Week 14: Residuals and Regression Inference

Data 8 Tutoring

# 1 Residuals

# Key Concepts

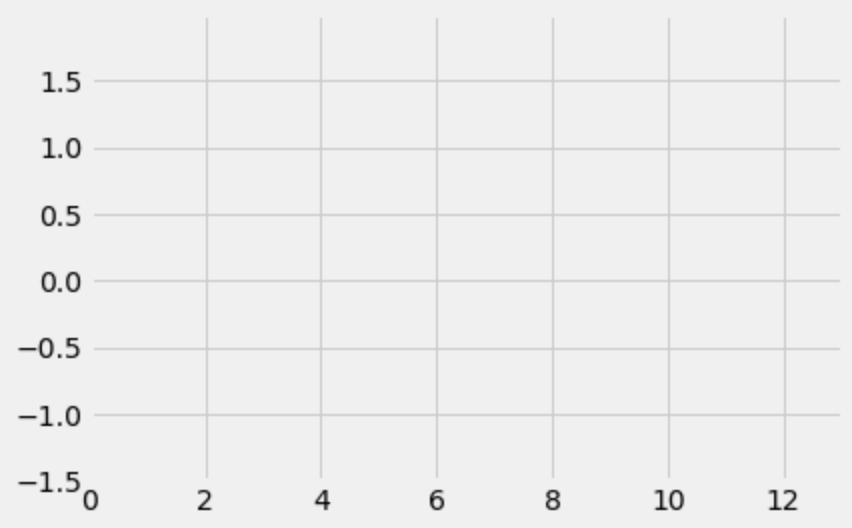
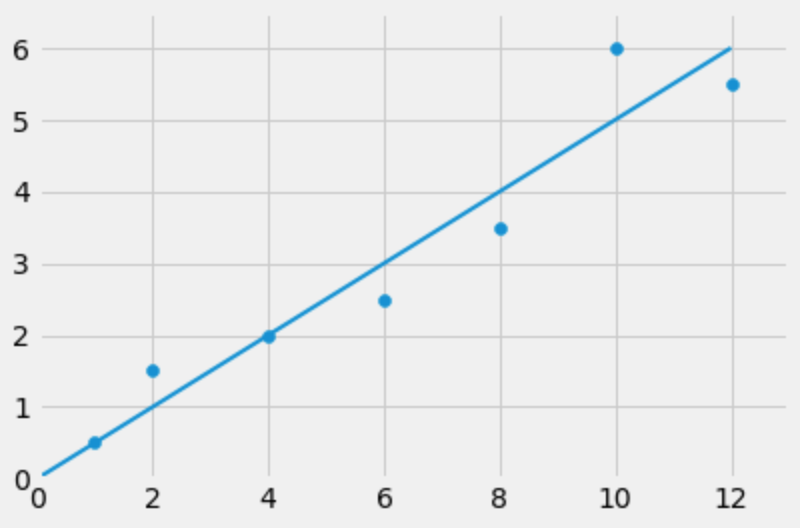
## **Definition/Properties**

* A residual is the difference between an observed value and its corresponding regression estimate.
* Visually, residuals are also the vertical difference between each observed point and its corresponding estimated point.
* The larger the absolute value of the residual, the further away our estimate is from our actual data point. If the estimates are equal to our data points, then all of our residuals will be equal to 0.
* residual = *y* − estimated value of *y* = *y* − height of regression line at *x*
* The sum of all residuals is 0.
* SD of residuals = SD of y

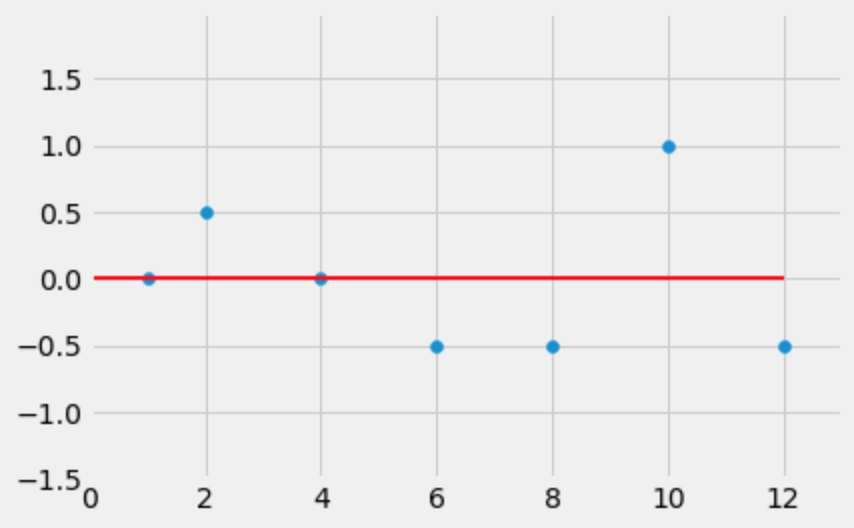
## **The relationship between Residual Plots and Regression**

* Residual plots are used to visually diagnose how well our regression line fits the data.
* First of all, the goal of least squares is to choose a line of best fit that minimizes error. We can use least squares to help us calculate the best slope and intercept to fit our data.
* Secondly, we should see that the residuals are random, and that there are no other non-linear relationships (e.g. quadratic, exponential) to explain our data.

Practice Problems

**1.1** Given the following regression line, draw an approximate residual plot.

# SOLUTION



**1.2** Answer whether the following questions are **True** or **False**.

**a)** If we perform linear regression on two variables, the residual plot never has a trend.

True

**b)** No matter what the shape of the scatter diagram, the average of the residuals is 0.

True

**c)** No matter the shape of the scatter plot, the SD of the residuals is less than or equal to the SD of the y values.

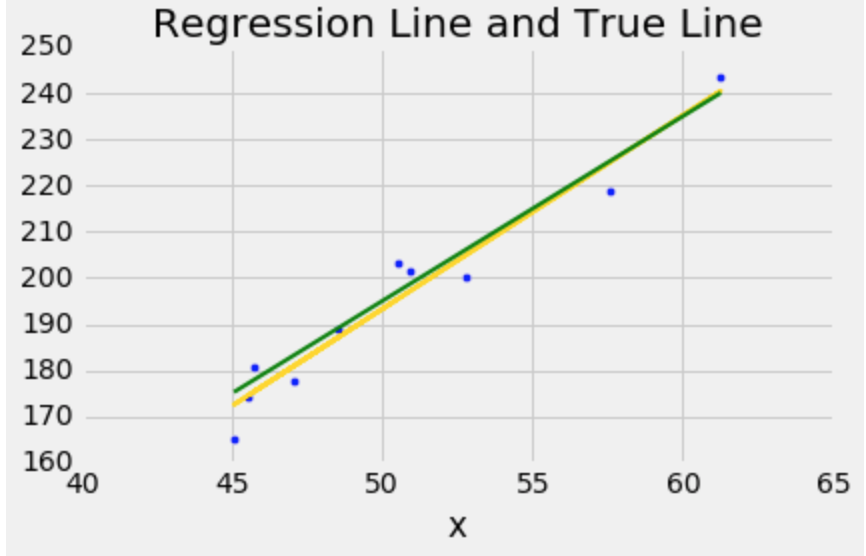
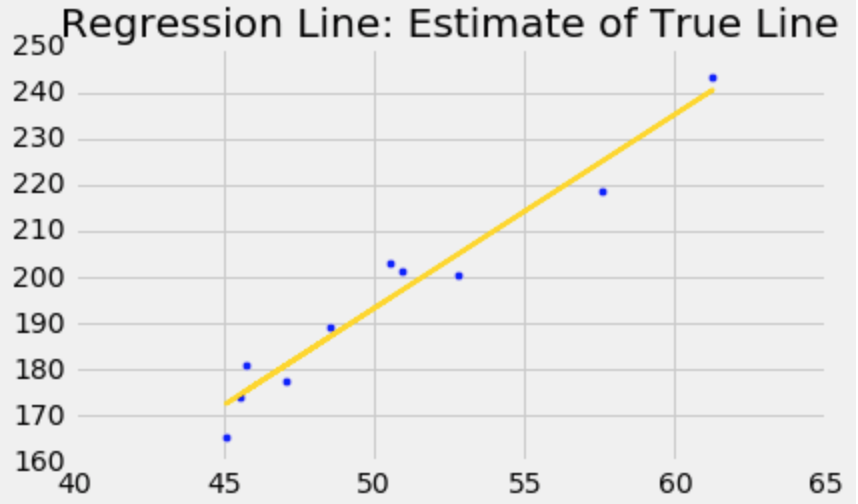
True

# 2 Regression Inference

Key Concepts

**Inference for the True Slope**

* Thinking about the bigger picture, let’s assume that we have a dataset that is completely linear and begin pushing the points away from the line at random, symmetrically on both sides. You end up with a data set that is clustered around the “true line.”



## **Key Mathematical Definitions**

* correlation coefficient
* *slope* =
* *intercept = average of Y – slope \* average of X*

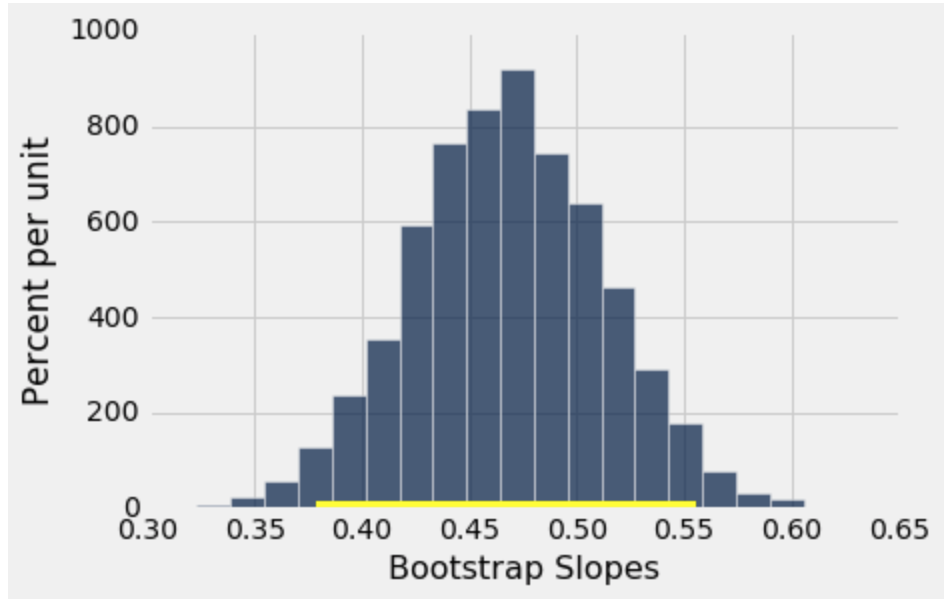
Practice Problems

**2.1** True or False:

**a)** The regression line’s *x* and *y* values are always measured in standard units.

False, it can be, but not necessarily/always

In the real world, we want to estimate a true line for a given population, but we only have access to a small number of observations. We can **bootstrap** to create confidence intervals for our true value of y. We are interested in the confidence interval for the slope since it helps us determine whether there really is a linear relationship between x and y.



It is important to bootstrap entire rows from a table to keep the data intact because we want to maintain the relationship between *x* and *y*. Splitting those up will eliminate that relationship. So we use tbl.sample(), which samples with replacement the same number of rows from the table.

**2.2** Why do we need to bootstrap slopes to test if a true slope is 0 or not? Why can’t we just tell from the sample slope we get?

We need to bootstrap slopes because it is highly unlikely that a sample slope would be exactly the same as the true slope. Most likely, if the true slope was 0, we would get sample slopes that are slightly off from 0, so creating a confidence interval helps us understand if we can or cannot reject the null hypothesis of the true slope being 0 based on our sample.

Let’s try to estimate a real slope using this bootstrap technique. We’ll be pulling an example from the textbook so you can look back at it for future reference. We have a dataset on babies’ birth weights and corresponding gestational days (days in the womb before birth).

One might guess that the longer a baby is in the womb, the heavier it is when it is born - since it has more time to grow.

The table baby is as follows:

|  |  |
| --- | --- |
| **Gestational Days** | **Birth Weight (oz)** |
| 284 | 120 |
| 282 | 113 |
| 279 | 128 |

1147 rows omitted…

**2.3** State the null and alternative hypotheses.

Null hypothesis:

There is not a linear relationship between gestational days and birth weight. The slope of the “true line” is 0.

Alternative hypothesis:

The slope of the “true line” is not 0.

Next, we will bootstrap our sample and repeat the regression process to estimate the variability of the regression slope.

Assume we already have a function correlation(tbl, x, y) that returns the correlation coefficient given a table and two columns. The following slope function might be helpful.

def slope(table, x, y):

r = correlation(table, x, y)

return r \* np.std(table.column(y))/np.std(table.column(x))

**2.4** Complete the function below to make one bootstrapped sample of the baby table, and calculate the slope of the best fit line of that bootstrapped sample.

Hint: You can use the slope function defined above!

def one\_slope():

bootstrapped\_baby\_table = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

slope = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

return \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

def one\_slope():

bootstrapped\_baby\_table = baby.sample()

slope = slope(bootstrapped\_baby\_table, ‘Gestational Days’, ‘Birth Weight (oz)’)

return slope

**2.5** Using the one\_slope function defined in 2.4, populate the array slopes with 10,000 bootstrapped slopes from the baby table.

slopes = make\_array()

for i in np.arange(10000):

bootstrapped\_slope = one\_slope()

slopes = np.append(slopes, bootstrap\_slope)

**2.6** Find the endpoints of the 98% confidence interval for our bootstrapped slopes.

lower\_bound = percentile(1, slopes)

upper\_bound = percentile(99, slopes)

**2.7** Let’s say we get the 98% confidence interval (0.356, 0.585)

1. What is the p-value cutoff associated with our level of confidence? Do we reject or fail to reject the null hypothesis at this cutoff value?

At a p-value cutoff of 2%, we reject the null hypothesis, since our confidence interval does not contain 0.

1. At a p-value cutoff of 5%, are we able to make conclusions about the null hypothesis? If so, do we reject or fail to reject the null hypothesis?

At a p-value cutoff of 5%, we reject the null hypothesis, since a 95% CI would be narrower and would not contain 0.

1. At a p-value cutoff of 1%, are we able to make conclusions about the null hypothesis? If so, do we reject or fail to reject the null hypothesis?

At a p-value cutoff of 1%, we cannot make conclusions since we do not know how much wider a 99% CI would be.