Regression_course_project

Summary:

The objective of this project is to use skills learned from the Coursera regression course to analyze the mtcars dataset. I have used single and multivariate linear regression to address whether there the difference between an automatic and manual transmission affects the fuel consumption (measured in miles per gallon) and if so, by how much. My analysis shows that considering only transmission type (am) suggests that manual transmissions increase fuel economy by 7 mpg. However, when also considering other covariables, including vehicle weight, the number engine cylinders and the engine displacement, transmission type is not a significant predictor of fuel economy.

Data exploration:

After loading the data and viewing the content of the variables, I converted the discrete variables to factors.

```
#load data
library(datasets)
data(mtcars)
head(mtcars)
```

```
##                    mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

```
#copy data and add labels to am factor variable
cars <- within(mtcars, {
        vs <- factor(vs, labels = c("V", "S"))
        am <- factor(am, labels = c("automatic", "manual"))
        cyl  <- ordered(cyl)
        gear <- ordered(gear)
        carb <- ordered(carb)
})
```

I then plotted the fuel economy (mpg) as a function of transmission type (am) (plot 1). The mean of manual transmissions is about 7 mpg larger than that for automatic transmissions.

```
#what are the means
aggregate(mpg~am, cars, mean)
```

```
##          am      mpg
## 1 automatic 17.14737
## 2    manual 24.39231
```

Inferential analysis:

In order to determine if this is a meaningful difference, I performed a linear regression. The slope coefficient (ammanual) is equal to the difference in the means shown above. The means that switching from an automatic to manual transmission increases fuel economy by about 7.2 mpg. Since the p-value for the slope coefficient is much less than alpha = 0.05, this suggests that this difference does not occur by chance. However, the R^2 value shows that transmission type only accounts for about 36% percent of the variability in fuel economy. Indeed, a plot of the residuals (plot 2) show quite a bit of scatter. The range was -10 to 10 mpg.

```
#what is the relationship
fit <- lm(mpg~am, cars)
summary(fit)
```

```
##
## Call:
## lm(formula = mpg ~ am, data = cars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3923 -3.0923 -0.2974  3.2439  9.5077
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   17.147      1.125  15.247 1.13e-15 ***
## ammanual       7.245      1.764   4.106 0.000285 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.902 on 30 degrees of freedom
## Multiple R-squared:  0.3598, Adjusted R-squared:  0.3385
## F-statistic: 16.86 on 1 and 30 DF,  p-value: 0.000285
```

So, I was curious if other variables might also influence fuel economy. I performed an analysis of variance of the multivariate linear regression using all variables as regressors to predict mpg. Vehicle weight (wt), the number engine cylinders (cyl) and the engine displacement (disp) all showed significant main effects suggesting that they may also have influence on fuel economy.

```
#what other variable may be correlated
anova(lm(mpg~., cars))
```

```
## Analysis of Variance Table
##
## Response: mpg
##           Df Sum Sq Mean Sq F value    Pr(>F)
## cyl        2 824.78  412.39 51.3766 1.943e-07 ***
## disp       1  57.64   57.64  7.1813   0.01714 *
## hp         1  18.50   18.50  2.3050   0.14975
## drat       1  11.91   11.91  1.4843   0.24191
## wt         1  55.79   55.79  6.9500   0.01870 *
## qsec       1   1.52    1.52  0.1899   0.66918
## vs         1   0.30    0.30  0.0376   0.84878
## am         1  16.57   16.57  2.0639   0.17135
## gear       2   5.02    2.51  0.3128   0.73606
## carb       5  13.60    2.72  0.3388   0.88144
## Residuals 15 120.40    8.03
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Therefore, I added cyl, disp, and wt to my model. The Rˆ2 value shows that the new model explains ~83% of the variability in mpg. Looking at the 95% confidence interval gives me confidence that the calculated coefficients accurately represent the population. Moreover, a plot of the residuals (plot 3) shows that residuals are normally distributed around the predicted values and there are unlikely to be any influential outliers.

```r
#what if we model the data using additional covariates
newfit <- lm(mpg~am+cyl+disp+wt, cars)
summary(newfit)
```

```
##
## Call:
## lm(formula = mpg ~ am + cyl + disp + wt, data = cars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5029 -1.2829 -0.4825  1.4954  5.7889
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 30.275005   3.290562   9.201 1.17e-09 ***
## ammanual     0.141212   1.326751   0.106   0.9161
## cyl.L       -4.467788   1.872177  -2.386   0.0246 *
## cyl.Q        0.935362   1.052358   0.889   0.3822
## disp         0.001632   0.013757   0.119   0.9065
## wt          -3.249176   1.249098  -2.601   0.0151 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.652 on 26 degrees of freedom
## Multiple R-squared:  0.8376, Adjusted R-squared:  0.8064
## F-statistic: 26.82 on 5 and 26 DF,  p-value: 1.73e-09
```

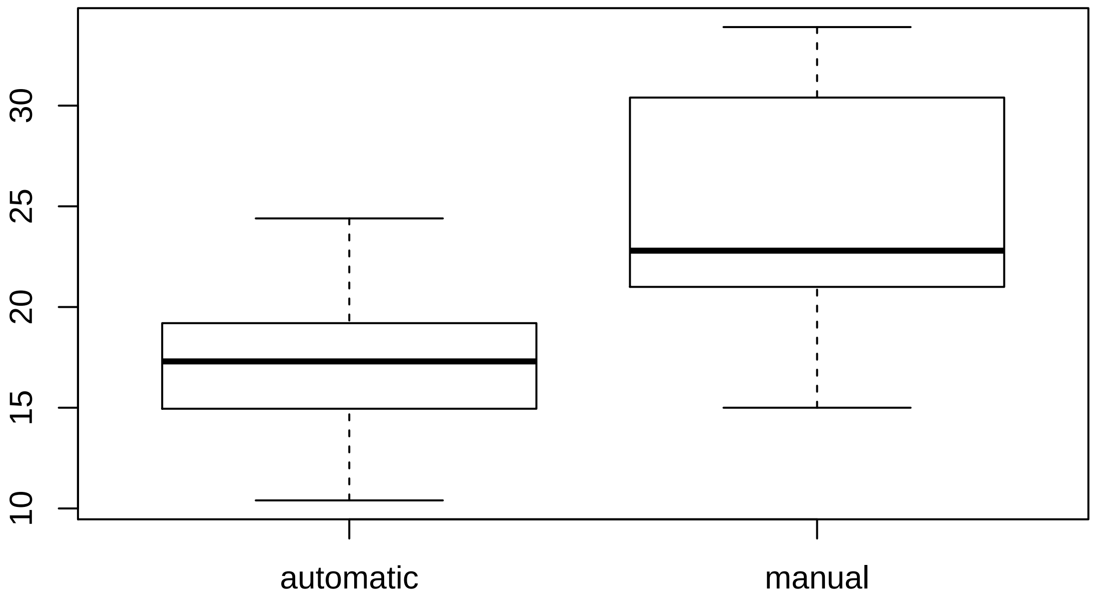
```r
confint(newfit)
```

```
##                   2.5 %      97.5 %
## (Intercept) 23.51115785 37.03885122
## ammanual    -2.58596405  2.86838805
## cyl.L       -8.31610165 -0.61947338
## cyl.Q       -1.22779038  3.09851407
## disp        -0.02664563  0.02990995
## wt          -5.81673442 -0.68161740
```

Appendix:

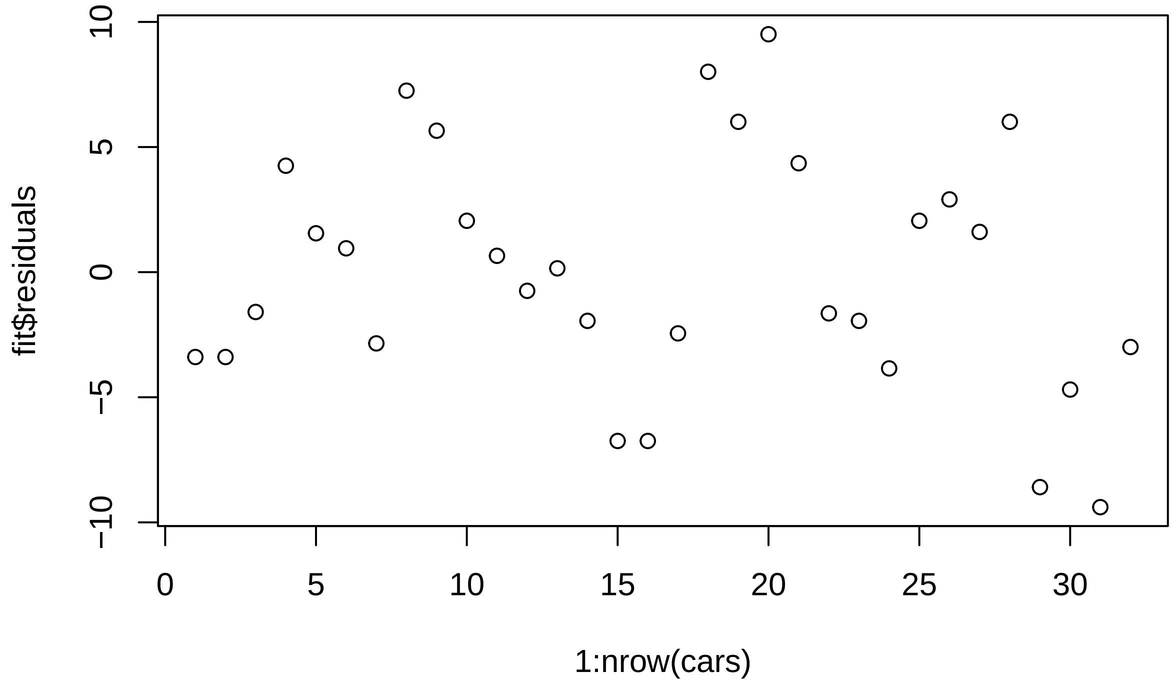
Plot 1: Plot of fuel economy (mpg) as a function of transmission type (am).

```
boxplot(mpg~am, cars)
```



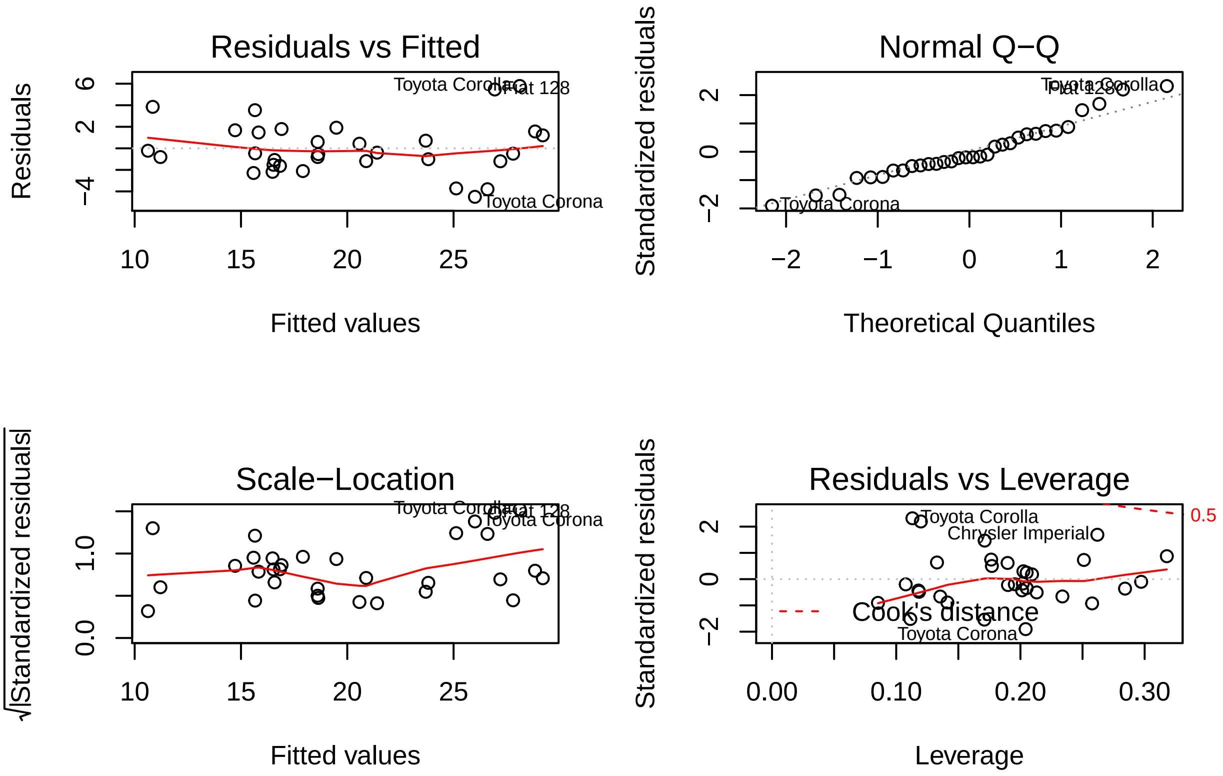
Plot 2: Residual plot of single variable linear regression (mpg~am).

```
plot(1:nrow(cars), fit$residuals)
```



Plot 3: Residual diagnositics of multivariate linear regression.

```
par (mfrow = c(2,2))
plot(newfit)
```


Residuals vs Fitted
Residuals
6
2
−4
Toyota Corolla
Fiat 128
Toyota Corona
10
15
20
25
Fitted values
Normal Q−Q
Standardized residuals
2
0
−2
Toyota Corolla
Fiat 128
Toyota Corona
−2
−1
0
1
2
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
1.0
0.0
Toyota Corolla
Fiat 128
Toyota Corona
10
15
20
25
Fitted values
Residuals vs Leverage
Standardized residuals
2
0
−2
Toyota Corolla
Chrysler Imperial
Toyota Corona
Cook's distance
0.5
0.00
0.10
0.20
0.30
Leverage