

Vocabulary

Term	Definition
Adjusted R^2	An adjustment to the R^2 statistic that attempts to allow for the number of predictors in the model. It is sometimes used when comparing regression models with different numbers of predictors.
ANOVA	An analysis method for testing equality of means across treatment groups.
ANOVA	The Analysis of Variance table that is ordinarily part of the multiple regression results offers an F -test to test the null hypothesis that the overall regression is no improvement over just modeling y with its mean: $H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0$ If this null hypothesis is not rejected, then you should not proceed to test the individual coefficients.
ANOVA model	The model for a one-way (one response, one factor) ANOVA is $y_{ik} = \mu_k + \epsilon_{ik}$ Estimating with $y_{ik} = y_k \text{ bar} + e_{ik}$ gives predicted values $y_{ik} = y_k \text{ bar}$ and residuals $e_{ik} = y_{ik} - y_k \text{ bar}$.
ANOVA table	The ANOVA table is convenient for showing the degrees of freedom, the treatment mean square, the error mean square, their ratio, the F -statistic, and its P -value. There are usually other quantities of lesser interest included as well.
Assumptions for ANOVA (and conditions to check)	<ul style="list-style-type: none"> • Independence Assumption. (Think about the design of the experiment or, if an observational study, how the data were collected.) • Equal Variance Assumption. (Similar Variance Condition. Look at side-by-side boxplots to check for similar spreads, or look at residuals vs. predicted to see if the plot thickens.) • Normal Population Assumption. (Nearly Normal Condition. Check a histogram or Normal probability plot of the residuals.)
Assumptions for inference in regression (and conditions to check for some of them)	<ul style="list-style-type: none"> • Linearity. Check that the scatterplots of y against each x are straight enough and that the scatterplot of residuals against predicted values has no obvious pattern. (If we find the relationships straight enough, we may fit the regression model to find residuals for further checking.) • Independent errors. Think about the nature of the data. Check a residual plot. Any evident pattern in the residuals can call the assumption of independence into question. • Constant variance. Check that the scatterplots show consistent spread across the ranges of the x-variables and that the residual plot has constant variance too. A common problem is increasing spread with increasing predicted values—<i>the plot thickens!</i> • Normality of the residuals. Check a histogram or a Normal probability plot of the residuals.
Balance	An experiment's design is balanced if each treatment level has the same number of experimental units. Balanced designs make calculations simpler and are generally more powerful.
Bonferroni method	One of many methods for adjusting the length of the ME when testing the differences between several group means.

Vocabulary

Term	Definition
Cell	A cell of a two-way table is one element of the table corresponding to a specific row and a specific column. Table cells can hold counts, percentages, or measurements on other variables. Or they can hold several values.
Chi-square component	The components of a chi-square calculation are $(Observed - Expected)^2 / Expected$ found for each cell of the table.
Chi-square models	Chi-square models are skewed to the right. They are parameterized by their degrees of freedom, and become less skewed with increasing degrees of freedom.
Chi-square statistic	The chi-square statistic is found by summing the chi-square components. Chi-square tests can be used to test goodness-of-fit, homogeneity, or independence.
Conditions for inference in regression (and checks for some of them)	1. Straight Enough Condition for linearity. (Check that the scatterplot of y against x has linear form and that the scatterplot of residuals against predicted values has no obvious pattern.) 2. Independence Assumption. (Think about the nature of the data. Check a residuals plot.) 3. Does the Plot Thicken? Condition for constant variance. (Check that the scatterplot shows consistent spread across the range of the x -variable, and that the residuals plot has constant variance too. A common problem is increasing spread with increasing predicted values - the <i>plot thickens!</i>)
* Confidence interval for a predicted mean value	Different samples will give different estimates of the regression model and, so, different predicted values for the same value of x . We find a confidence interval for the mean of these predicted values at a specified x -value, x_v , as $y_v \text{ hat} \pm t_{n-2}^* \times SE(\mu_v \text{ hat})$ where $SE(\mu_v \text{ hat}) = \text{the square root of } [SE^2(b_1) \cdot (x_v - \bar{x})^2 + (s_e^2/n)]$. The critical value, t_{n-2}^* , depends on the specified confidence level and the Student's t -model with $n - 2$ degrees of freedom.
Confidence interval for the regression slope	When the assumptions are satisfied, we can find a confidence interval for the slope parameter from $b_1 \pm t_{n-2}^* \times SE(b_1)$. The critical value, t_{n-2}^* , depends on the confidence interval specified and on Student's t -model with $n - 2$ degrees of freedom.
Contingency table	A two-way table that classifies individuals according to two categorical variables is called a <i>contingency table</i> .
Error mean square (MSE)	The error mean square (MS_E) is the estimate of the error variance obtained by <i>pooling</i> the variances of each treatment group. The square root of the MS_E is the estimate of the error standard deviation, s_p .
F -distribution	The F -distribution is the sampling distribution of the F -statistic when the null hypothesis that the treatment means are equal is true. It has two degrees of freedom, one for the numerator, $k - 1$, and one for the denominator, $N - k$, where N is the total number of observations and k is the number of groups.
F -statistic	The F -statistic is the ratio MS_T/MS_E . When the F -statistic is sufficiently large, we reject the null hypothesis that the group means are equal.

Vocabulary

Term	Definition
<i>F</i> -test	The <i>F</i> -test tests the null hypothesis that all the group means are equal against the one-sided alternative that they are not all equal. We reject the hypothesis of equal means if the <i>F</i> -statistic exceeds the critical value from the <i>F</i> -distribution corresponding to the specified significance level and degrees of freedom.
Goodness-of-fit	A test of whether the distribution of counts in one categorical variable matches the distribution predicted by a model is called a test of goodness-of-fit. A chi-square test of goodness-of-fit finds $X^2 = \sum (\text{all cells}) (Obs - Exp)^2 / Exp$, where the expected counts comes from the predicting model. It finds a P-value from a chi-square model with $n - 1$ degrees of freedom, where n is the number of categories in the categorical variable.
Homogeneity	A test comparing the distribution of counts for two or more groups on the same categorical variable is called a test of <i>homogeneity</i> . A chi-square test of homogeneity finds $X^2 = \sum (\text{all cells}) (Obs - Exp)^2 / Exp$, where the expected counts are based on the overall frequencies, adjusted for the totals in each group. We find a P-value from a chi-square distribution with $(\#Rows - 1) \times (\#Cols - 1)$ degrees of freedom, where $\#Rows$ gives the number of categories and $\#Cols$ gives the number of independent groups.
Independence	A test of whether two categorical variables are independent examines the distribution of counts for one group of individuals classified according to both variables. A chi-square test of <i>independence</i> uses the same calculation as a test of homogeneity. We find a P-value from chi-square distribution with $(\#Rows - 1) \times (\#Cols - 1)$ degrees of freedom, where $\#Rows$ gives the number of categories in one variable and $\#Cols$ gives the number of categories in the other.
Least significant difference (LSD)	The standard margin of error in the confidence interval for the difference of two means is called the least significant difference. It has the correct Type I error rate for a single test, but not when performing more than one comparison.
Least squares	We still fit multiple regression models by choosing the coefficients that make the sum of the squared residuals as small as possible. This is called the method of least squares.
Minimum significant difference (MSD)	The Bonferroni method's ME for the confidence interval for the difference of two group means difference (MSD) is called the minimum significant difference. This can be used to test differences of several pairs of group means. If their difference exceeds the MSD, they are different at the overall α rate.
Multiple comparisons	If we reject the null hypothesis of equal means, we often then want to investigate further and compare pairs of treatment group means to see if they differ. If we want to test several such pairs, we must adjust for performing several tests to keep the overall risk of a Type I error from growing too large. Such adjustments are called methods for multiple comparisons.

Vocabulary

Term	Definition
Multiple regression	A linear regression with two or more predictors whose coefficients are found to minimize the sum of the squared residuals is a least squares linear multiple regression. But it is usually just called a multiple regression. When the distinction is needed, a least squares linear regression with a single predictor is called a simple regression. The multiple regression model is $y = \beta_0 + \beta_1x_1 + \dots + \beta_kx_k + \varepsilon$.
Partial regression plot	The partial regression plot for a specified coefficient is a display that helps in understanding the meaning of that coefficient in a multiple regression. It has a slope equal to the coefficient value and shows the influences of each case on that value. A partial regression plot for a specified x displays the residuals when y is regressed on the <i>other</i> predictors against the residuals when the specified x is regressed on the other predictors.
* Prediction interval for an individual	Different samples will give different estimates of the regression model and, so, different predicted values for the same value of x . We can make a confidence interval to capture a certain percentage of the entire distribution of predicted values. This makes it much wider than the corresponding confidence interval for the mean. The confidence interval takes the form $y_v \text{ hat} \pm t_{n-2}^* \times SE(y_v \text{ hat})$, where $SE(y_v \text{ hat}) =$ the square root of $[SE^2(b_1) \cdot (x_v - x \text{ bar})^2 + (s_e^2/n) + (s_e^2)]$. The critical value, t_{n-2}^* , depends on the specified confidence level and the Student's t -model with $n - 2$ degrees of freedom.
Residual standard deviation	The spread of the data around the regression line is measured with the residual standard deviation, $s_e =$ The square root of $[(\sum y - \hat{y})^2 / (n - 2)] =$ the square root of $(\sum e^2) / (n - 2)$.
Residual standard deviation	The residual standard deviation, $s_p =$ the square root of $MSE =$ the square root of $[(\sum e^2) / (N - k)]$ gives an idea of the underlying variability of the response values.
Scatterplot matrix	A scatterplot matrix displays scatterplots for all pairs of a collection of variables, arranged so that all the plots in a row have the same variable displayed on their y -axis and all plots in a column have the same variable on their x -axis. Usually, the diagonal holds a display of a single variable such as a histogram or Normal probability plot, and identifies the variable in its row and column.
Standardized residual	In each cell of a two-way table, a standardized residual is the square root of the chi-square component for that cell with the sign of the <i>Observed - Expected</i> difference: $(\text{Observed} - \text{Expected}) /$ the square root of <i>Expected</i> . When we reject a chi-square test, an examination of the standardized residuals can sometimes reveal more about how the data deviate from the null model.
t -ratios for the coefficients	The t -ratios for the coefficients can be used to test the null hypotheses that the true value of each coefficient is zero against the alternative that it is not.

Vocabulary

Term	Definition
Treatment mean square (MS_T)	The treatment mean square (MS_T) is the estimate of the error variance under the assumption that the treatment means are all equal. If the (Null) Assumption is not true, the MS_T will be larger than the error variance.
<i>t</i> -test for the regression slope	When the assumptions are satisfied, we can perform a test for the slope coefficient. We usually test the null hypothesis that the true value of the slope is zero against the alternative that it is not. A zero slope would indicate a complete lack of linear relationship between y and x . To test $H_0: \beta_1 = 0$ we find: $t = (b_1 - 0)/(SE(b_1))$ where $SE(b_1) = s_e/[(\text{square root of } n - 1)s_x]$, s_e = the square root of $[(\sum(y - \hat{y})^2)/(n - 2)]$, n is the number of cases, and s_x is the standard deviation of the x -values. We find the P-value from the Student's <i>t</i> -model with $n - 2$ degrees of freedom.
Two-way table	Each <i>cell</i> of a two-way table shows counts of individuals. One way classifies a sample according to a categorical variable. The other way can classify different groups of individuals according to the same variable or classify the same individuals according to a different categorical variable.