Notebooks 19 & 20: Scatter Plots and Linear Regression
Course Notes: Statistics 1401

UNG Mathematics

November 2024

1 What is Linear Regression?

A **Linear Regression** hypothesis test compares two *numeric variables* to determine whether or not a linear relationship exists between the two.

We first inspect a scatter plot using the **.scatter** method on the table in question:

table_name.scatter(col_1,col_2)

After checking for associations in the scatter plot (e.g. evidence of a linear relationship) we run the regression. The output of this hypothesis test provides 3 things:

- Correlation (r)
- p-value
- Equation of the fit-line

The value r is called the correlation coefficient. Correlation ranges from $-1 < r < 1$, with $r \approx 0$ indicating no relationship (zero correlation). The p-value works like the t-tests. When we reject the null, we have evidence of a linear relationship. If we fail to reject the null, we have no evidence of a linear relationship.

2 Linear Regression

The inputs must be arrays, so **array1** is the x-variable values, and **array2** is the y-variable values.

2.1 Code

The following command will work if we have array data for both numeric variables.

stats.linregress(array1,array2)

2.2 Results

The output from the **stats.linregress** function is rather extensive, for example:

LinregressResult(slope=0.99214, intercept=7.3617,
rvalue=0.98476, pvalue=6.93252e-39, stderr=0.025035)

Focus specifically on the **rvalue** and the **pvalue**.

3 Examples: Scatter Plots

In Notebook 19, we have several examples:

- Baby birth weights and gestational days
- Hybrid (cars) table example with mpg and msrp
- Correlation demo

After these examples, we can transition to Notebook 20 and show examples with all steps including the **stats.linregress** hypothesis test.

4 Notebook 20, Example 1: SAT Scores

Do the scores on the SAT Math section correlate with the scores on the Critical Reading section?

4.1 Hypothesis

$$H_0 : \rho = 0$$
$$H_a : \rho \neq 0$$

The value ρ is the lower case Greek letter (pronounced "row"). We are using the correlation coefficient to estimate ρ, so we have

$$r \to \rho$$

As usual, the Law of Large Numbers tells us that this estimate becomes more and more accurate as sample size increases.

4.2 Run Test

Recall that, prior to the hypothesis test, we use the **.scatter** method to display a scatter plot. Once we have confirmed evidence of a linear relationship exists, then we can run the hypothesis test. The **stats.linregress** command produces a ton of output.

stats.linregress(sat2014.column("Critical Reading"), sat2014.column("Math"))
LinregressResult(slope=0.99214, intercept=7.3618,
rvalue=0.984756, pvalue=6.93252e-39, stderr=0.025035)

4.3 Reporting Out

Because $p = 6.93 \times 10^{-39} < 0.05 = \alpha$, we reject the null. Thus, we have evidence of a positive linear relationship between the scores on Math and Critical Reading.

5 Example 2: COVID Masks

Is mask compliance linked to **higher** COVID-19 death rates in Europe?

5.1 Hypothesis

Given how the reearch question is worded, notice that we are testing whether there is a significant **positive** linear relationship. Thus, we will use the "greater than" symbol in the alternative hypothesis.

$$H_0 : \rho = 0$$
$$H_a : \rho > 0$$

Run Test

stats.linregress(mask.column("Compliance"), mask.column("Deaths"))
LinregressResult(slope=837.6193464140961, intercept=540.9381857649194, rvalue=0.30209139971917576, pvalue=0.07777239516786949, stderr=460.1202833007869)

5.2 Reporting Out with the COVID Example

Since $p = 039 < 0.05 = \alpha$, we reject the null. Evidence suggests there was a positive correlation between compliance with mask mandates and COVID mortality rates.

Whoa! Do these results mean that, in Europe, wearing masks increased one's risk of mortality due to COVID?

No, but we have evidence that, in Europe, higher compliance with mask mandates did not improve mortality rates.

5.3 Key Takeaway

Correlation does not imply causation.