# ST_503 HW2

Halid Kopanski

2022-06-01

#1 Model A

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad E(y_i) = \bar{y} = \beta_0 + \beta_1 \bar{x}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad \hat{\beta}_1 = \frac{\Sigma (x_i - \bar{x})(y_i - \bar{y})}{\Sigma (x_i - \bar{x})^2}$$

$$\begin{aligned} y_i - \bar{y} &= \beta_0 + \beta_1 x_i + \varepsilon_i - \bar{y} \\ &= \bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_i + \varepsilon_i - \bar{y} \\ &= \hat{\beta}_1 (x_i - \bar{x}) + \varepsilon_i \end{aligned}$$

---

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(\hat{\beta}_1 (x_i - \bar{x}) + \varepsilon_i)}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} \beta_1 (x_i - \bar{x})^2 + \varepsilon_i (x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$= \beta_1 + \frac{\sum_{i=1}^{n} (x_i - \bar{x}) \varepsilon_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$E(\hat{\beta}_1) = \beta_1 + E\left( \frac{\Sigma (x_i - \bar{x})}{\Sigma (x_i - \bar{x})^2} \varepsilon_i \right) = \beta_1 \quad \text{since } E(\varepsilon_i) = 0$$

---

$$Var(\hat{\beta}_1) = Var(\beta_1) + Var\left( \frac{\Sigma (x_i - \bar{x})}{\Sigma (x_i - \bar{x})} \varepsilon_i \right) = \left( \frac{\Sigma (x_i - \bar{x})}{\Sigma (x_i - \bar{x})^2} \right)^2 Var\, \varepsilon_i$$

$$= \frac{SS_{xx}}{SS_{xx}^2} \sigma^2 = \frac{\sigma^2}{SS_{xx}}$$

---

$$\hat{\beta}_0 = \beta_0 + \beta_1 \bar{x} - \hat{\beta}_1 \bar{x} \qquad E(\hat{\beta}_0) = \beta_0 + \beta_1 \bar{x} - \bar{x}\, E(\hat{\beta}_1) = \beta_0 + \beta_1 \bar{x} - \beta_1 \bar{x}$$

$$= \beta_0$$

---

$$Var(\hat{\beta}_0) = Var(-\hat{\beta}_1 \bar{x}) = \bar{x}^2\, Var(\hat{\beta}_1) = \frac{\bar{x}^2 \sigma^2}{SS_{xx}}$$

#1 Model B

$$y_i = \beta_1 x_i + \varepsilon_i$$

$$\hat{\beta}_1 = \frac{\sum x_i y_i}{\sum x_i^2}$$

$$E(\hat{\beta}_1) = E\left(\frac{\sum\left[x_i(\beta_1 x_i + \varepsilon_i)\right]}{\sum x_i^2}\right)$$

$$= E\left(\frac{\sum(\beta_1 x_i^2 + x_i \varepsilon_i)}{\sum x_i^2}\right) = E\left(\beta_1 + \frac{\sum x_i \varepsilon_i}{\sum x_i^2}\right)$$

$$= \beta_1 + E\left(\frac{\sum x_i}{\sum x_i^2}\right) E(\varepsilon_i) = \beta_1$$

$$Var(\hat{\beta}_1) = Var\left(\frac{\sum x_i \varepsilon_i}{\sum x_i^2}\right) = \frac{\sum x_i^2 \, Var(\varepsilon_i)}{\left(\sum x_i^2\right)^2} = \frac{Var(\varepsilon_i)}{\sum x_i^2}$$

$$= \frac{\sigma^2}{n \overline{x^2}}$$

Model C

$$E(\hat{\beta}_0) = E(\bar{y}) = E\left(\frac{1}{n}\Sigma y_i\right)$$

$$= \frac{1}{n}E\left(\Sigma \beta_0 + \varepsilon_i\right)$$

$$= \frac{1}{n}\Sigma E(\beta_0) + \frac{1}{n}\Sigma E(\varepsilon_i)$$

$$= \beta_0 + 0$$

$$= \beta_0$$

$$Var(\hat{\beta}_0) = Var\left(\frac{1}{n}\Sigma y_i\right)$$

$$= \frac{1}{n^2}\sum_{i=1}^{n} Var\, y_i$$

$$= \frac{1}{n^2}\sum_{i=1}^{n} Var\left(\beta_0 + \varepsilon_i\right)$$

$$= \frac{1}{n^2}\left[\sum_{i=1}^{n} Var\,\beta_0 + Var\,\varepsilon_i\right]$$

$$= \frac{n}{n^2}\left(0 + \sigma^2\right)$$

$$= \frac{\sigma^2}{n}$$

2. $Cov(\hat{\beta}_1, \bar{Y})$

$$= E(\hat{\beta}_1 \bar{Y}) - E(\hat{\beta}_1)E(\bar{Y}) = E\left[(\hat{\beta}_1 - \beta_1)(\bar{Y} - E(\bar{Y}))\right]$$

$$= \bar{Y}E(\hat{\beta}_1) - \bar{Y}E(\hat{\beta}_1)$$

$$= 0$$

$Cov(\hat{\beta}_1, \hat{\beta}_0)$

$$= E(\hat{\beta}_1 \hat{\beta}_0) - E(\hat{\beta}_0)E(\hat{\beta}_1) = E\left[(\hat{\beta}_1 - \beta_1)(\hat{\beta}_0 - \beta_0)\right]$$

$$\hat{\beta}_0 - \beta_0 = \beta_1 \bar{x} - \hat{\beta}_1 \bar{x} = -\bar{x}(\hat{\beta}_1 - \beta_1)$$

$$Cov(\hat{\beta}_1, \hat{\beta}_0) = E\left[(\hat{\beta}_1 - \beta_1)\left(-\bar{x}(\hat{\beta}_1 - \beta_1)\right)\right]$$

$$= E\left[-\bar{x}(\hat{\beta}_1 - \beta_1)^2\right]$$

$$= -\bar{x}\,E\left[(\hat{\beta}_1 - \beta_1)^2\right] = -\bar{x}\,Var(\hat{\beta}_1)$$

$$= \frac{-\bar{x}\sigma^2}{SS_{xx}}$$

# #3

## Model B

3a) $y_i = \beta_1 x_i + \varepsilon_i$

$$\hat{\beta}_2 = \frac{\bar{y}}{\bar{x}} \qquad \hat{\beta}_3 = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{x_i}$$

$$E(\hat{\beta}_2) = E\left(\frac{\frac{1}{n}\sum_{i=1}^{n} y_i}{\bar{x}}\right) = E\left(\frac{\sum_{i=1}^{n}\beta_1 x_i + \varepsilon_i}{\sum_{i=1}^{n} x_i}\right)$$

$$= E\left(\beta_1 + \frac{\varepsilon_i}{n\bar{x}}\right) = E(\beta_1) + E\left(\frac{\varepsilon_i}{n\bar{x}}\right) = \beta_1 + 0$$

$E(\hat{\beta}_2) = \beta_1$, this is an unbiased estimator

3b) $$\hat{\beta}_2 = \frac{\frac{1}{n}\sum_{i=1}^{n} y_i}{\bar{x}} = \sum_{i=1}^{n}\frac{1}{n\bar{x}}\, y_i$$

$$c_i = \frac{1}{n\bar{x}}$$

3.c) $$Var(\hat{\beta}_2) = Var\left(\frac{\bar{y}}{\bar{x}}\right) = Var\left(\beta_1 + \frac{\sum \varepsilon_i}{\sum x_i}\right)$$

$$= 0 + \frac{\sigma^2}{n\bar{x}} = \frac{\sigma^2}{n\bar{x}}$$

3.d) $Var(\hat{\beta}_3)$ would be a function of $Var\left(\frac{y_i}{x_i}\right)$. Since $Var(\hat{\beta}_2)$ is zero, the $Var(\hat{\beta}_3)$ would be larger since it is non zero.

3e) $E(\hat{\beta}_3) = E\left(\frac{1}{n}\Sigma \frac{y_i}{x_i}\right)$

$$= \frac{1}{n}\sum E\left(\frac{y_i}{x_i}\right) = \frac{1}{n}\sum E\left(\frac{\beta_1 x_i + \varepsilon_i}{x_i}\right)$$

$$= \frac{1}{n}\sum E\left(\beta_1 + \frac{\varepsilon_i}{x_i}\right) = \frac{1}{n}\sum E(\beta_1) = \beta_1$$

$\hat{\beta}_3$ is unbiased.

3f) $\hat{\beta}_3 = \Sigma \frac{1}{n x_i} y_i$

$$c_i = \frac{1}{n x_i}$$

3g) $Var(\hat{\beta}_3) = Var\left(\frac{1}{n}\sum \frac{y_i}{x_i}\right)$

$$= \frac{1}{n^2}\sum Var\left(\frac{y_i}{x_i}\right) = \frac{1}{n^2}\sum Var\left(\frac{\beta_1 x_i + \varepsilon_i}{x_i}\right)$$

$$= \frac{1}{n^2} Var\left(\sum\left(\frac{\varepsilon_i}{x_i}\right)\right) = \frac{1}{n^2}\sum \frac{1}{x_i^2} Var(\varepsilon_i) = \frac{\sigma^2}{n^2}\sum_{i=1}^{n}\frac{1}{x_i}$$

3h) $Var(\hat{\beta}_3)$ is smaller than $Var(\hat{\beta}_1)$ due to the $\frac{1}{n^2}$ term

**LMR 3.4 a)**

```
##
## Call:
## lm(formula = total ~ expend + ratio + salary, data = sat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -140.911  -46.740   -7.535   47.966  123.329
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1069.234    110.925   9.639 1.29e-12 ***
## expend        16.469     22.050   0.747   0.4589
## ratio          6.330      6.542   0.968   0.3383
## salary        -8.823      4.697  -1.878   0.0667 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 68.65 on 46 degrees of freedom
## Multiple R-squared:  0.2096, Adjusted R-squared:  0.1581
## F-statistic: 4.066 on 3 and 46 DF,  p-value: 0.01209
```

```
## Analysis of Variance Table
##
## Response: total
##           Df Sum Sq Mean Sq F value   Pr(>F)
## expend     1  39722   39722  8.4276 0.005658 **
## ratio      1   1143    1143  0.2424 0.624795
## salary     1  16631   16631  3.5285 0.066668 .
## Residuals 46 216812    4713
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

| df_vector | SS | MS | F_stat | p_value |
|---:|---:|---:|---:|---:|
| 3 | 57495.74 | 19165.248 | 4.066203 | 0.0120861 |
| 46 | 216811.94 | 4713.303 | NA | NA |
| 49 | 274307.68 | NA | NA | NA |

There is not enough evidence to reject the null hypothesis that $\beta_{salary}$ is zero. The p value is above the critical value of 0.025 for a two sided t test.

There is sufficient evidence to reject the null hypothesis that states $\beta_{expend} = \beta_{ratio} = \beta_{salary}$. The f statistic and the corresponding p value are 4.0662033 and 0.0120861 respectively.

Given the large p values and low $R^2$, we cannot say that the predictors had a large effect on the response variable.

**LMR 3.4 b)**

```
##
## Call:
## lm(formula = total ~ expend + ratio + salary + takers, data = sat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -90.531 -20.855  -1.746  15.979  66.571
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1045.9715    52.8698  19.784  < 2e-16 ***
## expend         4.4626    10.5465   0.423    0.674
## ratio         -3.6242     3.2154  -1.127    0.266
## salary         1.6379     2.3872   0.686    0.496
## takers        -2.9045     0.2313 -12.559 2.61e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 32.7 on 45 degrees of freedom
## Multiple R-squared:  0.8246, Adjusted R-squared:  0.809
## F-statistic: 52.88 on 4 and 45 DF,  p-value: < 2.2e-16
```

In the above output, we can see that the t test produces a very small p value for the new model that includes the taker predictor.

```
## Analysis of Variance Table
##
## Model 1: total ~ expend + ratio + salary
## Model 2: total ~ expend + ratio + salary + takers
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     46 216812
## 2     45  48124  1    168688 157.74 2.607e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Analysis of Variance Table
##
## Response: total
##           Df Sum Sq Mean Sq  F value    Pr(>F)
## expend     1  39722   39722  37.1436 2.260e-07 ***
## ratio      1   1143    1143   1.0685 0.3068088
## salary     1  16631   16631  15.5514 0.0002779 ***
## takers     1 168688  168688 157.7379 2.607e-16 ***
## Residuals 45  48124    1069
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The ANOVA table between the two models shows that the taker model significantly improves the model and should be accepted in place of the model without the taker predictor. The large f statistic gives the same conclusion as the previous t test.

| df_vector | SS | MS | F_stat | p_value |
|---:|---:|---:|---:|---:|
| 4 | 226183.8 | 56545.95 | 52.87534 | 0 |
| 45 | 48123.9 | 1069.42 | NA | NA |

| df_vector | SS | MS | F_stat | p_value |
|---|---|---|---|---|
| 49 | 274307.7 | NA | NA | NA |

We have sufficient evidence to reject the null hypothesis that $\beta_{takers} = 0$ at the $\alpha$ level of 0.05 (p value in this case is close to zero).

**LMR 3.7 a)**

```
##
## Call:
## lm(formula = Distance ~ RStr + LStr + RFlex + LFlex, data = punting)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.941  -8.958  -4.441  13.523  17.016
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -79.6236    65.5935  -1.214    0.259
## RStr          0.5116     0.4856   1.054    0.323
## LStr         -0.1862     0.5130  -0.363    0.726
## RFlex         2.3745     1.4374   1.652    0.137
## LFlex        -0.5277     0.8255  -0.639    0.541
##
## Residual standard error: 16.33 on 8 degrees of freedom
## Multiple R-squared:  0.7365, Adjusted R-squared:  0.6047
## F-statistic:  5.59 on 4 and 8 DF,  p-value: 0.01902
```

None of the predictors are significant at the 5% level

**LMR 3.7 b)**

```
## Analysis of Variance Table
##
## Model 1: Distance ~ 1
## Model 2: Distance ~ RStr + LStr + RFlex + LFlex
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     12 8093.3
## 2      8 2132.6  4    5960.7 5.5899 0.01902 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

| df_vector | SS | MS | F_stat | p_value |
|---:|---:|---:|---:|---:|
| 4 | 5960.668 | 1490.1669 | 5.589941 | 0.0190248 |
| 8 | 2132.641 | 266.5801 | NA | NA |
| 12 | 8093.308 | NA | NA | NA |

According to the value of the f statistic and the p value, there is evidence that at least one of the $\beta$ values is not zero. Therefore indicating that at least one of the predictors has an effect on the response variable.

**LMR 3.7 c)**

Summary and ANOVA for reduced model

```
##
## Call:
## lm(formula = Distance ~ RStr + LStr, data = punting)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.280  -9.583   3.147  10.266  26.450
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  12.8490    33.0334   0.389    0.705
## RStr          0.7208     0.4913   1.467    0.173
## LStr          0.2011     0.4883   0.412    0.689
##
## Residual standard error: 17.24 on 10 degrees of freedom
## Multiple R-squared:  0.6327, Adjusted R-squared:  0.5592
## F-statistic: 8.611 on 2 and 10 DF,  p-value: 0.00669
```

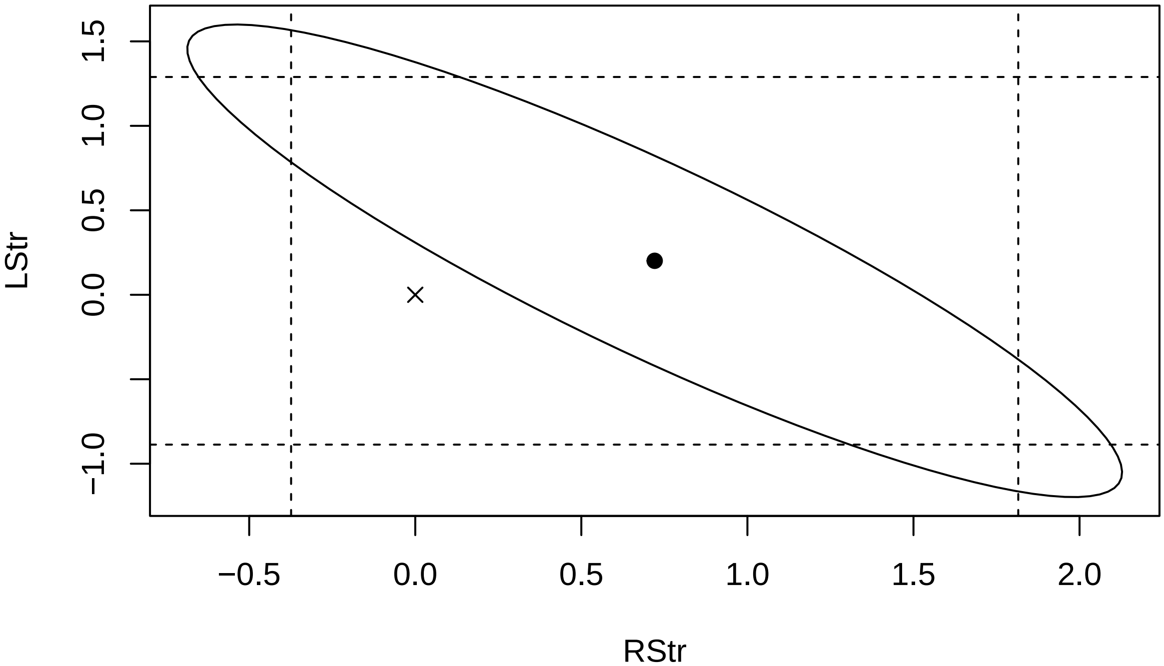
ANOVA table for reduced model

| df_vector | SS | MS | F_stat | p_value |
|---:|---:|---:|---:|---:|
| 2 | 5120.236 | 2560.1178 | 8.611016 | 0.0066896 |
| 10 | 2973.073 | 297.3073 | NA | NA |
| 12 | 8093.308 | NA | NA | NA |

According to the p values, neither RStr nor LStr have a significant effect on the the response variable. They are both above the $\alpha$ level of 0.05. Strictly comparing the p values for the two predictors, RStr does appear to have a greater effect on distance than LStr.

**LMR 3.7 d)**

```
##                    2.5 %    97.5 %
## (Intercept) -60.7540194 86.452019
## RStr         -0.3738903  1.815490
## LStr         -0.8868949  1.289095
```




The origin is not within the ellipse, therefore we can reject $H_0$ of $\beta_{RStr} = \beta_{LStr}$. Additionally, the origin lies within the boundaries of the 95% confident limits for both variables so we cannot reject the either null hypothesis of $\beta_{RStr}$ = 0 or $\beta_{LStr} = 0$. This supports what was was found in Part c). In regards to which predictor has a greater effect on the response, we can see that the origin is well within the 95% CL for LStr, but closer to the limit for RStr.

## Code

```
library(faraway)
library(tidyverse)
library(ellipse)
null_model1 <- lm(total ~ 1, data = sat)
#summary(null_model1)
#knitr::kable(anova(null_model1))

sat_model1 <- lm(total ~ expend + ratio + salary, data = sat)
summary(sat_model1)
print(anova(sat_model1))

SST <- sum((sat$total - mean(sat$total))^2)
SSR <- sum((sat_model1$fitted.values - mean(sat$total))^2)
SSE <- sum((sat$total - sat_model1$fitted.values)^2)
MSR <- SSR/(sat_model1$rank - 1)
MSE <- SSE/(nrow(sat) - sat_model1$rank)
df_vector <- c(sat_model1$rank - 1, nrow(sat) - sat_model1$rank, nrow(sat) - 1)
SS <- c(SSR, SSE, SST)

anova_table <- as_tibble(cbind(df_vector, SS))

anova_table <- anova_table %>% mutate(MS = SS/df_vector)

anova_table[3,3] <- NA

F_stat <- c(as.numeric(anova_table[1,3] / anova_table[2,3]), NA, NA)

p_value <- c(1 - (pf(F_stat[1], (sat_model1$rank - 1) , (nrow(sat) - sat_model1$rank))), NA, NA)

anova_table <- cbind(anova_table, F_stat, p_value)

knitr::kable(anova_table)

sat_model2 <- lm(total ~ expend + ratio + salary + takers, data = sat)
summary(sat_model2)
print(anova(sat_model1, sat_model2))
print(anova(sat_model2))

SST <- sum((sat$total - mean(sat$total))^2)
SSR <- sum((sat_model2$fitted.values - mean(sat$total))^2)
SSE <- sum((sat$total - sat_model2$fitted.values)^2)
MSR <- SSR/(sat_model2$rank - 1)
MSE <- SSE/(nrow(sat) - sat_model2$rank)
df_vector <- c(sat_model2$rank - 1, nrow(sat) - sat_model2$rank, nrow(sat) - 1)
SS <- c(SSR, SSE, SST)

anova_table <- as_tibble(cbind(df_vector, SS))

anova_table <- anova_table %>% mutate(MS = SS/df_vector)

anova_table[3,3] <- NA

F_stat <- c(as.numeric(anova_table[1,3] / anova_table[2,3]), NA, NA)

p_value <- c(1 - (pf(F_stat[1], (sat_model2$rank - 1) , (nrow(sat) - sat_model2$rank))), NA, NA)
```

```r
anova_table <- cbind(anova_table, F_stat, p_value)

knitr::kable(anova_table)

pun_model1 <- lm(Distance ~ RStr + LStr + RFlex + LFlex, data = punting)
summary(pun_model1)
null_model1 <- lm(Distance ~ 1, data = punting)
anova(null_model1, pun_model1)

SST <- sum((punting$Distance - mean(punting$Distance))^2)
SSR <- sum((pun_model1$fitted.values - mean(punting$Distance))^2)
SSE <- sum((punting$Distance - pun_model1$fitted.values)^2)
MSR <- SSR/(pun_model1$rank - 1)
MSE <- SSE/(nrow(punting) - pun_model1$rank)
df_vector <- c(pun_model1$rank - 1, nrow(punting) - pun_model1$rank, nrow(punting) - 1)
SS <- c(SSR, SSE, SST)

anova_table <- as_tibble(cbind(df_vector, SS))

anova_table <- anova_table %>% mutate(MS = SS/df_vector)

anova_table[3,3] <- NA

F_stat <- c(as.numeric(anova_table[1,3] / anova_table[2,3]), NA, NA)

p_value <- c(1 - (pf(F_stat[1], (pun_model1$rank - 1) , (nrow(punting) - pun_model1$rank))), NA, NA)

anova_table <- cbind(anova_table, F_stat, p_value)

knitr::kable(anova_table)
pun_model2 <- lm(Distance ~ RStr + LStr, data = punting)
summary(pun_model2)
SST <- sum((punting$Distance - mean(punting$Distance))^2)
SSR <- sum((pun_model2$fitted.values - mean(punting$Distance))^2)
SSE <- sum((punting$Distance - pun_model2$fitted.values)^2)
MSR <- SSR/(pun_model2$rank - 1)
MSE <- SSE/(nrow(punting) - pun_model2$rank)
df_vector <- c(pun_model2$rank - 1, nrow(punting) - pun_model2$rank, nrow(punting) - 1)
SS <- c(SSR, SSE, SST)

anova_table <- as_tibble(cbind(df_vector, SS))

anova_table <- anova_table %>% mutate(MS = SS/df_vector)

anova_table[3,3] <- NA

F_stat <- c(as.numeric(anova_table[1,3] / anova_table[2,3]), NA, NA)

p_value <- c(1 - (pf(F_stat[1], (pun_model2$rank - 1) , (nrow(punting) - pun_model2$rank))), NA, NA)

anova_table <- cbind(anova_table, F_stat, p_value)

knitr::kable(anova_table)
confint(pun_model2)

plot(ellipse(pun_model2, c(2,3)), type = 'l')
points(pun_model2$coefficients[2], pun_model2$coefficients[3], pch = 19)
```

```r
points(0,0, pch = 4)
abline(v = confint(pun_model2)[2, ], lty = 2)
abline(h = confint(pun_model2)[3, ], lty = 2)
```