Model-Level Inference

2019-06-28

Introduction and Research Question

In the last set of notes, we carried out an inferential analysis on the regression coefficients; testing whether the parameters were equal to zero and also computing confidence intervals to estimate the uncertainty in the coefficient estimates. In this set of notes, we will again consider statistical inference, but this time at the model level. To do so, we will again examine the question of whether time spent on homework is related to GPA using the *keith-gpa.csv* data (see the data codebook). To begin, we will load several libraries and import the data into an object called `keith`.

Preparation

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

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

```
# A tibble: 6 x 3
    gpa homework parent_ed
  <dbl>    <dbl>     <dbl>
1    78        2        13
2    79        6        14
3    79        1        13
4    89        5        13
5    82        3        16
6    77        4        13
```

```
# Fit regression model
lm.1 = lm(gpa ~ 1 + homework, data = keith)
```

Sometimes you are interested in the model as a whole, rather than the individual parameters. For example, you may be interested in whether a set of predictors *together* explains variation in the outcome. Recall that the model-level information is displayed using the `glance()` output from the **broom** package.

```
# 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     0.107        0.0981  7.24      11.8 8.85e-4     2  -339.  684.  691.
# ... with 2 more variables: deviance <dbl>, df.residual <int>
```

The `r.squared` column indicates the proportion of variation in the outcome explained by differences in the predictor *in the sample*. Here, differences in time spent on homework explains 10.7% of the variation in students' GPAs for the 100 students in the sample.

Model-Level Inference

The inferential question at the model level is: *Does the model explain variation in the outcome, in the population?* This can formally be expressed in a statistical hypothesis as,

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

To test this, we need to be able to obtain the sampling distribution of R^2 to estimate the uncertainty in the sample estimate. The thought experiment for this 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. There is NO relationship between these two attributes; $\rho^2 = 0$. Randomly sample n observations from the population. Fit the regression and compute the R^2 value. Repeat the process an infinite number of times.

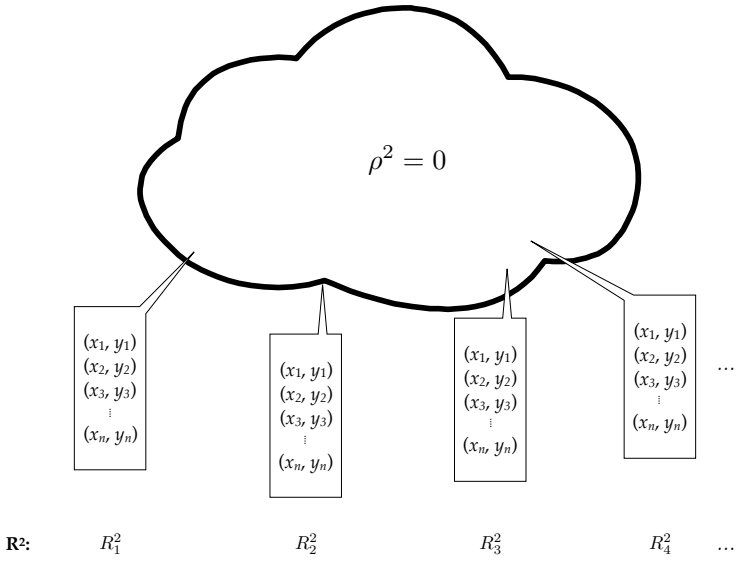

Figure 1. Thought experiment for sampling samples of size n from the population to obtain the sampling distribution of R-squared.

Below is a density plot of the sampling distribution for R^2 based on 1,000 random samples. (Not an infinite number of draws, but large enough that we should have an idea of what the distribution might look like.)

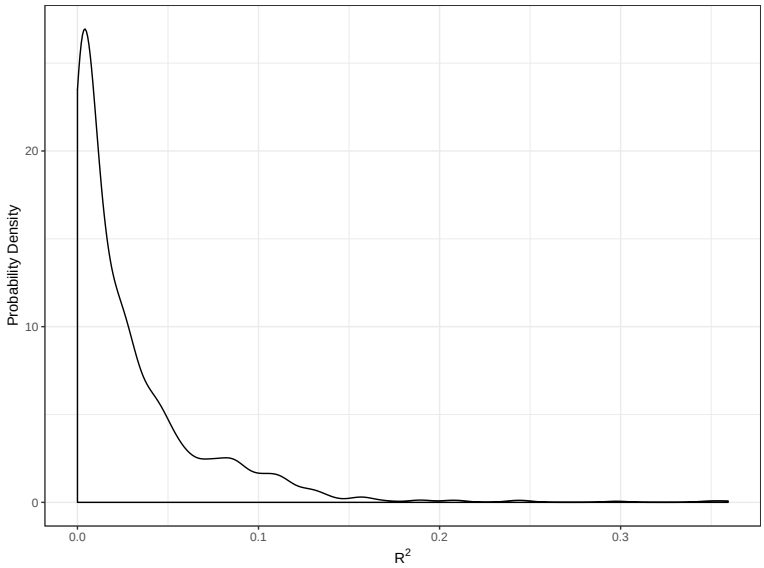

Figure 2. Sampling distribution based on 1000 simple random samples of size 32 drawn from a population where rho-squared = 0.

This sampling distribution is right-skewed. (WHY???) This means that we cannot use a t-distribution to model this distribution—remember the t-distribution is symmetric around zero. It turns out that this sampling distribution is better modeled using an F-distribution.

The F-Distribution

In theoretical statistics the F-distribution is the ratio of two chi-squared statistics,

$$F = \frac{\chi^2_1/df_1}{\chi^2_2/df_2}$$

where df_1 and df_2 are the degrees of freedom associated with each of the chi-squared statistics, respectively. For our purposes, we don't need to pay much attention to this other than to the fact that an F-distribution is defined using TWO parameters: df_1 and df_2. Knowing these two values completely parameterize the F-distribution (they give the shape, expected value, and variation).

In regression analysis, the F-distribution associated with model-level inference is based on the following degrees of freedom:

$$\begin{aligned} df_1 &= p \\ df_2 &= df_{\text{Total}} - p \end{aligned}$$

where p is the number of predictors used in the model and Total is the total degrees of freedom in the data used in the regression model ($\text{Total} = n - 1$). In our example, $df_1 = 1$ and $df_2 = 99 - 1 = 98$. Using these values, we have defined the $F(1, 98)$-distibution.

The F-distribution is the sampling distribution of F-values (not R^2-values). But, it turns out that we can easily convert an R^2-value to an F-value using,

$$F = \frac{R^2}{1 - R^2} \times \frac{df_2}{df_1}$$

In our example,

$$\begin{aligned} F &= \frac{0.107}{1 - 0.107} \times \frac{98}{1} \\ &= 0.1198 \times 98 \\ &= 11.74 \end{aligned}$$

Thus, our observed F-value is: $F(1, 98) = 11.74$. To evaluate this under the null hypothesis, we find the area under the $F(1, 98)$ density curve that corresponds to F-values *at least as extreme* as our observed F-value of 11.74.

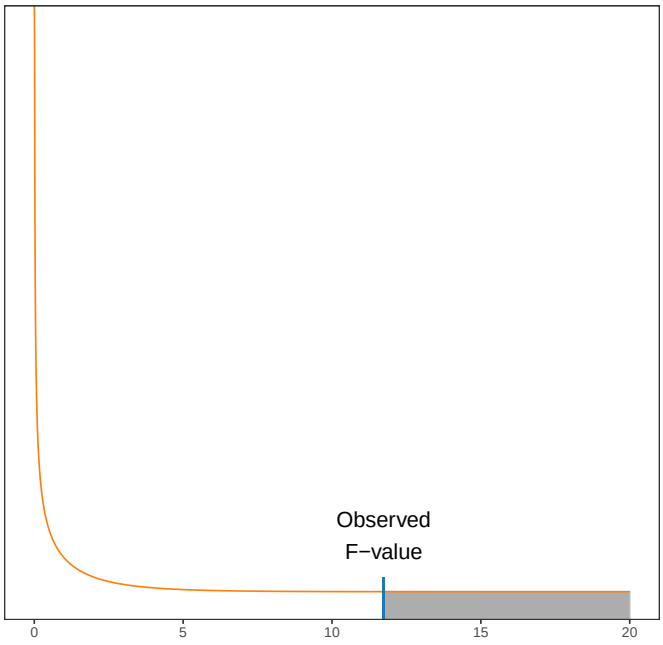

Figure 3. Plot of the probability curve for the F(1,98) distribution. The shaded area under the curve represents the p-value for a test evaluating whether the population rho-squared is zero using an observed F-value of 11.74.

This area (which is one-sided in the F-distribution) corresponds to the p-value. In our case this p-value is 0.000885. The probability of observing an F-value at least as extreme as we the one we observed ($F = 11.74$) under the assumption that the null hypothesis is true is 0.000885. This suggests that the empirical data are inconsistent with the hypothesis that $\rho^2 = 0$, and it is unlikely that the model explains no variation in students' GPAs.

Using the F-distribution in Practice

In practice, all of this information is provided in the output of the `glance()` function.

```
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     0.107        0.0981  7.24      11.8 8.85e-4     2  -339.  684.  691.
# ... with 2 more variables: deviance <dbl>, df.residual <int>
```

The observed F-value is given in the `statistic` column and the associated degrees of freedom are provided in the `df` and `df.residual` columns. Lastly, the p-value is given in the `p.value` column.

ANOVA Decomposition

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

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

```
Analysis of Variance Table

Response: gpa
          Df Sum Sq Mean Sq F value    Pr(>F)
homework   1  616.5  616.54  11.763 0.0008854 ***
Residuals 98 5136.4   52.41
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Note that the two *df* values 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 98) is the residual degrees-of-freedom. Note the p-value is the same as that from the `glance()` function.

This ANOVA decomposition also breaks out the sum of squared values into the variation explained by the model (616.5) and that which is unexplained by the model (residual variation; 5136.4). Summing these two values will give the total amount of variation which can be used to compute R^2; $R^2 = {\text{SS}_{\text{Model}}}/{\text{SS}_{\text{Total}}}$.

This decomposition also gives us another way to consider the F-statistic. Recall that F had a direct relationship to R^2

$$F = \frac{R^2}{1 - R^2} \times \frac{df_2}{df_1}$$

Using algebra, we could also express this as a ratio of two fractions

$$F = \frac{\frac{R^2}{df_1}}{\frac{1-R^2}{df_2}}$$

Since $R^2 = {\text{SS}_{\text{Model}}}/{\text{SS}_{\text{Total}}}$ we can rewrite this as

$$F = \frac{\frac{\text{SS}_{\text{Model}}}{\text{SS}_{\text{Total}}}}{1 - \frac{\text{SS}_{\text{Model}}}{\text{SS}_{\text{Total}}}} \times \frac{df_2}{df_1}$$

Using simple algebra,

$$\begin{aligned} F &= \frac{\frac{\text{SS}_{\text{Model}}}{\text{SS}_{\text{Total}}}}{\frac{\text{SS}_{\text{Total}}}{\text{SS}_{\text{Total}}} - \frac{\text{SS}_{\text{Model}}}{\text{SS}_{\text{Total}}}} \times \frac{df_2}{df_1} \\ &= \frac{\frac{\text{SS}_{\text{Model}}}{\text{SS}_{\text{Total}}}}{\frac{\text{SS}_{\text{Total}} - \text{SS}_{\text{Model}}}{\text{SS}_{\text{Total}}}} \times \frac{df_2}{df_1} \\ &= \frac{\text{SS}_{\text{Model}}}{\text{SS}_{\text{Total}} - \text{SS}_{\text{Model}}} \times \frac{df_2}{df_1} \\ &= \frac{\frac{\text{SS}_{\text{Model}}}{df_1}}{\frac{\text{SS}_{\text{Total}} - \text{SS}_{\text{Model}}}{df_2}} \end{aligned}$$

This expression of F helps us see two things: F is a ratio of the explained and unexplained variances, and F is distributed as a ratio of two chi-squared values.

F is a Ratio of the Explained and Unexplained Variances

First, the numerator is a function of the explained variation and the denominator is a function of the unexplained variation. The two degrees of freedom are also related to the model (explained) and residual/error (unexplained). In fact, df_1 is often referred to as df_{Model}, and df_2 is often referred to as df_{Error}. Furthermore, since $\text{SS}_{\text{Total}} - \text{SS}_{\text{Model}} = \text{SS}_{\text{Error}}$, F is often written as

$$F = \frac{\frac{\text{SS}_{\text{Model}}}{df_{\text{Model}}}}{\frac{\text{SS}_{\text{Error}}}{df_{\text{Error}}}}$$

In statistical theory, a sum of squares divided by a degrees of freedom is referred to as a *mean squared* value—the *average* amount of variation. Thus the F value here is the ratio of the average variation explained by the model and the average variation that remains unexplained. In our example

$$\begin{aligned} \text{MS}_{\text{Model}} &= \frac{616.5}{1} = 616.5 \\ \text{MS}_{\text{Error}} &= \frac{5136.4}{98} = 52.41 \end{aligned}$$

These values are also printed in the `anova()` output.

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

```
Analysis of Variance Table

Response: gpa
          Df Sum Sq Mean Sq F value    Pr(>F)
homework   1  616.5  616.54  11.763 0.0008854 ***
Residuals 98 5136.4   52.41
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The observed F-value of 11.76 indicates that the average explained variation is 11.76 times that of the average unexplained vartiation. There is an awful lot more explained variation than unexplained variation, on average.

Another name for a mean squared value is a *variance estimate*; the average amount of variation (in the squared metric) is quantified as a variance. For example, go back to the introductory statistics formula for variance

$$s^2_Y = \hat{\sigma}^2_Y = \frac{\sum(Y_i - \bar{Y})^2}{n-1}$$

This numerator is a sum of squares; namely the SS_{Total}. The denomiator is the df_{Total},

$$s^2_Y = \hat{\sigma}^2_Y = \frac{\text{SS}_{\text{Total}}}{df_{\text{Total}}} = \text{MS}_{\text{Total}}$$

Note that the MS_{Total} is not printed in the `anova()` output. However, it can be computed from the values that are printed. The SS_{Total} is just the sum of the printed sum of squares, and likewise the

$$df_{\text{Total}}$$

is the sum of the *df* values.

$$\begin{aligned}\text{SS}_{\text{Total}} &= 616.5 + 5136.4 = 5752.9\\ df_{\text{Total}} &= 1 + 98 = 99\end{aligned}$$

Then the MS_{Total} is the ratio of these values,

$$\text{MS}_{\text{Total}} = \frac{5752.9}{99} = 58.11$$

Since this is a variance estimate, we could also compute the sample variance of the outcome variable, `gpa`, using the `var()` function.

```
var(keith$gpa)
```

```
[1] 58.1102
```

The F-Distribution is the Ratio of Two Chi-Squared Distributions

Because mean square values are variance estimates, F can also be expressed as

$$F = \frac{\hat{\sigma}^2_{\text{Model}}}{\hat{\sigma}^2_{\text{Error}}}$$

Now that we know the numerator and denominator of the F-value are variance estimates, we can turn to the second thing: namely that the F-distribution is the ratio of two χ^2-distributions. Stat theory tells us that the sampling distribution for a variance is χ^2-distributed with a particular *df*. The model explained variance estimate ($\hat{\sigma}^2_{\text{Model}}$) is χ^2-distributed with $df_{\text{Total}} - p$ degrees of freedom, while the unexplained variance estimate ($\hat{\sigma}^2_{\text{Error}}$) is χ^2-distributed with p degrees of freedom.

Relationship Between Coefficient-Level and Model-Level Inference

Lastly, we point out that in simple regression models (models with only one predictor), 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.

```
# Model-level inference
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     0.107        0.0981  7.24      11.8 8.85e-4     2  -339.  684.  691.
# ... with 2 more variables: deviance <dbl>, df.residual <int>
```

```
# Coefficient-level inference
tidy(lm.1)
```

```
# A tibble: 2 x 5
  term        estimate std.error statistic  p.value
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)    74.3      1.94      38.3  1.01e-60
2 homework        1.21     0.354      3.43 8.85e- 4
```

That is because the model is composed of a single predictor, so asking whether the model accounts for variation in income level **is the same as** asking whether income level is different, on average, for employees with a one-unit difference in education level. *Once we have multiple predictors in the model, the model-level results and predictor-level results will not be the same.*

Confidence Envelope 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: $\hat{Y} = \beta_0 + \beta_1(X)$. Randomly 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.

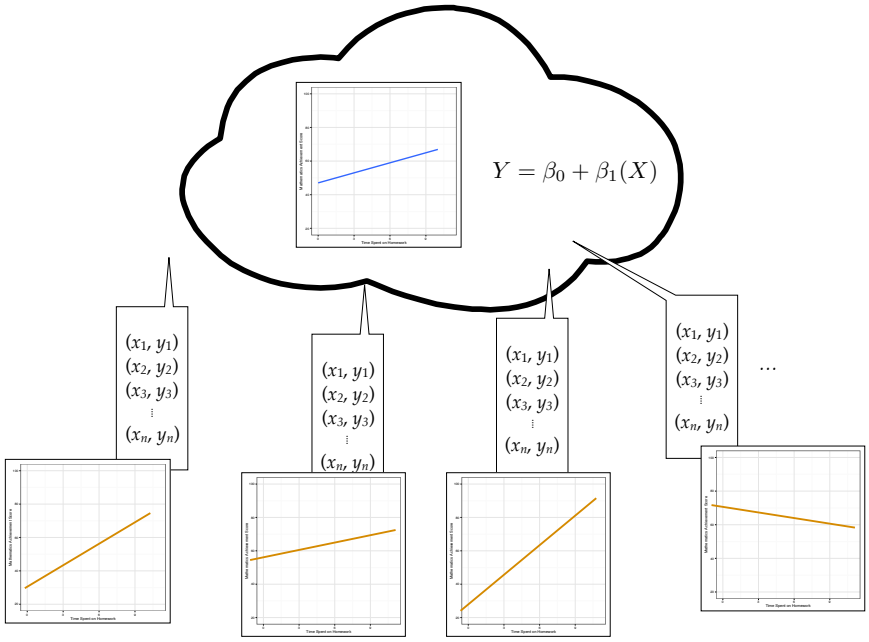

Figure 4. Thought experiment for sampling samples of size n from the population to obtain the fitted regression line.

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

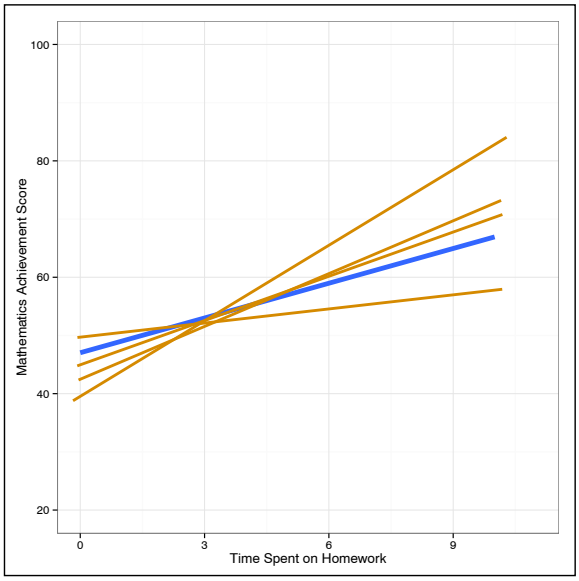

Figure 5. Plot showing the fitted regression lines for many, many random samples of size n.

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 using the layer `stat_watercolor_smooth()` from the **educate** package. You need to install the **educate** package from syntax since it is not currently available on CRAN. To do this you need to use the `install_github()` function from the **devtools** package. The syntax to install **educate** is:

```
library(devtools)
install_github("zief0002/educate")
```

Once **educate** is installed and loaded we have access to the `stat_watercolor_smooth()` function. This function, which we will use as a layer in `ggplot()` will be used to create the confidence enevelope. We also include the argument `method="lm"` in this layer. Finally, we add the regression line from the observed data using `geom_abline()`.

```
ggplot(data = keith, aes(x = homework, y = gpa)) +
  stat_watercolor_smooth(method = "lm") +
  geom_abline(intercept = 74.3, slope = 1.21) +
  xlab("Time spent on homework") +
  ylab("GPA (on a 100-pt. scale)") +
  theme_bw()
```

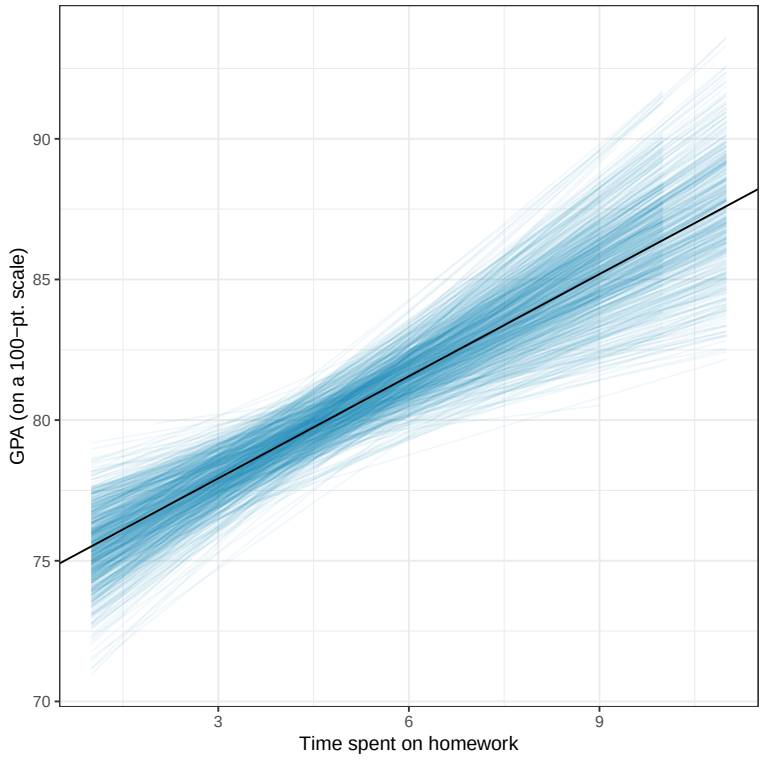

Figure 6. GPA plotted as a function of time spent on homework. The OLS regression line (black) and regression lines for 700 bootstrapped samples (blue) are also displayed.

Note that the confidence envelope is made up of many different regression lines. These are lines that we could have obtained had we drawn a different sample. Because of this, sometimes this plot is referred to as a *hypothetical outcomes plot*. These lines are based on fitted lines to a set of re-sampled, or bootstrapped, data. The level of transparency gives us an indication of the uncertainty based on the observed data. More probable locations for the regression lines are darker, while less probably locations are lighter.