

Multiple Regression

Andrew Zieffler



This work is licensed under a
[Creative Commons Attribution
4.0 International License](https://creativecommons.org/licenses/by/4.0/).

Prepare

```
# Load the data (homework-achievement.csv)
> city = read.csv("~/epsy-8251/riverside_final.csv")

# Load libraries; Note: you may need to install them first
> library(sm)
> library(ggplot2)
```

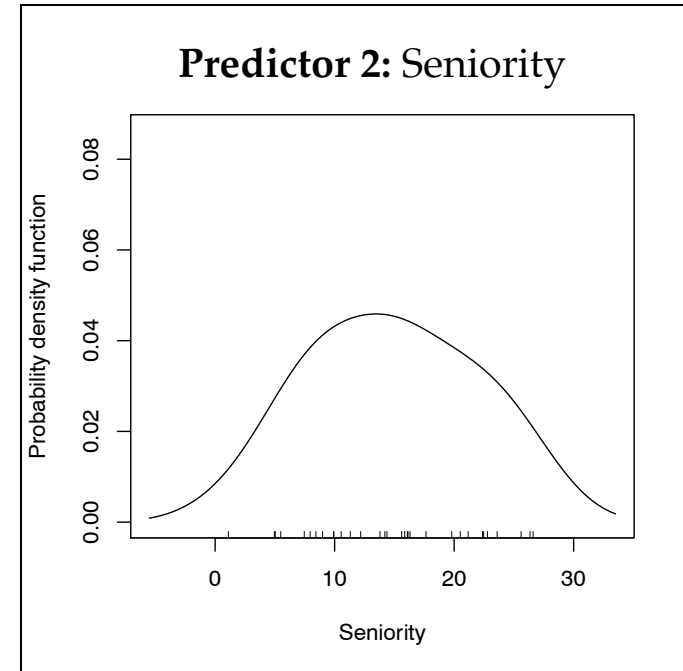
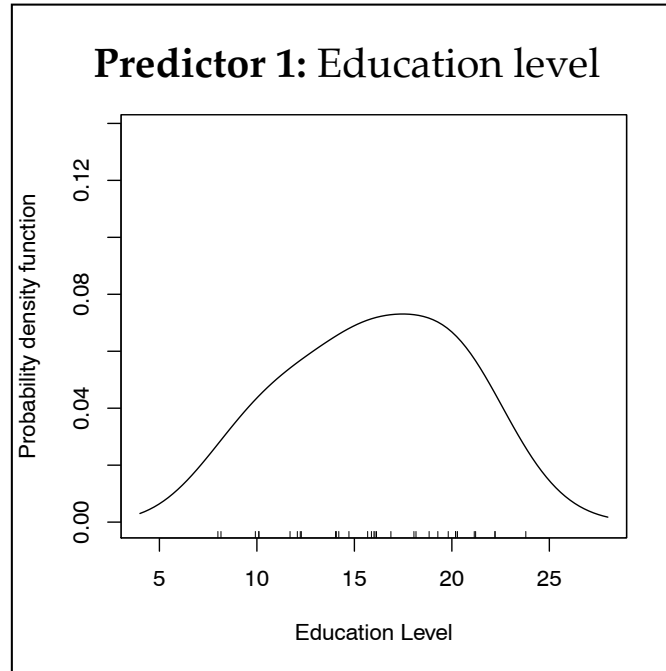
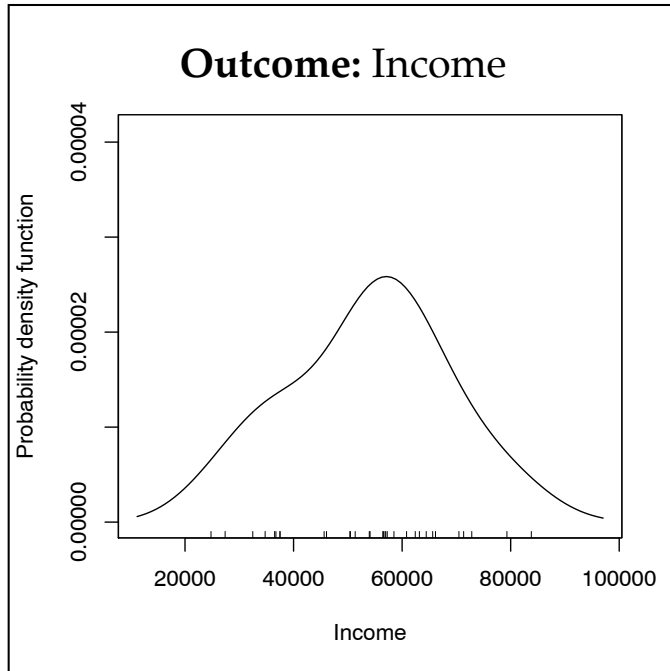
Back to the Drawing Board

**Do differences in education
level explain variation in
incomes?**

**After accounting for
differences in seniority, do
differences in education level
explain variation in incomes?**

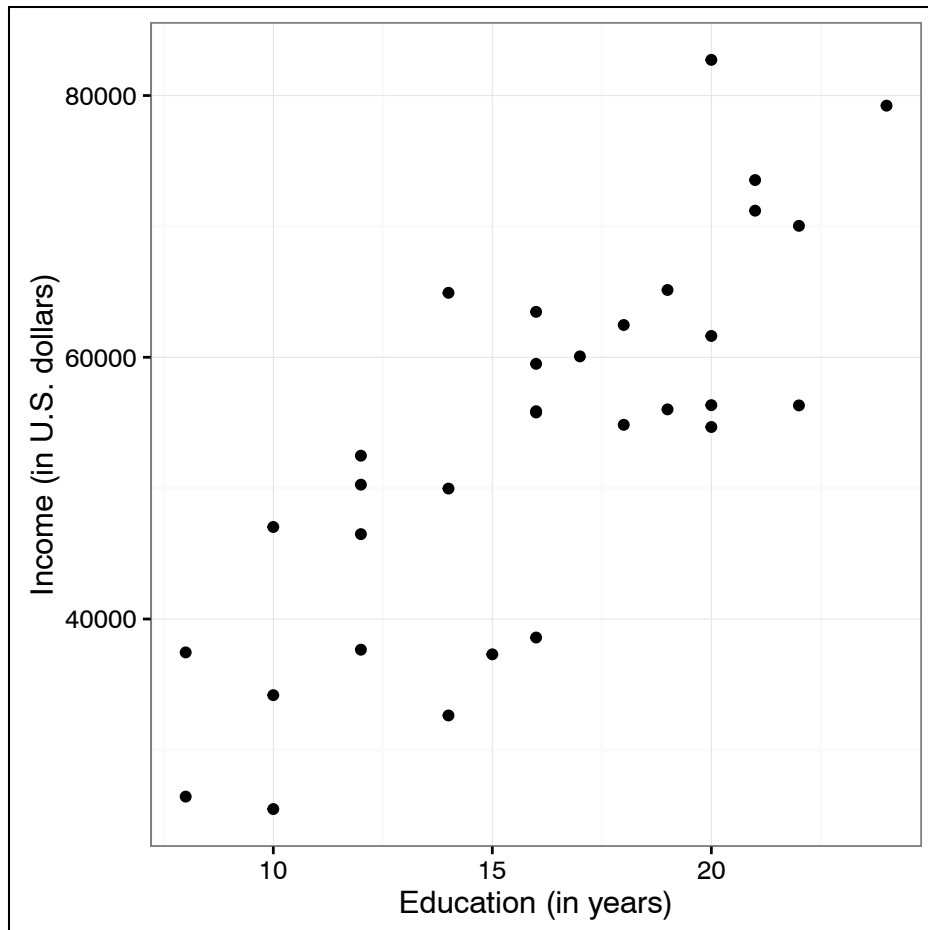


Examine the Outcome and Predictors

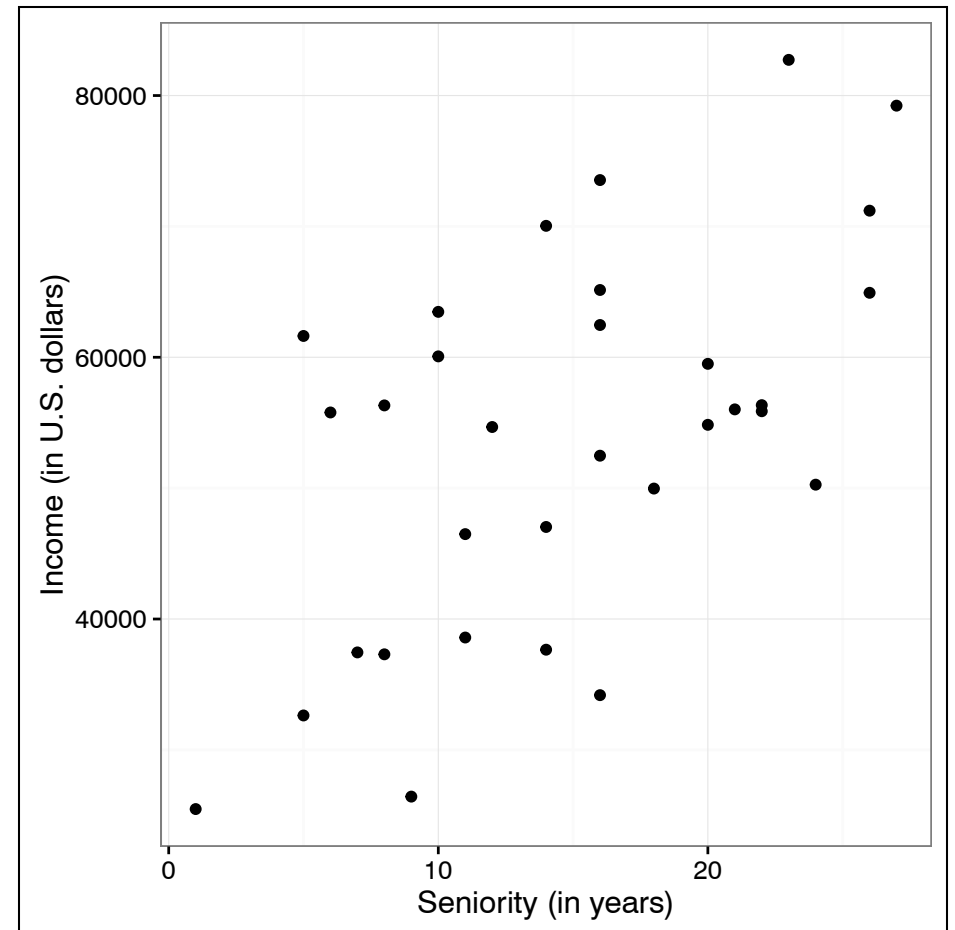


Variable	<i>M</i>	<i>SD</i>
Income	\$53,742	\$14,553
Education level	16	4.4
Seniority	14.8	6.9

Examining the Conditional Distributions of Income



The plot suggests a positive, strong, linear relationship between education level and income. There do not look to be any outlying observations in the plot.



The plot suggests a positive, weak-to-moderate, linear relationship between seniority and income. There do not look to be any outlying observations in the plot.

Correlation Matrix

```
> cor( city[ , c("income", "edu", "senior")] )
```

	income	edu	senior
income	1.0000000	0.7947847	0.5819032
edu	0.7947847	1.0000000	0.3394469
senior	0.5819032	0.3394469	1.0000000

The correlations between income and each of the predictors are both positive, and income's relationship with educational level is slightly stronger than that with seniority. We also note that there is a positive relationship between seniority and education level.

Fit the Simple Regression Model: Education Level

```
> lm.1 = lm(income ~ 1 + edu, data = city)
> summary(lm.1)
```

Coefficients:

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

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

$$\hat{\text{Income}} = 11321 + 2651(\text{Education Level})$$

Differences in education level explains 63.2% of the variation in income. This is statistically reliable, $F(1, 30) = 51.45, p < .001$.

Fit the Simple Regression Model: Seniority

```
> lm.2 = lm(income ~ 1 + senior, data = city)
> summary(lm.2)
```

Coefficients:

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

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

$$\hat{\text{Income}} = 35690 + 1219(\text{Seniority})$$

Differences in seniority explains 33.9% of the variation in income. This is statistically reliable,
 $F(1, 30) = 15.36, p < .001$.

What Do We Know?

At this point we can answer our first RQ: *Do differences in education level explain variation in incomes?* Yes. Education level predicts a variation in incomes (63.2%), and this was statistically reliable.

But: Seniority also predicts variation in incomes (33.9%) and education level and seniority are also related. So maybe the relationship we are seeing between education level and income is due to seniority (e.g., employees with more education tend to have higher incomes, but they are also the same employees who tend to have more seniority).

Is education level still an important predictor of income variation after we account for seniority? To answer this we need to include both predictors in the model simultaneously.

Fit the Multiple Regression Model

```
> lm.3 = lm(income ~ 1 + edu + senior, data = city)
> summary(lm.3)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	6769.2	5372.9	1.260	0.21776
edu	2251.8	334.6	6.729	0.00000022 ***
senior	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

Note that the p -value associated with the model-level test is different from all of the parameter-level tests.

Model-Level Inference

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

The model explains 74.2% of the variation in incomes (in the sample).

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

Given the small p -value, we would reject this hypothesis; the regression is statistically reliable, $F(2, 97) = 8.70, p < 0.001$. It is likely that the model explains variation in income, in the population.

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_p = 0$$

Given the small p -value, we would reject this hypothesis; the regression is statistically reliable, $F(2, 97) = 8.70, p < 0.001$. This suggests it is likely that *at least* one of the regression parameters is not 0.

Regression Coefficients

Coefficients:

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

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

$$\hat{\text{Income}} = 6769 + 2252(\text{Education Level}) + 739(\text{Seniority})$$

The slopes (of which there are now more than one) are referred to as *partial regression slopes* or *partial effects*.

Interpretation of the Intercept

$$\hat{\text{Income}} = 6769 + 2252(\text{Education Level}) + 739(\text{Seniority})$$

The intercept, \$6769, is the average (predicted) income for all employees whose education level is 0 years **and** who have 0 years of seniority.

Danger: This is prediction falls outside the range of the data we used to fit the model; our lowest education level in the data was 8 and our lowest seniority value was 1.

Interpretation of the Partial Effects

$$\hat{\text{Income}} = 6769 + 2252(\text{Education Level}) + 739(\text{Seniority})$$

Each one-year difference in education level is associated with a \$2252 difference in income, on average, ...**controlling for differences in seniority.**

Each one-year difference in seniority is associated with a \$739 difference in income, on average, ...**controlling for differences education level.**

Parameter-Level Inference

Coefficients:

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

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

The test for the intercept, $t(29) = 1.26$, $p = 0.218$, is not statistically reliable. This suggests that the population intercept is not different than 0.

The test for the partial effect of educational level, $t(29) = 6.73$, $p < 0.001$, is statistically reliable. This suggests that differences in education level explain variation in incomes, **after accounting for the variation explained by differences seniority**, in the population.

The test for the partial effect of seniority, $t(29) = 3.52$, $p = 0.001$, is statistically reliable. This suggests that in the population, differences in seniority explain variation in incomes, **after accounting for the variation explained by differences in education level**.

$$H_0 : \beta_0 = 0$$


$$H_0 : \beta_1 = 0$$

$$H_0 : \beta_2 = 0$$

If a partial effect is statistically reliable it means that predictor is statistically important in explaining variation in the outcome **above and beyond what the other predictors in the model explain**.

Multiple Regression Model

The multiple regression model says that a case's outcome (Y) is a function of two or more predictors (X_1, X_2, \dots, X_k) and some amount of error.

$$Y = \boxed{\beta_0 + \beta_1(X_1) + \beta_2(X_2) + \dots + \beta_k(X_k)} + \epsilon$$


$$\hat{Y} = \beta_0 + \beta_1(X_1) + \beta_2(X_2) + \dots + \beta_k(X_k)$$

We estimate the regression coefficients using the sample data to get the observed regression equation,

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1(X_1) + \hat{\beta}_2(X_2) + \dots + \hat{\beta}_k(X_k)$$

Which Predictor has a Stronger Effect on GPA?

$$\hat{\text{Income}} = 6769 + 2252(\text{Education Level}) + 739(\text{Seniority})$$

Based on the values for the estimated regression coefficients, you might suggest that time spent on education level has the bigger effect on income...

...This is true, but generally we **cannot** make that judgment by looking at the size of the unscaled regression coefficients. Often, different predictors are measured using different scales and remember that the magnitude of a regression coefficient is influenced by the unit of measurement.

To compare the relative influence of the predictors in a model, we typically examine the **standardized regression coefficients**, or the **beta weights**.

This time, we will use the `scale()` function to create the z-scores for each variable. Also, rather than include the z-scores as additional predictors in the data frame, we will create them directly in the model.

```
> lm.5 = lm( scale(income) ~ 1 + scale(edu) + scale(senior), data = city)
> summary(lm.5)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.00000000000000001396	0.0928746160831607770	0.000	1.00000
scale(edu)	0.6750404701087127091	0.1003170063937006140	6.729	0.00000022 ***
scale(senior)	0.3527627996989472492	0.1003170063937006418	3.516	0.00146 **

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

Residual standard error: 0.5254 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

$$\hat{z}_{\text{Income}} = 0 + 0.675(z_{\text{Education Level}}) + .353(z_{\text{Seniority}})$$

$$\hat{z}_{\text{Income}} = 0 + 0.675(z_{\text{Education Level}}) + .353(z_{\text{Seniority}})$$

The standardized income for employees with an average education level ($z = 0$) and an average level of seniority ($z = 0$) is 0 (average), on average.

Each **one-standard deviation difference in education level** is associated with a 0.675-standard deviation difference in income, on average, controlling for differences in seniority.

Each **one-standard deviation difference in seniority level** is associated with a 0.353-standard deviation difference in income, on average, controlling for differences in education level.

Standardizing puts all the predictors on the same scale; they can be compared. Education level has **more** influence on income than seniority. A good question might be whether this difference in the effects is statistically reliable...we will not worry about that here, but you can actually test that if it is of interest.

Using the Unstandardized vs. the Standardized Coefficients

Both types of coefficients can be useful to applied researchers, but perhaps for different parts of the interpretational process.

Rules of Thumb for When to Interpret b vs. β -Weights

Interpret b (unstandardized coefficients)

When variables are measured in a meaningful metric

To develop intervention or policy implications

To compare effects *across different* studies or samples

Interpret β -weights (standardized coefficients)

When the variables are not measured in a meaningful metric

To compare the relative effects of predictors *in the same* sample

Policy Decisions/Interventions

Advice for the school board is probably more interpretable if you use the unstandardized coefficients. e.g., What are the effects on GPA if we increase the amount of homework by 5 hours a week? (Note: This assumes the metric for the variables is meaningful...)

Comparing Across Studies

In different studies the variable you want to compare will often have a different distribution. For example, it is likely that the mean and SD for time spent on homework will be different across the studies (even if it is measured in hours/week in all the studies you are comparing). These differences in the distribution affect the magnitude of the β -Weights, not the b 's. Because of this it is better to interpret the unstandardized coefficients if your goal is to compare effects across studies.