title: “Day 2 Exercises” author: “Dr Tania Prvan” date: “4 July 2019” output: html_document

EXERCISE 1

Estimating volume of trees

(from The Theory of Linear Models by Bent Jorgensen (1993))

In forestry there is an evident need to predict the volume of timber in a given area of forest. It is not easy to measure the volume of a standing tree directly. The best, or at least easiest, method for estimating a tree’s volume is to measure its diameter and height, and calculate an estimate of the volume based on these measurements. The file **TREES.csv** contains a sample of measurements for black cherry trees in Allegheny National Forest, Pennsylvannia (from Ryan, Joiner and Ryan (1985) pp. 328-329).

The following measurements were made; diameter in inches at 4.5 feet above ground level (diameter), height in feet (height) and volume in cubic feet (volume). To convert to metric, height is multiplied by 0.3048 and diameter by 0.0254 to obtain values in metres, and Volume multiplied by to obtain cubic metres. The metric quantities are denoted by d, h and v.

1.1 We can get some data on the diameter, height and volume of black cherry trees from the file **TREES.csv** in the usual location. Since the data are from America, the measurements are in inches, feet and cubic feet. What effect will this have on the relationships between the variables? Use appropriate graphs (boxplots or histograms) to investigate each variable individually. Do there seem to be any outliers in any variables?

1.2 Obtain a matrix plot of the variables. Obtain the correlations between the three variables. Describe the nature of the associations. Make sure to refer to the matrix plot as well as correlations.

1.3 Obtain the least squares regression line of volume on diameter. Obtain the diagnostic plots. Write down the equation of the regression line and interpret the value of Adjusted R Squared. Do the regression coefficients have any physical interpretation in this situation? If so, what interpretation?

1.4 Look at the plot of theresiduals versus predicted values (you produced it in 1.3). What does the pattern of the residuals show? Do the residuals seem to follow a normal distribution?

1.5 Obtain the regression of volume on height and write down the details. If you wanted to estimate volumes of cherry trees, would you prefer to use diameter or height as a predictor? Give reasons for your answer.

1.6 Obtain the regression of volume on diameter and height. Write down the details. Check the model assumptions. Is this model better than that you chose in **1.5**?

EXERCISE 2 Redoing lecture example for yourself.

As part of a test of solar thermal energy, the total heat flux from homes is measured. Researchers wish to examine whether total heat flux (Heatflux) can be predicted by insulation (Insulation), by the position of the focal points in the east (East), south (South), and north (North) directions, and by the time of day (Time). The data are from the book by D.C. Montgomery and E.A. Peck, published by John Wiley & Sons in 1982, titled “Introduction to Linear Regression Analysis”. You can find this data in the file **heatflux.csv**.

2.1 Perform a regression of Heatflux on Insulation, East, South, North, and Time. Write down the equation of the model fitted to the data. Check that the model assumptions hold.

(HINT: For model assumptions to hold the Residual (standardized) versus Fitted Values plot should look like a random scatter about zero and the Normal Probability Plot of the Residuals (standardized) should be approximately linear)

2.2 Test whether the regression is significant. Make sure to write down your hypotheses.

2.3 Write down the adjusted R^2 and comment.

2.4 Obtain 95% confidence intervals for the parameters. Comment on each confidence interval.

2.5 Use the sequential sum of squares to suggest an appropriate model. Write this best model down.

2.6 Use the partial sum of squares to suggest an appropriate model. Reproduce the relevant part of the R output needed. Write this best model down.

2.7 Use forward stepwise regression to obtain the best model. Write this model down.

2.8 Use backward stepwise regression to obtain the best model. Write this best model down.

2.9 Use stepwise regression to obtain the best model. Write this best model down.

2.10 Obtain the best model using Mallow's C_p . Explain why you have chosen this model. Obtain the best model using Adjusted R^2. Explain why you have chosen this model. Obtain the best model using s. Explain why you have chosen this model.

EXERCISE 3

Good Cholesterol

High-density lipoproteins (HDLs) are lipoprotein complexes which are often referred to as the "good cholesterol" because they take cholesterol from peripheral tissues back to the liver and assist in lowering the total serum cholesterol. The data we will be looking at is from Applied Regression and Other Multivariate methods by Kleinbaum, Kupper, Muller and Nizam published by Duxbury Press in 1998.

An experiment involved a quantitative analysis of factors found in high-density lipoprotein (HDL) in a sample of human blood serum. Three variables thought to be predictive of or associated with HDL measurements (Y) were the total cholesterol (X1) and total triglyceride (X2) concentrations in the sample, plus the presence or absence of a certain sticky component called sinking pre-beta, or SPB (X3).

The file **HDL.csv** contains the data.

3.1 Obtain a matrix plot of the four variables. Comment briefly.

3.2 Test whether X1, X2 or X3 alone significantly helps in predicting Y. That is; regress Y on X1. Write down the regression equation and relevant summary statistics. Check the diagnostic plots and write a comment in the space below. Then regress Y on X2. Write down the regression equation and relevant summary statistics. Check the diagnostic plots and write a comment in the space below. Then regress Y on X3. Write down the regression equation and relevant summary statistics. Check the diagnostic plots and comment on them. What is your conclusion? How did you reach it?

3.3 Test whether X1, X2 and X3 taken together significantly help in predicting Y. Make sure to check your model assumptions and write down your conclusions. Also write down the regression equation and relevant summary statistics.

3.4 Informally test whether there is a significant interaction between X1 and X3. To do this first unstack the data according to X3. You can do this outside R. Regress Y on X1 for X3=0. Write down the regression equation. Regress Y on X1 for X3=1. Write down the regression equation. Look at appropriate graphical displays. What is your conclusion regarding interaction between X1 and X3?

3.5 Formally test whether there is a significant interaction between X2 and X3.

3.6 Assume now that the predictor of interest is X3, and that no interaction occurs. Determine whether either X1, X2 or both need to be included in the model because of confounding. To do this fit the following models in the table below to the data. Then fill in the table below. Write down your conclusion.

Model	p-value for X3 term
$X3$	
$X3, X1$	
$X3, X2$	
$X3, X1, X2$	

EXERCISE 4

Fuel Consumption

The data set **Fuel.csv** was used in **Example 2.1** from the book *Applied Linear Regression SECOND EDITION* by Sanford Weisberg (1985). The file consists of six columns of data for each of six contiguous US states, of the following values:

POP = 1971 popn, in thousands

TAX = 1972 motor fuel

NLIC = 1971 thousands of licensed drivers

INC = 1972 per capita income in thousands of dollars

ROAD = 1971 thousands of miles of federal-aid primary highways

FUELC = 1972 fuel consumption

As noted by Weisberg (1985) that before "beginning the analysis it is useful if we can combine or rescale the variables." As he points out the variables "NLIC and FUELC are measured for whole states and will vary with state size, while INC is measured per individual so it is not sensitive to state size."

The following variables are considered

X1=tax

X2=DLIC=100XNLIC/POP=% of popn with driver's licenses

X3=INC, average income ($'000s)

X4=ROAD ('000s miles)

Y=FUEL=1000*FUELC/POP=motor fuel consumption (gallons per person).

4.1 Fit a linear regression of Y on X1, X2, X3, and X4. Obtain the usual diagnostic plots. Write down the regression equation and comment on the plots.

4.2 Identify any possible outliers. The routine **lm** provides diagnostic plots. Look at the last plot. PUT SOME EXPLANATION IN ABOUT THIS PLOT

4.3 Identify any data points that may have high leverage. Do do this . . . .

4.4 Identify any points that are influential. Do do this . . .

EXERCISE 5

Pricing the C's of Diamond Stones

A diamond's beauty, rarity and price depends upon cut, clarity, carats, and colour. These attributes are called the 4Cs and are used throughout the world to classify the rarity of diamonds. Diamonds with the highest 4C ratings are rare and hence more expensive. Carat refers to the weight of the diamond, one carat equals 200 milligrams. Clarity refers to the presence of inclusions in a diamond (inclusions are natural identifying characteristics or fractures appearing when diamonds are first formed). Clarity is coded as IF ("internally flawless"), VVS1 ("very very slightly included 1"), VSS2 ("very very slightly included 2"), VS1 ("very slightly included 1"), or VS2 ("very slightly included 2") which are in descending order. Colour refers to the degree to which a diamond is colourless (grading goes from D which denotes colourless to Z). Cut refers to the angles and proportions of a diamond and also refers to shape (round, square, pear, heart etc.). We want to look at the relative pricing of caratage and the different grades of clarity and colour.

The data come an advertisement in Singapore's Business Times edition of February 18, 2000 and can be found in the file **diamonds.csv**. Only round diamonds have been considered because this is the most popular cut. Three certification bodies were mentioned in the advertisement: New York based Gemological Institute of America (GIA), Antwerp based International Gemological Institute (IGI) and Hoge Raad Voor Diamant (HRD).

This lab has been modified from Pricing the C's of Diamond Stones by Singfat Chu published in the Journal of Statistics Education, Volume 9, Number 2 (2001). Additional information on diamonds came from the website www.adiamondisforvever.com.

We will be fitting General Linear Models (GLMs). These allow us to have categorical predictors with more than 2 levels. When we have a categorical predictor with 2 levels we can treat it as continuous which was done in the workshop examples.

5.1 Plot Price against Carat. Comment.

5.2 Plot ln(Price) against Carat. Comment. (lnPrice<-ln(Price))

5.3 Why are the variables Clarity and Colour Ordinal?

5.4 Fit a linear regression model to ln_price with predictors Carat, Clarity, Colour, and Certification. Explain why Clarity, Colour and Certification are factors. Use Clarity grade VS2 as the referent category, Colour I as the referent category and Certification IGI as the referent category when fitting the linear regression model to the data. Obtain some of the diagnostic residual plots discussed today. Check the model assumptions. Does your model fit well? Support your answer.

EXERCISE 6

We can also consider transforming the response variable when:

The relationship between y and x is not linear. For example, species abundance models are common in biology. Ecologists usually model the number of species of a genus (y) in a fixed area (e.g. island) as:

$$y = \alpha x^{\beta}$$

where x is the surface area. Taking logs of both sides, we have

$$\log y = \log \alpha + \beta \log x$$

which is a linear relationship between $\log y$ and $\log x$.

Heteroscedasticity: i.e. non-constant variance of the errors ε_i, violating the model assumption $var(\varepsilon_i) = \sigma^2$ (homogeneity of variance or homoscedasticity) This will be picked up by residual plots exhibiting fanning or funnelling.

When the homogeneity of variance assumption is violated, we usually also find that the residuals are non-normal. Fortunately transformations which remedy the nonconstancy of variance generally also lead to normally distributed residuals.

We usually need a transformation when the distribution of either y or x is highly skewed.

Ladder of transformations

We can think of all transformations of the response variable as being of the form Y^{λ}. We then have the following ladder of transformations -

λ	Y^{λ}	Description
-1	$\frac{1}{Y}$	Inverse or Reciprocal
-0.5	$\frac{1}{\sqrt{\lambda}}$	Inverse square root
0	$\log Y$	Logarithmic
0.5	$\sqrt{Y}$	Square root
0	Y	*None*

The least severe transformation is the the square root, and the most severe is the inverse.

Other values of λ can be used.

Note that applying a logarithmic transformation to y and x implies a multiplicative, rather than additive, relationship.

Example: The species-area problem

Case (1975) analysed the number of species of lizards on 24 islands in the Gulf of California. Each observation was an island; the response variable of interest was the number of lizard species on the island; and explanatory variables were area of island, nearest distance to mainland, maximum elevation, perennial plant species and plant volume diversity. The aim of the study was to explain variation in the number of lizard species as a function of the explanatory variables.

We will only consider the species-area relationship. This data can be found in **Species.csv**.

6.1 Obtain a plot of Species against Area. Comment.

6.2 Obtain a histogram of Species and a separate histogram of Area. Comment.

6.3 Plot the square root of Species against the square root of Area. Comment.

6.4 Obtain a histogram of square root of Species and a separate histogram of square root Area. Comment.

6.5 Plot the logarithm of Species against logarithm of Area. Comment.

6.6 Obtain a histogram of logarithm of Species and a separate histogram of logarithm Area. Comment.

Normality of the covariates is not a model assumption; however we usually find that highly skewed x's lead to other model assumptions (such as homogeneity of variance) being violated. It is therefore usually beneficial to find a transformation which normalizes the covariates, and this transformation will often also linearize the relationship between y and x.

6.7 Regress logarithm of Species on logarithm of Area. Assess how well the model fits.