FIYEP Week 05 Exercises

1. IQ scores and Income. Continuing studying of IQ scores. In the data set is a listing from the National Longitudinal Study of Youth (NLSY79) with annual income in 2005 (in U.S. dollars, as recorded in a 2006 interview) and scores on the Word Knowledge, Paragraph Comprehension, Arithmetic Reasoning, and Mathematics Knowledge portions of the Armed Forces Vocational Test, which is based on a linear combination of the four components and which is sometimes used as a general intelligence test score. In this exercise we will consider the dependence of 2005 income with AFQT score. First study if or how well the AFGT score predicts the 2005 income. However, a much more interesting question is whether there might be some better linear combination of the four components than AFQT for predicting income. Investigate this by seeing whether the component scores are useful predictors of ”Income2006” in addition to AFQT. Also see whether AFQT is a predictor in addition to the four test scores. Which test scores seem to be the most important predictors of 2005 income?
   Solution:

```r
iq <- read.csv('data/IQ scores and income.csv')
# make model
afqt_model <- lm(Income2005 ~ AFQT, data=iq)
summary(afqt_model)
```

```
##
## Call:
## lm(formula = Income2005 ~ AFQT, data = iq)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -72004 -24075  -7815  12447 643506
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 21181.66    1925.59   11.00   <2e-16 ***
## AFQT          518.68      31.51   16.46   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 44460 on 2582 degrees of freedom
## Multiple R-squared:  0.09496,    Adjusted R-squared:  0.09461
## F-statistic: 270.9 on 1 and 2582 DF,  p-value: < 2.2e-16
```

AFQT very clearly predicts income. The p-value is extremely small (<2e-16), so we’re basically guaranteed that there’s a correlation. We can see that a persons yearly income is around $519 higher for every point on AFQT. It’s important to note that I have not done any tests to ensure the assumption about data (such as homoscedasticity or linearity), but I am assuming that the assumptions are met.

Now I set up a model that uses the individual test scores (no interaction).

```r
test_model <- lm(Income2005 ~ Arith + Word + Parag + Math, data=iq)
summary(test_model)
```

```
##
## Call:
## lm(formula = Income2005 ~ Arith + Word + Parag + Math, data = iq)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -77163 -23451  -7462  12036 650259
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  12494.9     3526.2   3.543 0.000402 ***
## Arith         1463.6      227.0   6.446 1.36e-10 ***
## Word            14.3      211.4   0.068 0.946076
## Parag         -663.2      453.0  -1.464 0.143306
## Math          1188.6      249.3   4.768 1.96e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 43990 on 2579 degrees of freedom
## Multiple R-squared:  0.115,  Adjusted R-squared:  0.1136
## F-statistic: 83.76 on 4 and 2579 DF,  p-value: < 2.2e-16
```

So Arithmetic Reasoning and Mathematics Knowledge seem to relate to income, with very low p-values (under 0.05), but Paragraph Comprehension and especially Word Knowledge have pretty high p-values, which means random data might as well have been used to generate the same results for those predictors. Let's remove those then.

```r
lim_test_model <- lm(Income2005 ~ Arith + Math, data=iq)
summary(lim_test_model)
```

```
##
## Call:
## lm(formula = Income2005 ~ Arith + Math, data = iq)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -76536 -23588  -7461  11906 648899
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   8948.6     2410.3   3.713 0.000209 ***
## Arith         1346.6      211.1   6.378 2.12e-10 ***
## Math          1094.4      240.5   4.551 5.59e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 44000 on 2581 degrees of freedom
## Multiple R-squared:  0.114,  Adjusted R-squared:  0.1133
## F-statistic:   166 on 2 and 2581 DF,  p-value: < 2.2e-16
```

These p-values are much better. As we can see, Arithmetic Reasoning has a slightly higher estimate. Now lets try to incorporate AFQT too, first with the two bad predictors.

```
high_model <- lm(Income2005 ~ Arith + Word + Parag + Math + AFQT, data=iq)
summary(high_model)
```

```
##
## Call:
## lm(formula = Income2005 ~ Arith + Word + Parag + Math + AFQT,
##     data = iq)
##
## Residuals:
##    Min      1Q Median      3Q     Max
## -77378 -23507  -7488  12137 649283
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 17347.70    5512.12   3.147  0.00167 **
## Arith        1316.83     260.68   5.052 4.69e-07 ***
## Word          -94.96     231.96  -0.409  0.68228
## Parag        -993.26     536.86  -1.850  0.06441 .
## Math          954.68     322.21   2.963  0.00308 **
## AFQT          142.99     124.84   1.145  0.25216
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 43990 on 2578 degrees of freedom
## Multiple R-squared:  0.1154, Adjusted R-squared:  0.1137
## F-statistic: 67.28 on 5 and 2578 DF,  p-value: < 2.2e-16
```

We’ve learned that AFQT also has a p-value too high for us to know for sure that it is correlated, most definitely due to the fact that the Word and Parag predictors themselves have high p-values. In fact, including AFQT doesn’t make much sense, as it is multicollinear with the other predictors, since AFQT is a measure of the other scores. That means the limited score model was the best one.