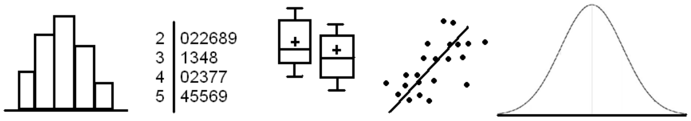


CMED 402: Statistical Science in Medical Research

Seminar 11: Simple Linear Regression

Preparation: Required Reading:

Rosner B. Fundamentals of Biostatistics. 6th ed. California: Thomson Brooks/Cole; 2006.

11 Regression and Correlation Methods, pp. 464 – 487, 492 - 496

Assigned Readings Outline Guide

There is often an approximately linear relationship between variables from a population. Simple linear regression allows us to quantify such relationships.

The full linear regression model takes the form: $y = \alpha + \beta x + e$

We fit the linear model to our data to obtain: $\hat{y} = E\left[y|x\right] = a + bx$

Match the simple linear regression component to its definition.

y _____ A. The value of the independent variable

α _____ B. The estimate of the intercept (regression coefficient)

β _____ C. The random error

x _____ D. The value of the dependent variable

e _____ E. The slope

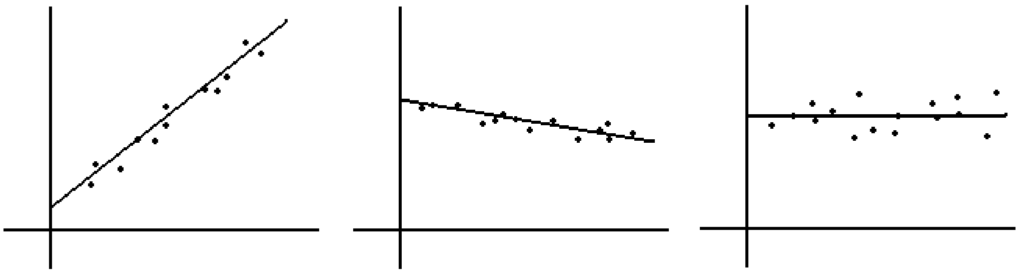
$\hat{y}$ _____ F. The estimate of the slope (regression coefficient)

$E[y|x]$ _____ G. The average value of y for a given value of x

a _____ H. The average value of y for a given value of x

b _____ I. The intercept

The interpretation of the regression line for different value of β



β___ *0* β___ *0* β___ *0*

For any sample point (x_i, y_i), the ___________ component of that point about the regression line is defined by $y_i - \hat{y}_i$ .This is the difference between the actual value and the predicted value. The ___________ component of that point about the regression line is defined by $\hat{y}_i - \bar{y}$. The best fitting regression line has large regression components and small residual components. The worst fitting regression line has small regression components and large residual components. One approach to quantifying how good a regression line fits the data is to square the deviations about the mean, $y_i - \bar{y}$, sum them up over all points, and decompose this sum of squares into regression and residual components.

Decomposition of the Total Sum or Squares into Regression and Residual Components:

$\sum_{i=1}^{n}(y_i - \bar{y})^2$ is the _________ sum of squares (Total SS)

$\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ is the _________ sum of squares (Reg SS)

$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ is the _________ sum of squares (Res SS)

R^2 can be thought of as the proportion of the variance of *y* that can be explained by *x*.
If R^2 = ____, then all variation in *y* can be explained by variation in *x*, and all data points fall on the regression line. If R^2 = ____, then *x* gives no information about *y*, and the variance of *y* is the same with or without knowing *x*.

The simple linear regression model is a great tool, but its answers will only be useful if it is the right model for the data. We need to check the assumptions before using the model. The four assumptions are:

(1)
(2)
(3)
(4)

The main question of interest in simple linear regression is whether or not a linear relationship exists between *x* and *y*. This can be tested by the slope parameter.

H$_0$: $\beta = 0$ (No linear association between *x* and *y*)
H$_1$:

Compute the test statistic: $t = \dfrac{b}{SE(b)}$ where $SE(b) = \sqrt{\dfrac{\text{Res SS}/n-2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$

Compare to critical value: $t_{n-2,1-\alpha/2}$

If H$_0$: $\beta = 0$ is not rejected, what is the best "guess" for the value of the response variable for any value of the predictor variable?

Confidence Intervals for the Slope of the Regression Line:

Lower Bound:

Upper Bound:

One important use for regression lines is in making predictions. For a given value of *x*, an estimate from the regression line is denoted $\hat{y} = ax + b$. Frequently, the accuracy of these predictions must be assessed. How accurate a prediction is depends on whether we are making predictions for **one specific subject** or for the **mean value of all subjects of a given *x***. Intervals for an individual from the population are called _______________ intervals. Intervals for the mean of the population are called confidence intervals.

Both are of the form: $\hat{y} \pm t_{n-2,1-\alpha/2} \cdot SE(\hat{y})$

but when predictions are made from regression lines for individual observations,

$$SE(\hat{y}) = \sqrt{\frac{\text{Res SS}}{n-2}\left[1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right]}$$

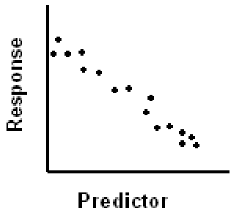
and when predictions are made from regression lines for the mean value of *y* for a given *x*,

$$SE(\hat{y}) = \sqrt{\frac{\text{Res SS}}{n-2}\left[\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right]}$$

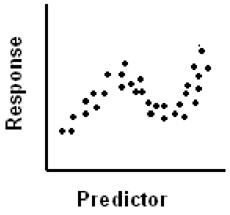
Are prediction or confidence intervals wider? Explain.

Simple Linear Regression focuses on predicting one dependent variable (*y*) from an independent variable (*x*). Often we are interested not in predicting one variable from another but rather in investigating whether or not there is a relationship between two variables. The _______________ ________________ is a useful tool for quantifying the **linear** relationship between two variables and is better suited for this purpose than the regression coefficient.

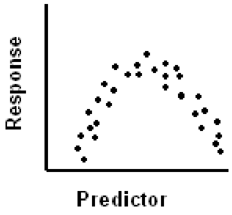
Label each scatter diagram as a linear, nonlinear or no relation between predictor and response.



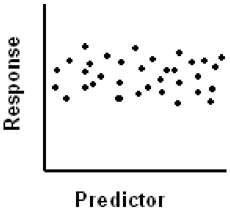
A. ___________



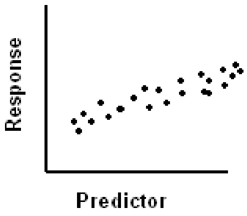
B. ___________



C. ___________



D. ___________



E. ___________

Properties of the Linear Correlation Coefficient

- Always between ___ and ___, inclusive
- If r = ___, there is a perfect positive linear relation between the two variables
- If r = ___, there is a perfect negative linear relation between the two variables
- If $r > 0$, then the variables are said to be positively correlated – as x increases, y tends to ___________, whereas as x decreases, y tends to ___________
- If $r < 0$, then the variables are said to be negatively correlated – as x increases, y tends to ___________, whereas as x decreases, y tends to ___________
- If r is close to ___, this implies no *linear* relation between the two variables.

True False For the least-squares regression model, we require that the independent variable, x, be normally distributed.

True False R^2 is defined as Res SS / Total SS

True False The point $(\bar{x}, \bar{y})$ falls on the regression line.

True False A regression coefficient is another name for a correlation coefficient.

True False A correlation coefficient of 0 means no relation

True False The sample correlation coefficient is only meaningful if the two variables are normally distributed