Author: Brenda Gunderson, Ph.D., 2015

License: Unless otherwise noted, this material is made available under the terms of the Creative Commons Attribution-NonCommercial-Share Alike 3.0 Unported License: http://creativecommons.org/licenses/by-nc-sa/3.0/

The University of Michigan Open.Michigan initiative has reviewed this material in accordance with U.S. Copyright Law and have tried to maximize your ability to use, share, and adapt it. The attribution key provides information about how you may share and adapt this material.

Copyright holders of content included in this material should contact open.michigan@umich.edu with any questions, corrections, or clarification regarding the use of content.

For more information about how to attribute these materials visit: http://open.umich.edu/education/about/terms-of-use. Some materials are used with permission from the copyright holders. You may need to obtain new permission to use those materials for other uses. This includes all content from:

Attribution Key

For more information see: http:://open.umich.edu/wiki/AttributionPolicy

Content the copyright holder, author, or law permits you to use, share and adapt:

Creative Commons Attribution-NonCommercial-Share Alike License

Public Domain – Self Dedicated: Works that a copyright holder has dedicated to the public domain.

Make Your Own Assessment

Content Open.Michigan believes can be used, shared, and adapted because it is ineligible for copyright.

Public Domain – Ineligible. Works that are ineligible for copyright protection in the U.S. (17 USC §102(b)) *laws in your jurisdiction may differ.

Content Open.Michigan has used under a Fair Use determination
Fair Use: Use of works that is determined to be Fair consistent with the U.S. Copyright Act (17 USC § 107) *laws in your jurisdiction may differ.

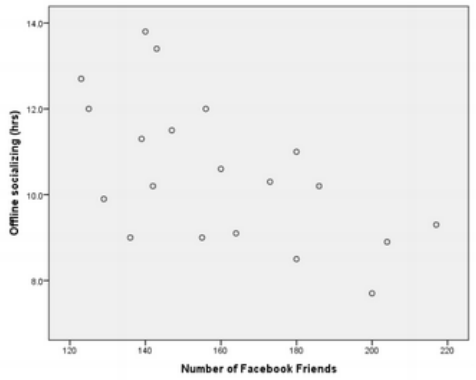
Our determination DOES NOT mean that all uses of this third-party content are Fair Uses and we DO NOT guarantee that your use of the content is Fair. To use this content you should conduct your own independent analysis to determine whether or not your use will be Fair.

Lab 10: Exploring Linear Regression

Objective: In this lab, you will examine relationships between two quantitative variables using a graphical tool called a scatterplot. You will interpret scatterplots in terms of form, direction, and strength of the relationship, and use it to assess the appropriateness of using a linear regression model to describe the relationship between the two variables. If appropriate, you can then perform a linear regression analysis to produce an estimated model that can be used to predict the value of the response *y* for a given value of the predictor *x*.

Application: Pam believes that the number of Facebook friends a person has could interfere with the number of hours they spends socializing offline. She wants to be able to predict the number of hours a person spends socializing offline for a given number of Facebook friends. She will collect data from her friends and examine the relationship between these two variables. Pam will be able to fit a linear regression model if the relationship between these variables is linear.

Overview: A regression model describes how the mean of one variable is thought to depend on the value of one or more other variables. If we think the number of Facebook friends may explain changes in the amount of time spent socializing online, we call the number of Facebook friends an **explanatory** variable (or **predictor** variable or **independent** variable) and the amount of time spent socializing offline is called the **response** variable (or **dependent** variable).



To start, we use a scatterplot to display the relationship between two quantitative variables, plotting the number of Facebook friends on the *x*-axis and the amount of time spend socializing offline on the *y*-axis. In examining the relationship, Pam looks at the overall pattern showing the **form**, which appears to be linear. She notes that the **direction** of the form is negative and **strength** of the relationship is moderate. She notes that there are no apparent **outliers**.

One of the many misconceptions about regression arises from the concept of association. Scatterplots can show the association between variables, but Pam should remember that <u>*correlation does not imply causation*</u>. For example: weekly flu medication sales and weekly sweater sales for an area with extreme seasons would exhibit a positive association because both tend to go up in winter and down in summer. However, neither causes the other. The observed association between two variables is sometimes due to other factors, such as confounding variables.

This correlation value, **r**, explains the strength of the linear relationship between x and y. The correlation can take on values between -1 and 1. The sign of the correlation also describes the direction of the linear relationship. The correlation between number of Facebook friends and time spent socializing offline is reported to be -0.59, which confirms what Pam observed in her scatterplot. The square of this correlation is known as the coefficient of determination and also has an important interpretation. Pam calculates this coefficient to be 0.348 and she can interpret this value by stating that 34.8% of the variation in the amount of time spent socializing online can be explained by the linear relationship between number of Facebook friends and time spent socializing online.

Since Pam's scatterplot suggested that the dependence of amount of time spent socializing online on the number of Facebook friends can be summarized by a straight line, the **least squares regression line** can be calculated. The least squares regression line is the line that minimizes the sum of the squared vertical distances of the data points to the line – hence the name **least squares**. This fitted line can be used to describe the linear relationship between the amount of time spent socializing offline and the number of Facebook friends and to predict the amount of time spent socializing online for a given number of Facebook friends.

The distances from the observed amount of time spend socializing offline to the predicted amount of time spent socializing offline are known as the **residuals**. These residuals are estimates of the true error terms associated with Pam's model.

Pam has fit a **linear regression model,** which is what will also be fit during the lab. Pam's regression is based on a linear model relating the number of Facebook friends to the mean amount of time spend socializing offline as follows: E(*time spent socializing offline*) = $\beta_0 + \beta_1$(*number of Facebook friends)*. Here β_0 and β_1 are parameters – fixed but unknown constants. Specifically, β_0 is the population *y*-intercept (the amount of time spent socializing offline when the number of Facebook friends is zero) and β_1 is the population slope (the change in the mean time spent socializing offline for every additional Facebook friend). These two values are unknown, but can be estimated using the least squares criterion. The resulting estimated regression line is generally written as: $\hat{y} = b_0 + b_1(x)$. The estimates, b_0 and b_1 are referred to as the **least squares estimates** of β_0 and β_1.

Formula Card:

Regression

Linear Regression Model

Population Version:

Mean: $\mu_Y(x) = E(Y) = \beta_0 + \beta_1 x$

Individual: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

where ε_i is $N(0, \sigma)$

Sample Version:

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

Individual: $y_i = b_0 + b_1 x_i + e_i$

Parameter Estimators

$$b_1 = \frac{S_{XY}}{S_{XX}} = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sum (x-\bar{x})^2} = \frac{\sum (x-\bar{x})y}{\sum (x-\bar{x})^2}$$

$$b_0 = \bar{y} - b_1\bar{x}$$

Residuals

$e = y - \hat{y}$ = observed y – predicted y

Correlation and its square

$$r = \frac{S_{XY}}{\sqrt{S_{XX}S_{YY}}}$$

$$r^2 = \frac{SSTO - SSE}{SSTO} = \frac{SSREG}{SSTO}$$

where $SSTO = S_{YY} = \sum (y-\bar{y})^2$

Estimate of σ

$$s = \sqrt{MSE} = \sqrt{\frac{SSE}{n-2}} \quad \text{where } SSE = \sum (y-\hat{y})^2 = \sum e^2$$

Standard Error of the Sample Slope

$$\text{s.e.}(b_1) = \frac{s}{\sqrt{S_{XX}}} = \frac{s}{\sqrt{\sum (x-\bar{x})^2}}$$

Confidence Interval for β_1

$b_1 \pm t^*\text{s.e.}(b_1)$ df $= n-2$

***t*-Test for β_1**

To test $H_0 : \beta_1 = 0$

$$t = \frac{b_1 - 0}{\text{s.e.}(b_1)} \quad \text{df} = n-2$$

or $F = \frac{MSREG}{MSE}$ df $= 1, n-2$

Confidence Interval for the Mean Response

$\hat{y} \pm t^*\text{s.e.(fit)}$ df $= n-2$

where $\text{s.e.(fit)} = s\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{XX}}}$

Prediction Interval for an Individual Response

$\hat{y} \pm t^*\text{s.e.(pred)}$ df $= n-2$

where $\text{s.e.(pred)} = \sqrt{s^2 + (\text{s.e.(fit)})^2}$

Standard Error of the Sample Intercept

$$\text{s.e.}(b_0) = s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}}}$$

Confidence Interval for β_0

$b_0 \pm t^*\text{s.e.}(b_0)$ df $= n-2$

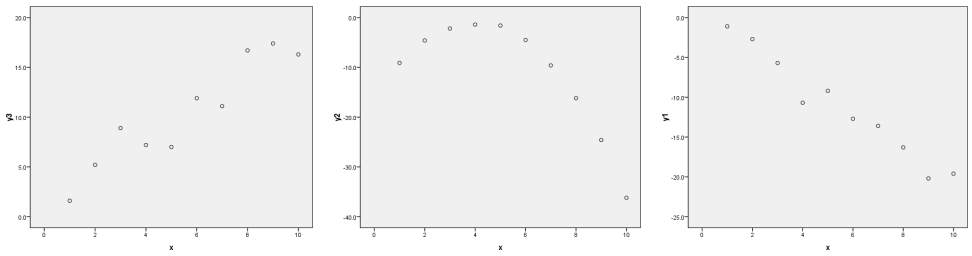
***t*-Test for β_0**

To test $H_0 : \beta_0 = 0$

$$t = \frac{b_0 - 0}{\text{s.e.}(b_0)} \quad \text{df} = n-2$$

Warm-Up: Check Your Understanding

For which scatterplot(s) does a linear regression analysis seem appropriate?



Plot 1 **Plot 2** **Plot 3**

ILP: Using a Scatterplot and Correlation to Analyze the Relationship

Background: As we approach the final exam, it may be of interest to you to know if there is a relationship between your existing exam and homework scores and your final exam score. To analyze this, we have a sample of exam and homework scores from Stats 250 students in a previous term, found in the **Stats250Data.Rdata** file.

***Task*:** First, we would like to determine what measure predicts Final Exam scores best – Homework, Exam 1, or Exam 2? To do this, we will examine scatterplots and correlations to ensure that a linear relationship is appropriate, and determine which model is strongest.

Procedure: We will have **two variables** being measured – Final Exam scores (which is **quantitative** and playing the role of the **response**) and your chosen predictor variable (which is also **quantitative** and plays the role of the **explanatory** variable). The goal is to assess the **relationship** between these two quantitative variables. The appropriate inference procedure for this scenario is **linear regression**.

1. Open the data set and produce a scatterplots for each of the potential predictor variables (AvgHW, Exam1, Exam2) against the Final Exam score, using **Graphs > Scatterplot**. Under **Options**, you will want to uncheck **Show Spread** and **Smooth Line**. Write a few sentences to describe each scatterplot in terms of:
    A. form (does there appear to be a linear relationship?)
    B. direction (positive or negative)?
    C. strength (generally weak, moderate, or strong?)
    D. outliers (any unusual observations or outliers present?)

    Interpret the Scatterplot for Final Exam versus AvgHW

Interpret the Scatterplot for Final Exam versus Exam 1

Interpret the Scatterplot for Final Exam versus Exam 2

2. Let's now assess the strength of the relationship via the correlation, or R (not the software, but the value!). To get R values for each pair of variables, select **Statistics > Summaries > Correlation Matrix**. This will display an array of values, where each value is the correlation of the variables given in the row and column headers. Report the following correlation values:

AvgHW and Final Exam:_____ Exam 1 and Final Exam:_____ Exam 2 and Final Exam:_____

3. Based on the scatterplots and correlation values, which variable will you use to predict Final Exam scores? Explain.

AvgHW Exam 1 Exam 2

ILP: Describing a Linear Relationship with a Regression Line

Now that we have chosen an appropriate linear model to predict final exam score, we are ready to estimate that model and use it to predict Final Exam scores.

***Task*:** Fit a linear model to the data. If you have questions about the regression output after the activity, refer to Supplement 6 in this workbook for more details.

1. Obtain the linear regression output first by creating a model using **Statistics > Fit Models > Linear Regression**, making sure to enter the appropriate response and explanatory variables. Give your model an appropriate name, and click OK. A summary of the model will appear in the output. Report the estimated regression line (the predicting equation or least squares regression line):

   _____

2. Interpret the estimated slope b_1 in terms of the change in your explanatory variable.

3. Report the coefficient of determination, r^2, and interpret it: r^2 = _____
   Interpretation:

4. Use your regression line to predict the Final Exam score for a student with score given in the table for your predicting (explanatory) variable (include units).

Variable	Value
Exam 1	58
Exam 2	50
Avg HW	25.6

5. The student above actually had a final exam score of 92 out of 100 points. What is this student's residual (include units)?

Cool-Down: Check Your Understanding

Could you use this regression line to predict your final exam score in this class?

What are some potential issues with the prediction this line gives you?