Correlation and Regression

**Our next topic involves methods for dealing with relationships between two variables. The concepts of correlation and regression are discussed.**

**Our goal is to make predictions based on statistical models. For instance, we may wish to predict the failure time of overloaded resistors. Given an overload (volts) we would predict the time (minutes) to failure of the resistor.**

**A correlation exists between two variables when one of them is related to the other.**

**How do we expect resistance to be related to time to failure?**

**The linear correlation coefficient r measures the strength of the linear relationship between the paired x and y values in a sample.**

**Sometimes the linear correlation is called the Pearson correlation coefficient in honour of Karl Pearson (1857-1936)**

****

**We use r to measure the linear correlation coefficient of a sample and ρ (Greek rho) to represent the population linear correlation coefficient.**

**Assumptions**

1. **The sample of paired (x,y) data is a random sample.**
2. **The pairs of (x,y) data have a bivariate normal distribution. This means that for any fixed value of x, the corresponding values of y have a normal distribution, and for any fixed value of y, the values of x have a distribution that is normal.**

**Properties of the linear Correlation Coefficient, *r***

1. **The value of *r* is always between -1 and 1.**

**-1 ≤ *r* ≤ 1**

1. **The value of *r* does not change if all values of either variable are converted to a different scale.**
2. **The value of *r* is not affected by the choice of *x* or *y*. Interchange all *x* and *y* values and the value of *r* will not change.**
3. ***r* measures the strength of a linear relationship. It is not designed to measure the strength of a relationship that is not linear.**
4. ***r* = 1 ⇒ perfect positive linear correlation**

***r* "close to" 1 ⇒ strong positive linear correlation**

***r* = 0 ⇒ no linear correlation**

***r* "close to" -1 ⇒ strong negative linear correlation**

***r* = -1 ⇒ perfect negative linear correlation**

**IMPORTANT NOTE:**

**Correlation does not imply that one variable causes the other.**

**Example:**

**The size of a person's foot has a strong positive linear correlation with the height of the person. Having a large foot does NOT cause a person to be tall.**

**Example:**

**The following data pertain to resistance (ohms) and the failure time (minutes) of certain overloaded circuits.**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Resistance (ohms)** | **Failure Time (Minutes)** | |  |  |
|  | **x** | **y** | **xy** | **x2** | **y2** |
|  | **43** | **32** | **1376** | **1849** | **1024** |
|  | **29** | **20** | **580** | **841** | **400** |
|  | **44** | **45** | **1980** | **1936** | **2025** |
|  | **33** | **35** | **1155** | **1089** | **1225** |
|  | **33** | **22** | **726** | **1089** | **484** |
|  | **47** | **46** | **2162** | **2209** | **2116** |
|  | **34** | **28** | **952** | **1156** | **784** |
|  | **31** | **26** | **806** | **961** | **676** |
|  | **48** | **37** | **1776** | **2304** | **1369** |
|  | **34** | **33** | **1122** | **1156** | **1089** |
|  | **46** | **47** | **2162** | **2116** | **2209** |
|  | **37** | **30** | **1110** | **1369** | **900** |
| **Total**      ***r* = 0.8324**  **Fairly strong positive linear correlation.** | **459** | **401** | **15907** | **18075** | **14301** |

**Computing r on your calculator:**

**Get into STAT/LINE mode:**

**Enter the data:**

**Find r:**

**We can be *pretty sure* of a linear correlation between 2 variables if r is *close enough* to ±1.**

**But how can we know how close to ±1 is close enough? And for a certain “close” value of r, how sure can we be that a linear correlation exists?**

**…Does this sound familiar?**

**We can conduct a formal hypothesis test to determine whether there is a significant linear correlation between two variables. Note this is a slightly different method than that given in your text (Section 11.6)**

**H0: ρ = 0, Test Statistic: with df =** *n* **– 2**

**Example:**

**Test for significant linear correlation between our resistance variable and our time to failure variable at a significance level of 0.05.**

**Now that we’ve analyzed paired data with the goal of determining whether there is a significant linear relationship between two variables, we wish to describe the relationship by finding the equation of the straight line that best fits the data.**

**Given a collection of paired sample data, the regression equation**

****

**describes the relationship between two variables. The graph of the regression equation is called the regression line (or line of best fit, or least squares line)**

***x* is called the independent or predictor variable**

***y* is called the dependent or response variable.**

**Notation for Regression Equation**

|  |  |  |
| --- | --- | --- |
|  | **Population** | **Sample** |
| **y-intercept of regression equation** | **α** | **a** |
| **slope of regression equation** | **β** | **b** |
| **Equation of the regression line** | **y = α + β*x*** |  |

**Assumptions:**

1. **We are investigating only linear relationships.**
2. **For each *x* value, *y* is a random variable having a normal distribution. All of these *y* distributions have the same variance. Also, for a given *x*, the distribution of *y* has a mean that lies on the regression line.**

**Our goal is to find estimates, *a* and *b*, that will form a line that is the minimum distance from all of the data points.**

**The derivation of these formulas involves a calculus procedure referred to as least squares. The process of least squares involves finding the minimum of the squared distance between the data points and the line.**

****

** Slope**

** y-intercept**

**Example:**

**The following data pertain to resistance (ohms) and the failure time (minutes) of certain overloaded circuits. Find the line that best fits the data.**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Resistance (ohms)** | **Failure Time (Minutes)** | |  |  |
|  | **x** | **y** | **xy** | **x2** | **y2** |
|  | **43** | **32** | **1376** | **1849** | **1024** |
|  | **29** | **20** | **580** | **841** | **400** |
|  | **44** | **45** | **1980** | **1936** | **2025** |
|  | **33** | **35** | **1155** | **1089** | **1225** |
|  | **33** | **22** | **726** | **1089** | **484** |
|  | **47** | **46** | **2162** | **2209** | **2116** |
|  | **34** | **28** | **952** | **1156** | **784** |
|  | **31** | **26** | **806** | **961** | **676** |
|  | **48** | **37** | **1776** | **2304** | **1369** |
|  | **34** | **33** | **1122** | **1156** | **1089** |
|  | **46** | **47** | **2162** | **2116** | **2209** |
|  | **37** | **30** | **1110** | **1369** | **900** |
| **Total** | **459** | **401** | **15907** | **18075** | **14301** |

****

****

****

**Note: Many of your calculators will do these calculations automatically.**

**Knowing the model (equation) that best fits the data allows us to make predictions for values that were not measured.**

**Suppose we wish to estimate the time to failure of a circuit with 40 ohms of resistance.**

****

**Note: It is NOT wise to make predictions for values that are well beyond your data. Our x values ranged from 29 to 47. For instance, we have no right to predict values for x = 70 because we have no idea whether the linear trend will continue beyond the range of our data.**

**Now that we have an estimate, , for y when x = 40, we can construct confidence interval for the true value of y.**

**Before constructing our prediction interval we need to learn about a value called the Standard Error of Estimate.**

**The Standard Error, Se, is a measure of the differences (or distances) between the observed sample y values and the predicted values that are obtained using the regression equation.**

****

**Think of this value as a measure of the average distance of the data from the line.**

**How do we expect r to be related to Se?**

**Example: Construct the standard error for our example involving time to failure of resistors.**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Resistance (ohms)** | **Failure Time (Minutes)** | |  |  |
|  | **X** | **y** | **xy** | **x2** | **y2** |
|  | **43** | **32** | **1376** | **1849** | **1024** |
|  | **29** | **20** | **580** | **841** | **400** |
|  | **44** | **45** | **1980** | **1936** | **2025** |
|  | **33** | **35** | **1155** | **1089** | **1225** |
|  | **33** | **22** | **726** | **1089** | **484** |
|  | **47** | **46** | **2162** | **2209** | **2116** |
|  | **34** | **28** | **952** | **1156** | **784** |
|  | **31** | **26** | **806** | **961** | **676** |
|  | **48** | **37** | **1776** | **2304** | **1369** |
|  | **34** | **33** | **1122** | **1156** | **1089** |
|  | **46** | **47** | **2162** | **2116** | **2209** |
|  | **37** | **30** | **1110** | **1369** | **900** |
| **Total** | **459** | **401** | **15907** | **18075** | **14301** |

****

**To construct a (1 - α)100% prediction interval we use the following:**

** - E ≤ *y* ≤  + E**

****

** has n - 2 degrees of freedom**

**Example: Suppose we wish to calculate a 95% prediction interval for y when x is 40 for our time to failure of resistors problem.**

**We can also construct confidence intervals for the parameters α and β in our regression model.**

**Finding a confidence interval for the true slope of the regression equation, β is similar to find other confidence intervals that we have created. We find an estimate for the population parameter and then use the estimate to construct a confidence interval.**

**To construct a (1- α)100% Confidence Interval for β, we use the following:**





**df = n - 2**

**Hypothesis Test For β**

**If β = 0, the regression line is horizontal and the mean of y does not depend on the linearity of x. It is common to perform hypothesis tests for this case.**

**The test statistic is:**



**The one sided critical value, tα, and the two sided critical value, tα/2, have df = n - 2**

**Example:**

**Continue our example involving time to failure of resistors. Test the hypothesis that the slope of the regression line is not zero at a significance level of 0.05.**