Class 23: Regression and Hypothesis Testing (Text: Sections 10.1)

Review of Regression (from Chapter 2)
We fit a line to data to make projections.

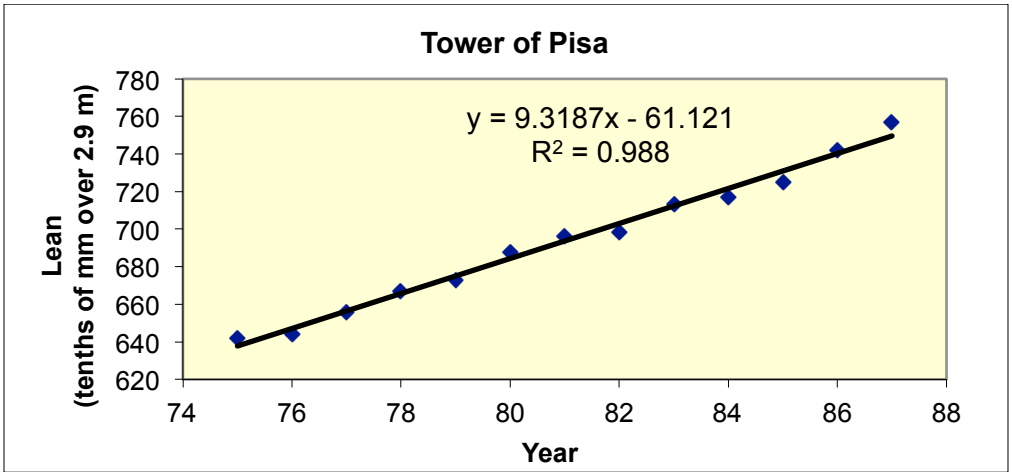
The Tower of Pisa is leaning more each year. The measurements below show the lean in tenths of millimeter beyond 2.9 meters. Thus in 1975, the Tower was leaning 2.9642 meter from the vertical. (http://torre.duomo.pisa.it/towersposters/english_version/)

Obs	Year	Lean (in tenths of mm over 2.9 m)
1	75	642
2	76	644
3	77	656
4	78	667
5	79	673
6	80	688
7	81	696
8	82	698
9	83	713
10	84	717
11	85	725
12	86	742
13	87	757

1. Does it look like the lean of the Tower is increasing with time? If so, how fast is the lean increasing with time?
2. Is there evidence that the lean changes significantly with time?

What are we doing now with regression that we did not do before?
Before: We fit a line to make projections
Now: We ask if the changes predicted by the line are significant. Could the predicted changes just be the result of random variation in the sample, while, in fact, no changes are happening?

Ex: Use the scatterplot and regression equation to describe how the lean changing with time. Interpret the slope and vertical intercept.



The scatter plot shows the lean is increasing with time.
The equation of the line is $y = -61.12 + 9.32x$, where lean is y and year is x.
The correlation is $\sqrt{0.988} = 0.994 = 99.4\%$.
The lean is increasing at a average rate of 9.32 tenths of a millimeter per year.

Notice the "real" axes are not shown; the real ones will go through the origin.
The -61.12 represents the lean the Tower would have had in year 0, that is 1900, if it has been growing at the same rate throughout the century. Recall that the lean is measured beyond 2.9 meters, so a lean of -61.12 means a lean of 61.12 tenths of a millimeter *less* that 2.9 meters.

How can we decide if the lean is increasing significantly with time?
Do a hypothesis test. The regression equation,

$$y = -61.12 + 9.32x,$$

was derived from a sample. It is a ***sample regression line***.

The slope and intercepts in this sample regression equation are estimates, based on the particular sample we had. If we took another sample (for example, if we measured the tower at different times of year), we would probably get a different slope and a different intercept. They are ***sample statistics***.

What we are really interested are the population parameters—that is, the slope and intercept of the line that fits the whole population, not just these sample values. This is the ***population regression line***.

Is it possible that the population regression equation has

- A different slope than the one we found, 9.32?
- A slope of different sign?
- A slope of 0?

It is certainly likely we'd get a different slope. The 9.32 is an estimate, and if we took a different sample, we'd expect to get a different slope. Whether or not we'd get negative slope depends on whether the lean really is increasing. We'd like to know if the slope is significant different that 0. We do a hypothesis test on the sample slope to find out.

Hypothesis Test that the Slope of Population Regression Line is Significantly Different than 0

Step1: Null hypothesis: Slope of regression line is 0: There is no relation between lean and year.
Alternate hypothesis: Slope of regression line is different from 0. There means there is a relationship between lean and year. (Note: Two-sided test.)
Step 2: Test statistic: Turns out to be a t statistic, with $n-2$ degrees of freedom, where n is the number of data points.
Step 3: *P*-value: is found using a computer to be $6.5 \cdot 10^{-12}$, which is tiny.
Step 4: Conclusion: The probability of seeing, by chance, a slope as different from 0 as we did if there was no relationship between the lean and time is $6.5 \cdot 10^{-12}$. Thus, we reject the null the idea that there is no relationship between lean and year; we conclude there ***is*** a relationship.

Regression on Excel
PC: Use the Data Analysis ToolPak (under Data menu), find Regression and fill in the dialog box.
Mac: Use StatPlus, under Statistics and fill in dialog box.

Ex: Interpret the Excel Regression output

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.99397169
R Square	0.98797972
Adjusted R Square	0.98688696
Standard Error	4.18097112
Observations	13

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	15804.4835	15804.4835	904.119785	6.50337E-12
Residual	11	192.285714	17.4805195		
Total	12	15996.7692			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-61.120879	25.1298185	-2.4322054	0.03327937	-116.4312647	-5.8104935
Year	9.31868132	0.3099142	30.0685847	6.5034E-12	8.636564422	10.0007982

The regression equation can be found from the table; the intercept and slope are given in the bottom two lines (left end). The standard errors of these estimates are next to them. The R^2 value of 0.988 is listed near the top. Notice the *P*-value =6.5034E-12 for the coefficient of the year.

Ex: How are the t-statistics and P-values for the slope calculated?
Since

$$t = \frac{\text{Value} - \text{Mean}}{\text{Standard error}},$$

so for the slope

$$t = \frac{9.31868132 - 0}{0.3099142} = 30.06858.$$

The P-values come from the table or from the calculator. Since the test is two-sided, we have $2 \cdot \text{tcdf}(30.07, 1000, 11) = 6.5 \cdot 10^{-12}$.

Ex: What is the confidence interval for the slope? What does it tell us?
The interval is $(8.637, 10.001)$. It tells us that the population slope is very likely (95% probability) to be between 8.637 and 10.001. In other words the slope is not likely to be 0 or negative. The fact that the interval does *not* contain 0 corresponds to the fact that we rejected the null hypothesis.

Ex: How does the standard error of the slope, 0.3099, relate to the confidence interval?
For $df = 11$, we have $t = 2.201$ for a 95% confidence interval, which is

$$(9.319 - 2.201 \cdot 0.309, 9.319 + 2.201 \cdot 0.309) = (8.64, 10.00).$$

Ex: What does the Excel output say about the value of the intercept?
We do not often have a practical reason to test the values of the intercept. However, we see that for the intercept

$$t = \frac{-61.12 - 0}{25.1298} = -2.4322.$$

and $P = 0.0332 = 3.32\%$. Thus we see that we would reject the hypothesis that the intercept is 0. The confidence interval, $(-116.4, -5.8)$, does not contain 0, supporting this conclusion. The intercept is most likely to be between -116.4 and -5.8, so the intercept is likely to have a negative value.

In ANOVA block:
Ex: What is the F-value and its P-value telling us? (P-value is called Significance F)
The F-value and its P-value tells us whether the regression equation is a significant predictor of the response variable. Since $P = 6.5 \cdot 10^{-12}$ is so small, we have evidence of a significant prediction.

A regression with only one independent variable (like this one) is called a **bivariate regression**. In such a case, the P-value of the F-statistics and the P-value of the hypothesis test for the slope are equal (here $6.5 \cdot 10^{-12}$. For a **multiple regression** with more independent variables, there is one P-value for the F-statistic, which tells us how well the whole regression predicts. In addition, the slope coefficient of each independent variable has its own P-value, showing whether that variable is contributing significantly to the prediction.

In Regression Statistics Block:
Ex: What does the standard error, 4.18, in the Regression Statistics tell us?
This is the standard deviation of the errors in the predictions. It tells us how far, on average, the predicted y-values are from the actual y-values.

Note: There are two standard errors in this page:

- 0.3099 is the SD of the estimates for the slope
- 4.18 is the SD of the data around the line

Since there is more variability in the individual data values than in the predicted means, the second SD is larger.

Regression and Hypothesis Tests

The notation for the ***sample regression line*** is

$$\hat{y} = b_0 + b_1 x$$

The "hat" on the y means it was estimated. The coefficients in the sample regression equation are given by

$$b_1 = r\frac{s_y}{s_x} \text{ and } b_0 = \bar{y} - b_1\bar{x}.$$

Thus, the point $(\bar{x}, \bar{y})$ is on the sample regression line.

The ***population regression line*** is written as

$$\mu_y = \beta_0 + \beta_1 x$$

(As before, Greek letters are population parameters; ordinary letters are the sample statistics.)
The μ_y is the mean value of y for that particular x. The sample regression line, which is found by the least squares method, is an estimate of the population regression line.

Suppose there are n data points, $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$. Even if we had the population regression line, the data points would probably not lie exactly on it. Random variation means that the points will around the line, and a point (x_i, y_i) will satisfy

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

where ε_i is the ***error*** or ***residual***. We assume the residuals are normally distributed with mean 0 and standard deviation σ. We assume that the residuals for different values of x are independent and the standard deviation σ is the same for all values of x.

The ***population parameters*** for this model are β_0, β_1, and σ.

Is there evidence for a relationship between x and y?
Notice that if $\beta_1 = 0$, then the population regression equation becomes

$$\mu_y = \beta_0 + 0 \cdot x$$

That is

$$\mu_y = \beta_0.$$

In other words, if $\beta_1 = 0$, the value of μ_y no longer depends on x, so there is no relationship.
Similarly, if $\beta_1 \neq 0$, then values of μ_y vary as x varies, and there *is* a relationship.

To test if there is evidence of a relationship between x and y, we test the null hypothesis that $\beta_1 = 0$ against the alternate hypothesis $\beta_1 \neq 0$. The test statistic is

$$t = \frac{b_1 - 0}{SE_{b_1}}$$

which has the T-distribution with $n - 2$ degrees of freedom, where n is the number of data points.

The errors, ε_i, have standard deviation σ, assumed independent of x. To estimate σ, we use an average of the squared residuals, giving the ***standard error*** listed in the Regression Statistics:

$$s = \sqrt{\frac{\sum \varepsilon_i^2}{n-2}}.$$

Large values of s tell us the regression gives imprecise predictions; small values tell us predictions are more accurate. The value of s tells us how far from the predicted values the observed data may be.
(Optional: We divide by $n - 2$ because there are $n - 2$ degrees of freedom.)