Chapter 3 Problem 7

Faraway, Julian J. Linear Models with R, Second Edition (Chapman & Hall/CRC Texts

Bruce Campbell

08 September, 2017

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.

- (a) 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?
- (b) Use an F-test to determine whether collectively these four predictors have a relationship to the response.
- (c) Relative to the model in (a), test whether the right and left leg strengths have the same effect.
- (d) Construct a 95% confidence region for (??RStr,??LStr). Explain how the test in (c) relates to this region - not required
- (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 leg strengths.
- (f) Relative to the model in (a), test whether the right and left leg flexibilities have the same effect.
- (g) Test for left-right symmetry by performing the tests in (c) and (f) simultaneously.
- (h) 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.

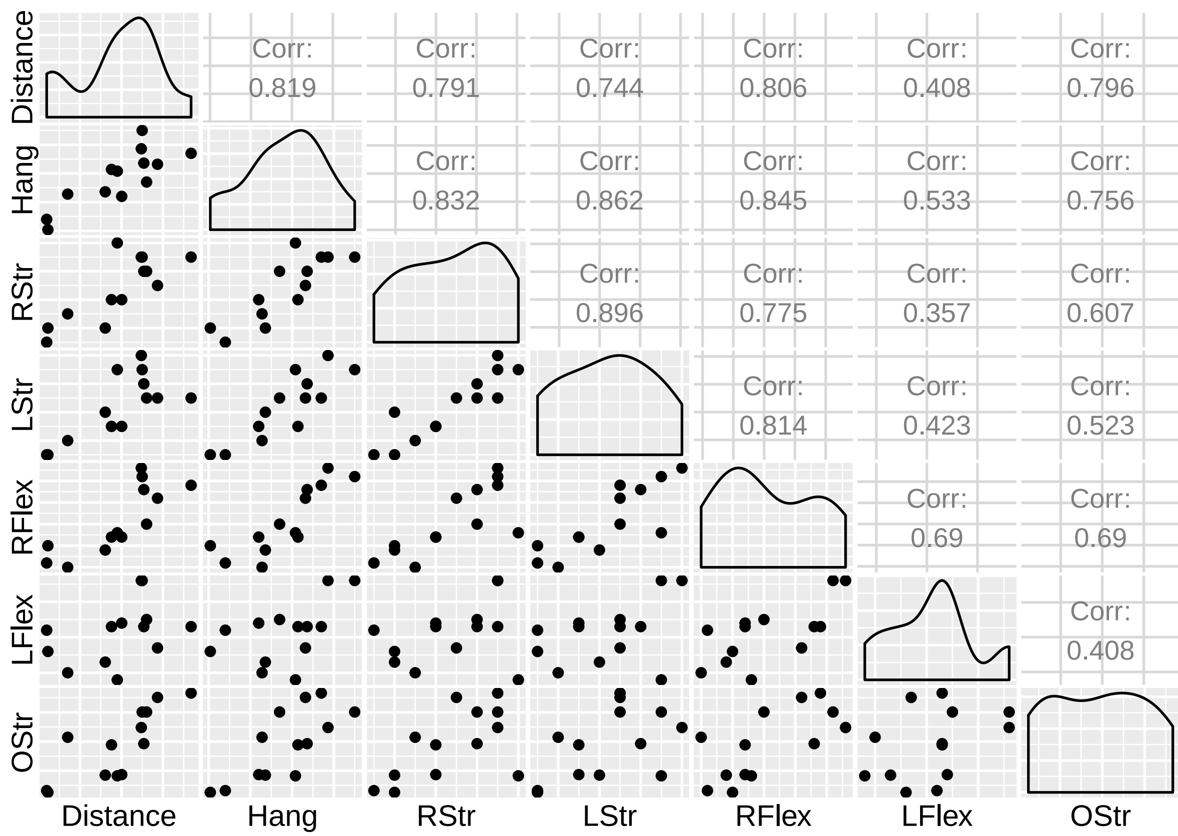
First we load and inspect the data.

```
data(punting, package = "faraway")
head(punting)
```

```
##   Distance Hang RStr LStr RFlex LFlex   OStr
## 1   162.50 4.75  170  170   106   106 240.57
## 2   144.00 4.07  140  130    92    93 195.49
```

```
## 3    147.50 4.04  180  170    93    78 152.99
## 4    163.50 4.18  160  160   103    93 197.09
## 5    192.00 4.35  170  150   104    93 266.56
## 6    171.75 4.16  150  150   101    87 260.56
```

```
ggpairs(data = punting, axisLabels = "none")
```



a) 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

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

summary(lm.fit)
```

```
##
## 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
```

```r
# Uncomment for diagnostic plots. plot(lm.fit)
```

We see that none of the predictors are significant at the $5% level for this model.

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

The test we want to perform is

$$H_0 : \beta_{Rstr} = \beta_{LStr} = \beta_{RFlex} = \beta_{LFlex} = 0$$

versus the alternative that one or more of the coefficients is not zero. The likelihood ratio test for the full model versus the null model $Y \sim \beta_0 + \epsilon$ works out to be an F-test.

```r
lm.fit.null <- lm(Distance ~ 1, data = punting)

anova(lm.fit.null, lm.fit)
```

```
## 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
```

Based on the p-value we have enough evidence to reject the null hypothesis at a significance of 5% in this case and claim that collectively the four predictors have a predictive relationship with the response.

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

The test we want to perform in this case is

$$H_0 : \beta_{Rstr} = \beta_{LStr}$$

versus the alternative that the effect is not the same.

```
lm.fit.subspace <- lm(Distance ~ I(RStr + LStr) + RFlex + LFlex, data = punting)

anova(lm.fit.subspace, lm.fit)
```

```
## 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
```

Based on this p-value we do not have enough evidence to reject the null hypothesis that the right and left leg strength have the same effect.

(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 leg strengths.

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

```
##
## 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
```

```
lm.fit.strength.sum <- lm(Distance ~ I(RStr + LStr), data = punting)
summary(lm.fit.strength.sum)
```

```
##
## 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(lm.fit.strength.sum, lm.fit.strength)
```

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

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

```
lm.fit.subspace <- lm(Distance ~ RStr + LStr + I(RFlex + LFlex), data = punting)

anova(lm.fit.subspace, lm.fit)
```

```
## 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
```

Based on this p-value we do not have enough evidence to reject the null hypothesis that the right and left leg flexibility have the same effect.

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

The test we want to perform is

$$H_0 : \beta_{Rstr} = \beta_{LStr} \ , \ \beta_{RFlex} = \beta_{LFlex}$$

```
lm.fit.subspace <- lm(Distance ~ I(RStr + LStr) + I(RFlex + LFlex), data = punting)

anova(lm.fit.subspace, lm.fit)
```

```
## 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
```

Based on this p-value we can not reject the null hypothesis of right-left symmetry.

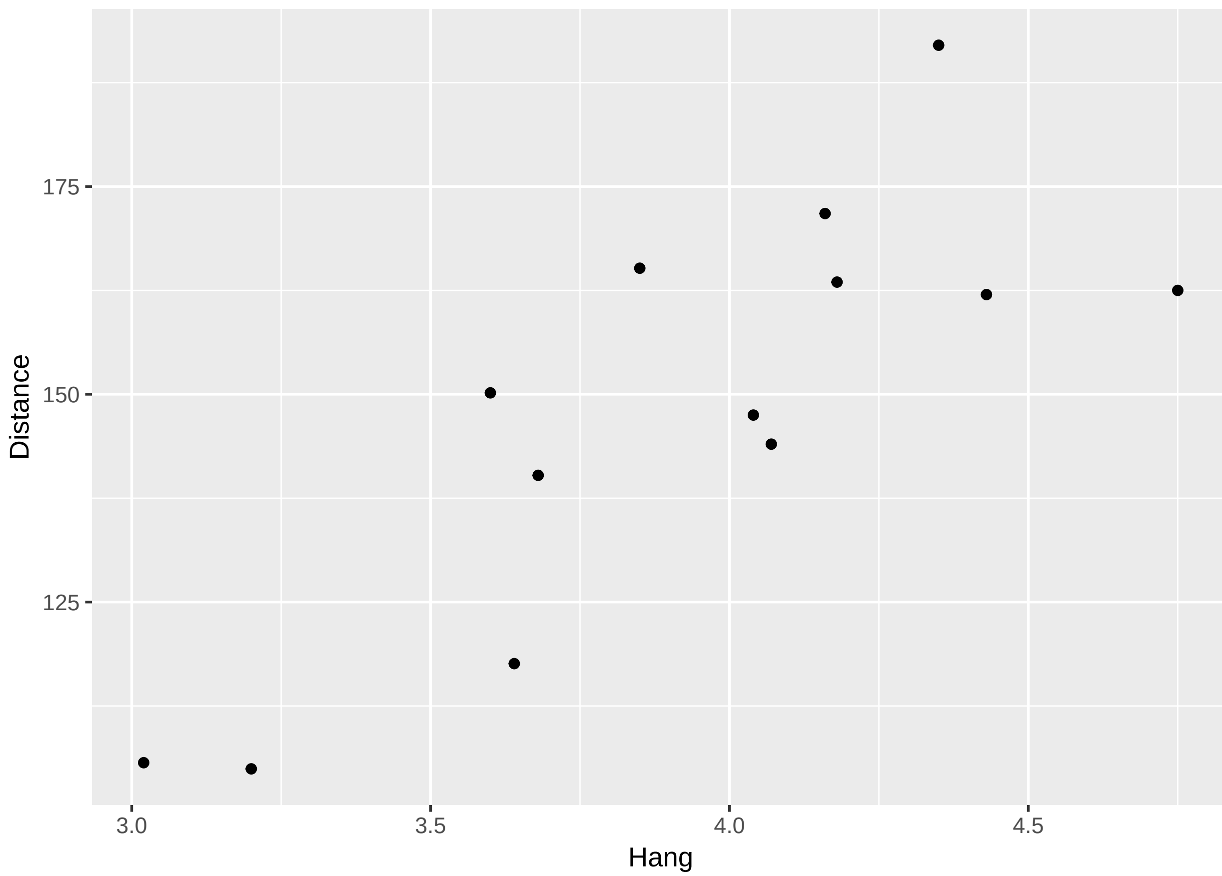
(h) Fit a model with Hang as the response and the same four predictors. Can we make a test to compare this model to that used in (a)? Explain.

```r
lm.fit <- lm(Hang ~ RStr + LStr + RFlex + LFlex, data = punting)

summary(lm.fit)
```

```
##
## 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
```

We see a higher R^2 for this model. Here is a plot of hang verus distance



It is not clear what the criteria is for comparison in this case. We know we can't use an F-test - the models are not nested. We could build a full model with all the variables and look at interactions, but that's not a test. We also don't have enough data to consider all the interactions in $Distance \sim Hang * RStr * LStr * RFlex * LFlex$