HW2 - Josh Janda

Josh Janda

19 February, 2020

```
#imports
library(faraway)
library(tidyverse)
library(knitr)
library(ellipse)
library(doParallel)
```

Problem 1

Using the sat data from the faraway library:

A.

Fit a model with total sat score as the response and expend, ratio and salary as predictors. Test the hypothesis that $\beta_{salary} = 0$. Test the hypothesis that $\beta_{salary} = \beta_{ratio} = \beta_{expend} = 0$. Do any of these predictors have an effect on the response?.

```
prob_1_a_model = lm(total ~ expend + ratio + salary, data = sat)

prob_1_a_model_summary = summary(prob_1_a_model)
prob_1_a_fstat = prob_1_a_model_summary$fstat[1]
prob_1_a_pval = pf(prob_1_a_model_summary$fstat[1],
                   prob_1_a_model_summary$fstat[2],#numerator df
                   prob_1_a_model_summary$fstat[3],#denominator df
                   lower.tail = FALSE)

prob_1_a_model_summary
```

```
##
## 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
```

Looking at the output above for testing the hypothesis that $\beta_{salary} = 0$, we get a t-statistic of -1.8784372 with an associated p-value of 0.0666677. Therefore, we cannot reject the null hypothesis that $\beta_{salary} = 0$ at $\alpha = 0.05$.

Also considering the output above, for testing the hypothesis that $\beta_{salary} = \beta_{ratio} = \beta_{expend} = 0$, we get an F-statistic of 4.0662033 with an associated p-value of 0.0120861. Therefore, we can reject the null hypothesis that $\beta_{salary} = \beta_{ratio} = \beta_{expend} = 0$ at $\alpha = .05$.

Judging by the summary output above, while all of these variables are jointly significant, no specific variable is statistically significant looking at their t-statistics and associated p-values.

B.

Now add takers to the model. Test the hypothesis that $\beta_{takers} = 0$. Compare this model to the previous one using an F-test. Demonstrate that the F-test is equivalent to the t-test.

```r
prob_1_b_model = lm(total ~ expend + ratio + salary + takers, data = sat)

prob_1_b_model_summary = summary(prob_1_b_model)
prob_1_b_anova = anova(prob_1_a_model, prob_1_b_model)

prob_1_b_takers_tval = prob_1_b_model_summary$coefficients[5, 3]
prob_1_b_takers_pval = prob_1_b_model_summary$coefficients[5, 4]
prob_1_b_model_summary
```

```
##
## 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
```

For testing the null hypothesis that $\beta_{takers} = 0$, we get a t-statistic of -12.5593745 with an associated p-value of $2.6065588 \times 10^{-16}$. Therefore, we can reject the null hypothesis at $\alpha = .05$ that $\beta_{takers} = 0$.

For comparing this model to the previous model using an F-Test, I will use ANOVA.

```
prob_1_b_anova
```

```
## 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
```

```
prob_1_b_anova_pval = prob_1_b_anova$`Pr(>F)`[2]
```

Looking at the ANOVA table above for testing between the first model (reduced model), and the second model (full model), we get an F-statistic of 157.7378885 with an associated p-value of $2.6065588 \times 10^{-16}$. Recall that the t-statistic for the hypothesis that $\beta_{takers} = 0$ was -12.5593745. Since we only added one new variable to the original reduced model, the F-statistic obtained from this ANOVA test is actually the squared value of the t-statistic from the newly added variable ($t^2 = F$).

With that said, it can be shown that the F-test is equivalent to the t-test in this scenario. (takers t-statistic squared) 157.7378885 = 157.7378885 (ANOVA F-statistic), with the associated p-values also being equivalent, (takers p-value) $2.6065588 \times 10^{-16} = 2.6065588 \times 10^{-16}$ (ANOVA p-value).

Therefore, when adding one additional variable to a model, it can be shown that the F-test is equivalent to the t-test.

Problem 2

For the prostate data from the faraway library, fit a model with lpsa as the response and the other variables as predictors:

```
prob_2_model = lm(lpsa ~ ., data = prostate)
```

A.

Compare 90% and 95% CIs for the parameter associated with age.

```
prob_2_age_90_ci= confint(prob_2_model, "age", level = .90)
prob_2_age_95_ci= confint(prob_2_model, "age", level = .95)

prob_2_age_cis = data.frame(Lower = c('.90 (90%)' = prob_2_age_90_ci[1],
                                      '.95 (95%)' = prob_2_age_95_ci[1]),
                            Upper = c('.90 (90%)' = prob_2_age_90_ci[2],
                                      '.95 (95%)' = prob_2_age_95_ci[2]))
kable(prob_2_age_cis, col.names = c("Lower Bound", "Upper Bound"))
```

	Lower Bound	Upper Bound
.90 (90%)	-0.0382102	-0.0010642
.95 (95%)	-0.0418406	0.0025663

The table above displays the lower and upper bounds for the 90% and 95% confidence intervals for the variable `age`. For the 90% confidence interval, we cannot reject the null hypothesis that $\beta_{age} = 0$ at $\alpha = .10$. This can be concluded since this confidence interval *does not* include zero. For the 95% confidence interval, we can reject the null hypothesis that $\beta_{age} = 0$ at $\alpha = .05$. This can be concluded since this confidence interval *does* include zero. This makes sense since a 95% confidence interval has a larger range, due to the increased confidence levels.

Overall, comparing these confidence intervals tells us that the p-value for β_{age} lies between .05 and .10, and is statistically significant at $\alpha = .10$.

B.

Remove all predictors that are not significant at the 5% level. Test this model against the original model. Which model is preferred?

```r
summary(prob_2_model)
```

```
##
## Call:
## lm(formula = lpsa ~ ., data = prostate)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7331 -0.3713 -0.0170  0.4141  1.6381
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.669337   1.296387   0.516  0.60693
## lcavol       0.587022   0.087920   6.677 2.11e-09 ***
## lweight      0.454467   0.170012   2.673  0.00896 **
## age         -0.019637   0.011173  -1.758  0.08229 .
## lbph         0.107054   0.058449   1.832  0.07040 .
## svi          0.766157   0.244309   3.136  0.00233 **
## lcp         -0.105474   0.091013  -1.159  0.24964
## gleason      0.045142   0.157465   0.287  0.77503
## pgg45        0.004525   0.004421   1.024  0.30886
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7084 on 88 degrees of freedom
## Multiple R-squared:  0.6548, Adjusted R-squared:  0.6234
## F-statistic: 20.86 on 8 and 88 DF,  p-value: < 2.2e-16
```

The variables that are not significant at $\alpha = .05$ are *age*, *lbph*, *lcp*, *gleason*, and *pgg45*. I will be removing these predictors and creating a reduced model.

```
prob_2_reduced_model = lm(lpsa ~ lcavol + lweight + svi,
                          data = prostate)

prob_2_anova = anova(prob_2_reduced_model, prob_2_model)
prob_2_anova_fval = prob_2_anova$F[2]
prob_2_anova_pval = prob_2_anova$`Pr(>F)`[2]

prob_2_anova
```

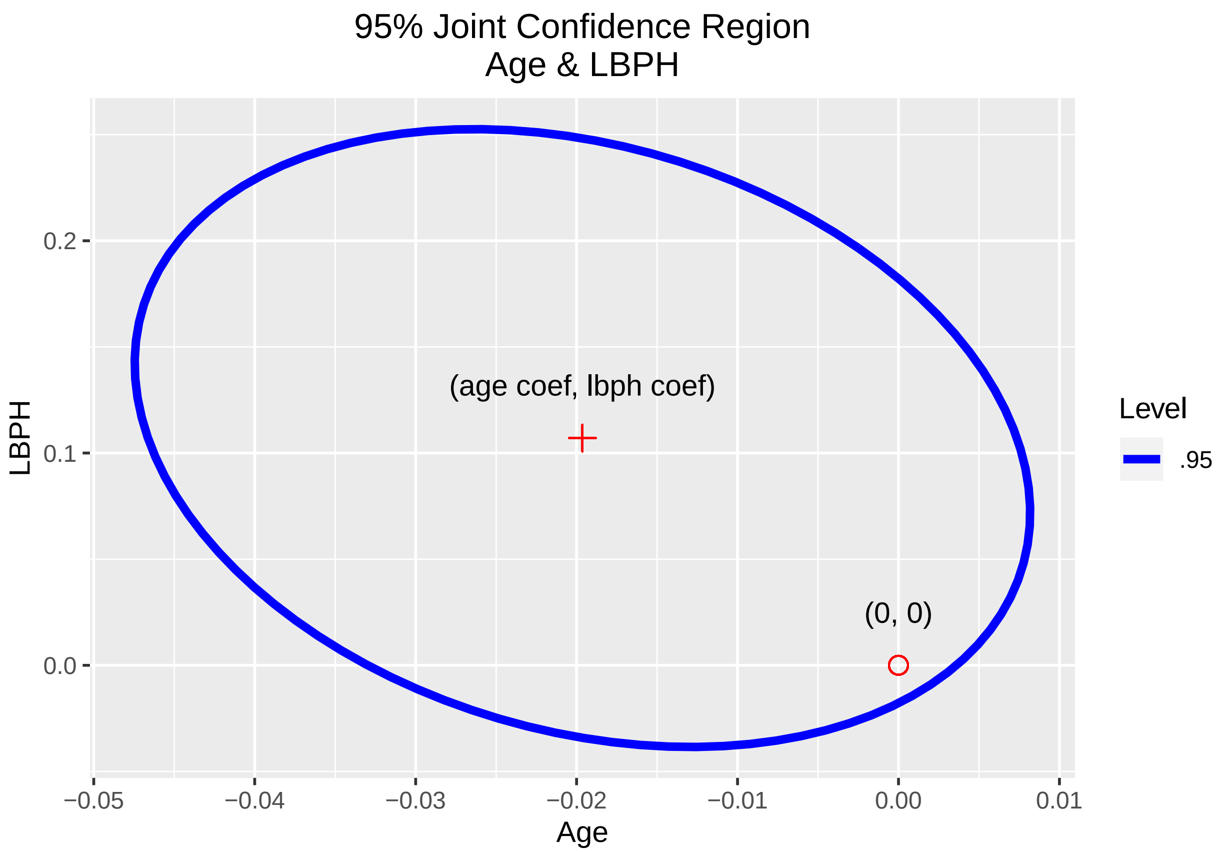
```
## Analysis of Variance Table
##
## Model 1: lpsa ~ lcavol + lweight + svi
## Model 2: lpsa ~ lcavol + lweight + age + lbph + svi + lcp + gleason +
##     pgg45
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     93 47.785
## 2     88 44.163  5    3.6218 1.4434 0.2167
```

Looking at the ANOVA table above which compares the original model to the reduced model that does not include any variables that are not significant at $\alpha = .05$, we get an F-statistic of 1.4433869 and an associated p-value of 0.2167334. Therefore, we cannot reject the null hypothesis that $\beta_{age} = \beta_{lbph} = \beta_{lcp} = \beta_{gleason} = \beta_{pgg45} = 0$. I can therefore conclude that the reduced model is the preferred model using the F-test.

C.

Compute and display a 95% joint confidence region for the parameters associated with age and lbph. Plot the origin on this display. The location of the origin on the display tell us the outcome of a certain hypothesis test. State that test and its outcome.

```
prob_2_c_cr = data.frame(ellipse(prob_2_model, c(4, 5), level = .95))

ggplot(data=prob_2_c_cr, aes(x=age, y=lbph)) +
  geom_path(aes(linetype=".95"), color = "blue", size=1.5) +
  geom_point(x=coef(prob_2_model)[4], y=coef(prob_2_model)[5], shape=3, size=3, colour='red') +
  geom_point(x=0, y=0, shape=1, size=3, colour='red') +
  scale_linetype_manual(values = c(".95" = 1)) +
  labs(x = "Age", y = "LBPH", title = "95% Joint Confidence Region\nAge & LBPH",
       linetype = "Level") +
  theme(plot.title = element_text(hjust = 0.5)) +
  annotate(geom = "text",
           x = coef(prob_2_model)[4],
           y = coef(prob_2_model)[5] + .025,
           label = "(age coef, lbph coef)") +
  annotate(geom = "text",
           x = 0,
           y = .025,
           label = "(0, 0)")
```



Looking at the plot above of the 95% joint confidence region between the variables *age* and *lbph*, we can conclude that the point (0,0) is inside the ellipsoid. The point (0,0) being inside the 95% ellipsoid region tells us that we cannot reject the null hypothesis that *age* and *lbph* are not statistically significant at $\alpha = .05$.

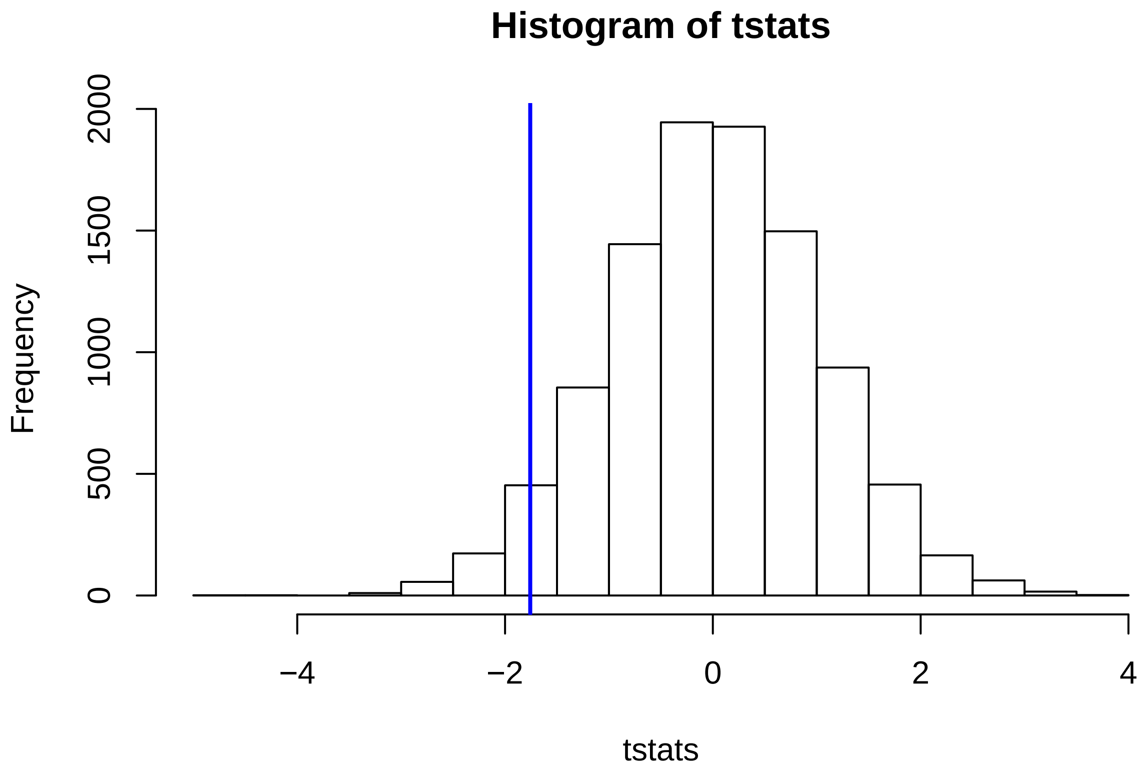
D.

In class we discussed a permutation test corresponding to the F −test for the significance of a set of predictors. Execute the permutation test corresponding to the t-test for age in this model.

```
iterations = 10000
prob_2_age_tstat = summary(prob_2_model)$coefficients[4, 3]
cl = makeCluster(detectCores(logical = FALSE))
registerDoParallel(cl)

tstats = foreach(i = 1:iterations, .combine = "rbind", .packages = "faraway") %dopar% {

        prostate_data = prostate
        prostate_data$age = prostate[sample(97), "age"]
        permu_mod = lm(lpsa ~ ., data = prostate_data)
        summary(permu_mod)$coefficients[4, 3]

}
```

```
prob_2_permu_pval = length(tstats[abs(tstats) > abs(prob_2_age_tstat)]) / iterations
```

```r
hist(tstats)
abline(v = prob_2_age_tstat, col = "blue", lwd = 2)
```



The p-value for this permutation test corresponding to the t-test for age is 0.0856. Therefore, we cannot reject the null hypothesis at $\alpha = .05$ that the *age* coefficient variable is statistically significantly different than zero.

Problem 3

In the punting data from the faraway library we find average distance and hang times of 10 punts of an American football as related to various measures of leg strength for 13 volunteers.

A.

Fit a regression model with Distance as the response, and the right and left strengths and flexibilities as predictors. Which predictors are significant at the 5% level?

```r
prob_3_model = lm(Distance ~ LStr + RStr + RFlex + LFlex, data = punting)
prob_3_model_fstat = summary(prob_3_model)$fstat[1]
prob_3_model_pval = pf(prob_3_model_fstat, 4, 8, lower.tail = FALSE)
summary(prob_3_model)
```

```
##
```

```
## Call:
## lm(formula = Distance ~ LStr + RStr + 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
## LStr         -0.1862     0.5130  -0.363    0.726
## RStr          0.5116     0.4856   1.054    0.323
## 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
```

Looking at the summary of the model above, no predictors are significant at the 5% level (or even the 10% level).

B.

Use an F-test to determine whether collectively these four predictors are significant at the 5% level.

Using the same summary table above from part A, we get an F-statistic of 5.5899409 with an associated p-value of 0.0190248. Therefore, we can reject the null hypothesis that $\beta_{LStr} = \beta_{RStr} = \beta_{RFlex} = \beta_{LFlex} = 0$ at $\alpha = .05$ and can conclude that the four predictors are jointly significant.

C.

Relative to the model in (a), test whether the right and left strength have the same effect.

The null hypothesis for this test is $\beta_{LStr} = \beta_{RStr}$ or $\beta_{LStr} - \beta_{RStr} = 0$.

```r
prob_3_c_model = lm(Distance ~  I(LStr - RStr) + RFlex + LFlex, data = punting)
prob_3_c_anova = anova(prob_3_c_model, prob_3_model)
prob_3_c_fstat = prob_3_c_anova$F[2]
prob_3_c_pval = prob_3_c_anova$`Pr(>F)`[2]
prob_3_c_anova
```

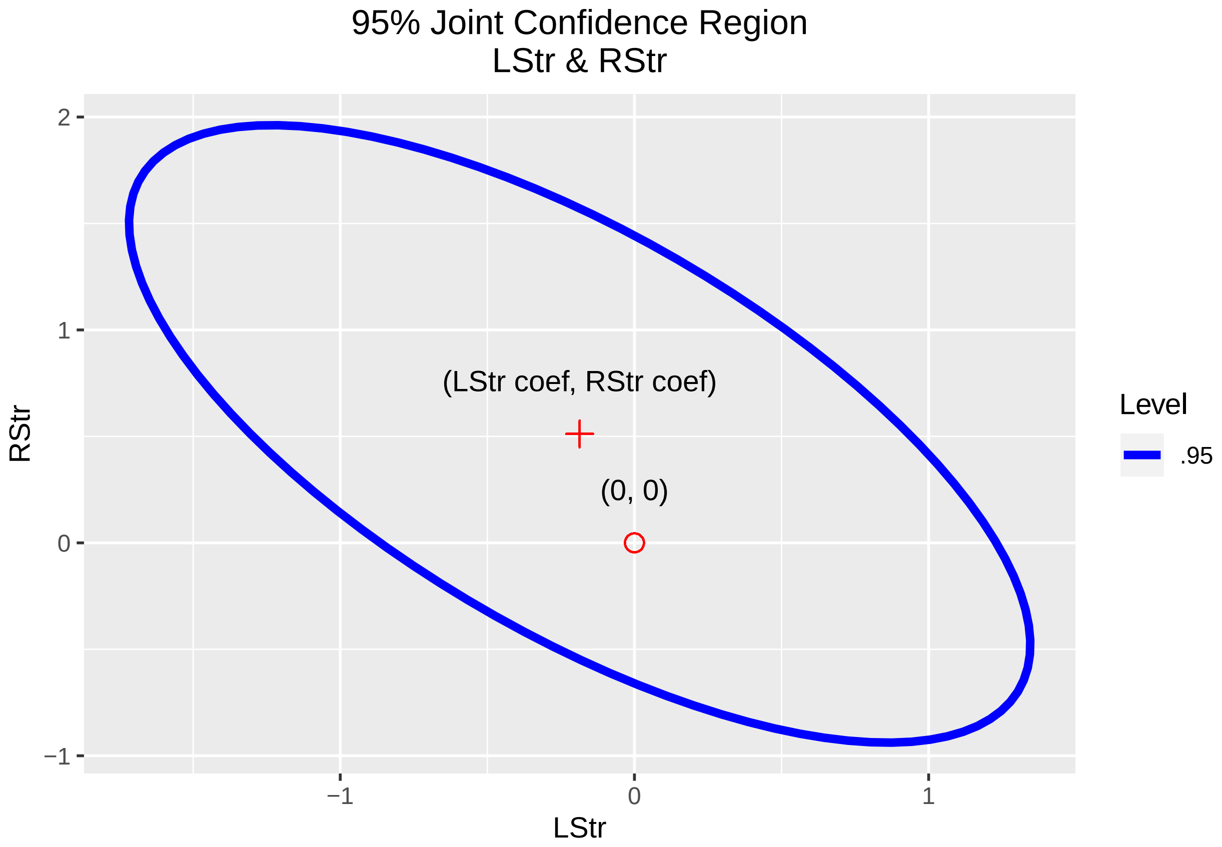
```
## Analysis of Variance Table
##
## Model 1: Distance ~ I(LStr - RStr) + RFlex + LFlex
## Model 2: Distance ~ LStr + RStr + RFlex + LFlex
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1      9 2310.5
## 2      8 2132.6  1    177.86 0.6672 0.4377
```

Looking at the output above, the F-statistic obtained from the null hypothesis $\beta_{LStr} = \beta_{RStr}$ is 0.6671876 with an associated p-value of 0.4376796. Therefore, we cannot reject the null hypothesis that right and left strength have the same effect on punting distance at $\alpha = .05$.

D.

Construct a 95% confidence region for $(\beta_{RStr}, \beta_{LStr})$. Explain how the test in (c) relates to this region.

```r
prob_3_d_cr = data.frame(ellipse(prob_3_model, c(2, 3), level = .95))

ggplot(data=prob_3_d_cr, aes(x=LStr, y=RStr)) +
  geom_path(aes(linetype=".95"), color = "blue", size=1.5) +
  geom_point(x=coef(prob_3_model)[2], y=coef(prob_3_model)[3], shape=3, size=3, colour='red') +
  geom_point(x=0, y=0, shape=1, size=3, colour='red') +
  scale_linetype_manual(values = c(".95" = 1)) +
  labs(x = "LStr", y = "RStr", title = "95% Joint Confidence Region\nLStr & RStr",
       linetype = "Level") +
  theme(plot.title = element_text(hjust = 0.5)) +
  annotate(geom = "text",
           x = coef(prob_3_model)[2],
           y = coef(prob_3_model)[3] + .25,
           label = "(LStr coef, RStr coef)") +
  annotate(geom = "text",
           x = 0,
           y = .25,
           label = "(0, 0)")
```



Looking at the confidence region above, we can see that the point (0, 0) lies inside the confidence region meaning that *LStr* and *RStr* are jointly insignificant at $\alpha = .05$. With both variables being jointly insignificant, we can conclude that the variable *LStr* has the same effect as *RStr* since we cannot say that the variables are significantly different at $\alpha = .05$.

E.

Fit a model to test the hypothesis that it is total leg strength defined by adding the right and left leg strengths that is sufficient to predict the response, in comparison to using individual left and right strengths

```
prob_3_e_model_full = lm(Distance ~ LStr + RStr, data = punting)
prob_3_e_model_rdc = lm(Distance ~ I(LStr + RStr), data = punting)
prob_3_e_anova = anova(prob_3_e_model_rdc, prob_3_e_model_full)

prob_3_e_tstat = summary(prob_3_e_model_rdc)$coefficients[2, 3]
prob_3_e_pval = summary(prob_3_e_model_rdc)$coefficients[2, 4]

summary(prob_3_e_model_rdc)
```

```
##
## Call:
## lm(formula = Distance ~ I(LStr + RStr), data = punting)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -27.632 -11.531   2.171   8.443  30.672
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     14.0936    31.8838   0.442  0.66703
## I(LStr + RStr)   0.4601     0.1082   4.252  0.00136 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.68 on 11 degrees of freedom
## Multiple R-squared:  0.6217, Adjusted R-squared:  0.5874
## F-statistic: 18.08 on 1 and 11 DF,  p-value: 0.001361
```

Looking at the summary output above, we get a t-statistic of 4.2521435 and an associated p-value of 0.0013608. Therefore, we can reject the null hypothesis that $\beta_{LStr+RStr} = 0$ at $\alpha = .05$. In other words, total leg strength is sufficient enough to predict punting distance compared to using individual leg strengths.

Problem 5

For the prostate data, fit a model with lpsa as the response and the other variables as predictors.

```
prob_5_mod = lm(lpsa ~ ., data = prostate)
```

A.

Suppose a new patient with the following values arrives:

```
patient_vals = data.frame(lcavol = 1.44692, lweight = 3.62301, age = 65,
                          lbph = .3001, svi = 0, lcp = -.79851, gleason = 7,
                          pgg45 = 15)
```

Predict lpsa for this patient along with an appropriate 95% CI.

```
patient_preds = predict(prob_5_mod, newdata = patient_vals, level = .95,
                        interval = "confidence")

kable(patient_preds, col.names = c("Predicted Value",
                                   "Lower 95% Value",
                                   "Upper 95% Value"))
```

Predicted Value	Lower 95% Value	Upper 95% Value
2.389053	2.172437	2.605669

Looking at the table above, we get a predicted point value of *lpsa* equal to ~2.389 with a 95% lower bound of ~2.172 and a 95% upper bound of ~ 2.606.

B.

Predict the last question for a patient with the same values except that he is age 20. Explain why the CI is wider.

```
patient_vals2 = data.frame(lcavol = 1.44692, lweight = 3.62301, age = 20,
                          lbph = .3001, svi = 0, lcp = -.79851, gleason = 7,
                          pgg45 = 15)

patient_preds2 = predict(prob_5_mod, newdata = patient_vals2, level = .95,
                         interval = "confidence")

kable(patient_preds2, col.names = c("Predicted Value", "Lower 95% Value",
                                    "Upper 95% Value"))
```

Predicted Value	Lower 95% Value	Upper 95% Value
3.272726	2.260444	4.285007

Looking at the table above, we get a predicted point value of *lpsa* equal to ~3.273 with a 95% lower bound of ~2.260 and a 95% upper bound of ~ 4.285. The confidence interval for this younger patient is much wider than the older patient holding all other values constant. Note that he *age* variable itself is statistically insignificant. Also note that our data contains no observations with an age less than 50, so when predicting on an age of 20 it is a value that our model has not seen and cannot accurately predict (due to being far from mean value), which will result in a larger confidence interval.

C.

For the model of the previous question, remove all predictors that are not significant at the 5% level. Now recompute the predictions of the previous question. Are the CIs wider or narrower? Which predictions would you prefer?

```r
prob_5_c_mod = lm(lpsa ~ lcavol + lweight + svi, data = prostate)

patient_vals3 = data.frame(lcavol = 1.44692, lweight = 3.62301, svi = 0)
patient_preds3 = predict(prob_5_c_mod, newdata = patient_vals3, level = .95,
                         interval = "confidence")

kable(patient_preds3, col.names = c("Predicted Value",
                                    "Lower 95% Value",
                                    "Upper 95% Value"))
```

Predicted Value	Lower 95% Value	Upper 95% Value
2.372534	2.197274	2.547794

The confidence interval obtained above is narrower than the previous confidence intervals with the full model. This is due to removing all insignifant variables at $\alpha = .05$ from the model, which will slightly increase the standard error of the model. Due to reducing the number of insignificant variables in the model, we will in turn reduce the standard error of the predicted point and ultimately create a narrower confidence interval.