This tutorial is designed to give you practice in the following:

1. Calculating VIF’s
2. Checking a model for linearity (planar data), and selecting transformations
3. Checking for outliers

In this tutorial we will be using the car data available from the website and cecil. These data were used in Assignment 2, 2002. The tutorial will help you towards attempting Assignment 2.

In this tutorial you will explore a data set consisting of various measurements on 138 cars that were taken from Road and Track’s “The Complete ’99 Car Buyer’s Guide”. The variables in this data set are:

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CITY</td>
<td>mileage (miles per gallon) in city driving (response)</td>
</tr>
<tr>
<td>PRICE</td>
<td>price in USD</td>
</tr>
<tr>
<td>WEIGHT</td>
<td>weight in pounds</td>
</tr>
<tr>
<td>DISP</td>
<td>displacement in cubic centimetres</td>
</tr>
<tr>
<td>COMP</td>
<td>compression ratio as value goes to 1</td>
</tr>
<tr>
<td>HP</td>
<td>horsepower at 6300rpm</td>
</tr>
<tr>
<td>TORQ</td>
<td>torque at 5200rpm</td>
</tr>
<tr>
<td>TRANS</td>
<td>transmission (1 = automatic, 0 = manual)</td>
</tr>
<tr>
<td>CYL</td>
<td>number of cylinders</td>
</tr>
</tbody>
</table>

**Task 1: Read the data**

Download the cars data from the web. Make a data frame cars.df. The response variable of interest is CITY, and we want to model this in terms of the other variables. The variable names in this example are UPPER CASE.

**Task 2: Make a pairs plot**

Make a pairs plot of the data. Are any of the explanatory variables closely related to each other? Calculate the correlations between the variables to confirm what you see in the plot. Do the relationships seem linear?

```
pairs(cars.df)
X <- cars.df[,-1] # why do we do this?
round(cor(X),3)
```
Task 3: Fit a model and compute the VIF’s

Fit a model not using column 1.

cars.lm1 <- lm(CITY~.,data=cars.df[,,-1])
summary(cars.lm1)

Compute VIF’s

diag(solve(cor(cars.df[-1])))

There are a few high VIF’s. Which variables are affected? Can you explain why some variables are quite highly correlated with the response but are not significant in the regression?

Task 4: Some model building

The usual remedy for collinearity is to remove variables from the model. We start by removing TORQ as it has the highest VIF and is not significant in the model.

cars.lm2 <- lm(CITY~.,data=cars.df[,,-c(1,8)])
summary(cars.lm2)
diag(solve(cor(cars.df[,,-c(1,8)])))

We next remove DISP for highest VIF and because it is still non-significant.

cars.lm3 <- lm(CITY~.,data=cars.df[,,-c(1,5,8)])
summary(cars.lm3)
diag(solve(cor(cars.df[,,-c(1,5,8)])))

After this removal, the VIF values are alright, but we still have four variables with non-significant coefficients, which we will remove until all variables are significant. This leaves us with a model of only three variables.

cars.lm4 <- lm(CITY~.,data=cars.df[,,-c(1,5,6,8,9,10)])
summary(cars.lm4)

Task 5: Examine the residuals

par(mfrow=c(2,2)) # 2x2 layout of plots
plot(cars.lm4)

Main points:
- Residuals/fitted values plot confirms the non-linearity
- Outliers (pts 30, 47, 125, 69?)

Sometimes we get a lot of outliers when the data are non-planar, as is the case here.

Task 6: Transform the response

The relationship between the response CITY and the other variables does not appear linear. Instead of checking for polynomial fits of the covariates we will use the Box-Cox plot to select a suitable transformation of the response.

```r
library(MASS) # Don't forget to load MASS
boxcox(cars.lm4)
```

We go for a fit with $\lambda = -1/2$.

```r
cars.lm5 <- lm(1/sqrt(CITY)~PRICE+WEIGHT+HP,data=cars.df)
summary(cars.lm5)
```

The transformation seems good and improves the fit. However, as the residuals show, we still have some curvature

```r
plot(cars.lm5,which=1)
```

Task 7: Transforming the covariates

We use GAM plots to assess whether any one of PRICE, WEIGHT or HP need transforming.

```r
library(mgcv) # do the GAM plots, do not forget to load the mgcv library
par(mfrow=c(1,3)) # Three covariates
plot(gam(1/sqrt(CITY)~s(PRICE)+s(WEIGHT)+s(HP),data=cars.df))
```

PRICE does not need transforming as the majority of points is in the left corner, and the crazy line in based on only a few observations. WEIGHT gets a square, and HP should be fine.

```r
cars.lm6 <- lm(1/sqrt(CITY)~PRICE+WEIGHT+I(WEIGHT^2)+HP,data=cars.df)
summary(cars.lm6)
```
**Task 8: Re-examine residuals**

The curvature has been alleviated, and the uprising on the right seems to correlate with the pattern for \textit{PRICE}, so we leave the model building at this stage. Let us see what the residuals say.

\[\text{par(mfrow} = c(2,2))\]
\[\text{plot(cars.lm6,which} = c(1,3,4,5))\]

Residual plots look OK, except for the outliers 46 (Dodge Viper RT/10) 47 (Ferrari F355 Berlinetta). Remove this

\[\text{cars.lm7} \leftarrow \text{lm}(1/\sqrt{\text{CITY}} \sim \text{PRICE} + \text{WEIGHT} + I(\text{WEIGHT}^2) + \text{HP}, \text{data} = \text{cars.df[-c(46,47),]})\]

At this point we have a reasonable model: the $R^2$ is now 92% and the residual plots look OK.

**Task 9: Assess the size of the largest studentised residual**

We noted in class that the studentised residuals are approximately normally distributed, so lie between $\pm 2$ with 95% probability.

However, the \textit{largest} residual is typically bigger than 2 in magnitude. The bigger the sample size, the bigger the largest residual will be.

To quantify this, we can do a small simulation: repeatedly draw normal samples, and calculate and record the largest value. We can then draw a histogram of these largest values to get an idea what is typical.

For example, for the cars data, there are $n = 136$ observations, excluding the Dodge Viper RT/10 and Ferrari F355 Berlinetta.

The following \texttt{R} code draws $N = 1000$ random normal samples of size $n = 136$, and records the 1000 maximum values in a vector \texttt{result}:

\[N \leftarrow 1000\]
\[n \leftarrow 136\]
\[\text{result} \leftarrow \text{numeric}(N)\]
\[\text{for}\{i \text{ in } 1:N\}\{\]
\[\quad \text{result}[i] \leftarrow \text{max(abs(rnorm(n)))}\]
\[\}\]

Draw a histogram of the results. The largest studentised residual from the regression fit had magnitude 2.578969, so this is not atypical.

\[\text{max(abs(rstudent(cars.lm7)))}\]