QAC 201: Applied Data Analysis

More with interpretations of models

1. We would like to explore the relationship between work experience (exper) and wages earned per hour for women in the workforce in 1975.

```
model1<-lm(wage~exper, data=mroz)
summary(model1)
```

```
##
## Call:
## lm(formula = wage ~ exper, data = mroz)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1056 -1.8940 -0.6859  0.7095 21.0717
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.88308    0.30455  12.750   <2e-16 ***
## exper        0.02260    0.01988   1.137    0.256
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.309 on 426 degrees of freedom
##   (325 observations deleted due to missingness)
## Multiple R-squared:  0.003024,   Adjusted R-squared:  0.0006837
## F-statistic: 1.292 on 1 and 426 DF,  p-value: 0.2563
```

The conclusion here would be that there is not a significant linear relationship between experience and hourly wage (B=0.0226, p-value=0.256)

2. Now suppose we are looking to explore the relationship between whether they live in the city (0=no, yes=1) and hourly wage. While we could explore this relationship with an appropriate bivariate test (here, that would be ANOVA), suppose instead we wish to construct an appropriate regression model.

```
model2<-lm(wage~factor(city), data=mroz)
summary(model2)
```

```
##
## Call:
## lm(formula = wage ~ factor(city), data = mroz)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.3026 -1.8704 -0.6991  0.8706 20.5265
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     3.6513     0.2652  13.770   <2e-16 ***
```

```
## factor(city)1   0.8223     0.3314   2.481   0.0135 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.29 on 426 degrees of freedom
##   (325 observations deleted due to missingness)
## Multiple R-squared:  0.01425,    Adjusted R-squared:  0.01193
## F-statistic: 6.157 on 1 and 426 DF,  p-value: 0.01348
```

The model suggests that there is a significant relationship between city status and wage (B=0.8223, p-value=0.0135). In particular, working in the city is associated with an hourly wage that is expected to be $0.82 higher than someone not working in the city.

3. We can continue to build the model to include additional control variables. Suppose now we wish to examine the relationship between city status and wage after controlling for experience and spouse's wage.

```
model3<-lm(wage~factor(city)+exper+huswage, data=mroz)
summary(model3)
```

```
##
## Call:
## lm(formula = wage ~ factor(city) + exper + huswage, data = mroz)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.2236 -1.7447 -0.7237  0.8070 20.7787
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2.17470    0.46406   4.686 3.75e-06 ***
## factor(city)1  0.28927    0.34984   0.827    0.409
## exper          0.03189    0.01957   1.630    0.104
## huswage        0.19402    0.04736   4.097 5.02e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.23 on 424 degrees of freedom
##   (325 observations deleted due to missingness)
## Multiple R-squared:  0.05447,    Adjusted R-squared:  0.04778
## F-statistic: 8.141 on 3 and 424 DF,  p-value: 2.784e-05
```

We can now conclude that city status is not significantly related to hourly wage (B=0.29, p-value=0.409) after controlling for work experience and spouse wage. It appears that these variables confound the relationship. You can continue making other interpretations in your model as well. For instance, it appears that spouse's wage (labeled huseduc in data) is significantly and positively associated with hourly wage (B=0.19, p-value<0.001) after controlling for city and experience.

4. Now we are looking to explore the relationship between maximum educational attainment level (defined as High School or Less, Some College, or Grad School) and hourly wage for this same sample.

```
model<-lm(wage~education, data=mroz)
summary(model)
```

```
##
## Call:
## lm(formula = wage ~ education, data = mroz)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.0775 -1.6647 -0.5171  0.9116 21.4165
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     4.0369     0.3856  10.468  < 2e-16 ***
## educationGrad School            2.4233     0.5240   4.625 4.98e-06 ***
## educationHigh School or Less   -0.4533     0.4281  -1.059     0.29
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.133 on 425 degrees of freedom
##   (325 observations deleted due to missingness)
## Multiple R-squared:  0.1085, Adjusted R-squared:  0.1044
## F-statistic: 25.88 on 2 and 425 DF,  p-value: 2.488e-11
```

Notice that even though we have only a single explanatory variable (education) there are two slope terms to consider for education. Every categorical explanatory variable that you use in a model will have a reference level and each level that appears in your model will be compared against that reference level. In this case, the reference level is 'College Education'.

In this model we can conclude that those who have a graduate school education make significantly more money on average ***than people with a maximum educational attainment level of Undergraduate College*** *(B=2.42, p-value<0.001). This model estimates that those who have a graduate school education make an average of $2.42 more per hour on average than those with only a maximum educational attaintment level of undergradute college.*

In this model we we do not have enough evidence to show that those who have have a maximum educational attainment of High School education or less make a significantly different amount on average ***compared to those with only an undergraduate college education*** *(B=-0.45, p-value=0.290). While the model estimates that those who have a high school education or less make an average of $0.42 less per hour than those with only an undergraduate education, the difference is not deemed significant.*

You may now be wondering, is there a difference between hourly salaries of those with Grad School educations compared to those with High School or Less? Our current model does not allow you to test this, so if it is of interest, you would need to change the reference level. We only recommend you change the reference level, if it is interesting to your research question.

```r
# Code below changes the reference level to Grad School
mroz$education<-as.factor(mroz$education)
mroz$education<-relevel(mroz$education, ref="Grad School")

# Re-run your model code and your output will give you
# two different slope terms
model<-lm(wage~education, data=mroz)
summary(model)
```

```
##
## Call:
## lm(formula = wage ~ education, data = mroz)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.0775 -1.6647 -0.5171  0.9116 21.4165
```

```
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     6.4601     0.3547  18.212  < 2e-16 ***
## educationCollege               -2.4233     0.5240  -4.625 4.98e-06 ***
## educationHigh School or Less  -2.8766     0.4005  -7.183 3.08e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.133 on 425 degrees of freedom
##   (325 observations deleted due to missingness)
## Multiple R-squared:  0.1085, Adjusted R-squared:  0.1044
## F-statistic: 25.88 on 2 and 425 DF,  p-value: 2.488e-11
```

There is enough evidence to suggest that those who have a maximum educational attainment level of High School or Less make significantly less money on average compared to those with graduate school education (B=-2.877, p-value<0.001). The model estimates that those with High School or Less make an average of $2.87 less per hour than those with graduate school education.

This next part should sound very familiar (look at our previous model). There is enough evidence to suggest that those who have a maximum educational attainment level of undergraduate college make significantly less money on average compared to those with graduate school education (B=-2.42, p-value<0.001). The model estimates that those with a maximum attainment level of undergraduate college make an average of $2.42 less per hour than those with graduate school education.

Running a third model will not be particularly useful since there are no more comparisons to make.

The models that you obtain by changing the reference are algebraically equivalent. Why?

Our model in this example can be written out as:

$Wage = 6.46 - 2.42College - 2.88HighSchoolOrLess$

Therefore, if I wanted to predict the hourly wage for someone whose maximum Education level was high school or less, I would plug in a dummy code of 1 for HighSchoolOrLess in the model and a 0 for College.

$Wage = 6.42 - 2.42(0) - 2.88(1) = 3.42$

If I wanted to predict the hourly wage for someone whose maximum educational level was undergraduate college, I would then plug in a 1 for College and a 0 for HighSchool or Less.

$Wage = 6.42 - 2.42(1) - 2.88(0) = 4.00$

Finally, if I wanted to predict the hourly wage for someone whose maximum educational level was graduate school, I would then plug in a 0 for Colelge and a 0 for HighSchool or Less.

$Wage = 6.42 - 2.42(0) - 2.88(0) = 6.42$

Can you see how you would have obtained the same predictions in the model before this one?