Simple Linear Regression Using R

Lu Qian, MS
Cancer Epidemiology
Moffitt Cancer Center, Tampa FL

November 7, 2018
ASA USF Student Chapter

MOFFITT CANCER CENTER

Outlines

- Review: Hypothesis Testing
- Explanatory Analysis
  - Data Transformation
- Simple Linear Regression
- ANOVA F Test

Hypothesis Testing

Review

One Mean	Two Means (unpaired)	Two Means (paired)	Several Means
t-test (N<30)	t-test	Paired t test	ANOVA
z-test (N>=30)	z-test	z-test	
Chi-square test	F-test		
Wilcoxon Signed Rank Test		Wilcoxon Signed Rank Test with continuity correction	

Simple Linear Regression

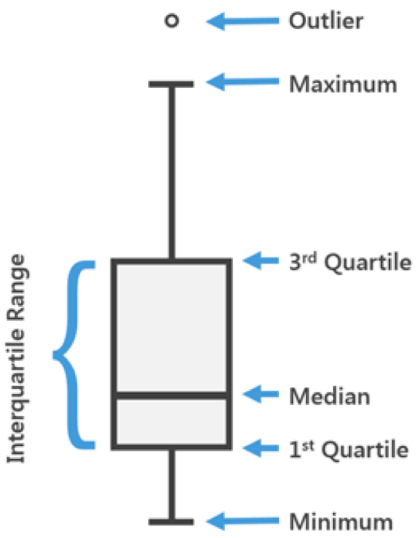
- In general, we have n pairs of sample points (X_1,Y_1), $(X_2,Y_2),\ldots,(X_n,Y_n)$, where *X* is referred to as the explanatory variable and Y is the response variable.
- In the case of a linear trend (i.e., the value of one outcome tends to increase or decrease linearly with an increase in another variable), we can fit a model that describes that trend.
- Note that this does NOT imply that *X* necessarily ***causes*** Y, although that is possible. X and Y may simply be associated, without any causative effect. We typically want to explain changes in the average of *Y* due to a difference in *X*.

Explanatory Analysis

- Use summary statistics and univariate charts (e.g., histograms, boxplots) to understand their marginal or individual distributions.
- We're interested in how the variables relate to each other, a key subsequent step is to construct a two-dimensional ***scatterplot,*** treating *Y* as a function of *X*.

Graphical Analysis

- Scatter Plot
  - Visualize linear relationship between dependent (response) variable and independent (predictor) variable
- Box Plot
  - Check for outliers



Picture credit: https://pro.arcgis.com/en/pro-app/help/analysis/geoprocessing/charts/box-plot.htm

MOFFITT CANCER CENTER

Data Transformation

- When the relationship between the response and explanatory variables does not show linear
- Used to correct violations of model assumptions such as constant error variance and normality
- Common transformation techniques:
  - Log transformation
  - Power transformation
  - Box-cox transformation

The Model

- If *X* and *Y* appear to be *linearly* associated, where *Y* on average increases or decreases linearly with an increase in *X*, then we may posit the **linear model**

$$\boldsymbol{Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i}$$

- where the intercept β_0 and slope β_1 determine the line, and ε_i models the variability around the line.

The Error Term

- The error term ε_i in the model accounts for this variability around the line $\beta_0 + \beta_1 X_i$.
- Note that $\beta_0 + \beta_1 X_i$ is fixed, not random. We further typically assume that $\varepsilon_i \sim N(0, \sigma^2)$.
- In other words, given the value X_i, we have
  - $E(Y_i) = \beta_0 + \beta_1 X_i$, and $Var(Y_i) = \sigma^2$.
  - Therefore,

$$Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$$

The Model Parameters

- The intercept β_0 represents the average of Y for X = 0.
- The slope β_1 is generally the focus of inference: it represents the change in the average of Y for every one unit increase in X.
- Since we are interested in how Y changes with X, then a nonzero slope indicates that Y and X are linearly associated.
- The variance term σ^2 represents the variability of the data around the line.

Residual

- Definition: the distance from a given value Y_i and its associated point on the line is given

$$Y_i - (\beta_0 + \beta_1 X_i)$$

- We compute estimates of the slope, intercept, and variance that minimize the sum of the squared residuals. The resulting estimates of the slope and intercept are given by

$$\beta_1 = \frac{\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}}{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2}, \beta_0 = \bar{Y} - \overline{\beta_1 X}$$

Interpreting the Model Fit

- Fitted value for Y_i is given by

$$\hat{Y}_i = \beta_0 + \beta_1 X_i$$

- There are two ways of viewing such a fitted value
  - the fitted value is our **predicted** Y_i of the given X_i
  - The fitted value is our estimate of the **average** Y_i of the given X_i

ANOVA F Test for Regression Coefficients

ANOVA approach provides a useful way of testing coefficients and comparing models in a variety of setting (particularly multiple regression with several variables).

For simple linear regression, the ANOVA *F* statistic for testing $H_0\colon \beta_1 = 0$ versus $H_1\colon \beta_1 \neq 0$ is given by

$$F = \frac{MSR}{MSE}$$

where MSR is the mean squared error due to regression; and MSE is the mean squared error s^2

ANOVA Table

A typical form of ANOVA table:

Source	Degrees of Freedom	Sum of Squares	Mean Squares	F-statistic	p-value
Regression	1	SSR= $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$	MSR=$\frac{SSR}{1}$	F=$\frac{MSR}{MSE}$	
Error	n-2	SSE= $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	MSE=$\frac{SSE}{N-2}$		
Total	n-1	SSTO			