Week 9 - Review

Tuesday, March 06, 2018
3:44 AM

- ## Regression Analysis
  - The process of constructing a mathematical or function that can be used to predict or determine one variable by another variable or other variables.
    - **Simple Regression (or Bivariate Regression)** – involving two variables in which one variable is predicted by another variable. In a simple regression, only a straight-line relationship between two variables is examined.
    - **Residual** - the difference between the actual y values and the predicted values is the error of the regression line at any given point, $y - \hat{y}$
      - It is the sum of squares of these residuals that is minimized to find the least squares line.
      - The sum of the residuals is always zero (approximately due to rounding).
    - How do you determine the equation of the line for a SLRL (simple linear regression line)?
    - **Outliers** - data points that lie apart from the rest of the points.
      - Outliers can produce residuals with large magnitudes and are usually easy to identify on scatter plots.
      - Outliers can be the result of misrecorded or miscoded data, or they may simply be data points that do not conform to the general trend.
      - The equation of the regression line is influenced by every data point used in its calculation in a manner similar to the arithmetic mean.
      - Outliers sometimes can unduly influence the regression line by "pulling" the line toward the outliers.
  - ### Model Specifications in Regression Analysis
    - Multiple linear regression relates a dependent variable to one or more independent variables
    - Stages of regression analysis
      - Exploratory Data Analysis
      - Model Specifications
      - Estimation of the parameters of the model
      - Diagnostic checking and validation
      - Interpretation of the parameters
    - EDA, theoretical considerations and prior experience contribute to model specification
      - Model Specification Questions:
        - Are the right independent variables included in the model?
        - Are unnecessary variables excluded from the model?
        - Are the variables expressed in proper functional form?
    - Specification errors can lead to problems of estimation, interpretation and erroneous prediction.
  - ### Using Residuals to Test the Assumptions of the Regression Model

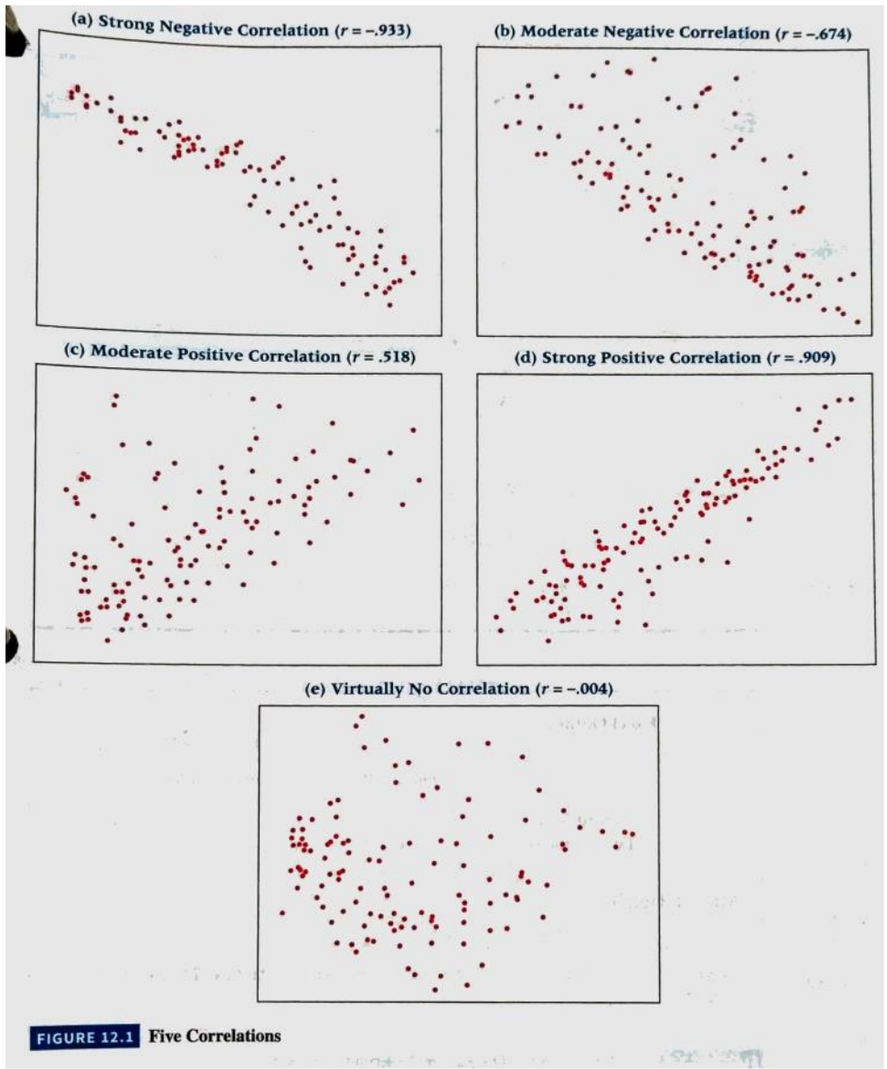


FIGURE 12.1 Five Correlations

Equation of the Simple Regression Line

- Specification errors can lead to problems of estimation, interpretation and erroneous prediction.

- **Using Residuals to Test the Assumptions of the Regression Model**
  - Assumptions to Test Regression Model
    - The model is linear
    - The error terms have constant variances
    - The error terms are independent
    - The error terms are normally distributed
  - **Residual Plot** - a type of graph in which the residuals for a particular regression model are plotted along their associated value of x as an ordered pair, $(x, y - \hat{y})$
    - Residual plots are more meaningful with larger data sets.
    - Smaller data sets tend to be problematic and are typically subject to over-interpretation
  - **Homoscedasticity** - the condition that occurs when the error variances created by a regression model are **constant.**
  - **Heteroscedasticity**- the condition that occurs when the error variances created by a regression model are **not constant.**
- **Standard Error of the Estimate**
  - Standard deviation of the error and therefore we can apply the Imperial Rule (Std1: 68%, Std2: 95%, Std3: 99%) .
  - Residuals represent errors of estimate for individual points.

  - **Sum of Squares of Error**
    - The total of residuals squared
    - $SSE = \sum (y - \hat{y})^2$
  - **Computational Formula for SSE**
    - $SSE = \sum y^2 - b_0 \sum y - b_1 \sum xy$
  - **Standard Error of the Estimate** - standard deviation of error of a regression model
    - $s_e = \sqrt{\frac{SSE}{n-2}}$
  - **How is the standard error of the estimate used?**
    - The standard error of the estimate is a standard deviation of error.
    - If data are approximately normally distributed, the empirical rule states that about 68% of all values are within μ ± 1σ and that about 95% of all values are within μ ± 2σ.
    - One of the assumptions for regression states that for a given x the error terms are normally distributed.
    - Because the error terms are normally distributed, s_e is the standard deviation of error, and the average error is zero, approximately 68% of the error values (residuals) should be within $0 \pm 1s_e$ and 95% of the error values (residuals) should be within $0 \pm 2s_e$.
    - By having knowledge of the variables being studied and by examining the value of s_e, the researcher can often make a judgment about the fit of the regression model to the data by using

Equation of the Simple Regression Line

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

where

b_0 = the sample y intercept
b_1 = the sample slope

Slope of the Regression Line

$$b_1 = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\Sigma(x - \bar{x})^2} = \frac{\Sigma xy - n\bar{x}\,\bar{y}}{\Sigma x^2 - n\bar{x}^2} = \frac{\Sigma xy - \dfrac{(\Sigma x)(\Sigma y)}{n}}{\Sigma x^2 - \dfrac{(\Sigma x)^2}{n}} \qquad (12.2)$$

y Intercept of the Regression Line

$$b_0 = \bar{y} - b_1\bar{x} = \frac{\Sigma y}{n} - b_1\frac{(\Sigma x)}{n} \qquad (12.4)$$

Using the Computer for Residual Analysis

Some computer programs contain mechanisms for analyzing residuals for violations of the regression assumptions. Minitab has the capability of providing graphical analysis of residuals. Figure 12.13 displays Minitab's residual graphic analyses for a regression model developed to predict the production of carrots in the United States per month by the total production of sweet corn. The data were gathered over a time period of 168 consecutive months (see Wiley-PLUS for the agricultural database).

FIGURE 12.9 Nonlinear Residual Plot

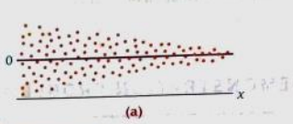
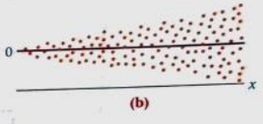


FIGURE 12.10 Nonconstant Error Variance



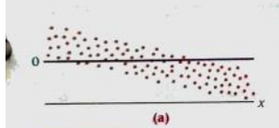


FIGURE 12.11 Graphs of Nonindependent Error Terms

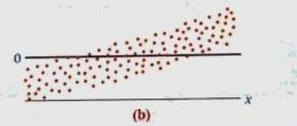


FIGURE 12.12 Healthy Residual Graph

Sum of Squares of Error

$$SSE = \Sigma(y - \hat{y})^2$$

and 95% of the error values (residuals) should be within $0 \pm 2s_e$.

- By having knowledge of the variables being studied and by examining the value of s_e, the researcher can often make a judgment about the fit of the regression model to the data by using s_e.

Coefficient of Determination (R^2)

- The coefficient of determination is the proportion of variability of the dependent variable (y) accounted for or explained by the independent variable (x).
  - The coefficient of determination ranges from 0 to 1.
  - An r^2 of zero means that the predictor accounts for none of the variability of the dependent variable and that there is no regression prediction of y by x.
  - An $r^2$of 1 means perfect prediction of y by x and that 100% of the variability of y is accounted for by x. Of course, most r^2values are between the extremes.
  - The researcher must interpret whether a particular $r^2$is high or low, depending on the use of the model and the context within which the model was developed.
- **Sum of Squares of *y (SSyy)***
  - $SSyy = \sum(y - \bar{y})^2 = \sum y^2 - \frac{(\sum y)^2}{n}$
  - This variation can be broken into two additive variations:
    - ***The explained variation*** - measured by the sum of squares of regression (SSR)
    - ***The unexplained variation*** - measured by the sum of squares error (SSE)
      - $SSyy = SSR + SSE$
  - **Coefficient of Determination**
    - This reflects the proportion of variance accounted for using a least squares regression line and X to predict Y.
    - $r^2 = 1 - \frac{SSE}{SSyy} = 1 - \frac{SSE}{\sum y^2 - \frac{(\sum y)^2}{n}}$
    - *Note: $0 \leq r^2 \leq 1$*
  - **Computational Formula for r^2**
    - $r^2 = \frac{b_1^2 SS_{xx}}{SS_{yy}}$
  - **Adjusted R^2**
    - This takes into consideration both the additional information for each new independent variables brings to the regression model and the changed degrees of freedom of regression.

Hypothesis Tests for the Slope of the Regression Model and Testing the Overall Model

- **Testing the Slope**
  - A hypothesis test can be conducted on a sample slope of the regression model to determine whether the population slope is significantly different from zero.
  - This test is another way to determine how well a regression model fits the data.
  - An alternative approach might be to average the *y* values and use $\bar{y}$ hat as the predictor for *y* for all values of *x*
- ***t* Test the Slope**
  - This is a two-tailed test and the null hypothesis can be rejected if the slope is either negative or positive

Sum of Squares of Error

$$SSE = \Sigma(y - \hat{y})^2$$

Computational Formula for SSE

$$SSE = \Sigma y^2 - b_0 \Sigma y - b_1 \Sigma xy$$

Standard Error of the Estimate

$$s_e = \sqrt{\frac{SSE}{n-2}}$$

Coefficient of Determination

$$r^2 = 1 - \frac{SSE}{SS_{yy}} = 1 - \frac{SSE}{\Sigma y^2 - \frac{(\Sigma y)^2}{n}} \qquad (12.5)$$

Note: $0 \leq r^2 \leq 1$

Computational Formula for r^2

$$r^2 = \frac{b_1^2 SS_{xx}}{SS_{yy}}$$

***t* Test of Slope**

$$t = \frac{b_1 - \beta_1}{s_b}$$

where

$$s_b = \frac{s_e}{\sqrt{SS_{xx}}}$$

$$s_e = \sqrt{\frac{SSE}{n-2}}$$

$$SS_{xx} = \Sigma x^2 - \frac{(\Sigma x)^2}{n}$$

β_1 = the hypothesized slope
$df = n - 2$

positive.
    - A negative slope indicates an inverse relationship between *x* and *y*.
    - To determine if there is a significant **positive relationship**
        - $H_0: \beta_1 = 0$
        - $H_a: \beta_1 > 0$
    - To determine if there is a significant **negative relationship**
        - $H_0: \beta_1 = 0$
        - $H_a: \beta_1 < 0$
- *t* Test of Slope
    - $t = \frac{b_1 - \beta_1}{s_b}$
    - Where
        - $s_b = \frac{s_e}{\sqrt{SS_{xx}}}$
        - $s_e = \sqrt{\frac{SSE}{n-2}}$
        - $SS_{xx} = \sum x^2 - \frac{(\sum x)^2}{n}$
        - $\beta_1 = the\ hypothesized\ slope; df = n - 2$
- **Testing the Overall Model**
    - The *F* test for overall significance is testing the same thing as the t test in simple regression.
        - $F = \frac{\frac{SS_{reg}}{df_{reg}}}{\frac{SS_{err}}{df_{err}}} = \frac{MS_{reg}}{MS_{err}}$
        - Where
            - $df_{reg} = k$
            - $df_{err} = n - k - 1$
            - $k = the\ number\ of\ independent\ variables$
        - The values of the sum of squares (SS), degrees of freedom (df), and the mean squares (MS) are obtained from the analysis of variance table, which is produced with standard regression output from statistical software packages.
    - In simple regression, the relationship between the critical t value to test the slope and the critical F value of overall significance is
        - $t^2_{\frac{\alpha}{2}, n-2} = F_{\alpha,1,n-2}$
- **Testing Strength of a Regession Model**
    - Residuals
        - Represent the observed y value and predicted y value.
        - Smaller residuals are better so that we have less error in our model
    - Standard Error
        - Standard deviation of error
        - The smaller the standard error is the better fit is the model
    - R^2

Extending this notion to multiple regression gives the general equation for the probabilistic multiple regression model.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k + \in$$

where

y = the value of the dependent variable
β_0 = the regression constant
β_1 = the partial regression coefficient for independent variable 1
β_2 = the partial regression coefficient for independent variable 2
β_3 = the partial regression coefficient for independent variable 3
β_k = the partial regression coefficient for independent variable k
k = the number of independent variables

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + \cdots + b_k x_k$$

where

$\hat{y}$ = the predicted value of y
b_0 = the estimate of the regression constant
b_1 = the estimate of regression coefficient 1
b_2 = the estimate of regression coefficient 2
b_3 = the estimate of regression coefficient 3
b_k = the estimate of regression coefficient k
k = the number of independent variables

Formulas

The *F* value

$$F = \frac{MS_{reg}}{MS_{err}} = \frac{SS_{reg}/df_{reg}}{SS_{err}/df_{err}} = \frac{SSR/k}{SSE/(N-k-1)}$$

Sum of squares of error

$$SSE = \Sigma(y - \hat{y})^2$$

Standard error of the estimate

$$s_e = \sqrt{\frac{SSE}{n-k-1}}$$

Coefficient of multiple determination

$$R^2 = \frac{SSR}{SS_{yy}} = 1 - \frac{SSE}{SS_{yy}}$$

Adjusted R^2

$$\text{Adjusted } R^2 = 1 - \frac{SSE/(n-k-1)}{SS_{yy}/(n-1)}$$

        - The higher the value of R^2, the better fit is the model

- ## Correlation
    - Product Moment Correlation Coefficent
        - $r = b_1 \frac{s_x}{s_y}$
    - The value of *r* is commonly used to summarize the association between two variables and dates back to at least the year 1846.
    - **Population correlation *ρ (aka "rho")***
        - Rho is the value of *r* if all individuals of interest could be measures.
        - A fundamental property of the population correlation coefficient is that if X and Y are independent, then *ρ = 0.*
        - This property is of practical importance because if persuasive empirical evidence, based on *r*, indicates that $\rho \neq 0$, then it is reasonable to conclude that X and Y are dependent.
    - ### Interpreting r
        - A good understanding of *r* requires an understanding not only of what it tells us about an association, but also what it does not tell us.
        - It is important not to read more into the value of *r* than is warranted.
        - There are five major features that impact the magnitude
            - The slope of the line around which points are clustered
            - The magnitude of the residuals
            - Outliers
            - Restricting the range of the X values, which can cause *r* to go up or down
            - Curvature
    - ### How to use r
        - Pearson's correlation, *r*, has two useful functions.
            - 1.) It can be used to establish dependence between two variables by testing and rejecting the hypothesis that *ρ* is equal to zero .
            - 2.) r^2 reflects the extent to which the least squares regression estimates of Y, namely Y hat, improves upon the sample mean, Y bar, in terms of predicting Y
    - ### Extrapolation can be dangerous
        - If the results you are trying to predict are not within the same range of data you used to create the least square slope and intercept, there is a chance that highly inaccurate results might be obtained.