Categorical Predictors

Lab 5 Handout Solutions

Statistics 139

Problem 2: Categorical predictors with multiple levels

The Prevention of REnal and Vascular END-stage Disease (PREVEND) study took place between 2003 and 2006 in the Netherlands. Clinical and demographic data for the 4,095 participants are in the `prevend.csv` data set.

Is RFFT score associated with educational attainment? The variable `Education` indicates the highest level of education that an individual completed: primary school (0), lower secondary school (1), higher secondary school (2), or university (3).

a) Add a variable to the `prevend` data frame that recodes `Education` as a factor variable. The original numeric version of the variable will be used in part (d).

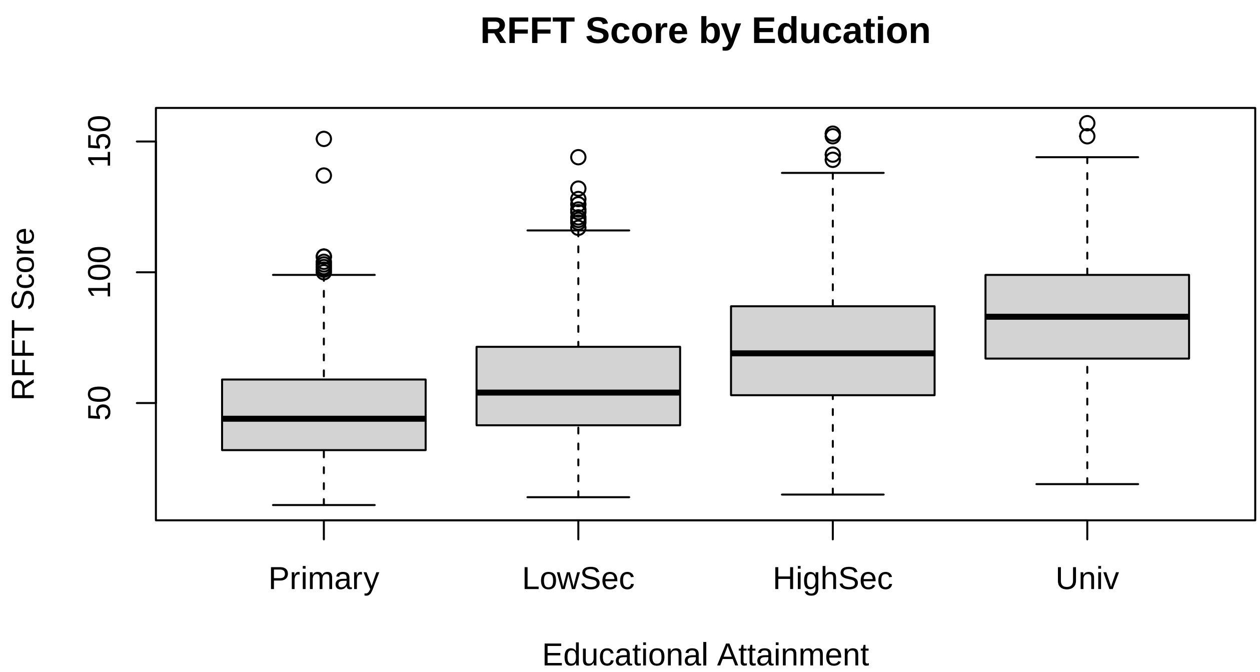
```
#load the data
prevend = read.csv("data/prevend.csv")

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

b) Create a plot that shows the association between RFFT score and educational attainment. Describe what you see.

A higher educational level is associated with higher median RFFT score. The boxplots almost look like they incraese linearly from one group to the next.

```
boxplot(prevend$RFFT ~ prevend$Education.Factor,
        main = "RFFT Score by Education", xlab = "Educational Attainment",
        ylab = "RFFT Score")
```



c) Apply the ANOVA procedure to explore whether RFFT score is associated with educational attainment. For the purposes of part d), do not apply a correction for multiple testing.

There is sufficient evidence to reject the overall F-test and reject the alternative there is a difference in some linear combination of mean RFFT scores for these groups ($F = 385.26,\ p < 0.0001$). All pairwise tests are highly significant, supporting the conclusion that the population mean RFFT for each group is different. The observed data suggest that as education level increases, population mean RFFT score increases.

```
#omnibus F test
model <- aov(RFFT ~ Education.Factor, data = prevend)
anova(model)
```

```
## Analysis of Variance Table
##
## Response: RFFT
##                    Df  Sum Sq Mean Sq F value    Pr(>F)
## Education.Factor    3  611381  203794  385.26 < 2.2e-16 ***
## Residuals        4091 2164057     529
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#pairwise tests
pairwise.t.test(prevend$RFFT, prevend$Education.Factor,
                p.adj = "none")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  prevend$RFFT and prevend$Education.Factor
##
```

```
##         Primary LowSec  HighSec
## LowSec  6.8e-14 -       -
## HighSec < 2e-16 < 2e-16 -
## Univ    < 2e-16 < 2e-16 < 2e-16
##
## P value adjustment method: none
```

d) Fit a linear model that regresses RFFT score on education level.

i. Fit the model using the factor version of `Education`. Interpret the coefficients, including the intercept. How do the values of the coefficients and associated p-values relate to the output from part c)?

```r
#means by group
tapply(prevend$RFFT, prevend$Education.Factor, mean)
```

```
##  Primary   LowSec  HighSec     Univ
## 47.42356 57.40347 71.01451 82.68791
```

```r
#linear model, part i.
summary(lm(RFFT ~ Education.Factor, data = prevend))
```

```
##
## Call:
## lm(formula = RFFT ~ Education.Factor, data = prevend)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -63.688 -16.403  -1.424  15.312 103.576
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)               47.424      1.151  41.187  < 2e-16 ***
## Education.FactorLowSec     9.980      1.327   7.518  6.8e-14 ***
## Education.FactorHighSec   23.591      1.344  17.558  < 2e-16 ***
## Education.FactorUniv      35.264      1.307  26.977  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23 on 4091 degrees of freedom
## Multiple R-squared:  0.2203, Adjusted R-squared:  0.2197
## F-statistic: 385.3 on 3 and 4091 DF,  p-value: < 2.2e-16
```

The intercept is the mean RFFT score for individuals who at most completed primary school (47.42). Each of the slope coefficients represents the change in mean RFFT score as compared to the baseline category, `Primary`. For `LowSec`, the mean RFFT score is 9.98 points higher; for `HighSec`, it is 23.59 points higher; for `Univ`, it is 35.26 points higher.

The p-values for the slope coefficients match the p-values in the first column of the pairwise t-test output; the regression compares all other groups with the baseline category, `Primary`.

ii. Fit the model using the numeric version of `Education`. How does the interpretation of this model differ from the interpretation of the model in part i.? Which model is preferable?

```r
#linear model, part ii.
summary(lm(RFFT ~ Education, data = prevend))
```

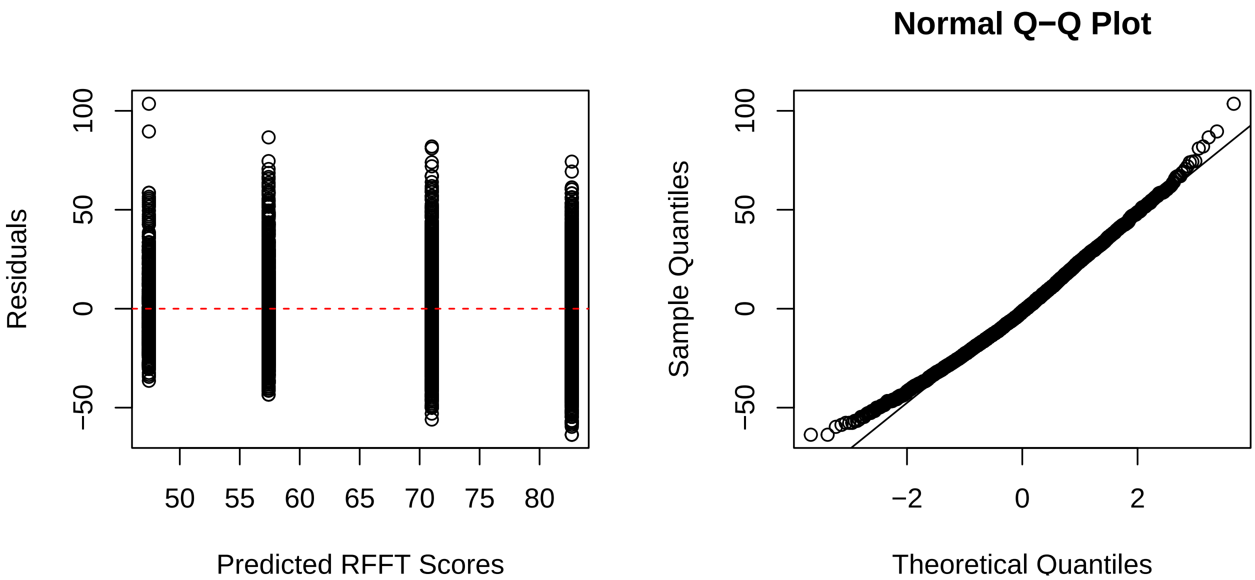
```
##
## Call:
## lm(formula = RFFT ~ Education, data = prevend)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -63.702 -16.296  -1.499  15.298 104.908
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  46.0925     0.7550   61.05   <2e-16 ***
## Education    12.2031     0.3596   33.94   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23 on 4093 degrees of freedom
## Multiple R-squared:  0.2196, Adjusted R-squared:  0.2194
## F-statistic:  1152 on 1 and 4093 DF,  p-value: < 2.2e-16
```

This model assumes that the estimated change in mean RFFT score between each level is equal. That is, going from Primary to Lower Secondary has an equivalent change as going between Higher Secondary and University. In general, a model where the categorical predictor is coded as a factor is preferable. For this particular setting, it is not detrimental to assume the equivalent step increase between levels; the difference in means between groups does seem roughly similar. Note, however, that blindly fitting the numeric variable can also lead to misinterpretation if the numeric codes do not correspond to the natural ordering of the factor level.

iii. Check the assumptions for the model in part i. Briefly comment on whether the assumptions seem reasonably satisfied.

```r
#check assumptions
par(mfrow = c(1, 2))
edu.model = lm(RFFT ~ Education.Factor, data = prevend)
plot(resid(edu.model) ~ fitted(edu.model), ylab = "Residuals",
     xlab = "Predicted RFFT Scores")
abline(h = 0, col = "red", lty = 2)

qqnorm(resid(edu.model))
qqline(resid(edu.model))
```



Linearity is automatically satisfied for categorical predictors. Constant variability seems reasonable across groups. The Q-Q plot shows the residuals are approximately normally distributed, with only slight deviations in the tails.

e) Is there evidence that mean RFFT score varies across levels of educational attainment? Perform a formal hypothesis test.

Test the null hypothesis $H_0 : \beta_1 = \beta_2 = \beta_3$ against the alternative that at least one $\beta_j \neq 0$. It yields the same F-statistic as the ANOVA (F=385.26, with $p < 2e - 16$. There is sufficient evidence to reject the null hypothesis and conclude that mean RFFT score varies across levels of educational attainment.

f) Let's consider two nested models for predicting RFFT score. The variables of interest are statin use (`Statin`), age (`Age`), and educational attainment (`Education.Factor`).

- Model 1: statin use, age
- Model 2, statin use, age, educational attainment

Formally compare the two models to assess whether educational attainment is a useful predictor.

```r
#fit the models
model1 <- lm(RFFT ~ Statin + Age, data = prevend)
model2 <- lm(RFFT ~ Statin + Age + Education.Factor, data = prevend)

#view the summary output
summary(model1)
```

```
##
## Call:
## lm(formula = RFFT ~ Statin + Age, data = prevend)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -68.740 -15.448  -0.842  14.630  78.249
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 130.96291    1.69809  77.124  < 2e-16 ***
## Statin       -4.04354    0.87320  -4.631 3.76e-06 ***
## Age          -1.12435    0.03124 -35.990  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.2 on 4092 degrees of freedom
## Multiple R-squared:  0.2733, Adjusted R-squared:  0.273
## F-statistic: 769.6 on 2 and 4092 DF,  p-value: < 2.2e-16
```

```r
summary(model2)
```

```
##
## Call:
## lm(formula = RFFT ~ Statin + Age + Education.Factor, data = prevend)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -65.688 -14.192  -1.099  13.416  88.125
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             104.80442    2.14657  48.824  < 2e-16 ***
## Statin                   -2.25250    0.81696  -2.757  0.00586 **
## Age                      -0.90174    0.03073 -29.346  < 2e-16 ***
```

```
## Education.FactorLowSec    5.87304    1.20140   4.888 1.06e-06 ***
## Education.FactorHighSec  13.75717    1.24935  11.011  < 2e-16 ***
## Education.FactorUniv     24.17763    1.22948  19.665  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.69 on 4089 degrees of freedom
## Multiple R-squared:  0.3694, Adjusted R-squared:  0.3686
## F-statistic: 479.1 on 5 and 4089 DF,  p-value: < 2.2e-16
```

```r
#run ESS F-test
anova(model1, model2)
```

```
## Analysis of Variance Table
##
## Model 1: RFFT ~ Statin + Age
## Model 2: RFFT ~ Statin + Age + Education.Factor
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1   4092 2016823
## 2   4089 1750156  3    266667 207.68 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value of the ESS F-test comparing the nested models, one without educational attainment and one with, is highly significant ($F = 207.68$, $p < 0.001$). There is evidence that including educational attainment in the model to predict RFFT score is worthwhile. There is a substantial increase in proportion of variation (in RFFT) explained from adding educational attainment.