**SLR overview**

|  |  |
| --- | --- |
| Simple Linear Regression Model | |
| Basic | For a quantitative response variable Y and a single quantitative explanatory variable X, the simple linear regression model is    Where t he errors are independent of each other and follow a normal distribution, that is |
| Assumption | 1. We expect the average error *ei* to be zero so the regression line passes through the conditional mean of *Y*. 2. The *ei* have constant variance *σe2.* 3. The *ei* are normally distributed. 4. The *ei* are independent. |

**Bivariate Data:**

|  |
| --- |
| When you collect two variables on one subject, you get bivariate data (Etc. Height vs.Weight)  If both values are quantitative, you can make a scatter plot letting one variable become explanatory and the other one become responsive |

**Scatter plot**

|  |
| --- |
| \*scatter plot shows relationship between two quantitative variables |
| Examine the pattern of the scatter plot  1) Form: Is there a straight line relationship between the variables? Does the graph curve slightly or sharply either up or down? Can you see a pattern  2) Direction: A pattern that runs from the upper left to the lower right has negative direction and a pattern that runs the other way has a positive direction.  3) Strength: The less scatter the stronger the relationship between the variables. Graphs with a lot of scatter show a weak relationship between the points |

**Simple linear Model**

|  |
| --- |
| The SIMPLE LINEAR MODEL relating Y and X is  Y = b0 + b1X.  b0 is the Y-INTERCEPT of the model, the point where the line crosses the Y-axis.  b1 is the SLOPE of the model, the change in Y for a given unit change in X (“rise” over “run”). |
| Correlation shows the strength and direction of a linear relationship. It is measured by the Correlation Coefficient    1) -1 ≤ r ≤ +1  2) r = -1 represents perfect negative linear association.  3) r = +1 represents perfect positive linear association.  4) r = 0 represents no linear association.  \*r is very sensitive to extreme observations |
| The Coefficient of Determination, r2   * r2  is the percentage of variation in the Y variable that is explained by the least squares regression on X.   r2 has values between 0 and 100%  r = 0.944, Hence r2 = 0.8905  Etc. Hence 89% of the variation in snake weight is explained by variation in snake length. |

**In ANOVA table**

|  |
| --- |
| Recall the formulae for the intercept and slope can be found by      SST = 9990  SSR = 12372/172=8896.331  SSE = 9990-8896.331=1093.6686 |
| = 8896.331/9990=0.8905  r = square root of r2 and the sign must match the sign of the slope = 0.9436 |

|  |
| --- |
| Assumptions Concerning the Population Regression Line   1. We expect the average error *ei* to be zero so the regression line passes through the conditional mean of *Y*. 2. The *ei* have constant variance *σe2.* 3. The *ei* are normally distributed. 4. The *ei* are independent.   Residual = observed value – predicted value    Observed weight is 136 g when length is 60 cm  Predicted weight is 130.42 when length is 60 cm  Residual = 136- 130.42 = 5.58 which shows an underestimate  Checking Model Adequacy  The plot of residuals vs x-values should show NO interesting features. If you see something your model may be missing something  If the residuals show no interesting patterns when we plot them against x, we can look at how big they are as we are trying to make them as small as possible. Since their mean is always zero it is only sensible to look at how much they vary. The se gives us a measures of how much the points vary around the regression line.  This leads to the assumption check of Equal Variance Assumption.  Standard Deviation of the Residuals    The divisor *n-2* used in the previous calculation follows a general rule that degrees of freedom are sample size – the number of estimates we make (*b0* and *b1*)before estimating the variance.   * A normal probability plot of the residuals shows that the residuals are normally distributed * Residuals vs fitted value plots can also be used to check for problems with linearity, constant variance or both. |

Check Assumption:

|  |
| --- |
| * Assumption #1 * If we construct the line of best fit, the residuals must sum to zero. * Assumption #2 * Examine a Q-Q plot * Do the points appear to follow in a linear pattern? * Yes – Good! This implies the errors are normally distributed * Assumption # 3 * Examine a residual plot. * Do you see a pattern? * No – good! This implies the constant variance assumption has been met * Assumption #4 * We must look at how the sample was selected. Did we use a simple random sample? Are the subjects independent? |

Inferences about *β0* and *β1*

|  |
| --- |
| * We use our sample data to estimate *β0* by *b0* and *β1* by *b1.* If we had a different sample, we would not be surprised to get different estimates. * Understanding how much they would vary from sample to sample is an important part of the inference process. * We use the assumptions, together with our data, to construct the sampling distributions for *b0* and *b1*. |
| The most common test is that a change in the *X* variable does not induce a change in *Y*, which can be stated:  *H0:* *β1 = 0*  *Ha:* *β1 ≠ 0*  If H0 is true it implies the population regression equation is a flat line; that is, regardless of the value of *X*, *Y* has the same distribution. |
| *t*-Distribution for Inferences for *b*1 |
| Inference for the slope; Test Statistic =  P-value = 0.000 at df = 9-2=7 |
| Confidence interval for slope    7.192 ± 2.3646(.953)  (4.938,9.445)  We are 95% confident that the slope of the line lies between 4.938 and 9.445 grams. |
| Confidence Intervals about the Regression Line   * Suppose we wish to determine a confidence interval for the predicted value of Y for a given value of X. * To make an interval estimate, we need some kind of standard error. * Because our point estimate is a function of the random variables *b0* and *b1* their standard errors enter into our computation. |
| Obtain a 95% confidence interval for the predicted mean value of Y at the point X = 62 cm  Point estimate for ŷ = -301.087+(7.192)(62)  = 144.817  Previously, Sxx = 172  se = 12.4995  Mean of X = 63    144.817 ± 10.1066  We are 95% confident that the predicted value of y when x=62 cm lies between 134.7104 to 154.9236 grams. |
| Outlier  Data points that diverge in a big way from the overall pattern are called [outliers](http://stattrek.com/Help/Glossary.aspx?Target=Outlier). There are four ways that a data point might be considered an outlier.   * It could have an extreme X value compared to other data points. * It could have an extreme Y value compared to other data points. * It could have extreme X and Y values. * It might be distant from the rest of the data, even without extreme X or Y values.   An influential point is an [outlier](http://stattrek.com/Help/Glossary.aspx?Target=Outlier) that greatly affects the slope of the regression line. One way to test the influence of an outlier is to compute the regression equation with and without the outlier. |

Multiple Regression:

|  |
| --- |
| For *k* explanatory variables, we can express the population mean response (***my***) as a linear equation:  ***E(y) = my* = *b*0 + *b*1*x*1 … + *bkxk***  The statistical model for *n* sample data (*i =* 1, 2, *… n*) is then:  Data = fit + residual  *yi* = (*b*0 + *b*1*x*1*i* … + *bkxki*) + (*ei*)  Where the *ei* are independent and normally distributed *N* (0, *s*). |
| *b*0 *= E(y)* when *x1 = x2 = … = xk =* 0.  *b*1, *b*2 , … , *b*k are called *partial regression coefficients.*  Controlling for other predictors in model, there is a linear relationship between *E(y)* and *x1* with slope *b*1.  i.e., consider case of *k* = 2 explanatory variables,  *E(y) = b0* + *b*1x1 + *b*2x2  If *x1* goes up 1 unit with x2held constant, the change in *E(y)* is  [*b0* + *b*1(x1 + 1)+ *b*2x2] – [*b0* + *b*1x1 + *b*2x2] = *b*1. |
| * *We can write the most general type of linear model in variables x1, x2,…xk in the form*   *ŷ = b0+ b1 x1+ b2 x2 + … + bk xk*     * *This model is called a ‘first-order model’ with k explanatory variables.* * *The method of least squares chooses the b’s that make the sum of squares of the residuals as small as possible. The b’s must vary simultaneously to minimize the sum of squared deviations.* |
| Assumption   * Assumptions of a regression model’s errors:   + Normal distribution of the errors, i.e., *N*(0, σ).   + Constant variance (homoscedasticity) for all values of x.   + Independence. * Assumptions are evaluated graphically.   + Check normality with a histogram of residuals   + Check constant variance with a residuals plot (residuals vs. predicted values)   + Check independence with a plot of the residuals versus time order. |
| Example:  Y=Bo + B1x1+B2x2+B3x3+e  Where  y= sales price (in dollars)  x1= appraised land value (dollars)  x2=appraised value of the improvements ($)  x3=area of living space (square feet)   * Note: A first order model includes only quantitative variables and does not contain any higher order terms   First,   * Construct scatterplots to examine the bivariate relationships between the response variable and each of the explanatory variables. * Describe each relationship * Interpret all three scatterplots: Form, direction and strength   Next,  use the method of least squares to find estimate of the unknown parameters      Now,  Interpret the parameters     * We *estimate* the mean sales price to increase 0.8145 dollars for every $1 increase in appraised land value when both appraised improvements and area are held fixed. * Similar interpretations can be made for both estimate for appraised improvements and additional square foot of living area. * But there is no meaningful interpretation for the intercept value of 1,470.28 |
| To test the coefficients for one of the predictors in a multiple regression the hypotheses are  Ho: Bi = 0 versus Ho: Bi ≠ 0  With test statistic  t = parameter estimate  standard error of the estimate  Similarly, confidence intervals for regression coefficients     * Where t\* is the critical value for the t distribution with df=n-k-1 where k is the number of predictors |
| Evaluating Overall Model Utility of a Multiple Regression Model   * 1. Use the F-test to conduct a test of the adequacy of the overall model * 2. Conduct t-test on those parameters in which you are particularly interested. CAUTION: conducting a series of t-tests can lead to a high overall type I error rate. * 3. Examine the values of R2adjusted and 2s to evaluate how well, numerically, the model fits the data |
| Global F-test  To test  *H*0: *b*1 = *b*2 = … = *bk*= 0  *H*a: at least one of these coefficients is not zero  *F* statistic: *F* = MSM / MSE  Use to test if any of the predictors is effective in the model. Does not tell us which one as that as the role for the individual t-tests. |
| Coefficient of Determination, r2  Recall, r2 is a measure of the percentage of total variability explained by the regression model    From example, 89.74% of the variability in sales price can be explained by the regression model based on appraised land value (dollars), appraised value of the improvements ($) and area of living space (square feet)   * The adjusted R2 helps account for the number of predictors in the model. * This measure is needed as r2 tends to increase as new predictors are added to a model   Requirements of a Good Model   * Statistical significance of all parameters:   + ANOVA p-value < α     - Coefficient of determination (r2) is high   + β parameter p-values < α   + All required assumptions are met.   + Which are …?     - Errors are normally distributed     - Errors have constant variance (homoscedasticity) |
| Variable Selection Techniques - Which Model?   * One method to determine if multicollinearity needs to be dealt with is to use the VIF. * VIF values that exceed 5 (or 10) might be in indication that the associated regression coefficients are poorly estimated.   With a large number of potential *X* variables, it may be best to use one of the iterative selection methods. |
| 1. Backwards Elimination   1. Start with all variables in the equation. 2. Examine the variables in the model for significance and identify the least significant one. 3. Remove this variable if it does not meet some minimum significance level. 4. Run a new regression and repeat until all remaining variables are significant.   2. Forward Selection   * At each stage, it looks at the *X* variables not in the current equation and tests to see if they will be significant if they are added. * In the first stage, the *X* with the highest correlation with *Y* is added.   At later stages it is much harder to see how the next *X* is selected  3. Stepwise Regression   * A limitation with the backwards procedure is that a variable that gets eliminated is never considered again. * With forward selection, variables entering stay in, even if they lose significance. * Stepwise regression corrects these flaws. A variable entering can later leave. A variable eliminated can later come back in. |
|  |
|  |