Executive summary

Problem description
You work for Motor Trend, a magazine about the automobile industry. Looking at a data set of a collection of cars, they are interested in exploring the relationship between a set of variables and miles per gallon (MPG) (outcome). They are particularly interested in the following two questions:

- Is an automatic or manual transmission better for MPG?
- Quantify the MPG difference between automatic and manual transmissions.

The data
The data for this analysis is the data set mtcars of the R dataset package. It was extracted from the 1974 Motor Trend US magazine and it contains characteristics like fuel consumption, automobile design and performance for 32 automobiles. The complete list of variables is:

Name	Description
mpg	Miles/(US) gallon
cyl	Number of cylinders
disp	Displacement (cu.in.)
hp	Gross horsepower
drat	Rear axle ratio
wt	Weight (lb/1000)
qsec	1/4 mile time
vs	V/S
am	Transmission (0 = automatic, 1 = manual)
gear	Number of forward gear
carb	Number of carburetors

The variables are all numeric.

Solution
After conducting some exploratory analysis, we have fit different linear regression models and we have discovered that manual transmission cars are better for mpg than automatic ones but there are other variables like weight and cylinder that are better predictor for MPG and the transmission type depends on them.

Solution details

Load the data
After loading the data, the transmission type (am) is transformed in a factor for convenience.

```
data(mtcars)
mtcars$am = factor(mtcars$am, label = c("automatic", "manual"))
```

Exploratory analysis
From the summary below and the boxplot in the appendix (figure 1), it is possible to see that MPG statistics (mean, median, ...) are greater for manual transmission than automatic transmission. This suggests that manual transmission could be better than automatic transmission for MPG.

```
tapply(mtcars$mpg, mtcars$am, summary)
```

```
## $automatic
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    10.4    15.0    17.3    17.1    19.2    24.4
##
## $manual
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    15.0    21.0    22.8    24.4    30.4    33.9
```

Fit multiple models and model selection
The first model considered is the simplest model with transmission type (am) as the only one explanatory variable.

```
fit.simple <- lm(mpg ~ am, data = mtcars)
summary(fit.simple)$coeff
```

```
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept)   17.147      1.125  15.247 1.134e-15
## ammanual       7.245      1.764   4.106 2.850e-04
```

Interpreting the coefficients of the summary above, the model estimates an expected increase of 7.245 MPG for a change from automatic transmission to manual transmission, on average.

After calculating confidence intervals,

```
confint(fit.simple)
```

```
##              2.5 % 97.5 %
## (Intercept) 14.851  19.44
## ammanual     3.642  10.85
```

it is possible to affirm with 95% confidence that a change from automatic transmission to manual transmission results in a 3.642 to 10.85 increase in MPG. Nevertheless, this model is not very expressive because the adjusted R-squared is 0.3385 and this means that the model explains only 33.85% of the variance. We use adjusted R-squared instead of R-squared to measure the regression model because adjusted R-squared does not increase if we add variables that doesn't really provide any new information. A general approach to find a better model is to consider the most complete model with all the variables of the data set as explanatory variables and then applying stepwise model selection with backwards elimination technique.

Before fitting the model, we transform the numeric variables cyl, vs, gear and carb in factors for convenience because they can assume only few integer values.

```
mtcars$vs = factor(mtcars$vs)
mtcars$cyl = factor(mtcars$cyl)
mtcars$gear = factor(mtcars$gear)
mtcars$carb = factor(mtcars$carb)
```

Final model
The explanatory variables of the final model are cylinders (cyl), gross horsepower (hp), weight (wt) and transmission type (am).

```
fit.total <- lm(mpg ~ ., data = mtcars)
fit.final <- step(fit.total)
```

```
summary(fit.final)$coeff
```

```
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept) 33.70832    2.60489 12.9404 7.733e-13
## cyl6        -3.03134    1.40728 -2.1540 4.068e-02
## cyl8        -2.16368    2.28425 -0.9472 3.523e-01
## hp          -0.03211    0.01369 -2.3450 2.693e-02
## wt          -2.49683    0.88559 -2.8194 9.081e-03
## ammanual     1.80921    1.39630  1.2957 2.065e-01
```

The adjusted R-squared is now 0.8401 so this model is more expressive than the first one.

Holding the remaining variables constant, the new model estimates an expected increase of 1.809 MPG for a change from automatic transmission to manual transmission, on average . Therefore transmission type appears to have a lower impact on MPG than if all the other variables are disregarded as in the first model. Moreover, the transmission type is not statistically significant for this model if we use 0.05 as a type I error rate significance benchmark (the p-value of transmission type is 0.2065 and so larger than 0.05). In the new model indeed, the most significant variable is the weight and the model estimates an expected decrease of 2.497 MPG for every one ton increase in weight, holding the remaining variables constant.

After calculating confidence intervals,

```
confint(fit.final)
```

```
##                2.5 %    97.5 %
## (Intercept) 28.35390 39.062744
## cyl6        -5.92406 -0.138632
## cyl8        -6.85902  2.531671
## hp          -0.06025 -0.003964
## wt          -4.31718 -0.676478
## ammanual    -1.06093  4.679356
```

it is possible to affirm with 95% confidence that an increase of one ton in weight results in a 0.6765 to 4.3172 decrease in MPG. To summarize, manual transmission cars have higher MPG but probably due to their weight (and number of cylinders, not discussed here for reason of space).

Residual diagnostic
The residuals diagnostic plots in the appendix (figure 2) confirm a good regression fit. In more details, in the Residuals vs Fitted plot the points are randomly scattered with no particular pattern and in the Normal Q-Q plot the points follow the line so it is possible to conclude that the residuals are normally distributed around 0 with constant variability. In the other two plots, the points are more or less grouped near the center.
To summarize, the conditions for applying the final linear regression model are satisfied.

Appendix

Figure 1. Relationship between MPG and transmission type

```
boxplot(mtcars$mpg ~ mtcars$am, xlab = "Trasmission type", ylab = "MPG Miles/(US) gallon")
```

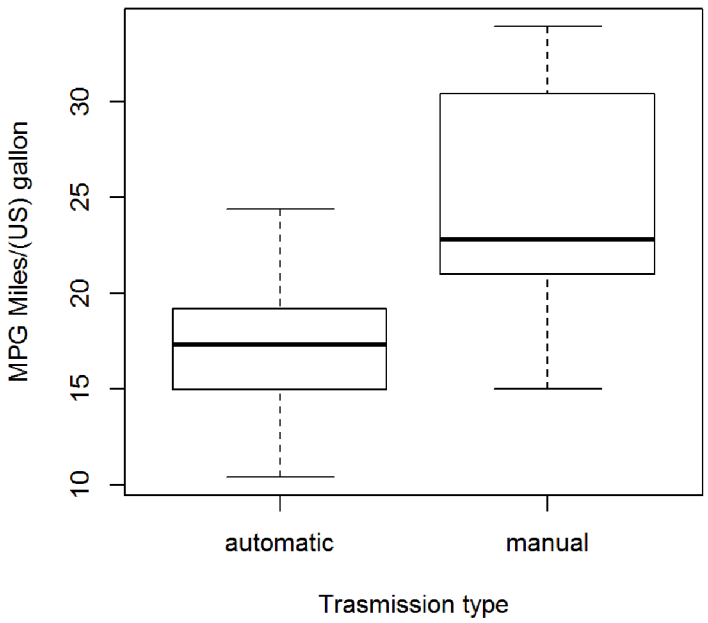
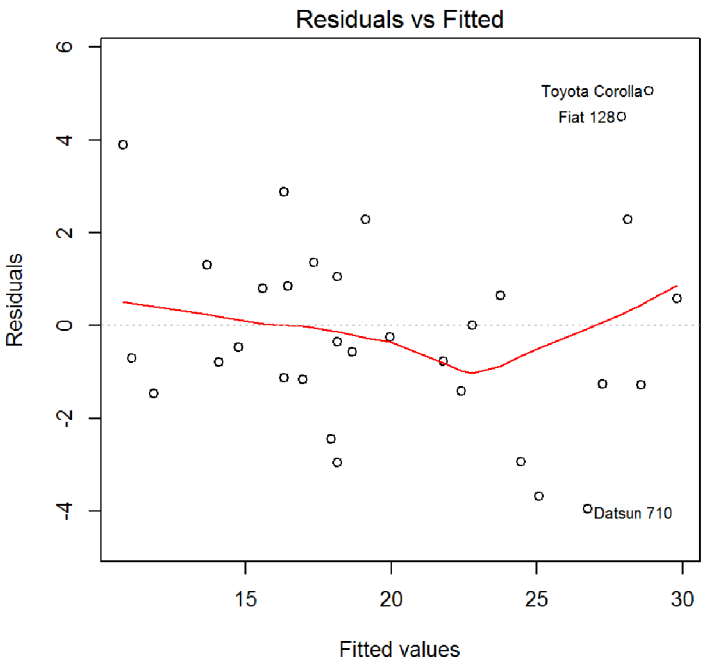
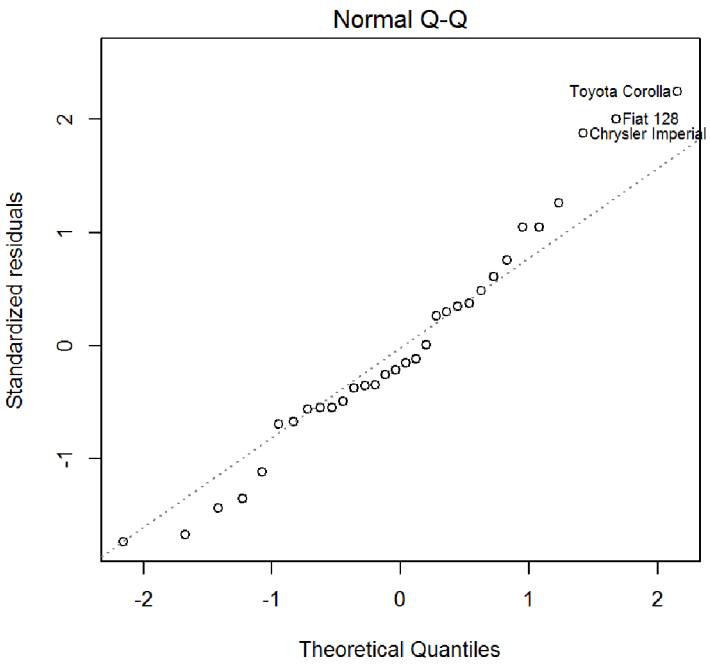
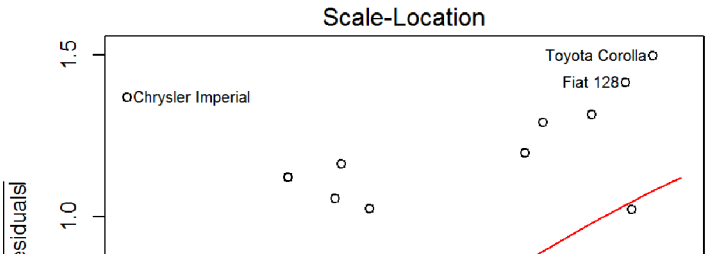


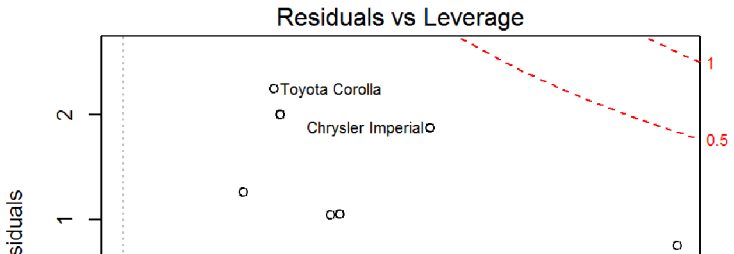
Figure 2. Residual diagnostics

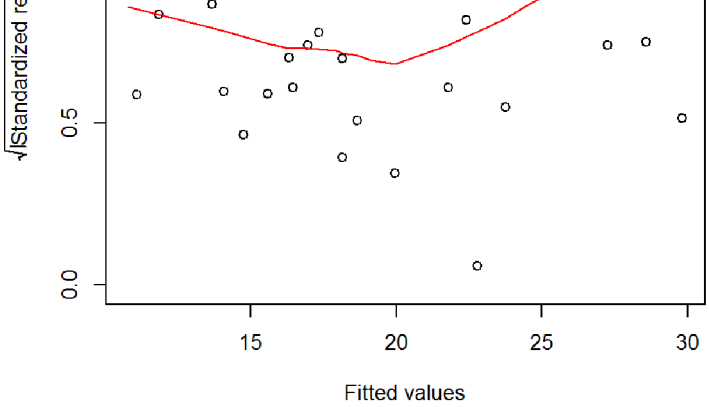
```
par(mfrow = c(2, 2))
plot(fit.final)
```

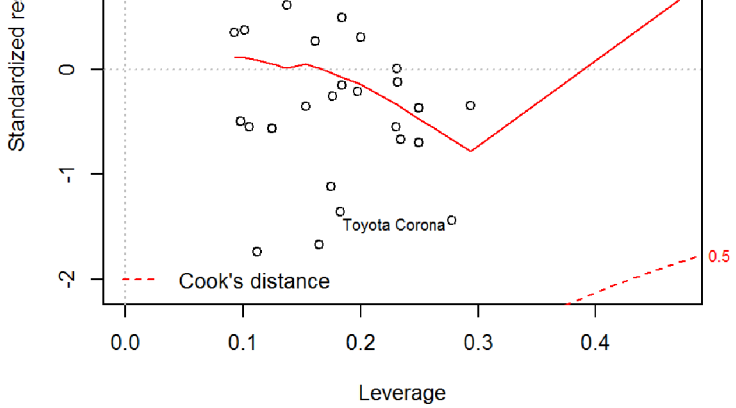









√|Standardized re
0.0
0.5
15
20
25
30
Fitted values


Standardized re
0
-1
-2
Toyota Corona
Cook's distance
0.5
0.0
0.1
0.2
0.3
0.4
Leverage