Question 4: Interactions

Lab 7 Handout Solutions

Statistics 139

```r
#load the data
prevend = read.csv("data/prevend.csv")

prevend$Education.Factor = factor(prevend$Education, levels = 0:3,
                                  labels = c("Primary", "LowSec", "HighSec", "Univ"))
```

Question 4: Interactions

This problem investigates the relationship between RFFT score (`RFFT`), age (`Age`), and diabetes (`DM`).

a) Fit a linear model that regresses RFFT score on age and diabetes status.

```r
#fit model
model.statin.dm = lm(RFFT ~ Age + DM, data = prevend)
summary(model.statin.dm)
```

```
##
## Call:
## lm(formula = RFFT ~ Age + DM, data = prevend)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -68.436 -15.634  -0.827  14.733  78.647
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 130.90436    1.68904   77.50  < 2e-16 ***
## Age          -1.13019    0.03055  -36.99  < 2e-16 ***
## DM           -8.26679    1.46563   -5.64 1.81e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.17 on 4092 degrees of freedom
## Multiple R-squared:  0.2752, Adjusted R-squared:  0.2748
## F-statistic: 776.7 on 2 and 4092 DF,  p-value: < 2.2e-16
```

i. According to the model, how does the average RFFT score for a 60-year-old compare to that of a 50-year-old, if both have diabetes?

The change in mean RFFT score can be determined directly from the coefficient for age, if diabetes status is held constant. An increase in one year of age is associated with a 1.13 point decrease in mean RFFT score; thus, an increase in ten years of age is associated with a 11.3 point decrease in mean RFFT score.

ii. According to the model, how does the average RFFT score for a 60-year-old compare to that of a 50-year-old, if both do not have diabetes?

This calculation does not differ from the one in part i. According to the model, the relationship between RFFT score and age is consistent whether diabetes status is held constant at 'diabetic' or at 'non-diabetic'.

b) Fit a linear model for RFFT score from age, diabetes status, and the interaction term between age and diabetes status.

```r
#fit interaction model
model.statin.dm.int = lm(RFFT ~ Age*as.factor(DM), data = prevend)
summary(model.statin.dm.int)
```

```
##
## Call:
## lm(formula = RFFT ~ Age * as.factor(DM), data = prevend)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -68.776 -15.571  -1.033  14.627  78.759
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        132.42948    1.72258  76.879  < 2e-16 ***
## Age                 -1.15842    0.03119 -37.143  < 2e-16 ***
## as.factor(DM)1     -48.51672    9.49994  -5.107 3.42e-07 ***
## Age:as.factor(DM)1   0.63364    0.14777   4.288 1.84e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.13 on 4091 degrees of freedom
## Multiple R-squared:  0.2784, Adjusted R-squared:  0.2779
## F-statistic: 526.1 on 3 and 4091 DF,  p-value: < 2.2e-16
```

i. Write the overall estimated model equation.

I indicates the indicator function below:

$$\widehat{RFFT} = 132.43 - 1.158(Age) - 48.52(I_{DM=1}) + 0.634(Age \times I_{DM=1})$$

ii. Simplify the model equation for diabetics. Simplify the model equation for non-diabetics.

$$\begin{aligned}\widehat{RFFT} =& 132.43 - 1.158(Age) - 48.52(I_{DM=1}) + 0.634(Age \times I_{DM=1}) \\ =& 132.43 - 1.158(Age) - 48.52(1) + 0.634(Age \times 1) \\ =& (132.43 - 48.52) + (-1.158 + 0.634)(Age) \\ =& 83.91 - 0.524(Age)\end{aligned}$$

$$\begin{aligned}\widehat{RFFT} =& 132.43 - 1.158(Age) - 48.52(I_{DM=1}) + 0.634(Age \times I_{DM=1}) \\ =& 132.43 - 1.158(Age) - 48.52(0) + 0.634(Age \times 0) \\ =& 132.43 - 1.158(Age)\end{aligned}$$

iii. How does fitting an interaction term change the model? Specifically, how do the interpretations from parts a) i. and ii. change when the model has an interaction term?

Fitting an interaction term allows for the association between RFFT score and age to be different between diabetics and non-diabetics. In this model, it is possible to make predictions based on the observed trend that the association between RFFT score and age is less negative for diabetics than for non-diabetics.

c) Fit a model to predict RFFT score from age, educational attainment, and the interaction between the two. Formally test whether the interaction term(s) provide a statistically significant improvement in prediction accuracy as measured by R^2 (you will need to fit a second model). Create a plot for the interaction model and summarize the model results.

There is a negative association between RFFT score and age for each level of educational attainment. In the plot below, primary school is represented as blue, lower secondary school as red, higher secondary school as green, and university as orange. From the ESS F-test, there is evidence that the interaction terms contribute to the model.

The slope for primary education is not significantly different from 0. Interestingly, the negative association between cognitive score and age is stronger among the two groups with the highest level of educational attainment (higher secondary school and university).

```r
#fit the model
edu.interact = lm(RFFT ~ Education.Factor*Age, data = prevend)
summary(edu.interact)
```

```
##
## Call:
## lm(formula = RFFT ~ Education.Factor * Age, data = prevend)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -65.910 -14.249  -1.393  13.641  89.329
##
## Coefficients:
```

```
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   95.04751    6.15117  15.452  < 2e-16 ***
## Education.FactorLowSec         9.97123    6.93510   1.438 0.150570
## Education.FactorHighSec       26.46364    6.83174   3.874 0.000109 ***
## Education.FactorUniv          40.42818    6.74464   5.994 2.22e-09 ***
## Age                           -0.75856    0.09658  -7.854 5.10e-15 ***
## Education.FactorLowSec:Age    -0.05756    0.11062  -0.520 0.602852
## Education.FactorHighSec:Age -0.20813    0.11146  -1.867 0.061932 .
## Education.FactorUniv:Age      -0.27674    0.11024  -2.510 0.012096 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.68 on 4087 degrees of freedom
## Multiple R-squared:  0.3701, Adjusted R-squared:  0.369
## F-statistic:   343 on 7 and 4087 DF,  p-value: < 2.2e-16
```

```r
#ESS F-Test
edu.age = lm(RFFT ~ Education.Factor + Age, data = prevend)
summary(edu.age)
```

```
##
## Call:
## lm(formula = RFFT ~ Education.Factor + Age, data = prevend)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -65.459 -14.101  -1.178  13.407  86.280
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             105.34425    2.13934  49.241  < 2e-16 ***
## Education.FactorLowSec    5.88488    1.20236   4.894 1.02e-06 ***
## Education.FactorHighSec  13.86215    1.24977  11.092  < 2e-16 ***
## Education.FactorUniv     24.38332    1.22821  19.853  < 2e-16 ***
## Age                      -0.92257    0.02981 -30.950  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.71 on 4090 degrees of freedom
## Multiple R-squared:  0.3682, Adjusted R-squared:  0.3676
## F-statistic:   596 on 4 and 4090 DF,  p-value: < 2.2e-16
```

```r
anova(edu.age, edu.interact)
```

```
## Analysis of Variance Table
##
## Model 1: RFFT ~ Education.Factor + Age
## Model 2: RFFT ~ Education.Factor * Age
##   Res.Df     RSS Df Sum of Sq      F   Pr(>F)
```

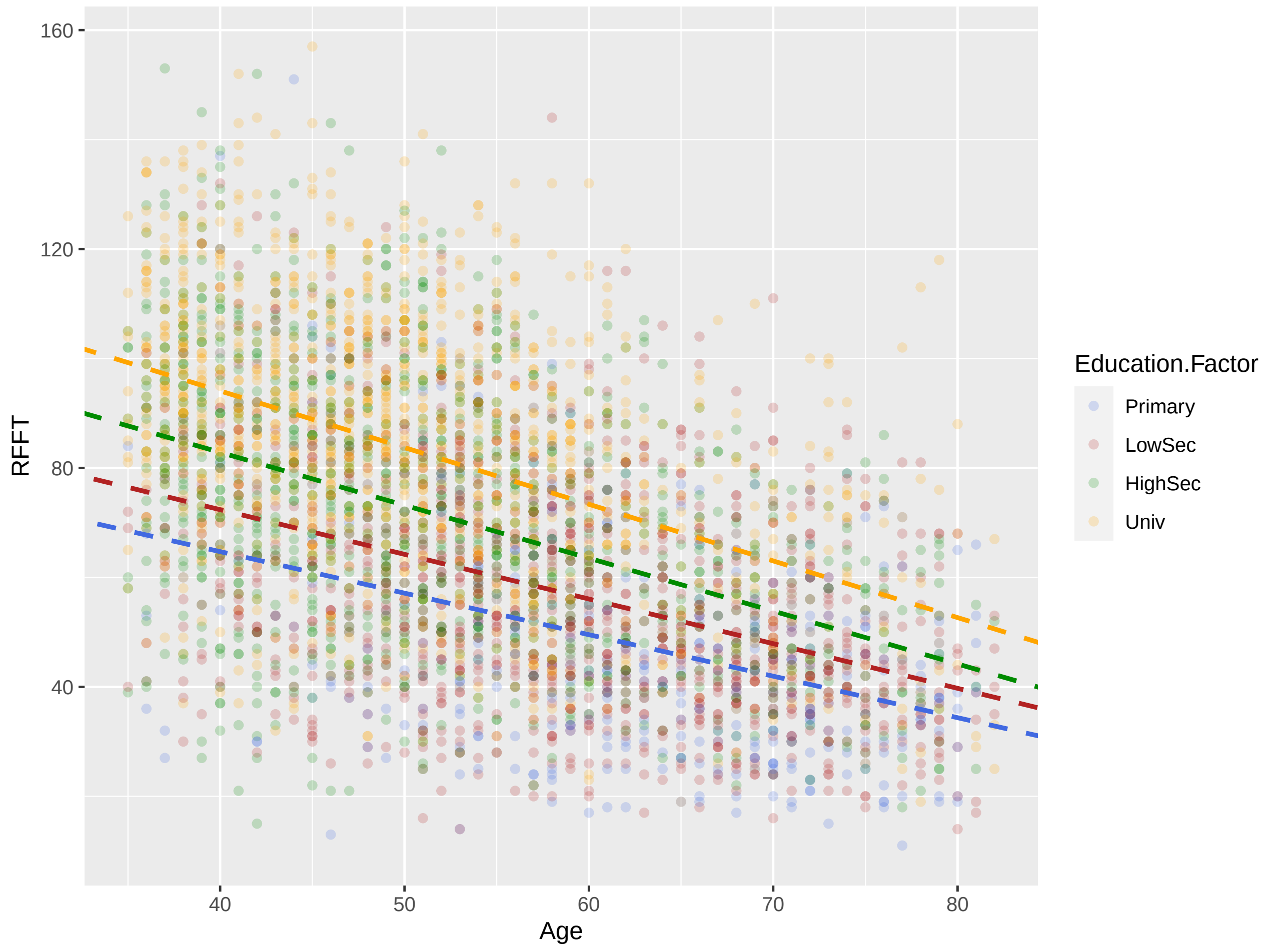
```
## 1   4090 1753410
## 2   4087 1748317  3    5092.9 3.9685 0.007771 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#create a plot
primary = (prevend$Education.Factor == "Primary")
lowsec = (prevend$Education.Factor == "LowSec")
highsec = (prevend$Education.Factor == "HighSec")
univ = (prevend$Education.Factor == "Univ")

library(tidyverse)
```

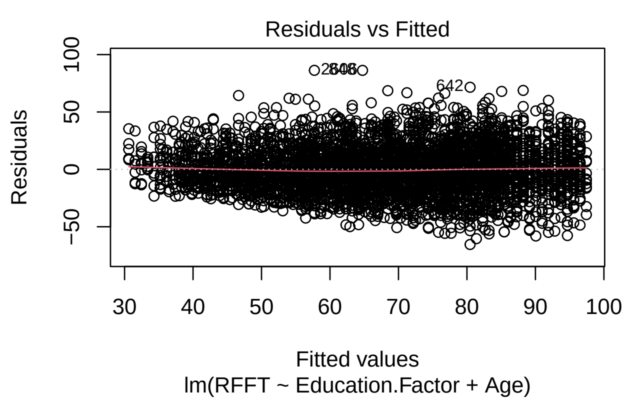
```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to beco
```

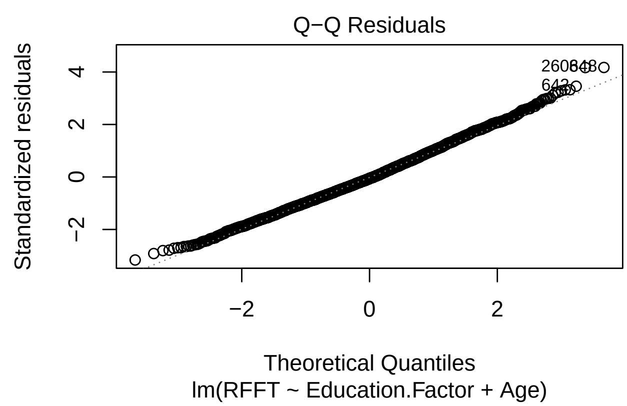
```
library(ggplot2)
prevend %>% ggplot(mapping = aes(x=Age, y=RFFT, color=Education.Factor)) +
  geom_point(alpha=0.2) +
  scale_color_manual(values = c("royalblue", "firebrick","green4", "orange"))  +
  geom_abline( intercept=edu.interact$coef[1], slope = edu.interact$coef[5], linetype = "dashed
  geom_abline( intercept=edu.interact$coef[1] + edu.interact$coef[2], slope = edu.interact$coef
  geom_abline( intercept=edu.interact$coef[1] + edu.interact$coef[3], slope = edu.interact$coef
  geom_abline( intercept=edu.interact$coef[1] + edu.interact$coef[4], slope = edu.interact$coef
```

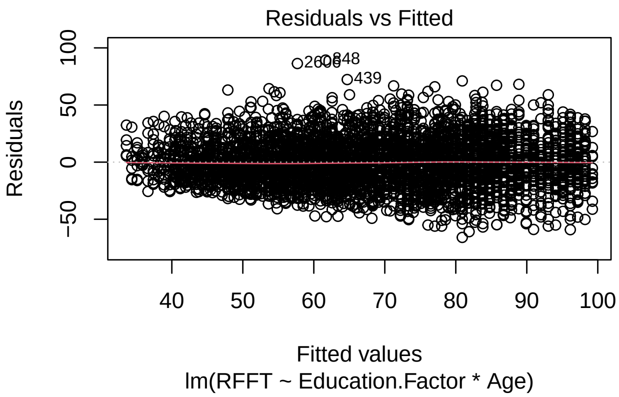


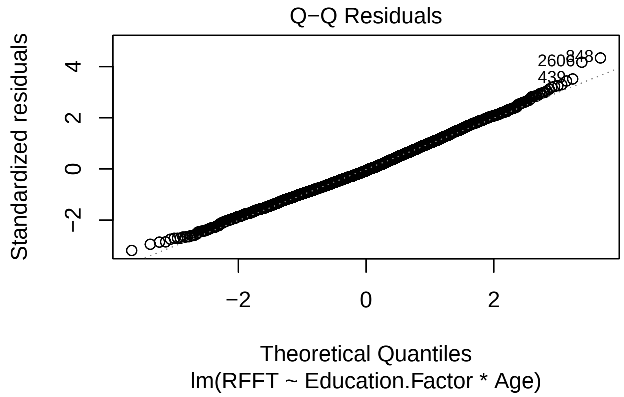
d) Visually assess the linearity assumption for the two models you used in the test in the previous part. How do they compare?

```r
plot(edu.age,which=c(1,2))
plot(edu.interact,which=c(1,2))
```









The QQ plots above show that the normality assumption is pretty similar in both models (no concerns). The residual vs. fitted plots suggest that there is likely a little bit of non-constant variance in both models. The non-linearity present in the non-interactive (aka, additive) model is potentially fixed in the model with interaction. In the additive model, the residual scatterplot suggests that at low values of $\hat{y}$ (around 30-40), the points are mostly above the zero horizontal line: the residuals are more likely to be positive in this range, thus the observations are being underestimated. This issue seems to go away in the interactive model.