Print

Summary: Correlation and Regression

To go to the video where you learned a task or concept, select a link.

What Is Correlation?

Correlation is a measure of the strength of the linear relationship between two variables. The sample correlation coefficient, r, is used to quantify the strength of the linear association between two variables. Correlation is a unit-free measure that ranges from -1 to +1.

- The closer the correlation is to +1, the stronger the positive linear relationship.
- The closer the correlation is to -1, the stronger the negative linear relationship.
- The closer the correlation is to zero, the weaker the linear relationship, or association.

Interpreting Correlation

A strong correlation between variables does not necessarily mean that a change in the values of one variable is the cause of the change in the values of the other variable. Correlation does not imply causation. There might be a hidden, lurking variable that is not studied, but it makes the relationship seem stronger (or weaker) than it actually is.

A correlation coefficient of zero or near zero does not necessarily mean that there is no relationship between the variables. It simply means that there is no linear relationship.

Outliers can have a significant influence on the correlation. It's important to always plot the data before you interpret the correlation.

Introduction to Regression Analysis

In regression, and in statistical modeling in general, you want to model the relationship between an output variable, or a response, and one or more input variables, or factors. The term *regression* describes a general collection of techniques that are used when you model a response as a function of predictors.

When only one continuous predictor is used, you refer to the modeling procedure as *simple linear regression*. In simple linear regression, both the response and the predictor are continuous. Regression creates a statistical model that enables you to predict a response at different values of the predictor, including values of the predictor that are not included in the original data.

The Simple Linear Regression Model

In simple linear regression, you can assume that, for a fixed value of a predictor X, the mean of the response Y is a linear function of X. The regression line that you fit to the data is an estimate of this unknown function. You don't know the true line, and you fit a line for the one sample. This fitted line is only one estimate for the true model.

The Method of Least Squares

When you fit a regression line to a set of points, you assume that there is some unknown linear relationship between Y and X, and that for every one-unit increase in X, Y increases by some set amount on average.

The fitted regression line enables you to predict the response, Y, for a given value of X. However, for any specific observation, the actual value of Y can deviate from the predicted value. The deviations between the actual and predicted values are called *errors*, or *residuals*. The better the line fits the data, the smaller the residuals (on average).

If you add all the errors, the sum will be zero. To find the best regression line, square the errors and find a line that minimizes this sum of the squared errors. This method, the *method of least squares*, finds values of the intercept and slope coefficient that minimize the sum of the squared errors.

Regression Model Assumptions

You make a few assumptions when you use linear regression to model the relationship between a response and a predictor. These assumptions are essentially conditions that should be met before you make inferences regarding the model estimates or before you use a model to make a prediction.

- Assume that the relationship really is linear, and that the errors, or residuals, are simply random fluctuations around the true line.
- Assume that the variability in the response doesn't increase as the value of the predictor increases. This is the *assumption of equal variance*.
- Also assume that the observations are independent of one another.

To check regression assumptions, use residual plots to graph the residuals and look for any unusual patterns. The most useful graph is a *residual by predicted* graph. If the assumptions are met, the residuals are randomly scattered around the center line of zero, with no obvious pattern. The residuals resemble an unstructured cloud of points.

Interpreting Regression Results

All the variation in the response can be separated into either model sum of squares or error sum of squares. The sums of squares are reported in the ANOVA table. In the context of regression, the *p*-value reported in this table provides an overall test for the significance of the model. The *p*-value is used to test the hypothesis that there is no relationship between the predictor and the response.

The estimates in the Parameter Estimates table are the coefficients in the fitted model. The confidence interval for the slope coefficient provides an additional test for the size of the slope coefficient. This might be easier to interpret and explain than a *p*-value.

You can construct two types of intervals using the model: *confidence intervals* and *prediction intervals*. Confidence intervals, which are displayed as confidence curves, provide a range of values for the predicted mean for a given value of the predictor. Prediction intervals provide a range of

values where you can expect future observations to occur for a given value of the predictor.

RSquare provides a measure of the strength of the linear relationship between the response and the predictor. This statistic, which is between 0 and 1, measures the proportion of the total variation explained by the model. The closer RSquare is to 1, the more variation that is explained by the model. However, if you rely too heavily on RSquare, there are some cautions that are should discussed.

Fitting a Model with Curvature

A good practice, before you interpret statistical output, is to look at the graphical displays of the data and the residuals. The residual plot also provides insights into how you might improve the model. For example, to explain curvature in the residuals you could fit a second-order polynomial model to the data.

However, if you see a non-random pattern in the residuals you might need to apply a transformation to the response or the predictor, or you might be overlooking other important effects that explain the relationship.

Fitting the Multiple Linear Regression Model

In simple linear regression the method of least squares is used to find the best-fitting line for the observed data. When you have more than one predictor, this same least squares approach is used to estimate the values of the model coefficients.

In multiple linear regression, the significance of each term in the model depends on the other terms in the model, so it can be difficult to determine which predictors are important.

RSquare (R^2) is a measure of the variability in the response that is explained by the model. *RSquare Adjusted* is used when you fit multiple regression models.

The *root mean square error*, or RMSE, is a measure of the unexplained variation in the model. When the root mean square error is lower, the points are generally closer to the fitted line. For a predictive model, this corresponds to a model that predicts more precisely.

Interpreting Results in Explanatory Modeling

In explanatory modeling, you are generally interested in identifying the predictors that tell you the most about the response, and in understanding the magnitude and direction of the model coefficients. You want to know how the response values change as you change the values of a given predictor.

In multiple regression, you test the null hypothesis that all the regression coefficients are zero versus the alternative that at least one slope coefficient is nonzero. The *F* ratio is a statistical signal-to-noise ratio. It is a ratio of the variation explained by the model (mean square model) and the unexplained variation (mean square error).

When there is no relationship between the response and any of the predictors, the model does not explain much of the variation in the response. Mean square model and mean square error are approximately the same, and the *F* ratio is close to 1. If the alternative hypothesis is true, at least one coefficient is nonzero. The model explains at least some of the variation in the response. Mean square model is greater than mean square error, and the *F* ratio is greater than 1.

The *F* ratio and the corresponding *p*-value are reported in the ANOVA table. The ANOVA table enables you to make decisions about the significance of the model on the whole, but it doesn't tell you which predictors are significant. For this, you use the information that is reported in

the Effects Test table. This information is also reported in the Effect Summary table.

The F ratios and *p*-values provide information about whether each individual predictor is related to the response. These tests are known as *partial tests* because each test is adjusted for the other predictors in the model.

Residual Analysis and Outliers

Conduct a residual analysis to verify that the conditions for drawing inferences about the coefficients in a linear model are met.

Residual by predicted plots, *normal quantile* plots, and *residual by row number* plots are used to check these assumptions. Studentized residuals are more effective for detecting outliers and for assessing the equal variance assumption. The *Studentized Residual by Row Number* plot essentially conducts a *t* test for each residual. Studentized residuals that are outside the red limits are potential outliers.

An observation is considered an outlier if it is extreme, relative to other response values. Some outliers or high leverage observations exert influence on the fitted regression model. This creates a bias about the model estimates. A statistic referred to as *Cook's D*, or *Cook's distance*, helps identify influential points. Cook's D measures how much the model coefficient estimates changes if an observation were removed from the data set. The higher the Cook's D value, the greater the influence. Generally accepted rules are that Cook's D values above 1.0 indicate influential values, and any values that obviously differ from the rest might also be influential.

Multiple Linear Regression with Categorical Predictors

To integrate a two-level categorical variable into a regression model, create one indicator or dummy variable with two values. Assign 1 for the first shift and -1 for the second shift. Notice that, instead of using -1/1 effect coding, many software applications apply 0/1 dummy coding. That is, they assign a zero for the first shift and a one for the second shift. These two coding schemes result in the same model predictions. However, from an explanatory perspective, the interpretation of the coefficients is different.

For a k-level categorical predictor, the software computes *k*-1 coefficients.

RSquare can be inflated by adding more terms to the model, even if these new terms are not significant. So, in multiple linear regression situations, you use *RSquare Adjusted* when you compare different models with the same data instead of using RSquare. RSquare Adjusted applies a penalty for each additional term, *p*, that is added to the model. If a term is added to the model that does not explain variation in the response, the RSquare Adjusted value decreases.

Multiple Linear Regression with Interactions

It is possible that the effect of one predictor on the response is dependent on the values of another predictor. This dependency is known in statistics as an *interaction effect*.

You can visualize interactions using interaction plots. Each interaction plot in this matrix shows the interaction of the row effect with the column effect. For each pair of variables, there are two interaction plots. They enable you to visualize the interactions from different perspectives. Understanding interactions is important because it provides additional insights into the response. The Prediction Profiler is extremely useful for exploring and interpreting models with interactions.

Variable Selection

This task of identifying the best subset of predictors to include in the model, among all possible subsets of predictors, is referred to as *variable selection*.

One approach is to fit a full model by starting with the term with the highest *p*-value and slowly removing terms one at a time. This is referred to as *backward selection*.

An alternative to backward selection is *forward selection*. Instead of starting with a full model, you start with a model that contains only the intercept. Then, starting with the predictor with the lowest *p*-value, you slowly add terms to the model, one at a time. This continues until all the remaining terms that are not included in the model are greater than a specified *p*-value threshold.

A third classic variable selection approach is *mixed selection*. This is a combination of forward selection (for adding significant terms) and backward selection (for removing nonsignificant terms).

Another approach is *best subsets regression*, or *all possible models*. Here, you fit all possible models from the combinations of the potential predictors. In best subsets regression, you fit all potential models from the set of predictors and then compare the models to choose the one "best" model.

Multicollinearity

The term *collinearity*, or *multicollinearity*, refers to the condition in which two or more predictors are highly correlated with one another. Collinearity can make it difficult to determine the effect of each predictor on the response and can make it challenging to determine which variables to include in the model.

One method for detecting whether collinearity is a problem is to compute the *variance inflation factor*, or *VIF*. This is a measure of how much the standard error of the estimate of the coefficient is inflated due to multicollinearity. The smallest possible value of VIF is 1.0, and that indicates a complete absence of collinearity. A VIF of 5 or 10 indicates that the collinearity might be problematic.

What Is Logistic Regression?

Logistic regression is used to model categorical response data. When multiple predictors are used, the predictors can be continuous or categorical.

The logistic regression model predicts the probability of a particular response category rather than predicting a mean response value. Because the model predicts probabilities, the predicted value for a category must fall between zero and one, and, the predicted probabilities for the different response categories must sum to one.

In logistic regression with a binary response, observations can be one of two possible categories. Logistic models estimate probabilities of membership in one of these two categories.

The Simple Logistic Model

The logistic model relates the log of the odds, or simply the log-odds, to a linear model in the predictors. This is called the log-odds of *p*, or the *logit*. The logistic model predicts the log-odds of an event as a linear function of the predictors, but you can reorganize this formula to calculate the predicted probability of the event.

You can graph the logistic model to visualize the probabilities for different values of the intercept and the slope. The s-shaped or sigmoidal curve in the graph, which is called the logistic curve, shows how the predicted probability changes as you increase the value of the predictor.

This curve provides a graphical measure of the strength of the relationship between the predictor and the categorical response variable. The stronger the relationship, the steeper the logistic curve.

Interpreting Logistic Regression Results

In linear regression, you used the method of least squares to find the best fitting model and estimate the model coefficients. Least squares is actually a special case of a more general approach for estimating statistical models, the *method of maximum likelihood*. The basic idea behind maximum likelihood is to find estimates for the model coefficients that are the most consistent with the observed data.

The estimated coefficients that are reported in the Parameter Estimates table are for the log odds, or the logit. You can use this formula to compute the probabilities for the outcome for a given value of the predictor. From the predicted probabilities, you can make classifications of the most likely outcome. Based on the predicted probability of membership in one of the two classes, the predicted class is the one with the highest probability.

Comparison of the actual class to the predicted class provides a measure of the performance of the model. The results are often summarized in a table that is referred to as a confusion matrix. You use the *misclassification* rate as a measure of the predictive performance of logistic regression models.

Multiple Logistic Regression

As you saw with multiple linear regression, you can include many predictors in one model. The predictors can be continuous or categorical, and you can also include interaction terms. When the response is categorical, the modeling procedure is called multiple logistic regression.

A test for the significance of the model as a whole is reported in the Whole Model Test report. Here, you are testing that the full model with the specified predictors is better than the reduced model with only the intercept and no predictors.

To judge model performance, you use the confusion matrix and the misclassification (or *error*) rate. You can also state the performance in terms of correct classifications. This is the *accuracy rate*.

The Parameter Estimates table reports test statistics and *p*-values for the model coefficients.

Common Issues

When you fit logistic models with categorical predictors and interactions you can encounter problems with estimating model coefficients if you don't have enough data. If you don't have data for all the combinations of the response and predictors that are involved in the interactions, you can't properly estimate the coefficients. How much data you need can get a bit complicated. It depends on the structure of the data set and the proportion of the observations in the target category.

Continuous variables have much more information content than categorical variables. In general, regression models with continuous responses require much less data than models with categorical responses.

Another common issue in logistic regression is called *separation*. This occurs if a variable is a perfect, or a near perfect, predictor of the response. The logistic curve is a nearly vertical line that separates the zeros from the ones. When this occurs, the slope coefficient approaches

infinity, and it can't properly be estimated. The Parameter Estimates table reports both the slope and the intercept as unstable.