Math 146 10.2 - Regression

A regression equation is an equation that best models bivariate data, meaning it can be used as a predictor or estimator of the dependent data.

A **linear regression** equation is a linear equation, such as $y = mx + b$, that best fits a collection of paired data.

NOTATION

Linear Regression Equation for a *sample*: $\hat{y} = b_0 + b_1 x$, (b_0 is the y-intercept, b_1 is the slope) ($\hat{y}$ is called y-hat)

Linear Regression Equation for the *population*: $y = \beta_0 + \beta_1 x$

REQUIREMENTS

1) The sample of paired (x, y) data is a random sample.
2) A linear relation is confirmed using a scatterplot.
3) Outliers have a strong effect on the regression equation. Try and determine if an outlier is a known error or not a known error.

FORMULAS

Slope: $b_1 = r \frac{s_y}{s_x}$ **y-intercept:** $b_0 = \overline{y} - b_1 \overline{x}$

EXAMPLE 1 Use our data from 10.1 and the above formulas to calculate the linear regression equation. Create a scatter plot and graph the regression line. See the text for other ways of calculating the regression equation.

x	1	2	3	4	5
y	2	4	5	8	8

EXAMPLE 2 **Predicting Temperature From Cricket Chirps**

The following data is the number of cricket chirps in 15 seconds, and the outside temperature at that time:

Cricket Chirps	20	16	20	18	17	16	15	17	15	16	15	17	16	17	14
Temperature (F)	89	72	93	84	81	75	70	82	69	83	80	83	81	84	76

a) Make a scatter plot to determine if there is a linear correlation, and to see if there are any outliers affecting our calculations.

b) Calculate the correlation coefficient and regression equation.

c) Explain the meaning of r^2 in context of the problem.

d) Use a t-test to determine if the correlation is statistically significant.

e) For what values is the regression equation most reliable for?

f) What is the best predicted temperature for 19 chirps per 15 seconds?

g) What is the estimated number of chirps one would expect to record at a temperature of 80°F?

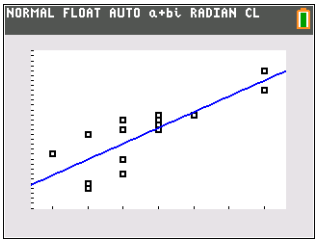
h) **EXTRA** If we switch the data variables (LinReg L_2, L_1) and find a new regression equation to predict the number of chirps from temperature we get a slightly different predicted value than we get in part (g). Why?

EXTRA EXTRA Using a single point estimate prediction for $\hat{y}$ does not come with a lot of confidence. However, we can create a *prediction interval* (similar to a confidence interval. See Section 10.3.) to calculate a range of values for our prediction based on a confidence level. This gives us a better idea of possible values for our prediction of $\hat{y}$, or more precisely, an estimate of what the population mean $\overline{y}$ is.

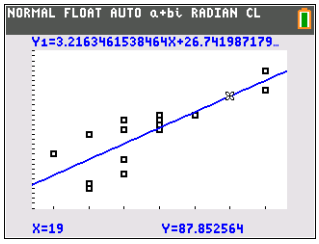
How To Calculate A Prediction Interval:

Let $SE_{\hat{y}} = s\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{s_x^2 (n-1)}}$ (this is called the standard error of the prediction), where x_0 is our new x data value, and s is the standard error of the regression line, and is calculated when doing **LinRegTTest**. The prediction interval for $\overline{y}$ is: $\hat{y} \pm t_{\alpha/2} \cdot SE_{\hat{y}}$ for a $1 - \alpha$ confidence level. These values can be found in the VARS Statistics menus. **Note**: the degrees of freedom for finding $t_{\alpha/2}$ is $df = n - 2$. The last two screen shots show the prediction intervals for a 90% prediction interval.

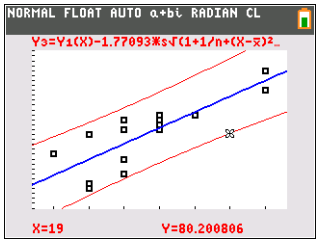
Linear Regression Line



Prediction for $x = 19$



Lower Prediction Level



Upper Prediction Level