

Regression inference: main concepts

- The slope and intercepts we compute in least squares regression are **statistics**, based upon our sample data. They are estimates of corresponding parameters; namely, the slope and intercept we would observe if the relationship between our variables were truly linear and we had data on the entire population.
- The regression model is more than just the equation of a line; it is a model that tells us what reality should look like. Remember: just because the model claims to reflect reality doesn't mean this is so.
- The regression model says (1) the relationship between x and the mean of y is truly linear; (2) the distributions of x -values about this linear relationship are normal; (3) the standard deviations of these various normal distributions are the same for every x -value; and (4) for each x -value, the corresponding y -values are independent.
- Under the four conditions described above, our statistics (sample slope and intercept) are unbiased estimators of the corresponding parameters.
- Under those same conditions, we can perform statistical inference procedures on the slope and intercept, though inference on the intercept is far less common (and less important).
- In particular, we can perform a “model utility test,” which considers the null hypothesis that the population slope is actually zero. If this hypothesis is true, then our linear model is not “useful,” in the sense that our explanatory variable does not help us predict the value of our response variable.
- The predicted y -value given by the regression line is interpreted as the mean value of all possible y 's that we could observe for that particular x value (assuming the model is good, of course).
- Even in statistical inference, association (correlation) does not imply causation. If we were to reject the aforementioned null hypothesis and conclude that our linear model is “useful,” we still could not conclude that our explanatory variable causes the observed responses.

Regression inference: teaching tips

- You might want to review the section on regression from earlier in the course. Back then, we were concerned with describing relations between two variables. Now we're interested in inferring relations that exist in the population based on a random sample.
- There are (at least) two purposes for regression. One is prediction: for a given x value, how well can I predict the y -value that I'll see? The other purpose is to estimate the value of the slope and, in particular, to see if it is non-zero.
- The output that Fathom provides for regression is idiosyncratic and doesn't look like the output that students will be expected to interpret on the AP test. Make sure they see examples of output from other programs (such as Minitab).
- A good exercise for students, to help them learn to read regression output tables, is to give them incomplete tables and ask them to re-construct the missing output.
- Use examples where many of the x values are the same. Students need to see scatter plots where multiple y values correspond to the same x value to absorb the idea of responses having a **distribution** for each x value.
- As with all previous inference procedures, you must check certain conditions before you proceed; otherwise, the confidence interval/hypothesis test calculations performed by the computer are worthless. You can check three of the underlying conditions for hypothesis tests of a slope:
 1. Are the data linear? Look at the residual plot for curvature or other violations of linearity.
 2. Is the variance constant across all x values? Look at the residual plot for fanning or bulging.
 3. Is the distribution of the responses normal? Look at a normal probability plot of the residuals.
- By-hand computations of the slope, let alone the standard error of the slope, are unpleasant and distracting. Students should practice interpretation rather than computation.

Regression inference: sampling variability

This activity will (hopefully) reinforce the idea of slope and intercept as statistics, each varying from sample to sample. We will use Beth Chance's regression sampling applet:

<http://statweb.calpoly.edu/chance/applets/regcoeff/regcoeff.html>

The applet allows you to select the population slope and intercept, which in turn determine the population regression line (in yellow). You can also choose a mean and standard deviation for the x -values, and finally the population standard deviation for the responses about the regression line. For now, let's all be consistent:

- Set the population slope to 1.5 and the population intercept to 2. Keep all other values the same.
- Click the Set Population button to create your new population of data. (Note: If you would like to see more of the graph, you can change the window frame using the gray boxes along the 4 sides of the graph.)

At the bottom of the page, you should see the equation $y = 1.50x + 2$. That is the **population equation**.

We will now sample from the population displayed on the graph (the blue dots).

- Hit the Draw Samples button once. The applet randomly selects a sample of points (in red; $n = 80$ is the default). Then the applet calculates the least squares regression line for those n points and graphs that line in red. The equation of the line appears at the bottom of the page. Is it exactly the same as our population equation? Do the graphs line up exactly? Why not?
- Hit the Draw Samples button a few more times, just to see how the samples — and, hence, the resulting least squares regression lines — differ from sample to sample. This is an illustration of sampling variability.
- Change the “num samples” from 1 to 100, then click Draw Samples. The applet will superimpose all 100 sample least squares lines onto the graph (the “wave” of red) and launch a window with dot plots of the sampling distributions of the slope and intercept. Focus on the slopes: what do you see? Is the center of the dot plot reasonably close to 1.5? Do you notice a shape forming?
- Before you close the dot plot window, note the standard deviation for the slopes somewhere.

Now let's see how varying the other "parameters" of the applet changes things.

- Hit the Reset button.
- Change the value of sigma from .45 (the default) to 2.45, and click Set Population. What do you notice happened to the population graph? Remember, sigma is the standard deviation of the y -values about the regression line.
- Once again, take 100 samples of size $n = 80$ and look at the sampling distribution of the slopes. What happened to their spread? (That is, did the standard deviation of the slopes increase or decrease, compared to the value you noted earlier?) Is this what you would expect to happen for a larger value of sigma?
- Change sigma back to .45 (the default) and click Set Population. For our next illustration, change the sample size from 80 to 20. Again, take 100 samples and look at the resulting slopes. What happened to the spread of the slopes this time? Is this what you would expect to happen for a smaller sample size?
- Finally, change the sample size back to 80 (the default). How do the x -values play a role? To find out, change "x std" (the standard deviation of the x -values) from 1.84 to 4.84. With sample size back at 80, take 100 samples again. What happened to the spread of the slopes? Does this result surprise you?

Assuming the simulations went according to plan, we should have found three patterns among the variety of lines provided by the variety of random samples:

- (1) The larger the standard deviation of the responses about the line, the more widely-varying our estimates of the slope will be.
- (2) The variability of the sampling distribution of the slopes is larger for smaller sample sizes.
- (3) Slopes across different samples are *less* variable when the x -values are more variable.

Regression inference: example

The data below show information on price (in thousands of dollars), horse power, and gas mileage for a random sample of vehicle models commonly sold in the United States. You will use this data for the practice problems.

<u>Model</u>	<u>price</u>	<u>hp</u>	<u>mpg</u>
Chevy Cavalier	13.4	110	36
Chevy Lumina APV	16.3	170	23
Chevy Astro	16.6	165	20
Dodge Shadow	11.3	93	29
Dodge Caravan	19.0	142	21
Eagle Vision	19.3	214	28
Ford Probe	14.0	115	30
Hyundai Elantra	10.0	124	29
Lexus SC300	35.2	225	23
Mazda RX-7	32.5	255	25
Oldsmobile Achieva	13.5	155	31
Pontiac Grand Prix	18.5	200	27
Suzuki Swift	8.6	70	43
Volkswagen Fox	9.1	81	33
Volvo 850	26.7	168	28

1. Test the hypothesis that horse power is a useful linear predictor of gas mileage. Be sure you check the necessary conditions for the hypothesis test.
2. Find and interpret a 95% confidence interval for the slope relevant to Question 1.
3. Find and interpret a 95% confidence interval for the slope for predicting price from horse power, if this is reasonable. If not, explain why not.

Regression inference: example

SOLUTIONS

1. Before any analysis, make a scatter plot of the data, with horse power as the explanatory (x) variable and miles per gallon as the response (y) variable. The scatter plot looks reasonably linear, so a linear model should be appropriate here.

For the hypothesis test, our parameter is β , the true slope for predicting gas mileage (mpg) from horse power across all motor vehicles. The hypotheses are $H_0: \beta = 0$ versus $H_a: \beta \neq 0$.

Make a residual plot and a normal quantile plot of the residuals. While not perfect, the plots do not suggest a severe violation of linearity, constant variance, or normality. We will proceed with the test.

Using our computer, we find the sample slope is $b = -0.06940$, the standard error of the slope is $s_b = 0.02329$, the test statistic is $t = (-0.06940 - 0)/0.02329 = -2.98$ with $df = n - 2 = 15 - 2 = 13$, and the P -value is 0.011. Since this is less than the standard $\alpha = 0.05$, we reject the null hypothesis.

At the 5% significance level, we conclude that horse power is a useful linear predictor of gas mileage (mpg).

2. At 95% confidence, the t critical value at 13 df is 2.16, and so a 95% confidence interval for β is $b \pm 2.16 \times s_b = -0.06940 \pm 2.16 \times 0.02329 = (-.1197, -.0191)$. Hence, we are 95% confident that a positive difference of 1 horse power corresponds, on average, to a negative difference of between .0191mpg and .1197mpg.

3. We should again check that the requisite assumptions are met. A scatter plot shows a moderate positive relationship between price and horse power. However, a careful examination of the residuals reveals two problems: (1) the residual plot clearly shows non-constant variance, and (2) a normal quantile plot shows the residuals might be non-normal, although the latter problem is less pronounced. Hence, a blind calculation based on the standard t formula would not be valid here.