Week 14: Residuals and Regression Inference

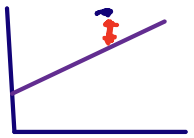
Data 8 Tutoring

---

1 Residuals

Residual := $y_{observed} - y_{expected}$

↳ The Error



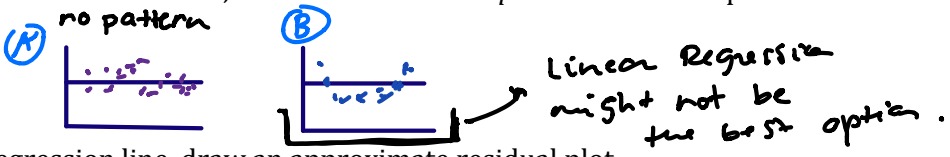
Key Concepts

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. ⟶ Intuition
- residual = y − estimated value of y = y − height of regression line at x
- The sum of all residuals is 0.
- SD of residuals = $\sqrt{1 - r^2} * \text{SD of } y$ ↳ If time left, we can derive this formula.

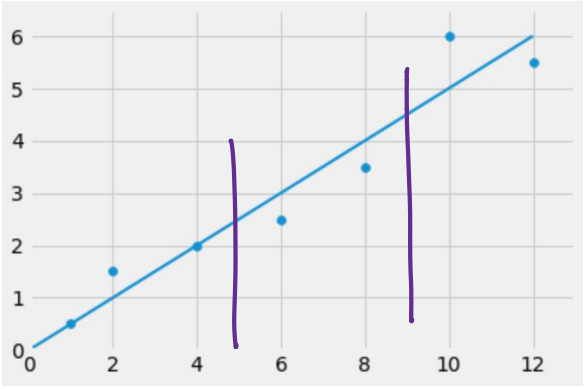
The relationship between Residual Plots and Regression

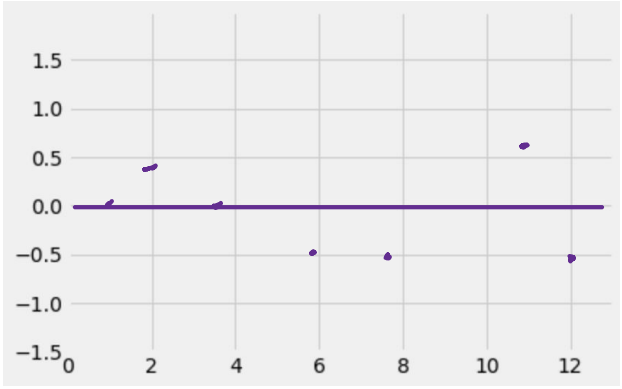
- Residual plots are used to visually diagnose how well our regression line fits the data.
- First of all, the goal of least squares is to choose a line of best fit that minimizes error. We can use least squares to help us calculate the best slope and intercept to fit our data.
- If linear regression is a good method to use, then there will be *no pattern* in the our plot of residuals.



Practice Problems

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





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; Trend refers to correlation. Correlation CAN'T be applied to residuals. Conversely, patterns can occur.

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

True; $\frac{\text{sum(res)}}{\text{nums}} = \frac{\emptyset}{\text{nums}} = \emptyset$;

c) No matter what the shape of the scatter plot, the SD of the residuals is less than or equal to the SD of the true y values. → [0, 1]

True; $\sqrt{1-r^2}$ $SD_y = SD_{residuals}$

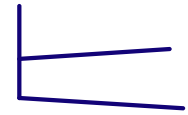
↳ $r = [-1, 1]$

2 Regression Inference

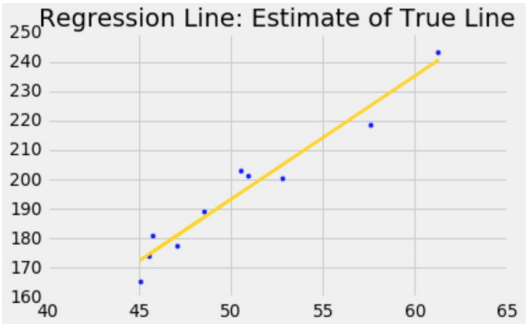
Key Concepts

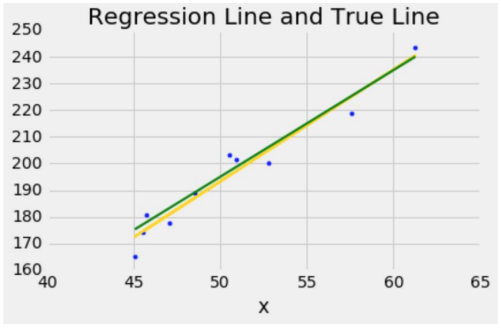
Inference for the True Slope

- 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."



r = 0.01





Key Mathematical Definitions

- correlation coefficient $r = mean(x_{standard\ units} * y_{standard\ units})$
- $slope = r * \frac{SD\ of\ Y}{SD\ of\ X}$
- $intercept = average\ of\ Y - slope * average\ of\ X$

Practice Problems

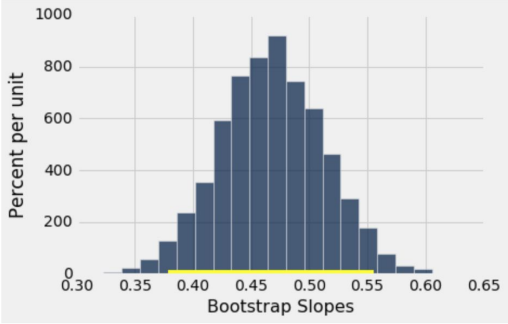
2.1 True or False:

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

It doesn't always have to be.

Previously in this class, you all constructed confidence intervals using the bootstrap to predict population parameters, such as the population mean. When investigating the correlation between two variables, you can actually construct a confidence interval for the slope of the best fit line using the same method!

Performing this calculation can allow you to determine if you believe there actually exists a correlation between two variables, or if the relationship you observe between the variables is due to random variation in the sample.



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?

→ 0.5

→ -0.5

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…`

You're interested in constructing a test to determine whether there exists a relationship between the number of gestational days and the birth weight of the baby. This would suggest that the slope of the best fit line when using the gestational days to predict the birth weight is nonzero.

2.3 State the null and alternative hypotheses.

Null hypothesis: True slope is zero. There's no relationship b/w the two variables.

Alternative hypothesis: True slope is nonzero.

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 between two columns, `x` and `y`, in the table `tbl`. Additionally, we have already defined the following `slope` function.

```
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!

with_replacement = True
size = n

```
def one_slope():
    bootstrapped_baby_table = baby.sample()
    slope = slope(b-b-t, 'gest', 'bw')
    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(10_000):
    bootstrapped_slope = one_slope()
    slopes = np.append(slopes, bt_s)
```

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)`

a) 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?

0.02; null hypothesis: True slope is ø
↳ ø not in 98% CI → Reject Null

b) 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?

95% < 98% → Reject Null Hypothesis.
narrower than

c) 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?

We don't know
↳ we need to calculate 99% Confidence Interval