean is $E[Y] = np$. We use the transformation

$$\sin^{-1}\sqrt{\frac{Y+c}{n+2c}}$$

where the optimal value of c is $\frac{3}{8}$ if $E[Y]$ and $n - E[Y]$ are large. The variance is approximately $\frac{1}{4}(n+\frac{1}{2})^{-1}$.

6.2 Transformations to Linearize the Model

Sometimes when we plot Y against X, the plot does not look linear. Then we can apply transformations to the response variable Y to make the plot look more linear. Below are some examples:

6.2.1 Exponential Model

If the true model is

$$Y = \beta_0 e^{\beta_1 X}\epsilon$$

then we can transform it by taking the logarithms. Here is what the model would look like:

$$\ln(Y) = \ln(\beta_0) + \beta_1 X + \ln(\epsilon)$$

We still have to make the usual assumptions for a linear model and then verify them.

6.2.2 Reciprocal Model

The model $Y = \beta_0 + \beta_1 X^{-1} + \epsilon$ can be linearized using the trainsformation $X^* = X^{-1}$. This is a transformation on the predictor variable.

The model $\frac{1}{Y} = \beta_0 + \beta_1 X + \epsilon$ can be linearized using the reciprocal transformation $Y^* = Y^{-1}$. This is a transformation of the response variable.

The model $Y = \frac{X}{\beta_0+\beta_1 X}$ can be linearized using the reciprocal transformation in 2 steps. First, we use $Y^* = Y^{-1}$ and $X^* = X^{-1}$. Then we have

$$\begin{aligned} Y &= \frac{X}{\beta_0 + \beta_1 X} \\ &= \frac{1}{\beta_0 X^{-1} + \beta_1} \end{aligned}$$

Consequently, we get that $Y^{-1} = \beta_0 X^{-1} + \beta_1$. Then we obtain $Y^* = \beta_1 + \beta_0 X^*$.

6.3 Box-Cox Transformations

The data may not appear to be normally distributed sometimes. Then Box and Cox suggested a power transformation of the type

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

where

$$\dot{Y} = \ln^{-1}[\frac{\sum \ln(Y_i)}{n}]$$

The value of λ can be estimated using trial and error. We fit a model for $Y^{(\lambda)}$ for various values of λ and selecting the one which mimimizes the error sum of squares from a graphical plot. We can also construct a confidence interval for λ. Using R, we can estimate λ using maximum likelihood.

The theory behind the transformation is as follows:

The original model was assumed to be $\boldsymbol{Y} \sim N_n(\boldsymbol{X\beta}, \sigma^2\boldsymbol{I})$

Then the transformed model is $Y^{(\lambda)}$ and has likelihood function given by

$$\frac{1}{(2\pi)^{n/2}\sigma^n} exp(-\frac{(\boldsymbol{y}^{(\lambda)} - \boldsymbol{X\beta})^T(\boldsymbol{y}^{(\lambda)} - \boldsymbol{X\beta})}{2\sigma^2})J(\lambda; \boldsymbol{y})$$

with parameters $\boldsymbol{\beta}$ and the Jacobian for the transformation

$$J(\lambda; \boldsymbol{y}) = \prod_{i=1}^{n} |\frac{\partial y_i^{(\lambda)}}{\partial y_i}|$$

The maximum likelihood estimator of the variance is given by

$$\hat{\sigma^2} = \frac{\boldsymbol{(Y^{(\lambda)})^T[I - X(X^T X)^{-1} X^T]Y^{(\lambda)}}}{n}$$

The maximum log likelihood for fixed λ is

$$L_{max} = -\frac{n}{2}\log(\hat{\sigma^2}) + \log(J(\lambda; \boldsymbol{y})) = -\frac{n}{2}\log(\hat{\sigma^2}) + (\lambda - 1)\sum \log(y_i)$$

We then plot $L_{max}(\lambda)$ against λ to find the value of λ which yields the maximum.

6.4 Weighted Least Squares

This approach is for when the constancy of variance assumption is not reasonable. We assume that $Var[e_i] = \sigma_i^2$. Instead of minimizing the sum of the squared errors, we can minimize the sum of the weighted squared errors

$$\sum w_i e_i^2$$

where the weights satisfy

$$Var[\sqrt{w_i} e_i] = \sigma^2$$

. In other words, we're to minimize

$$\sum w_i [Y_i - \beta_0 - \beta_1 X_i]$$

with respect to β_0 and β_1.

6.4.1 Estimation and Fitting

With ordinary least squares regression, we have that the linear regression model is

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$$

We define the matrix $\boldsymbol{W}$ as the matrix of weights

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

Then the original model goes to

$$\boldsymbol{W^{1/2}Y} = \boldsymbol{W^{1/2}X\beta} + \boldsymbol{W^{1/2}\epsilon}$$

, where

$$\boldsymbol{W}^{1/2} = \begin{pmatrix} w_1^{1/2} & 0 & \dots & 0 \\ 0 & w_2^{1/2} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & w_n^{1/2} \end{pmatrix}$$

Then we get that the least squares estimate for $\boldsymbol{\beta}$ is

$$\begin{aligned} \boldsymbol{b_W} &= ((\boldsymbol{W^{1/2}X})^{\boldsymbol{T}}(\boldsymbol{W^{1/2}X}))^{\boldsymbol{-1}}(\boldsymbol{W^{1/2}X})(\boldsymbol{W^{1/2}Y}) \\ &= (\boldsymbol{X^T W^{1/2} W^{1/2} X})^{\boldsymbol{-1}}\boldsymbol{X^T W^{1/2} W^{1/2} Y} \\ &= (\boldsymbol{X^T W X})^{\boldsymbol{-1}}\boldsymbol{X^T W Y} \end{aligned}$$

The MSE becomes

$$\begin{aligned} MSE_W &= \frac{\sum w_i(Y_i - \hat{Y}_i)^2}{n-p} \\ &= \frac{\sum w_i(e_i)^2}{n-p} \end{aligned}$$

6.4.2 Choosing weights

We want to now choose the appropriate weights for weighted least squares. There are many ways we can do this. For example, we can take $w_i = \sqrt{X_i}$ or $w_i = \sqrt{Y_i}$. However, there is one way that we can commonly use. We first divide up the data into clusters using the X variable. Then, we estimate the sample variances of the Y_i for each group, denote them as s_i^2. Then we compute the averages of the X_i values in each cluster. We fit a regression model of the variances of Y_i against the averages of X_i in each cluster. We then estimate the variances by substituting all the X_i values into the equation of the regression function of the variances against the averages. The weights are the inverses of the estimated variances.

For example, we use the turkey data. We group the data into clusters, then we obtain the averages of the X_i in each group and the variances of Y_i in each group. We obtain the regression function $s^2 = 1.5329 - 0.7334\overline{X} + 0.0883\overline{X}^2$. Then we substitute the individual X_i in the data to find $\hat{s_i^2} = 1.5329 - 0.7334X_i + 0.0883X_i^2$. Then we obtain the weights as $\frac{1}{s_i^2}$. With the usual linear regression model, we get that the regression function is

$$Y = -0.579 + 1.14X$$

and the weighted regression is

$$Y = -0.892 + 1.16X$$

6.5 Breusch-Pagan Test

The Breusch-Pagan Test is used for checking for constancy of variance. It assumes the model

$$\log(\sigma_i^2) = \gamma_0 + \gamma_1 X_i$$

. We want to test the hypothesis $H_0 : \gamma_1 = 0$. To do this, we regress the squared residuals e_i^2 against the predictors X_i. Then we can calculate the regression sums of squares of the model SSR^*. Let SSE represent the error sum of squares when we fit a model for Y against X. The test statistic is then

$$\chi^2_{BP} = \frac{\frac{SSR^*}{2}}{\frac{SSE}{n}} \sim \chi^2(1)$$

We reject H_0 for large values of the test statistic.

We can also use R to do the test. For the purpose, we use the Turkey data. Then we get that

```
turkey = read.table("/System/Volumes/Data/MAT 3375/Summary Sheet/Turkey data.txt", header=TRUE , sep='\
names(turkey)
```

```
## [1] "Age"    "Weight" "Origin" "Z1"     "Z2"
```

```
y <- turkey$Weight
x <- turkey$Age
fit <- lm(y~x)
bptest(fit)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  fit
## BP = 2.5466, df = 1, p-value = 0.1105
```

We obtain a test statistic of 2.5466 and a p-value of 0.11. This means that we fail to reject H_0. We can assume that the variance of error terms are constant.