Categorical Predictors with Several Levels and Inference in Regression

Chapter 7, Lab 3: Solutions

OpenIntro Biostatistics

Topics

- Categorical predictors with several levels
- Inference in multiple regression

This lab expands on the topics introduced in Chapter 6, Lab 4 (Categorical Predictors with Two Levels and Inference in Regression) by discussing categorical predictors with more than two levels and generalizing inference in regression to the setting where there are several slope parameters.

The material in this lab corresponds to Sections 7.4 - 7.6 and 7.9 in *OpenIntro Biostatistics*.

Introduction

Categorical predictors with several levels

Fitting a regression model with a categorical predictor that has several levels is analogous to comparing the means of several groups, where the groups are defined by the categorical variable. The equation of the regression line has intercept b_0, which equals the mean of one of the groups, and slopes $b_1, b_2, \ldots, b_{p+1}$, where $p+1$ equals the number of groups and each slope b_k for $k = 1, 2, \ldots, p+1$ equals the difference in means between the reference group and group k.

Inference in multiple regression

The observed data $(y_i, x_{i1}, x_{i2}, \ldots, x_{ip})$ for $i = 1, 2, \ldots, n$ cases are assumed to have been randomly sampled from a population where the response variable Y and p explanatory variables $X_1, X_2, \ldots, X_p$ follow a population model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon,$$

where $\epsilon \sim N(0, \sigma)$. Under this assumption, the intercept and slopes of the regression line, b_0 and $b_1, b_2, \ldots, b_p$, are estimates of the population parameters β_0 and $\beta_1, \beta_2, \ldots, \beta_p$.

In multiple regression, the coefficient β_j of a predictor X_j denotes the change in the response variable Y associated with a one unit change in X_j when the values of the other predictors are held constant.

Hypothesis tests and confidence intervals for regression population parameters have the same basic structure as tests and intervals about population means. Inference is usually done about the slope parameters, $\beta_1, \beta_2, \ldots, \beta_p$.

The F-statistic is used in an overall test of the model to assess whether the predictors in the model, considered as a group, are associated with the response.

Categorical predictors with several levels

The variable `Education` in the PREVEND data indicates the highest level of education that an individual completed in the Dutch educational system: primary school, lower secondary school, higher secondary education, or university education. This following questions step through exploring the association between RFFT score (`RFFT`) and educational level (`Education`) in `prevend.samp`, a random sample of $n = 500$ individuals from the PREVEND data.

1. Load `prevend.samp` and convert `Education` to a factor variable. The variable currently takes on values of either `0`, `1`, `2`, or `3`, where `0` denotes at most a primary school education, `1` a lower secondary school education, `2` a higher secondary education, and `3` a university education.

```
#load the data
library(oibiostat)
data("prevend.samp")

#convert Education to a factor
prevend.samp$Education = factor(prevend.samp$Education,
                                levels = c(0, 1, 2, 3),
                                labels = c("Primary", "LowerSecond",
                                           "HigherSecond", "Univ"))
```

2. Identify how many individuals are in each level of `Education`.

   51 individuals have completed at most primary school, 157 have completed at most lower secondary school, 134 have completed at most higher secondary school, and 158 have completed a university education.

```
table(prevend.samp$Education)
```

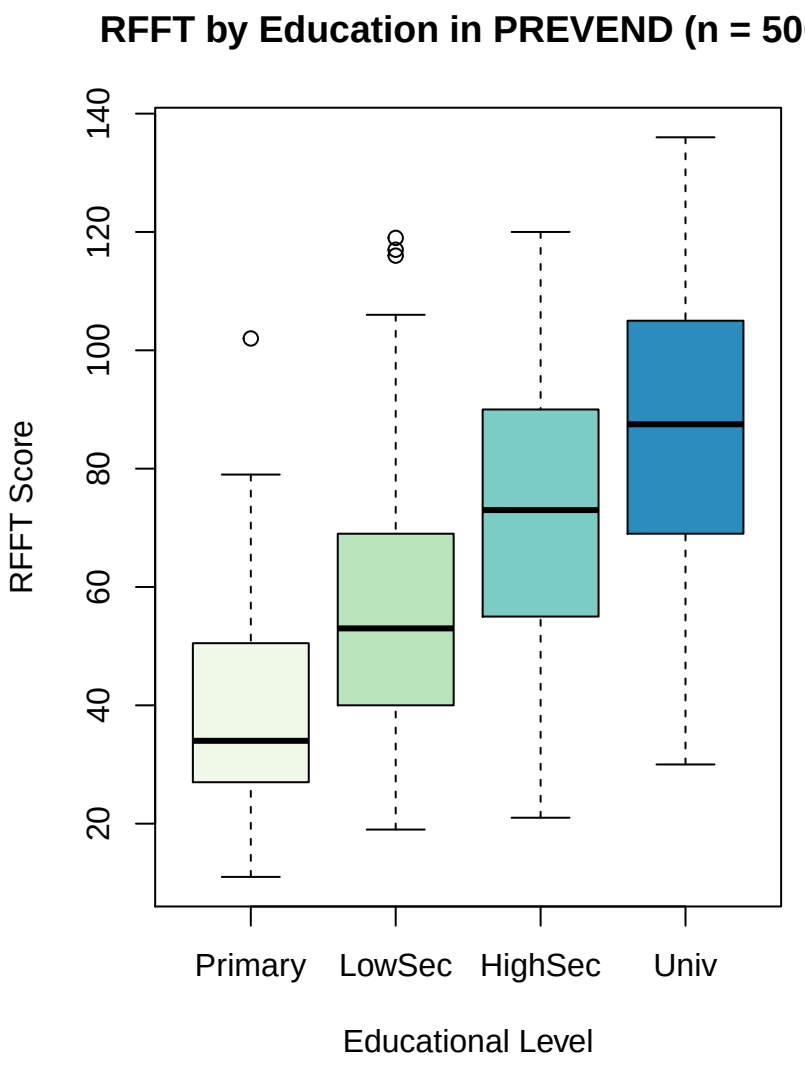
```
##
##      Primary  LowerSecond HigherSecond         Univ
##           51          157          134          158
```

3. Create a plot that shows the association between RFFT score and educational level. Describe what you see.

   There is a positive association between RFFT score and educational level. Individuals who have completed a higher level of education have higher median RFFT score; for example, median RFFT in the primary school group is about 35, while median RFFT score among those who have completed a university degree is about 85. The variability in RFFT score seems equal between the groups. In the primary and lower secondary school groups, there are a few upper outliers indicating individuals who had unusually high RFFT scores relative to others in the same educational group.

```
#load package for color palette
library(RColorBrewer)

#create plot
plot(RFFT ~ Education, data = prevend.samp,
 xlab = "Educational Level", ylab = "RFFT Score",
 main = "RFFT by Education in PREVEND (n = 500)",
```

```
names = c("Primary", "LowSec", "HighSec", "Univ"),
col = brewer.pal(4, "GnBu"))
```



4. Calculate mean RFFT score for each educational attainment group.

   Mean RFFT is 40.94 in the primary school group, 55.72 in the lower secondary school group, 73.07 in the higher secondary school group, and 85.91 in the university group.

```
#calculate means
tapply(prevend.samp$RFFT, prevend.samp$Education, mean)
```

```
##      Primary  LowerSecond HigherSecond         Univ
##     40.94118     55.71975     73.07463     85.90506
```

5. Fit a linear regression model relating RFFT score and educational attainment.

```
#fit a model
model.RFFTvsEdu = lm(RFFT ~ Education, data = prevend.samp)
coef(model.RFFTvsEdu)
```

```
##           (Intercept)  EducationLowerSecond EducationHigherSecond
##              40.94118              14.77857              32.13345
##         EducationUniv
##              44.96389
```

a) Write the equation of the least-squares line in terms of the variable names (e.g., *RFFT*).

$$\widehat{RFFT} = 40.94 + 14.78(EduLowSec) + 32.13(EduHighSec) + 44.96(EduUniv)$$

b) Based on part a), solve for the four possible values of $\widehat{RFFT}$ and interpret the values.

The values of the predictors can all be 0, or one predictor can be 1.

When all predictors are 0, $\widehat{RFFT}$ is 40.94. This is the mean RFFT score in the primary school group.

When $EduLowSec$ is 1, $\widehat{RFFT} = 40.94 + 14.78 = 55.72$. This is the mean RFFT score in the lower secondary school group.

When $EduHighSec$ is 1, $\widehat{RFFT} = 40.94 + 32.13 = 73.07$. This is the mean RFFT score in the higher secondary school group.

When $EduUniv$ is 1, $\widehat{RFFT} = 40.94 + 44.96 = 85.90$. This is the mean RFFT score in the university group.

c) Confirm that the numbers obtained in part b) match those from Question 4.

The numbers match, with a small rounding discrepancy (85.90 versus 85.91) for the mean RFFT score in the university group.

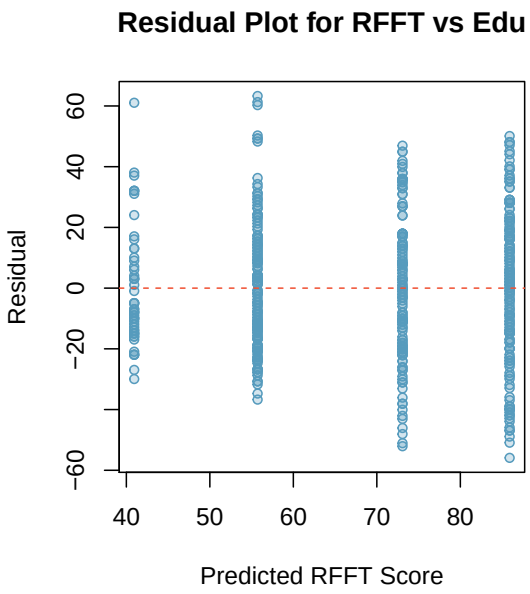
d) Using a residual plot and a Q-Q plot, check the assumptions for linear regression. It is reasonable to assume that these observations are independent. Why is it not necessary to check the linearity assumption for the predictors in this model?

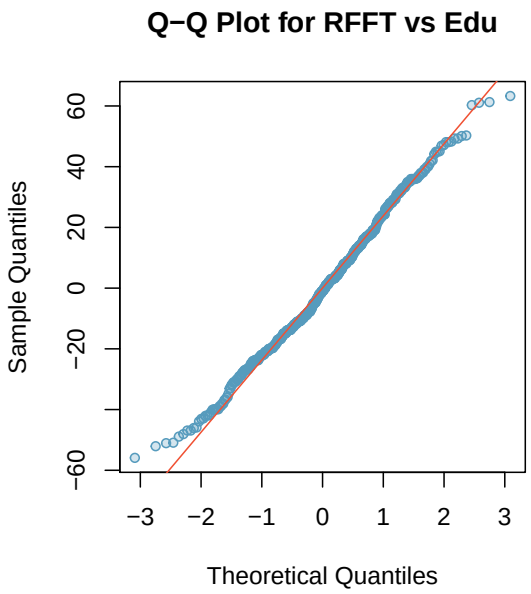
The plots show that the model fits the data reasonably well. The residuals are roughly constant across predicted RFFT score; although variability seems somewhat lower for individuals in the primary school group (on the left side of the plot), this may be an artifact from having relatively few individuals in this group. The residuals are approximately normal, with only small deviations from normality in the tails.

Each predictor in the model can be thought of as a binary variable, since each represents a group being compared to the reference group. Linearity is automatically satisfied for binary variables.

```
#load color package
library(openintro)
data(COL)
```

```r
par(mfrow = c(1, 2))

#residual plot
plot(resid(model.RFFTvsEdu) ~ fitted(model.RFFTvsEdu),
 xlab = "Predicted RFFT Score",
 ylab = "Residual",
 main = "Residual Plot for RFFT vs Edu",
 pch = 21, col = COL[1], bg = COL[1, 4],
 cex = 0.8)
abline(h = 0, col = COL[4], lty = 2)

#Q-Q plot
qqnorm(resid(model.RFFTvsEdu),
   main = "Q-Q Plot for RFFT vs Edu",
   pch = 21, col = COL[1], bg = COL[1, 4],
   cex = 0.8)
qqline(resid(model.RFFTvsEdu),
   col = COL[4])
```





Inference in regression

The t-statistic for a null hypothesis $H_0 : \beta_k = \beta_k^0$ has degrees of freedom $df = n - p - 1$, where n is the number of cases and p is the number of predictors in the model. The value β_k^0 equals 0 when the null hypothesis is one of no association.

$$t = \frac{b_k - \beta_k^0}{\text{s.e.}(b_k)} = \frac{b_k}{\text{s.e.}(b_k)}$$

A 95% confidence interval for β_k has the following formula, where $t^\star$ is the point on a t-distribution with $n - p - 1$ degrees of freedom and $\alpha/2$ area to the right.

$$b_k \pm \left(t^\star \times \text{s.e.}(b_k)\right)$$

The F-statistic in multiple regression is used to test hypotheses similar to those in ANOVA. The null hypothesis $H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$ is tested against the alternative that at least one of the slope coefficients is not 0. A significant p-value for the F-statistic is evidence that the predictor variables in the model, when considered as a group, are associated with the response variable.

6. Carry out inference based on the linear model in Question 5.

```
#use summary(lm( ))
summary(model.RFFTvsEdu)
```

```
##
## Call:
## lm(formula = RFFT ~ Education, data = prevend.samp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -55.905 -15.975  -0.905  16.068  63.280
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)             40.941      3.203  12.783  < 2e-16 ***
## EducationLowerSecond    14.779      3.686   4.009 7.04e-05 ***
## EducationHigherSecond   32.133      3.763   8.539  < 2e-16 ***
## EducationUniv           44.964      3.684  12.207  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.87 on 496 degrees of freedom
## Multiple R-squared:  0.3072, Adjusted R-squared:  0.303
## F-statistic:  73.3 on 3 and 496 DF,  p-value: < 2.2e-16
```

a) Identify the t-statistics and p-values for each slope coefficient in the model. Interpret the p-values in the context of the data.

The t-statistics and p-values for the slope coefficients $EduLowSec$, $EduHighSec$, and $EduUniv$ are 14.78 ($p < 0.0001$), 32.13 ($p < 0.0001$), and 44.96 ($p < 0.0001$). The three p-values are each smaller than $\alpha = 0.05$, indicating there is sufficient evidence to conclude

that, for each group, the mean RFFT is significantly different from the mean RFFT for the baseline group, primary school education. In each case, the mean RFFT is higher than mean RFFT for individuals who completed at most a primary school education.

b) Calculate and interpret the 95% confidence interval for the slope coefficient of X_3, university education.

The 95% confidence interval is (37.72, 52.20). We are 95% confident that the interval (37.72, 52.50) captures the amount by which mean RFFT score for individuals who have completed a university education is higher than mean RFFT score for those who have completed at most a primary school education.

```
#calculate confidence interval
confint(model.RFFTvsEdu, level = 0.95)
```

```
##                             2.5 %   97.5 %
## (Intercept)            34.648633 47.23372
## EducationLowerSecond    7.535743 22.02139
## EducationHigherSecond 24.739787 39.52711
## EducationUniv         37.726683 52.20109
```

c) From the F-statistic, determine whether there is evidence of a significant association between RFFT score and educational level.

The F-statistic is 73.3, with an associated p-value that is less than 0.0001. There is sufficient evidence to conclude that as a group, these variables are associated with the response variable; i.e., there is evidence of a significant association between RFFT score and educational level.

7. Conduct ANOVA to compare mean RFFT score among the four educational levels. Compare the results of inference based on the linear model to those based on ANOVA. For comparison purposes, leave the p-values uncorrected.

The F-statistic from ANOVA and from linear regression are the same (73.3). The null hypotheses that all the group means are equal (ANOVA) and that all the model slopes are equal to 0 (regression) are equivalent. Consider that each coefficient in the linear model is an estimate of the difference in mean RFFT for a particular educational level versus the baseline category. A significant F-statistic in regression indicates that at least one of the slope parameters does not equal 0; i.e., that at least one group mean is different from the mean of the reference category.

Inference based on the linear model tests the differences in means for each group relative to the reference group. Thus, the p-values in the first column of the table produced by `pairwise.t.test()` correspond to those from the regression approach, since these are for comparisons relative to the primary school group. The p-values provided by a summary of the regression model are unadjusted for multiple comparisons.

For a more detailed discussion of the connection between ANOVA and regression, refer to Section 7.9 in *OpenIntro Biostatistics*.

```
#ANOVA F-test
summary(aov(RFFT ~ Education, data = prevend.samp))
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## Education     3 115041   38347    73.3 <2e-16 ***
## Residuals   496 259469     523
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
#pairwise testing
pairwise.t.test(prevend.samp$RFFT, prevend.samp$Education,
            p.adj = "none")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  prevend.samp$RFFT and prevend.samp$Education
##
##              Primary LowerSecond HigherSecond
## LowerSecond  7.0e-05 -           -
## HigherSecond < 2e-16 2.6e-10     -
## Univ         < 2e-16 < 2e-16     2.3e-06
##
## P value adjustment method: none
```

8. Suppose that the linear model in Question 5 had been fit with the original version of Education that had not been converted to a factor.

   a) Load the prevend.samp data to return Education to its original coding as an integer vector.

```r
#load the data
data("prevend.samp")

#fit the model
lm(RFFT ~ Education, data = prevend.samp)
```

```
##
## Call:
## lm(formula = RFFT ~ Education, data = prevend.samp)
##
## Coefficients:
## (Intercept)    Education
##       41.15        15.16
```

   b) Fit a linear model predicting RFFT from educational level without converting Education to a factor.

      i. Interpret the slope coefficient of the model.

         According to this model, a one unit change in Education is associated with an increase in mean RFFT score of 15.16 points.

      ii. What does this model imply about the change in mean RFFT between groups? Explain why this model is flawed.

This model implies that the change in mean RFFT score associated with a one unit change in Education is necessarily equal regardless of the identity of the groups.

This model is flawed because it is not reasonable to assume that the difference in mean RFFT score when comparing, for example, the primary school group to the lower secondary group, will be equal to the difference in mean RFFT score between the higher secondary group and the university group.

Another point to be careful about with this model is that it would not provide consistent results if the numerical codes were altered. The numerical codes assigned to the groups are simply short-hand labels, and are assigned arbitrarily.

Reanalyzing the PREVEND data

The following questions return to examining the association between cognitive decline and statin use, after adjusting for potential confounders.

In addition to age, there are two natural candidates for potential confounders: educational level and presence of cardiovascular disease (CVD). Individuals with more education tend to have higher incomes and consequently, better access to health care and medication; also, individuals with more education may be more comfortable with assessments like the RFFT. Individuals with cardiovascular disease are often prescribed statins to lower cholesterol; cardiovascular disease can lead to vascular dementia and cognitive decline.

9. Fit the multiple regression model relating RFFT (RFFT) with statin use (Statin), adjusting for the potential confounders age (Age), educational level (Education), and presence of CVD (CVD). The variable CVD is coded 0 if CVD is absent and 1 if CVD is present.

```
#convert Education to a factor
prevend.samp$Education = factor(prevend.samp$Education,
                            levels = c(0, 1, 2, 3),
                            labels = c("Primary", "LowerSecond",
                                       "HigherSecond", "Univ"))

#convert Statin to a factor
prevend.samp$Statin = factor(prevend.samp$Statin,
                         levels = c(0, 1),
                         labels = c("NonUser", "User"))

#convert CVD to a factor
prevend.samp$CVD = factor(prevend.samp$CVD,
                      levels = c(0, 1),
                      labels = c("Absent", "Present"))

#fit the model
model1 = lm(RFFT ~ Statin + Age + Education + CVD, data = prevend.samp)
summary(model1)
```

```
##
## Call:
```

```
## lm(formula = RFFT ~ Statin + Age + Education + CVD, data = prevend.samp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -56.348 -15.586  -0.136  13.795  63.935
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)             99.03507    6.33012  15.645  < 2e-16 ***
## StatinUser               4.69045    2.44802   1.916  0.05594 .
## Age                     -0.92029    0.09041 -10.179  < 2e-16 ***
## EducationLowerSecond    10.08831    3.37556   2.989  0.00294 **
## EducationHigherSecond   21.30146    3.57768   5.954 4.98e-09 ***
## EducationUniv           33.12464    3.54710   9.339  < 2e-16 ***
## CVDPresent              -7.56655    3.65164  -2.072  0.03878 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.71 on 493 degrees of freedom
## Multiple R-squared:  0.4355, Adjusted R-squared:  0.4286
## F-statistic: 63.38 on 6 and 493 DF,  p-value: < 2.2e-16
```

10. Based on the model from Question 9, summarize the evidence for an association between statin use and decreased cognitive function.

    After adjusting for age, educational level, and the presence of cardiovascular disease, statin use is associated with a 4.7 point higher mean RFFT score (relative to non-use, when all other variables are held constant). The p-value for the slope coefficient is 0.056, which suggests moderate evidence of an association (significant at $\alpha = 0.10$, but not at $\alpha = 0.10$). These data do not support an association between statin use and decreased cognitive function.

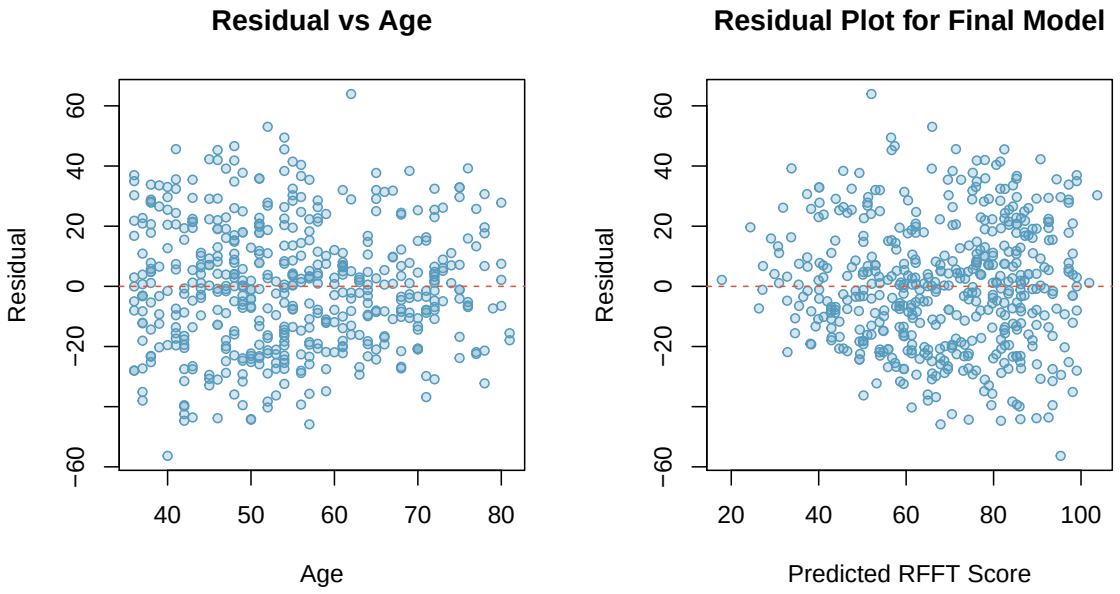
11. Evaluate the fit of the model.

    a) Assess the assumptions for linear regression.

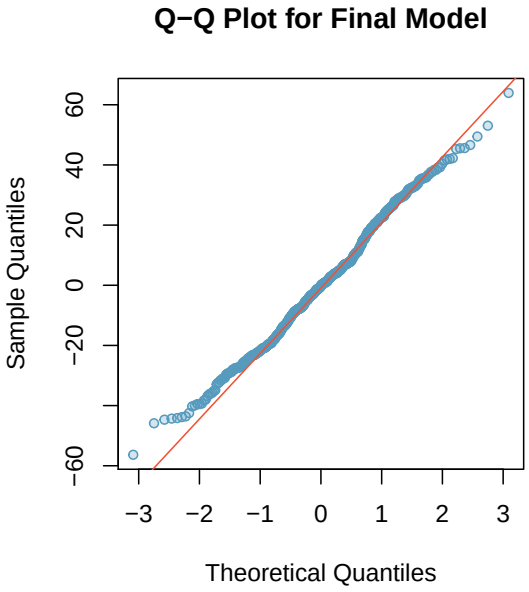
    Linearity only needs to be checked with respect to age. There does not appear to be remaining non-linearity with respect to age once the model is fit; the residuals scatter evenly across the $y = 0$ line with no apparent pattern. The residual plot indicates that constant variability of the residuals is reasonable; there is only a slight increase in variability for lager predicted values. The normal probability plot shows that the residuals depart slightly from normality in the tails. Overall, the assumptions for linear regression are reasonably satisfied.

```r
#load color package
library(openintro)
data(COL)

par(mfrow = c(1, 2))

#linearity with respect to age
```

```r
plot(resid(model1) ~ prevend.samp$Age,
 xlab = "Age", ylab = "Residual",
 main = "Residual vs Age",
 pch = 21, col = COL[1], bg = COL[1, 4], cex = 0.8)
abline(h = 0, col = COL[4], lty = 2)

#residual plot
plot(resid(model1) ~ fitted(model1),
 xlab = "Predicted RFFT Score",  ylab = "Residual",
 main = "Residual Plot for Final Model",
 pch = 21, col = COL[1], bg = COL[1, 4], cex = 0.8)
abline(h = 0, col = COL[4], lty = 2)
```




```r
#Q-Q plot
qqnorm(resid(model1),
   main = "Q-Q Plot for Final Model",
   pch = 21, col = COL[1], bg = COL[1, 4], cex = 0.8)
qqline(resid(model1),
   col = COL[4])
```



b) Comment on the R^2 and adjusted R^2 of the model. For comparison, the adjusted R^2 from the model including only age as a potential confounder is 0.282.

The R^2 of the model is 0.436; the model explains 43.6% of the observed variability in RFFT score. The adjusted R^2 is 0.429; the additional predictors relative to the model with only age as a confounder increase the strength of the model enough to justify the additional complexity.