Regression, Inference, and Model Building

Scatter Plots and Correlation

Correlation coefficient, r

$-1 \leq r \leq 1$

If r is positive, then the scatter plot has a positive slope and variables are said to have a positive relationship

If r is negative, then the scatter plot has a negative slope and variables have negative relationships

If r = 0, then no linear relationship exists between the two variables

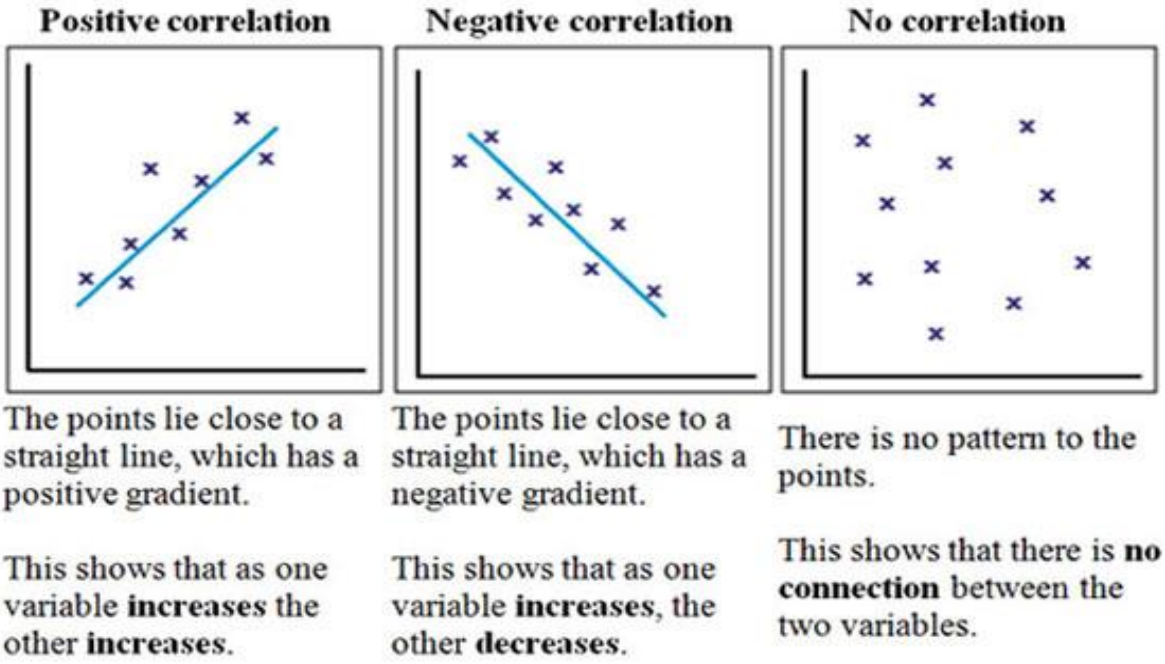
If r = 1, then the data falls in a perfect line with a positive slope

If r = -1, the data falls in a perfect line with a negative slope

Strength of the correlation is $|r|$

The larger $|r|$ is, the stronger the correlation

$$r = \frac{(n\sum xy)-(\sum x)(\sum y)}{\left(\sqrt{(n\sum x^2)*(\sum x)^2}\right)*\left(\sqrt{(n\sum y^2)*(\sum y)^2}\right)}$$



Significant Linear Relationship (two-tailed test):

Ho: $p = 0$ (implies there is no significant linear relationship)

Ha: $p \neq 0$ (implies there is a significant linear relationship)

Negative Linear Relationship (left-tailed test):

Ho: $p \geq 0$

Ha: $p < 0$

Created by: Megan Krou
Spring 2017
STUDENT LEARNING ASSISTANCE CENTER (SLAC)
Texas State University

Positive Linear Relationship (right-tailed test):
Ho: $p \leq 0$
Ha: $p > 0$

Test statistic:

$$t = \frac{r}{\sqrt{\frac{1 - r^2}{n - 2}}}$$

Degrees of freedom = n-2

Significant Linear Relationship (two-tailed tests):
Reject Ho if $|t| \geq \left(t_{\frac{\alpha}{2}}\right)$

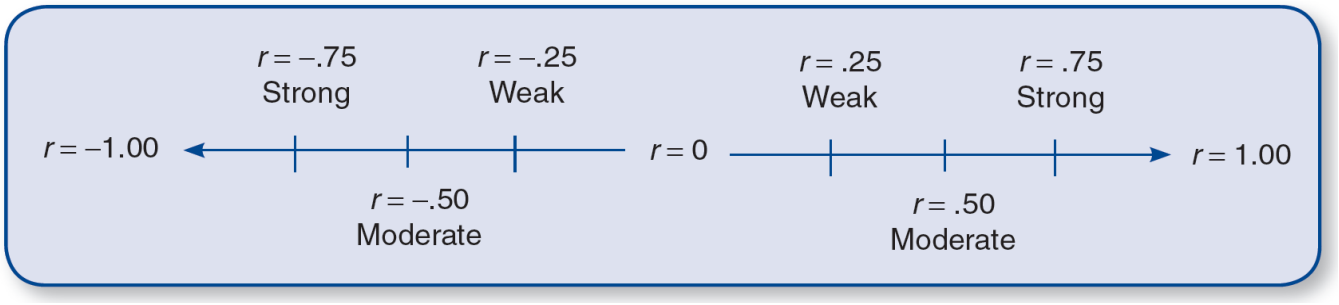
Negative Linear Relationship (left-tailed test):
Reject Ho if $t \leq -t_{\alpha}$

Positive Linear Relationship (right-tailed test):
Reject Ho if $t \geq t_{\alpha}$

Using critical values to determine statistical significance:
The correlation coefficient, r, is statistically significant if the absolute value of the correlation is greater than the critical value in the table.
$|r| > r_{\alpha}$

The coefficient of determination, r^2, is the measure of the amount of variation in y explained by the variation in x.

Figure 13.4 General Guidelines for Interpreting the Magnitude of Correlation Coefficients



Fitting a Linear Model

A linear relationship is graphically described as a line.
$y = b_0 + b_1 x$

b_0= y-intercept b_1= slope of the line $\hat{y}$= predicted y n = number of data values

Error = observed Y – predicted Y

Error = y - $\hat{y}$
$y = b_0 + b_1 x + error$
$\hat{y} = b_0 + b_1 x$

Sum of Squared Errors (SSE):

$$SSE = \sum (error_i)^2 = \sum (y_i - \hat{y_i})^2 = \sum \left(y_i - (b_0 + b_1 x_1)\right)^2$$

Defining the least squares line:

$$Slope = b_1 = \frac{\sum xy - \frac{(\sum x)(\sum y)}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}$$

$$y - intercept = b_0 = \frac{1}{n}\left(\sum y - b_1 \sum x\right)$$

$$r^2 = \left(\frac{(n\sum xy) - (\sum x)(\sum y)}{\left(\sqrt{(n\sum x^2) * (\sum x)^2}\right) * \left(\sqrt{(n\sum y^2) * (\sum y)^2}\right)}\right)^2$$

The parameter $r^2$is defined as the fraction of the total variation explained by the least squares line. In other words, r^2 measures how well the least squares line fits the sample data. If the total variation is explained completely, then we have $r^2 = 1$ and we say that there is a perfect linear correlation. On the other hand, if the total variation is all unexplained, then $r^2 = 0$.

Regression Analysis

Variance of Errors

$$(S_e)^2 = \frac{SSE}{n-2}$$

Variance of Slope:

$$\left(S_{b_1}\right)^2 = \frac{(S_e)^2}{\sum x^2 - \frac{(\sum x)^2}{n}}$$

Created by: Megan Krou
Spring 2017
STUDENT LEARNING ASSISTANCE CENTER (SLAC)
Texas State University

Bounds of confidence interval:

$$b_1 \pm \left(t_{\frac{\alpha}{2}}\right)\left(S_{b_1}\right)$$

b_1=slope
S_{b_1}=standard deviation of slope
$\left(t_{\frac{\alpha}{2}}\right)$=look up in table $\frac{\alpha}{2}$ and df = n-2

Multiple Regression

Multiple regression model

$\hat{y} = b_0 + b_1x_1 + b_2x_2 + \cdots + b_kx_k$

Where $x_1, x_2, \ldots, x_k$ are the k independent variables in the model and $b_1, b_2, \ldots, b_k$ are the corresponding coefficients of the independent variables. These values, $b_1, b_2, \ldots, b_k$, are the sample estimates of the corresponding population parameters, $\beta_1, \beta_2, \ldots, \beta_k$. The y-intercept of the multiple regression equation is b_0, which is the sample estimate of the population parameter, β_0.

The multiple coefficient of determination, R^2

Test the claim that at least one of the independent variables' coefficients is not equal to 0
Ho: $\beta_1 = \beta_2 = \cdots = \beta_k = 0$
Ha: At least one coefficient does not equal 0

p-value = Significance F (from running the ANOVA test in excel)
p-value < alpha reject the null
p-value ≥ alpha fail to reject the null

Test if specific variables are significant
Ho: $\beta_1 = 0$
Ha: $\beta_1 \neq 0$
*If reject null then variable is significant

ANOVA Regression

Analysis of Variance provides information about how well an estimated regression model fits the data.

ANOVA table divides the total variation in Y into variation that can be explained by the model and variation that cannot be explained by the model. The division of variation takes place in the column labeled "SS."

Source	SS	DF	MS	F
Regression	SSR	k	$MSR = \frac{SSR}{k}$	$F = \frac{MSR}{MSE}$
Error	SSE	n-(k+1)	$MSE = \frac{SSE}{n-(k+1)}$	
Total	TotalSS	n-1		

Created by: Megan Krou
Spring 2017
STUDENT LEARNING ASSISTANCE CENTER (SLAC)
Texas State University

Total variation in Y = variation in Y explained by the model + variation in Y not explained by the model
TotalSS = SSR + SSE

SSR = Sum of Squared Regression = amount of variation in Y explained by the model
SSR = TotalSS - SSE

SSE = Sum of Squared Error = variation in Y that the model could not explain

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y})^2$$

TotalSS = Total Sum of Squared Deviations about the mean of the dependent variable Y

$$TotalSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

*Note: that if this expression were divided by (n-1), it would be the sample variance of

k = number of independent variables in the model, called **DFR** – Degrees of Freedom among Regression
n-(k+1) = degrees of freedom associated with unexplained variation in the model, called **DFE** – Degrees of Freedom among Error
n-1 = total degrees of freedom, where n is the number of observations, called **DFT** – Total Degrees of Freedom

$MSR = \frac{SSR}{k}$, average amount of variation explained per independent variable
$MSE = \frac{SSE}{n-(k+1)}$, variance of the error terms
$F - statistic = \frac{MSR}{MSE}$, large F values are desirable since that would indicate that explained variation is relatively larger than the unexplained.

Created by: Megan Krou
Spring 2017
STUDENT LEARNING ASSISTANCE CENTER (SLAC)
Texas State University