

13.7: The Purpose of the Four Parts of a Regression Analysis

So far we have reviewed the types of hypotheses and data which are a good fit to regression, what prediction means in regression, and the logic behind using the regression line to predict a Y -value. However, a regression line should only be used to make predictions when it is sufficiently useful in doing so. Thus, we need to understand how to test a hypothesis to know *if* X is useful for predicting Y before we use it to do so.

Regression is a complex technique which uses three sets of analyses to test a hypothesis and establish whether X is useful in predicting Y . These three sets of analyses are: Correlation, ANOVA, and a t -Test. When it is determined that X significantly predicts Y , a fourth component can then be used to make predictions. This component is the linear equation of

$$\hat{Y} = b_0 + b_1x$$

Thus, regression is actually a technique that draws from other existing techniques and puts them together to serve the new purpose of predicting.

It can be helpful to have a broad idea of the major parts of regression and their purposes before going into details about each. Therefore, we will start with a brief overview of the role of each of the components of regression before going into detail about how each is computed.

The four components are:

1. Correlations to establish whether X relates to Y ,
2. ANOVA to test whether using X significantly reduces error in predicting Y ,
3. t -testing to assess whether it is the slope of the line that is reducing the error in predictions and, when warranted,
4. The linear equation to make predictions.

Each of these four components answers a different question we are posing when using regression.

Correlation in Regression

Correlation is used in regression to answer the question:

Does X relate to Y ?

Statisticians will often check whether there is a significant relationship between an X -variable and a Y -variable before progressing to testing predictions. When X relates to Y , the coefficient of determination may also be computed and reported as part of a test using regression.

Importantly, the scatterplot with the fit line from correlation is a visual depiction of a regression model (Recall that a fit line in regression is called a *regression line*.). “Regression model” simply refers to how X is being used to predict Y with the linear equation. The regression line is the foundation from which predictions are made in regression. Thus, when the correlation is significant between X and Y , it offers support for a regression model which uses X to predict Y . For a detailed review of correlation, see Chapter 12.

Though correlations are often checked in regression, they are not always reported for bivariate regression. This is because the results of the correlation are redundant to the results of the ANOVA and t -test when there is only one X -variable (predictor) being tested in a regression model. Instead, the coefficient of determination (r^2) is often checked and reported for simple regression and r can be checked but does not generally need to be reported.

ANOVA in Regression

ANOVA is used in regression to answer the question:

Does using X significantly improve predictions of Y ?

ANOVA is a very important part of a regression because it is used to test the central aspect of the hypothesis: whether X predicts Y . The ANOVA portion of regression is used to compute how good X is at predicting Y by assessing whether it significantly reduces the error of predictions compared to an alternative prediction model. Thus, improvement in predictions is defined and estimated as reduction in errors in predictions when using the regression model. By *regression model* we simply mean a model where X is used to predict Y . The greater the reduction of error when using X to predict Y , the more useful the regression model is and, thus, the larger the ANOVA F -value will be.

The regression model must be compared to an alternative model. The default alternative way to predict any Y -value is to simply use the mean of Y for all predictions. The symbol for the mean of Y is \bar{Y} . When the regression ANOVA is significant, it indicates that using X significantly improves predictions of Y over using \bar{Y} as the prediction for all Y -values. Thus, when the ANOVA is significant, it offers support for a regression model which uses X to predict Y . This concept is central to testing a regression hypothesis and, thus, we will focus on this in quite a bit of detail in this chapter.

t-Test in Regression

A t -test is used in regression to answer the question:

Does the slope of the regression line significantly improve the predictions of Y ?

The slope of the regression line is what allows the line to get as close to all the data points as possible, on average. Regression models use one or more X -variables to predict a Y -variable. For this chapter we are focused on bivariate regression where there is only one X -variable. However, when multiple X -variables are used to predict Y (which is common in research), they will each contribute a different slope. In regression, t -tests are used to assess which of those slopes significantly improved predictions and which did not. Think of this as the post-hoc part of a regression; if the regression ANOVA is significant, the t -tests are used to see which X -variables had slopes that significantly contributed to the accuracy of predictions. The improvements gained by the slope of each X -variable are tested with separate t -tests. However, in bivariate linear regression, only one X -variable is being tested meaning only one slope is being used to predict Y . In this case, if the ANOVA is significant, the t -test will also be significant. Therefore, assessing the significance of the t -test is less essential for a bivariate regression than in multivariate regression. However, it should still be assessed and reported as part of a complete regression analysis.

In addition, the slope and y -intercept for the regression formula are computed when using SPSS and can then be reported with the t -test results. Thus, we will review how to read and interpret results for the t -test component of regression to get necessary information about the slope and the y -intercept.

Making Predictions in Regression

When a regression is significant, it establishes that using the regression line and X -values are useful in predicting Y . It follows then that the regression equation should be used to predict Y when regression results are significant. Thus, the last part of the regression is interpreting how the slope predicts Y and creating the corresponding regression equation which can be used to predict Y . Predictions of Y are made using the linear equation as follows:

$$\hat{Y} = b_0 + b_1 x$$

In this equation, a predicted Y value (\hat{Y}) is computed by multiplying the slope of the regression line (b_1) by an X -value (x) and then adding the y -intercept (b_0). Slopes and intercepts can be computed by hand or using SPSS. Later in this chapter, we will review how to find and interpret these using SPSS. In addition, slopes are typically interpreted and included as part of a complete APA-formatted results paragraph for regression. The regression equation can also be constructed and used to predict Y -values, if desired, but it is not always included in a results paragraph.

1. What is correlation used to check in a regression?
2. What does ANOVA compare when used to test a regression model?
3. What part of the regression model is tested by the t -test in a regression?
4. Under what conditions and for what purpose is the linear equation used with regression?

This page titled [13.7: The Purpose of the Four Parts of a Regression Analysis](#) is shared under a [CC BY-NC-SA 4.0](#) license and was authored, remixed, and/or curated by .