Introduction to Simple Linear Regression

STAT 215
Fall 2013

Modeling - Overview

- In many research studies several variables are observed simultaneously and the intent is to study their relationships.

- Recall:
  - *Explanatory*: Variable that is thought to affect ("explain") another variable
  - *Response*: Variable that is thought to be affected by ("respond to") the explanatory variable(s)

- Different statistical methods exist, to describe relationships between one or more explanatory variables and a response.
  - Which method to use depends on how many variables there are and whether they are *quantitative* or *categorical*.

Statistical Methods - Overview

- One-sample z or t-test (Chapter 8)
  - Compare a population mean to some theoretical value
  - One quantitative variable
- Two-sample z or t-test (Chapter 9)
  - Models relationship between
  - One categorical explanatory variable (with two categories)
  - One quantitative response
- One-Way ANOVA (Chapter 10)
  - Models relationship between
  - One categorical explanatory variable (with more than two categories)
  - One quantitative response

Statistical Methods - Overview

- Simple Linear Regression (Chapter 12)
  - Models *linear* relationship between
  - One quantitative explanatory variable
  - One quantitative response variable

- Multiple Linear Regression (Chapter 13)
  - Models *linear* relationship between
  - Several quantitative explanatory variables
  - One quantitative response variable

Simple Linear Regression (SLR)

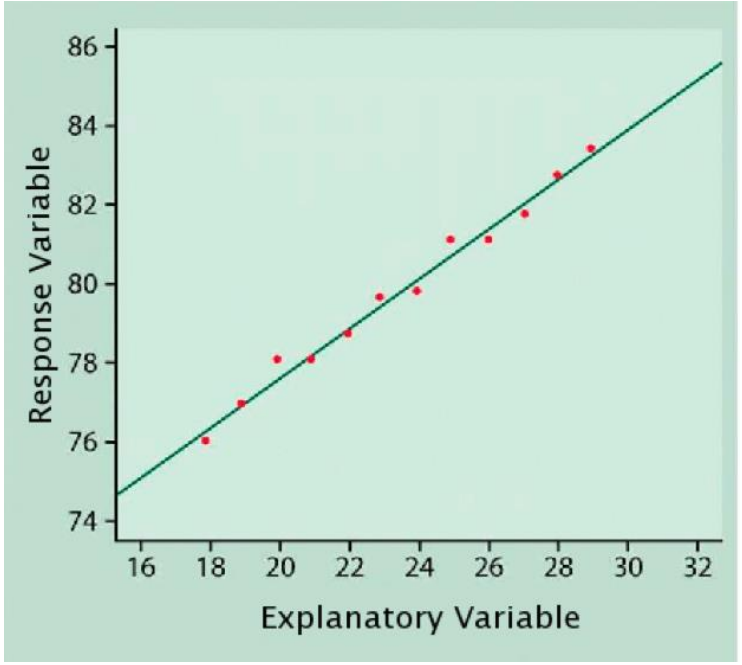
- Suppose we have observations on pairs of quantitative variables (one explanatory, one response).

- Example:
  - Height and weight of patients
  - Duration of exercise and heart rate
  - Temperature and soil moisture
  - Amount of water a plant receives and height of plant

- We want to investigate the *linear* relationship between these variables

Simple Linear Regression (SLR)

- The graphical display for two quantitative variables is called a *scatterplot*.
- Usually the explanatory variable is plotted on the x-axis and the response on the y-axis.
- A measure of association for two quantitative variables is the *correlation coefficient* r.
- If two variables are strongly associated, the points in their scatter plot will lie close to a line.



Least Squares Regression

- Even if the association between two variables is high, we cannot expect all points to lie exactly on a line - there is always variation!
- But we can use a line to “model” the relationship between the variables.
- Which straight line approximates a given set of data points “best”?
- A standard approach for estimating the line is called *Least Squares Regression*.
- The line is chosen, so that the sum of squared vertical distances between points and the line is minimized.

Vocabulary

- The (x,y)-coordinates actually measured in the experiments are the ***observed*** values. These represent the available data for the study.

  Notation: y_i - Observed response for individual i

  x_i - Value of explanatory variable for individual i

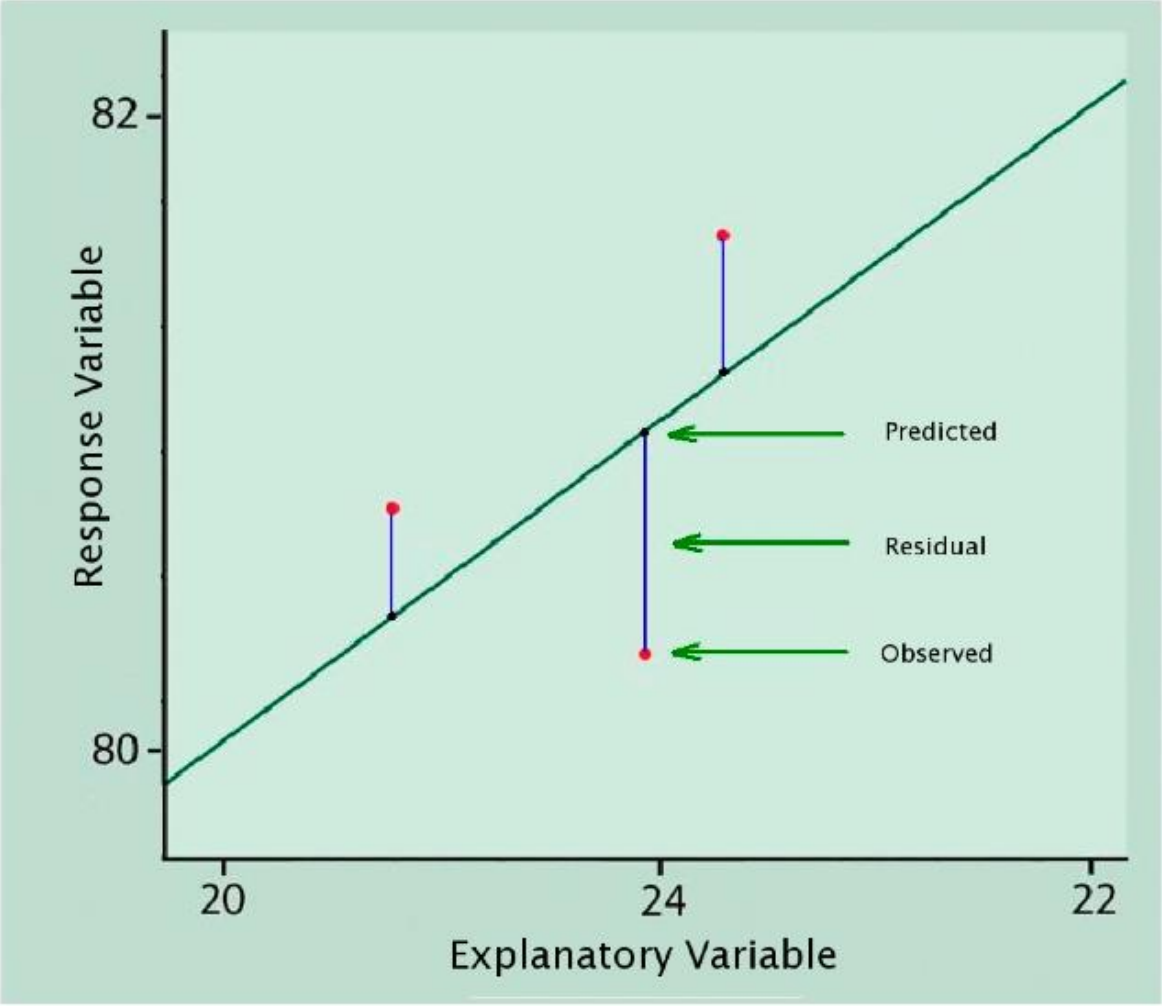
- The points that lie on the regression line vertically above (or below) an observed value are called the ***predicted or fitted*** values.

  Notation: $\hat{y}_i$

- The (vertical) distances between observed and predicted values are called the ***residuals***.

  Notation: $e_i = y_i - \hat{y}_i$

Vocabulary


82
80
Response Variable
20
24
22
Explanatory Variable
Predicted
Residual
Observed

SLR Model

Statement of Model:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \text{ where } \begin{cases} i = 1,2,\ldots,n \\ \varepsilon_i \sim N(0,\sigma^2) \end{cases}$$

- Remember y_i, x_i are our observed values for the response and explanatory variables for individual i

- Model Parameters (unknown)
  - β_0= intercept
  - β_1 = slope
  - σ^2 = error variance

Assumptions

- *Linear* relationship between X and Y

- Model assumes that the *error terms* are
  - Independent
  - Normal
  - Have constant variance.

- *Residuals* may be used to explore the legitimacy of these assumptions.

Interpretation of Regression Coefficients

- We call the intercept (β_0) and slope (β_1) parameters the *regression coefficients*

- The intercept (β_0) is the average value of the response variable when X = 0.

- The slope (β_1) represents the increase (or decrease if negative) in the mean response for a 1-unit increase in the value of X.

Fitted Regression Line

- The parameters $\beta_0, \beta_1, \sigma^2$ must be estimated from the data
- Estimates denoted: $\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2$
- Fitted (or estimated) regression line is:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

- The “hat” notation denotes an estimate of a parameter (or a fitted value for the response)

Correlation Coefficient (r)

- Describes the *strength* and *direction* of the *linear* relationship between the explanatory and response variables
- We discussed this in chapter 5.
- Ranges between -1 and +1
- Sign: Positive or negative slope (relationship)
- Strength: The closer to ± 1, the stronger the linear relationship.
  - The points in the scatter plot will lie close to the regression line.
  - http://istics.net/stat/correlations/

Coefficient of Determination (r^2)

- The square of the correlation coefficient (r^2) is the proportion of variation in the response variable y that is explained by the regression of *y* on *x*.

- Typical to multiply by 100 and express as a percentage (the percentage of variation explained by the model).

Example: Old Faithful

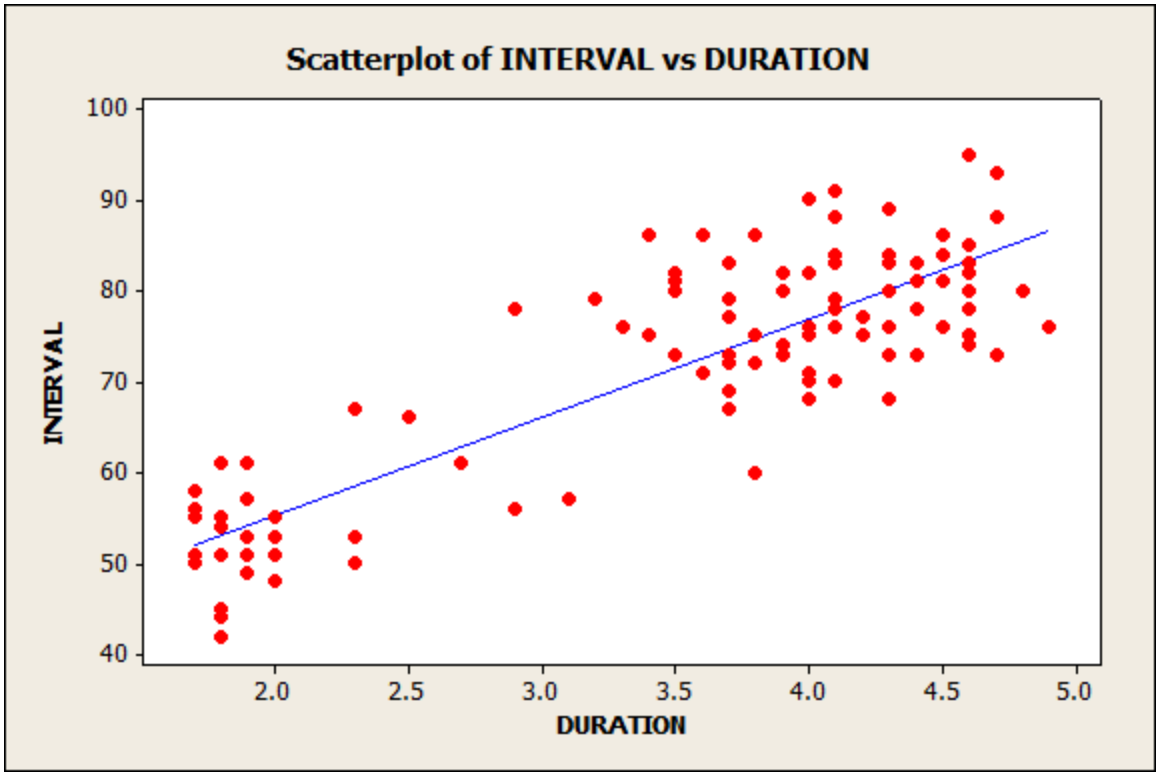
- Can we predict when the next eruption of old faithful will be based upon the duration of the previous eruption?

- Historical data:
- A park geologist collected measurements over an 8 day period from August 1-8, 1978
- Data on n=107 eruptions during this time period

- Variables measured:
  - Duration of eruption in minutes
  - Interval of time (in min) until the next eruption starts

- Which is the response? Which is the explanatory variable?

Example: Old Faithful Scatterplot



- Direction of relationship? Strength?

Example : Old Faithful Minitab Output

Regression Analysis: INTERVAL versus DURATION

The regression equation is

INTERVAL = 33.8 + 10.7 DURATION

Predictor	Coef	SE Coef	T	P
Constant	33.828	2.262	14.96	0.000
DURATION	10.7410	0.6263	17.15	0.000

S = 6.68261 R-Sq = 73.7% R-Sq(adj) = 73.4%

Example : Old Faithful Model Interpretation

- From the output we see that:
  - $\hat{\beta}_0$= 33.828
  - $\hat{\beta}_1$=10.741
  - r^2=0.737 or 73.7%, r=+$\sqrt{r^2}$=0.858 (+ since relationship is positive)
- So the fitted regression line is: $\hat{y} = 33.828 + 10.741x$
- 73.7% of the variation in interval of time until the next eruption is explained by the duration of the previous eruption.
- There is a strong positive linear relationship between interval between interval and duration.

Example: Old Faithful Slope Parameter

- Slope interpretation: As the eruption duration increases by 1 minute, the interval until the next eruption is expected to increase by 10.741 minutes

- To test whether there is a significant linear relationship between the response and the explanatory variable, we test whether the slope coefficient is zero (null hypothesis) or not (alternative hypothesis).

- If we have evidence that the slope is not zero (reject null) this means that there is a significant linear relationship between the two variables.

- The p-value for the slope here is <0.001. This means we have sufficient evidence that the slope is nonzero and that there is a significant linear relationship between interval and duration.

Example: Old Faithful Intercept Parameter

- Intercept interpretation: For a duration of 0 minutes, the average interval until the next eruption will be 33.828 minutes.

- Does this make sense?

Scope of Model

- The *scope of a regression model* is the range of X-values over which we actually have data.
- Predicting the values of the mean response for X-values within this range is called *interpolation*
- *Example:* Predicting average interval until next eruption for a previous duration of 2 minutes

Scope of the Model

- Using a model to look at X-values outside the scope of the model is called *extrapolation*
- Example: Predicting average interval until next eruption for a previous duration of 0 minutes or 10 minutes
- Extrapolation is dangerous since the model may not be valid over a wider range of values for the explanatory variable.
- Often, this is why we may not necessarily be interested in the interpretation of the intercept.

Example: Old Faithful Prediction

- According to our model, what is the predicted value for the interval until the next eruption if the previous duration is 2 minutes?

$$\hat{y} = 33.828 + 10.741 * 2 = 55.31 \text{ minutes}$$

- We can also calculate a “prediction interval” to give us a range of plausible values for when the next eruption will be with a certain level of confidence.

Final Note: Remember Correlation Does not Imply Causation!

- For regression models it is tempting to say that the explanatory variable “causes” changes in the response variable.
- But, we can only discuss causation if we have an experiment where randomization was employed.
- Otherwise, we can say there is an “association” at best and must be aware of possible confounding variables.
- http://xkcd.com/552/