

Introduction to Multiple Regression

2017-01-29

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, *Do differences in education level explain variation in incomes?* and *Do differences in education level explain variation in incomes even after accounting for differences in seniority?*

Preparation

```
# Read in data
city = read.csv(file = "~/Google Drive/andy/epsy-8251/data/riverside.csv")
head(city)
```

	education	income	seniority	gender	male	party
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

```
# Load libraries
library(dplyr)
library(ggplot2)
library(sm)
```

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)
summary(lm.1)
```

Call:

```
lm(formula = income ~ 1 + education, data = city)
```

Residuals:

Min	1Q	Median	3Q	Max
-15808	-5783	2088	5127	18379

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	11321.4	6123.2	1.849	0.0743 .
education	2651.3	369.6	7.173	0.0000000556 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8978 on 30 degrees of freedom

Multiple R-squared: 0.6317, Adjusted R-squared: 0.6194

F-statistic: 51.45 on 1 and 30 DF, p-value: 0.00000005562

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 (in years)")
```

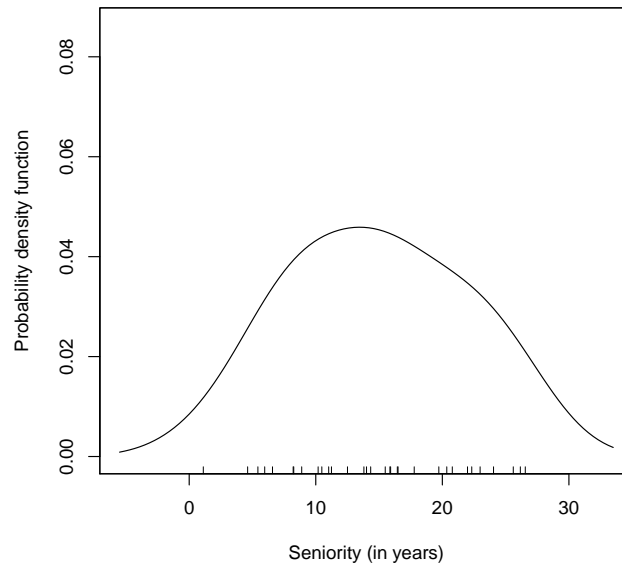


Figure 1. Density plot of the marginal distribution of seniority.

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

```
      M      SD
1 14.8125 6.948834
```

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 marginal 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)")
```

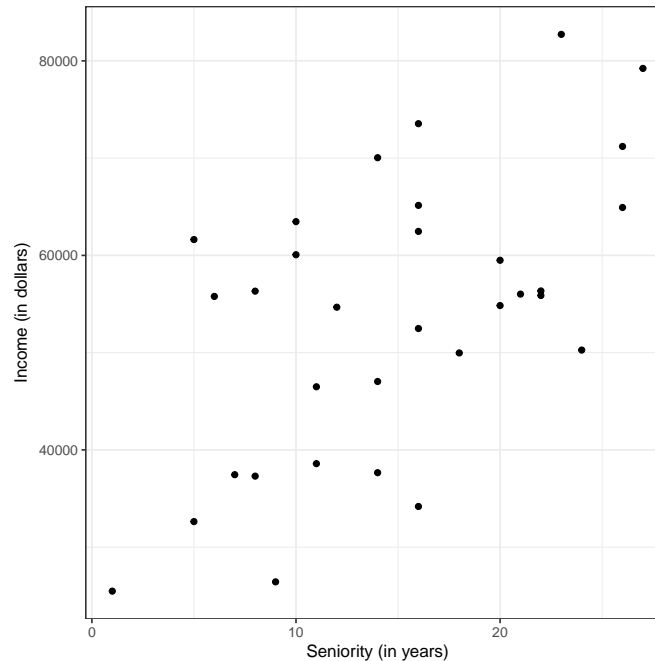


Figure 2. Scatterplot showing the relationship between seniority level and income.

```
# Correlation matrix
cor(city[, c("income", "education", "seniority")])
```

```
      income education seniority
income  1.0000000 0.7947847 0.5819032
education 0.7947847 1.0000000 0.3394469
seniority 0.5819032 0.3394469 1.0000000
```

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 ($r = 0.33$), 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)
summary(lm.2)
```

Call:

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

Residuals:

Min	1Q	Median	3Q	Max
-21007	-7778	-2659	10695	19845

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	35690	5074	7.035	0.0000000807 ***
seniority	1219	311	3.919	0.000477 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12030 on 30 degrees of freedom

Multiple R-squared: 0.3386, Adjusted R-squared: 0.3166

F-statistic: 15.36 on 1 and 30 DF, p-value: 0.0004767

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)
summary(lm.3)
```

Call:

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

Residuals:

Min	1Q	Median	3Q	Max
-12340	-6119	-1396	7002	13928

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	6769.2	5372.9	1.260	0.21776
education	2251.8	334.6	6.729	0.00000022 ***
seniority	738.8	210.1	3.516	0.00146 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7646 on 29 degrees of freedom

Multiple R-squared: 0.7418, Adjusted R-squared: 0.724

F-statistic: 41.65 on 2 and 29 DF, p-value: 0.000000002977

Model-Level Inference

To interpret multiple regression results, begin with the model-level information. Together, differences in education level AND seniority explain 74.2% of the variation in income. This is statistically significant ($p < .001$).

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 these two levels are different. The formal model-level hypothesis that is being tested 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. Equivalently, 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 asking whether the effect of every predictor included in the model is 0. Although these two expressions of the model-level hypothesis look quite different, they are answering the same question, namely whether the model is worthwhile in predicting variation in income.

Interpreting the Coefficients

Now we turn to the coefficient-level information produced in the `summary()` output. 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.
- The partial effect of seniority is 739. This indicates that a one year difference in seniority 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 average 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 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. In plain English, the statistical question being asked is: After accounting for ALL the other predictors included in the model, is there an effect of X on Y ? The mathematical expression of the hypothesis is

$$H_0 : \beta_k = 0.$$

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 asking is there an effect of education level on income after accounting for differences in seniority? 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. (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.)

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, \dots, X_k) and some amount of error. Mathematically it can be written as

$$Y_i = \beta_0 + \beta_1(X1_i) + \beta_2(21_i) + \dots + \beta_k(Xk_i) + \epsilon_i$$

As with simple regression we are interested in estimating the values for each of the regression coefficients, namely, $\beta_0, \beta_1, \beta_2, \dots, \beta_k$. 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).

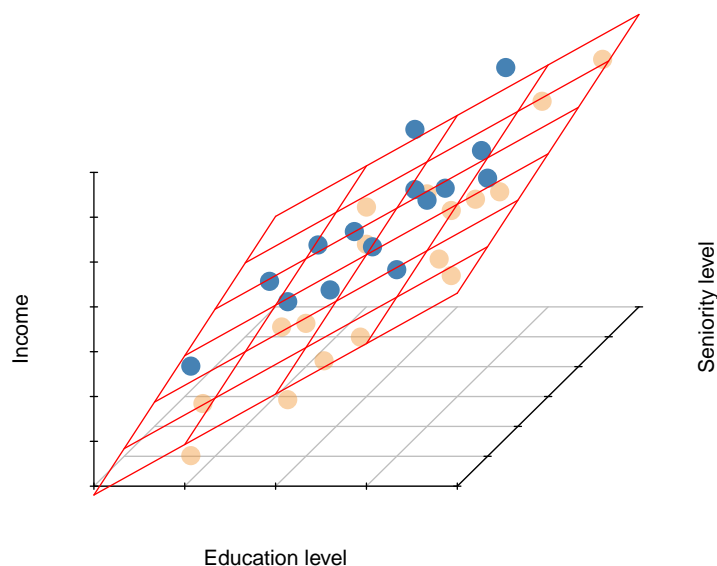


Figure 3. Three-dimensional scatterplot showing the relationship between education level, seniority, and income. The fitted regression plane is also shown. Blue observations have a positive residual and yellow observations have a negative residual.

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

```
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 $SS_{\text{Residuals}} = 1,695,313,285$. 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*** (370)		2,252*** (335)
Seniority		1,219*** (311)	739*** (210)
Intercept	11,321.380* (6,123)	35,690.300*** (5,073)	6,769.172 (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 germane to answer the RQs, Model 2 could just as easily be omitted from the table.)

References

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