ST552 Homework 4

Nick Sun

February 12, 2019

Part 1

1) Overall regression F statistic

$H_0 : \beta_0 = \beta_1 = \beta_2 = \beta_3 = 0$

K =

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

and m =

$$\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

2)

$H_0 : \beta_1 = 0$

K =

$$\begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}$$

and m = 0

3)

$H_0 : \beta_1 = 0$ in model 2 $y_i = \beta_0 + \beta_1 Acetic_i + \epsilon_i$

K =

$$\begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

and m = 0

4)

$H_0 : \beta_0 = \beta_1 = 0$ in full model

K =

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

and m =

$$\begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

5)

$H_0 : \beta_1 = \beta_2 \rightarrow H_0 : \beta_1 - \beta_2 = 0$

K =

$$\begin{pmatrix} 0 & 1 & -1 & 0 \end{pmatrix}$$

and m = 0

Question 2

Now we consider the following regression model from `teengamb`:

$$gamble_i = \beta_0 + \beta_1 male_i + \beta_2 status_i + \beta_3 income_i + \beta_4 verbal_i + \beta_5 statusxmale + \beta_6 incomexmale + \beta_7 verbalxmale + \epsilon_i$$

for $i = 1, \ldots, 47$

Part 1

Since income is a continuous variable, the effect of income can be interpreted as the change in the mean response given a unit change in income.

In the case of income, the effect can be interpreted as β_3 for all females and $\beta_3 + \beta_6$ for all males.

Part 2

We can answer if the effect of income depends on sex by conducting a t-test on the interaction term between $male_i$ and $income_i$.

If we reject $H_0 : male_i x income_i$, then we have evidence to say that there is some difference between the effect income has between males and females.

Part 3

If I was interested in seeing if there was a positive or negative relationship between acmount gambled and income for males, I examine the parameter estimate for $income_i x male_i$, β_6.

If β_6 is postive and significant, that tells us that males on average gamble more if their income increases.

If β_6 is negative and significant, that tells us that males on average gamble less if their income increases.

Part 4

Since sex is coded as *male* and is a categorical variable, the effect of *sex* can be interpreted as the difference in average amount gambled between males and females. The parameter estimate corresponding to this difference in mean amount is β_1.

There are also interaction terms that are at play in this model, $\beta_5, \beta_6, \beta_7$, which involve *sex*. These effects are prevalent when there is a unit change in *status*, *income*, or *verbal* score respectively, holding all other variables constant.

The total effect sex on amount gambled for this model can be written as:

$$effect_{\text{sex}} = \beta_1 + \beta_5(status) + \beta_6(income) + \beta_7(verbal)$$

Part 5

I would conduct an F-test between a full model containing the *sex* categorical variable coded as *male* plus all the interaction terms against a reduced model with none of these terms present:

$$\text{Reduced Model: } \hat{gamble}_i = \beta_0 + \beta_2 status_i + \beta_3 income_i + \beta_4 verbal_i$$

The $H_0 : \beta_1 = \beta_5 = \beta_6 = \beta_7 = 0$ can be tested using this F-test framework.

Part 2

Faraway 3.7

Part a

Here is a summary of the model:

```
##
## 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 these predictors have a relationship to the response according to the individual t-tests! However, we notice that the p-value for the F-test is significant, telling us that the model at least performs better than the null model.

Part b.

From the above, the F-statistic is 5.59 and the associated p-value against the null distribution is .01902, thereby rejecting the null hypothesis.

This tells us that at least one of these predictors is significant to the model. Perhaps there are some interaction terms that are going on here.

Part c.

We can test that the effects of right and left leg strength are equal using the `I()` function. When the model is created, only one parameter estimate will be made for *RStr* and *LStr*.

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

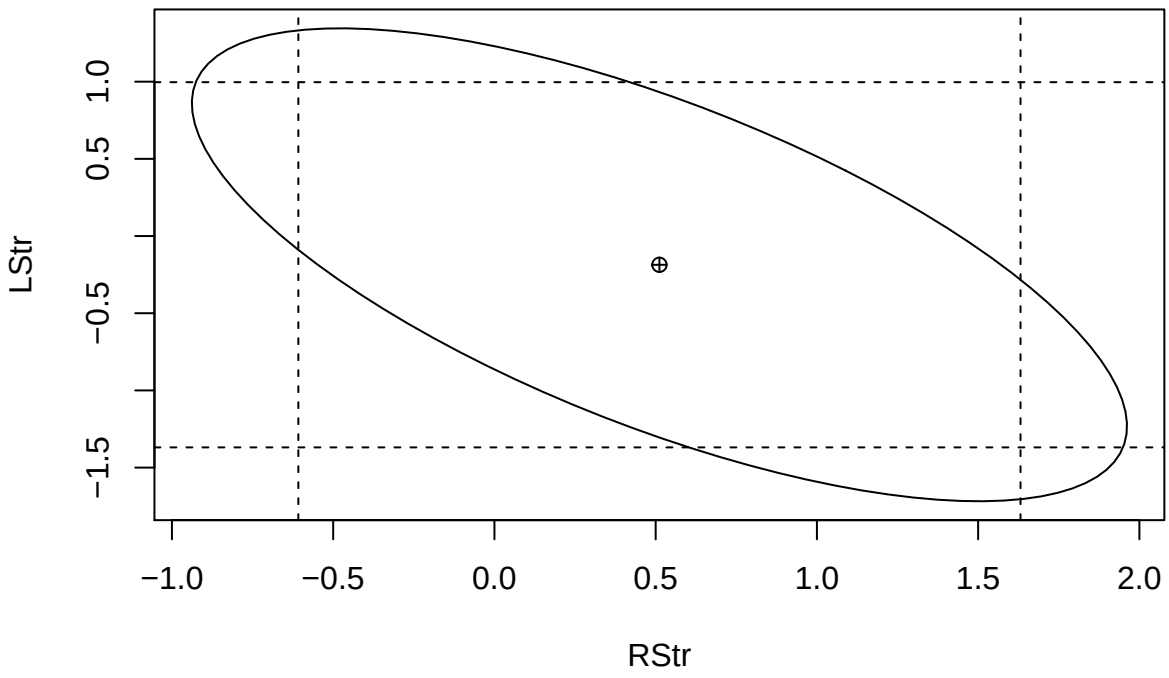
The F-test between the full model where *RStr* and *LStr* have their own parameter estimates and the reduced model where they share a parameter estimate produces a p-value of .468, telling us that we cannot reject H_0. According to this F-test,we should prefer to reduced model. This means that using the reduced model with an effect for *RStr* and *LStr* is justifiable.

Part d.

Constructing a 95% Confidence region for $(\beta_{RStr}, \beta_{LStr})$ is related to part (c) since if the confidence region contains the line where $\beta_{RStr} = \beta_{LStr}$, then we know that we cannot reject the null hypothesis that the effects are the same.

```
##
## Attaching package: 'ellipse'

## The following object is masked from 'package:graphics':
##
##     pairs
```



This is relevant to part c. because if a point falls on the line for $\beta_{RStr} = \beta_{LStr}$ **and** inside this confidence region, we will fail to reject the null hypothesis that it is equal to 0. In our case, the parameter estimate for $\beta_1 = .1741$, which is inside the region. This verifies our result from the F-test in part c.

```
summary(punt_lm2)
```

```
##
## 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
```

Part e.

We can do this using the `I()` function. This model is similar to part c. but it excludes the variables for flexibility.

```
punt_lm3 <- lm(Distance ~ I(RStr + LStr), data = punting)
punt_lm3b <- lm(Distance ~ RStr + LStr, data = punting)
anova(punt_lm3, punt_lm3b)
```

```
## 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
```

The F-test output fails to reject the null hypothesis that the reduced model adequately predicts the distance of punts. Therefore, we are justified in using the reduced model and considering a single parameter esitmate for the sum of **RStr + LStr**.

Part f.

```
anova(punt_lm, punt_lm3)
```

```
## Analysis of Variance Table
##
## Model 1: Distance ~ RStr + LStr + RFlex + LFlex
## Model 2: Distance ~ I(RStr + LStr)
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1      8 2132.6
## 2     11 3061.3 -3   -928.71 1.1613 0.3827
```

Part g.

Performing the tests from (c) and (f) simultaneously involves checking if the effects of right and left leg strengths are the same and if the effects of right and left leg flexibility are the same.

```
##
## 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
```

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

We cannot reject the null hypothesis for the F statistics so we have no significant evidence to say that the full model with separate effects for **RStr, LStr, RFlex, LFlex** is better than the reduced model. We're better off using the model that has a combined estimate for the leg strengths and the leg flexibilities.

Part h.

```
##
## 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 cannot perform an F-test to compare this model against the model from part a. since these models are not nested in one another - they model completely different responses.

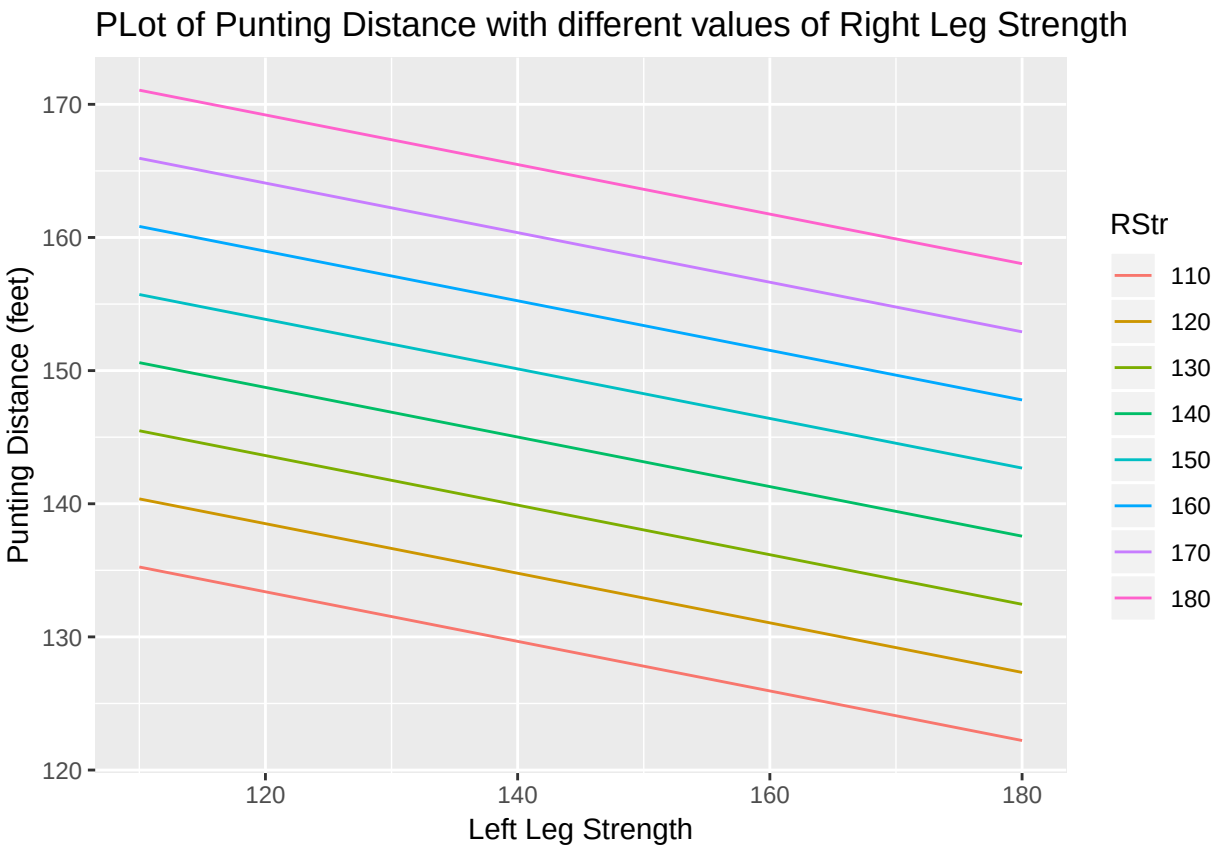
Part i.

Our goal is to create a plot showing the predicted value when you use different values of *RStr*.

```
## [1] 104
```

```
##   Intercept RStr LStr    RFlex    LFlex predicted_values
## 1         1  110  170 95.69231 91.23077         124.0784
## 2         1  120  170 95.69231 91.23077         129.1948
## 3         1  130  170 95.69231 91.23077         134.3112
```

```
## 4         1   140   170 95.69231 91.23077         139.4276
## 5         1   150   170 95.69231 91.23077         144.5439
## 6         1   160   170 95.69231 91.23077         149.6603
```

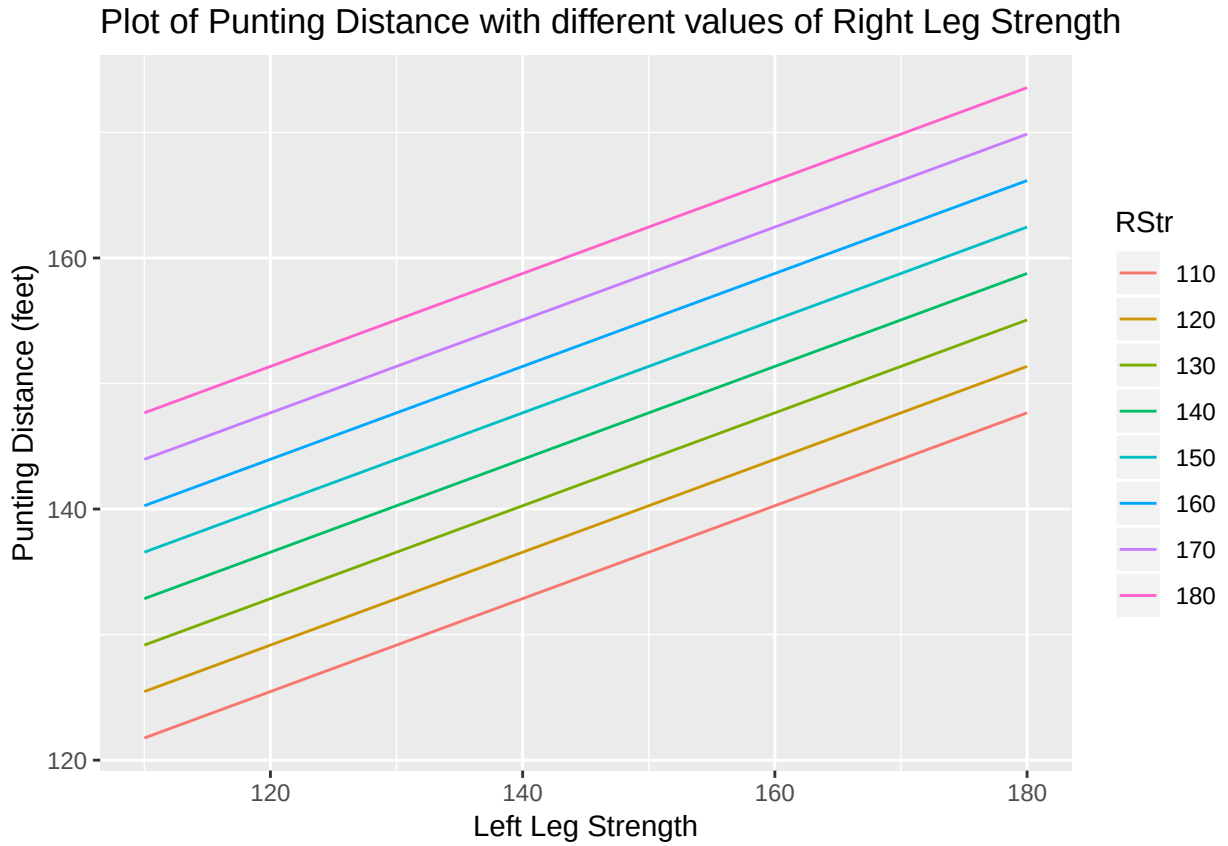


Neat plot!

Part j.

I am a fan of the model with two parameter estimates, one for the combined leg strengths and another for the combined leg flexibilities. I think it is parsimonious while still having relatively good performance.

```
ggplot(aes(x = LStr, y = predicted_values2, color = RStr), data = plot_pred2) + geom_line() + ggtitle("
  xlab("Left Leg Strength") +
  ylab("Punting Distance (feet)")
```



Weirdly enough, this plot has punting distance moving in the opposite direction of the previous plot as we change left leg strength.

Faraway 4.5

Part a.

```
##
## Call:
## lm(formula = brozek ~ age + weight + height + neck + chest +
##     abdom + hip + thigh + knee + ankle + biceps + forearm + wrist,
##     data = fat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.264  -2.572  -0.097   2.898   9.327
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -15.29255   16.06992  -0.952  0.34225
## age           0.05679    0.02996   1.895  0.05929 .
## weight       -0.08031    0.04958  -1.620  0.10660
## height       -0.06460    0.08893  -0.726  0.46830
## neck         -0.43754    0.21533  -2.032  0.04327 *
## chest        -0.02360    0.09184  -0.257  0.79740
```

```
## abdom         0.88543    0.08008  11.057  < 2e-16 ***
## hip          -0.19842    0.13516  -1.468  0.14341
## thigh         0.23190    0.13372   1.734  0.08418 .
## knee         -0.01168    0.22414  -0.052  0.95850
## ankle         0.16354    0.20514   0.797  0.42614
## biceps        0.15280    0.15851   0.964  0.33605
## forearm       0.43049    0.18445   2.334  0.02044 *
## wrist        -1.47654    0.49552  -2.980  0.00318 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.988 on 238 degrees of freedom
## Multiple R-squared:  0.749,  Adjusted R-squared:  0.7353
## F-statistic: 54.63 on 13 and 238 DF,  p-value: < 2.2e-16
```

```
## Analysis of Variance Table
##
## Model 1: brozek ~ age + weight + height + neck + chest + abdom + hip +
##     thigh + knee + ankle + biceps + forearm + wrist
## Model 2: brozek ~ age + weight + height + abdom
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1    238 3785.1
## 2    247 4205.0 -9    -419.9 2.9336 0.002558 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

It is not justifiable to use the smaller model since our p-value from the F-test is below the $\alpha = .05$ level. Therefore we reject the H_0 that the reduced model is appropriate.

Part b.

Compute a 95% prediction interval for median predictor values and compare to the results to the interval for the full model.

```
x0 <- model.matrix(reduced_fat)
x1 <- model.matrix(full_fat)

y0 <- apply(x0, 2, median)
y1 <- apply(x1, 2, median)

predict(reduced_fat, new=data.frame(t(y0)), interval="prediction")
```

```
##        fit      lwr      upr
## 1 17.84028 9.696631 25.98392
```

```
predict(full_fat, new=data.frame(t(y1)), interval="prediction")
```

```
##        fit     lwr      upr
## 1 17.49322 9.61783 25.36861
```

The intervals are pretty close - they don't differ by any appreciable amount.

Part c.

For the reduced model, examine all the observations from case numbers 25 to 50. Which two observations seem particularly anomalous?

```
colnames(x0)
```

```
## [1] "(Intercept)" "age"         "weight"      "height"      "abdom"
```

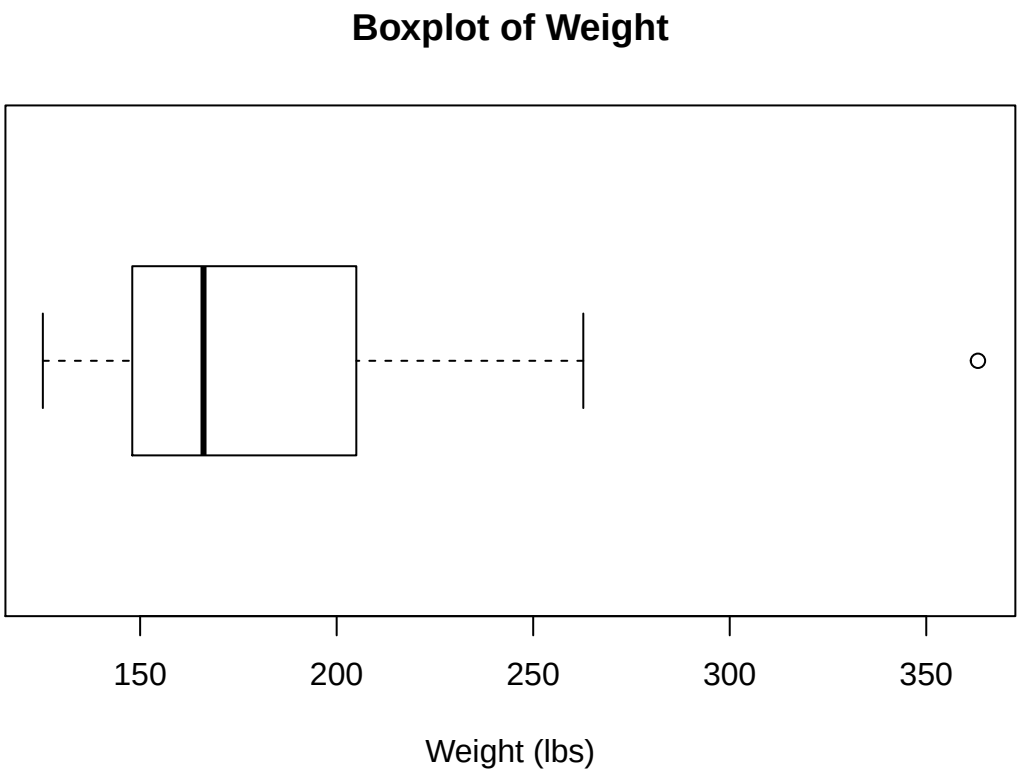
```
x0 <- as.data.frame(x0[25:50,])
dim(x0)
```

```
## [1] 26  5
```

```
summary(x0)
```

```
##   (Intercept)      age            weight          height
##  Min.   :1    Min.   :27.00   Min.   :125.2   Min.   :29.50
##  1st Qu.:1    1st Qu.:31.25   1st Qu.:148.1   1st Qu.:67.50
##  Median :1    Median :40.50   Median :166.1   Median :68.88
##  Mean   :1    Mean   :38.92   Mean   :182.6   Mean   :67.93
##  3rd Qu.:1    3rd Qu.:45.00   3rd Qu.:204.5   3rd Qu.:71.25
##  Max.   :1    Max.   :50.00   Max.   :363.1   Max.   :73.75
##      abdom
##  Min.   : 70.40
##  1st Qu.: 79.20
##  Median : 86.60
##  Mean   : 93.25
##  3rd Qu.:104.30
##  Max.   :148.10
```

```
boxplot(x0$weight,
        main = "Boxplot of Weight",
        xlab = "Weight (lbs)",
        horizontal= TRUE)
```



From the summary and the boxplot of weight, we see that there are two anomalous points: one outlier for weight on the far right end of the scale and another outlier for height on the far left.

```
x0[x0$height == min(x0$height),]
```

```
##    (Intercept) age weight height abdom
## 42           1  44    205   29.5 104.3
```

```
x0[x0$weight == max(x0$weight),]
```

```
##    (Intercept) age weight height abdom
## 39           1  46 363.15  72.25 148.1
```

Part d.

Recompute the 95% prediction interval for median predictor values after these two anomalous cases that have been excluded from the data. Did this make a difference?

```
new_data <- fat[c(-39, -42),]

new_full_fat <- lm(brozek ~ age + weight + height + neck + chest + abdom + hip + thigh + knee + ankle +

new_reduced_fat <- lm(brozek ~ age + weight + height + abdom, data = new_data)
```

```r
x1 <- model.matrix(new_full_fat)
x0 <- model.matrix(new_reduced_fat)

y0 <- apply(x0, 2, median)
y1 <- apply(x1, 2, median)

predict(new_reduced_fat, new=data.frame(t(y0)), interval="prediction")
```

```
##       fit      lwr      upr
## 1 17.9033 9.887851 25.91874
```

```r
predict(new_full_fat, new=data.frame(t(y1)), interval="prediction")
```

```
##        fit      lwr      upr
## 1 17.54174 9.754495 25.32898
```

Even after removing the anomalous cases, the prediction intervals don't actually change that much!