SMM637 Answers 4 - Regression with Dummy Variables

1. (a)

```
install.packages("wooldridge")
library(wooldridge)
data(wage1)
```

(b)

```
wage.lm <-lm(wage ˜ female + educ + exper + tenure, data = wage1)
summary(wage.lm)

Call:
lm(formula = wage ˜ female + educ + exper + tenure, data = wage1)

Residuals:
    Min      1Q  Median      3Q     Max
-7.7675 -1.8080 -0.4229  1.0467 14.0075

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.56794    0.72455  -2.164   0.0309 *
female      -1.81085    0.26483  -6.838 2.26e-11 ***
educ         0.57150    0.04934  11.584  < 2e-16 ***
exper        0.02540    0.01157   2.195   0.0286 *
tenure       0.14101    0.02116   6.663 6.83e-11 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 2.958 on 521 degrees of freedom
Multiple R-squared:  0.3635,    Adjusted R-squared:  0.3587
F-statistic:  74.4 on 4 and 521 DF,  p-value: < 2.2e-16
```

(c) The negative intercept is the average wage for men when all values for `educ`, `exper`, and `tenure` are set to zeroes, which in this case is not very meaningful.

(d) The coefficient on `female` is statistically different and it measures the average difference in hourly wage between a man and a woman who have the same levels of `educ`, `exper`, and `tenure`. If we take a woman and a man with the same levels of education, experience, and tenure, the woman earns, on average, $1.81 less per hour than the man.

(e)

```
> wage1.lm <-lm(wage ˜ female, data = wage1)
> summary(wage1.lm)

Call:
lm(formula = wage ˜ female, data = wage1)

Residuals:
    Min      1Q  Median      3Q     Max
-5.5995 -1.8495 -0.9877  1.4260 17.8805

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.0995     0.2100  33.806  < 2e-16 ***
female       -2.5118     0.3034  -8.279 1.04e-15 ***
```

```
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 3.476 on 524 degrees of freedom
Multiple R-squared:  0.1157,    Adjusted R-squared:  0.114
F-statistic: 68.54 on 1 and 524 DF,  p-value: 1.042e-15
```

(f) The intercept is the average wage for men in the sample (let female = 0), so men earn \$7.10 per hour on average. The coefficient on female is the difference in the average wage between women and men. Thus, the average wage for women in the sample is $7.10 - 2.51 = 4.59$, or \$4.59 per hour. (Incidentally, there are 274 men and 252 women in the sample.)

(g) The estimated wage differential between men and women is larger in (e) than in (b) because model in (e) does not control for differences in education, experience, and tenure, and these are lower, on average, for women than for men in this sample. The model fitted in (b) gives a more reliable estimate of the ceteris paribus gender wage gap; it still indicates a very large differential.

(h)

```
> t.test(wage1$wage[wage1$female==1], wage1$wage[wage1$female==0],
+        var.equal = TRUE)

        Two Sample t-test

data:  wage1$wage[wage1$female == 1] and wage1$wage[wage1$female == 0]
t = -8.2787, df = 524, p-value = 1.042e-15
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.107878 -1.915782
sample estimates:
mean of x mean of y
 4.587659  7.099489
```

Model fitted in (e) provides a simple way to carry out a comparison-of-means test between the two groups, which in this case are men and women. The estimated difference, -2.51, has a t-statistic of -8.2787, which is very statistically significant (and, of course, \$2.51 is economically large as well). Generally, simple regression on a constant and a dummy variable is a straightforward way to compare the means of two groups. For the usual t-test to be valid, we must assume that the homoskedasticity assumption holds, which means that the population variance in wages for men is the same as that for women.

(i) The new model is

$$\texttt{wage} = \alpha + \beta_1\texttt{female} + \beta_2\texttt{educ} + \beta_3\texttt{exper} + \beta_4\texttt{tenure} + \beta_5\texttt{female} * \texttt{educ} + \epsilon$$

```
wage2.lm <-lm(wage ~ female * educ + exper + tenure, data = wage1)
summary(wage2.lm)

Call:
lm(formula = wage ~ female * educ + exper + tenure, data = wage1)

Residuals:
    Min      1Q  Median      3Q     Max
-7.8607 -1.7730 -0.4345  1.0240 14.0358

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.27461    0.87227  -2.608  0.00938 **
female      -0.06001    1.23480  -0.049  0.96125
educ         0.62577    0.06186  10.116  < 2e-16 ***
exper        0.02563    0.01156   2.217  0.02702 *
tenure       0.14233    0.02116   6.727 4.58e-11 ***
female:educ -0.13974    0.09626  -1.452  0.14721
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 2.954 on 520 degrees of freedom
Multiple R-squared:  0.3661,    Adjusted R-squared:   0.36
F-statistic: 60.07 on 5 and 520 DF,  p-value: < 2.2e-16
```

This model is not supported because the coefficient associated with the interaction term `female:educ` is not significant.