Econ 5100 - Quantitative Methods and Applications
In-class problem set 2

What determines wages?

What factors determine a person's wage is of just importance for policy. Today we are do some simple analyses using a dataset on individual characteristics and wage outcomes in 1980. You can find the CVS dataset under data in Canvas (Wages).

The variables in the data set are:

- 1. wage monthly earnings
- 2. hours average weekly hours
- 3. IQ IQ score
- 4. KWW knowledge of world work score
- 5. educ years of education
- 6. exper years of work experience
- 7. tenure years with current employer
- 8. age age in years
- 9. married =1 if married
- 10. black =1 if black
- 11. south =1 if live in south
- 12. urban =1 if live in SMSA
- 13. sibs number of siblings
- 14. brthord birth order
- 15. meduc mother's education
- 16. feduc father's education
- 17. lwage natural log of wage

The dependent variable is going to be monthly earnings. We are going to use a combination of average weekly hours, IQ score, years of education, years of work experience, years with current employer, age in years, mother's education, and father's education.

1. Do the normal descriptive statistics for the variables we are going to use. Anything stand out among these numbers?

   Answer: There is substantial number of missing for both parental education variables. There is relatively little variation in age.

2. Find the correlations between the variables we will use. What stands out here?

   Answer: None of the correlations are very strong. The strongest correlations are between IQ and education, and between age and experience.

3. Run a regression with hours and IQ score as the explanatory variables. What are the coefficients for both and their interpretation (remember to include whether they are statistically significant).

   Answer: [See model summary at end]. Every additional hour worked per week is associated with a reduction of $1.82 in the monthly wage, holding IQ constant, although this effect is not statistically significant. Increasing IQ by one point is associated with an increase of $8.37 in monthly wage, holding hours worked constant. This effect is statistically significant at the 1% level.

4. Run the same regression as in Q3, but now add education as an additional explanatory variable. What are the coefficients and their interpretations (remember to include whether they are statistically significant).

   Answer: [See model summary at end]. The effect of hours is still not statistically significant, but the effect is now a reduction of $2.51 per extra hour worked. IQ is now associated with a $5.18 increase in wage per extra IQ point and this is statistically significant at the 1% level. Finally, every additional year of education is associated with an increase of $42.65 in monthly wage, with the effect statistically significant at the 1% level.

5. Explain why the coefficients change between the two regression. Are the changes in the direction you would expect?

   Answer: Both of the original coefficients are closer to zero than without education. This happens because somebody with more education tend to both work longer hours and to have a higher IQ score. I am not surprised that higher IQ and higher education go hand-in-hand, while I did not have a strong prior expectation about the relationship between hours worked and education (except for what I learned from the correlation above). The main point of interest is that the original IQ scored captured both the effect of IQ and part of the effect of education and was therefore likely too high.

6. Run a regression with hours, IQ score, education, age, experience, and tenure as the explanatory variables. Interpret the results.

   Answer: [See model summary at end]

7. Explain why the coefficients change from the model in Q4. Does this make sense and why? (Hint: think what is behind the correlations you observed)

   Answer: The effect of IQ does not change substantially, which is not surprising since there should be relatively little relation between your age, how long you worked overall or at a specific company, and your IQ score. What is more interesting is what happens to the education coefficient, which now is larger than before. One way to explain this is hold age constant, and think about experience. For a given age, the only way to have more experience is to leave school earlier! We know that more experience is associated with higher wage. Hence, the original education coefficient was lower, exactly because the only way to get more education is to give up experience, which was then incorporated into the original estimate.

8. What does the F-test tell you about the model in Q6? What about the R^2 and adjusted R^2?

   Answer: We clearly reject the null hypothesis that all the explanatory variables have a zero effect (F-stat of 33.37). This is not surprising given that all but hours have statistically significant effects. We explain about 17.8% of the variation in the monthly with our model. The adjusted R^2 is almost identical, which again is not surprising with the statistically significant effects.

9. Can we drop tenure and hours? (Do a joint test). What are the results and how do we interpret them?

   No! The F-statistics is 4.1133, and we reject the null hypothesis that both are jointly equal to zero at the 5% level. (Note: whether this approach is a good idea is a different matter; we will get back to that later in the quarter).

10. Can you rule out that the effect of experience overall and tenure at the current job have the same effect on wage?

    We cannot rule out that the effects are the same, since the F-statistics is 0.60, which is nowhere close to where is should be to reject the null hypothesis. (You can get a ballpark idea of this simply by looking at the standard errors for each coefficient, but be careful, it is a solid test)

11. If you have time, try to add parental education to the last model. Think about what changes. Also, try some joint hypotheses (for example, parental education have similar effect; difference between age and experience).

    Answer: [See R code]

Model 1:

```
Call:
lm(formula = wage ~ hours + IQ, data = df)

Residuals:
   Min     1Q Median     3Q    Max
-888.1 -257.0  -46.4  204.3 2046.9

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 190.3847   110.9182   1.716   0.0864 .
hours        -1.8194     1.7474  -1.041   0.2981
IQ            8.3675     0.8386   9.977   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 384.7 on 932 degrees of freedom
Multiple R-squared:  0.09659, Adjusted R-squared:  0.09465
F-statistic: 49.82 on 2 and 932 DF,  p-value: < 2.2e-16
```

Model 2:

```
Call:
lm(formula = wage ~ hours + IQ + educ, data = df)

Residuals:
   Min     1Q Median     3Q    Max
-873.5 -256.5  -39.7  200.6 2075.4

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -31.1321   113.7586  -0.274    0.784
hours        -2.5096     1.7132  -1.465    0.143
IQ            5.1821     0.9557   5.422 7.50e-08 ***
educ         42.6527     6.5584   6.504 1.28e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 376.5 on 931 degrees of freedom
Multiple R-squared:  0.1358,  Adjusted R-squared:  0.1331
F-statistic: 48.78 on 3 and 931 DF,  p-value: < 2.2e-16
```

Model 3:

```
Call:
lm(formula = wage ~ hours + IQ + educ + age + exper + tenure,
    data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-846.40 -242.68  -44.56  189.42 2159.86

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -736.7637   171.6533  -4.292 1.96e-05 ***
hours         -2.2974     1.6796  -1.368  0.17170
IQ             5.1015     0.9391   5.432 7.10e-08 ***
educ          52.4772     7.2788   7.210 1.16e-12 ***
age           12.5442     4.7210   2.657  0.00802 **
exper          9.7844     3.7399   2.616  0.00904 **
tenure         6.0879     2.5000   2.435  0.01507 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 367.9 on 928 degrees of freedom
Multiple R-squared:  0.1775,  Adjusted R-squared:  0.1722
F-statistic: 33.37 on 6 and 928 DF,  p-value: < 2.2e-16
```