

Regression models Course Assignment

Jean-Baptiste Poulet

February 16, 2015

Introduction and executive summary

The goal of this document is to respond to 2 main questions: “is an automatic or manual transmission better for MPG”, “what is the MPG difference between automatic and manual transmissions”. The data are briefly described in Section **Material**. We explain how we build our models in Section Methods and results. Responses to the above mentioned questions are given in Section Discussion.

Material

The *mtcars* data are the input data. More information can be found about this data set by typing `?mtcars` in R.

Methods and results

We first explore the data.

Using a scatterplot matrix (see Appendix), one can see that most of the predictors seem to have some impact on MPG. At this point, we do not see obvious outliers. Let’s fit a linear model with all variables.

```
fit <- lm(mpg ~ ., data=df)
summary(fit)
```

None of the variables shows a P-value smaller than 0.05. Let’s select a subset of these variables using the AIC stepwise selection and the `regsubsets` function from the R *leaps* package.

```
# AIC stepwise selection (both direction)
library(MASS)
stepB <- stepAIC(fit, direction="both")
# confirming our variable selection with a second method
library(leaps)
leaps <- regsubsets(mpg ~ ., data = df, nbest = 10)
```

Both methods recommend to use the variables *wt*, *am*, *hp* and *cyl* as predictors in the model, where we retrieve our variable of interest *am* (see Appendix for the results of the `regsubsets` function).

```
fitR <- lm(mpg ~ wt + am + hp + cyl, data=df)
summary(fitR)
```

The intercept and the variables *wt*, *hp* and *cyl* show p-values smaller than 0.05 (more details in the appendix). Let’s have a look at the residuals plots to assess how well this model fits the data. Based on the residuals plots, it seems that the model has some difficulty to fit the data with low or high MPG values. Let’s see how we may correct this based on our scatterplot matrix (see Appendix).

There seems to be some non linear function between MPG and WT. Let’s see whether a $\log(wt)$ instead of *wt* may improve the model and how the residual plots have changed.

```
fitR2 <- lm(mpg ~ log(wt) + hp + cyl + am, data=df)
anova(fitR2)
```

The curvature of the residuals vs fitted values has been reduced. Similarly, the squared of the standardized residuals plot does not show anymore an increasing slope. The normal Q-Q plots seem to show less deviation from the normality for the error term. Based on the plot “standardized residuals vs leverage”, none of the points is above a Cook’s distance threshold of 0.5, which indicates that none of the point distorts the outcome and accuracy of our regression model. See Appendix for more details.

The hat values obtained with the influence function describes the influence each observed value has on each fitted value. 1 point seems to have more influence on the fitted values: “Maserati Bora. However this point is not dramatically high (see Appendix). The final selected model is $mpg \sim \log(wt) + hp + cyl + am$.

Discussion

In this section both questions mentioned in the Introduction are answered based the model built in Section Methods and results. Summary of the model is given the Appendix.

Here is how to understand the coefficients of the model:

Numerical variables

- if $\log(wt)$ changes by 1 (or wt changes by 10) the *mpg* value changes by -10.133,
- if *hp* changes by 1 the *mpg* value changes by -0.027,

Factor variables

- if we have 6 cylinders (*cyl6*), *mpg* changes by -2.205 compared to having 4 cylinders,
- if we have 8 cylinders (*cyl8*), *mpg* changes by -1.789 compared to having 4 cylinders,
- if we have a manual transmission (*am1*), *mpg* changes by 0.867 compared to having an automatic transmission.

```
(ci <- confint(fitR2))
```

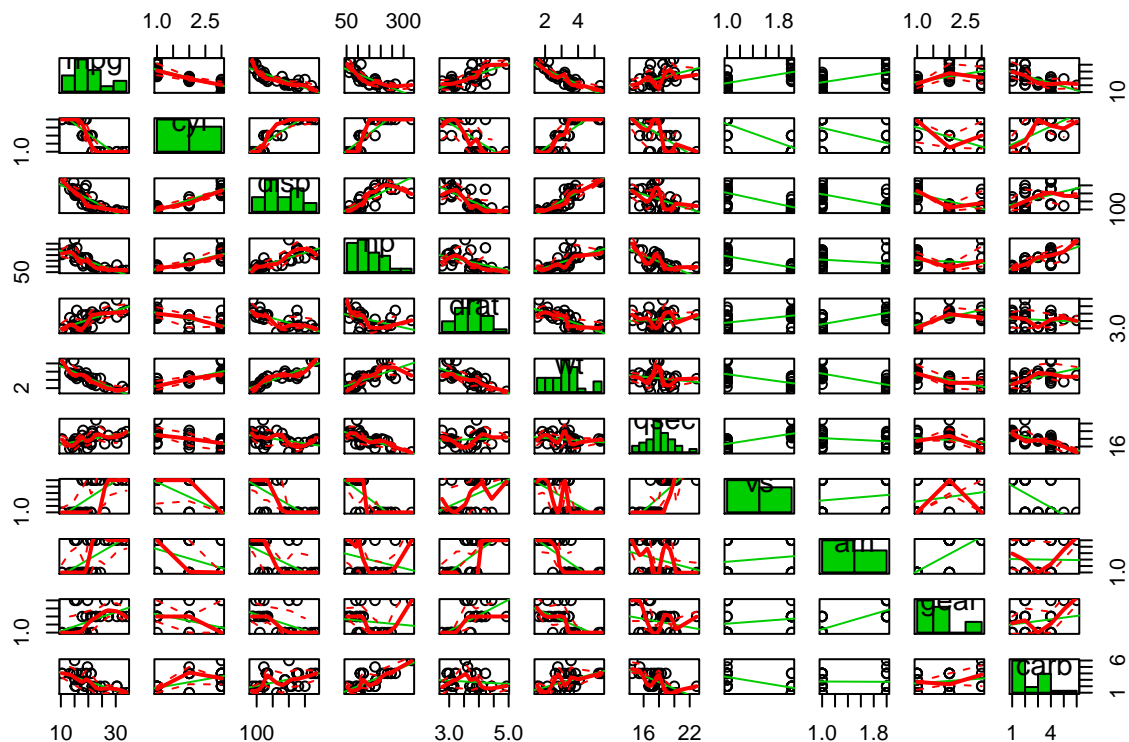
```
##              2.5 %          97.5 %
## (Intercept) 30.52014642 42.2557456987
## log(wt)     -16.07153044 -4.1945572966
## hp          -0.05441425 -0.0003648742
## cyl6        -5.04344839  0.6333139289
## cyl8        -6.22770611  2.6496674261
## am1         -2.03562296  3.7688902021
```

Looking at the 95%-confidence interval of the estimates, one can see that the AM variable shows the interval [-2.04,3.77]. So it is not possible to tell whether an automatic transmission is better for MPG than a manual one. With the likelihood ratio test, we can say at least that adding the AM term in our model is not significantly better for estimating MPG as shown in the Appendix.

Appendix

Scatterplot matrix of all variables including MPG.

```
scatterplotMatrix(df,diagonal='histogram')
```



Fit using all variables.

```
summary(fit)
```

```
##
## Call:
## lm(formula = mpg ~ ., data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.5087 -1.3584 -0.0948  0.7745  4.6251
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  23.87913   20.06582   1.190  0.2525
##      cyl6       -2.64870    3.04089  -0.871  0.3975
##      cyl8       -0.33616    7.15954  -0.047  0.9632
##      disp      0.03555    0.03190   1.114  0.2827
##      hp       -0.07051    0.03943  -1.788  0.0939 .
##      drat      1.18283    2.48348   0.476  0.6407
##      wt       -4.52978    2.53875  -1.784  0.0946 .
##      qsec      0.36784    0.93540   0.393  0.6997
##      vs1       1.93085    2.87126   0.672  0.5115
##      am1       1.21212    3.21355   0.377  0.7113
```

```
## gear4      1.11435    3.79952    0.293    0.7733
## gear5      2.52840    3.73636    0.677    0.5089
## carb2     -0.97935    2.31797   -0.423    0.6787
## carb3      2.99964    4.29355    0.699    0.4955
## carb4      1.09142    4.44962    0.245    0.8096
## carb6      4.47757    6.38406    0.701    0.4938
## carb8      7.25041    8.36057    0.867    0.3995
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.833 on 15 degrees of freedom
## Multiple R-squared:  0.8931, Adjusted R-squared:  0.779
## F-statistic:  7.83 on 16 and 15 DF,  p-value: 0.000124
```

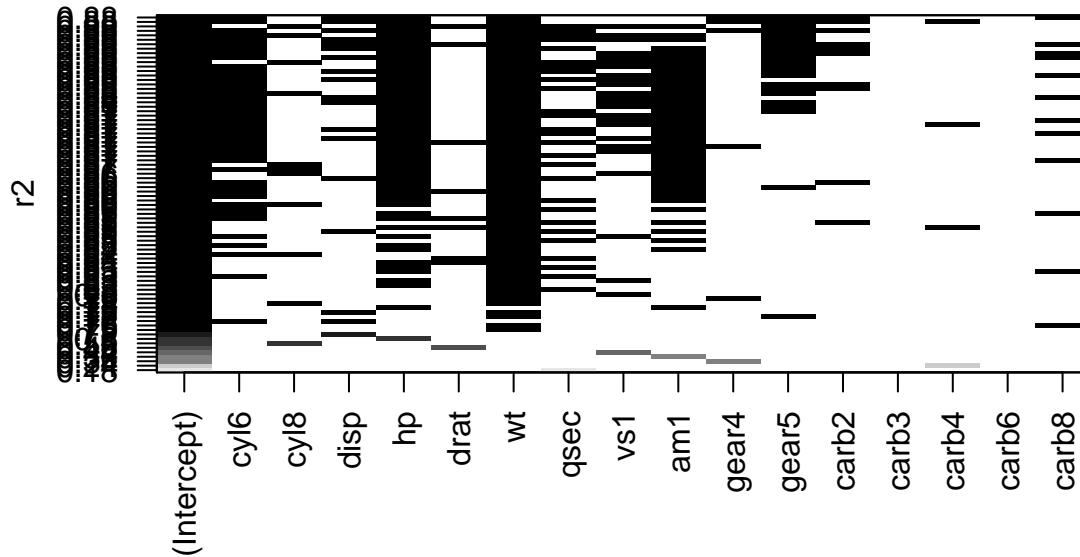
Variable selection using the stepAIC from the MASS package and the regsubsets function from the leaps package.

```
stepB$anova # display results
```

```
## Stepwise Model Path
## Analysis of Deviance Table
##
## Initial Model:
## mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
##
## Final Model:
## mpg ~ cyl + hp + wt + am
##
##
```

	Step	Df	Deviance	Resid. Df	Resid. Dev	AIC
## 1				15	120.4027	76.40339
## 2 - carb	5	13.5988573		20	134.0015	69.82769
## 3 - gear	2	5.0215145		22	139.0230	67.00492
## 4 - drat	1	0.9672159		23	139.9903	65.22678
## 5 - disp	1	1.2473996		24	141.2377	63.51066
## 6 - qsec	1	2.4420033		25	143.6797	62.05921
## 7 - vs	1	7.3459298		26	151.0256	61.65483

```
plot(leaps,scale="r2")
```



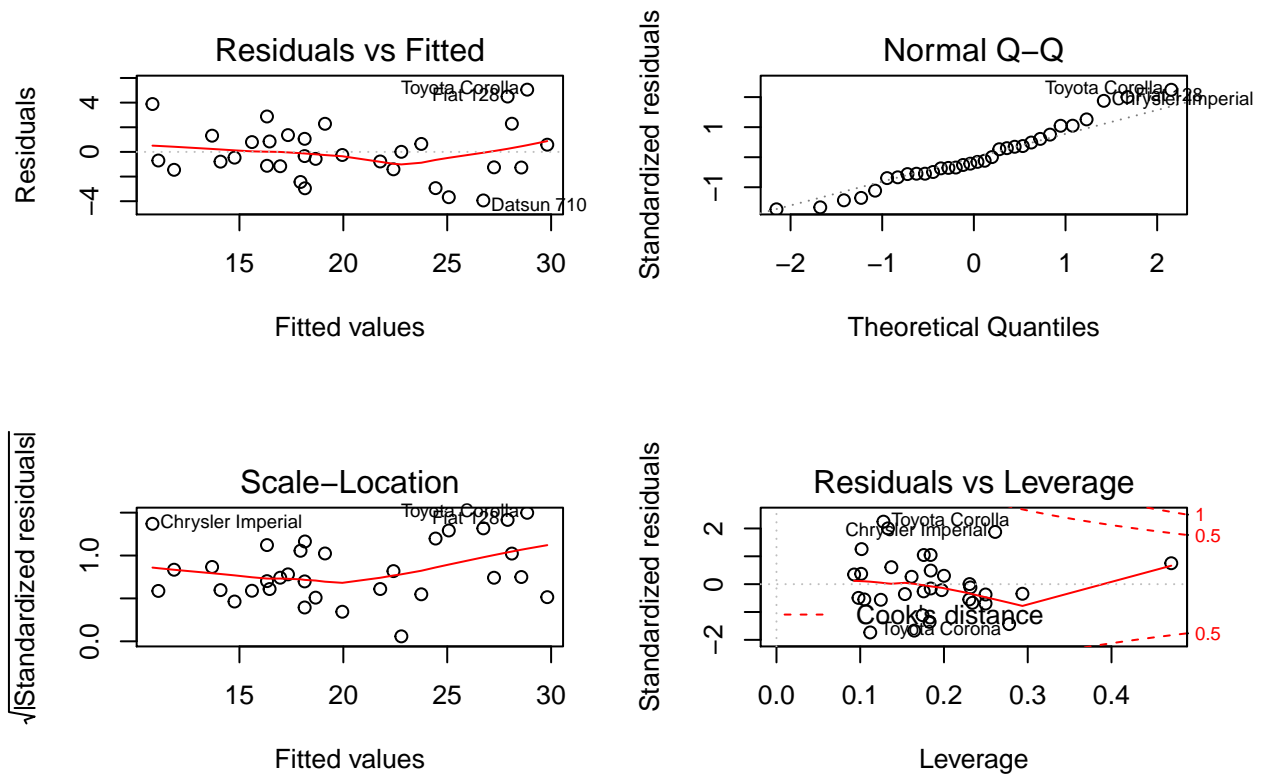
Summary of the model $mpg \sim wt + am + hp + cyl$

```
summary(fitR)
```

```
##
## Call:
## lm(formula = mpg ~ wt + am + hp + cyl, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.9387 -1.2560 -0.4013  1.1253  5.0513
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  33.70832    2.60489   12.940 7.73e-13 ***
## wt           -2.49683    0.88559   -2.819  0.00908 **
## am1            1.80921    1.39630    1.296  0.20646
## hp           -0.03211    0.01369   -2.345  0.02693 *
## cyl6          -3.03134    1.40728   -2.154  0.04068 *
## cyl8          -2.16368    2.28425   -0.947  0.35225
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.41 on 26 degrees of freedom
## Multiple R-squared:  0.8659, Adjusted R-squared:  0.8401
## F-statistic: 33.57 on 5 and 26 DF, p-value: 1.506e-10
```

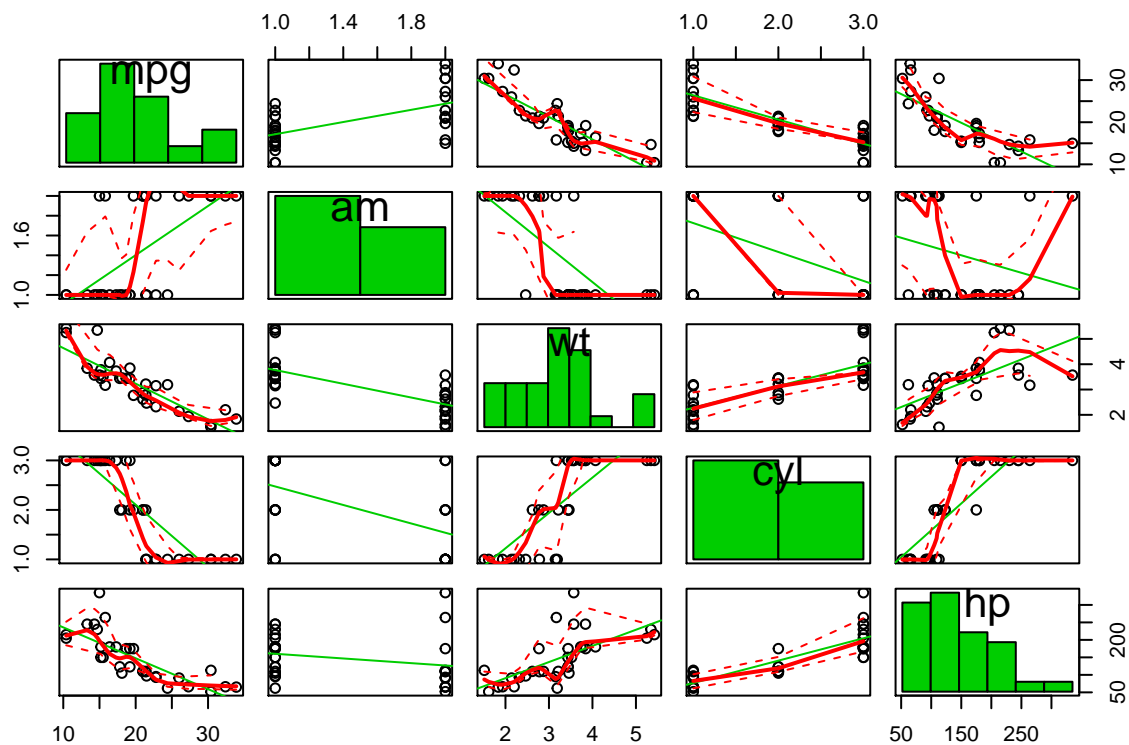
Residuals for the model $mpg \sim wt + am + hp + cyl$

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



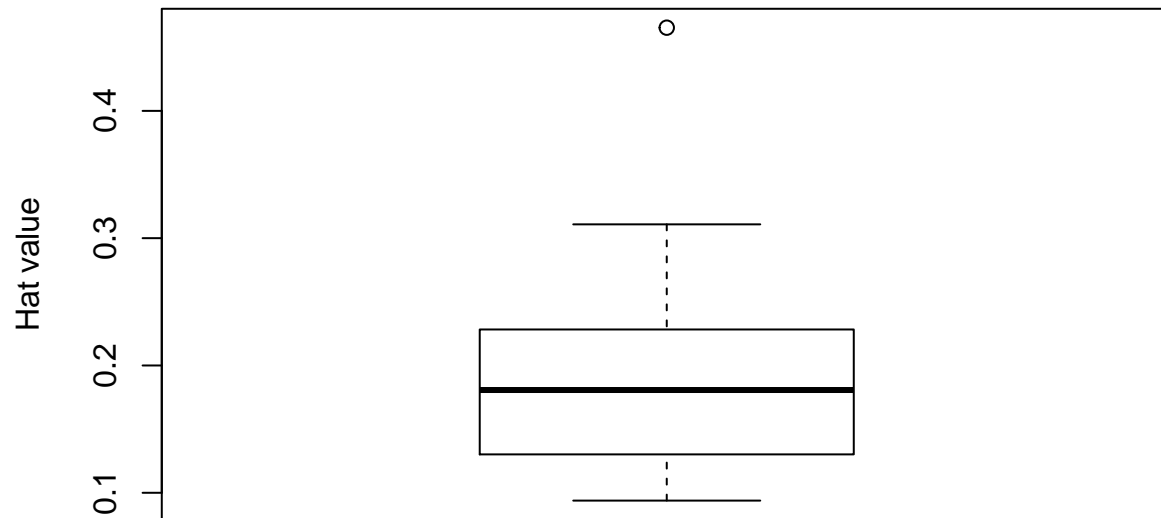
Scatterplot matrix for the model $mpg \sim wt + am + hp + cyl$

```
scatterplotMatrix(df[,c('mpg', 'am', 'wt', 'cyl', 'hp')], diagonal='histogram')
```



Boxplot of the hat values for the model $mpg \sim \log(wt) + am + hp + cyl$

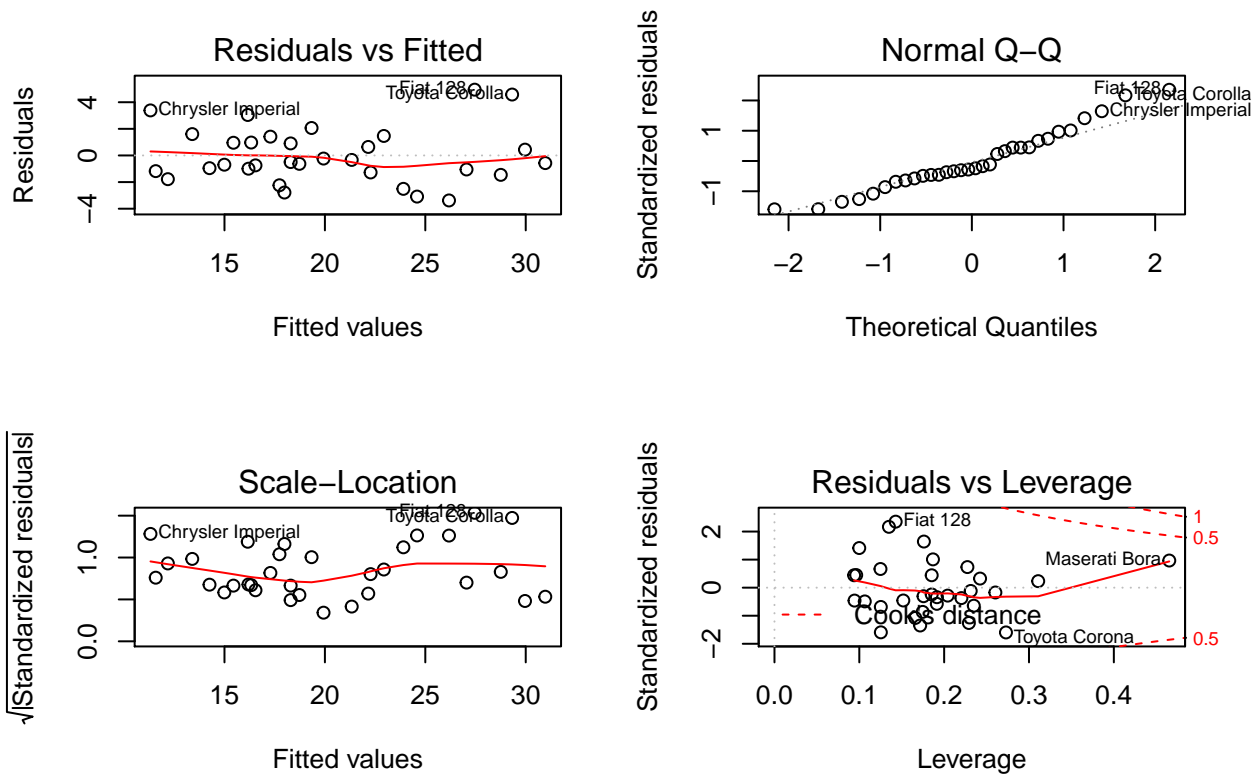
```
boxplot(hat,ylab="Hat value")
```



```
#identify(rep(1, length(hat)), hat, labels = names(hat))
```

Residuals for the model $mpg \sim \log(wt) + am + hp + cyl$

```
par(mfrow=c(2,2))  
plot(fitR2)
```



Fit summary for the model $mpg \sim \log(wt) + am + hp + cyl$

```
summary(fitR2)
```

```
##
## Call:
## lm(formula = mpg ~ log(wt) + hp + cyl + am, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.380 -1.202 -0.534  1.081  4.943
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  36.38795    2.85464  12.747 1.09e-12 ***
## log(wt)      -10.13304    2.88903  -3.507  0.00166 **
## hp           -0.02739    0.01315  -2.083  0.04720 *
## cyl16        -2.20507    1.38085  -1.597  0.12237
## cyl18        -1.78902    2.15939  -0.828  0.41494
## am1           0.86663    1.41193   0.614  0.54468
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.269 on 26 degrees of freedom
## Multiple R-squared:  0.8811, Adjusted R-squared:  0.8583
## F-statistic: 38.54 on 5 and 26 DF,  p-value: 3.214e-11
```


Comparing the models with and without *am* ($mpg \sim \log(wt) + am + hp + cyl$ VS $mpg \sim \log(wt) + hp + cyl$)

```
## Likelihood ratio test
##
## Model 1: mpg ~ log(wt) + hp + cyl + am
## Model 2: mpg ~ log(wt) + hp + cyl
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1    7 -68.303
## 2    6 -68.533 -1  0.4604    0.4975
```