Chapter 5 Homework Answers

John Fox

Soc. 740, Winter 2012

Exercise 5.6 (optional)

When an explanatory variable is added to the model, the residual sum of squares cannot get larger. Because the regression coefficients are calculated to minimize the residual sum of squares, the worst that can happen is that the coefficient of the new variable will be identically zero, and the residual sum of squares will not get smaller. Since the squared multiple correlation is (TSS - RSS)/TSS, the correlation can only get bigger or remain the same when a variable is added. (The total sum of squares depends only on the Y-values, not on the explanatory variables in the model; it is the same regardless of the model fit to the data — it is the decomposition of TSS into RSS and RegSS that changes.)

Exercise D5.3

Regression of TFR on GDP Per Capita:

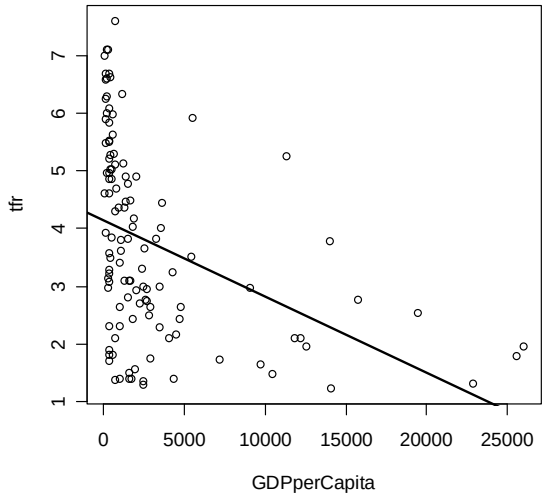
```
> UN <- read.table(
+ "http://socserv.mcmaster.ca/jfox/Books/Applied-Regression-2E/datasets/UnitedNations.txt",
+   header=TRUE)
> UN <- na.omit(UN[,c("tfr", "GDPperCapita", "illiteracyFemale", "contraception")])
> plot(tfr ~ GDPperCapita, data=UN)
> mod.gdp <- lm(tfr ~ GDPperCapita, data=UN)
> abline(mod.gdp, lwd=2)
> summary(mod.gdp)

Call:
lm(formula = tfr ~ GDPperCapita, data = UN)

Residuals:
    Min      1Q  Median      3Q     Max
-2.6758 -1.1183 -0.1419  1.0434  3.5492

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.147e+00  1.663e-01  24.939  < 2e-16 ***
GDPperCapita -1.320e-04  2.793e-05  -4.725 6.51e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.539 on 116 degrees of freedom
Multiple R-squared: 0.1614,     Adjusted R-squared: 0.1542
F-statistic: 22.33 on 1 and 116 DF,  p-value: 6.51e-06
```



Interpretation:

- $A = 4.147$: When GDP per capita is 0 dollars, the average total fertility rate is 4.1 children per woman. Of course, per capita GDP can't be 0, but there are some countries quite close to 0.
- $B = -0.000132$: When GDP per capita increases by 1 dollar, on average the TFR decreases by about 0.0001 child per woman; put another way, when GDP per capita increases by \$10,000, on average the TFR decreases by about 1 child per woman.
- $S_E = 1.539$: A typical residual is about 1.5 children per woman from the regression line, a substantial amount.
- $r = -\sqrt{.1614} = -.402$: The regression accounts for about 16 percent of the variation in the TFR among the countries.

The least squares line is not a good summary of the relationship between TFR and per capita GDP, which is highly nonlinear with some substantial outliers.

Note: I've filtered the data set for missing data, so that all of the regressions will be on the same subset of 118 observations with valid data for all four variables .

Regression of TFR on Female Illiteracy:

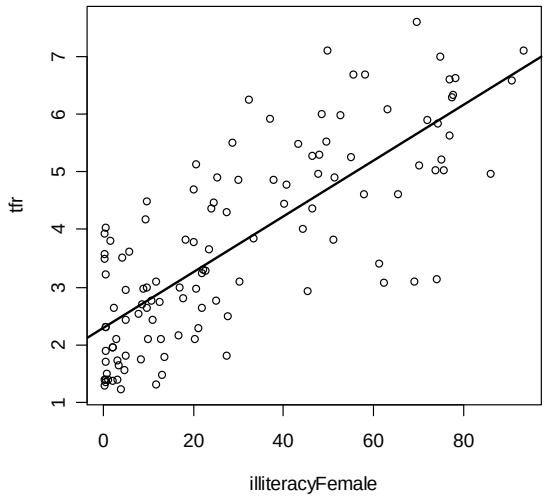
```
> plot(tfr ~ illiteracyFemale, data=UN)
> mod.illit <- lm(tfr ~ illiteracyFemale, data=UN)
> abline(mod.illit, lwd=2)
> summary(mod.illit)

Call:
lm(formula = tfr ~illiteracyFemale, data = UN)

Residuals:
     Min       1Q   Median       3Q      Max
-2.72565 -0.80639 -0.07995  0.74146  2.40063

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       2.28938    0.14569   15.71   <2e-16 ***
illiteracyFemale  0.04839    0.00363   13.33   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.056 on 116 degrees of freedom
Multiple R-squared: 0.6051,     Adjusted R-squared: 0.6017
F-statistic: 177.7 on 1 and 116 DF,  p-value: < 2.2e-16
```



Interpretation:

- $A = 2.28938$: When female illiteracy is 0 percent, the average total fertility rate is 2.3 children per woman.
- $B = 0.04839$: When female illiteracy increases by 1 percent, on average the TFR increases by about 0.05 children per woman.
- $S_E = 1.056$: A typical residual is about one child per woman from the regression line, a moderate error.

- $r = +\sqrt{.6051} = .778$: The regression accounts for about 61 percent of the variation in the TFR among the countries.

The least squares line is a reasonable of the roughly linear relationship between TFR and female illiteracy.

Regression of TFR on Contraception:

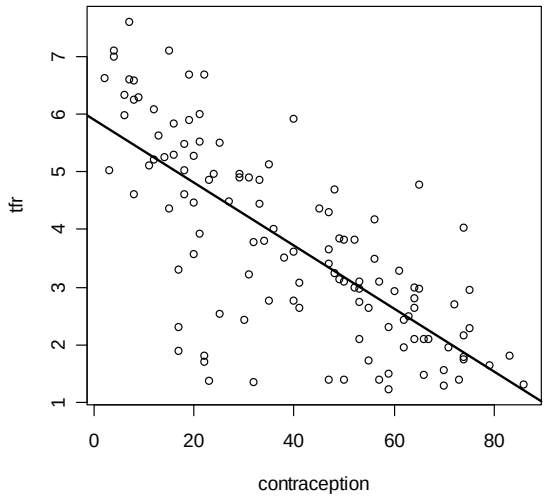
```
> plot(tfr ~ contraception, data=UN)
> mod.con <- lm(tfr ~ contraception, data=UN)
> abline(mod.con, lwd=2)
> summary(mod.con)

Call:
lm(formula = tfr ~contraception, data = UN)

Residuals:
     Min       1Q   Median       3Q      Max
-3.26855 -0.52558  0.07253  0.71491  2.41757

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    5.905947   0.212777   27.76   <2e-16 ***
contraception -0.054669   0.004666  -11.71   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.137 on 116 degrees of freedom
Multiple R-squared: 0.542,      Adjusted R-squared: 0.538
F-statistic: 137.3 on 1 and 116 DF,  p-value: < 2.2e-16
```



Interpretation:

- $A = 5.905947$: When contraceptive use is 0 percent, the average total fertility rate is nearly 6 children per woman.
- $B = -0.054669$: When contraceptive use increases by 1 percent, on average the TFR decreases by about 0.05 children per woman.
- $S_E = 1.137$: A typical residual is about 1.1 children per woman from the regression line, a moderate error.

- $r = -\sqrt{.542} = -.736$: The regression accounts for about 54 percent of the variation in the TFR among the countries.

The least squares line is a reasonable summary of the roughly linear relationship between TFR and contraception.

Exercise D5.5

```
> mod <- lm(tfr ~ GDPperCapita + illiteracyFemale + contraception, data=UN)
> summary(mod)

Call:
lm(formula = tfr ~ GDPperCapita + illiteracyFemale + contraception,
    data = UN)

Residuals:
    Min      1Q  Median      3Q     Max
-2.0259 -0.5864  0.0830  0.5195  2.1703

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       4.021e+00  2.837e-01  14.175  < 2e-16 ***
GDPperCapita     -2.998e-05  1.771e-05  -1.693   0.0933 .
illiteracyFemale  3.188e-02  3.905e-03   8.165 4.83e-13 ***
contraception    -2.884e-02  4.800e-03  -6.008 2.30e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8952 on 114 degrees of freedom
Multiple R-squared: 0.7212,     Adjusted R-squared: 0.7138
F-statistic: 98.28 on 3 and 114 DF,  p-value: < 2.2e-16
```

Interpretation:

- $A = 4.021$: When GDP per capita, female illiteracy, and contraceptive use are all 0, the average total fertility rate is about 4 children per woman. There are no countries in the data set with this combination of explanatory-variable values.
- $B_1 = -0.00002998$: When GDP per capita increases by \$1, holding female illiteracy and contraception constant, on average the TFR decreases by about 0.00003 children per woman, or by about .3 children for a \$10,000 increase in per capita GDP; this is a small effect, and about one quarter the magnitude of the marginal regression coefficient for GDP per capita, -0.000132.
- $B_2 = 0.03188$: When female illiteracy increases by 1 percent, holding GDP per capita and contraception constant, on average the TFR increases by about 0.03 children per woman; this is somewhat smaller than the marginal regression coefficient for female illiteracy, 0.04839.
- $B_3 = -0.02884$: When contraception increases by 1 percent, holding GDP per capita and female illiteracy constant, on average the TFR decreases by about 0.03 children per woman; this is about half the magnitude of the marginal regression coefficient for contraception, -0.054669.
- $S_E = 0.8952$: A typical residual is about 0.9 children per woman from the regression line, a moderate error.
- $R = \sqrt{.7212} = .849$: The regression accounts for about 72 percent of the variation in the TFR among the countries.