Introduction to Multiple Regression

2018-07-24

Introduction and Research Question

In this set of notes, you will continue your foray into regression analysis. To do so, we will again examine the question of whether education level is related to income using the *riverside.csv* data from C. Lewis-Beck & Lewis-Beck (2016). Specifically we will ask,

(1) Do differences in education level explain variation in incomes? *and*
(2) Do differences in education level explain variation in incomes even after accounting for differences in seniority?

Preparation

```
# Load libraries
library(broom)
library(corrr)
library(dotwhisker)
library(dplyr)
library(ggplot2)
library(readr)
library(sm)

# Read in data
city = read_csv(file = "~/Documents/github/epsy-8251/data/riverside.csv")
head(city)
```

```
# A tibble: 6 x 6
  education income seniority gender  male party
      <int>  <int>     <int> <chr>  <int> <chr>
1         8  37449         7 male       1 Democrat
2         8  26430         9 female     0 Independent
3        10  47034        14 male       1 Democrat
4        10  34182        16 female     0 Independent
5        10  25479         1 female     0 Republican
6        12  46488        11 female     0 Democrat
```

Answering the First Research Question

In previous notes, we fitted a model regressing employees' incomes on education level.

```
# Fit regression model
lm.1 = lm(income ~ 1 + education, data = city)

# Obtain model-level results
glance(lm.1)
```

```
  r.squared adj.r.squared    sigma statistic          p.value df    logLik
1 0.6316828     0.6194055 8978.116  51.45152 0.00000005562116  2 -335.6549
       AIC      BIC   deviance df.residual
1 677.3097 681.7069 2418196934          30
```

```
# Obtain coefficient-level results
tidy(lm.1)
```

```
         term  estimate std.error statistic          p.value
1 (Intercept) 11321.379 6123.2350  1.848921 0.07434601558004
2   education  2651.297  369.6232  7.172972 0.00000005562116
```

The fitted equation,

$$\hat{\text{Income}} = 11,321 + 2,651(\text{Education Level}),$$

suggests that the estimated mean income for employees with education levels that differ by one year varies by \$2,651. We also found that differences in education level explained 63.2% of the variation in income, and that this was statistically significant, $p < .001$. All this suggests that education level is related to income.

Examining the Seniority Predictor

Let's do some analysis on the seniority predictor.

```
# Examine the marginal distribution
sm.density(city$seniority, xlab = "Seniority level (in years)")
```

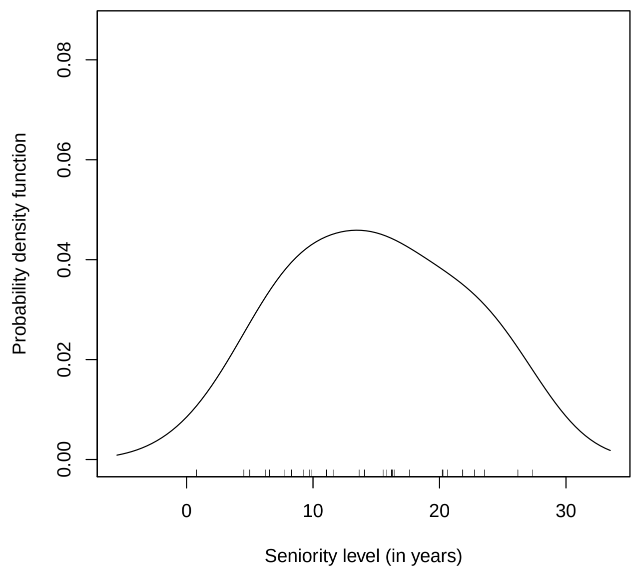

Figure 1: Density plot of the marginal distribution of seniority.

```
# Compute mean and standard deviation
city %>%
  summarize(
    M = mean(seniority),
    SD = sd(seniority)
    )
```

```
# A tibble: 1 x 2
      M    SD
  <dbl> <dbl>
1  14.8  6.95
```

Seniority is symmetric with a typical employee having roughly 15 years of seniority. There is quite a lot of variation in seniority, however, with most employees having between 8 and 22 years of seniority. After we examine the mariginal distribution, we should examine the relationships among all of three variables we are considering in the analysis. Typically researchers will examine the scatterplots between each predictor and the outcome (to evaluate the functional forms of the relationships with the outcome) and also examine the correlation matrix.

```
# Relationship between income and seniority
ggplot(data = city, aes(x = seniority, y = income)) +
  geom_point() +
  theme_bw() +
  xlab("Seniority (in years)") +
  ylab("Income (in dollars)")
```

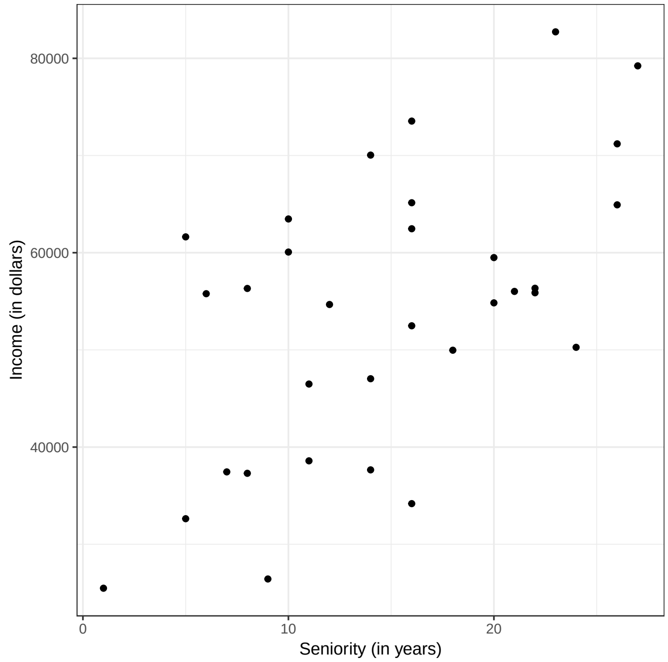

Figure 2: Scatterplot showing the relationship between seniority level and income.

```
# Correlation matrix
city %>%
  select(income, education, seniority) %>%
  correlate()
```

```
# A tibble: 3 x 4
  rowname   income education seniority
  <chr>      <dbl>     <dbl>     <dbl>
1 income    NA         0.795     0.582
2 education  0.795    NA         0.339
3 seniority  0.582     0.339    NA
```

The relationship between seniority and income seems linear and positive ($r = 0.58$). This suggests that employees with more seniority also tend to have higher incomes. Education level and seniority are also modestly correlated ($r = 0.34$), indicating that employees with higher education levels tend to also have more seniority.

Because the correlation between the two predictors is not 0, this calls into question our previous findings about whether there actually is a relationship between education level and income. It might be that this relationship is spurious. That really it is the fact that the reason we saw that employees with higher education levels tended to have higher incomes is that they also tend to have more seniority. What we need to know is whether **after we account for differences in seniority** is there is still a relationship between education level and income. To answer this question, we will need to fit a model that includes both predictors.

Simple Regression Model: Seniority as a Predictor of Income

Before we fit the model with both predictors, we will first fit the simple regression model using seniority as a predictor of variation in income.

```
lm.2 = lm(income ~ 1 + seniority, data = city)

# Model-level results
glance(lm.2)
```

```
  r.squared adj.r.squared    sigma statistic      p.value df    logLik
1 0.3386114     0.3165651 12031.02  15.35911 0.0004767296  2 -345.0212
       AIC      BIC   deviance df.residual
1 696.0424 700.4396 4342365212          30
```

```
# Coefficient-level results
tidy(lm.2)
```

```
         term  estimate std.error statistic          p.value
1 (Intercept) 35690.298 5073.4526  7.034716 0.00000008074761
2   seniority  1218.689  310.9638  3.919070 0.00047672958164
```

The fitted equation,

$$\hat{\text{Income}} = 35,690 + 1,219(\text{Seniority Level}),$$

suggests that the estimated mean income for employees with seniority levels that differ by one year varies by \$1,219. We also find that differences in seniority level explain 33.9% of the variation in income, and that this is statistically significant, $p < .001$. All this suggests that seniority level is related to income.

Multiple Regression Model: Education Level and Seniority as a Predictors of Income

To fit the multiple regression model, we will just add (literally) additional predictors to the right-hand side of the `lm()` formula.

```
lm.3 = lm(income ~ 1 + education + seniority, data = city)
```

Model-Level Results

To interpret multiple regression results, begin with the model-level information.

```
# Model-level results
glance(lm.3)
```

```
  r.squared adj.r.squared    sigma statistic           p.value df
1 0.7417857     0.7239778 7645.854   41.6549 0.000000002976563  3
     logLik      AIC      BIC   deviance df.residual
1 -329.9724 667.9448 673.8077 1695313285          29
```

Together, differences in education level AND seniority explain 74.2% of the variation in income, in the sample. We can test whether together these predictors explain variation in the population. The formal model-level null hypothesis that tests this can be written mathematically as,

$$H_0 : \rho^2 = 0.$$

This is a test of whether *all the predictors together* explain variation in the outcome variable. The results of this test, $F(3, 29) = 41.65$, $p < .001$, which is statistically significant, suggest that we should reject the null hypothesis; it is likely that together education level and seniority level explain variation in the population.

Equivallently, we can also write the hypothesis as a function of the predictor effects, namely,

$$H_0 : \beta_{\text{Education Level}} = \beta_{\text{Seniority}} = 0.$$

In plain English, this is akin to stating that there is NO EFFECT for every predictor included in the model. Rejection of this null hypothesis suggests that AT LEAST ONE of the predictor effects is likely not zero.

Although the two expressions of the model-level null hypothesis look quite different, they are answering the same question, namely whether the model is worthwhile in predicting variation in income.

Coefficient-Level Results

Now we turn to the coefficient-level information produced in the `tidy()` output.

```
# Coefficient-level results
tidy(lm.3)
```

```
         term  estimate std.error statistic         p.value
1 (Intercept) 6769.1720 5372.8914  1.259875 0.2177593428983
2   education 2251.8456  334.6443  6.729073 0.0000002202903
3   seniority  738.7965  210.0954  3.516481 0.0014597774256
```

First we will write the fitted multiple regression equation,

$$\hat{\text{Income}} = 6,769 + 2,252(\text{Education Level}) + 739(\text{Seniority Level}).$$

The slopes (of which there are now more than one) are referred to as *partial regression slopes* or *partial effects*. They represent the effect of the predictor *AFTER* accounting for the effects of the other predictors included in the model. For example,

- The **partial effect of education level** is 2,252. This indicates that a one year difference in education level is associated with a $2,252 difference in income (on average), after accounting for differences in seniority level.
- The **partial effect of seniority** is 739. This indicates that a one year difference in seniority level is associated with a $739 difference in income (on average), after accounting for differences in education level.

The language "after accounting for" is not ubiquitous in interpreting partial regression coefficients. Some researchers instead use "controlling for", "holding constant", or "partialling out the effects of". For example, the education effect could also be interpreted these ways:

> A one year difference in education level is associated with a $2,252 difference in income (on average), after controlling for differences in seniority.
>
> A one year difference in education level is associated with a $2,252 difference in income (on average), after holding the effect of seniority constant.
>
> A one year difference in education level is associated with a $2,252 difference in income (on average), after partialling out the effects of seniority.

Lastly, we can also interpret the intercept:

> The average income for all employees with 0 years of education AND 0 years of seniority is estimated to be $6,769.

This is the predicted avergage Y value when ALL the predictors have a value of 0. As such, it is often an extrapolated prediction and is not of interest to most applied researchers. For example, in our data, education level ranges from 8 to 24 years and seniority level ranges from 1 to 27 years. We have no data that has a zero value for either predictor, let alone for both. This makes prediction tenuous.

Coefficient-Level Inference

At the coefficient-level, the hypotheses being tested are about each individual predictor. The mathematical expression of the hypothesis is

$$H_0 : \beta_k = 0.$$

In plain English, the statistical null hypothesis states: After accounting for ALL the other predictors included in the model, there is NO EFFECT of X on Y. These hypotheses are evaluated using a t-test. For example, consider the test associated with the education level coefficient.

$$H_0 : \beta_{\text{Education Level}} = 0$$

This is akin to stating there is NO EFFECT of education level on income after accounting for differences in seniority level. The null hypothesis would be rejected, $t(29) = 6.73$, $p < .001$, suggesting that there is indeed an effect of education on income after controlling for differences in seniority level. (Note that the *df* for the t-test for all of the coefficient tests is equivalent to the error, or denominator, *df* for the model-level F-test.)

It is important to note that the p-value at the model-level is different from any of the coefficient-level p-values. This is because when we include more than one predictor in a model, the hypotheses being tested at the model- and coefficient-levels are different. The model-level test is a simultaneous test of all the predictor effects, while the coefficient-level tests are testing the added effect of a particualr predictor.

Multiple Regression: Statistical Model

The multiple regression model says that each case's outcome (Y) is a function of two or more predictors (X_1, X_2, $\ldots$, X_k) and some amount of error. Mathematically it can be written as

$$Y_i = \beta_0 + \beta_1(X1_i) + \beta_2(X2_i) + \ldots + \beta_k(Xk_i) + \epsilon_i$$

As with simple regression we are interested in estimating the values for each of the regression coefficients, namely, β_0, β_1, β_2, $\ldots$, β_k. To do this, we again employ least squares estimation to minimize the sum of the squared error terms.

Since we have more than one X term in the fitted equation, the structural part of the model no longer mathematically defines a line. For example, the fitted equation from earlier,

$$\hat{Y} = 6,769 + 2,252(X1) + 739(X2),$$

mathematically defines a regression plane. (Note we have three dimensions, Y, $X1$, and $X2$. If we add predictors, we have four or more dimensions and we describe a hyperplane.)

The data and regression plane defined by the education level, seniority level, and income for the City of Riverside employees is shown below. The regression plane is tilted up in both the education level direction (corresponding to a positive partial slope of education) and in the seniority level direction (corresponding to a positive partial slope of seniority). The blue points are above the plane (employees with a positive residual) and the yellow points are below the plane (employees with a negative residual).

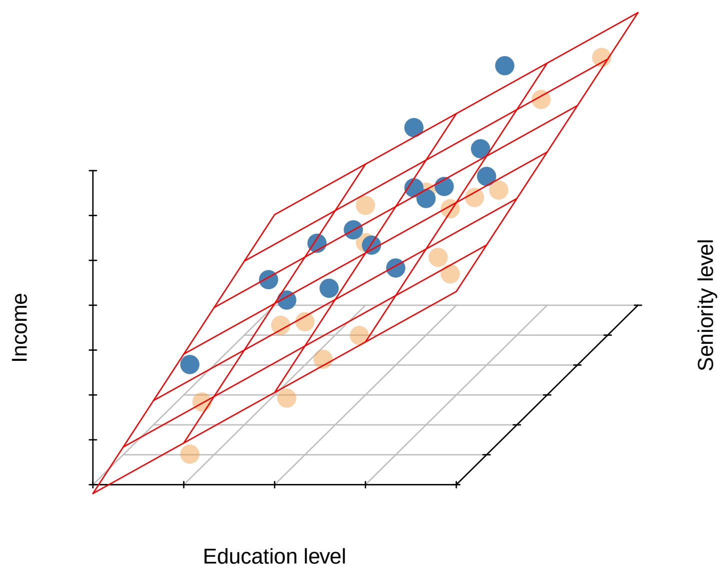

Figure 3: Three-dimensional scatterplot showing the relationship between education level, seniority, and income. The fitted regression plane is also shown. Blue observations have a positive residual and yellow observations have a negative residual.

The residual sum of squares can be obtained using the `anova()` function to give the ANOVA decomposition of the model.

```r
anova(lm.3)
```

```
Analysis of Variance Table

Response: income
          Df     Sum Sq    Mean Sq F value         Pr(>F)
education  1 4147330492 4147330492  70.944 0.000000002781 ***
seniority  1  722883649  722883649  12.366        0.00146 **
Residuals 29 1695313285   58459079
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Here the $\text{SS}_{\text{Residuals}} = 1,695,313,285$. Any other plane (i.e., different coefficient values for the intercept and predictors) would produce a higher sum of squared residuals value. Note that the *df* value in the `Residuals` row of the ANOVA output is another way to find the *df* associated with the *t*-tests for the coefficient tests we presented earlier.

Presenting Results

It is quite common for researchers to present the results of their regression analyses in table form. Different models are typically presented in different columns and predictors are presented in rows. (Because it is generally of less substantive value, the intercept is often presented in the last row.)

Table 1. *Regression Models Fitted to City Employee Data (n = 32) Using Education Level and Seniority to Predict Income*

	Model 1	Model 2	Model 3
Education level	2,651***		2,252***
	(370)		(335)
Seniority		1,219***	739***
		(311)	(210)
Intercept	11,321.380*	35,690.300***	6,769.172
	(6,123)	(5,073)	(5,373)
R^2	0.632	0.339	0.742
RMSE	8,978	12,031	7,646

Note: *p<0.1; **p<0.05; ***p<0.01

Based on these fitted models, we can now go back and answer our research questions. Do differences in education level explain variation in incomes? Based on Model 1 the answer is yes. Is this true even after accounting for differences in seniority? Model 3 suggests that, again, the answer is yes. (Since it is not germaine to answerig the RQs, Model 2 could just as easily be omitted from the table.)

Coefficient Plot

The `dw_plot()` function from the **dotwhisker** package automates much of the creation of regression coefficient plots. For example, to create the coefficient plot for Model 1 (`lm.1`), we (1) create the `tidy()` model object and then (2) submit that tidy object as an argument to the `dw_plot()` function. To also display the intercept, we also include the argument `show_intercept=TRUE`.

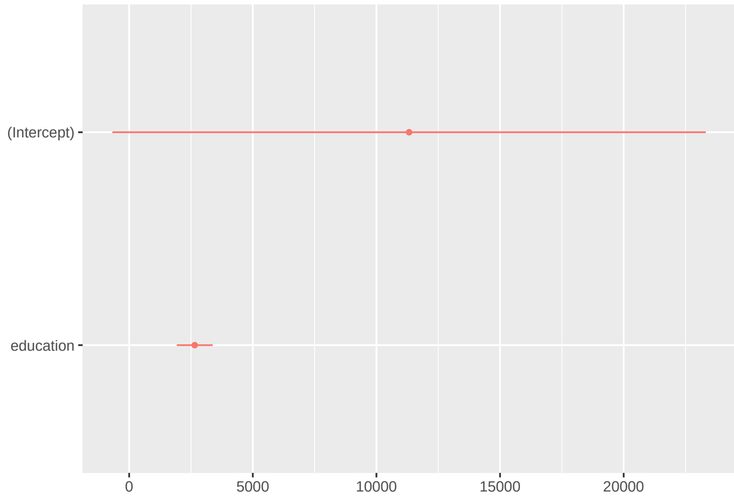

Figure 4: Coefficient plot for the model regressing income on education. Uncertainty based on the 95% confidence intervals are displayed.

We can also re-arrange the order of the variables displayed by `dw_plot()` and, since the output is a ggplot object, we can customize it by adding ggplot layers.

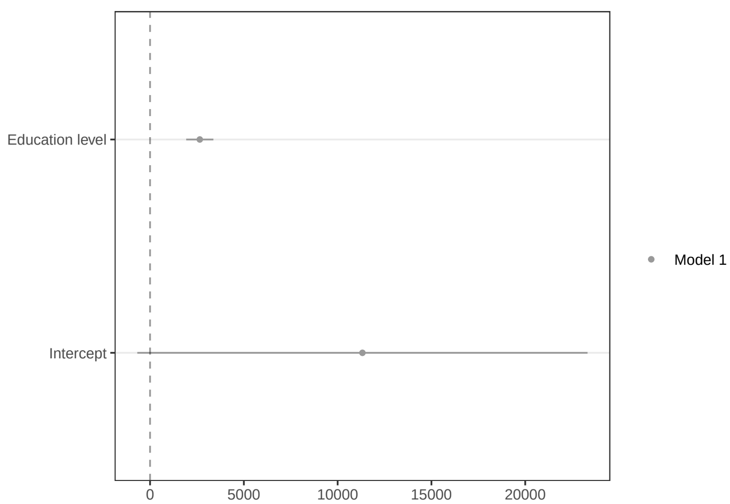

Figure 5: Coefficient plot for the model regressing income on education. Uncertainty based on the 95% confidence intervals are displayed.

It is critical when you are changing labels that you double-check the actual `tidy()` output so that you don't errantly label the coefficients. Here for example, the `tidy()` output indictaes that the intercept coefficient is 11,321 and the education coefficient is 2,651. This corresponds to what we see in the plot.

We can also give the `dw_plot()` function tidy objects from multiple models. To do so, we (1) create each tidy model object, (2) bind the tidy model objects into a single object, and (3) use this object ib the `dw_plot()` function. Below we do not display the intercept as it is not of substantive interest.

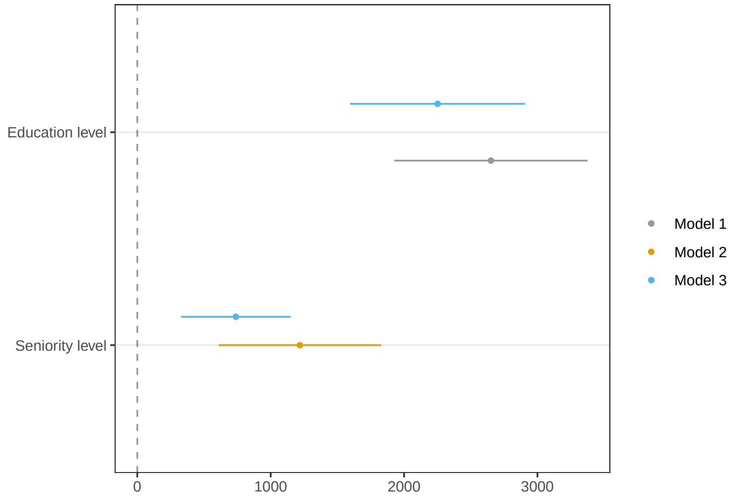

Figure 6: Coefficient plot for three models regressing income on education and seniority. Uncertainty based on the 95% confidence intervals are displayed.

References

Lewis-Beck, C., & Lewis-Beck, M. (2016). *Applied regression: An introduction* (2nd ed.). Thousand Oaks, CA: Sage.