Contents

Determining Regression Estimates

The calculations required for determining the values of the parameter estimates in a multiple regression equation and the associated standard error values are quite complex and generally involve matrix algebra. However, computer software, such as `R`, is widely available for carrying out such calculations.

1 SLR Example

The data give the yields of cotton and irrigation levels in the Salt River Valley for different plots of land. Each plot was on Maricopa sandy loam soil. The variables are as follows:

- **Irrigation** The amount of irrigation water applied in feet per acre. This is the predictor variable.
- **Yield** The yield of Pima cotton in pounds per acre. This is the response variable.

Observation	Irrigation	Yield	Observation	Irrigation	Yield
1	1.8	260	8	1.5	280
2	1.9	370	9	1.5	230
3	2.5	450	10	1.2	180
4	1.4	160	11	1.3	220
5	1.3	90	12	1.8	180
6	2.1	440	13	3.5	400
7	2.3	380	14	3.5	650

Table 1: Add caption

Descriptive Statistics

	Mean	Std. Deviation	N
Yield	306.4286	149.6461	14
Irrig	1.971429	0.754911	14

Next we are given the output from the correlation analysis and the regression ANOVA. The intercept and slope estimate are determined by examining the "coefficients".

2 Correlation

Pearson's correlation coefficient (r) is a measure of the strength of the 'linear' relationship between two quantitative variables. A major assumption is the normal distribution of variables. If this assumption is invalid (for example, due to outliers), the non-parametric equivalent Spearman's rank correlation should be used.

2.1 Formal test of Correlation

2.2 Lurking variables and Spurious Correlation

Spurious Correlations. Although you cannot prove causal relations based on correlation coefficients, you can still identify so-called spurious correlations; that is, correlations that are due mostly to the influences of "other" variables. For example, there is a correlation between the total amount of losses in a fire and the number of firemen that were putting out the fire; however, what this correlation does not indicate is that if you call fewer firemen then you would lower the losses. There is a third variable (the initial size of the fire) that influences both the amount of losses and the number of firemen. If you "control" for this variable (e.g., consider only fires of a fixed size), then the correlation will either disappear or perhaps even change its sign. The main problem with spurious correlations is that we typically do not know what the "hidden" agent is. However, in cases when we know where to look, we can use partial correlations that control for (partial out) the influence of specified variables.

2.3 Simpson's Paradox

2.4 Rank correlation

Spearman's Rank correlation coefficient

2.5 Partial Correlation

Partial correlation analysis involves studying the linear relationship between two variables after excluding the effect of one or more independent factors.

3 Multiple Linear Regression

Multiple regression: To quantify the relationship between several independent (predictor) variables and a dependent (response) variable. The coefficients $(a, b_1 tob_i)$ are estimated by the least squares method, which is equivalent to maximum likelihood estimation. A multiple regression model is built upon three major assumptions:

1. The response variable is normally distributed,
2. The residual variance does not vary for small and large fitted values (constant variance),
3. The observations (explanatory variables) are independent.

3.1 Dummy Variables

A dummy variable is a numerical variable used in regression analysis to represent subgroups of the sample in your study. In research design, a dummy variable is often used to distinguish different treatment groups. In the simplest case, we would use a 0,1 dummy variable where a person is given a value of 0 if they are in the control group or a 1 if they are in the treated group. Dummy variables are useful because they enable us to use a single regression equation to represent multiple groups. This means that we don't need to write out separate equation models for each subgroup.

3.2 Estimates

Assumptions

The assumptions of multiple linear regression analysis are similar to those of the simple case involving only one independent variable. For point estimation, the principal assumptions are that

(1) the dependent variable is a random variable,

(2) the relationship between the several independent variables and the one dependent variable is linear. Additional assumptions for statistical inference (estimation or hypothesis testing) are that

(3) the variances of the conditional distributions of the dependent variable, given various combinations of values of the independent variables, are all equal,

(4) the conditional distributions of the dependent variable are normally distributed, and

(5) the observed values of the dependent variable are independent of each other. Violation of this assumption is called autocorrelation.

4 Training and validation

Using Validation and Test Data

In some cases you might want to use only training and test data. For example, you might decide to use an information criterion to decide what effects to include and when to terminate the selection process. In this case no validation data are required, but test data can still be useful in assessing the predictive performance of the selected model. In other cases you might decide to use validation data during the selection process but forgo assessing the selected model on test data.

5 Multiple Linear Regression

5.1 What is Multiple Linear Regression

Multiple regression is a statistical technique that allows us to predict a numeric value on the response variable on the basis of the observed values on several other independent variables.

Suppose we were interested in predicting how much an individual enjoys their job. Variables such as salary, extent of academic qualifications, age, sex, number of years in full-time employment and socioeconomic status might all contribute towards job satisfaction. If we collected data on all of these variables, perhaps by surveying a few hundred members of the public, we would be able to see how many and which of these variables gave rise to the most accurate prediction of job satisfaction. We might find that job satisfaction is most accurately predicted by type of occupation, salary and years in full-time employment, with the other variables not helping us to predict job satisfaction.

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \dots$$

- $\hat{y}$ is the ***fitted value*** for the dependent variable Y, given a linear combination of values for the independent valriables.
- x_1 is the value for independent variable X_1.
- b_o is the constant regression estimate (commonly known as the **Intercept Estimate** in the case of simple linear regression).

Multiple Linear Regression

Multiple regression analysis is an extension of simple regression analysis, as described previously, to applications involving the use of two or more ***independent variables*** (predictors) to estimate the value of the ***dependent variable*** (response variable). In the case of two independent variables, denoted by X1 and X2, the linear algebraic model is

$$\hat{Y}_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \varepsilon_I$$

The definitions of the above terms are equivalent to the definitions in previous classes for simple regression analysis, except that more than one independent variable is involved in the present case.

Based on sample data, the linear regression equation for the case of two independent variables is

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2$$

The multiple regression equation identifies the best-fitting line based on the method of least squares. In the case of multiple regression analysis, the best-fitting line is a line through n-dimensional space (3-dimensional in the case of two independent variables).

6 Terminology

6.1 Beta (standardised regression coefficients)

The beta value is a measure of how strongly each predictor variable influences the response variable. The beta is measured in units of standard deviation. For example, a beta value of 2.5 indicates that a change of one standard deviation in the predictor variable will result in a change of 2.5 standard deviations in the response variable. Thus, the higher the beta value the greater the impact of the predictor variable on the response variable.

The Standardized Beta Coefficients give a measure of the contribution of each variable to the model. A large value indicates that a unit change in this predictor variable has a large effect on the criterion variable. The t and Sig (p) values give a rough indication of the impact of each predictor variable a big absolute t value and small p value suggests that a predictor variable is having a large impact on the criterion variable.

7 ANOVA

In ANOVA we are trying to determine how much of the variance is accounted for by our manipulation of the independent variables (relative to the percentage of the variance we cannot account for).

8 Information Criterions

We define two types of information criterion: the Bayesian Information Criterion (BIC) and the Akaike Information Criterion (AIC). In AIC and BIC, we choose the model that has the minimum value of:

$$AIC = 2log(L) + 2m,$$

$$BIC = 2log(L) + mlogn$$

where

- L is the likelihood of the data with a certain model,
- n is the number of observations and
- m is the number of parameters in the model.

8.1 AIC

The Akaike information criterion is a measure of the relative **goodness of fit** of a statistical model.

When using the AIC for selecting the parametric model class, choose the model for which the AIC value is lowest.