Multiple Linear Regression Review

Jack Cunningham

```r
data <- read.csv("HoustonChronicle.csv")
head(data)
```

```
    District X.Repeating.1st.Grade X.Low.income.students Year   County
1      Alvin                   4.1                  49.7 2004 Brazoria
2      Alvin                   5.8                  41.1 1994 Brazoria
3   Angleton                   7.1                  44.2 2004 Brazoria
4   Angleton                   6.7                  30.2 1994 Brazoria
5 Brazosport                   7.3                  49.4 2004 Brazoria
6 Brazosport                   2.6                  33.7 1994 Brazoria
```

```r
attach(data)
```

To answer these three questions we use the model:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

Where Y is the percentage of students repeating first grade, x_1 is the percentage of low income students, and x_2 is a dummy variable that is equal to 1 when Year is 2004 and 0 when year is 1994.

```r
data$year_dummy <- ifelse(Year == 2004, 1,0)
```

```r
fit <- lm(X.Repeating.1st.Grade ~ X.Low.income.students + year_dummy +
            X.Low.income.students*year_dummy, data = data)
summary(fit)
```

```
Call:
lm(formula = X.Repeating.1st.Grade ~ X.Low.income.students +
    year_dummy + X.Low.income.students * year_dummy, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-8.1606 -2.6121 -0.5576  1.7495 11.6014

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                       3.27194    1.22347   2.674  0.00855 **
X.Low.income.students             0.06080    0.03093   1.966  0.05167 .
year_dummy                       -0.38956    1.76109  -0.221  0.82532
X.Low.income.students:year_dummy  0.01903    0.03949   0.482  0.63066
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.845 on 118 degrees of freedom
Multiple R-squared:  0.1288,    Adjusted R-squared:  0.1066
F-statistic: 5.813 on 3 and 118 DF,  p-value: 0.0009689
```

a)

The association between the percentage of low income students and the percentage of students repeating first grade is low and not statistically significant. If we test the hypothesis:

$$H_0 : \beta_1 = 0, H_a : \beta_1 \neq 0$$

The null hypothesis cannot be rejected with a p-value barely over a significance of $\alpha = 0.05167$.

The effect size is low as well, $\hat{\beta}_1 = 0.0608$ means an increase in one percentage of low income students only increases the percentage of students repeating first grade by 0.06%.

b)

To review, when the year is 1994 we have:

$$Y = \beta_0 + \beta_1 x_1$$

When the year is 2004 we have:

$$Y = (\beta_0 + \beta_2) + (\beta_1 + \beta_3)x_1$$

In order to test if there has been an increase in the percentage of students repeating first grade between 1994-1995 and 2004-2005 let us first test whether there is a difference between the two years. We do this by testing:

$$H_0 : \beta_2 = \beta_3 = 0, H_a : \beta_2 \neq 0 \text{ or } \beta_3 \neq 0$$

We have test statistic:

$$F = \frac{(RSS_{reduced} - RSS_{full})/(df_{reduced} - df_{full})}{RSS_{full/}/df_{full}}$$

Where $df_{full} = 61 - 3 - 1$ and $df_{reduced} = 61 - 1 - 1$.

```
fit_reduced <- lm(X.Repeating.1st.Grade ~ X.Low.income.students, data = data)
summary(fit_reduced)
```

```
Call:
lm(formula = X.Repeating.1st.Grade ~ X.Low.income.students, data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.9845 -2.5072 -0.4184  1.8505 11.1067 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)            2.91419    0.83836   3.476 0.000709 ***
X.Low.income.students  0.07550    0.01823   4.141 6.47e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.821 on 120 degrees of freedom
Multiple R-squared:  0.125,	Adjusted R-squared:  0.1177 
F-statistic: 17.14 on 1 and 120 DF,  p-value: 6.472e-05
```

```
df_full = 61 - 3 - 1
df_reduced = 61 - 1 -1
```

```r
RSS_full = sum(fit$residuals^2)
RSS_reduced = sum(fit_reduced$residuals^2)

F = (RSS_reduced - RSS_full)/(df_reduced - df_full) /
  (RSS_full/df_full)
p_value = 1 - pf(F, df_reduced - df_full, df_full)
cat("F Statistic: ",F, " p value: ", p_value)
```

```
F Statistic:  0.1227336  p value:  0.8847324
```

With a p value of 0.8847324 we can cannot reject the null hypothesis that no increase in the percentage of students is due to the year being 1994-1995 or 2004-2005.

c)