

**Class 24: Regression and Hypothesis Testing (Text: Sections 10.1)****Regression**

The notation for the *sample regression line* is

$$\hat{y} = b_0 + b_1x$$

The “hat” on the  $\hat{y}$  means it was estimated from the data. The coefficients in the sample regression equation are given by

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

Thus the point  $(\bar{x}, \bar{y})$  is on the sample regression line. The *population regression line* is written as

$$\mu_y = \beta_0 + \beta_1x$$

(As before, Greek letters are population parameters; ordinary letters are the sample statistics.)

The  $\mu_y$  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 are  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ . 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

$$y_i = \beta_0 + \beta_1x_i + \varepsilon_i$$

where the  $\varepsilon_i$  are random variables, called *errors* or *residuals*. We assume

- Errors are normally distributed with mean 0 and standard deviation  $\sigma$ .
- Errors for different values of  $x$  are independent and the standard deviation  $\sigma$  is the same for all values of  $x$ .

The *population parameters* for this model are  $\beta_0, \beta_1$ , and  $\sigma$ .

**Hypothesis Test: 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,

$$\mu_y = \beta_0.$$

Since the value of  $\mu_y$  no longer depends on  $x$ , there is *no* relationship.

Similarly, if  $\beta_1 \neq 0$ , then values of  $\mu_y$  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,  $\varepsilon_i$ , have standard deviation  $\sigma$ , assumed independent of  $x$ . To estimate  $\sigma$ , we use an average of the squared residuals, giving the *standard error* listed in the Regression Statistics:

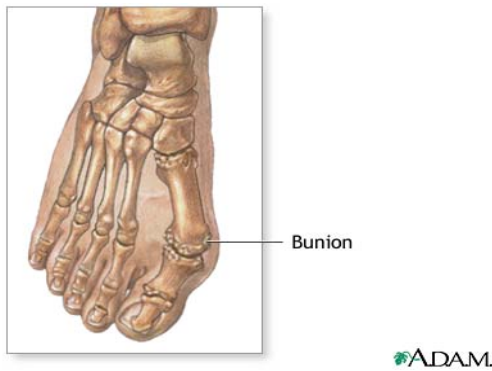
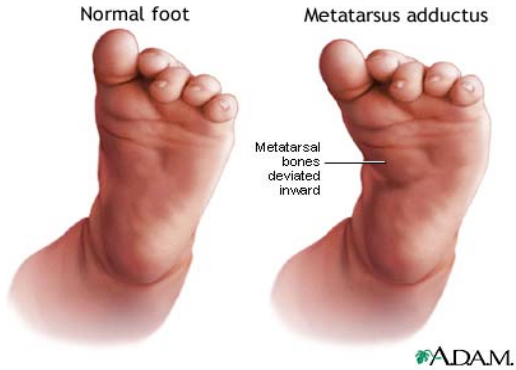
$$s = \sqrt{\frac{\sum \varepsilon_i^2}{n - 2}}.$$

Small values of  $s$  means the regression predicts well; large values of  $s$  mean worse predictions. (Optional: We divide by  $n - 2$  because there are  $n - 2$  degrees of freedom.)

**Foot Problems**

MA is a foot problem that usually corrects itself. (The front part of the foot is turned.) HAV is a more serious problem (deformation of the big toe) that usually requires surgery.

<http://www.nlm.nih.gov/medlineplus/ency/imagepages/9052.htm> and [http://www.nlm.nih.gov/medlineplus/ency/presentations/100005\\_2.htm](http://www.nlm.nih.gov/medlineplus/ency/presentations/100005_2.htm).



Severity of HAV	Severity of MA
28	18
32	16
25	22
34	17
38	33
26	10
25	18
18	13
30	19
26	10
28	17
13	14
20	20
21	15
17	16
16	10
21	7
23	11
14	15
32	12
25	16
21	16
22	18
20	10
18	15
26	16
16	10
30	12
30	10
20	10
50	12
25	25
26	30
28	22
31	24
38	20
32	37
21	23

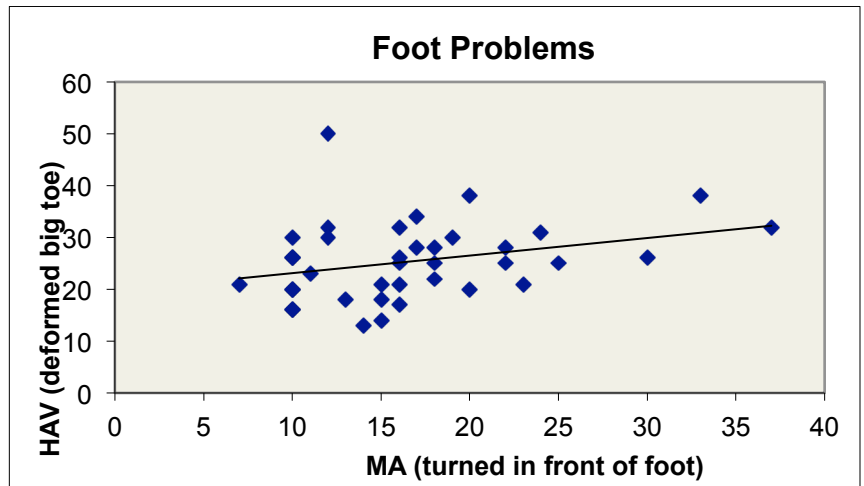
**Ex:** Numerical measurements of the severity of MA and HAV in patients who had both conditions are on the right. We use the MA readings to predict HAV.

- What is the equation of the regression line? Interpret the slope.
- What is the distribution of the test statistic for the slope coefficient?
- Fill in the missing values in the regression table on the next page.
- Is there evidence that MA can be used to predict HAV? Include the null and alternate hypotheses. Use 1%, 5% and 10% significance levels.
- Do we have evidence that MA causes HAV?

SUMMARY OUTPUT						
<i>Regression Statistics</i>						
Multiple R	0.30213679					
R Square						
Adjusted R Square	0.0660446					
Standard Error	7.22370645					
Observations	38					
<i>ANOVA</i>						
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
Regression	1	188.713503	188.713503			
Residual	36	1878.54965	52.1819349			
Total	37	2067.26316				
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	19.7232673	3.21716807	6.13063008	4.6513E-07	13.1985567	26.247978
MA	0.33883543	0.17817527			-0.0225203	0.70019115

**Ex:** We use the MA readings to predict HAV.

- What is the equation of the regression line? Interpret the slope.
- What is the distribution of the test statistic for the slope coefficient?
- Fill in the missing values in the regression table on the next page.
- Is there evidence that MA can be used to predict HAV? Include the null and alternate hypotheses. Use 1%, 5% and 10% significance levels.



**Correlation Coefficient  $r$  and Coefficient of Determination  $r^2$** 

The *correlation coefficient*,  $r$ , tells us how close the data is to the regression line. Writing  $s_x$  for the standard deviation of the  $x$ -values and  $s_y$  for the standard deviation of the  $y$ -values, the formula for the correlation coefficient is

$$r = \frac{1}{n-1} \sum_{i=1}^{i=n} \frac{(x_i - \bar{x})}{s_x} \cdot \frac{(y_i - \bar{y})}{s_y}.$$

**Ex: What is the relation between the sign of  $r$  and the slope of the regression line?**

*Coefficient of determination*,  $r^2$  or  $R^2$ , tells us how much of the variation in the  $y$ -values is predicted by the regression line. That is

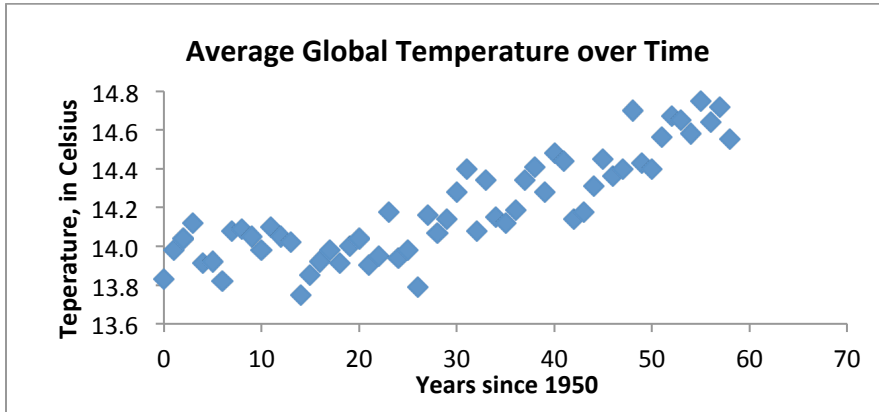
$$R^2 = \frac{\text{Variation in predicted } \hat{y} \text{ values from the mean } \bar{y}}{\text{Variation in observed } y \text{ values from the mean } \bar{y}} = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2}$$

**Ex: Find  $R^2$  in the foot problems table. Interpret it.**

**Ex: Find the  $F$ -statistic of the regression and its  $P$ -value. What do they tell us?**

**Ex: What does the standard error in the regression statistics tell us?**

**Climate Change: What Does the Data Say?<sup>1</sup> Average global temperature has been rising:**



SUMMARY OUTPUT		Temperature and Year				
<i>Regression Statistics</i>						
Multiple R	0.853943					
R Square	0.729218					
Adjusted R Square	0.724467					
Standard Error	0.139817					
Observations	59					
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	13.81171	0.035947	384.2215	5.5E-99	13.73973	13.8837
Years After 1950	0.013243	0.001069	12.38957	8.27E-18	0.011103	0.015384

How fast has temperature been rising?

What does the  $R^2$  value tell us?

What does  $P$ -value tell us?

**Carbon dioxide levels have been rising: Keeling curve**

SUMMARY OUTPUT		CO <sub>2</sub> and Year				
<i>Regression Statistics</i>						
Multiple R	0.987014					
R Square	0.974197					
Adjusted R Square	0.973744					
Standard Error	3.664276					
Observations	59					
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	303	0.942094	321.6238	1.38E-94	301.1135	304.8865
Years After 1950	1.299525	0.028013	46.38969	5.76E-47	1.24343	1.355621

How fast is CO<sub>2</sub> level rising?

What does the  $R^2$  value tell us?

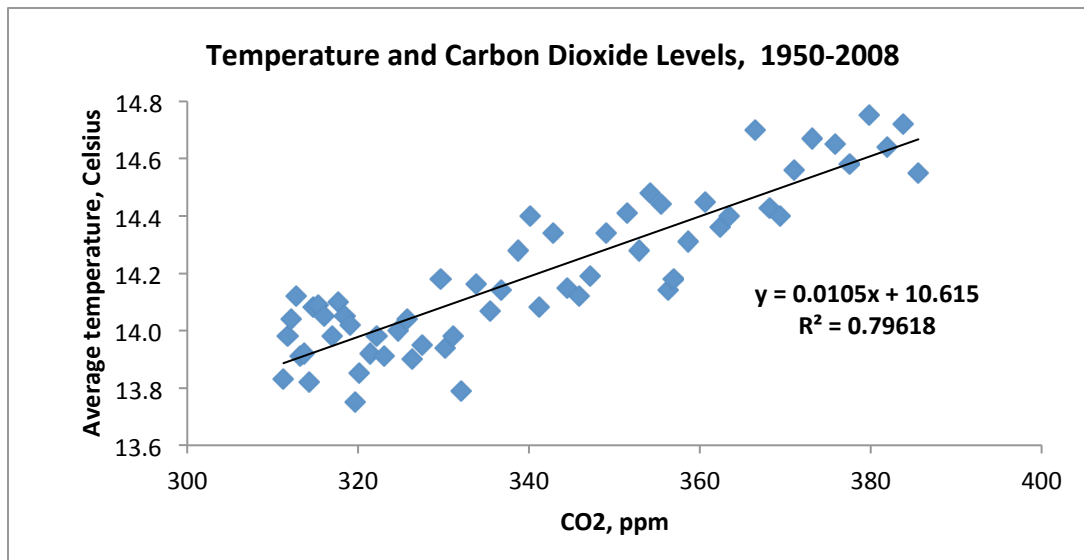
What does  $P$ -value tell us?

<sup>1</sup> <http://data.giss.nasa.gov/gistemp/tabledata/GLB.Ts.txt> and <http://www.esrl.noaa.gov/gmd/ccgg/trends/index.html#mlo>

**Does this data show that the rise in CO<sub>2</sub> levels caused the increase in temperature?**

**To show causation**, you need a scientific mechanism that explains how CO<sub>2</sub> affects temperature. Otherwise a randomized experiment—not possible here—or more sophisticated regression techniques.

SUMMARY OUTPUT		CO <sub>2</sub> and Temperature				
<i>Regression Statistics</i>						
Multiple R	0.892288					
R Square	0.796178					
Adjusted R Square	0.792602					
Standard Error	0.121304					
Observations	59					
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	10.61512	0.240482	44.14112	9.07E-46	10.13357	11.09668
CO2 parts per million	0.01051	0.000704	14.92166	2.42E-21	0.0091	0.011921



How fast is temperature been rising?

What does the  $R^2$  value tell us?

What does  $P$ -value tell us?

**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, Regression; fill in the dialog box.