# Contents

# 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

# 4 Multicollinearity

In multiple regression, two or more predictor variables are colinear if they show strong linear relationships. This makes estimation of regression coefficients impossible. It can also produce unexpectedly large estimated standard errors for the coefficients of the X variables involved.

This is why an exploratory analysis of the data should be first done to see if any collinearity among explanatory variables exists. Multicolinearity is suggested by non-significant results in individual tests on the regression coefficients for important explanatory (predictor) variables. Multicolinearity may make the determination of the main predictor variable having an effect on the outcome difficult.

## 4.1 How to Identify Multicollinearity

You can assess multicollinearity by examining **tolerance** and the **Variance Inflation Factor** (VIF) are two collinearity diagnostic factors that can help you identify multicollinearity. Tolerance is a measure of collinearity reported by most statistical programs such as SPSS; the variables tolerance is 1-R2. A small tolerance value indicates that the variable under consideration is almost a perfect linear combination of the independent variables already in the equation and that it should not be added to the regression equation. All variables involved in the linear relationship will have a small tolerance. Some suggest that a tolerance value less than 0.1 should be investigated further. If a low tolerance value is accompanied by large standard errors and nonsignificance, multicollinearity may be an issue.

## 4.2 The Variance Inflation Factor (VIF)

The Variance Inflation Factor (VIF) measures the impact of collinearity among the variables in a regression model. The Variance Inflation Factor (VIF) is 1/Tolerance, it is always greater than or equal to 1. There is no formal VIF value for determining presence of multicollinearity. Values of VIF that exceed 10 are often regarded as indicating multicollinearity, but in weaker models values above 2.5 may be a cause for concern. In many statistics programs, the results are shown both as an individual R2 value (distinct from the overall R2 of the model) and a Variance Inflation Factor (VIF). When those R2 and VIF values are high for any of the variables in your model, multicollinearity is probably an issue. When VIF is high there is high multicollinearity and instability of the b and beta coefficients. It is often difficult to sort this out.

You can also assess multicollinearity in regression in the following ways:

(1) Examine the correlations and associations (nominal variables) between independent variables to detect a high level of association. High bivariate correlations are easy to spot by running correlations among your variables. If high bivariate correlations are present, you can delete one of the two variables. However, this may not always be sufficient.

(2) Regression coefficients will change dramatically according to whether other variables are included or excluded from the model. Play around with this by adding and then removing variables from your regression model.

(3) The standard errors of the regression coefficients will be large if multicollinearity is an issue.

(4) Predictor variables with known, strong relationships to the outcome variable will not achieve statistical significance. In this case, neither may contribute significantly to the model after the other one is included. But together they contribute a lot. If you remove both variables from the model, the fit would be much worse. So the overall model fits the data well, but neither X variable makes a significant contribution when it is added to your model last. When this happens, multicollinearity may be present.

### 4.3 Variance Inflation Factor

The variance inflation factor (VIF) quantifies the severity of multicollinearity in a regression analysis.

The VIF provides an index that measures how much the variance (the square of the estimate's standard deviation) of an estimated regression coefficient is increased because of collinearity.

A common rule of thumb is that if the VIF is greater than 5 then multicollinearity is high. Also a VIF level of 10 has been proposed as a cut off value.

## 5 Law of Parsimony

Parsimonious: The simplest plausible model with the fewest possible number of variables.

## 6 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.

## 7 Multiple Linear Regression

### 7.1 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).

# 8 Terminology

## 8.1 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.

# 9 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).

# 10 Variable Selection Procedures

- Enter: This is the forced entry option. SPSS will enter at one time all specified variables regardless of significance levels.
- Forward: This method will enter variables one at a time, based on the significance value to enter.
- Backward: This enters all independent variables at one time and then removes variables one at a time based on a preset significance value to remove.
- Stepwise: This combines both forward and backward procedures. Since inter correlations are complex, the variance due to certain variables will change when new variables are entered into the equation. This is the most frequently used of the regression methods.
- Remove: This is the forced removal option. It requires an initial regression analysis usingthe Enter procedure. In the next block (Block 1 of 1) you may specify one or morevariables to remove. SPSS will then remove the specified variables and run the analysis again.

There are different ways that the relative contribution of each predictor variable can be assessed. In the simultaneous method (which SPSS calls the Enter method), the researcher specifies the set of predictor variables that make up the model. The success of this model in predicting the criterion variable is then assessed.

In contrast, hierarchical methods enter the variables into the model in a specified order. The order specified should reflect some theoretical consideration or previous findings. If you have no reason to believe that one variable is likely to be more important than another you should not use this method. As each variable is entered into the model its contribution is assessed. If adding the variable does not significantly increase the predictive power of the model then the variable is dropped.

In statistical methods, the order in which the predictor variables are entered into (or taken out of) the model is determined according to the strength of their correlation with the criterion variable. Actually there are several versions of this method, called forward selection, backward selection and stepwise selection.

## 10.1 Forward Selection

In Forward selection, SPSS enters the variables into the model one at a time in an order determined by the strength of their correlation with the criterion variable. The effect of adding each is assessed as it is entered, and variables that do not significantly add to the success of the model are excluded.

## 10.2 Forward Regression

We consider first SPSSs forward regression. In this procedure, once a predictor is selected into the model, it cannot be removed. Other predictors may be added at future steps, but predictors already in the model remain in the model. As we will see, this is in contrast to SPSSs stepwise regression, in which we can specify criteria for both adding and removing predictors at each step. We detail now a procedural description of SPSSs forward selection procedure.

**Step 1**
The predictor with the largest squared correlation with Y is entered into the model. Since this is the first step of the selection procedure, entering the predictor with the largest squared correlation is equivalent to entering the predictor with the largest squared semipartial correlation. It may seem trivial to bring up the idea of semipartial correlation at step 1 of the procedure, but we do so because at subsequent steps, the criterion for entrance into the regression equation will be the squared semipartial correlation (or equivalently, the amount of variance contributed by the new predictor over and above variables already entered into the equation).

**Step 2**
The predictor with the largest squared semipartial correlation with Y is selected. That is, the predictor with the largest correlation with Y after being adjusted for the first predictor, is entered if it meets entrance criteria in terms of preset statistical significance for entry, what SPSS refers to as PIN (probability of entry, or in) criteria. Be sure to note that even once this new predictor is entered at step 2, the predictor entered at step 1 remains in the equation, even if its new semipartial correlation with Y is now less than what it was at step 1. This is the nature of the forward selection procedure, it does not re-evaluate already-entered predictors into the model after adding new variables. That is, it only add predictors to the model. Again, this is in contrast to SPSSs stepwise procedure (to be discussed in some detail shortly) in which in addition to entrance criteria being specified for new variables, removal criteria is also specified at each stage of the variable-selection procedure. **Step 3**
The predictor with the largest squared semipartial correlation with Y is selected. That is, the predictor with the largest correlation with Y after being adjusted for both of the first predictors is entered. Be sure to note that the entrance of this variable is conditional upon its relationship with the previously entered variables at step 1 and step 2. Hence, for a variable to be entered at step 3, SPSS asks the question, Which among available variables currently not entered into the regression equation contribute most to variance explained in Y given that variables entered at steps 1 and 2 remain in the model? Translated into statistical language, what this question boils down to is selecting the variable that has the largest statistically significant squared semipartial correlation with Y.

**Steps 4, 5, 6,**
We do not detail subsequent steps for the reason that they mimic the preceding steps. It is worth noting that we didnt even really need to detail steps 2 and 3, and could have just stated the rule of forward regression by referring to the first step alone. We can state the general rule of forward regression as follows: Forward regression, at each step of the selection procedure from step 1 through subsequent steps, chooses the predictor variable with the greatest squared semipartial correlation with the response variable for entry into the regression equation. The given predictor will be entered if it satisfies entrance criteria (significance level, PIN) specified in advance by the researcher.

The above is the simplest way to describe the procedural routine of how forward regression operates. What is perhaps most noteworthy about the above rule is what is not included just as much as what is included in the statement. Notice that nowhere in the rule is there any mention of removal of predictors at any step of the selection process. Forward selection only

## 10.3 Backward Selection

In Backward selection, SPSS enters all the predictor variables into the model. The weakest predictor variable is then removed and the regression re-calculated. If this significantly weakens

the model then the predictor variable is re-entered otherwise it is deleted. This procedure is then repeated until only useful predictor variables remain in the model.

## 10.4 Stepwise Selection

Stepwise is the most sophisticated of these statistical methods. Each variable is entered in sequence and its value assessed. If adding the variable contributes to the model then it is retained, but all other variables in the model are then re-tested to see if they are still contributing to the success of the model. If they no longer contribute significantly they are removed. Thus, this method should ensure that you end up with the smallest possible set of predictor variables included in your model.

In addition to the Enter, Stepwise, Forward and Backward methods, SPSS also offers the Remove method in which variables are removed from the model in a block the use of this method will not be described here.

If you have no theoretical model in mind, and/or you have relatively low numbers of cases, then it is probably safest to use Enter, the simultaneous method. Statistical procedures should be used with caution and only when you have a large number of cases. This is because minor variations in the data due to sampling errors can have a large effect on the order in which variables are entered and therefore the likelihood of them being retained. However, one advantage of the Stepwise method is that it should always result in the most parsimonious model. This could be important if you wanted to know the minimum number of variables you would need to measure to predict the criterion variable. If for this, or some other reason, you decide to select a statistical method, then you should really attempt to validate your results with a second independent set of data. The can be done either by conducting a second study, or by randomly splitting your data set into two halves . Only results that are common to both analyses should be reported.

## 10.5 Stepwise Regression

Stepwise regression combines forward selection and backward elimination. At each step, the best remaining variable is added, provided it passes the significant at 5criterion, then all variables currently in the regression are checked to see if any can be removed, using the greater than 10% significance criterion. The process continues until no more variables are added or removed. This is the one we shall use. It is not guaranteed to find the best subset of independents but it will find a subset close to the best.

## 10.6 How to Identify Multicollinearity

You can assess multicollinearity by examining tolerance and the Variance Inflation Factor (VIF) are two collinearity diagnostic factors that can help you identify multicollinearity. Tolerance is a measure of collinearity reported by most statistical programs such as SPSS; the variables tolerance is 1-R2. A small tolerance value indicates that the variable under consideration is almost a perfect linear combination of the independent variables already in the equation and that it should not be added to the regression equation. All variables involved in the linear relationship will have a small tolerance. Some suggest that a tolerance value less than 0.1 should be investigated further. If a low tolerance value is accompanied by large standard errors and nonsignificance, multicollinearity may be an issue.

You can also assess multicollinearity in regression in the following ways:

1. Examine the correlations and associations (nominal variables) between independent variables to detect a high level of association. High bivariate correlations are easy to spot by running correlations among your variables. If high bivariate correlations are present, you can delete one of the two variables. However, this may not always be sufficient.

2. Regression coefficients will change dramatically according to whether other variables are included or excluded from the model. Play around with this by adding and then removing variables from your regression model.

3. The standard errors of the regression coefficients will be large if multicollinearity is an issue.

4. Predictor variables with known, strong relationships to the outcome variable will not achieve statistical significance. In this case, neither may contribute significantly to the model after the other one is included. But together they contribute a lot. If you remove both variables from the model, the fit would be much worse. So the overall model fits the data well, but neither X variable makes a significant contribution when it is added to your model last. When this happens, multicollinearity may be present.

# 11 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.

## 11.1 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.