# Practice Workbook

**Emilie Krutnes Engen**

Predictive Modelling
Carlos III University of Madrid

Universidad
Carlos III de Madrid

# Contents

# List of Tables



# List of Figures

# Part I

# Simple Linear Regression

## 1 Exercise 1

In this exercise we want to investigate the relationship between the birth rate for females between the age of 15 and 17 and the poverty rate. The data is collected from the district of Colombia in 2002. The birth rate is measured per 1000 females and the poverty rate is the percentage of the states population living in households with income below the federally defined poverty level. Our initial assumption is a linear relationship between the two variables and we therefore want to study the sustainability of a simple linear regression model for explaining the teenage birth rate, $Y$, as a function of the poverty rate, $X$.

Before fitting the model we want to investigate how the variables are related to each other. This can be done graphically by constructing a scatter plot. From the scatter plot in Figure 1 we have that the teenage birth rate can be expressed as a function of the poverty rate, using a simple linear regression. A simple linear regression model can be described by the equation

$$Y = \beta_0 + \beta_1 X + \varepsilon \tag{1}$$

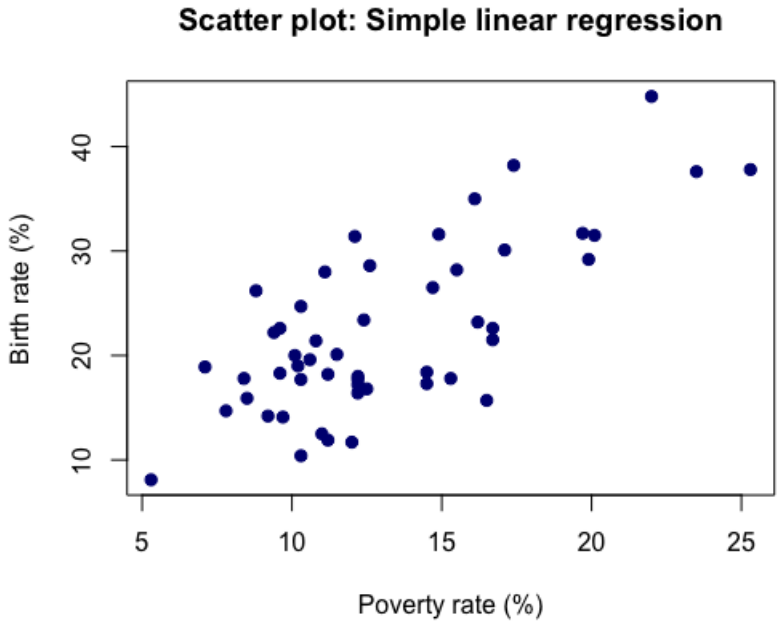


Figure 1: Scatter plot of birth rate for given poverty rates

In this context, the dependent variable $Y$ represents the teenage birth rate and the indepen-

dent variable $X$ represents the poverty rate. The variable $\varepsilon$ is the error term and represents other factors affecting $Y$. Further $\beta_0$ is the intercept parameter, also referred to as the constant term and $\beta_1$ is the slope parameter, describing the relationship between the dependent and independent variables, holding other factors fixed. Thus, $\beta_1$ describes the change in the teenage birth rate given by a change in the poverty rate, with $\varepsilon$ fixed Thomas (2005).

Notice, that the true regression is unknown, as it is impossible to derive the exact relationship between the variables using a regression line. Hence, the regression line is referred to as the best-fitting line. By using a sample of $N$ observations the expected value of $y_i$ for observation $i$

$$\hat{y_i} = E(y_i) = \hat{\beta_0} + \hat{\beta_1} x_i, \quad i \in N \tag{2}$$

is a linear function of $x_i$ (Thomas, 2005). These are called the fitted values, where $\hat{y_i}$ is the value on the least squared The model applies to the mean for all observations and thus holds an average. Having $\varepsilon$ uncorrelated with $X$ is desired. Hence, the expected value of $\varepsilon$ is assumed zero (Thomas, 2005).

$$E(\varepsilon_i) = 0 \quad i \in N \tag{3}$$

As $\hat{y_i}$ represent the estimated value of $y_i$, the residual $e_i$, representing the error of the estimated value should be sufficiently small.

The residual is defined as the difference between the $y_i$ and the estimated value $\hat{y_i}$ (Thomas, 2005).

$$e_i = y_i - \hat{y_i} \quad i \in N \tag{4}$$

Notice the difference between the estimated error $\varepsilon_i$ and the residual $e_i$. The residual explain the error within the model and the error term explains the variance outside the model (Thomas, 2005). Because the true regression is unknown, an estimated regression is needed. The estimated regression is obtained by predicting the dependent variable $\hat{Y_t}$ by using observations of $X_t$. This includes predicting the values of the parameters $\beta_0$ and $\beta_1$, using the least squares method. This method aims to estimate the intercept and slope parameter such that the regression line have the smallest possible sum of squares, the vertical distance from each point to the line (Becker and Wilks, 1988). The least squares regression line is given by

$$\hat{Y} = \hat{\beta_0} + \hat{\beta_1} X \tag{5}$$

where $\hat{\beta_0}$ and $\hat{\beta_1}$ are the estimated regression coefficients.

## (a) Regression Equation

Using a simple linear regression, we have that the teenage birthrate $Y$ can be explained by the poverty rate $X$ given by the following regression equation.

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X \tag{6}$$

Inserting for the estimated coefficients we get:

$$\hat{Y} = 4.2673 + 1.3733X \tag{7}$$

The R-squared for this regression is 0.5333, which suggest that the regression is not perfectly accurate in describing the variation in the data set. The result for this linear regression is presented in Table 1.

| Simple linear regression | | | | |
|---|---|---|---|---|
| Coefficient | Estimate | Std. Error | t-Test | p-value |
| $\beta_0$ | 4.2673 | 2.5297 | 1.69 | 0.0980 |
| $\beta_1$ | 1.3733 | 0.1835 | 7.48 | $0.0000^{***}$ |
| Significance codes: $0^{***}$, $0.001^{**}$, $0.05^{*}$ | | | | |

Table 1: Results from fitting the simple linear regression $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$

From the results it is evident that the intercept parameter $\hat{\beta}_0$ is not significant with a significance level of 5 %, as the p-value is 0.0980. However the slope parameter is significant, which suggest that the poverty rate is significant in describing changes in the birth rate. From the scatter plot in Figure 1 the linear relationship seems evident. The fitted linear regression model is presented in Figure 2. From this plot we can see that the residuals seem to be constant for different values of the poverty rate, $X$.


Figure 2: Simple linear regression model of birth rate for given poverty rates

The residual plots in Figure 3 support the assumption of a linear relationship between the two variables. In these plots the studentized residuals for the poverty rate, $X$ and the fitted values are plotted. The errors $\varepsilon_i$ are assumed independent $N(0, \sigma)$ random variables. The $\varepsilon_i$ are not observed, however the residuals $\epsilon_i = Y_i - \hat{Y}_i$ are the error in the fit of the regression line and thus serves to mimic $\varepsilon$ (Walpole et al., 2014). Ideally the residuals should show a truly random fluctuation around a value of zero. From the plot in Figure 3 no specific pattern occur in the residuals and we therefore conclude that a linear regression model is sustainable in describing the teenage birth rate in terms of the poverty rate.

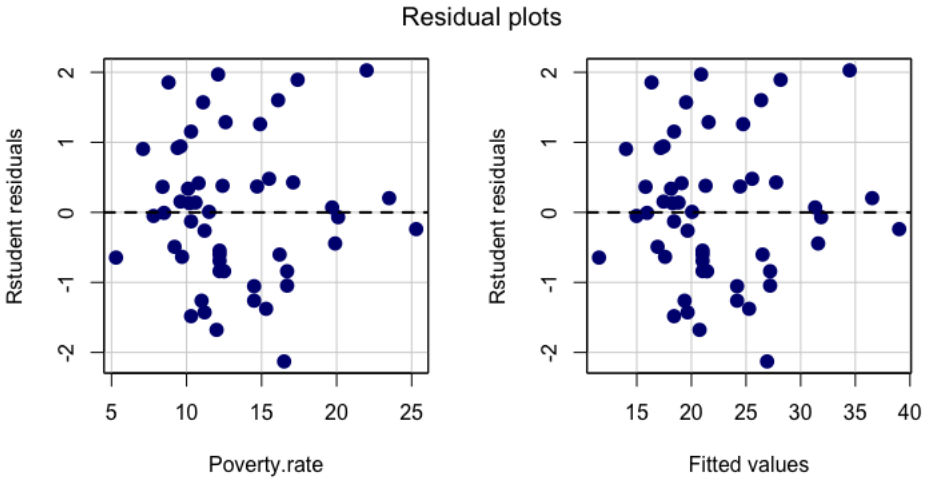


Figure 3: Studentized residual plots for poverty rate and fitted values

### (b) Confidence Interval for the Slope Parameter

A confidence interval is defined as a range of values where there is a specified probability that the value of the parameter lies within the given interval. A 95 % confidence interval for the slope parameter is thus the interval constructed such that there is a 95 % probability that the true value of $\beta_1$ lies within the interval (Walpole et al., 2014). In order to construct a 95 % confidence interval for the slope parameter $\beta_1$ we assume that the standard requirements for simple linear regression are met. These include having a linear relationship between the dependent variable $Y$ and the independent variable $X$. For any given value of $X$, the $Y$ values have to be independent and roughly normally distributed. We investigate the confidence interval of $\beta_1$, using the regression from Equation 6. The 95 % confidence interval for $\beta_1$ is given by [1.0045, 1.7421] with an estimated value of 1.3733.

### (c) Significance Level

In order to determine whether the poverty rate $X$ is significant in determining the teenage birth rate $Y$ we look at results provided in Table 1. With a significance level of 5 % the poverty rate is significant in determining the teenage birth rate, with a p-value close to zero, 0.0000 with four significant digits.

### (d) Estimation of the Slope Parameter

From Table 1 we have that $\beta_1$ is estimated to 1.3733. Hence the effect of an increase of the poverty rate leads to an estimated increase of 37.33 % in the teenage birth rate.

### (e) Prediction Interval with Specific Regressor Values

In this case we want to predict the birth rate given a poverty rate of 15 %. When predicting a value of a specific future observation, constructing a prediction interval is useful. Where the confidence interval only account for the variation due to estimating the mean, the prediction interval also account for variation in the future observation (Walpole et al., 2014). Hence the estimated birth rate for a poverty rate of 15 % is predicted by obtaining a prediction interval for $X = 15$. From this prediction we get $\hat{Y}_t = 24.8675$. With a poverty rate of 15 %, the estimated birth rate is approximately 25 %. From the fitted model in Figure 2 this prediction looks reasonable.

## 2 Exercise 2

In this exercise we want to investigate the relationship between the taxes paid when buying a house and the price of the house in \$1000. In this exercise we do not expect the relationship between the two variables to be linear as it would make the taxes very high for increasing prices. We therefore want to study whether a transformation of the simple linear regression model is needed for explaining the taxes, $Y$, as a function of the prices, $X$.

Before fitting the model we want to investigate how the variables are related to each other. This can be done graphically by constructing a scatter plot. From the scatter plot in Figure 4 the regression seems linear for lower price values. However the variation increases for higher prices. This might suggest that some log transformation is needed.

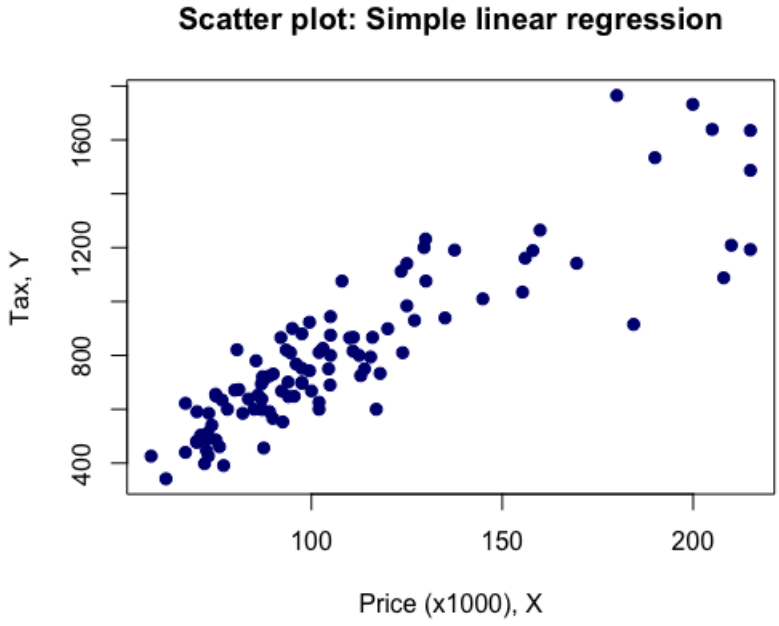


Figure 4: Scatter plot of tax for given house prices

## (a) Regression Equation

We first express the taxes as a function of the house prices, using a simple linear regression. The tax $Y$ can be explained by the price $X$ given by the following regression equation.

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X \tag{8}$$

Inserting for the estimated coefficients we get:

$$\hat{Y} = 61.2716 + 6.8763X \tag{9}$$

The R-squared for this regression is 0.7773, which suggest that the regression is quit good in describing the variation in the data set. The result for this linear regression is presented in Table 2.

| Simple linear regression | | | | |
|---|---|---|---|---|
| Coefficient | Estimate | Std. Error | t-Test | p-value |
| $\beta_0$ | 61.2716 | 41.9781 | 1.46 | 0.1475 |
| $\beta_1$ | 6.8763 | 0.3645 | 18.87 | $0.0000^{***}$ |
| Significance codes: $0^{***}$, $0.001^{**}$, $0.05^{*}$ | | | | |

Table 2: Results from fitting the simple linear regression $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$

From the results it is evident that the intercept parameter $\beta_0$ is not significant with a significance level of 5 %, as the p-value is 0.1475. The fitted linear regression model is presented in Figure 5. From this plot we can see that the residuals are not constant for higher price values, $X$.

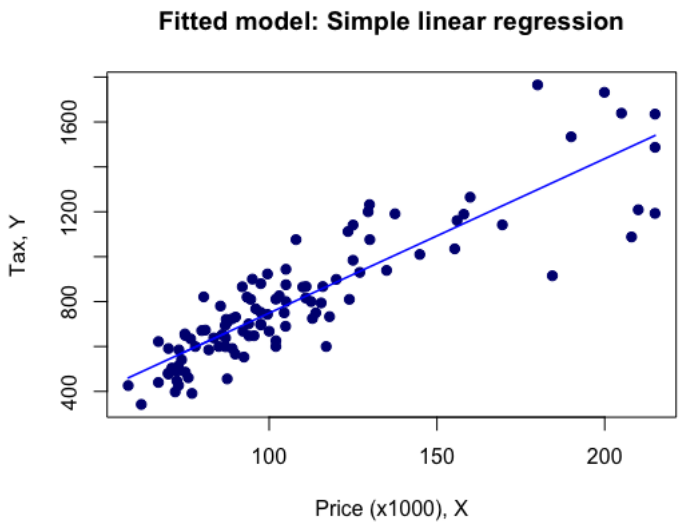


Figure 5: Simple linear regression model of taxes for given house prices

The residual plots in Figure 6 support the assumption of increasing residuals. In these plots the studentized residuals for the prices, $X$ and the fitted values are plotted. Having increasing residual terms might suggest that multiplicative error model is appropriate.

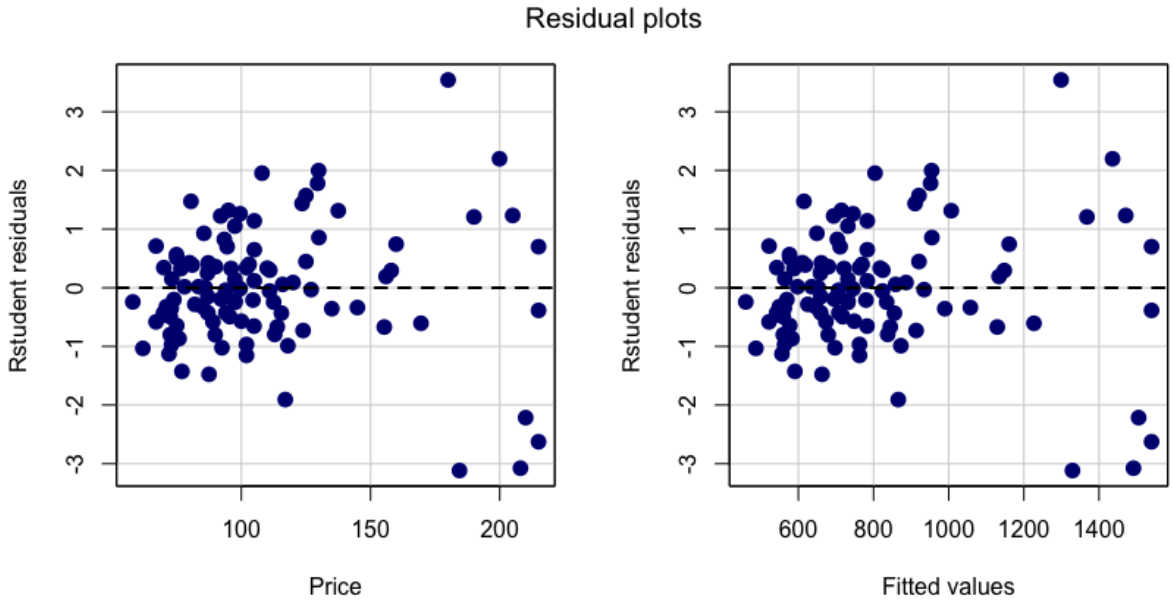


Figure 6: Studentized residual plots for poverty rate and fitted values

Because of the increasing variation in taxes for higher house prices, we want to test the model, using a log transformation in the regressor. The linear-log regression is expressed by the following equation.

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 log(X) \tag{10}$$

Inserting for the estimated coefficients we get:

$$\hat{Y} = -3132.9132 + 850.1198 log(X) \tag{11}$$

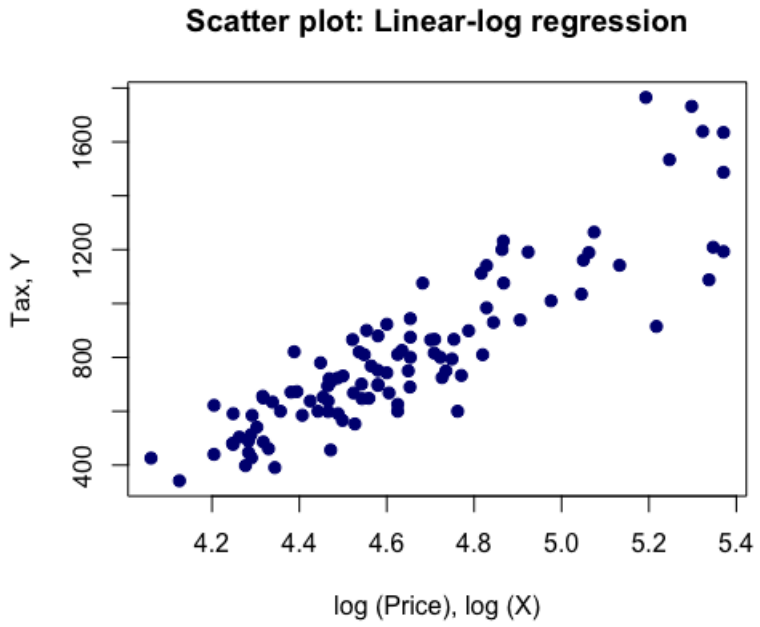


Figure 7: Scatter plot of tax for given log prices

The R-squared for this regression is 0.7874, which suggest a slightly improvement from the simple linear regression model.The result for this linear regression is presented in Table 3.

| Simple linear-log regression | | | | |
|---|---|---|---|---|
| Coefficient | Estimate | Std. Error | t-Test | p-value |
| $\beta_0$ | -3132.9132 | 203.2401 | -15.41 | 0.0000*** |
| $\beta_1$ | 850.1198 | 43.7390 | 19.44 | 0.0000*** |
| Significance codes: 0***, 0.001**, 0.05* | | | | |

Table 3: Results from fitting the linear-log regression $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 log(X)$

From the results it is evident that the both parameters are significant with a significance level of 5 %. However, when looking at the scatter plot, presented in Figure 7 the residuals are still increasing with increasing prices. This is also evident in the residual plots in Figure 8.

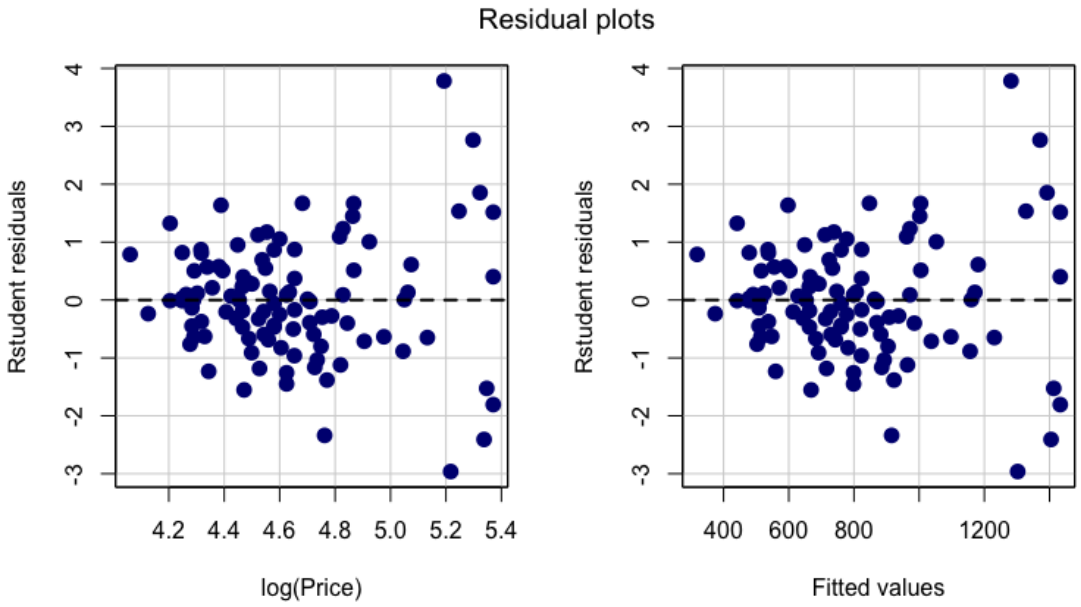


Figure 8: Studentized residual plots for log poverty rate and fitted values

This suggest that a multiplicative error model may provide a better fit. An alternative for such a model is the power model, expressed by the following equation.

$$\hat{Y} = \beta_0 X^{\beta_1} \varepsilon \tag{12}$$

We can linearize this model by taking the logarithm, and thus we obtain:

$$log(\hat{Y}) = \hat{\beta}_0 + \hat{\beta}_1 log(X) + log(\varepsilon) \tag{13}$$

where $\hat{\beta}_0 = log(\beta_0)$ and $\hat{\beta}_1 = log(\beta_1)$. We have here included the error terms to illustrate how the linearization of this regression provide an additive error term of the form $log(\varepsilon)$ contrary to a multiplicative error. Inserting for the estimated coefficients we get:

$$log(\hat{Y}) = 2.0764 + 0.9830 log(X) \tag{14}$$

As you can see from the regression equation the log-log model is linear in the parameters $\beta_0$ and $\beta_1$ and can thus be treated as a linear model. The results from this regression is presented in Table 4.

| Simple log-log regression | | | | |
|---|---|---|---|---|
| Coefficient | Estimate | Std. Error | t-Test | p-value |
| $\hat{\beta}_0$ | 2.0764 | 0.2370 | 8.76 | 0.0000*** |
| $\hat{\beta}_1$ | 0.9830 | 0.0510 | 19.28 | 0.0000*** |
| Significance codes: 0***, 0.001**, 0.05* | | | | |

Table 4: Results from fitting the log-log regression $log(\hat{Y}) = \hat{\beta}_0 + \hat{\beta}_1 log(X)$

The R-squared value for this regression is 0.7846, which implies that 78.46 % of the variance in the data set is described by the multiplicative regression. This is less than in the linear-log regression, but the fitted log-log model is presented in Figure 9 suggest that this model provide an better fit. This plot illustrate the improved fit obtained by the log transformation in both the response variable and the regressor. The variation seems approximately constant for different prices.

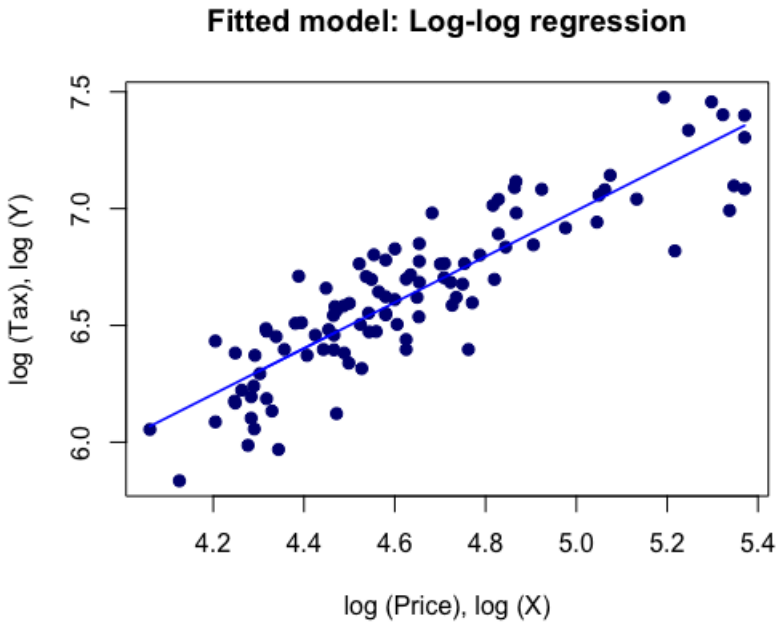


Figure 9: Simple linear regression model of log taxes for given log prices

The residual plots presented in Figure 10 support the assumption of a log-log relationship between taxes and prices. There is no clear pattern in the residuals, which suggest that this model is more appropriate.

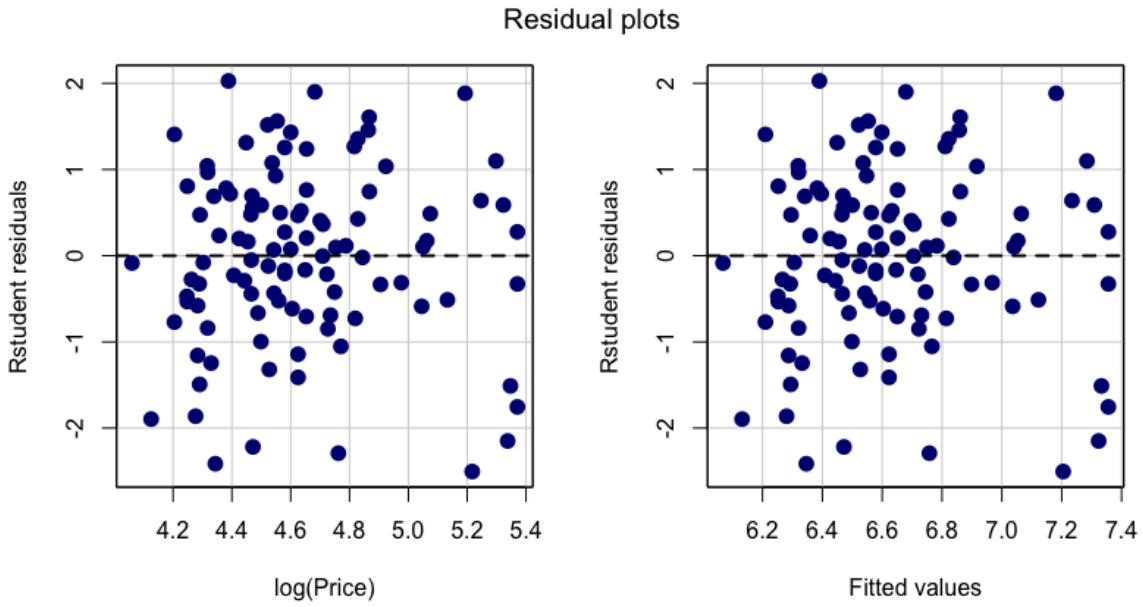


Figure 10: Studentized residual plots for log price and fitted values

### (b) Confidence Interval for the Slope Parameter

We investigate the confidence interval of $\beta_1$, using the following regression.

$$log(\hat{Y}) = \hat{\beta}_0 + \hat{\beta}_1 log(X) \tag{15}$$

The 95 % confidence interval for $\hat{\beta}_1$ is given by [0.8819, 1.0842] with a estimated value of 0.9830. This means that there is a 95 % probability that the an increase in the price lead to an increase between 0.88 % and 1.08 %.

### (c) Significance Level

In order to determine whether the price $X$ is significant in determining the amount of taxes paid $Y$ we look at results provided in Table 3. With a significance level of 5 % the price is significant in determining the taxes, with a p-value close to zero, $< 2e - 16$.

### (d) Estimation of the Slope Parameter

From Table 3 we have that $\beta_1$ is estimated to 0.9830. A more precise result is 0.983028. The slope parameter can be interpreted as the price elasticity of taxes, meaning the responsiveness in taxes due to a change in prices. The elasticity is approximately equal to 1, which suggest unitary elastic taxes.

### (e) Prediction Interval with Specific Regressor Values

The estimated taxes for a price of 100(×\$1000) is predicted by constructing a prediction interval for $X = 100$. The regression model is given by the following equation.

$$log(\hat{Y}_t) = \hat{\beta}_0 + \hat{\beta}_1 log(X_t), \quad t \in N \tag{16}$$

We therefore have to perform back-transformation to predict the value of $\hat{Y}$.

$$e^{log(\hat{Y})} = e^{\hat{\beta}_0} + e^{\hat{\beta}_1 log(X)} \tag{17}$$

$$e^{log(\hat{Y})} = e^{log(\beta_0)} + e^{log(\beta_1) log(X)} \tag{18}$$

$$\hat{Y} = \beta_0 X^{\beta_1} \tag{19}$$

From this prediction interval, with a price level equal to 100 (×\$1000), the predicted tax is given by $\hat{Y} = 737.6208$. The taxes for a house with a price of \$100.000 is approximately \$738. From the fitted model in Figure 5 this prediction seems reasonable.

## 3 Exercise 3

In this exercise we want to investigate the relationship between the number of advertisement pages in hundreds and the total revenue in million dollars. We want to study whether a simple linear regression model is suitable for explaining the revenue, $Y$, as a function of the number of pages, $X$.

### (a) Regression Equation

Before fitting the model we want to investigate how the variables are related to each other. This can be done graphically by constructing a scatter plot. From the scatter plot in Figure 11 the regression seems linear. However the plot suggest the occurrence of one or several outliers. An outlier is a point for which $y_i$ is far from the value predicted by the regression model. Outliers can occur for numerous reasons, such as incorrect recording during the data collection process (James et al., 2014). From looking at the scatter plot, the point to the right with almost 80,000 advertisement pages appear as an obvious outlier. When investigating the plot more closely we can see that there are two point with zero advertisement pages with a revenue of almost \$ 10 million. Intuitively this seems incorrect.

The scatter plot in Figure 11 suggest that the revenue can be expressed as a function of the number of advertisement pages, using simple linear regression.

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X \tag{20}$$

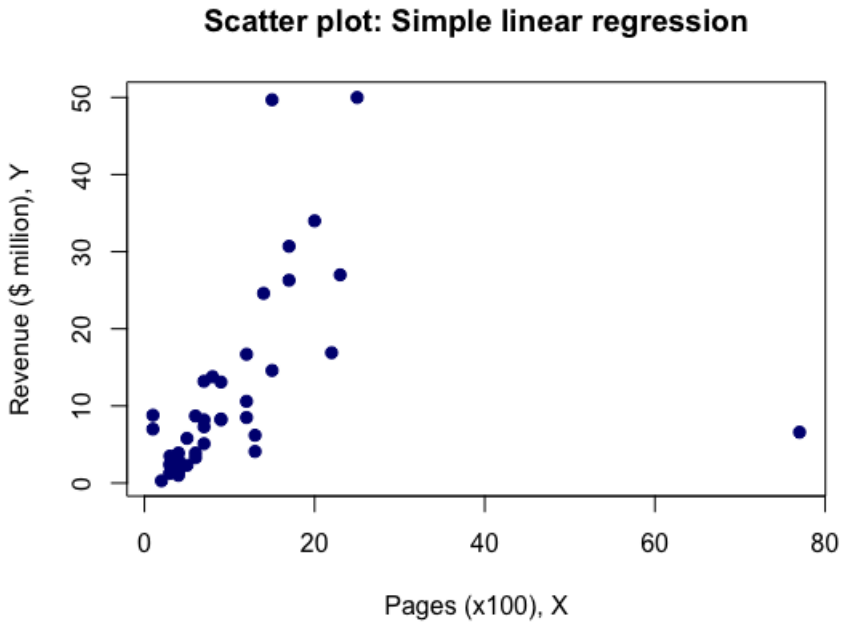


Figure 11: Scatter plot of revenue for given numbers of advertisement pages

Inserting for the estimated coefficients we get:

$$\hat{Y} = 7.6041 + 0.3527X \tag{21}$$

The R-squared for this regression is 0.1263, which suggest that the regression is not accurate in describing the variation in the data set. The result for this linear regression is presented in Table 5.

| Simple linear regression | | | | |
|---|---|---|---|---|
| Coefficient | Estimate | Std. Error | t-Test | p-value |
| $\beta_0$ | 7.6041 | 2.4061 | 3.16 | 0.0030* |
| $\beta_1$ | 0.3527 | 0.1486 | 2.37 | 0.0226* |
| Significance codes: 0***, 0.001**, 0.05* | | | | |

Table 5: Results from fitting the simple linear regression $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$

From the results it is evident that both the intercept parameter $\beta_0$ and the slope parameter $\beta_1$ are significant with a significance level of 5 %. The residual plots in Figure 12 further support the assumption of outliers.

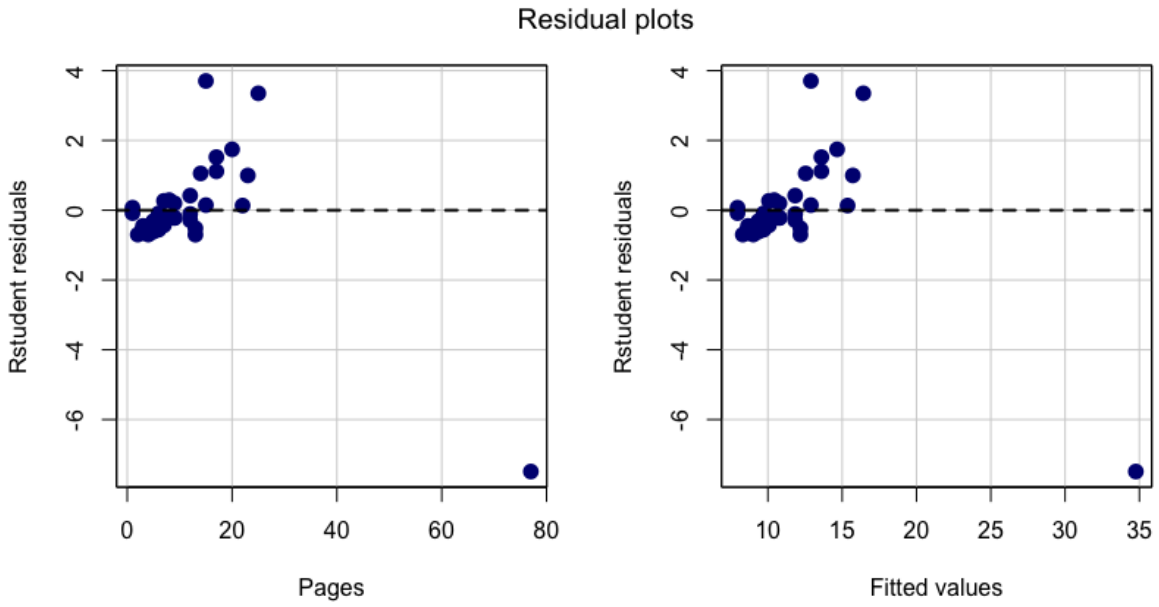


Figure 12: Studentized residual plots for number of pages and fitted values

We test the regression using a log transformation in the response variable, revenue. The log-linear model can be expressed by the following equation:

$$log(\hat{Y}) = \hat{\beta}_0 + \hat{\beta}_1 X \tag{22}$$

From the scatter plot in Figure 13 the log-linear regression does not seem appropriate. However this plot also suggest the occurrence of one or several outliers.

The results from this regression is presented in Table 6. The R-squared is 0.1574, which suggest an slight improvement from the linear regression.

| Simple log-linear regression | | | | |
|---|---|---|---|---|
| Coefficient | Estimate | Std. Error | t-Test | p-value |
| $\beta_0$ | 1.5007 | 0.2181 | 6.88 | $0.0000^{***}$ |
| $\beta_1$ | 0.0363 | 0.0135 | 2.70 | $0.0102^{*}$ |
| Significance codes: $0^{***}$, $0.001^{**}$, $0.05^{*}$ | | | | |

Table 6: Results from fitting the log-linear regression $log(\hat{Y}) = \hat{\beta}_0 + \hat{\beta}_1 X$

From the results it is evident that both the intercept parameter $\hat{\beta}_0$ and the slope parameter $\hat{\beta}_1$ are significant at a 5 % level. The residual plots in Figure 14 also support the assumption of outliers.

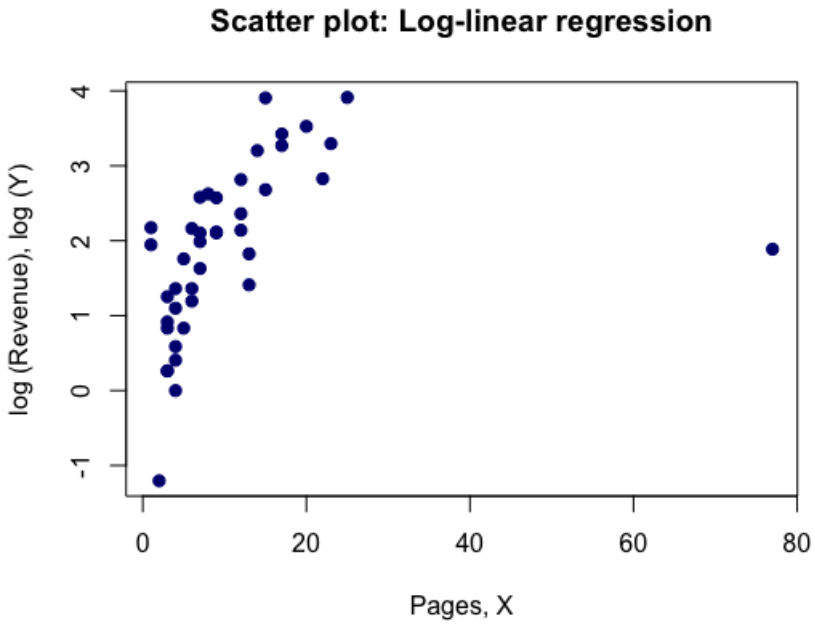


Figure 13: Scatter plot of log revenue for given numbers of advertisement pages

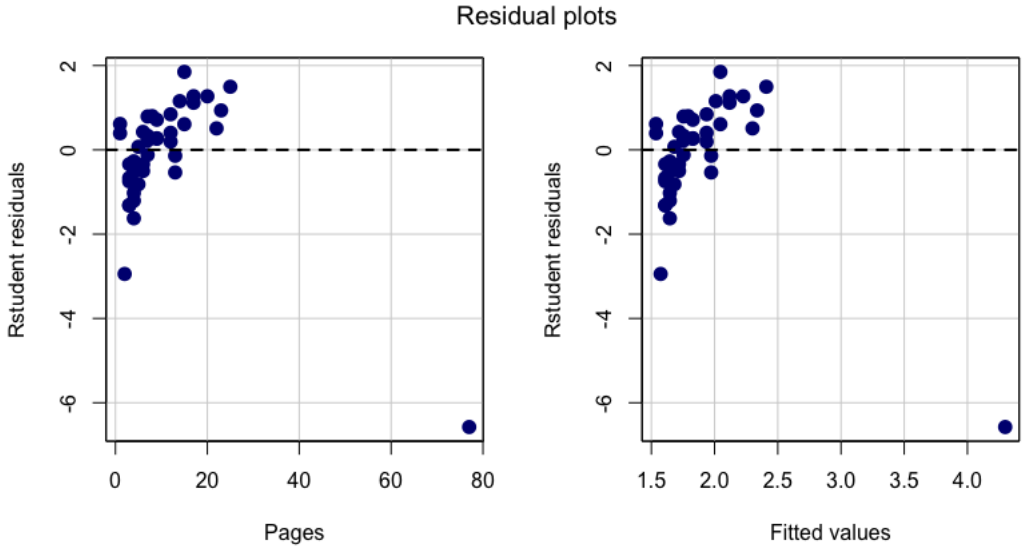


Figure 14: Studentized residual plots for number of pages and fitted values

Therefore we test the the linear regression, using a log transformation of both the dependent variable $Y$ and the independent variable $X$. The log-log regression model can be expressed

by the following equation:

$$log(\hat{Y}) = \hat{\beta}_0 + \hat{\beta}_1 log(X) \tag{23}$$

The scatter plot presented in Figure 15 suggest that the log-log regression model seems appropriate, as log pages and log revenue show a linear relationship. Here the three outliers described earlier occur quit clearly.

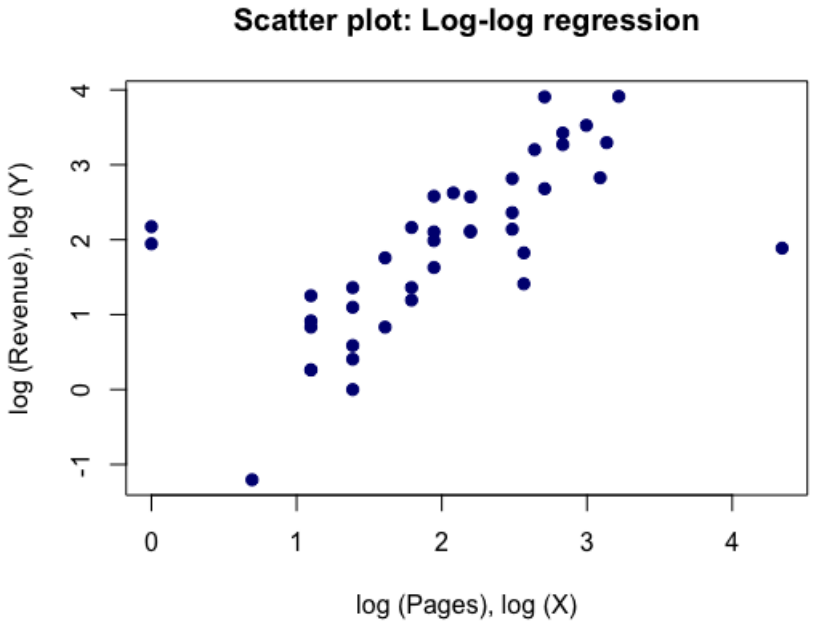


Figure 15: Scatter plot of log revenue given for log numbers of advertisement pages

The results from this regression is presented in Table 7. The R-squared is 0.4203, which suggest an improvement from the previous regression models.

| Simple log-log regression | | | | |
|---|---|---|---|---|
| Coefficient | Estimate | Std. Error | t-Test | p-value |
| $\beta_0$ | 0.2323 | 0.3398 | 0.68 | 0.4983 |
| $\beta_1$ | 0.8354 | 0.1571 | 5.32 | 0.0000*** |
| Significance codes: 0***, 0.001**, 0.05* | | | | |

Table 7: Results from fitting the log-log regression $log(\hat{Y}) = \hat{\beta}_0 + \hat{\beta}_1 log(X)$

From the results it is evident that the intercept parameter $\hat{\beta}_0$ is not significant with a significance level of 5 %, as the p-value is 0.4983. However the slope parameter $\hat{\beta}_1$ is significant with a 5 % level. From the residual plots in Figure 16 the outliers are also evident.

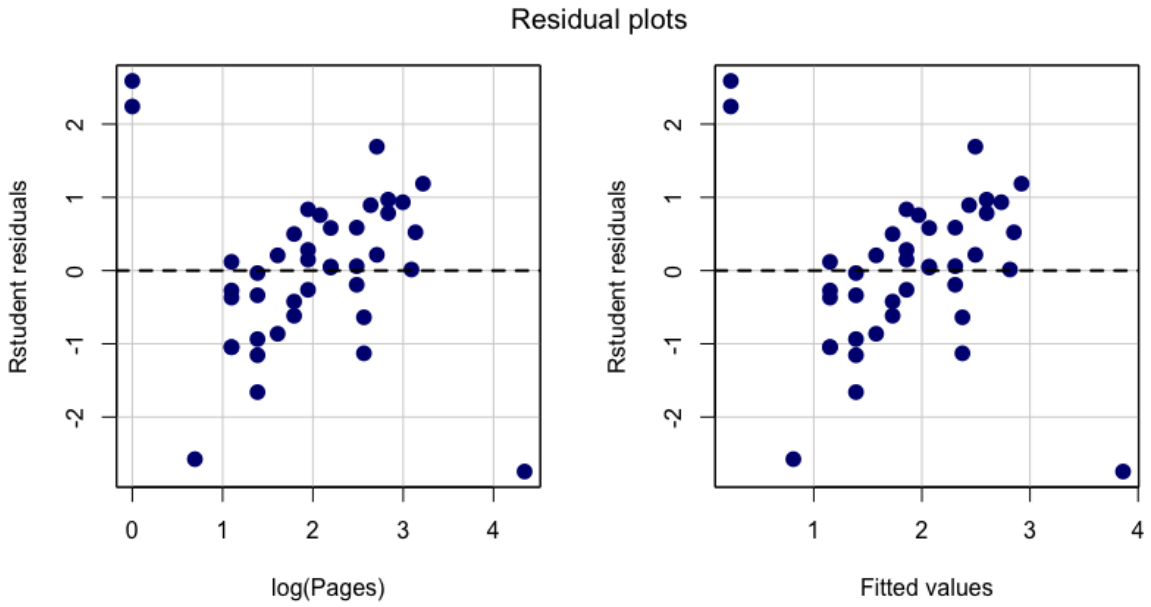


Figure 16: Studentized residual plots for log number of pages and fitted values

We now investigate the fit of the log-log regression when the identified outliers are excluded. The results from this presented in Table 8. The R-squared is 0.7910, which is an improvement from the previous regression models. This indicate that the log-log regression model is able to explain 79.10 % of the variation in the data.

| Simple log-log regression excluding outliers | | | | |
|---|---|---|---|---|
| Coefficient | Estimate | Std. Error | t-Test | p-value |
| $\beta_0$ | -1.2134 | 0.2793 | -4.34 | $0.0001^{***}$ |
| $\beta_1$ | 1.5278 | 0.1309 | 11.67 | $0.0000^{***}$ |
| Significance codes: $0^{***}$, $0.001^{**}$, $0.05^{*}$ | | | | |

Table 8: Results from fitting the log-log regression $log(\hat{Y}) = \hat{\beta}_0 + \hat{\beta}_1 log(X)$, excluding outliers

From the results it is evident that both the intercept parameter $\hat{\beta}_0$ and the slope parameter $\hat{\beta}_1$ are significant with a significance level of 5 %. The residual plots in Figure 17 excluding the outliers, support the assumption of a log-log regression, as we can not identify any pattern in the residuals.

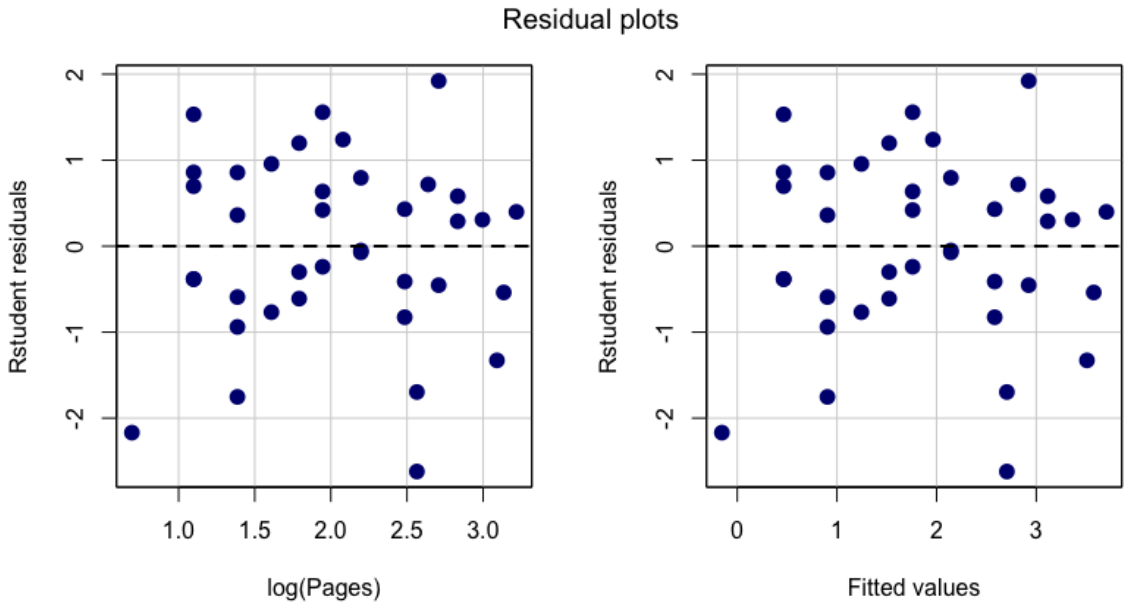


Figure 17: Studentized residual plots for log number of pages and fitted values excluding outliers

In Figure 18 we have obtained the fitted log-log model excluding the outliers. As we can see the model seems appropriate in fitting the data, when excluding the outliers.

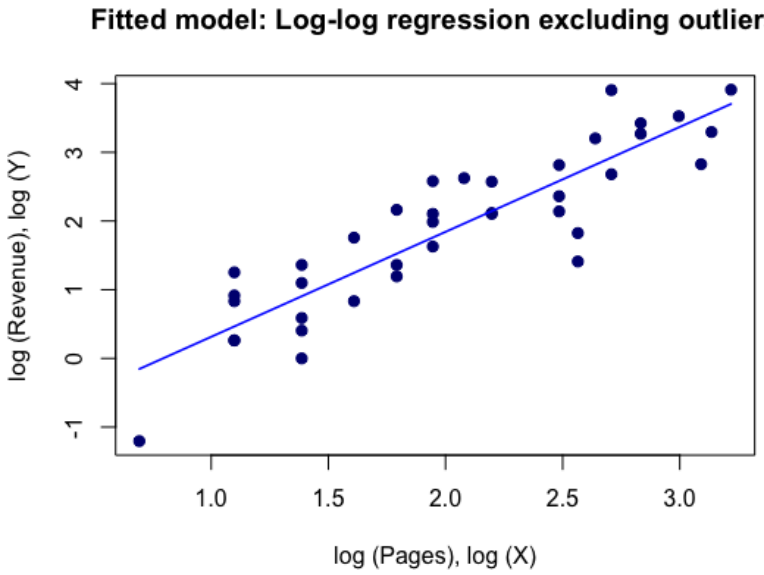


Figure 18: Simple linear regression model of log revenue for given log number of pages, excluding outliers

### (b) Confidence Interval for the Slope Parameter

We investigate the confidence interval of $\beta_1$, using the following regression:

$$log(\hat{Y}) = \hat{\beta}_0 + \hat{\beta}_1 log(X) \tag{24}$$

The 95 % confidence interval for $\beta_1$ is given by [1.2623, 1.7933] with a estimated value of 1.5278.

### (c) Significance Level

In order to determine whether the number of pages $X$ is significant in determining the revenue $Y$ we look at results provided in Table 8. With a significance level of 5 % the number of pages is significant in determining the revenue, with a p-value close to zero, 0.0000 with four significant digits.

### (d) Estimation of the Slope Parameter

From Table 7 we have that $\beta_1$ is estimated to 1.5278. A more precise result is 1.527807. The effect of an increase of the number of pages leads to an increase in revenue of about 53 %.

## Part II

# Multiple Linear Regression

A multiple linear regression is just an extension of the simple regression we saw in Chapter 1. Instead of trying to explain a response variable $Y$ with only one regressor $X$, we now investigate if a response variable can be expressed by several regressors. A regression with more than one explanatory variable is referred to as a multiple regression. A multiple linear regression model can expressed by the following equation:

$$\hat{Y} = \hat{\beta}_0 + \sum_{i=1}^{I} \hat{\beta}_i X_i \tag{25}$$

for all explanatory variables $i \in I$. In this chapter we will study different multiple linear regression model and investigate how to determine which variables to include in the regression model by looking at the significance of each of the regressors and the overall performance of the model.

## 4 Exercise 1

In this exercise we are analysing the costs of a given firm. We want to investigate how costs related to the production at different factories affect the unit cost of the production. The explanatory cost variables are depreciation of machinery and equipment, cost of raw material, energy cost and salary per hour. The response variable is the unit cost of production. Our initial assumption is that all regressors are positive related to the response variable. We therefore want to investigate whether a multiple linear regression is suitable for describing the given data set.

### (a) Scatter Plot

Before fitting the regression model we want to investigate how the variables are related. This can be done by constructing scatter plots of all pairwise combinations of the variables in the data set. Figure 19 show the scatter plot matrix relating all the variables in the data set.

From the scatter plot matrix we can see that there seems to be a positive relation between the response variable and the regressor variables. However, the plot indicate that a simple multiple regression might not be appropriate in describing the data, and suggest that some log transformation is needed. This sounds reasonable as the company, assuming a large production, should be able to take advantage in economies of scale in production.

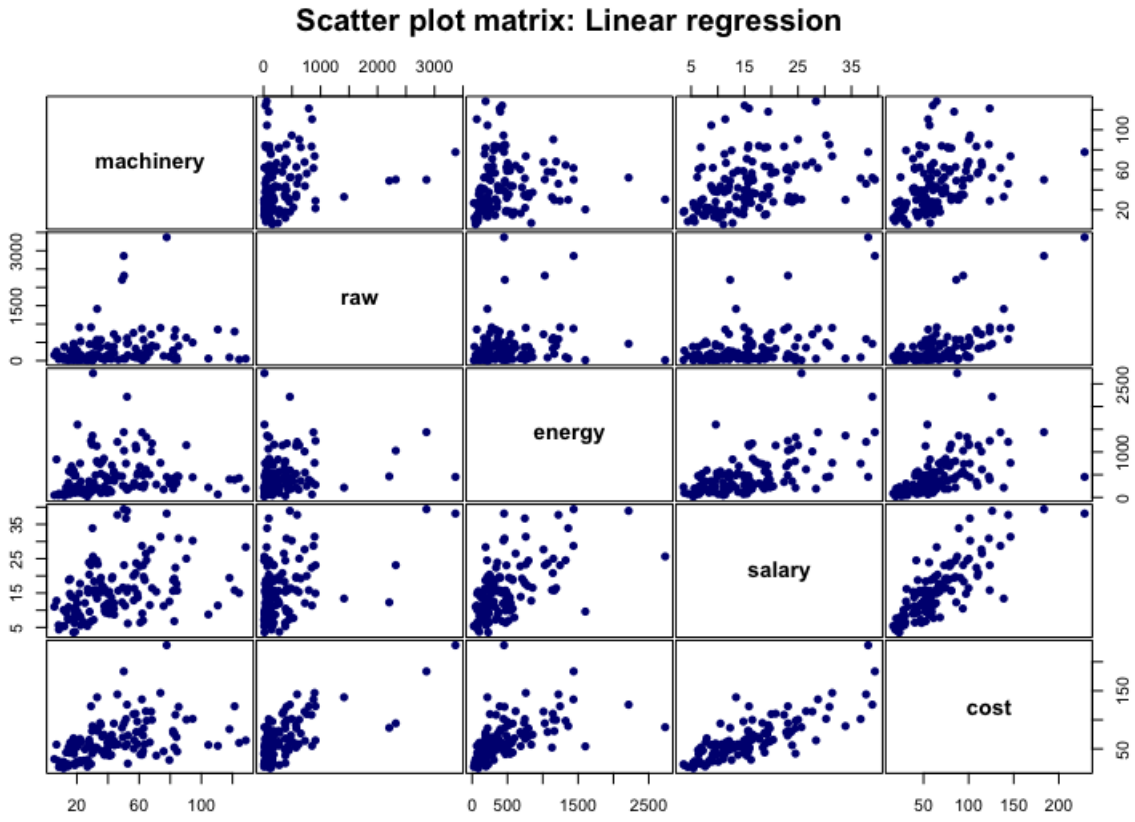


Figure 19: Scatter plot matrix showing the pairwise relation in the cost data set

### (b) Full Regression Model

By including all the explanatory cost variables we get

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3 + \hat{\beta}_4 X_4 \tag{26}$$

where $X_1$ is the machinery cost, $X_2$ is the material cost, $X_3$ is the energy cost and $X_4$ is the salary per hour. By inserting for the estimated parameters we get:

$$\hat{Y} = 9.1470 + 0.1906 X_1 + 0.0294 X_2 + 0.0139 X_3 + 2.0056 X_4 \tag{27}$$

The result for this regression is presented in Table 9. The results indicate that all variables are significant with a significance level of 5 %. The adjusted R-squared value for this regression is 0.8196, which indicate that this regression provides a good fit for the given data. Notice that we are considering the adjusted R-squared value instead of the R-squared value. While the R-squared value provide information about how well the model is in determining the variation in the model, the adjusted R-squared value is considered a better measure when comparing

multiple regression model with different including different numbers of explanatory variables. The adjusted R-square is a variation of R-squared that provides an adjustment for degrees of freedom. This implies that the adjusted R-squared model is designed to punish overfitted models, where R-squared always will favor the inclusion of additional regressors.

| The linear regression model | | | | |
|---|---|---|---|---|
| Coefficient | Estimate | Std. Error | t-Test | p-value |
| $\beta_0$ | 9.1470 | 3.1943 | 2.86 | $0.0049^{*}$ |
| $\beta_1$ | 0.1906 | 0.0528 | 3.61 | $0.0004^{**}$ |
| $\beta_2$ | 0.0294 | 0.0028 | 10.41 | $0.0000^{***}$ |
| $\beta_3$ | 0.0139 | 0.0037 | 3.73 | $0.0003^{**}$ |
| $\beta_4$ | 2.0056 | 0.2300 | 8.72 | $0.0000^{***}$ |
| Significance codes: $0^{***}$, $0.001^{**}$, $0.05^{*}$ | | | | |

Table 9: Results from fitting the multiple linear regression model

## (c) Residual Plots

We can further investigate if any transformation is needed in the regression model by constructing a diagnostic plot, plotting the studentized residuals for all explanatory variables. We take a look at the residual plots from the regression. From the residual plots in Figure 20 we can see that the linear regression fails to fit the data accurately for large values of the the material cost, $X_2$ an the energy cost, $X_3$ or there seems to be some outliers associated with these variables.

Because the variance seems to be non-constant, we want to try to fit the log-log regression model. As we have seen previously the multiplicative error model can be linearized by transforming it to the log-log regression, which is linear in the parameters. The scatter plot matrix for the log-log regression is presented in Figure 21. The plot suggest that this regression provide a better fit for the given data set.

The log-log regression model for a multivariate linear regression can be obtained by the following equation.

$$log(\hat{Y}) = \hat{\beta}_0 + \hat{\beta}_1 log(X_1) + \hat{\beta}_2 log(X_2) + \hat{\beta}_3 log(X_3) + \hat{\beta}_4 log(X_4) \tag{28}$$

Inserting for the parameters we obtain:

$$log(\hat{Y}) = 0.8539 + 0.1219 log(X_1) + 0.1244 log(X_2) + 0.1611 log(X_3) + 0.4570 log(X_4) \tag{29}$$

The result for this regression is presented in Table 10. The results indicate that all variables are significant with a significance level of 5 %.

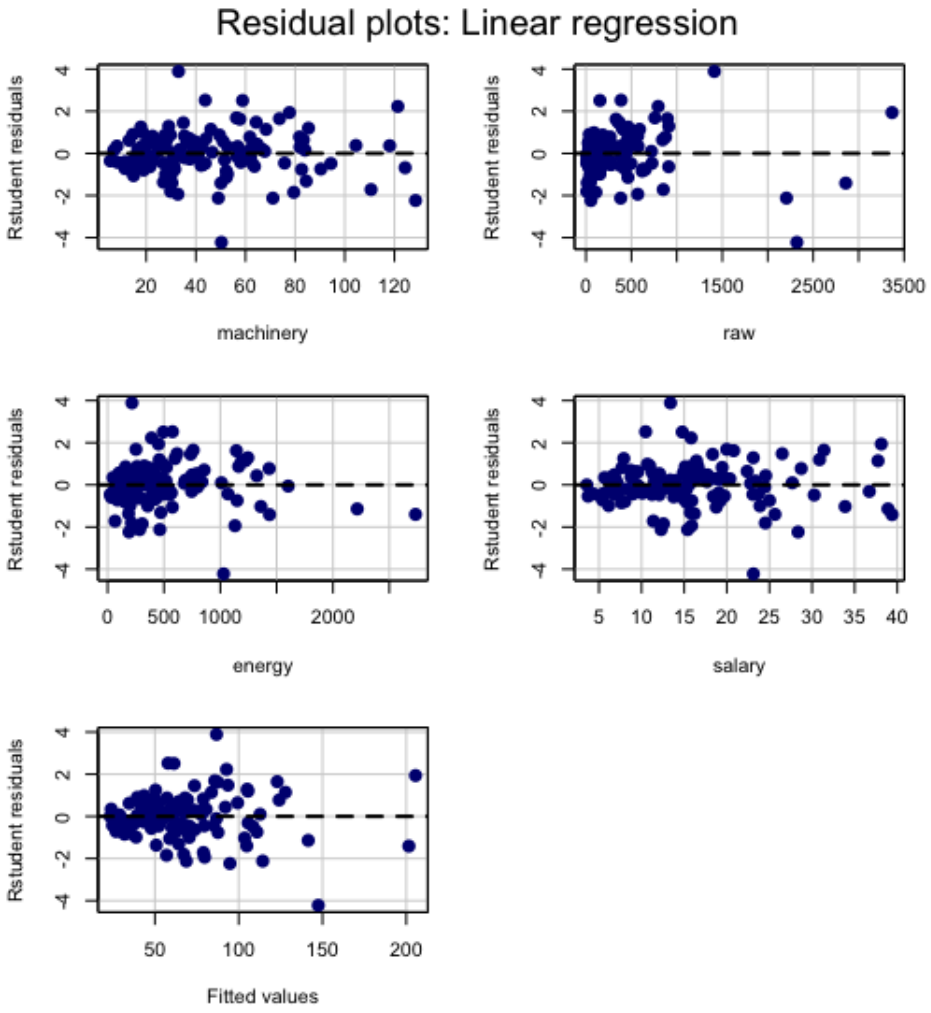


Figure 20: Residual plots for all explanatory variables with multiple linear regression

| The log-log regression model | | | | |
|---|---|---|---|---|
| Coefficient | Estimate | Std. Error | t-Test | p-value |
| $\hat{\beta}_0$ | 0.8539 | 0.1459 | 5.85 | 0.0000*** |
| $\hat{\beta}_1$ | 0.1219 | 0.0329 | 3.71 | 0.0003** |
| $\hat{\beta}_2$ | 0.1244 | 0.0158 | 7.86 | 0.0000*** |
| $\hat{\beta}_3$ | 0.1611 | 0.0262 | 6.15 | 0.0000*** |
| $\hat{\beta}_4$ | 0.4570 | 0.0492 | 9.29 | 0.0000*** |
| Significance codes: 0***, 0.001**, 0.05* | | | | |

Table 10: Results from fitting the multiple log-log regression model

The adjusted R-squared value for this regression is 0.8301, which indicate that the log-log model provide a better fit. We further investigate the log-log regression model by plotting the residuals. The residual plots are presented in Figure 22.

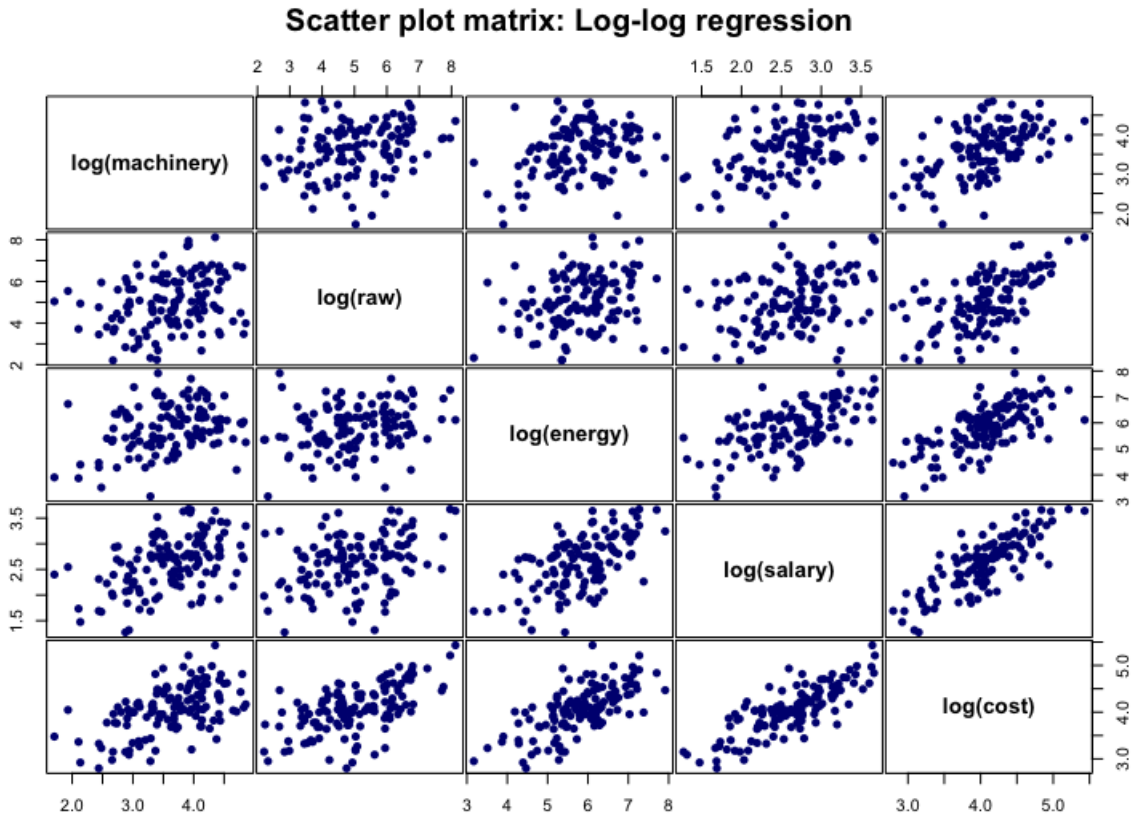


Figure 21: Scatter plot matrix showing the pairwise log-log relation in the cost data set

The residual plots support the assumption that the log-log regression is more appropriate in explaining the relation between the unit production cost and the regressor variables.

## (d) Inference

**(i) Regressor Significance** The results from the log-log regression is presented in Table 10. The results indicate that all variables are significant with a significance level of 5 %, as the p-values for all parameters are less than 0.05. From Table 10 we have with a 0.05 level that the machinery cost $X_1$ is significant in controlling the other variables, having a p-value of 0.0003. The other variables are also significant, raw cost $X_2$ with a p-value of 0.0000, the p-value for energy cost $X_3$ is 0.0000 and the salary cost $X_4$ is 0.0000.

**(ii) Confidence Interval** We further construct a 95 % confidence interval for the regression coefficients of the explanatory variables. The 95 % confidence interval for the machinery cost coefficient, $\beta_1$, is [0.0568, 0.1870], for material cost, $\beta_2$, [0.0930, 0.1557], for energy cost, $\beta_3$, [0.1092, 0.2130] and for salary cost, $\beta_4$, [0.3596, 0.5543]. As we can see, none of these intervals include zero, we can therefore conclude that all the explanatory variables have a

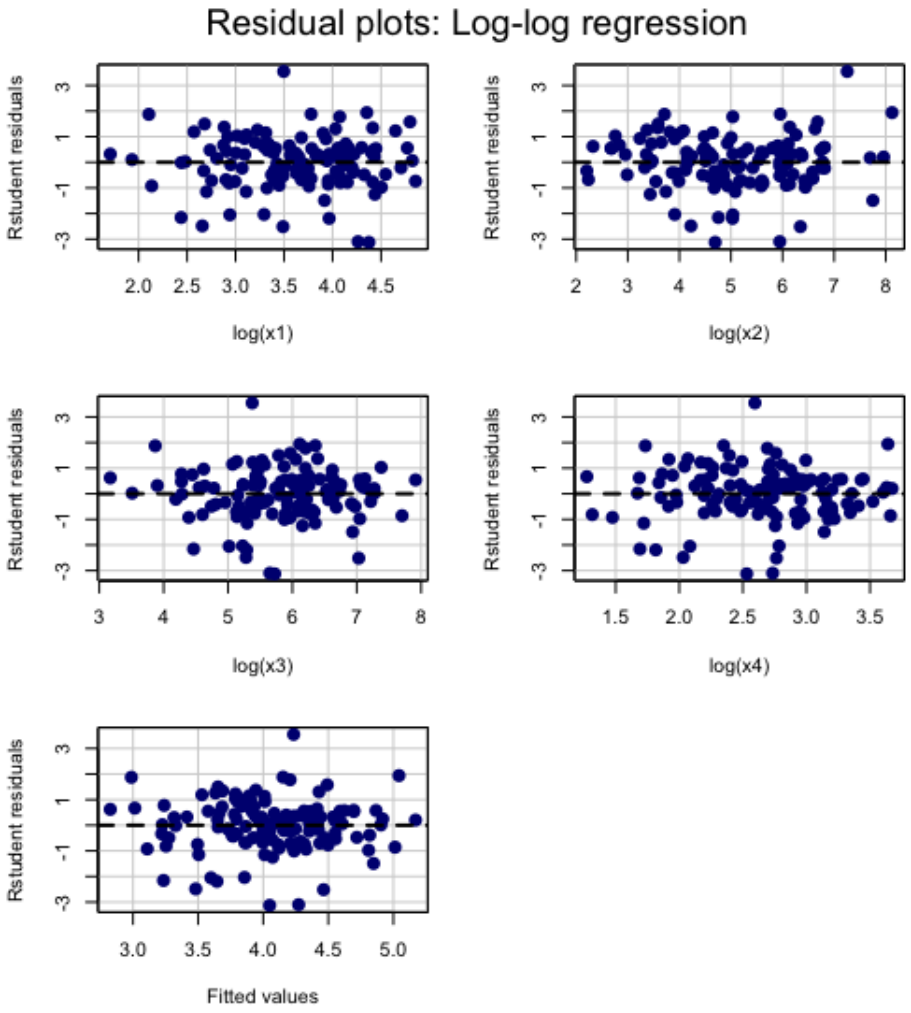


Figure 22: Residual plots for all explanatory variables with multiple log-log regression

positive impact on the unit cost of production, $Y$, in the sense that an increase in any of the explanatory cost variables, the cost of production will increase. In general salary cost seems to be the most influential cost, as the lower bound for this confidence interval is higher than the upper bound for all other cost variables. From an economic perspective this results are realistic.

**(iii) Predicted cost** We can also predict the regression model for specific values of the explanatory variables. The predicted cost can be obtained by constructing a 95 % prediction interval for machinery cost $X_1 = 80$, raw cost $X_2 = 2000$, energy cost $X_3 = 1000$ and salary cost $X_4 = 25$. Similarly to the exercise in the previous chapter, we have to perform back-transformation in order to obtain the estimated value of $\hat{Y}$. By doing this we get that the predicted unit cost of production for the specific values of the cost variables is thus $\hat{Y} = 136.5999$. Thus is this scenario we would expect the unit cost to be about \$136.6.

# 5 Exercise 2

In this exercise we want to investigate the consumption pattern of cigarettes, by analysing data from different states in the United States. The variables we are studying are presented in Table 11.

| # | Variable | Definition |
|---|---|---|
| $X_1$ | Age | Median age of a person living in a state |
| $X_2$ | HS | Percentage of people over 25 years of age in a state who had completed high school |
| $X_3$ | Income | Per capita personal income for a state (income in dollars) |
| $X_4$ | Black | Percentage of blacks living in a state |
| $X_5$ | Female | Percentage of females living in a state |
| $X_6$ | Price | Weighted average price (in cents) of a pack of cigarettes in a state |
| $Y$ | Sales | Number of packs of cigarettes sold in a state on a per capita basis |

Table 11: Variables in cigarette consumption data (Chatterjee and Hadi, 2015)

The response variable is the consumption of cigarettes measured in number of packs of cigarettes sold. Our initial assumption is that not all regressors affect the cigarette consumption and we therefore want to compare the suitability of different multiple regressions. We further expect that some logarithmic transformation is necessary to provide the best fit in the consumption data.

Before fitting the regression model we want to want to investigate how the variables are related. This can be done by constructing scatter plots of all pairwise combinations of the variables in the data set. Figure 23 show the scatter plot matrix for the given data set.

From the scatter plot its seems that age, high school education and income have a slightly positive impact on cigarette consumption, although these may be better explained in logs. Black and female does not seem to have a clear impact on consumption and the relation between price and consumption seems to be non-linear. We start by testing if the explanatory variables collectively have an effect on the response variable by constructing the following null hypothesis:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = 0 \tag{30}$$

The null hypothesis state that the slope variables of all explanatory variables are zero. If we can reject the null hypothesis we can proceed by analyzing the significance of all the regression coefficients while controlling for the other variables in the model. By performing a F-test, we get a p-value of 0.006857. With a 0.05 level we therefore reject the null hypothesis. A multiple linear regression model, including all regressors, can be expressed by the following equation:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3 + \hat{\beta}_4 X_4 + \hat{\beta}_5 X_5 + \hat{\beta}_5 X_5 + \hat{\beta}_6 X_6 \tag{31}$$

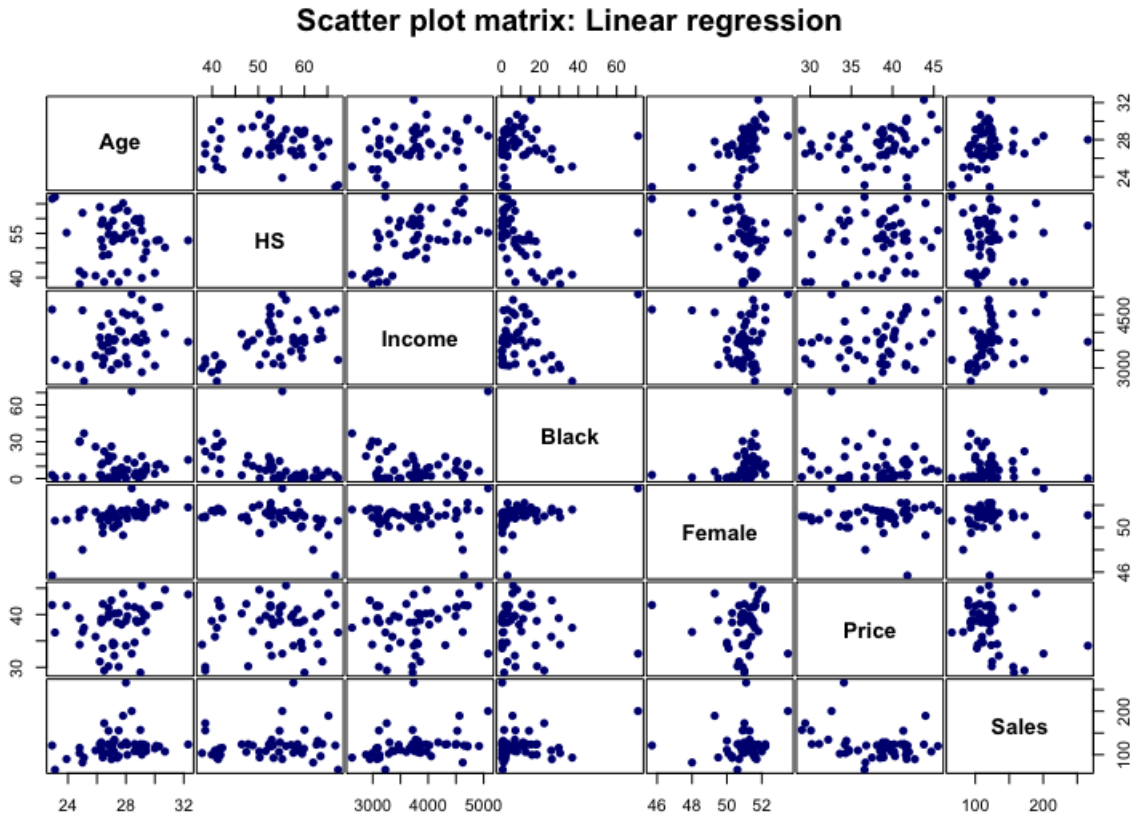


Figure 23: Scatter plot matrix showing the relationship between the variables in the consumption data set

where $\hat{Y}$ is the estimated sales explained by average age $X_1$, percentage with high school education $X_2$, average income $X_3$, percentage blacks $X_4$, percentage females $X_5$ and average price $X_6$. By fitting the full regression model we get the results presented in Table 12. This gives us the following regression.

$$\hat{Y} = 103.3449 + 4.5205X_1 - 0.0616X_2 + 0.0190X_3 + 0.3575X_4 - 1.0529X_5 - 3.2549X_6 \quad (32)$$

From this regression we can see that age, income and black seems to have a positive effect on the cigarette consumption and female, high school and price have a negative effect on consumption. The fact that age and income have a positive effect on cigarette consumption is reasonable. However the effect is quit small, as these number are average numbers for a given state. It is also reasonable that education and price have a negative effect on the consumption. However, the adjusted R-squared value for this regression is 0.2282, which suggest that this regression provide a week fit for the given data.

We can further investigate the full regression model by constructing a diagnostic plot, plotting the studentized residuals for all explanatory variables. The residual plots in Figure 20 support

| The full regression model | | | | |
|---|---|---|---|---|
| Coefficient | Estimate | Std. Error | t-Test | p-value |
| $\beta_0$ | 103.3449 | 245.6072 | 0.42 | 0.6760 |
| $\beta_1$ | 4.5205 | 3.2198 | 1.40 | 0.1674 |
| $\beta_2$ | -0.0616 | 0.8147 | -0.08 | 0.9401 |
| $\beta_3$ | 0.0190 | 0.0102 | 1.86 | 0.0704 |
| $\beta_4$ | 0.3575 | 0.4872 | 0.73 | 0.4670 |
| $\beta_5$ | -1.0529 | 5.5610 | -0.19 | 0.8507 |
| $\beta_6$ | -3.2549 | 1.0314 | -3.16 | 0.0029* |
| Significance codes: 0***, 0.001**, 0.05* | | | | |

Table 12: Results from fitting the full linear regression model

the assumption that excluding female might be appropriate. The residual plots also suggest the occurrence of some outliers in the data set.

## (a) The Reduced Regression Model I

The results presented in Table 12 suggest that female is not significant in controlling the other variables in the model, with a p-value of 0.85071. We are therefore interested in testing the regression, excluding the variable female. We proceed by fitting the regression, excluding the female variable. This implies testing whether the corresponding coefficient is equal to zero. This gives us the following hypothesis':

$$H_0 : \beta_5 = 0 \tag{33}$$

$$H_1 : \beta_5 \neq 0 \tag{34}$$

The null hypothesis in the two-tailed test, states that female is excluded from the regression by setting $\beta_5 = 0$ and the alternative hypothesis states that the coefficient associated with the female variable, $\beta_5$, either has a positive of negative impact on cigarette consumption. The test is performed using a partial F-test and involves comparing the sum of squared errors (SSE) from the reduced regression model to the SSE from the full regression model. The partial F-test is conducted by comparing the two fitted models with anova. By excluding the female variable we obtain the following reduced regression model:

$$\hat{Y} = \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3 + \hat{\beta}_4 X_4 + \hat{\beta}_6 X_6 \tag{35}$$

The results from the fitted model is presented in Table 13.

The adjusted R-squared value for this regression is 0.2448, which is a slight improvement from the previous model. The results further suggest that price, $X_6$, is significant in controlling the other variables. However, none of the other variables are significant.

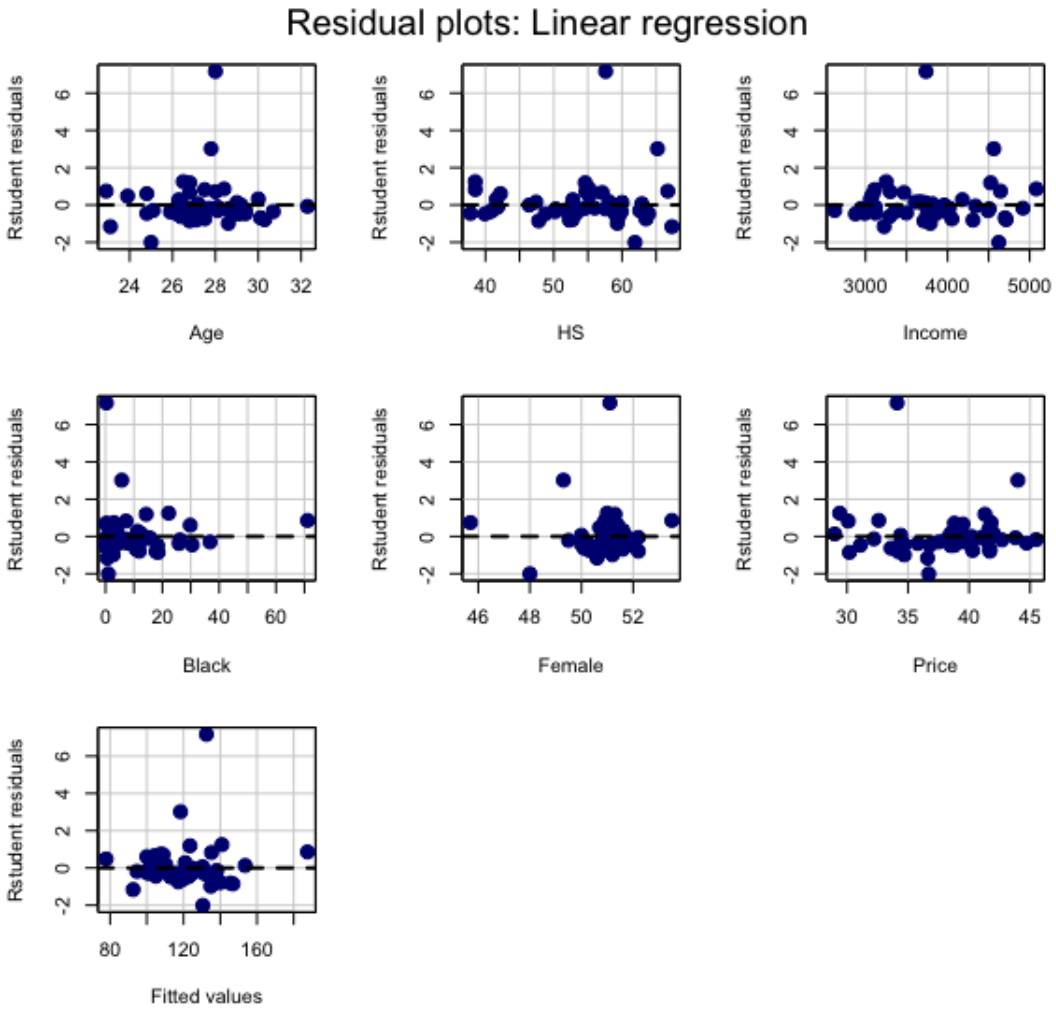


Figure 24: Residual plots for all explanatory variables with multiple linear regression

| The reduced regression model I | | | | |
|---|---|---|---|---|
| Coefficient | Estimate | Std. Error | t-Test | p-value |
| $\beta_0$ | 59.4633 | 80.3876 | 0.74 | 0.4633 |
| $\beta_1$ | 4.1178 | 2.3912 | 1.72 | 0.0919 |
| $\beta_2$ | -0.0668 | 0.8054 | -0.08 | 0.9343 |
| $\beta_3$ | 0.0195 | 0.0097 | 2.00 | 0.0519 |
| $\beta_4$ | 0.3115 | 0.41756 | 0.75 | 0.4596 |
| $\beta_6$ | -3.2520 | 1.0202 | -3.19 | 0.00261* |
| Significance codes: 0***, 0.001**, 0.05* | | | | |

Table 13: Results from fitting the reduced regression model, excluding the female variable

We compare the reduced regression with the reduced regression by performing an Anova analysis. From this analysis we obtain a p-value of 0.8507, which is quit high. With a 5

% significance level we will certainly not reject the null hypothesis $\beta_6 = 0$. In other word the variable female does not contribute significant information to the sales once the other variables are taken into consideration and should therefore be excluded from the regression.

### (b) The Reduced Regression Model II

We now want to test whether a certain subset of the coefficients simultaneously are equal to zero. From Table 13 we have that the p-value for high school education is quit high. We therefore want to test the regression, excluding both female and high school education. This implies testing whether the corresponding coefficients are equal to zero. This gives us the following hypothesis':

$$H_0 : \beta_2 = \beta_5 = 0 \tag{36}$$

$$H_1 : \beta_2 \neq 0, \beta_5 \neq 0 \text{ or both} \tag{37}$$

By excluding the female and high school variables we obtain the following reduced regression model:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_3 X_3 + \hat{\beta}_4 X_4 + \hat{\beta}_6 X_6 \tag{38}$$

The results from the fitted model is presented in Table 14.

| The reduced regression model II | | | | |
|---|---|---|---|---|
| Coefficient | Estimate | Std. Error | t-Test | p-value |
| $\beta_0$ | 55.3296 | 62.3953 | 0.89 | 0.3798 |
| $\beta_1$ | 4.1915 | 2.1955 | 1.91 | 0.0625 |
| $\beta_3$ | 0.0189 | 0.0069 | 2.75 | 0.0086* |
| $\beta_4$ | 0.3342 | 0.3121 | 1.07 | 0.2899 |
| $\beta_6$ | -3.2399 | 0.9988 | -3.24 | 0.0022* |
| Significance codes: 0***, 0.001**, 0.05* | | | | |

Table 14: Results from fitting the reduced regression model, excluding the female and high school education variables

The adjusted R-squared value for this regression is 0.2611, which is a slight improvement from the previous model. The p-values in Table 14 suggest that income, $X_3$, and price, $X_6$, are significant in controlling the other variables. However, age, $X_1$, and black, $X_4$, are still not significant.

When comparing the two regressions with Anova we get a p-value 0.9789. With a 5 % significance level we can certainly not to reject the null hypothesis $\beta_2 = \beta_5 = 0$. In other word the variables female and high school does not contribute significant information to the cigarette consumption, once the other variables are taken into consideration.

### (c) Confidence Interval

We now construct a 95 % confidence interval for the income coefficient $\beta_3$, using the reduced model, excluding both the variables female and high school. The interval obtained is [-0.0016, 0.0395]. This imply that it is not clear whether income have a positive or a negative effect on cigarette consumption. The effect from income on consumption is between -0.16 % and 3.95 %. This also imply that we can reject the null hypothesis that $\beta_3 = 0$, so one may consider removing income from the regression.

### (d) Variation from Income

The total variation in sales, when the variable income is removed from the regression can be found be looking at the R-squared value from the reduced regression, excluding income. This regression can be expressed by the following equation.

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_4 X_4 + \hat{\beta}_5 X_5 + \hat{\beta}_6 X_6 \tag{39}$$

This reduced regression have an R-squared value of 0.2678, which implies that the total variation in sales explained by the model, with income is excluded is 26.78 %.

### (e) Variation from Age, Income and Price

We can further analyze the situation when female, high school and black is excluded from the regression. From the initial scatter plot in Figure **??**, there is no clear relation between black and sales. The reduced regression can be expressed by the following equation.

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_3 X_3 + \hat{\beta}_6 X_6 \tag{40}$$

Here sales is expressed by age $X_{1t}$, income $X_{3t}$ and price $X_{6t}$. The R-squared value of this regression is 0.3032. Thus the percentage of total variation in sales accounted for by age, income and price is 30.32 %. Even though we are not suppose to use R-square to compare different models, we can see that although this regression have less explanatory variables than the one above, it is still able to describe a larger fraction of the total variation in the data set.

### (f) Variation from Income in Simple Regression

In the case when sales is regressed only by the variable income, we get the following regression equation.

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_3 X_3 \tag{41}$$

The R-squared value implies that the percentage of total variation in sales accounted for by income is 10.63 %. In other words, income alone is able to describe 10.63 % of the variability in the data set.

# 6 Exercise 3

In this exercise we want to investigate factors affecting the life expectancy, using data from the 50 different states in the United States (Becker and Wilks, 1988). The variables we are studying are presented in Table 15

| # | Variable | Definition |
|---|---|---|
| $X_1$ | Population | Population estimate (1975) |
| $X_2$ | Income | Per capita income (1974) |
| $X_3$ | Illiteracy | Illiteracy in percent of population (1970) |
| $X_4$ | Murder | Murder and non-negligent manslaughter rate per 100,000 population (1976) |
| $X_5$ | HS Grad | Percent high-school graduates (1970) |
| $X_6$ | Frost | Mean number of days with minimum temperature below freezing in capital or large city (1931–1960) |
| $X_7$ | Area | Land area in square miles |
| $Y$ | Life Exp | Life expectancy in years (1969–71) |

Table 15: Variables in the state data set (Becker and Wilks, 1988)

## (a) Full Regression Model

Before fitting the regression model we want to want to investigate how the variables are related. This can be done by constructing scatter plots of all pairwise combinations of the variables in the data set. Figure 25 show the scatter plots from the given data set.

From the scatter plot matrix we can see that multiple linear regression makes sense. Life expectancy seems to have a linear relationship with income, illiteracy, murder rate, high school graduates. For population and area the relation seem to be log-linear. However, there does not seems to be a clear relation between life expectancy and frost. Notice that when plotting a scatter plot matrix the values on the x and y axis' are fixed. This may affect the relation between the variables in the data set. In order to get more accurate result each scatter plot should be investigated in isolation. For the purpose of this report, we will use the scatter plot as a reference and not look at each scatter plot in detail. For the murder rate the residuals seems to be constant. However, for the other explanatory variables there is some variation in the residuals. This may suggest that a multiplicative error model is appropriate.

We first test if the explanatory variables collectively have an effect on the response variable

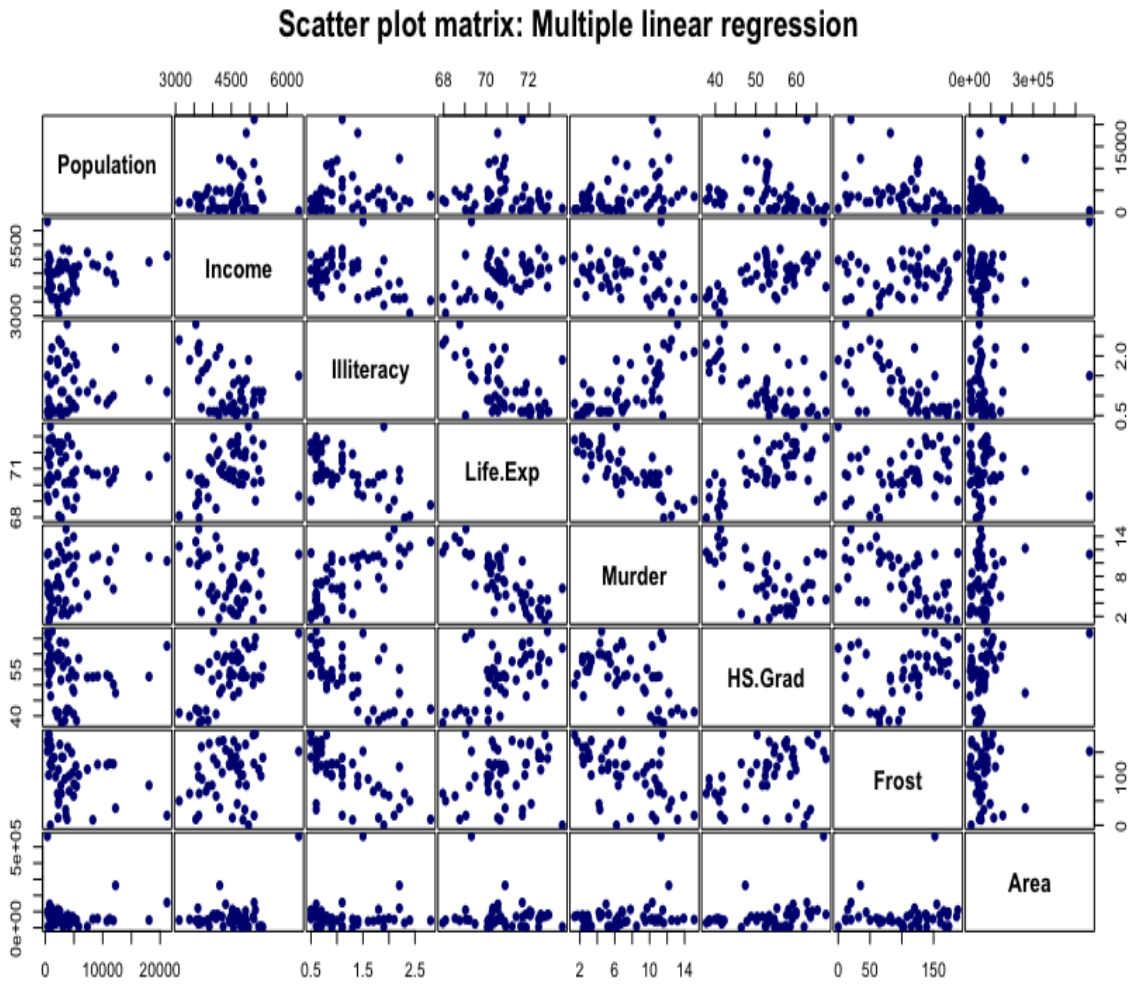


Figure 25: Scatter plot matrix showing the pairwise relation between the variables in the state data set

by constructing the following null hypothesis:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = 0 \tag{42}$$

The null hypothesis state that the slope variables of all explanatory variables are zero. If we can reject the null hypothesis we can proceed by analyzing the significance of all the regression coefficients while controlling for the other variables in the model. By performing a F-test, we get a p-value of 0.0000. With a 5 % significance level we therefore reject the null hypothesis. Thus the full multiple linear regression model is expressed by the following equation

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3 + \hat{\beta}_4 X_4 + \hat{\beta}_5 X_5 + \hat{\beta}_6 X_6 + \hat{\beta}_7 X_7 \tag{43}$$

where $\hat{Y}$ is the estimated life expectancy explained by population $X_1$, income $X_2$, illiteracy $X_3$, murder rate $X_4$, percentage of high school graduates $X_5$, number of days per year with frost $X_6$ and area of the state $X_7$. By fitting the full regression model we get the results presented in Table 16. This gives us the following regression.

$$\hat{Y} = 70.9432 + 0.0001X_1 + 0.0338X_3 - 0.3011X_4 + 0.0489X_5 - 0.0057X_6 \tag{44}$$

As we can see from this regression population have a small positive impact on life expectancy. Illiteracy and high school graduates have positive impact on life expectancy, while murder rate and frost have a negative impact on life expectancy. This seems reasonable. We further see that frost has a smaller impact than the murder rate which also is intuitive. From this regression, income and area has no impact on life expectancy. From the p-values we can see that the income, illiteracy and area have high p-values. We should therefore consider excluding these from the regression.

| The full regression model | | | | |
|---|---|---|---|---|
| Coefficient | Estimate | Std. Error | t-Test | p-value |
| $\beta_0$ | 70.9432 | 1.7480 | 40.59 | $0.0000^{***}$ |
| $\beta_1$ | 0.0001 | 0.0000 | 1.77 | 0.0832 |
| $\beta_2$ | -0.0000 | 0.0002 | -0.09 | 0.9293 |
| $\beta_3$ | 0.0338 | 0.3663 | 0.09 | 0.9269 |
| $\beta_4$ | -0.3011 | 0.0466 | -6.46 | $0.0000^{***}$ |
| $\beta_5$ | 0.0489 | 0.0233 | 2.10 | $0.0420^{*}$ |
| $\beta_6$ | -0.0057 | 0.0031 | -1.82 | 0.0752 |
| $\beta_7$ | -0.0000 | 0.0000 | -0.04 | 0.9649 |
| Significance codes: $0^{***}$, $0.001^{**}$, $0.05^{*}$ | | | | |

Table 16: Result from the fitting the full linear regression model

The adjusted R-squared value for this regression is 0.6922, which suggest that it provides quit a good fit.

## (b) Reduced Regression Model

We are now interested in testing the regression, excluding the variables income, illiteracy and area. This in order to test if the life expectancy essentially is a linear function of population, murder, high school graduates and frost. This implies testing whether the coefficients corresponding to income, illiteracy and area are equal to zero. This gives us the following hypothesis':

$$H_0 : \beta_2 = \beta_3 = \beta_7 = 0 \tag{45}$$

$$H_1 : \beta_2 \neq 0, \beta_3 \neq 0, \beta_7 \neq 0 \text{ or all} \tag{46}$$

The test is performed using a partial F-test and involves comparing the SSE from the reduced regression model to the SSE from the full regression model. The partial F-test is conducted by comparing the two fitted models with Anova. By excluding the income, illiteracy and area we obtain the following reduced regression model:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_4 X_4 + \hat{\beta}_5 X_5 + \hat{\beta}_6 X_6 \tag{47}$$

The results from the fitted model is presented in Table 17.

| The reduced regression model I | | | | |
|---|---|---|---|---|
| Coefficient | Estimate | Std. Error | t-Test | p-value |
| $\beta_0$ | 71.0271 | 0.9529 | 74.54 | 0.0000*** |
| $\beta_1$ | 0.0001 | 0.0000 | 2.00 | 0.0520 |
| $\beta_4$ | -0.3001 | 0.0366 | -8.20 | 0.0000*** |
| $\beta_5$ | 0.0466 | 0.0148 | 3.14 | 0.0030* |
| $\beta_6$ | -0.0059 | 0.0024 | -2.46 | 0.0180* |
| Significance codes: 0***, 0.001**, 0.05* | | | | |

Table 17: Results from fitting the reduced regression model, excluding income, illiteracy and area

The adjusted R-squared value for this regression is 0.7126, which suggest that the reduced regression provides a better fit than the full regression model. Inserting for the coefficients gives us the following regression:

$$\hat{Y} = 71.0271 + 0.0001 X_1 - 0.3001 X_4 + 0.0466 X_5 - 0.0059 X_6 \tag{48}$$

When comparing the two regressions using Anova we get a p-value 0.9993. With a 5 % significance level we can be quit certain not to reject the null hypothesis $\beta_2 = \beta_3 = \beta_7 = 0$. In other words the variable income, illiteracy and area does not contribute significant information to the life expectancy once the other variables are taken into consideration. As a result we can establish that life expectancy is essentially a linear function of the predictors population, murder, high school graduates and frost.

### (c) Log Transformation

In order to investigate whether the variables population and area are better expressed using a logarithmic transformation we look at the residual plots for the full regression, presented in Figure 26.

The plots suggest that population and area fails to fit the data, using a linear fit. We redefine the scatter plot, by taking introducing log in population and area. The scatter plot is presented in Figure 27.

The scatter plot suggest an improvement in population when performing the log transformation. However, there is still no clear linear relationship between life expectancy and area.

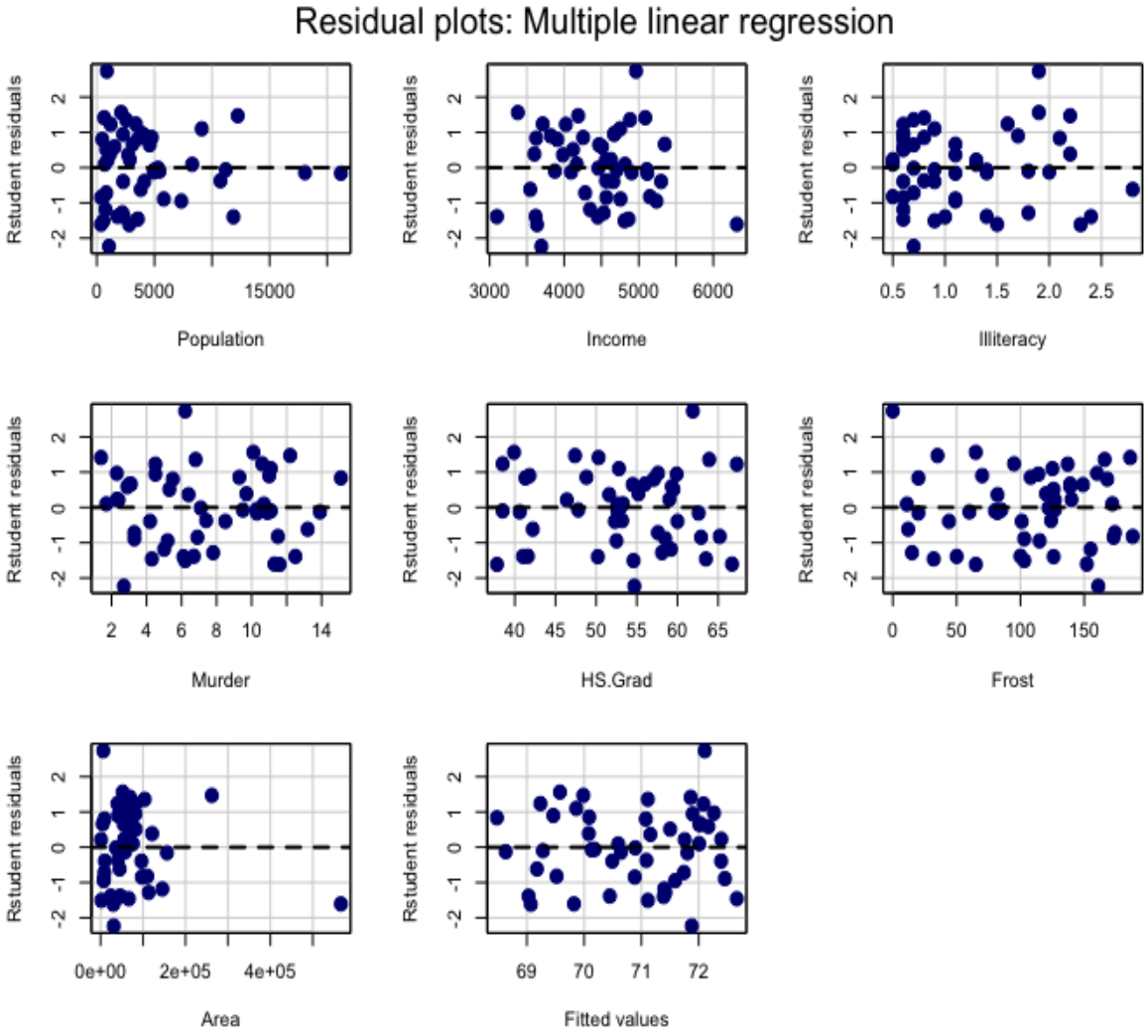


Figure 26: Scatter plot matrix showing the relationship between the variables in the state data

This indicate that area should be excluded from the model. We first introduce the regression model with a logarithmic transformation on both population and area. This gives us the following regression.

$$\hat{Y}_t = \hat{\beta}_0 + \hat{\beta}_1 log(X_1) + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3 + \hat{\beta}_4 X_4 + \hat{\beta}_5 X_5 + \hat{\beta}_6 X_6 + \hat{\beta}_7 log(X_7) \tag{49}$$

The result from this regression is presented in Table 18. Using a log transformation gives us an adjusted R-squared value of 0.7008, which is worse than the reduced model.

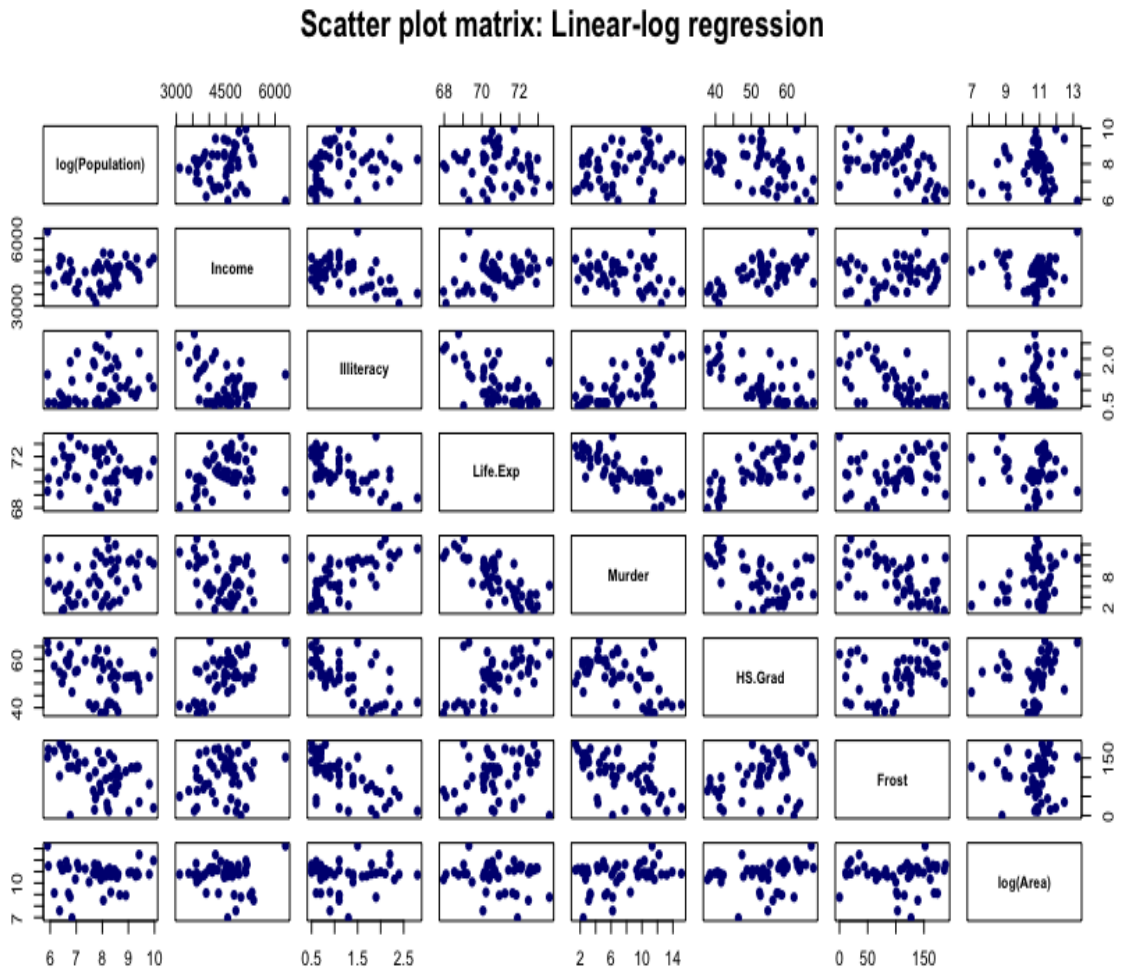


Figure 27: Scatter plot matrix showing the relationship between the variables in the state data

From the table we see that income, illiteracy and log area have high p-values. Therefore we again consider excluding these variables, keeping the log transformation in population. The residual plots for this regression is given in Figure 28. The plots shows that there still seems to be a pattern in the studentized residual for the log transformation of area.

We therefore try to exclude income, illiteracy and area from the regression model. This gives us the following regression.

$$\hat{Y}_t = \hat{\beta}_0 + \hat{\beta}_1 log(X_1) + \hat{\beta}_4 X_4 + \hat{\beta}_5 X_5 + \hat{\beta}_6 X_6 \tag{50}$$

| The full linear-log regression model | | | | |
|---|---|---|---|---|
| Coefficient | Estimate | Std. Error | t-Test | p-value |
| $\beta_0$ | 67.9529 | 2.0923 | 32.48 | 0.0000*** |
| $\beta_1$ | 0.2527 | 0.1351 | 1.87 | 0.0685 |
| $\beta_2$ | 0.0000 | 0.0002 | 0.06 | 0.9547 |
| $\beta_3$ | 0.1126 | 0.3507 | 0.32 | 0.7497 |
| $\beta_4$ | -0.3092 | 0.0471 | -6.57 | 0.0000*** |
| $\beta_5$ | 0.0528 | 0.0248 | 2.13 | 0.0394* |
| $\beta_6$ | -0.0049 | 0.0032 | -1.51 | 0.1373 |
| $\beta_7$ | 0.0686 | 0.1098 | 0.62 | 0.5354 |
| Significance codes: 0***, 0.001**, 0.05* | | | | |

Table 18: Results from fitting the full linear-log regression with log in population and area

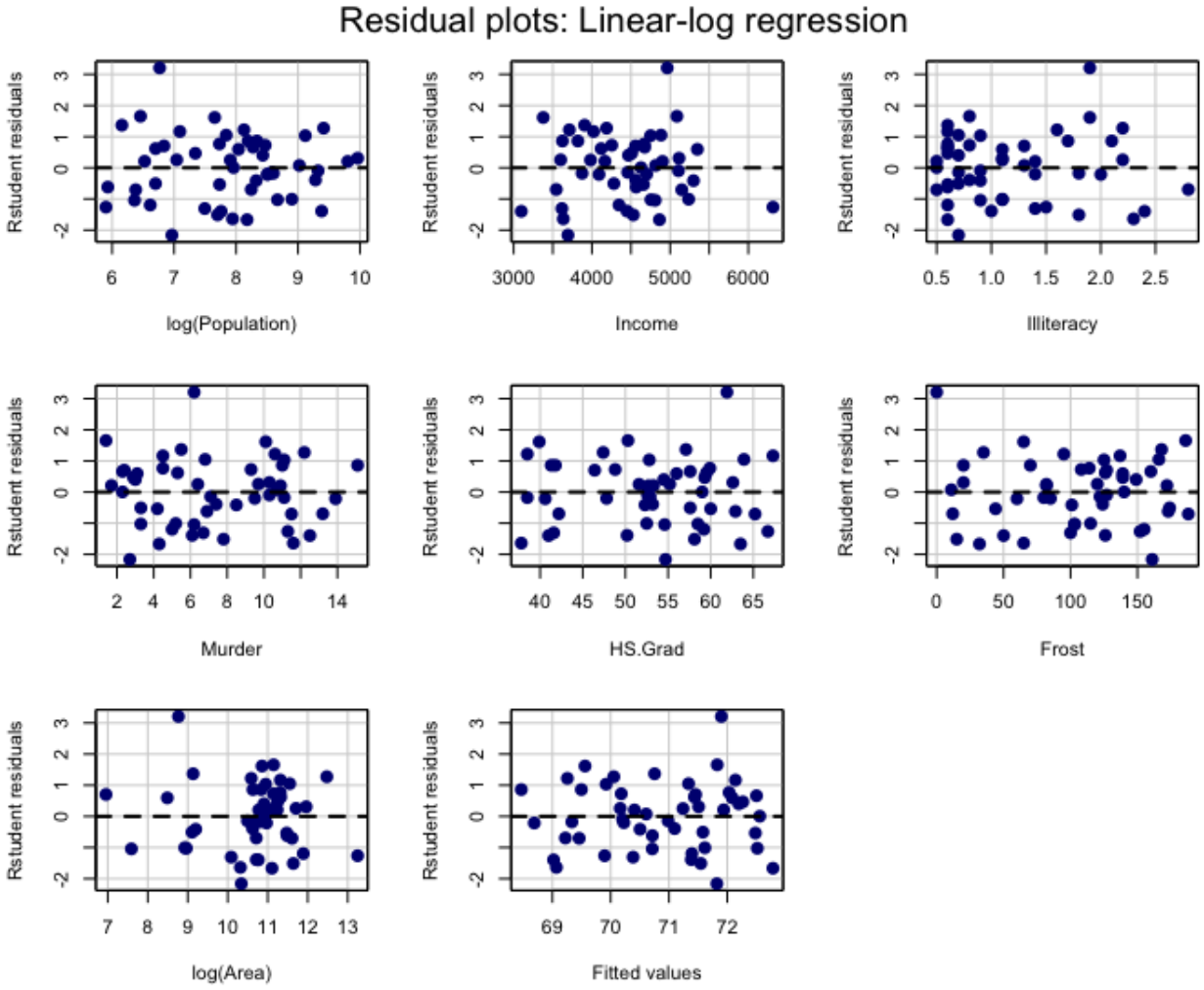


Figure 28: Residual plots after log transformation in population and area

The result from this regression is presented in Table 19. Excluding income, illiteracy and area result in an adjusted R-squared value of 0.7173, which is slightly better than the reduced model without a log transformation in population.

| The reduced regression model II | | | | |
|---|---|---|---|---|
| Coefficient | Estimate | Std. Error | t-Test | p-value |
| $\beta_0$ | 68.7208 | 1.4168 | 48.50 | $0.0000^{***}$ |
| $\beta_1$ | 0.2468 | 0.1125 | 2.19 | $0.0335^{*}$ |
| $\beta_4$ | -0.2900 | 0.0354 | -8.18 | $0.0000^{***}$ |
| $\beta_5$ | 0.0546 | 0.0148 | 3.70 | $0.0006^{**}$ |
| $\beta_6$ | -0.0052 | 0.0025 | -2.08 | $0.0428^{*}$ |
| Significance codes: $0^{***}$, $0.001^{**}$, $0.05^{*}$ | | | | |

Table 19: Results from fitting the reduced regression model with log(population), excluding income, illiteracy and area

With a 5 % significance level all the parameters are significant. When performing Anova to compare the reduced regression using log transformation in population and excluding income, illiteracy and area, with the full regression model including log transformation in both population and area we get a p-value equal to 0.9168. The p-value is high, which indicates that we can be quit certain to exclude income, illiteracy and area from our model. The adjusted R-squared, in addition to the interpretation of the plots further suggest that the reduced model with log transformation in population is the most appropriate model in describing the data.

# Part III

# General Linear Hypothesis

## 7 Exercise 1

Before fitting the regression model we want to investigate how the variables are related. This can be done by constructing scatter plots of all pairwise combinations of the variables in the data set. Figure **??** show the scatter plots from the given data set.

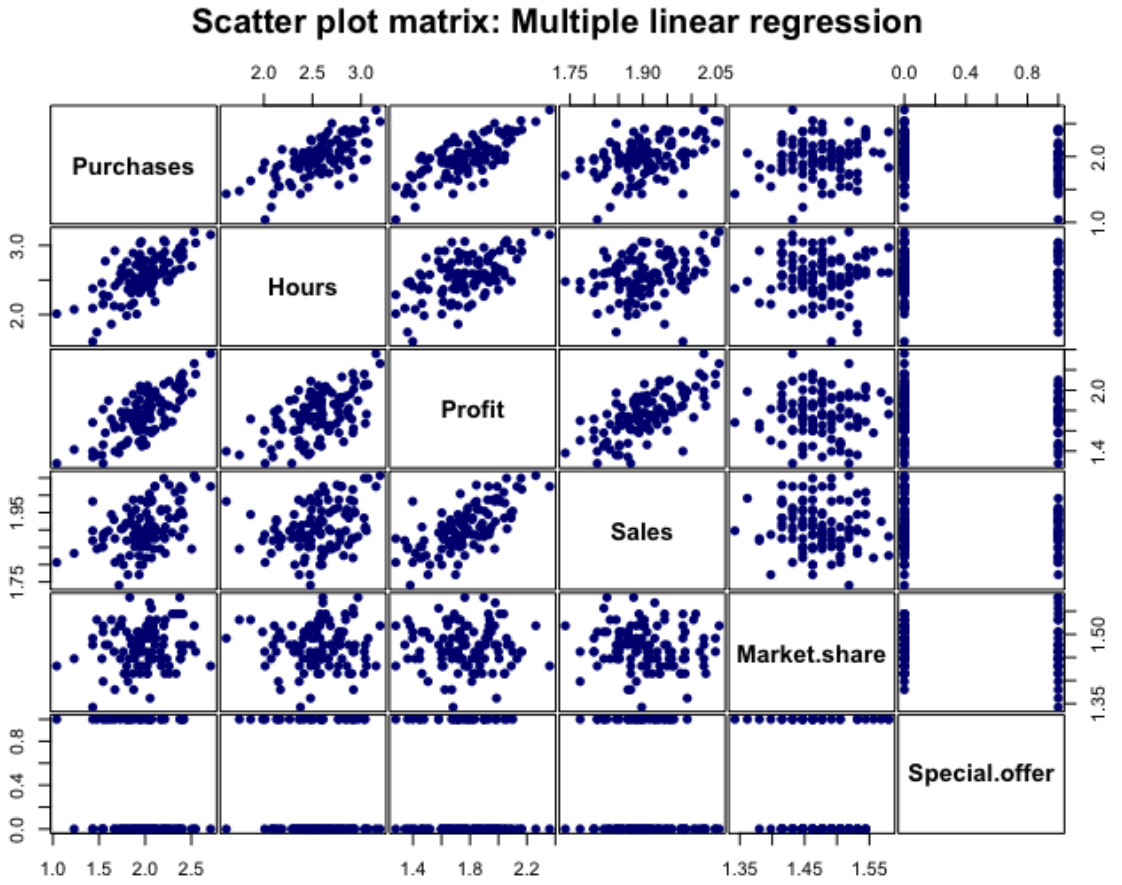


Figure 29: Scatter plot matrix showing the relationship between the variables in the profit data

### Full Regression Model

We first test if the explanatory variables, except the binary variable special offer, collectively have an effect on the response variable by constructing the following null hypothesis:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0 \tag{51}$$

The null hypothesis state that the slope variables of all explanatory variables are zero. If we can reject the null hypothesis we can proceed by analyzing the significance of all the regression coefficients while controlling for the other variables in the model. By performing a F-test, we get a p-value of $< 2.2e - 16$. With a 5 % significance level we therefore reject the null hypothesis. A multiple linear regression model, including all regressors can be expressed by the following equation:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3 + \hat{\beta}_4 X_4 \tag{52}$$

where $\hat{Y}$ is the estimated profit explained by purchases $X_1$, sales $X_2$, hours of operation $X_3$, and market share $X_4$. By fitting the full regression model we get the results presented in Table 20. This gives us the following regression:

$$\hat{Y} = -1.5638 + 0.3515X_1 + 1.4495X_2 + 0.0836X_3 - 0.2356X_4 \tag{53}$$

| The full regression model | | | | |
|---|---|---|---|---|
| Coefficient | Estimate | Std. Error | t-Test | p-value |
| $\beta_0$ | -1.56384 | 0.5721 | -2.73 | $0.0074^{**}$ |
| $\beta_1$ | 0.35148 | 0.0617 | 5.70 | $0.0000^{***}$ |
| $\beta_2$ | 1.44948 | 0.2008 | 7.22 | $0.0000^{***}$ |
| $\beta_3$ | 0.08356 | 0.0533 | 1.57 | 0.1202 |
| $\beta_4$ | -0.23562 | 0.2666 | -0.88 | 0.3789 |
| Significance codes: $0^{***}$, $0.001^{**}$, $0.05^{*}$ | | | | |

Table 20: Results from fitting the full regression model

The adjusted R-squared value for this regression is 0.6922, which indicate that this regression provide quit a good fit of the data. However when analyzing the significance of each of the variable we can see that hours and market share is not significant with a 5 % significance level. We therefore proceed by analyzing whether the model is improved by removing these variables from the regression.

### Reduced Regression Model

From the full regression model it seems as if the profit is mostly determined by purchases $X_1$ and sales $X_2$. Therefore we are now interested in testing the regression, excluding the variable operation $X_3$ and market share $X_4$. This implies testing whether the corresponding

coefficient $\beta_3$ and $\beta_4$ is equal to zero. This can be formalized using the follow general linear hypothesis':

$$H_0 : \beta_3 = \beta_4 = 0 \tag{54}$$

$$H_1 : \beta_3 \neq 0, \beta_4 \neq 0 \text{ or both} \tag{55}$$

The test is performed using a partial F-test and involves comparing the SSE from the reduced regression model to the SSE from the full regression model. The partial F-test is conducted by comparing the two fitted models with Anova. By excluding the female variable we obtain the following reduced regression model:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 \tag{56}$$

Inserting for the coefficients gives us the following regression.

$$\hat{Y} = -1.8685 + 0.4026 X_1 + 0.1987 X_2 \tag{57}$$

The results from the fitted model is presented in Table 21.

| The reduced regression model | | | | |
|---|---|---|---|---|
| Coefficient | Estimate | Std. Error | t-Test | p-value |
| $\beta_0$ | -1.8685 | 0.3464 | -5.39 | 0.0000*** |
| $\beta_1$ | 0.4026 | 0.0466 | 8.64 | 0.0000*** |
| $\beta_2$ | 1.4862 | 0.1987 | 7.48 | 0.0000*** |
| Significance codes: 0***, 0.001**, 0.05* | | | | |

Table 21: Results from fitting the reduced regression model, excluding hours and market share

The adjusted R-squared value for this regression is 0.6873, which is slightly lower than the previous model. We compare the two models by performing an Anova analysis. From Table **??** the results from the Anova is presented. When comparing the two models we get a p-value of 0.1631. This value is not very rigorous. However, with a significance level of 5 % we do not have evidence against the reduced model. We can therefore not reject the null hypothesis that hours and market share are excluded. In Table **??** we have that the residual sum of squares associated with the reduced model is 1.6183. We call this RSS(reduced). We thus have, RSS(reduced) = 1.6183. We will use this is our further analysis.

| Anova table | | | | | | |
|---|---|---|---|---|---|---|
| Model | Res. Df. | RSS | Df | Sum of Sq | F-Test | p-value |
| Reduced model | 107 | 1.6183 | | | | |
| Full model | 105 | 1.5634 | 2 | 0.0550 | 1.8452 | 0.1631 |

Table 22: Results from comparing the two regression models with Anova

# 8 Exercise 2

## (a) Residual Sum of Squares

We now consider the case of including the dummy variable special offer in the reduced model obtain in Exercise 1. Before fitting the regression model we want to investigate how the variables are related in the case where the dummy variable special offer is 1 and 0. This can be done by constructing scatter plots of all pairwise combinations of the variables in the data set. Figure **??** show the scatter plots from the given data set, where the dark blue color represent the case without a special offer.

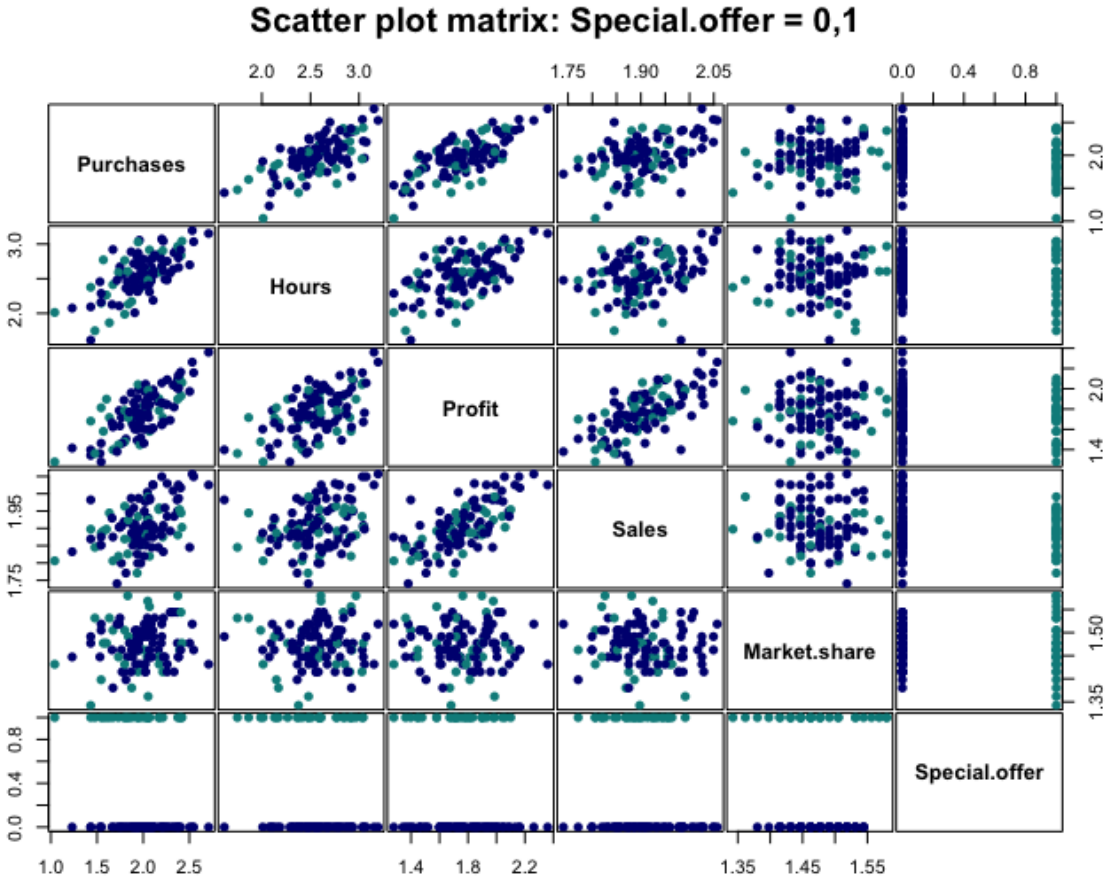


Figure 30: Scatter plot matrix showing the relationship between the variables Z=0 in blue

By including this dummy variable in the reduced model we obtain the following regression.

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 Z + \hat{\beta}_4 Z X_1 + \hat{\beta}_5 Z X_2 \tag{58}$$

where $Z$ is the dummy variable special offer. Inserting for the coefficients gives us the following

regression:

$$\hat{Y} = -1.7524 + 0.4672X_1 + 1.3522X_2 - 0.8418Z - 0.1563ZX_1 + 0.6247ZX_2 \tag{59}$$

The results from the fitted model is presented in Table 23.

| Regression including special offer | | | | |
|---|---|---|---|---|
| Coefficient | Estimate | Std. Error | t-Test | p-value |
| $\beta_0$ | -1.7524 | 0.3833 | -4.57 | 0.0000 |
| $\beta_1$ | 0.4672 | 0.0587 | 7.96 | 0.0000 |
| $\beta_2$ | 1.3522 | 0.2222 | 6.09 | 0.0000 |
| $\beta_3$ | -0.8418 | 0.8857 | -0.95 | 0.3441 |
| $\beta_4$ | -0.1563 | 0.0963 | -1.62 | 0.1075 |
| $\beta_5$ | 0.6247 | 0.4990 | 1.25 | 0.2134 |

Table 23: Results from fitting the regression including special offer

The R-squared value for this regression is 0.6937, which suggest a slightly improvement from both previous models. We can find the residual sum of squares by performing Anova on this regression. The results from the Anova is presented in Table 24.

| Anova table | | | | | |
|---|---|---|---|---|---|
| Variable | Df | Sum Sq | Mean Sq | F-Test | p-value |
| $X_1$ | 1 | 2.81 | 2.81 | 189.54 | 0.0000 |
| $X_2$ | 1 | 0.85 | 0.85 | 57.09 | 0.0000 |
| $Z$ | 1 | 0.03 | 0.03 | 2.18 | 0.1429 |
| $ZX_1$ | 1 | 0.02 | 0.02 | 1.48 | 0.2269 |
| $ZX_2$ | 1 | 0.02 | 0.02 | 1.57 | 0.2134 |
| Residuals | 104 | 1.54 | 0.01 | | |

Table 24: Results from performing Anova on the regression including special offer

From the Table we have that the residual sum of squares is 1.54, or more accurately RSS = 1.5409. We now proceed by testing the regression model excluding special offer on the subset where there is no special offer, $Z = 0$. This gives us the following regression.

$$\hat{Y}_{Z=0} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 \tag{60}$$

Inserting for the coefficients gives us the following regression:

$$\hat{Y}_{Z=0} = -1.7524 + 0.4672X_1 + 1.3522X_2 \tag{61}$$

The results from the fitted model is presented in Table 25.

We can find the residual sum of squares by performing Anova on this regression. The results from the Anova is presented in Table 26.

| Regression on subset Z=0 | | | | |
|---|---|---|---|---|
| Coefficient | Estimate | Std. Error | t-Test | p-value |
| $\beta_0$ | -1.7524 | 0.3727 | -4.70 | 0.0000 |
| $\beta_1$ | 0.4672 | 0.0571 | 8.19 | 0.0000 |
| $\beta_2$ | 1.3522 | 0.2161 | 6.26 | 0.0000 |

Table 25: Results from fitting the regression on subset where special offer is 0

| Anova table | | | | | |
|---|---|---|---|---|---|
| Variable | Df | Sum Sq | Mean Sq | F-Test | p-value |
| $X_1$ | 1 | 2.26 | 2.26 | 161.43 | 0.0000 |
| $X_2$ | 1 | 0.55 | 0.55 | 39.17 | 0.0000 |
| Residuals | 77 | 1.08 | 0.01 | | |

Table 26: Results from performing Anova on the subset where special offer is 0

From the Table we have that the residual sum of squares is 1.08, or more accurately RSS0 = 1.0787. We now proceed by testing the regression model excluding special offer on the subset where there is a special offer, $Z = 1$. This gives us the following regression.

$$\hat{Y}_{Z=1} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 \tag{62}$$

Inserting for the coefficients gives us the following regression:

$$\hat{Y}_{Z=1} = -1.7524 + 0.4672 X_1 + 1.3522 X_2 \tag{63}$$

The results from the fitted model is presented in Table 27.

| Regression on subset Z=0 | | | | |
|---|---|---|---|---|
| Coefficient | Estimate | Std. Error | t-Test | p-value |
| $\beta_0$ | -2.5942 | 0.8583 | -3.02 | 0.0054 |
| $\beta_1$ | 0.3109 | 0.0821 | 3.79 | 0.0008 |
| $\beta_2$ | 1.9768 | 0.4802 | 4.12 | 0.0003 |

Table 27: Results from fitting the regression on subset where special offer is 0

We can find the residual sum of squares by performing Anova on this regression. The results from the Anova is presented in Table 28.

From the Table we have that the residual sum of squares is 0.46, or more accurately RSS1 = 0.4622. We now proceed by testing the regression model including special offer, assuming that the slope parameter is the same but with different intercept parameter. This gives us the following regression.

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 Z \tag{64}$$

| Anova table | | | | | |
|---|---|---|---|---|---|
| Variable | Df | Sum Sq | Mean Sq | F-Test | p-value |
| $X_1$ | 1 | 0.60 | 0.60 | 35.08 | 0.0000 |
| $X_2$ | 1 | 0.29 | 0.29 | 16.95 | 0.0003 |
| Residuals | 27 | 0.46 | 0.02 | | |

Table 28: Results from performing Anova on the subset where special offer is 1

Inserting for the coefficients gives us the following regression:

$$\hat{Y} = -1.7524 + 0.4672X_1 + 1.3522X_2 \tag{65}$$

The results from the fitted model is presented in Table 29.

| Including special offer with no interaction | | | | |
|---|---|---|---|---|
| Coefficient | Estimate | Std. Error | t-Test | p-value |
| $\beta_0$ | -1.9362 | 0.3476 | -5.57 | 0.0000 |
| $\beta_1$ | 0.4108 | 0.0467 | 8.80 | 0.0000 |
| $\beta_2$ | 1.5077 | 0.1982 | 7.61 | 0.0000 |
| $\beta_3$ | 0.0391 | 0.0266 | 1.47 | 0.1448 |

Table 29: Results from fitting the regression including special offer, with no interaction

We can find the residual sum of squares by performing Anova on this regression. The results from the Anova is presented in Table 30.

| Anova table | | | | | |
|---|---|---|---|---|---|
| Variable | Df | Sum Sq | Mean Sq | F-Test | p-value |
| $X_1$ | 1 | 2.81 | 2.81 | 187.69 | 0.0000 |
| $X_2$ | 1 | 0.85 | 0.85 | 56.53 | 0.0000 |
| $Z$ | 1 | 0.03 | 0.03 | 2.16 | 0.1448 |
| Residuals | 106 | 1.59 | 0.01 | | |

Table 30: Results from performing Anova including special offer, with no interaction

From the Table we have that the residual sum of squares is 1.59, or more accurately RSS(dummy) = 1.5860. In Table 31 the residual sum of squares for all the regressions are presented.

| Residual sum of squares | | | | | | |
|---|---|---|---|---|---|---|
| | RSS(reduced) | RSS | RSS0 | RSS1 | RSS0 + RSS1 | RSS(dummy) |
| RSS | 1.6183 | 1.5409 | 1.0787 | 0.4622 | 1.5409 | 1.5860 |

Table 31: Table showing residual sum of squares for all the regression models

## References


Becker, R. A., C. J. M. and Wilks, A. R. (1988). *The New S Language*, Wadsworth & Brooks/Cole.

Chatterjee, S. and Hadi, A. S. (2015). *Regression analysis by example*, John Wiley & Sons.

James, G., Witten, D. and Hastie, T. (2014). *An Introduction to Statistical Learning: With Applications in R.*

Thomas, R. L. (2005). *Using statistics in economics*, McGraw-Hill.

Walpole, R. E., Myers, R. H., Myers, S. L. and Ye, K. (2014). *Probability and statistics for engineers and scientists*, Pearson Education.