Coefficient-Level Inference

EPsy 8251

2018-07-04

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

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

# Read in data
city = read_csv(file = "~/Dropbox/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
```

```
# 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.

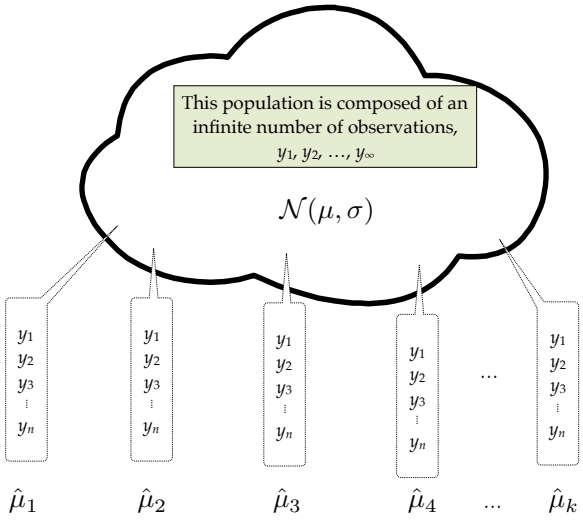

Figure 1: Thought experiment for sampling samples of size n from the population to obtain the sampling distribution of the mean.

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.

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

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.

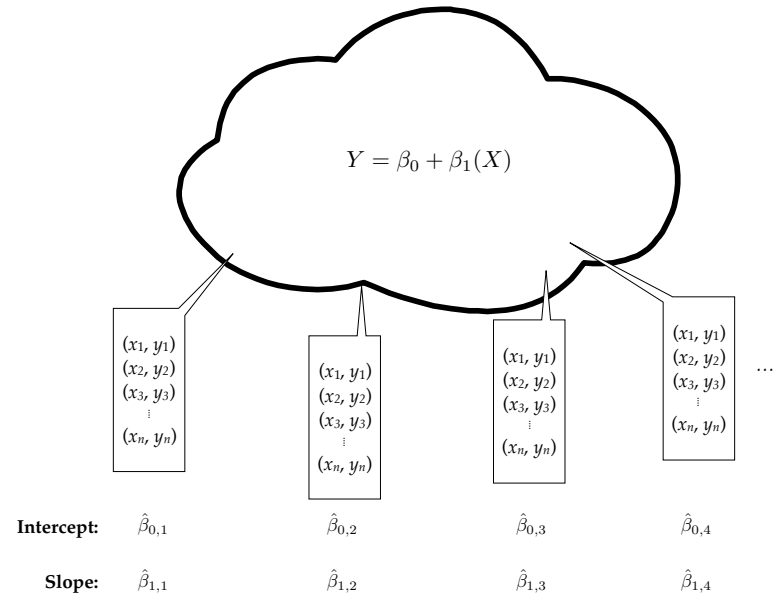

Figure 2: Thought experiment for sampling samples of size n from the population to obtain the sampling distribution of the regression coefficients.

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 `tidy()` function from the **broom** package to display the fitted regression output.

```
# Display the output
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
```

In 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, which indicates the precision (uncertainty) of our guess.

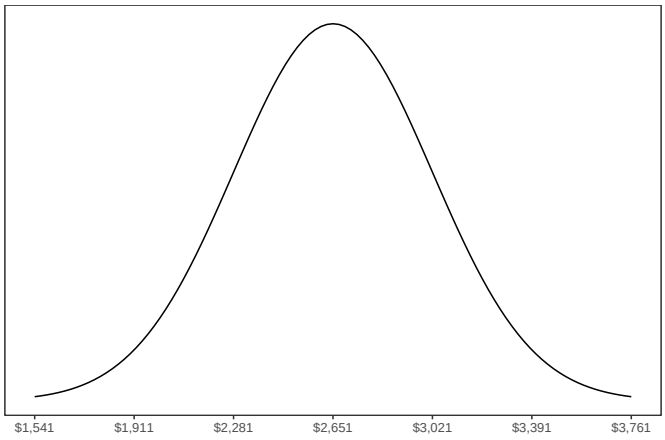

Figure 3: Sampling distribution of the slope coefficient. The distribution is normal with a mean of 2651 and a standard error of 370.

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,

> 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,

> 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. Before we talk about how to test this, we need to introduce one wrinkle into the procedure.

Estimating Variation from Sample Data: No Longer Normal

In theory, the sampling distribution for two regression coefficients were both normally distributed. This is the case when we know the variation parameters in the population. For example, for the sampling distribution of the slope to be normally distributed, we would need to know σ_ϵ and σ_x.

[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.

In practice these values are typically unknown and are estimated from the sample data. Anytime we are estimating things we introduce additional uncertainty. In this case, the uncertainty affects the shape of the sampling distribution.

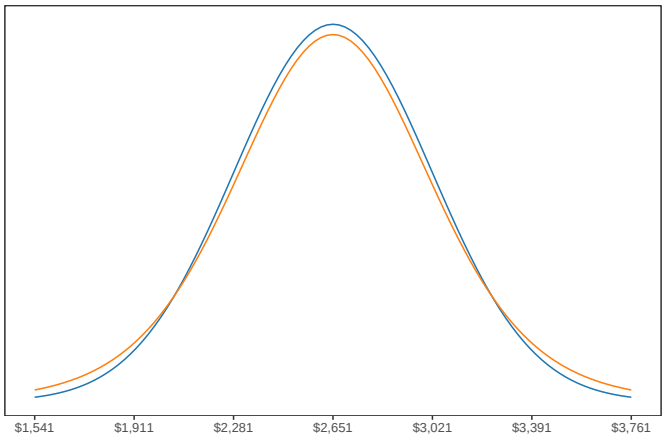

Figure 4: Comparison of two distributions. The normal distribution (blue) and one with additional uncertainty (orange).

Compare the normal distribution (blue) to the distribution with additional uncertainty (orange). From the figure you can see that the additional uncertainty slightly changed the shape of the distribution from normal.

- It is still symmetric and unimodal (like the normal distribution).
- The additional uncertainty makes more extreme values more likely than they are in the normal distribution.
- The additional uncertainty makes values in the middle less likely than they are in the normal distribution.

It is important to note that the amount of uncertainty affects how closely the shape of the distribution matches the normal distribution. And, that the sample size directly affects the amount of uncertainty we have. All things being equal, we have less uncertainty when we have larger samples. The following figure illustrates this idea.

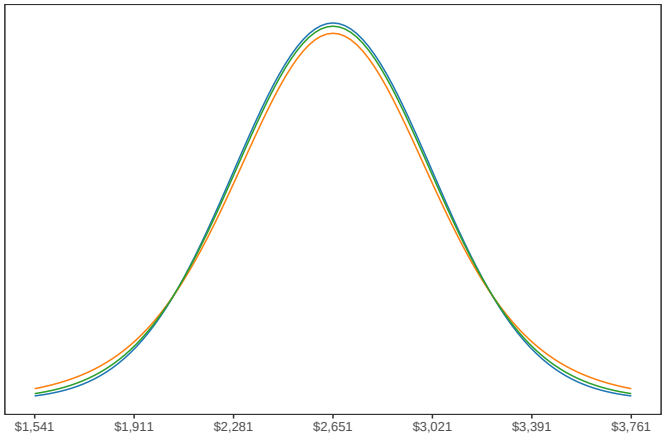

Figure 5: The normal distribution (blue) and two with additional uncertainty; one based on n=10 (orange), and the other based on n=30 (green).

The t-Distribution

As pointed out, the distributions with uncertainty introduced from using a sample of data are not normally distributed. Thus, it doesn't make sense to use a normal distribution as a model for describing the sampling variation. Instead, we will a t-distribution; a family of distributions that have several advantageous properties:

- The are unimodal and symmetric.
- The have more variation (uncertainty) than the normal distribution resulting in a distribution that has thicker tails and is shorter in the middle than a normal distribution.
- How thick the tails are and how short the middle of the distribution is, is related to the sample size.

Specifically, the t-distribution is unimodal and symmetric with a mean of 0. The variance of the distribution (which also specifies the exact shape), is

$$\text{Var} = \frac{df}{df - 2}$$

for $df > 2$ where df is referred to as the *degrees of freedom*.

Back to the Test

Recall that we are interested in testing

$$H_0 : \beta_1 = 0$$

To test this we compute the number of standard errors that our observed slope ($\hat{\beta}_1 = 2651$) is from the hypothesized value of zero (stated in the null hypothesis). Since we already obtained the standard error for the slope, 370, we just use some straight-forward algebra to compute this:

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

Interpreting this, we can say,

> The observed slope of 2,651 is 7.16 standard errors from the expected value of 0.

This value is referred to as the observed t-value. (It is similar to a z-value in the way it is computed; it is standardizing the distance from the observed slope to the hypothesized value of zero. But, since we had to estimate the SE using the data, we introduced additional uncertainty; hence a t-value.)

We can evaluate this t-value within the appropriate t-distribution. For regression coefficients, the t-distribution we will use for evaluation has degrees of freedom that are a function of the sample size and the number of coefficient parameters being estimated in the regression model, namely,

$$df = n - (\text{number of parameters}).$$

In our example the sample size (n) is 32, and the number of coefficient parameters being estimated in the regression model is two (β_0 and β_1). Thus,

$$df = 32 - 2 = 30$$

Based on this, we will evaluate our observed t-value of 7.16 using a t-distribution with 30 degrees of freedom. Using this distribution, we can compute the probability of obtaining a t-value (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 for the t-distribution that is greater than or equal to 7.16.[3] This is called the p-value.

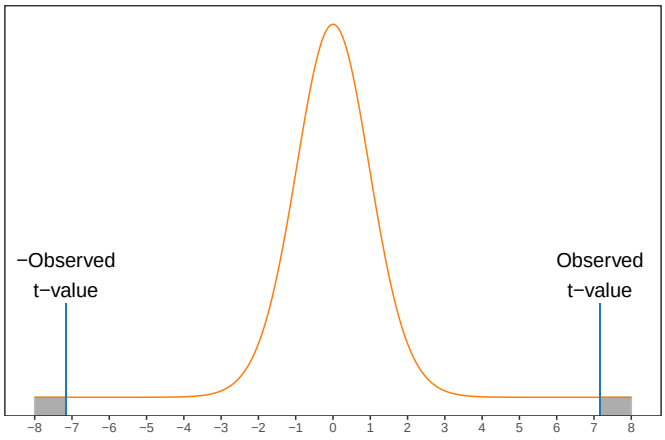

Figure 6: Plot of the probability curve for the t(30) distribution. The shaded area under the curve represents the p-value for a two-tailed test evaluating whether the population slope is zero using an observed t-value of 7.16.

The p-value is computed for us and displayed in the `tidy()` output, along with the t-value (provided in the `statistic` column). In our example, $p = 0.0000000556$. To interpret this we would say,

> The probability of observing a t-value of 7.16, or a t-value that is more extreme, under the assumption that $\beta_1 = 0$ is 0.0000000556.

This is equivalent to saying:

> The probability of observing 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. In other words, it is likely that $\beta_1 \neq 0$.

Note that the p-value might be printed in scientific notation. For example, it may be printed as `5.56e-08`, which is equivalent to 5.56×10^{-8}.

[3] We actually compute the area under the probability curve that is greater than or equal to 7.16 AND that is less than or equal to −7.16.

Testing the Intercept

The hypothesis being tested for the intercept is $H_0 : \beta_0 = 0$. The `tidy()` output also provides information about this test:

```
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 results indicate that the observed intercept of 11,321 is 1.85 standard errors from the hypothesized value of 0;

$$t = \frac{11321}{6123} = 1.85$$

Assuming the null hypothesis that $\beta_0 = 0$ is true, the probability of observing a sample intercept of 11,321, or one that more extreme, is 0.074. This is not overwhelming evidence against the hypothesized model.[4] Because of this, we would not reject the hypothesis that $\beta_0 = 0$; it may be that the intercept in the population is indeed zero.

Confidence Intervals as a Method of Testing

We can also use the confidence intervals (CIs) we computed earlier to test hypotheses.

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

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

This computation suggests that the population slope is likely between 1,896 and 3,406.

The null hypothesis was a test to decide whether β_1, the population slope, was equal to zero. Given the CI for the slope, the value of zero is not a likely candidate value for the slope; 0 is not between 1,896 and 3,406. Using this methodology, we can test any value we want.

We also tested whether the population intercept was equal to zero ($H_0 : \beta_0 = 0$). The CI for the intercept suggests that likely candidates for the population intercept are: $[-1184,\ 23826]$. Since zero is a likely candidate value, we cannot reject the hypothesis that the population intercept may be zero.

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

Coefficient Plot

One plot that gives a great deal of information about the estimates of the regression coefficients and the associated uncertainty is a *coefficient plot*. This plot, recommended by Gelman & Hill (2007), is a graphical representation of the information provided in the `tidy()` output.

```
         term  estimate std.error statistic          p.value lwr_limit
1 (Intercept) 11321.379 6123.2350  1.848921 0.07434601558004 -925.0908
2   education  2651.297  369.6232  7.172972 0.00000005562116 1912.0502
  upr_limit
1 23567.849
2  3390.543
```

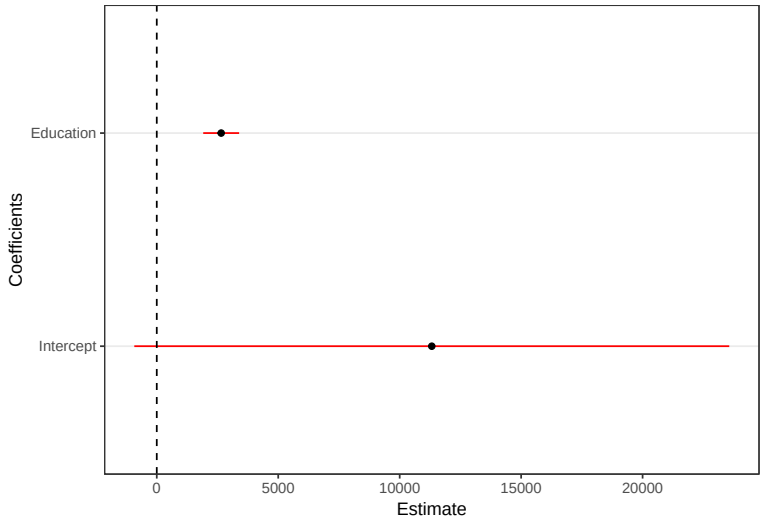

Figure 7: Coefficient plot for the model regressing income on education. Uncertainty based on the 95% confidence intervals are displayed.

The coefficient plot shows the estimates of the regression coefficients (dots) and the uncertainty in those estimates via the confidence intervals (red lines). It also indicates whether zero is a likely candidate value for each of the coefficients.

The pseudocode too create this plot is:

- Mutate new columns for the lower- and upper-limits for the CIs onto the `tidy()` output.
- Plot line segments that correspond to the CIs.
- Plot the estimates for the coefficients as a point.
- Add a line at 0.
- Spruce up the plot.

The syntax I used was:

```
# Add the lower- and upper-limits for the CIs
my_reg = tidy(lm.1) %>%
  mutate(
    lwr_limit = estimate - 2 * std.error,
    upr_limit = estimate + 2 * std.error
    )

# View the output
my_reg

# Create the coefficient plot
ggplot(data = my_reg, aes(x = estimate, y = term)) +
  geom_segment(
    aes(x = lwr_limit, xend = upr_limit, y = term, yend = term),
    color = "red") +
  geom_point() +
  geom_vline(
    xintercept = 0,
    linetype = "dashed"
    ) +
  theme_bw() +
  theme(
    panel.grid.major.x = element_blank(),
    panel.grid.minor.x = element_blank()
  ) +
  scale_y_discrete(name = "Coefficients", labels = c("Intercept", "Education")) +
  xlab("Estimate")
```

There are several variations of this plot. For example, some researchers show the 95% confidence interval and the 50% interval on the plot. For our example,

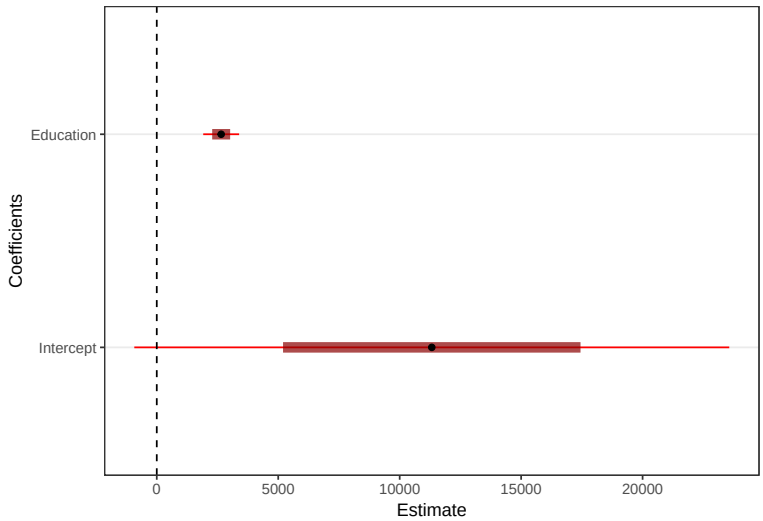

Figure 8: Coefficient plot for the model regressing income on education. Uncertainty based on the 68% and 95% confidence intervals are displayed.

The pseudocode to create this plot is:

- Mutate new columns for the lower- and upper-limits for the 95% CIs onto the `tidy()` output.
- Mutate new columns for the lower- and upper-limits for the 68% CIs onto the `tidy()` output.
- Plot line segments that correspond to the 95% CIs.
- Plot line segments that correspond to the 68% CIs. Make these thicker.
- Plot the estimates for the coefficients as a point.
- Add a line at 0.
- Spruce up the plot.

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.