Introduction to Multiple Regression

2019-06-28

Introduction and Research Question

In this set of notes, you will continue your foray into regression analysis. To do so, we will go back to the *riverview.csv* data (see the data codebook) and again examine the question of whether education level is related to income. 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(dplyr)
library(ggplot2)
library(readr)
library(tidyr)
library(ungeviz)

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

```
# A tibble: 6 x 6
  education income seniority gender  male party
      <dbl>  <dbl>     <dbl> <chr>  <dbl> <chr>
1         8   37.4         7 male       1 Democrat
2         8   26.4         9 female     0 Independent
3        10   47.0        14 male       1 Democrat
4        10   34.2        16 female     0 Independent
5        10   25.5         1 female     0 Republican
6        12   46.5        11 female     0 Democrat
```

Answering the First Research Question

In previous notes, we fitted a model regressing employees' incomes on education level. We will do that again, and also look at the inferential evidence.

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

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

```
# A tibble: 1 x 11
  r.squared adj.r.squared sigma statistic p.value    df logLik   AIC   BIC
      <dbl>         <dbl> <dbl>     <dbl>   <dbl> <int>  <dbl> <dbl> <dbl>
1     0.632         0.619  8.98      51.5 5.56e-8     2  -115.  235.  240.
# ... with 2 more variables: deviance <dbl>, df.residual <int>
```

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

```
# A tibble: 2 x 5
  term        estimate std.error statistic      p.value
  <chr>          <dbl>     <dbl>     <dbl>        <dbl>
1 (Intercept)    11.3      6.12       1.85 0.0743
2 education       2.65     0.370      7.17 0.0000000556
```

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 thousand dollars, on average. We also found that differences in education level explained 63.2% of the variation in income, and that the empirical evidence is inconsistent with the hypothesis that education level explains none of the variation in incomes ($p < .001$). All this suggests that education level is likely related to income.

Examining the Seniority Predictor

Before we can tackle the second research question, it behooves us to explore the seniority predictor. Below we examine the marginal distribution of seniority for the 32 employees in the sample.

```
# Examine the marginal distribution
ggplot(data = city, aes(x = seniority)) +
  geom_histogram(aes(y = ..density..), fill = "yellow", color = "black") +
  stat_density(geom = "line") +
  theme_bw() +
  xlab("Seniority level (in years)") +
  ylab("Probability density")
```

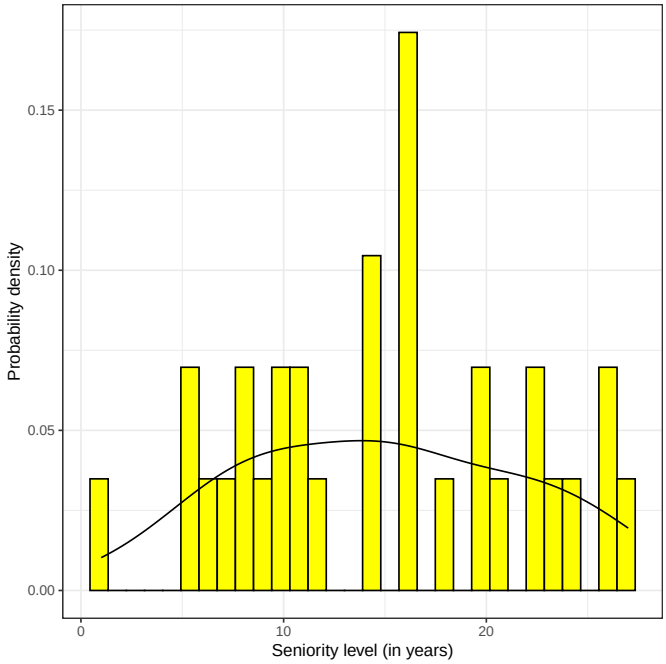

Figure 1. Density plot of the marginal distribution of seniority.

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

Table 1
Mean (M) and Standard Deviation (SD) for the Seniority Predictor

M	SD
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, 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. Since we have already looked at the scatterplot between education-level and income, we focus here on the relationship between seniority and income.

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

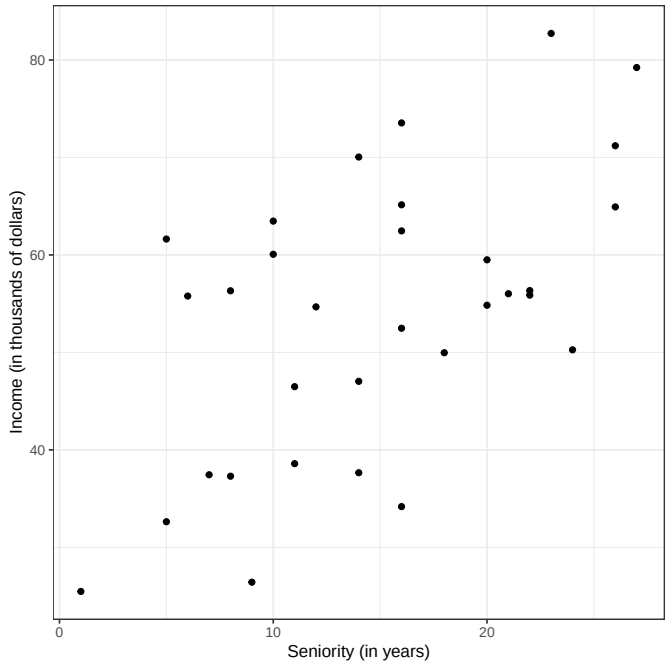

Figure 2. Scatterplot showing the relationship between seniority level and income.

The correlation matrix between all three variables is also examined.

```r
# 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 there is a positive correlation between seniority and income in the sample, it suggests that city employees with more seniority level tend to have higher incomes. The correlation between the two predictors (education level and seniority) is also positive suggesting that city employees with higher education levels tend to have more seniority.

Our research question is focused on examining the relationship between education level and income. Now that we have seen the correlation matrix, we may have some doubts about this relationship. It may be that employees with higher education levels have higher incomes, but that may be because the employees with higher education levels have more seniority which impacts their incomes.

What we need to know in order to determine the effect of education on incomes 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)
```

```
# A tibble: 1 x 11
  r.squared adj.r.squared sigma statistic p.value    df logLik   AIC   BIC
      <dbl>         <dbl> <dbl>     <dbl>   <dbl> <int>  <dbl> <dbl> <dbl>
1     0.339         0.317  12.0      15.4 4.77e-4     2  -124.  254.  258.
# ... with 2 more variables: deviance <dbl>, df.residual <int>
```

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

```
# A tibble: 2 x 5
  term        estimate std.error statistic      p.value
  <chr>          <dbl>     <dbl>     <dbl>        <dbl>
1 (Intercept)    35.7      5.07       7.03 0.0000000807
2 seniority       1.22     0.311      3.92 0.000477
```

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 thousand dollars. We also find that differences in seniority level explain 33.9% of the variation in income, and that this empirical evidence is inconsistent with the hypothesis that seniority explains none of the variation in incomes ($p < .001$).

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)
```

```
# A tibble: 1 x 11
  r.squared adj.r.squared sigma statistic p.value    df logLik   AIC   BIC
      <dbl>         <dbl> <dbl>     <dbl>   <dbl> <int>  <dbl> <dbl> <dbl>
1     0.742         0.724  7.65      41.7 2.98e-9     3  -109.  226.  232.
# ... with 2 more variables: deviance <dbl>, df.residual <int>
```

> 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, $F(3, 29) = 41.65$, $p < .001$, suggest that the empirical evidence is inconsistent with the null hypothesis; it is likely that together education level and seniority level do 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)
```

```
# A tibble: 3 x 5
  term        estimate std.error statistic     p.value
  <chr>          <dbl>     <dbl>     <dbl>       <dbl>
1 (Intercept)    6.77      5.37       1.26 0.218
2 education      2.25      0.335      6.73 0.000000220
3 seniority      0.739     0.210      3.52 0.00146
```

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 thousand dollar difference in income (on average), after accounting for differences in seniority level.
- The **partial effect of seniority** is 0.739. This indicates that a one year difference in seniority level is associated with a 0.739 thousand dollar 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 thousand dollar difference in income (on average), after controlling for differences in seniority.
>
> A one year difference in education level is associated with a 2.252 thousand dollar 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 thousand dollar 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 thousand dollars.

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 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 empirical evidence is inconsistent with this hypothesis, $t(29) = 6.73$, $p < .001$, suggesting that there is likely 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, ..., 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, ..., 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,

$$\hat{Y} = 6.769 + 2.252(X1) + 0.739(X2),$$

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).

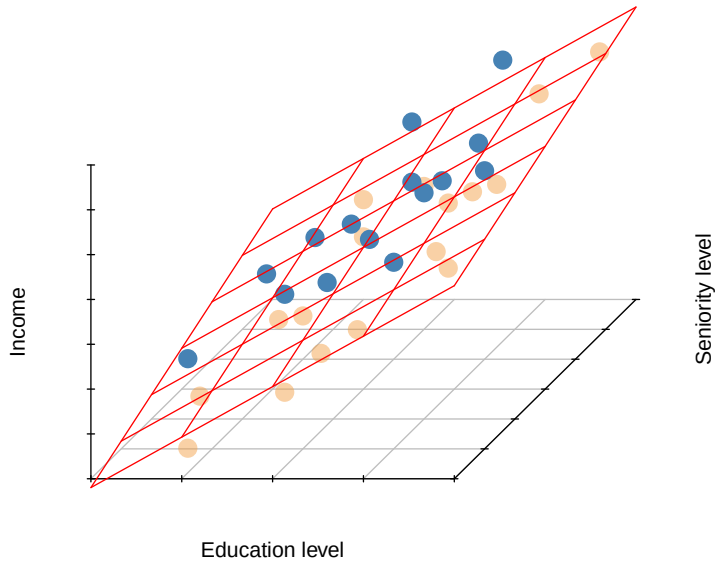

Figure 3. Three-dimensional scatterplot showing the relationship between education level, seniority, and income. The fitted regression plane is also shown. Blue observations have a positive residual and yellow observations have a negative residual.

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 4147.3  4147.3  70.944 0.000000002781 ***
seniority  1  722.9   722.9  12.366        0.00146 **
Residuals 29 1695.3    58.5
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Here the $\text{SS}_{\text{Residuals}} = 1695.3$. 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.)

Note that we **DO NOT INCLUDE stars to indicate "statistical significance"** as is the recommendation of the American Statistical Association. (Wasserstein, Schirm, & Lazar, 2019)

Table 2
Coefficients (and SEs) for the OLS Regression Models Using Education Level and Seniority to Predict Income

	Model 1	Model 2	Model 3
Education level	2.651		2.252
	(0.370)		(0.335)
Seniority level		1.219	0.739
		(0.311)	(0.210)
Constant	11.321	35.690	6.769
	(6.123)	(5.073)	(5.373)
R^2	0.632	0.339	0.742
RMSE	8.978 (df = 30)	12.031 (df = 30)	7.646

Note. All models used the city employee data ($n = 32$). RMSE = Root mean squared error.

Based on the results of fitting the three 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 empirical evidence suggests the answer is yes. Is this true even after accounting for differences in seniority? The empirical evidence from 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

To create a coefficient plot for a multiple regression, we will again use the `stat_confidence_density()` function from the **ungeviz** package. For example, to create the coefficient plot for Model 3 (`lm.3`), we (1) create the `tidy()` model object and filter out the intercept term, then (2) submit that filtered object as the data in `ggplot()`. (To also display the intercept, don't filter the `tidy()` object.)

```
# Create coefficient plot
coef_output = tidy(lm.3) %>%
  filter(term != "(Intercept)")

ggplot(data = coef_output, aes(x = estimate, y = term)) +
  stat_confidence_density(aes(moe = std.error, confidence = 0.68, fill = stat(ndensity)),
                          height = 0.15) +
  geom_point(aes(x = estimate), size = 2) +
  scale_fill_gradient(low = "#eff3ff", high = "#6baed6") +
  theme_bw() +
  theme(
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank()
    ) +
  scale_x_continuous(name = "Estimate", limits = c(-1, 4)) +
  scale_y_discrete(name = "Coefficients", labels = c("Education level", "Seniority"))
```

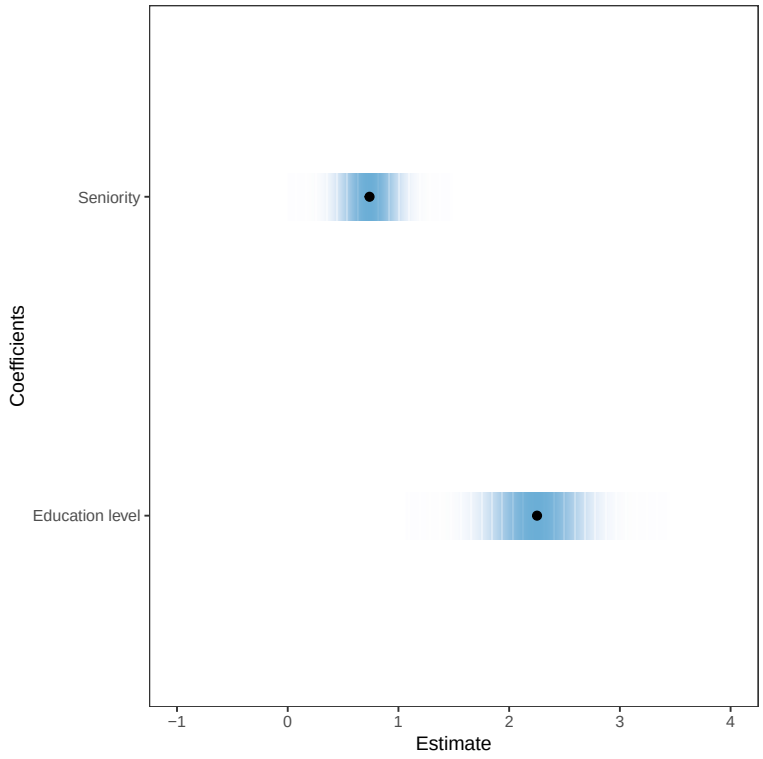

Figure 4. Coefficient plot for the model regressing income on education. Uncertainty based on the 95% confidence intervals are displayed.

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 indicates that the coefficient for education level is 2.25 and the seniority coefficient is 0.739. This corresponds to what we see in the plot.

Although it is typical to create a coefficient plot for only the "final" adopted model, we can create coefficient plots for 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, filter out the intercepts, and drop rows with missing values (not all models have the same subset of predictors) and (3) use this new combined object in the data argument of `ggplot()`. To get this to plot correctly, we also facet the plot on the `model` column.

```
# Create tidy() objects and identify each with a model column
m1 = tidy(lm.1) %>% mutate(model = "Model 1")
m2 = tidy(lm.2) %>% mutate(model = "Model 2")
m3 = tidy(lm.3) %>% mutate(model = "Model 3")

# Combine all three tidy() outputs, filter out intercepts, and drop missing values
all_models = rbind(m1, m2, m3) %>%
  filter(term != "(Intercept)") %>%
  drop_na()

# Create coefficient plots
ggplot(data = all_models, aes(x = estimate, y = term)) +
  stat_confidence_density(aes(moe = std.error, confidence = 0.68, fill = stat(ndensity)),
                          height = 0.15) +
  geom_point(aes(x = estimate), size = 2) +
  scale_fill_gradient(low = "#eff3ff", high = "#6baed6") +
  theme_bw() +
  theme(
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank()
    ) +
  scale_x_continuous(name = "Estimate", limits = c(-1, 4)) +
  scale_y_discrete(name = "Coefficients", labels = c("Education level", "Seniority")) +
  facet_wrap(~model)
```

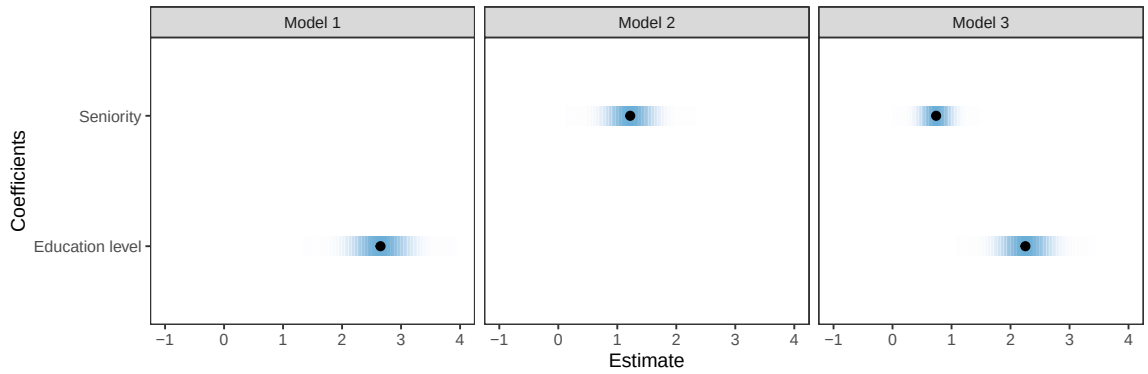

Figure 5. Coefficient plot for three models regressing income on education and seniority. Uncertainty based on the 95% confidence intervals are displayed. (The intercept is not displayed.)

References

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a world beyond "$p < 0.05$". *The American Statistician*, *73*(sup1), 1–19. doi:10.1080/00031305.2019.1583913