

BIOL 458 BIOMETRY

Lab 9 - Bivariate Regression

Introduction to Regression

The procedures discussed in the previous ANOVA labs are most useful in cases where we are interested in testing hypotheses about differences in the locations of several populations in terms of a single random variable. However, we may also be interested in examining the relationship between two different random variables, X and Y , measured on each subject sampled in a single population. This relationship is known as a bivariate relationship. If two variables are related in such a way that the value of one the variables X is useful in predicting the value of the variable Y , then there may be a significant linear regression of variable Y on variable X .

Linear regression and correlation analyses are useful in evaluating the association between variables and expressing the nature of their relationship. In regression, the relationship between a dependent variable, Y and an independent or predictor variable X is examined.

For example, consider the relationship between crop yield and precipitation. Yield (Y) is a function of precipitation (X), since we hypothesize that water availability affects plant growth, but on average plant growth does not affect precipitation. Using linear regression, we can quantify the observed relationship between the two variables. We might ask if there is a significant regression of yield on precipitation, indicating that yield can be predicted from knowledge of precipitation, or if no regression exists and yield cannot be predicted from knowledge of precipitation. In bivariate regression, we assume that the relationship between variables can be described by a straight line. The line relating two variables X and Y is described by the equation:

$$Y = b_0 + b_1x + \varepsilon$$

where, b_0 is called the intercept, corresponding to the point where $X = 0$ and the line intercepts the Y axis, and b_1 is the slope, the change in Y per unit change in X . The independent variable, X , is used to predict Y , the dependent variable. b_0 and b_1 are the regression coefficients or the parameters of a line which fix and define the linear relationship between Y and X . ε is the "error" or that component of the variation in the values of Y that cannot be predicted by the regression of Y on X . ε arises both because the fit of the regression equation or regression model to the data may be inadequate and because there is inherent variability in the values of Y observed at each value of X .

The differences (errors) between the actual values of Y and the values predicted by the regression equation - a line fitting the data - are called residuals. The estimation of the parameters is performed by finding the "best fit" regression line, the line that minimizes the sums - of - squares of the observed values from the predicted values. This method is called the method of Ordinary Least Squares (OLS).

Assumptions of Linear Regression

In using linear regression, a number of assumptions must be made; these are discussed in detail in lecture. In summary, it is assumed that:

1. each value of X is measured without error;
2. the set of observed (X, Y) values consists of n independent measures;
3. ε is a normally distributed error with a mean of zero and some standard deviation σ_e^2
4. Y is a linear function of X .

If we make these assumptions, fitting the regression model is very simple; however, to use the regression model to predict or estimate values of Y , we must test the assumptions we have made. The residuals are the difference between the observed and predicted values of Y . These deviations can be used as a tool to see if the necessary assumptions for regression have been met, and to further investigate the adequacy or goodness - of - fit of the regression model.

If the assumptions have been met, the residuals (errors) should be independent and, for each value of X , the set of possible residuals should be approximately normally distributed with a mean of zero and a variance that is not a function of X . If the residuals do not have these characteristics, then some of the assumptions made in fitting the model must be incorrect and the results of the regression cannot be deemed valid. Therefore, it is not reasonable to accept a regression without examining whether the assumptions are met.

There are many methods that can be used in evaluating the residuals from a regression equation. Some of these methods are objective hypothesis tests; the alternative is to graph the residuals versus the values of X and to evaluate them subjectively. The properties to be evaluated are: 1) independence, 2) normality, and 3) constant variance.

The property of independence can be tested in a variety of ways. Basically, though, we can assume that, if the errors are independent, the errors will not tend to have any pattern; they will be random. If they are not random, this fact will

be evident because there will be a few long series of either positive or negative values, instead of numerous shorter series. A hypothesis test known as a "runs" test can be used to evaluate the randomness of the residuals; alternately, a subjective evaluation can be performed. Other objective tests are also available.

Normality is another desirable property in the residuals; however, the assumptions underlying the regression model do not require that the set of residuals have a normal distribution. What is required is that for any particular value of X the set of possible errors should be normally distributed. Unless a large number of measurements of Y have been obtained for each of the values of X , there is no reasonable way of testing this assumption. Therefore, the normality of the residuals is usually not evaluated.

The requirement of a constant variance is very important in evaluating regression models. Of the four residual properties listed above, this is the one that is tested most often. However, unless the number of residuals is relatively large, a subjective visual evaluation is usually all that can be performed. Visually, residuals that suggest that the variances are constant look like a rectangular scatter that is evenly spread about the regression line all along its length. However, when samples sizes are small it is often difficult to make an accurate judgment about constancy of variances based on a visual inspection of the residuals.

Introduction to Correlation

A correlation coefficient measures the strength of association between two variables. It takes on values between -1 and +1 with the sign indicating negative or positive correlations, respectively. Correlation is related to Regression, but correlation analyses make different assumptions about the data. First in correlation, there is no independent or dependent variable, so one is not predicting Y from X as in Regression. For parametric correlation analyses one assumes independence and bi-variate normality of the coordinate pairs. NO assumptions are made about the variances in Y or X . Significance tests are available to establish that the estimated correlation is unlikely by chance at some α level.

Further Instructions for Lab 9

Data files for regression and correlation require that each subject be represented by a line in the data file and each column represents a variable. So, for correlation or bivariate regression (with two variable y and x), an **SPSS** data file need only have 2 columns of values. However, if you have more than two variables for a single set of subjects for which you want to calculate their correlations, just enter all the variables in separate columns and **SPSS** can calculate the correlations between the variables in each pair of columns - a correlation matrix.

Performing a Correlation Analysis in SPSS

To perform a correlation analysis in **SPSS**, go to the **Analyze** menu and select **Correlate**, and then select **Bivariate**. The **Bivariate Correlations** sub-window will open. The left hand box will list all variables in the data file. Click over to the right hand box all variable for which you want correlations. You must click over at least two variable names. Check the boxes below to choose which correlation coefficients you want, Pearson, Kendall, or Spearman (parametric and non-parametric). We usually want Pearson and/or Spearman's. Both can be obtained in one analysis.

The output table from the correlation analysis is a square 2 x 2 table with correlations between var 1 and var 2 and between var 2 and var1 (of course these are equal), along with the sample size and significance level. These values are on the "off diagonal" of the matrix. On the diagonal (from left to right), one finds the correlations between var2 and var 2 and var 1 and var 1 which by definition equal 1. Ignore these values on the diagonal. If you had 3 variables to correlate you would obtain a similar table, but with 3 x 3 cells in the table. The values above the diagonal would equal the values below ($\text{corr}(\text{var2}, \text{var3}) = \text{corr}(\text{var3}, \text{var2})$).

Bivariate Regression in SPSS

To perform a regression analysis in **SPSS**, from the **Analyze** menu choose **Regression**, and from the submenu choose **Linear**. The **Linear Regression** sub-window will open. This regression module is capable of performing bivariate (one Y and one X variable) and multiple regression (one Y, and 2 or more X variables), so we will deal with some of the options later. To perform a Bivariate regression click over the dependent or Y variable from the variables box to the dependent box. Click over the independent or X variable to the independent box. No other options or specifications have to be made to perform the regression, although you may decide to request some diagnostic information to help determine if the fitted regression line is a good fit. Click on the Plot button to obtain the Plot sub-window. Choose ***zresid** and click it into the Y box. Choose ***zpred** and click it into the X box. This will give you a plot of your standardized residuals versus the standardized predicted values. This can be used to determine if you meet the assumptions of linearity and homogeneity of variances. Click on **Continue** to return to the **Linear Regression** sub-window. Click on statistics and choose **Casewise Diagnostics**. This will give you information about whether any of your residuals are excessively large. Click **Continue** to return to the **Linear Regression** sub-window. Now you are ready to run the regression, so click on **Ok**. If you followed all the instructions I gave above, then you should get 5 tables and one graph from SPSS. The first table just tells you which X variables were in the regression model. And which variable was your

dependent variable. The second table gives you the estimates of r , of R^2 , and the standard error of the estimate. The third table gives you the ANOVA of the regression. The fourth table gives you the estimates of the regression coefficients and their standard errors (un-standardized coefficients) and t tests of whether or not the coefficients differ from 0. The line labeled "constant" will be the y-intercept estimate, and the line with the variable name will be the slope estimate. The last table will contain information about the residuals and predicted values. Only the residuals are of interest. Any residuals above 3 might suggest that your model is not adequate. The last item in the output will be the residual plot. Examine this plot for evidence that your data fail to meet the assumption of linearity or homogeneity of variances (patterns in the residuals).

LAB - 9 Assignment

PART 1: Introduction to Regression

The Bermuda Petrel is an oceanic bird spending most of its year on the open sea, only returning to land during the breeding season. Its nesting sites are on a small, uninhabited island of the Bermuda group, where careful hatching records have been kept over several years. The Bermuda Petrel feeds only upon fish caught in the open ocean waters far from land. Unfortunately, DDT is now so widespread, and is so concentrated by the biological amplification system known as the "food chain," that the Bermuda Petrel can no longer lay hard shelled eggs. Since DDT breaks down so slowly, it would appear that this beautiful bird is doomed to extinction (along with how many others?)

You have the following information about hatching rates.

- a) What is the independent variable? (Predictor, X Variable)
- b) What is the dependent variable? (Dependent, predicted variable)
- c) Use simple linear regression in **SPSS** to see if there is a significant relationship between the percent of clutches hatching over time. Interpret the output. Also produce a scatter plot of the relationship between hatching and year.

Year	1966	1967	1968	1969	1970	1971	1972	1973
% of Clutches Hatching	80	60	67	39	48	37	35	17

PART 2: Assumptions of simple linear regression

- a) Estimate the linear regression model for each of the three sample data sets ([req1](#), [req3](#), [req5](#)) using the **Regression** Command, option **Linear**, under the **Analyze** Menu is **SPSS**. Use data in Column 1 as the X -variate and column 2 as the Y -variate in each data file.
- b) Write the regression equation for at least two of the data sets.
- c) Reiterating from the lab, the null hypothesis to be tested in each instance states that Y is not a linear function of X , and thus X will not be a good predictor of Y . More specifically, under the null hypothesis we are testing that the slope, b_1 , will be equal to zero, since this would be indicative of no relationship between the two variables. At the $\alpha = 0.05$ level, based on the output of the regression alone (F - test) for which of the five data sets would you reject the null hypothesis?
- d) Based on the R^2 values, which model reveals the best fit?
- e) To see if the models are adequate, you must test to see if the assumptions of the regression have been met. For each data set, plot Y versus X . Looking at the plots; is it appropriate to assume a linear relationship between the variables of each data set? Alternatively, plot the standardized residual on the Y -axis ($zresid$) against the standardized predicted ($zpred$) values on the X -axis using the “plots” option in the **Regression - Linear** module. For which data sets is linear regression applicable, and for which data sets is it clear that a linear regression model should not be imposed on the data? (e.g. would some transformation of the data make these data linear?)
- f) Further checking the assumptions, examine plots of the standardized residuals in relation to the standardized predicted values. If the variance of the errors is constant, then no pattern (in other words the pattern will be random) will emerge on the plots. If this assumption has been violated, you would expect to see the spread of the residuals increasing or decreasing with values of the predicted variables or independent variables. Evaluate the randomness and the homoscedasticity of each set of residuals. Do any of the plots reveal that linear regression cannot be applied, i.e. the assumptions have been violated?
- g) In conclusion, for which data sets are the fitted regression models adequate and for which are they inadequate? Could these regressions be made to meet the assumptions if either or both Y and X variate were transformed?