Homework Two

Chiayu Tu (Louis Tu)

2022-10-16

Question One

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

(a)

Compute 90 and 95% CIs for the parameter associated with age. Using just these intervals, what could we have deduced about the p-value for age in the regression summary?

```
#creat a model
model_l <- lm(lpsa ~ ., data = prostate)
summary(model_l)
```

```
##
## 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
```

```r
#Compute 90% CIs
confint(model_l, level = 0.90)
```

```
##                       5 %         95 %
## (Intercept) -1.485718237  2.824391633
## lcavol       0.440867156  0.733176497
## lweight      0.171846568  0.737088281
## age         -0.038210200 -0.001064151
## lbph         0.009890745  0.204217317
## svi          0.360029029  1.172285623
## lcp         -0.256770899  0.045822373
## gleason     -0.216620186  0.306903382
## pgg45       -0.002824333  0.011874796
```

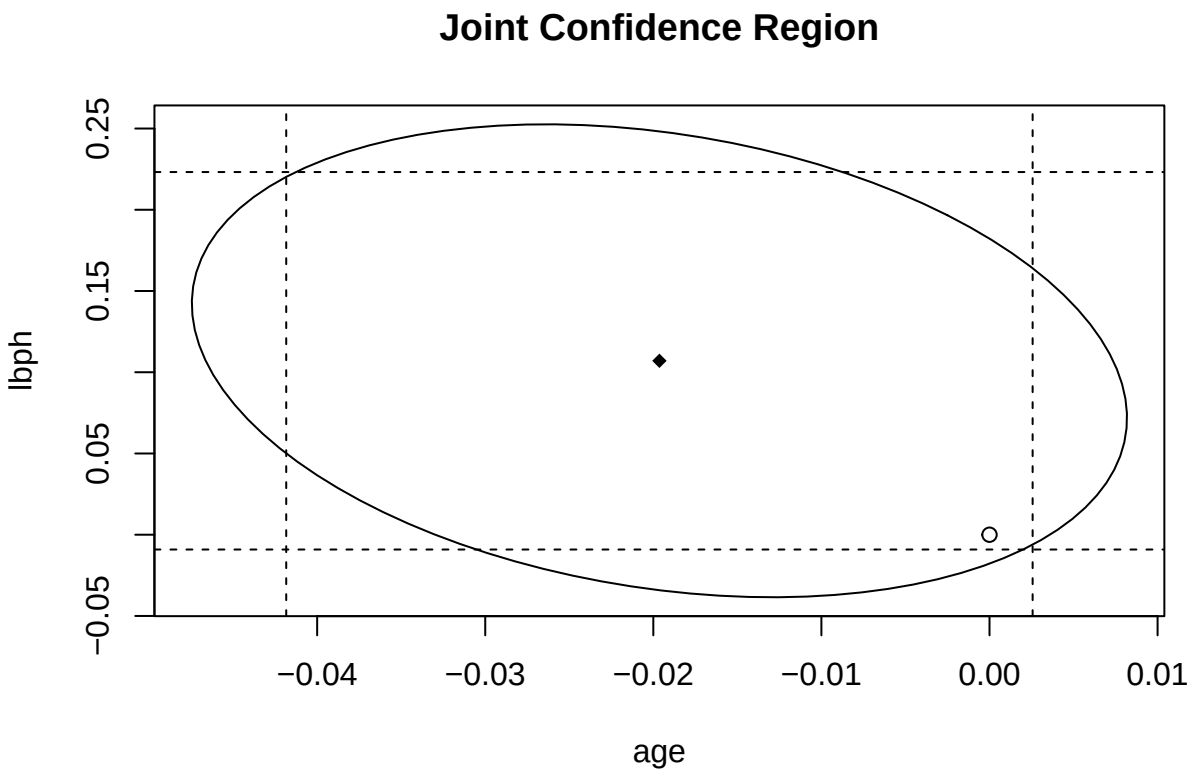
```r
#Compute 95% CIs
confint(model_l, level = 0.95)
```

```
##                     2.5 %      97.5 %
## (Intercept) -1.906960983 3.245634379
## lcavol       0.412298699 0.761744954
## lweight      0.116603435 0.792331414
## age         -0.041840618 0.002566267
## lbph        -0.009101499 0.223209561
## svi          0.280644232 1.251670420
## lcp         -0.286344443 0.075395916
## gleason     -0.267786053 0.358069248
## pgg45       -0.004260932 0.013311395
```

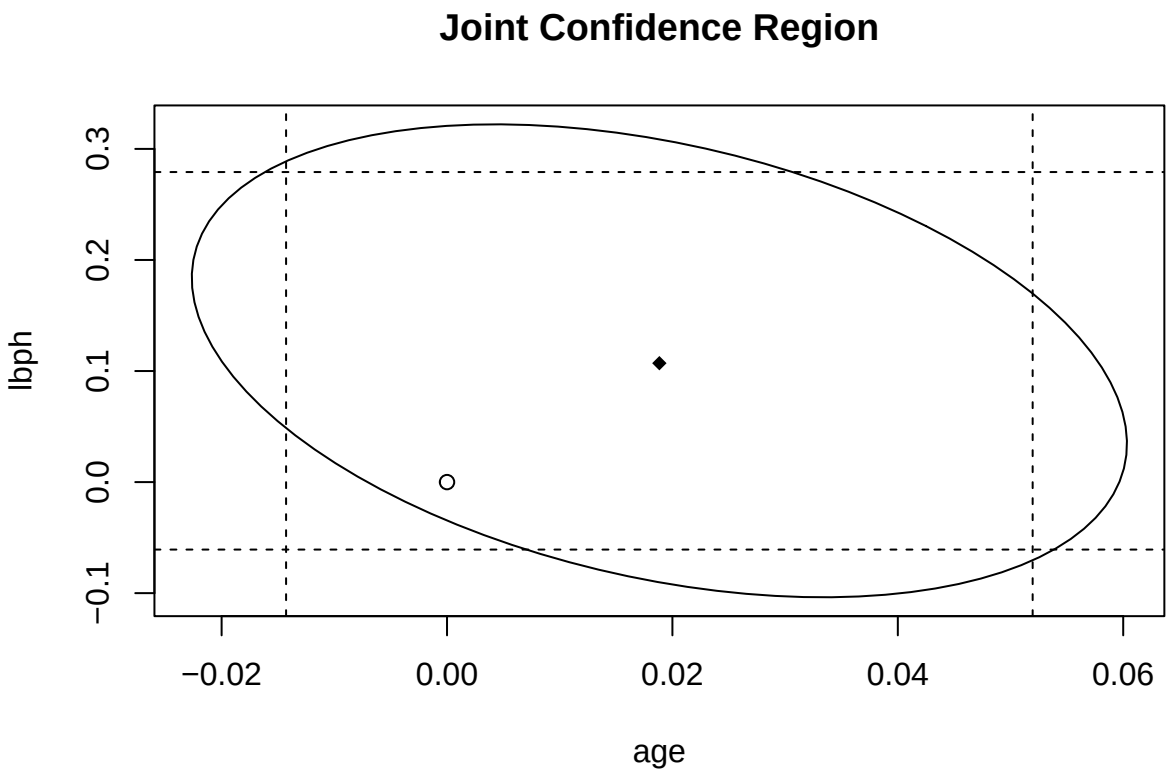
(b)

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 tells us the outcome of a certain hypothesis test. State that test and its outcome.

```r
#Origin plot
plot(ellipse(model_l, c("age", "lbph")),
     type = "l",
     main = "Joint Confidence Region")
points(coef(model_l)["age"], coef(model_l)["lbph"], pch = 18)
abline(v = confint(model_l)["age", ], lty = 2)
abline(h = confint(model_l)["lbph", ], lty = 2)
points(0, 0)
```




```
#Creat a model
model_l_1 <- lm(lpsa ~ age + lbph, data = prostate)
#Ellipsee plot
plot(ellipse(model_l_1, c("age", "lbph")),
     type = "l",
     main = "Joint Confidence Region")
points(coef(model_l_1)["age"], coef(model_l)["lbph"], pch = 18)
abline(v = confint(model_l_1)["age", ], lty = 2)
abline(h = confint(model_l_1)["lbph", ], lty = 2)
points(0, 0)
```




```
#
anova(model_l_1, model_l)
```

```
## Analysis of Variance Table
##
## Model 1: lpsa ~ age + lbph
## Model 2: lpsa ~ lcavol + lweight + age + lbph + svi + lcp + gleason +
##     pgg45
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1     94 122.124
## 2     88  44.163  6    77.961 25.891 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(d)

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

```
#Predictors are significant at the 5% level
model_l_2 <- lm(lpsa ~ lcavol + lweight + svi, data = prostate)
summary(model_l_2)
```

```
##
```

```
## Call:
## lm(formula = lpsa ~ lcavol + lweight + svi, data = prostate)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.72964 -0.45764  0.02812  0.46403  1.57013
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.26809    0.54350  -0.493  0.62298
## lcavol       0.55164    0.07467   7.388  6.3e-11 ***
## lweight      0.50854    0.15017   3.386  0.00104 **
## svi          0.66616    0.20978   3.176  0.00203 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7168 on 93 degrees of freedom
## Multiple R-squared:  0.6264, Adjusted R-squared:  0.6144
## F-statistic: 51.99 on 3 and 93 DF,  p-value: < 2.2e-16
```

```
anova(model_l, model_l_2)
```

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

Question Two

Thirty samples of cheddar cheese were analyzed for their content of acetic acid, hydrogen sulfide and lactic acid. Each sample was tasted and scored by a panel of judges and the average taste score produced. Use the cheddar data to answer the following:

(2a)

Fit a regression model with taste as the response and the three chemical contents as predictors. Identify the predictors that are statistically significant at the 5% level.

```
#fit the model using lm with taste as the response
model_c <- lm(taste ~ Acetic + H2S + Lactic, data = cheddar)
summary(model_c)
```

```
##
## Call:
## lm(formula = taste ~ Acetic + H2S + Lactic, data = cheddar)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -17.390  -6.612  -1.009   4.908  25.449
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -28.8768    19.7354  -1.463  0.15540
## Acetic        0.3277     4.4598   0.073  0.94198
## H2S           3.9118     1.2484   3.133  0.00425 **
## Lactic       19.6705     8.6291   2.280  0.03108 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.13 on 26 degrees of freedom
## Multiple R-squared:  0.6518, Adjusted R-squared:  0.6116
## F-statistic: 16.22 on 3 and 26 DF,  p-value: 3.81e-06
```

```
print("H2S and Lactic are the only two variables which are significant at 5% level of significance.")
```

```
## [1] "H2S and Lactic are the only two variables which are significant at 5% level of significance."
```

(2b)

Acetic and H2S are measured on a log scale. Fit a linear model where all three predictors are measured on their original scale. Identify the predictors that are statistically significant at the 5% level for this model.

```
model_c_1 <- lm(taste ~ log(Acetic) + log(H2S) + Lactic, data = cheddar)
summary(model_c_1)
```

```
##
## Call:
## lm(formula = taste ~ log(Acetic) + log(H2S) + Lactic, data = cheddar)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.465  -7.324  -0.798   5.458  24.647
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -40.338     34.429  -1.172  0.25197
## log(Acetic)   -2.441     24.995  -0.098  0.92296
## log(H2S)      23.245      7.526   3.089  0.00474 **
## Lactic        20.151      8.688   2.319  0.02849 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.18 on 26 degrees of freedom
## Multiple R-squared:  0.6481, Adjusted R-squared:  0.6075
## F-statistic: 15.96 on 3 and 26 DF,  p-value: 4.349e-06
```

```
print("log H2S and Lactic still are the only two variables which are significant at 5% level of signifi
```

```
## [1] "log H2S and Lactic still are the only two variables which are significant at 5% level of signifi
```

(2c)

Can we use an F-test to compare these two models? Explain. Which model provides a better fit to the data? Explain your reasoning.

```
print("These two models can not be compared by F-test, because these two models are not nested")
```

```
## [1] "These two models can not be compared by F-test, because these two models are not nested"
```

(2d)

If H2S is increased 0.01 for the model used in (a), what change in the taste would be expected?

```
model_c$coefficients
```

```
## (Intercept)      Acetic         H2S      Lactic
## -28.8767696   0.3277413   3.9118411  19.6705434
```

Question Three

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

(3a)

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

```
model_p <- lm(Distance ~ RStr + LStr + RFlex + LFlex, data = punting)
summary(model_p)
```

```
##
## 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
```

```
print("All variables' p-value are less than 0.05, so we can know that there are not prodictors are signi
```

```
## [1] "All variables' p-value are less than 0.05, so we can know that there are not prodictors are sign
```

(3b)

Use an F-test to determine whether collectively these four predictors have a relationship to the response.

```
model_p_1 <- lm(Distance ~ 1, data = punting)
anova(model_p_1, model_p)
```

```
## 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
```

(3c)

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

```
model_p_2 <- lm(Distance ~ I(RStr + LStr) + RFlex + LFlex, data = punting)
summary(model_p_2)
```

```
##
## Call:
## lm(formula = Distance ~ I(RStr + LStr) + RFlex + LFlex, data = punting)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -21.698  -9.494  -5.155   9.081  20.611
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -71.2694    63.1447  -1.129    0.288
## I(RStr + LStr)   0.1741     0.1940   0.898    0.393
## RFlex            2.3137     1.4013   1.651    0.133
## LFlex           -0.5772     0.8035  -0.718    0.491
##
## Residual standard error: 15.94 on 9 degrees of freedom
## Multiple R-squared:  0.7174, Adjusted R-squared:  0.6232
## F-statistic: 7.615 on 3 and 9 DF,  p-value: 0.00769
```

```
anova(model_p_2, model_p)
```

```
## Analysis of Variance Table
##
## Model 1: Distance ~ I(RStr + LStr) + RFlex + LFlex
## Model 2: Distance ~ RStr + LStr + RFlex + LFlex
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1      9 2287.4
## 2      8 2132.6  1    154.72 0.5804  0.468
```

(3d)

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

```
confint(model_p, c("RStr", "LStr"))
```

```
##           2.5 %    97.5 %
## RStr -0.6080871 1.6313618
## LStr -1.3690973 0.9966981
```

(3e)

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 leg strengths.

```
model_p_3 <- lm(Distance ~ RStr + LStr, data = punting)
summary(model_p_3)
```

```
##
## 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
```

```
model_p_4 <- lm(Distance ~ I(RStr + LStr), data = punting)
summary(model_p_4)
```

```
##
## Call:
## lm(formula = Distance ~ I(RStr + LStr), 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(RStr + LStr)   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
```

```
anova(model_p_3, model_p_4)
```

```
## Analysis of Variance Table
##
## Model 1: Distance ~ RStr + LStr
## Model 2: Distance ~ I(RStr + LStr)
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     10 2973.1
## 2     11 3061.3 -1   -88.281 0.2969 0.5978
```

(3f)

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

```
model_p_5 <- lm(Distance ~ RStr + LStr + I(RFlex + LFlex), data = punting)
summary(model_p_5)
```

```
##
## Call:
## lm(formula = Distance ~ RStr + LStr + I(RFlex + LFlex), data = punting)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.510 -13.417   2.165   7.988  23.316
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -44.26189   63.52790  -0.697    0.504
## RStr               0.70392    0.48904   1.439    0.184
## LStr               0.01518    0.51703   0.029    0.977
## I(RFlex + LFlex)   0.46194    0.43975   1.050    0.321
##
## Residual standard error: 17.15 on 9 degrees of freedom
## Multiple R-squared:  0.6728, Adjusted R-squared:  0.5637
## F-statistic: 6.168 on 3 and 9 DF,  p-value: 0.01451
```

```
anova(model_p_5, model_p)
```

```
## Analysis of Variance Table
##
## Model 1: Distance ~ RStr + LStr + I(RFlex + LFlex)
## Model 2: Distance ~ RStr + LStr + RFlex + LFlex
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1      9 2648.4
## 2      8 2132.6  1    515.72 1.9346 0.2017
```

(3g)

Test for left–right symmetry by performing the tests in (c) and (f) simultaneously

```
model_p_6 <- lm(Distance ~ I(RStr + LStr) + I(RFlex + LFlex), data = punting)
summary(model_p_6)
```

```
##
## Call:
## lm(formula = Distance ~ I(RStr + LStr) + I(RFlex + LFlex), data = punting)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.948 -13.929   1.020   9.795  29.111
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -36.1525    60.9655  -0.593    0.566
## I(RStr + LStr)     0.3700     0.1430   2.588    0.027 *
## I(RFlex + LFlex)   0.4093     0.4228   0.968    0.356
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.73 on 10 degrees of freedom
## Multiple R-squared:  0.6541, Adjusted R-squared:  0.585
## F-statistic: 9.457 on 2 and 10 DF,  p-value: 0.004948
```

```
anova(model_p_6, model_p)
```

```
## Analysis of Variance Table
##
## Model 1: Distance ~ I(RStr + LStr) + I(RFlex + LFlex)
## Model 2: Distance ~ RStr + LStr + RFlex + LFlex
##   Res.Df    RSS Df Sum of Sq    F Pr(>F)
## 1     10 2799.1
## 2      8 2132.6  2    666.43 1.25  0.337
```

```
print("Based on this p-value we can not reject the null hypothesis of right-left symmetry.")
```

```
## [1] "Based on this p-value we can not reject the null hypothesis of right-left symmetry."
```

(3h)

Fit a model with Hang as the response and the same four predictors. Can we make a test to compare this model to that used in (a)? Explain.

```
model_p_7 <- lm(Hang ~ RStr + LStr + RFlex + LFlex, data = punting)
summary(model_p_7)
```

```
##
## Call:
## lm(formula = Hang ~ RStr + LStr + RFlex + LFlex, data = punting)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.36297 -0.13528 -0.07849  0.09938  0.35893
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.225239   1.032784  -0.218    0.833
## RStr         0.005153   0.007645   0.674    0.519
## LStr         0.007697   0.008077   0.953    0.369
## RFlex        0.019404   0.022631   0.857    0.416
## LFlex        0.004614   0.012998   0.355    0.732
##
## Residual standard error: 0.2571 on 8 degrees of freedom
## Multiple R-squared:  0.8156, Adjusted R-squared:  0.7235
## F-statistic: 8.848 on 4 and 8 DF,  p-value: 0.004925
```

```
print("These two models can not be compared by F-test, because these two models are not nested")
```

```
## [1] "These two models can not be compared by F-test, because these two models are not nested"
```