Mathematical Statistics 2B

Practical 5: Linear Regression

Department of Statistics - University of Johannesburg

1. Introduction

Multiple linear regression analysis (MLR) is an extension of simple linear regression analysis. It is used to model the linear relationships between a single numerical response variable (dependent variable) and multiple predictor/explanatory variables (independent variables). The independent variables can be numerical of factor variables. For the purpose of this model, we will perform the model checking for MLR analyses consisting of numerical independent variables only (Section 3). In Section 4 we will look at how we compare the simple linear regression lines of two independent groups using MLR analysis. For this section, model checking is beyond the scope of the module.

2. Multiple Linear Regression Analysis

MLR analysis consists of more than just fitting a linear line through data points. Before getting into any of the model investigations, inspect and prepare your data. Check it for errors, treat any missing values, and inspect outliers to determine their validity. After you are comfortable that your dataset is correct, go ahead and apply the following steps:

1) Select the set of independent variables and run the regression.
2) Evaluate model fit.
3) Test model assumptions.
4) Address potential problems with the model.
5) Predict using the final model.

Step 1: Select the set of independent variables and run the regression

To select the right independent variable in MLR, you need to have a good understanding of your data and your problem statement. It makes no sense to include all possible variables in your model. Two criteria are used to achieve the best set of predictors, namely the meaningfulness of the variables to the problem statement and statistical significance of the variables in the model. We want a model that consists of the maximum amount of information from a minimum number of (relevant) variables.

MLR algorithms are built to assist in selecting the best set of independent variables. These include forward selection, backward selection and stepwise selection. You will learn more about this in Statistics 3. For this year, we will select the set of variables that are meaningful as input, fit the model and may remove variables that do not contribute to the model (i.e., insignificant) or are problematic (e.g., highly collinear) to create the final model.

Step 2: Evaluate model fit

We consider a number of different measures to evaluate the model fit, with the most commonly used as follows:

1) Global F-test

   Test the significance of your predictor variables (as a group) for predicting the response of your dependent variable using ANOVA. It tests the hypothesis H_0: there is no linear model vs. H_1: there is a linear model.

2) Residual standard error

   The root mean square error (RMSE) is a way to quantify how far the data points are from the regression line, namely the average amount that the real values of Y differ from the predictions provided by the regression line, using the formula:

   $$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{n}}$$

   Instead of dividing by the sample size *n*, we can divide by the degrees of freedom *df* to obtain an unbiased estimation of the standard deviation of the error term ε, where *df* is the sample size minus the number of parameters in the model (intercept plus number of independent variables). This quantity is called the residual standard error (RSE) which is automatically calculated in R using the formula:

   $$\text{RSE} = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{df}}$$

   A regression model that has a small RSE will have data points that are closely packed around the fitted regression line. The smaller the RSE, the better a regression model fits a dataset, and the higher the RSE, the worse a regression model fits a dataset.

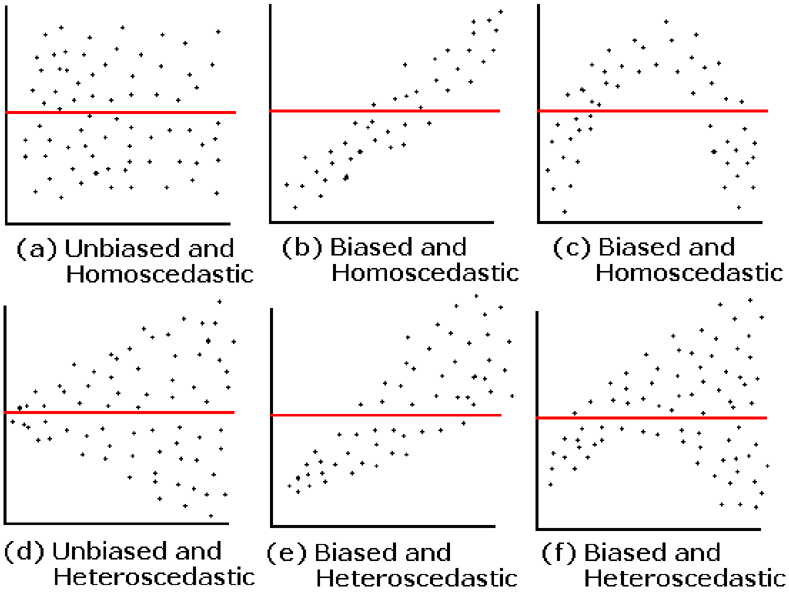
3) R-squared and adjusted R-squared

   R^2 represents the proportion of the variation in the dependent variable that is explained by the independent variables in the MLR model. The adjusted R^2 considers the number of predictors in the model and penalises excessive variables, providing a more accurate measure of the model's goodness-of-fit, especially when the model includes a large number of independent variables. When comparing MLR models, the model with the highest R^2 (or adjusted R^2) value is the better model for the data.

Step 3: Test model assumptions

The linear regression model relies on several assumptions. As part of the modelling process, we need to test that the data and model adhere to these assumptions. In general, we check the regression assumptions by examining the variability left over after we fit the regression line. We simply graph the residuals and look for any unusual patterns. Because we are fitting a linear model, we assume that the relationship really is linear. If a linear model makes sense, the residuals will have a constant variance (assumption of equal variances, also referred to as homoscedasticity or homogeneity of variance), be approximately normally distributed with a mean of zero, and be independent of one another. The most useful graph for analysing residuals is the *residual by predicted* plot. This is a graph of each residual value plotted against the corresponding predicted value.

If the assumptions are met, the residuals will be randomly scattered around the centre line of zero with no obvious pattern. If there is a non-random pattern, the nature of the pattern can pinpoint potential issues with the model. For example, if curvature is present in the residuals, then it is likely that there is curvature in the relationship between the response and the predictors that is not explained by our model, implying that the linear model is not appropriate. If the residuals fan out as the predicted values increase, then we have heteroscedasticity. This means that the variability in the response is changing as the predicted value increases.



(a) Unbiased and Homoscedastic

(b) Biased and Homoscedastic

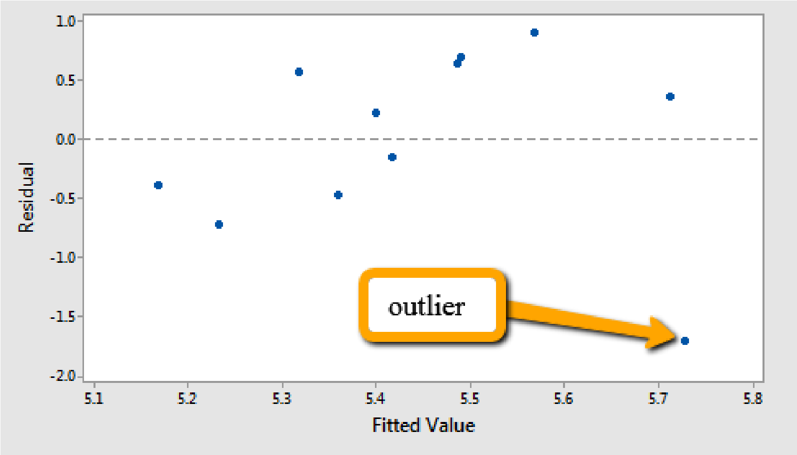
(c) Biased and Homoscedastic

(d) Unbiased and Heteroscedastic

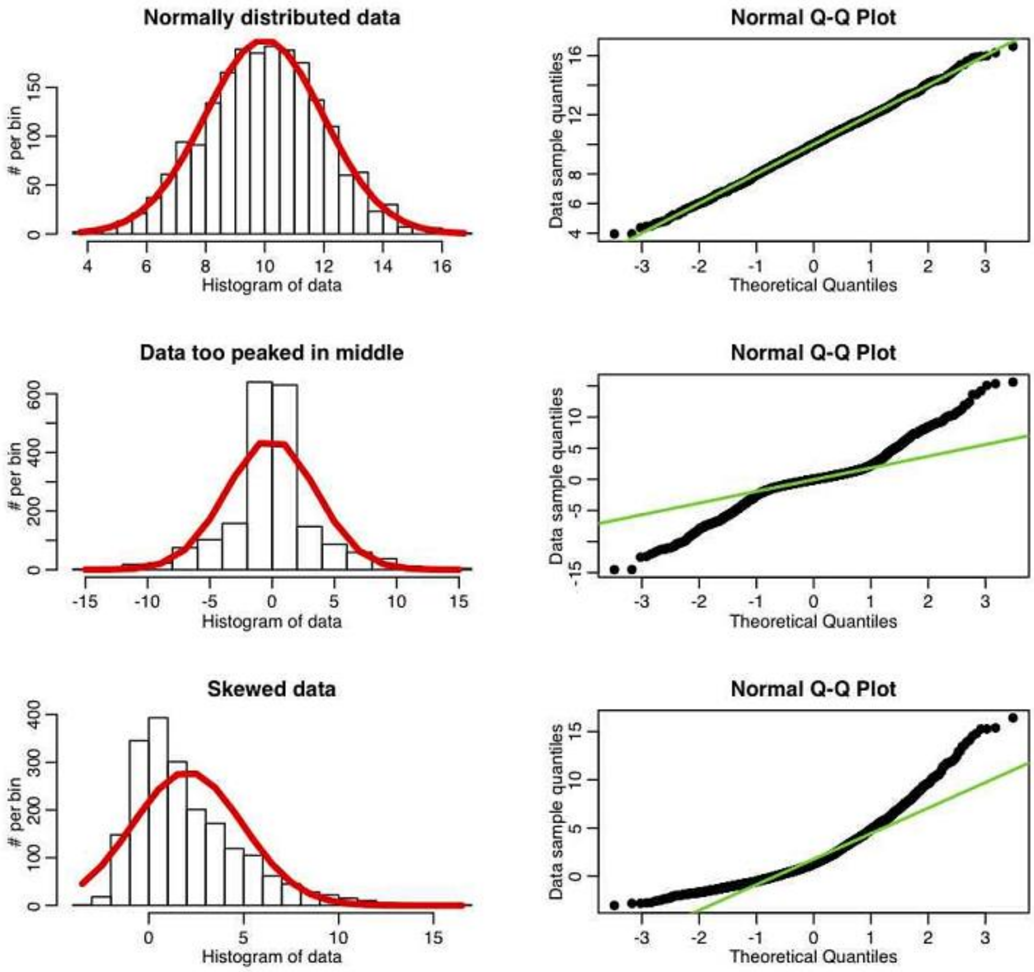
(e) Biased and Heteroscedastic

(f) Biased and Heteroscedastic

An unusual pattern might also be caused by an outlier. Outliers can have a big influence on the fit of the regression line. There are many formal measures to identify outliers in a regression model, but this is beyond the scope of the module.



In addition to the residual versus predicted plot, we use a histogram of residuals and a normal probability plot of residuals to check if our residuals are approximately normally distributed. Note that we check the residuals for normality, not the raw data.



If the independent variables are highly correlated it can cause problems when fitting and interpreting the regression model. This is referred to as multicollinearity. The simplest way to detect multicollinearity is through a correlation matrix of all the independent variables. More formally, the presence of multicollinearity is tested using the variance inflation factor (VIF). The VIF value starts at 1 and has no upper limit. A general rule of thumb for interpreting VIFs is as follows:

- VIF equal to 1 indicates no significant correlation between at least one predictor and the others
- VIF between 1 and 5 indicates a moderate correlation between at least one predictor and the others
- VIF greater than 5 indicates potentially severe correlation between at least one predictor and the others
    - In this case, the coefficient estimates and *p*-values in the regression output are likely unreliable

Step 4: Address potential problems with the model

It is not unusual that at least one of the model assumptions will be violated. In some cases, it is possible to fix or minimize the problem(s) that are in conflict with the assumptions.

- If the assumption of linearity is not met, we can transform the dependent and/or the independent variables, such as the log transform, if it makes sense to do. Another possibility to consider is adding another regressor that is a nonlinear function of one of the other variables.
- If the data are heteroscedastic, the log of the response variable is typically used to remove the heteroscedasticity.
- If there are severe outliers in the data, these have to be removed before fitting the model.
- If there is multicollinearity in the data, we can either remove one or more of the independent variables, or we can transform all the independent variables to uncorrelated variables through multivariate techniques such as Principal Component Analysis (this is beyond the scope of this module).
- If all else fails, perhaps consider an alternative multivariate technique for the data.

Step 5: Predict using the final model

Based on our decisions in Step 4, we will redo the analysis to find the final regression model, which can then be used for prediction. Always remember, we can only predict the dependent variables for valid values of the independent variables. In other words, each value of the independent variables used to predict the dependent variable must be within the valid observed range of the respective independent variable. This is referred to as interpolation. If a value of the independent variable is used for prediction that is outside the observed range of the variable, this is extrapolation and will lead to unreliable predictions.

3. Multiple Linear Regression In R

The *lm()* function is used to fit linear regression models in R. The function has many arguments, but for the purpose of this module we will only use the formula and data arguments. For a multiple linear regression model with three independent variables, the formula argument defines the model as $Y \sim X_1 + X_2 + X_3$ (this equation is adjusted for any number of independent variables). The data argument identifies the dataset. There are four plots as part of the regression output that we will evaluate this year, namely the residuals vs. fitted, normal Q-Q, scale-location, and Cook's distance.

Example

The annual data on advertising, promotions, sales expenses and sales for a sample of companies are given in the *sales.csv* dataset. All variables are measured in millions of dollars. Fit a MLR model to predict sales using the other three variables.

Regression analysis

```
sales_reg=lm(sales~advertising+promotion+expenses,data=sales)
summary(sales_reg)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)     5.9141      1.8445      3.206       0.00489 **
advertising     1.1127      0.6954      1.600       0.12701
promotion       4.5865      0.6455      7.106       1.27e-06 ***
expenses        22.5791     2.0326      11.108      1.73e-09 ***

Residual standard error: 1.292 on 18 degrees of freedom
Multiple R-squared:  0.9104,  Adjusted R-squared:  0.8954
F-statistic: 60.93 on 3 and 18 DF,  p-value: 1.264e-09
```

Based on the F-statistic, the model is highly significant. Based on the RSE we can say that the model accurately predicts sales with about 1.292 error, on average. Based on the R^2, approximately 91% of the variation in sales is explained by the variation in the three independent variables in the model. All this indicates a good model fit.

Check multicollinearity

cor(sales)

```
             sales        advertising  promotion    expenses
sales        1.0000000    -0.1402891   0.54185188   0.80956373
advertising  -0.1402891   1.0000000    -0.34021309  -0.10058307
promotion    0.5418519    -0.3402131   1.00000000   0.06258872
expenses     0.8095637    -0.1005831   0.06258872   1.00000000
```

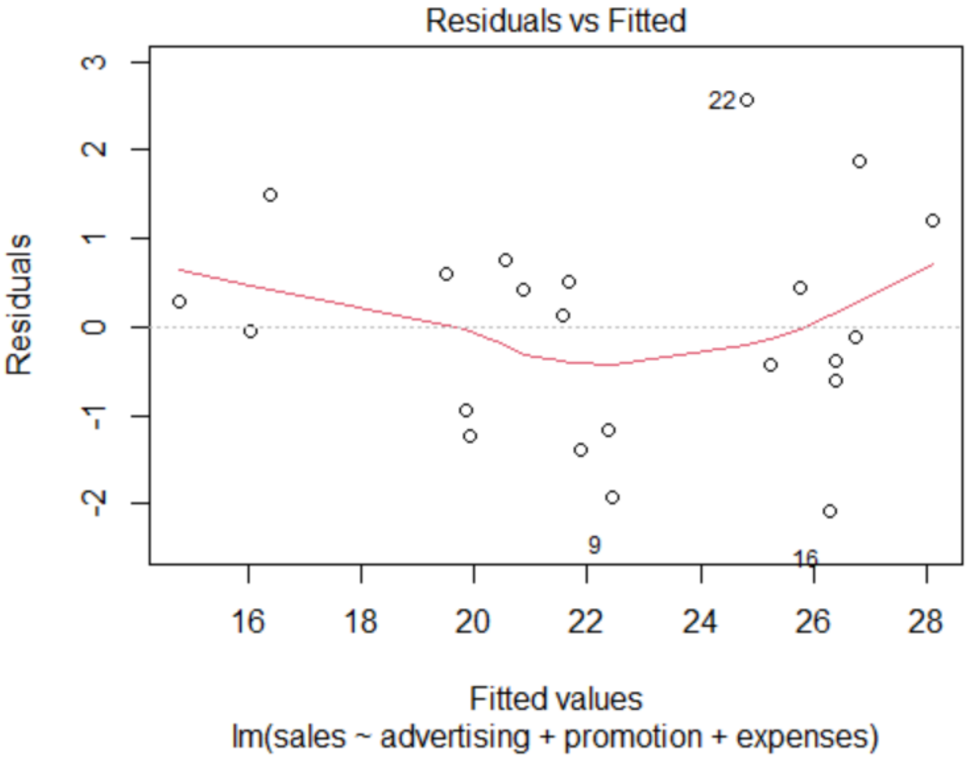
The VIF ranges between 1.01 and 1.14 (note, I will always provide the VIF if I want you to interpret it).

The three independent variables are not highly correlated, and the VIF values are relatively low. Therefore, there is no multicollinearity in the data.

Residual plot (residual vs. predicted)

This plot shows if residuals have nonlinear patterns. If the residuals are equally spread around a horizontal line without a distinct pattern, it is a good indication that there is a linear relationship.

> plot(sales_reg,1)

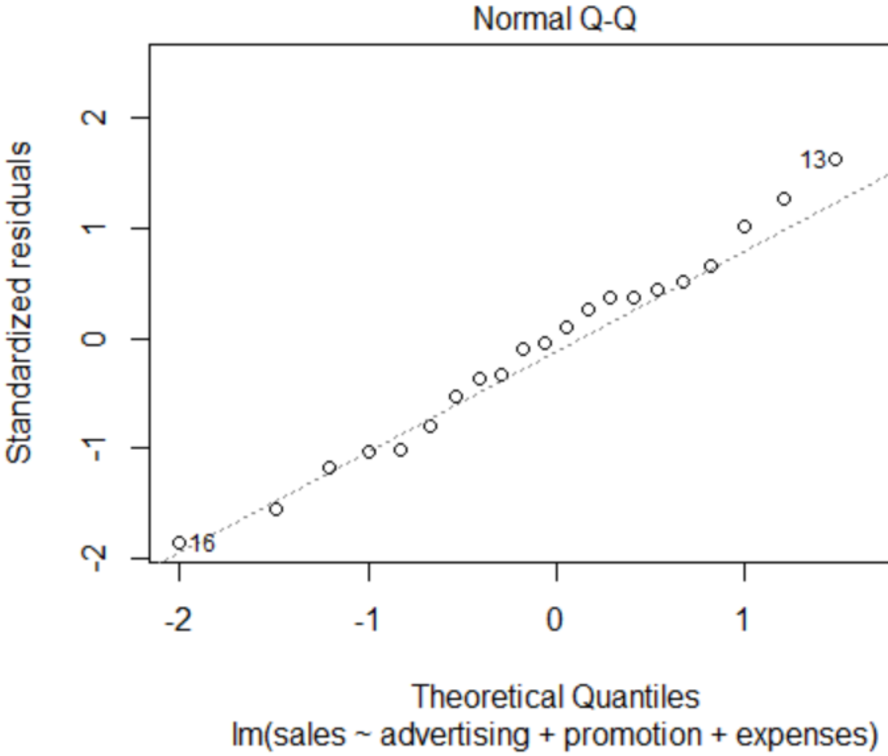


There appears to be some pattern in the residual plot indicating possible nonlinearity in the data. This may also be as a result of the possible outliers (#9, #16, #22). In general the graph seems reasonable.

Normal Q-Q

This plot shows if residuals are normally distributed. If the residuals follow a straight line, the assumption of normality of the residuals are met.

```
> plot(sales_reg,2)
```



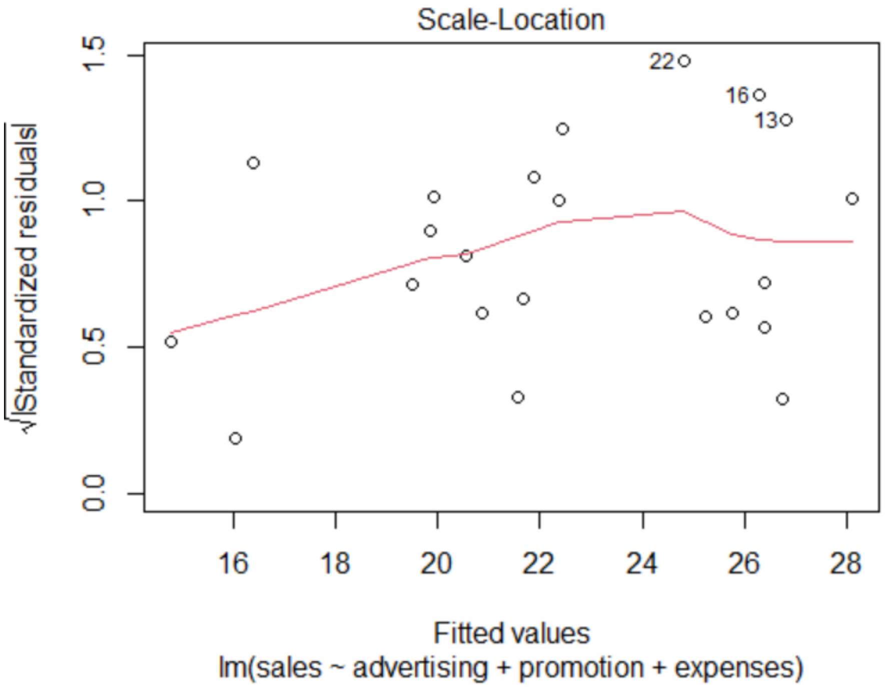
The Q-Q plot appears to show approximate normality, although observation #13 and #22 seem a little off, and could be outliers.

Scale-location

This plot shows if residuals are spread equally along the ranges of predictors. The plot, also called a spread-location plot, is used to check the assumption of equal variance (homoscedasticity). It is good if you see a horizontal line with equally (randomly) spread points.

From the graph on the next page, the residuals appear somewhat randomly spread but is still affected by the potential outliers (#13, #16, #22).

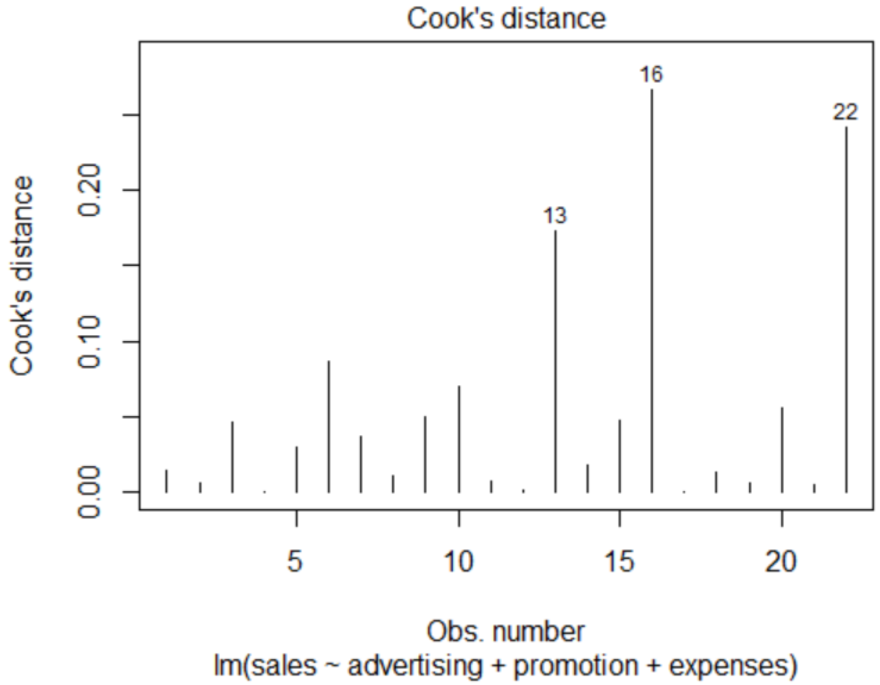
```
> plot(sales_reg,3)
```



Cook’s distance

This plot identifies potential outliers.

```
> plot(sales_reg,4)
```



The plot shows that observation 13, 16 and 22 are outliers. We could remove outliers, remove insignificant variables, or transform variables to try to improve the model. However, this is a very small dataset (*n* = 22) and we do not see much improvement when we do any of the above. For regression analysis, we use a rule of thumb of 10 observations for each independent variable. More data is better.

Example

Import the *marks.csv* file, consisting of the marks (in percentage) of 50 students for the two semester tests and the exam. Check the multicollinearity in the data using the correlation matrix and the VIF value, which is 9.795 for each of the two test marks.

```
#Create a correlation matrix
cor(marks)
              Test1        Test2        Exam
Test1         1.0000000    0.9475803    0.7593029
Test2         0.9475803    1.0000000    0.5872121
Exam          0.7593029    0.5872121    1.0000000
```

The two test mark variables are very highly correlated. Both test mark variables have a moderate to strong, positive relationship with exam mark. The VIF value of 9.795 indicates the presence of multicollinearity.

```
#Fit a MLR model to predict exam marks using the marks of the two tests (selected output are shown here)
test_reg=lm(Exam~Test1+Test2,data=marks)
summary(test_reg)
Coefficients:
              Estimate     Std. Error   t value      Pr(>|t|)
(Intercept)   55.87020     1.11593      50.066       < 2e-16 ***
Test1         0.39673      0.04576      8.670        2.59e-11 ***
Test2         -0.27323     0.04833      -5.654       8.96e-07 ***

Residual standard error: 0.4769 on 47 degrees of freedom
Multiple R-squared:  0.748,    Adjusted R-squared:  0.7372
F-statistic: 69.74 on 2 and 47 DF,  p-value: 8.603e-15
```

Based on the F-statistic, the model is significant. Based on the RSE we can say that the two semester test marks accurately predicts exam mark with about 0.48% error, on average. Based on the R^2, approximately 75% of the variation in exam marks is explained by the variation in the semester test marks in the model. All this indicates a good model fit. However, the estimate for test 2 is negative, implying that it has a negative correlation with exam marks, which is not what we have seen in the correlation matrix. This is due to the high multicollinearity in the data, which distorts the interpretation of the regression model.

4. Comparing Two Simple Linear Regression Lines

MLR has so far been used for the case where all the variables in the equation are continuous. However, the MLR model can include categorical predictor variables. In such cases, the model essentially develops a regression model for each of the levels of the categorical variable and allows us to compare the lines. We will restrict our analysis to a single numerical dependent variable, one numerical independent variable, and one binary independent variable, with values 0 and 1. The result is two simple linear regression lines, one for each level of the binary independent variable. We can then compare the slopes and intercepts of the two lines to see if the relationships are different for the two groups.

Consider a dataset consisting of a dependent variable *Y*, a numerical independent variable *X* and a binary independent variable IND. We will model the following MLR equation:

$$Y = \beta_0 + \beta_1 IND + \beta_2 X + \beta_3 \left(IND \times X \right) + \varepsilon$$

The estimated model is therefore:

$$\hat{Y} = b_0 + b_1 IND + b_2 X + b_3 \left(IND \times X \right)$$

If IND = 0, this equation reduces to $Y = \beta_0 + \beta_2 X + \varepsilon$, estimated as $\hat{Y} = b_0 + b_2 X$.

If IND = 1, this equation reduces to $Y = \left(\beta_0 + \beta_1 \right) + \left(\beta_2 + \beta_3 \right) X + \varepsilon$, estimated as $\hat{Y} = \left(b_0 + b_1 \right) + \left(b_2 + b_3 \right) X$.

To compare the two lines, we test whether the slopes and intercepts are different from one another by looking at the *p*-values of the parameter estimates, testing at the α% level of significance.

- If $\beta_1 = 0$ (*p*-value > α) → intercepts are equal, i.e., the two lines cut the *y*-axis at the same value.
- If $\beta_1 \neq 0$ (*p*-value < α) → intercepts are not equal, i.e., the two lines cut the *y*-axis at different values.
- If $\beta_3 = 0$ (*p*-value > α) → slopes are equal, i.e., the two lines are parallel.
- If $\beta_3 \neq 0$ (*p*-value < α) → slopes are not equal, i.e., the two lines are not parallel.

In R, we can either define the binary variable as a factor, or we can create the product between the *X* variable and the binary indicator variable, and define the model above with three independent variables in the *lm*() function.

Example

An engineer wants to test if the relationship between the effective life of a tool (Y) and its lathe speed (X) is the same or different for two tool types (Type A coded as 0 and Type B coded as 1). The R code is as follows:

```
# Import the tools.csv data
tools=read.csv("tools.csv",header = T)
```

```
 Create a new variable X times indicator, and attach it to the dataset under a new name
type_x=tools$x*tools$type
tools2=cbind(tools,type_x)
```

```
# Run the regression analysis
tools_reg=lm(y~type+x+type_x,data=tools2)
summary(tools_reg)
```

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	32.785331	4.527527	7.241	1.97e-06 ***	(this is b_0)
type	22.890349	6.614200	3.461	0.00322 **	(this is b_1)
x	-0.020982	0.005935	-3.535	0.00275 **	(this is b_2)
type_x	-0.010652	0.008640	-1.233	0.23542	(this is b_3)

For tool type A (type = 0): $\hat{y} = b_0 + b_2 x$

For tool type B (type = 1): $\hat{y} = \left(b_0 + b_1\right) + \left(b_2 + b_3\right) x$

- Since b_1 is significantly different from 0, the two intercepts of the two lines are different.
- Since b_3 is not significantly different from 0, the two slopes of the two lines are not different.
- Therefore, the relationship between the tool life and lathe speed is the same, but on average the tool life of one of the tools is different from the tool life of the other tool.

If you are interested in creating the graph of the two different lines, you can look at the R code on the next page. The graph is also shown. I do not expect you to be able to produce this graph in a test this year.

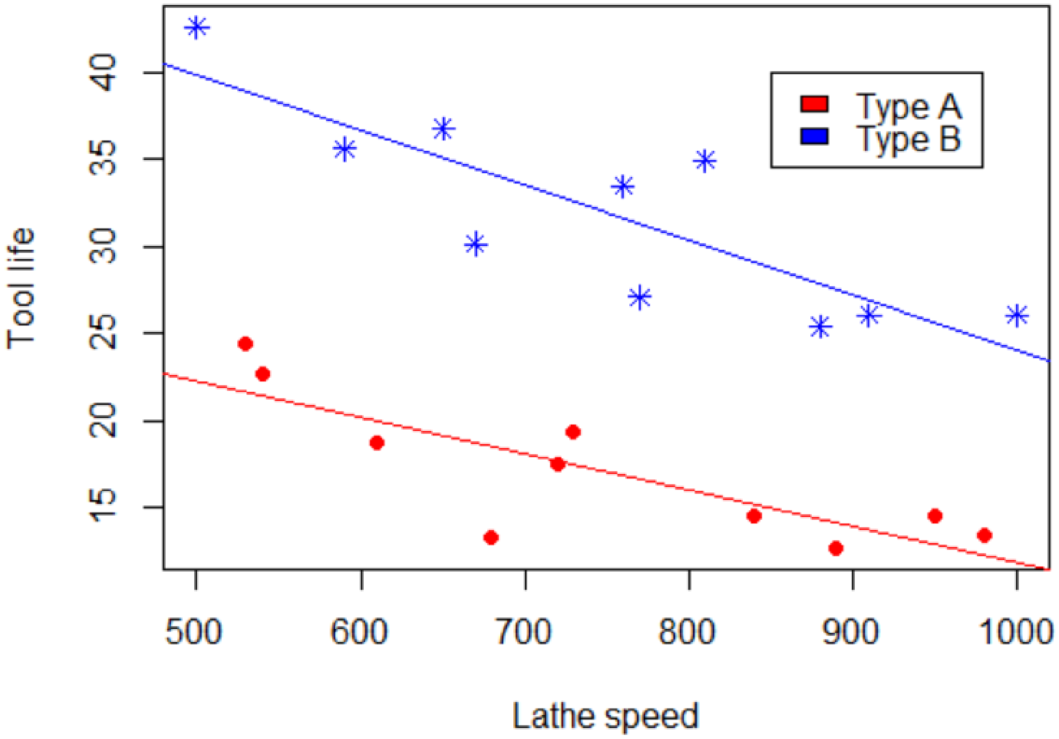
```
#Plot data for tool type A in red – specify the min and max axis values to be able to plot any value in the data
plot(x=tools$x[tools$type==0],y=tools$y[tools$type==0],pch=16,col="red",
    xlim=c(min(tools$x),max(tools$x)),ylim=c(min(tools$y),max(tools$y)),
    xlab="Lathe speed",ylab="Tool life")

#Add the regression line for tool A from the regression output in red
abline(a=tools_reg$coefficients[1],b=tools_reg$coefficients[3],col="red")

#Add the data for tool type B in blue to the same plot
points(x=tools$x[tools$type==1],y=tools$y[tools$type==1],pch=8,col="blue")’

#Add the regression line for tool B from the regression output in blue
abline(a=(tools_reg$coefficients[1]+tools_reg$coefficients[2]),
      b=(tools_reg$coefficients[3]+tools_reg$coefficients[4]),col="blue")

#Add a legend to the plot at the coordinate (850, 40) to identify the data for the two tool types
legend(850, 40, legend=c("Type A", "Type B"),fill=c("red", "blue"))
```



Tool type B seems to have a longer tool life than tool type A, on average. The lathe speed has the same impact on tool life, i.e., tool life seems to deteriorate at the same rate for increased lathe speed.

Exercises

1. The *mydata.csv* datafile consists of one dependent variable and three independent variables, all numerical.
   a) Create a correlation matrix and discuss all pairwise linear relationships.
   b) Fit a multiple linear regression model to predict *Y* using the three *X* variables.
   c) Plots the residual, normal and outlier graphs.
   d) Interpret the regression results.
   e) What could you do to improve the results?

2. A researcher collected air quality data, expressed as an air pollution index (*api*), wind velocity (*wind*) in km/h, temperature (*temp*) in °F and humidity (*humid*) in percentage, in a large city for 20 days during summer. The data are given in the *API.csv* dataset.
   a) Create a correlation matrix and discuss all pairwise linear relationships.
   b) Develop a MLR model to predict the air quality in a large city during the summer months, using wind velocity, temperature and percentage humidity.
   c) Plots the residual, normal and outlier graphs.
   d) Interpret the regression results.
   e) Predict the API value when wind velocity is equal to 10 km/h, temperature is equal to 95°F, and the percentage humidity is equal to 76%.
   f) What could you do to improve the results?

3. A nutrition expert wanted to investigate the impact of protein content in diet on the relationship between age and height of children. She collected the ages (in years) and height (in centimetres) for a sample of children with protein-rich (indicator DIET = 0) and protein-poor (indicator DIET = 1) diets, respectively. Create the interaction term for diet and age (diet × age). The data are given in the *diet.csv* file.
   a) Fit a MLR model where *Y* = height, and compare the slope and intercept of the regression line for the protein-rich diet with that of the protein-poor diet.
   b) Comment on the similarities and differences.