**STAT210/410**

**Review of Simple Linear Regression**

You may wish to refer to the text &/or the Powerpoint slides for Chapter 3.

One of the more challenging problems confronting the water pollution control field is presented by the tanning industry. Tannery wastes are chemically complex. They are characterised by high values of biochemical oxygen demand, volatile solids and other pollution measures. The experimental data were obtained from 33 samples of chemically treated waste. Readings on SRP, the percent reduction in total solids, and ODP, the percent reduction in chemical oxygen demand for the 33 samples were recorded.

Source: Walpole R.E, and Myers R.H., (1989), *Probability and Statistics for Engineers and Scientists*, 4th ed., Macmillan, New York, page 359.

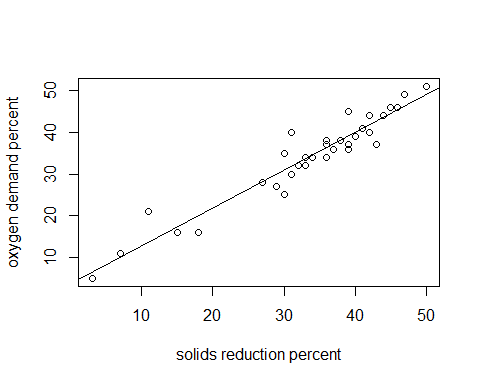
The data have been analysed and the R code and resultant output are given below.

* Work through the code and output, ensuring that you understand the process and the results. Ask for help as required.
* Answer the questions that are embedded in the output below.

1. Read in data and produce exploratory plot

options(digits=3,show.signif.stars=F) *# set no. of significant figures*

# read in data

dat1 <- read.table("CH03\_SLREx.txt",header=T)  
  
# fit simple linear regression model  
xy.lm <- lm(ODP~SRP, data=dat1)  
  
#plot data and fitted line  
plot(ODP~SRP,data=dat1, ylab = "oxygen demand percent", xlab="solids reduction percent")  
abline(xy.lm)

*# alternative plot using package ggplot2*

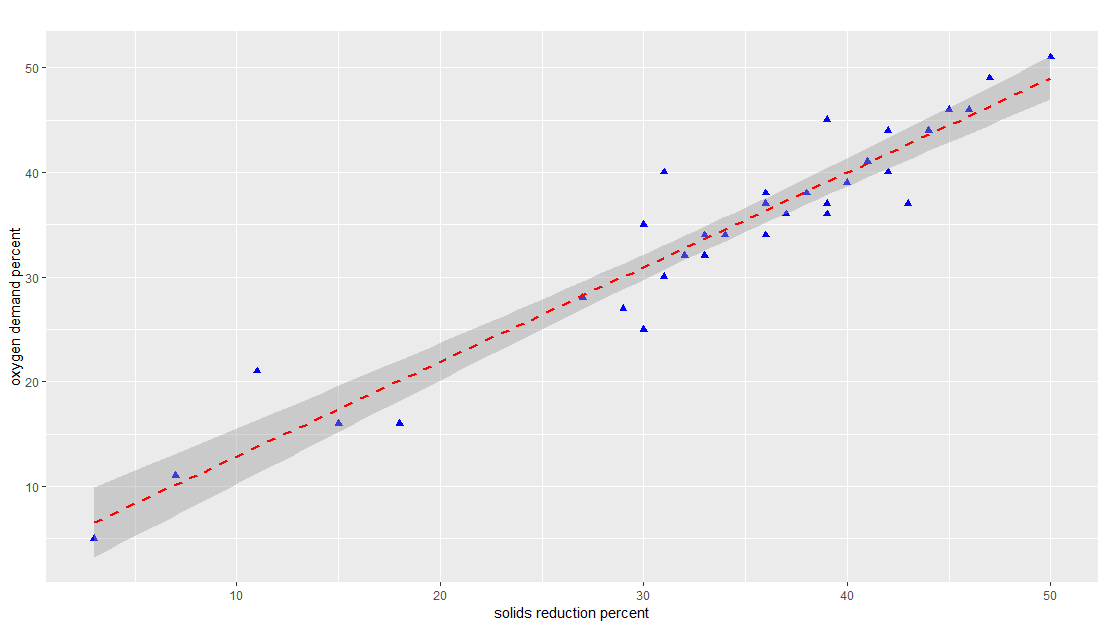
library(ggplot2)

ggplot(data=dat1, aes(x=SRP, y=ODP)) +

geom\_point(pch=17, color="blue", size=2) +

geom\_smooth(method="lm", color="red", linetype=2) +

labs(title="", x="SDP", y="ODP")



**Q1:** Describe the association between the two variables, stating the form, direction and strength of the association.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_­­­­­­­­­­­­­­­­\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

1. **Simple linear regression analysis to obtain estimates of regression coefficients, their std. errors and 95% CIs**

# Table of regression coefficients and analysis of variance table  
print(summary(xy.lm))

##   
## Call:  
## lm(formula = ODP ~ SRP, data = dat1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5.939 -1.783 -0.228 1.506 8.157   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 3.8296 1.7684 2.17 0.038  
## SRP 0.9036 0.0501 18.03 <2e-16  
##   
## Residual standard error: 3.23 on 31 degrees of freedom  
## Multiple R-squared: 0.913, Adjusted R-squared: 0.91   
## F-statistic: 325 on 1 and 31 DF, p-value: <2e-16

print(anova(xy.lm))

## Analysis of Variance Table  
##   
## Response: ODP  
## Df Sum Sq Mean Sq F value Pr(>F)  
## SRP 1 3391 3391 325 <2e-16  
## Residuals 31 323 10

#CI for regression coefficients  
confint(xy.lm)

## 2.5 % 97.5 %  
## 0.223 7.44  
## 0.801 1.01

**With reference to the output above:**

**Q2:** Give the equation of the line of best fit: ­­­­­­­­­­­­­­­­­­\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Q3:** Is SRP a useful predictor of ODP? Justify your response. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Q4:** What is the value of the correlation coefficient, r?

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Q5:** State and interpret the value of the coefficient of determination, r2.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Q6**: What is the estimate of σ2? Find this value from two different parts of the output.

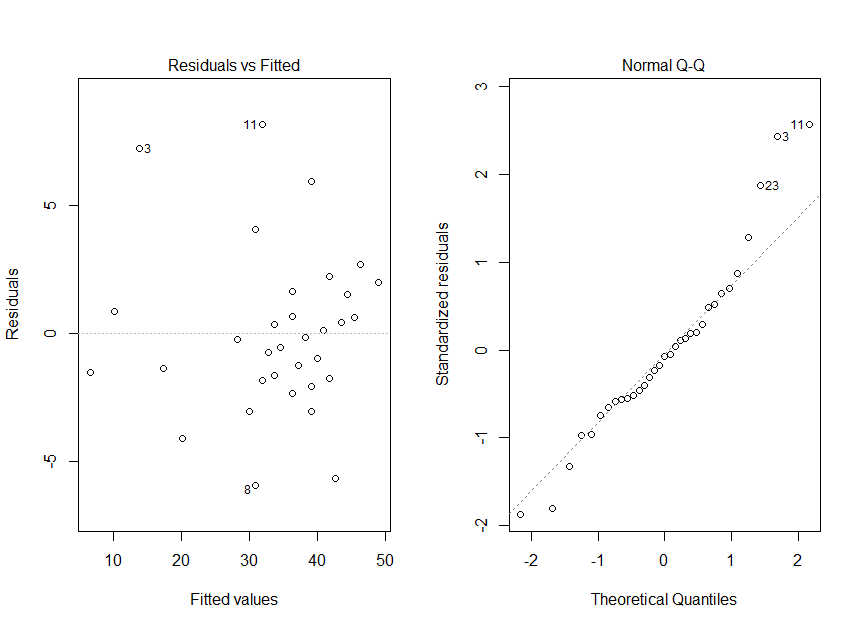
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Q7:** Give an informative interpretation of the 95% CI for β1.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

1. **Check Model Assumptions**

#diagnostic plots  
par(mfrow=c(1,2)) *# 2 plots to the page (1 row, 2 columns)*plot(xy.lm,which=1:2,add.smooth=F)

#Shapiro Wilk's Test of normality  
print(shapiro.test(xy.lm$residuals))

##   
## Shapiro-Wilk normality test  
##   
## data: xy.lm$residuals  
## W = 1, p-value = 0.1

**Q8:** State the assumptions of the linear model.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**With reference to the output above:**

**Q9:** Summarise the information available from the two residuals plots.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Q10**: What does the Shapiro-Wilk’s test imply?

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

1. **Predicting from the model**

# # predictions for 8 data points: SRP = 10,20,30,40,50,60,70,80  
# create new data frame of 8 observations

pred.df <- data.frame(SRP=c(10,20,30,40,50,60,70,80))  
CI <- predict(xy.lm,interval="confidence",newdata=pred.df,level=0.95)  
PI<- predict(xy.lm,interval="predict",newdata=pred.df,level=0.95)

with(pred.df, cbind(SRP, CI, PI))

SRP fit lwr upr fit lwr upr

10 12.9 10.2 15.5 12.9 5.76 20.0

20 21.9 20.1 23.7 21.9 15.08 28.7

30 30.9 29.7 32.1 30.9 24.24 37.6

40 40.0 38.6 41.3 40.0 33.26 46.7

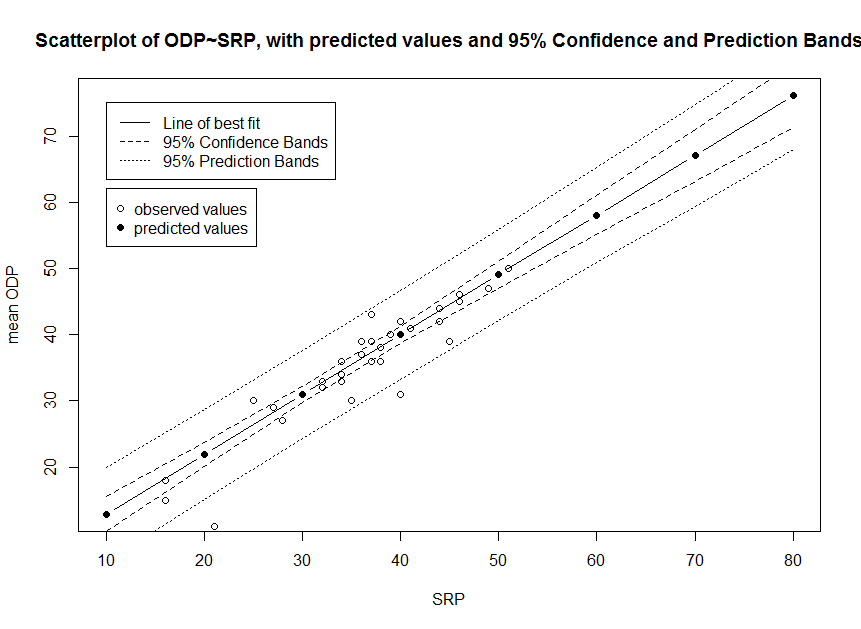
50 49.0 47.0 51.1 49.0 42.12 55.9

60 58.0 55.1 61.0 58.0 50.83 65.3

70 67.1 63.2 71.0 67.1 59.43 74.7

80 76.1 71.2 81.0 76.1 67.92 84.3

par(mfrow=c(1,1)) *# 1 plot per page*plot(pred.df$SRP,CI[,1],type="b", pch=16, xlab="SRP", ylab="mean ODP", main="Scatterplot of ODP~SRP, with predicted values and 95% Confidence and Prediction Bands")  
points(dat1$ODP, dat1$SRP)  
legend(10,50, lty=c(1, 2,3), legend=c("Line of best fit", "95% Confidence Bands", "95% Prediction Bands"))  
legend(10,42, pch=c(1,16), legend = c("observed values", "predicted values"))  
lines(pred.df$SRP,CI[,2],lty=2) *# lty indicates line type, lty= 2 -> dashed line*lines(pred.df$SRP,CI[,3],lty=2)  
lines(pred.df$SRP,PI[,2],lty=3) *# lty=3 produces a dotted line*  
lines(pred.df$SRP,PI[,3],lty=3)



**Q11:** With reference to the CI and PI for the predicted ODP for each of the eight values of SRP (previous page), and the plot shown above, comment on the reliability of using the SLR model for prediction.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**R script file**

Note that your script file should be structured in the following way:

* Comment giving brief summary of analysis, name of author
* Read in the data file
* Produce exploratory plots
* Fit various models using appropriate model selection method.
* Obtain relevant output (e.g. estimates of coefficients, CIs etc)
* Check model assumptions using diagnostic plots
* Use the model for prediction, inference etc.
* Annotate commands with comments (#) so that script is explained

####################################################

*# fit simple linear regression model to tannery data.*

*# Code developed by Jackie Reid*

*# options() allows the user to set and examine a variety of*

*# global options which affect the way in which*

*# R computes and displays its results.*

*# In the options command below, we set the number of digits to print*

*# but it is a suggestion only and not strictly adhered to.*

*# show.signif.stars=F removes some outdated display in*

*# output associated with p-values*

options(digits=3,show.signif.stars=F)

*# read in data*

*# header=T assigns variable names to the first row of the dataset*

*# dat1 is the name of the dataframe that contains the data, which is*

*# read from the text file “CH03-SLREx.txt”*

dat1 <- read.table("CH03\_SLREx.txt",header=T)

*# fit simple linear regression model using lm(y~x, data= …)*

*# lm stands for* ***l****inear* ***m****odel*

xy.lm <- lm(ODP~SRP, data=dat1)

*# plot data and fitted line using plot(y~x) or plot(x,y)*

plot(ODP~SRP,data=dat1, ylab = "oxygen demand percent", xlab="solids reduction percent")

*# add the line of best fit to the existing plot*

abline(xy.lm)

*# alternative plot using package ggplot2*

*library(ggplot2)*

*ggplot(data=dat1, aes(x=SRP, y=ODP)) +*

*geom\_point(pch=17, color="blue", size=2) +*

*geom\_smooth(method="lm", color="red", linetype=2) +*

*labs(title="", x="solids reduction percent", y="oxygen demand percent")*

*# Table of regression coefficients, std errors, t-tests and p-values;*

*# and aov table*

summary(xy.lm)

anova(xy.lm)

*# CI for regression coefficients*

confint(xy.lm)

*# diagnostic plots – check model assumptions*

*# par(mfrow=c(1,2)) sets up plotting window*

*# to allow one row and 2 columns for the plots (ie 2 plots in window)*

par(mfrow=c(1,2))

*# plot(model) produces diagnostic (residuals plots)*

*#which=1:2 plots the first 2 of 5 possible residuals plots*

*# the residuals vs fitted and the Normal QQ plot*

*# add.smooth=F removes a smoothing curve from the plot,*

*# which can sometimes be distracting or misleading.*

plot(xy.lm,which=1:2, add.smooth=F)

*# Shapiro Wilk's Test of normality*

*# Null hypothesis is that residuals come from a normal distribution*

print(shapiro.test(xy.lm$residuals))

*# predictions for 8 data points: SRP = 10,20,30,40,50,60,70,80*

*# create new data frame of 8 observations*

pred.df <- data.frame(SRP=c(10,20,30,40,50,60,70,80))

*# predict mean values and 95% CI for those predictions*

CI <- predict(xy.lm,interval="confidence",newdata=pred.df,level=0.95)

*# predict mean values and 95% prediction intervals*

PI<- predict(xy.lm,interval="predict",newdata=pred.df,level=0.95)

*# print SRP, fitted values 95%CIs and 95% PIs in columns (cbind)*

*# rbind() would print them in rows*

*# note use of with to identify dataframe used*

with(pred.df, cbind(SRP,CI,PI))

*# par(mfrow=c(1,1)) sets up plotting window*

*# to allow one row and 1 column for the plot(s)*

*# one plot in this case*

par(mfrow=c(1,1))

*# plot predicted values against SRP*

*# note use of $ sign to identify that SRP comes from pred.df data*

*# frame (as distinct from SRP in dat1 data frame)*

*# CI[,1] is the first column in the object labelled CI*

*# It contains the predicted or “fitted values”*

*# type=”b” plots both the points and lines joining the points*

*# other options include type=”l” (lines only), type=”p” (points only)*

*# pch = 16 plots the points as dots, pch=1 plots as an open circle*

plot(pred.df$SRP,CI[,1],type="b", pch=16, xlab="SRP", ylab="mean ODP", main="Scatterplot of ODP~SRP, with predicted values and 95% Confidence and Prediction Bands")

*# points() command adds points to an existing plot*

points(dat1$ODP, dat1$SRP)

*# legend(x,y,..) places legend on plot at coordinates (x,y)*

*# lty=c(1,2,3) puts 3 lines in the legend*

*# lty=1 is a solid line; lty=2 is a dashed line;*

*# lty=3 is a dotted line*

legend(10,75, lty=c(1, 2,3), legend=c("Line of best fit", "95% Confidence Bands", "95% Prediction Bands"))

legend(10, 62, pch=c(1,16), legend = c("observed values", "predicted values"))

*# lines() adds lines to an existing plot*

*# CI[,2] gives the lower bound of the CI;*

*# CI[,3] contains the upper bound*

*# similarly for PI*

lines(pred.df$SRP,CI[,2],lty=2)

lines(pred.df$SRP,CI[,3],lty=2)

lines(pred.df$SRP,PI[,2],lty=3)

lines(pred.df$SRP,PI[,3],lty=3)