Simple Regression: Inference

2017-01-22

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).

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(ggplot2)
library(sm)

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

```

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

Coefficients:
(Intercept)    education
      11321         2651
```

Answering the Research Question

In previous notes, we fitted a model regressing employees' incomes on education level. 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. All this suggests that education level is related to income. . . at least for the $n = 32$ employees in the sample.

Statistical Inference

What if we want to understand the relationship between education level and income for ALL city employees? The problem is that if we had drawn a different sample of $n = 32$ employees, all the regression estimates ($\hat{\beta}_0, \hat{\beta}_1$, and R^2) would be different than the ones we obtained from our sample. This makes it difficult to say, for example, how does the conditional mean income differ for employees with differing education levels. In our observed sample, $\hat{\beta}_1$ was \$2,651. But, had we sampled different employees, we might have found that $\hat{\beta}_1$ was \$1,500. And a different random sample of employees we might have produced a $\hat{\beta}_1$ of \$3,000.

This variation in the estimates arises because of the random nature of the sampling. One of the key findings in statistical theory is that the amount of variation in estimates under random sampling is completely predictable (this variation is called *sampling error*). Being able to quantify the sampling error allows us to provide a more informative answer to the research question. For example, it turns out that based on the quantification of sampling error in our example, we believe that the actual β_1 is between \$1,895 and \$3,406.

Statistical inference allows us to learn from incomplete or imperfect data Gelman & Hill (2007). In many studies, the primary interest is to learn about one or more characteristics about a population. These characteristics must be estimated from sample data. This is the situation in our example, where we have only a sample of employees and we want to understand the relationship between education level and income for ALL employees.

In the example, the variation in estimates arises because of sampling variation. It is also possible to have variation because of imperfect measurement. This is called *measurement error*. Despite these being very different sources of variation, in practice they are often combined (e.g., we measure imperfectly and we want to make generalizations). Regardless of the sources of variation, the goals in most regression analyses are two-fold:

1. Estimate the parameters from the observed data; and
2. Summarize the amount of uncertainty (e.g., quantify the sampling error) in those estimates.

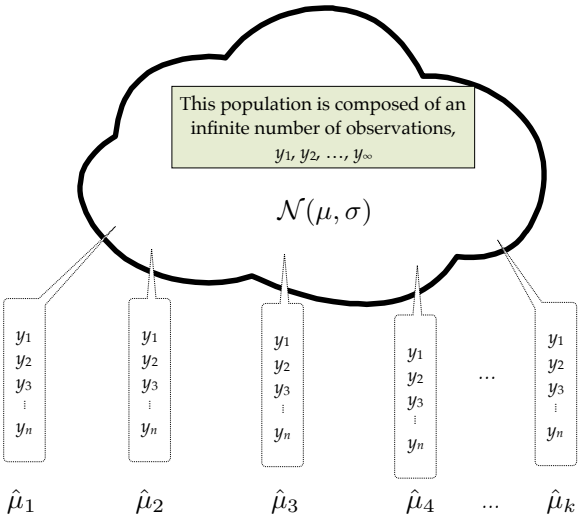
The first goal we addressed in the notes on description. It is the second goal that we will explore in these notes.

Quantification of Uncertainty

Before we talk about estimating uncertainty in regression, let me bring you back in time to your Stat I course. In that course, you probably spent a lot of time talking about sampling variation for the mean. The idea went something like this: Imagine you have a population that is infinitely large. The observations in this population follow some probability distribution. (This distribution is typically unknown in practice, but for now, let's pretend we know what that distribution is.) For our purposes, let's assume the population is normally distributed with a mean of μ and a standard deviation of σ.

Sample n observations from that population. Based on the n sampled observations, find the mean. We will call this $\hat{\mu}_1$ since it is an estimate for the population mean (the subscript just says it is the first sample). In all likelihood, $\hat{\mu}_1$ is not the exact same value as μ. It varies from the population mean because of sampling error.

Now, sample another n observations from the population. Again, find the mean. We will call this estimate $\hat{\mu}_2$. Again, it probably varies from μ, and may be different than $\hat{\mu}_1$ as well. Continue to repeat this process: randomly sample n observations from the population; and find the mean.



The distribution of the sample means, it turns out, is quite predictable using statistical theory. Theory predicts that the distribution of the sample means will be normally distributed. It also predicts that the mean, or *expected value*, of all the sample means will be equal to the population mean, μ.[1] Finally, theory predicts that the standard deviation of this distribution, called the *standard error*, will be equal to the population standard deviation divided by the square root of the sample size. Mathematically, we would write all this as,

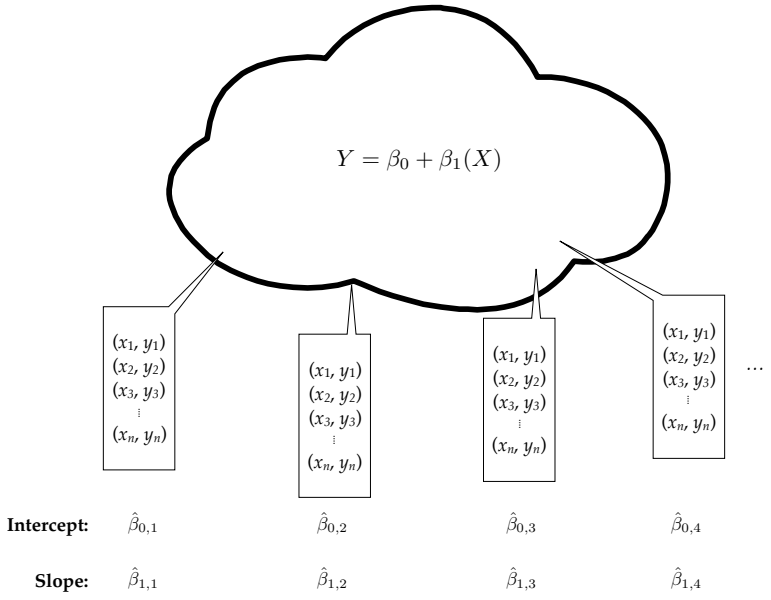
$$\hat{\mu}_n \sim \mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right).$$

The important thing isn't that you memorize this result, but that you understand that the process of randomly sampling from a known population can lead to predictable results in the distribution of statistical summaries (e.g., the distribution of sample means). The other crucial thing is that there the sampling variation can be quantified. The standard error is the quantification of that sampling error. In this case, it gives a numerical answer to the question of how variable the sample mean will be because of random sampling.

Quantification of Uncertainty in Regression

We can extend these ideas to regression. Now the thought experiment goes something like this: Imagine you have a population that is infinitely large. The observations in this population have two attributes, call them X and Y. The relationship between these two attributes can be expressed via a regression equation as: $\hat{Y} = \beta_0 + \beta_1(X)$. Randomly sample n observations from the population. This time, rather than computing a mean, regress the sample Y values on the sample X values. Since the sample regression coefficients are estimates of the population parameters, we will write this as: $\hat{Y} = \hat{\beta}_{0,1} + \hat{\beta}_{1,1}(X)$. Repeat the process. This time the regression equation is: $\hat{Y} = \hat{\beta}_{0,2} + \hat{\beta}_{1,2}(X)$. Continue this process an infinite number of times.

[1] Mathematically, we would write $E(\hat{\mu}) = \mu$.



Statistical theory again predicts the characteristics of the two distributions, that of $\hat{\beta}_0$ and that of $\hat{\beta}_1$. The distribution of $\hat{\beta}_0$ can be expressed as,

$$\hat{\beta}_0 \sim \mathcal{N}\left(\beta_0,\ \sigma_\epsilon\sqrt{\frac{1}{n} + \frac{\mu_X^2}{\sum(X_i - \mu_X)^2}}\right).$$

Similarly, the distribution of $\hat{\beta}_1$ can be expressed as,

$$\hat{\beta}_1 \sim \mathcal{N}\left(\beta_1,\ \frac{\sigma_\epsilon}{\sigma_x\sqrt{n-1}}\right).$$

Again, don't panic over the formulae. What is important is that theory allows us to quantify the variation in both $\hat{\beta}_0$ and $\hat{\beta}_1$ that is due to sampling error. In practice, our statistical software will give us the numerical estimates of the two standard errors.

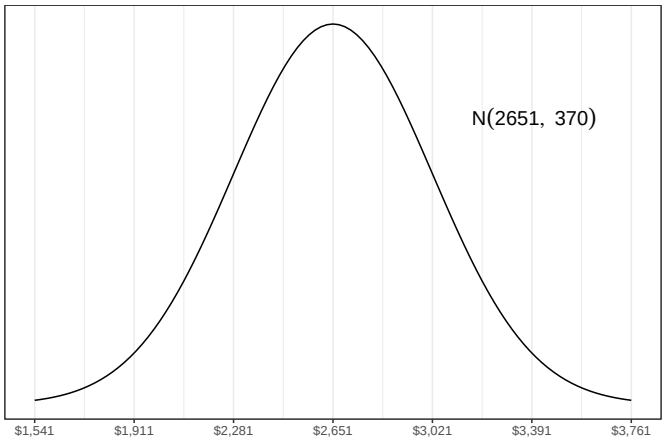
Obtaining SEs for the Regression Coefficients

To obtain the standard errors for the regression coefficients, we will fit the regression using the `lm()` function and save the output into an object, as we did previously. Now, however, we will use the `summary()` function to display the fitted regression output.

```r
# Display the output
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
```

In the *Coefficients* section of the displayed output, we now obtain the estimates for the standard errors in addition to the coefficient estimates. We can use these values to quantify the amount of uncertainty due to sampling error. For example, the estimate for the slope, \$2,651, has a standard error of \$370. One way to envision this is as a distribution. Our best guess (mean) for the slope parameter is \$2,651. The standard deviation of this distribution is \$370.



In the social sciences, it is typical to express uncertainty as $\pm 2(SE)$. Here we would say that becuase of sampling variation, the slope is likely between \$1,911 and \$3,391. Interpreting this, we might say that a one-year difference in education is associated with a difference in income between \$1,911 and \$3,391, on average, for all city employees. Similarly, we could express the uncertainty in the intercept as,

$$11,321 \pm 2(6,123) = [-925,\ 23,567]$$

Interpreting this, we might say that the average income for all city employees with zero years of education is between $-$\$925 and \$23,567.

We can use the `confint()` function to obtain these limits. We just provide the fitted regression object as the input to this function.[2]

```
confint(lm.1, level = 0.95)
```

```
                2.5 %    97.5 %
(Intercept) -1183.935 23826.693
education    1896.425  3406.168
```

Hypothesis Testing

Some research questions point to examining whether the value of some regression parameter differs from a specific value. For example, it may be of interest whether a particular population model (e.g., one where $\beta_1 = 0$) could produce the sample result of a particular $\hat{\beta}_1$. To test something like this, we state the value we want to test in a statement called a *hypothesis*. When the value we are testing is zero, the statement is referred to as a *null hypothesis*. For example,

$$H_0 : \beta_1 = 0$$

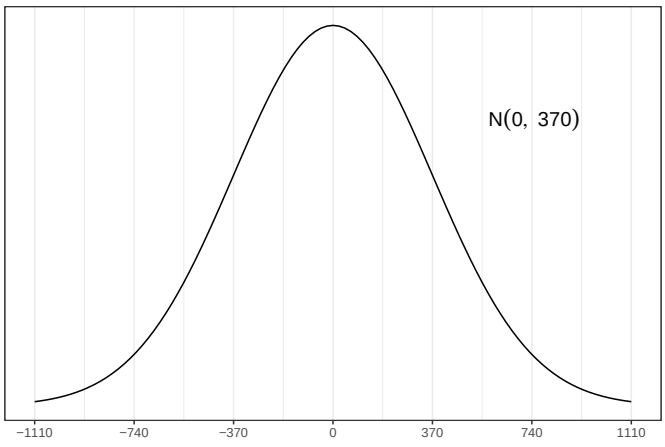
The hypothesis is a statement about the population. Here we hypothesize $\beta_1 = 0$. It would seem logical that one could just examine the estimate of the parameter from the observed sample to answer this question, but we also have to account for sampling uncertainty. The key is to quantify the sampling variation, and then see if the sample result is unlikely given the stated hypothesis.

One question of interest may be: Is there evidence that the average income differs for different education levels? In our example, we have a $\hat{\beta}_1 = 2651$. This is sample evidence, but does \$2,651 differ from 0 more than we would expect because of random sampling? If it doesn't, we cannot really say that the average income differs for different education levels. To test this, we make an assumption that there is no relationship between education level and income, in other words, the slope of the line under this assumption would be 0.

Now, recall that we know some things about the distribution of sample slopes ($\hat{\beta}_1$) under random sampling:

- The distribution is normally distributed.
- The expected value is equal to the population value of β_1, which in the assumed model is 0.
- The standard error is predictable (and is computed by our software)

We already computed the standard error as 370, so our distribution looks like this:



[2]The actual limits from the 'confint()' function are computed using a multiplier that is slightly different than two; thus the discrepancy between our off-the-cuff computation earlier and the result from R. Using a multiplier of two is often close enough for practical purposes, especially when the sample size is large.

Based on the assumed model ($\beta_1 = 0$), a sample value of $\hat{\beta}_1 = 2651$ is quite unlikely. Most sample slopes from this model would be somewhere between -740 and $+740$. Since the sample value of 2651 is actually observed evidence, this would supoort rejecting the model that $\beta_1 = 0$.

Statisticians and researchers take this one step further, and compute how many standard errors the observed value is from the hypothesized value. This is relatively easy algebra:

$$\frac{2651 - 0}{370} = 7.16$$

Interpreting this, we can say that the observed slope of 2,651 is 7.16 standard errors from the expected value of 0.

We can compute the probability of obtaining a sample slope (under random sampling) at least as extreme as the one in the data under the assumed model. This is equivalent to finding the area under the probability curve that is greater than or equal to 2,651.[3] This is called the p-value.

In our example, $p = 0.0000000556$. Obtaining a sample slope of 2,651, or a slope that is more extreme, under the assumption that $\beta_1 = 0$ is 0.0000000556. This is quite unlikely, so it serves as evidence against the hypothesized model. It is likely that $\beta_1 \neq 0$.

Both the distance from the hypothsized value of 0 (in the standard error metric) and the p-value are also provided in the `summary()` output for an `lm` object.

```
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
```

Note that the p-value for the slope is given in scientific notation ... `5.56e-08` is equivalent to 5.56×10^{-8}.

Testing the Intercept

The hypothesis being tested for the intercept is $H_0 : \beta_0 = 0$. The results indicate that the observed intercept of 11,321 is 1.85 standard errors from the hypothesized value of 0. The probability of observing a sample intercept of 11,321, or one that is even further away from 0, is 0.074. This is not overwhelming evidence against the hypothesized model.[4] Because of this, we would not reject the hypotheis; it may be that $\beta_0 = 0$ in the population.

[3] We actually compute the area under the probability curve that is greater than or equal to 2,651 AND that is less than or equal to -2651.

[4] Social science tends to say evidence against a hypothesized model is when the p-value is less than or equal to 0.05.

Model-Level Inference

Sometimes you may want to carry out inference for the model as a whole, rather than for the individual parameters. The statistical question at the model level is: *Does the model explain variation in the outcome?* This can formally be expressed in a statistical hypothesis as,

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

The model-level inferential information is shown at the bottom of the `summary()` output.

```
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
```

Based on the evidence ($p < .001$), we reject the hypothesis that the model does not explain variation in incomes in the population, $F(1,\ 30) = 51.45$. Our best guess for the amount of variation explained by the model is 63.2% (the Multiple R-squared value).

We can also get the model-level inferential information from the `anova()` output. This also gives us the ANOVA decomposition for the model.

```
anova(lm.1)
```

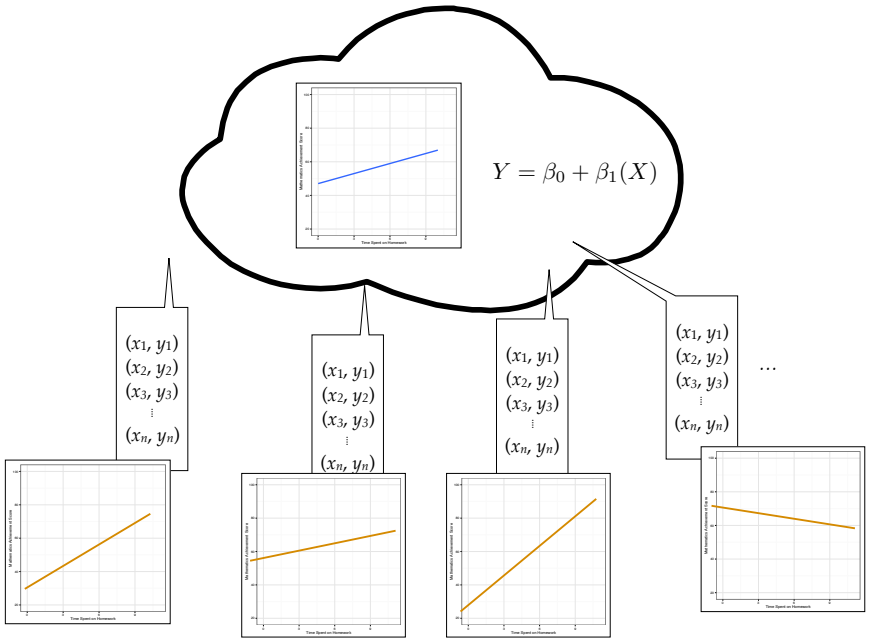
```
Analysis of Variance Table

Response: income
          Df     Sum Sq    Mean Sq F value        Pr(>F)
education  1 4147330492 4147330492  51.452 0.00000005562 ***
Residuals 30 2418196934   80606564
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Note that the two *df* for the model-level F-statistic correspond to the *df* in each row of the ANOVA table. The first *df* (in this case 1) is the model degrees-of-freedom, and the second *df* (in this case 30) is the residual degrees-of-freedom.

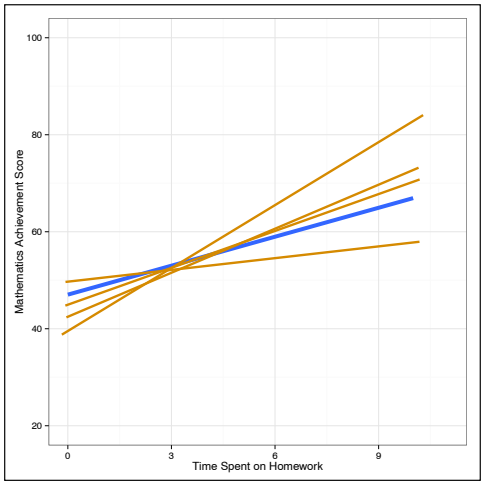
Lastly, we point out that in simple regression models, the results of the model-level inference (i.e., the p-value) is exactly the same as that for the coefficient-level inference for the slope. That is because the model is composed of a single predictor, so asking whether the model accounts for variation in achievement scores **is the same as** asking whether differences in time spent on homework account for variation in achievement scores. *Once we have multiple predictors in the model, the model-level results and predictor-level results will not be the same.*

Confidence Enevelope for the Model

Re-consider our thought experiment. Again, imagine you have a population that is infinitely large. The observations in this population have two attributes, call them X and Y. The relationship between these two attributes can be expressed via a regression equation as: $\hat{Y} = \beta_0 + \beta_1(X)$. Randomlly sampe n observations from the population, and compute the fitted regression equation, this time plotting the line (rather than only paying attention to the numerical estimates of the slope or intercept). Continue sampling from this population, each time drawing the fitted regression equation.



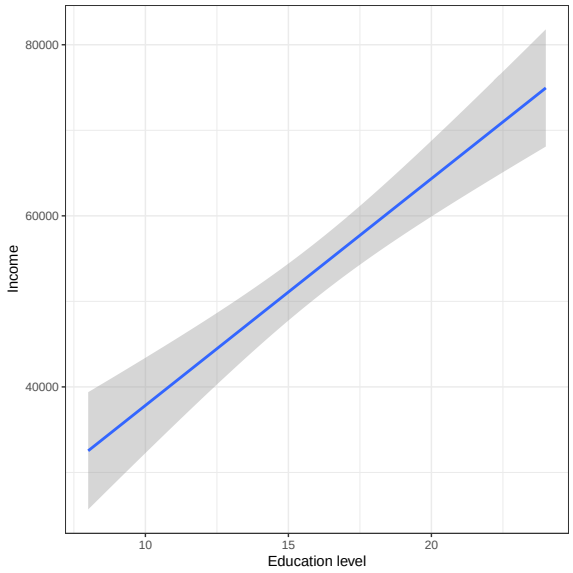
Now, imagine superimposing all of these lines on the same plot.



Examining where the sampled lines fall gives a visual interpretation of the uncertainty in the model. This two-dimensional display of uncertainty is referred to as a confidence envelope. In practice we estimate the uncertainty from the sample data and plot it around the fitted line from the sample.

For simple regression models, we can plot this directly in `ggplot` by including the `geom_smooth()` layer. We will use the arguments `method="lm"` and `se=TRUE`. This will use the method of regression and adds a confidence enevelope.

```
ggplot(data = city, aes(x = education, y = income)) +
  geom_smooth(method = "lm", se = TRUE) +
  xlab("Education level") +
  ylab("Income") +
  theme_bw()
```



References

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. New York: Cambridge University Press.

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