Which transmission better for MPG

Fridosj

02-07-2020

Synopsis

In this document, we will answer 2 questions below by taking mtcars dataset and using regression models and exploratory data analyses:

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

Data Processing

In mtcars dataset, the variable am is for the Transmission (0 = automatic, 1 = manual).

```
library(datasets);library(car);data(mtcars)
```

```
## Loading required package: carData
```

```
str(mtcars)
```

```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

From plots about mpg vs.am shown in appendix fig-1, we can see that the mean of mpg(miles per gallon) for manual transmission is bigger than mean of automatic, but when considering other variables, the difference is not constant.

Calculate the cor() between mpg and am with other variables.

```
cor(mtcars$mpg, mtcars[c(2:11)])
```

```
##            cyl       disp         hp      drat         wt     qsec        vs
## [1,] -0.852162 -0.8475514 -0.7761684 0.6811719 -0.8676594 0.418684 0.6640389
##             am      gear       carb
## [1,] 0.5998324 0.4802848 -0.5509251
```

```
cor(as.numeric(mtcars$am), mtcars[c(2:8,10:11)])
```

```
##            cyl      disp         hp      drat         wt       qsec        vs
## [1,] -0.522607 -0.591227 -0.2432043 0.7127111 -0.6924953 -0.2298609 0.1683451
##           gear       carb
## [1,] 0.7940588 0.05753435
```

The cor() result shows that mpg is correlative with all other variables, and am is too. So the am interact with other variables. And they are considered to be added into model together.

```
mtcars$am<-factor(mtcars$am)
fit1<-lm(mpg ~ am, data=mtcars)
fitall<-lm(mpg ~ ., data=mtcars);vif(fitall);summary(fitall)
```

```
##       cyl      disp        hp      drat        wt      qsec        vs        am
## 15.373833 21.620241  9.832037  3.374620 15.164887  7.527958  4.965873  4.648487
##      gear      carb
##  5.357452  7.908747
```

```
##
## Call:
## lm(formula = mpg ~ ., data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4506 -1.6044 -0.1196  1.2193  4.6271
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.30337   18.71788   0.657   0.5181
## cyl         -0.11144    1.04502  -0.107   0.9161
## disp         0.01334    0.01786   0.747   0.4635
## hp          -0.02148    0.02177  -0.987   0.3350
## drat         0.78711    1.63537   0.481   0.6353
## wt          -3.71530    1.89441  -1.961   0.0633 .
## qsec         0.82104    0.73084   1.123   0.2739
## vs           0.31776    2.10451   0.151   0.8814
## am1          2.52023    2.05665   1.225   0.2340
## gear         0.65541    1.49326   0.439   0.6652
## carb        -0.19942    0.82875  -0.241   0.8122
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.65 on 21 degrees of freedom
## Multiple R-squared:  0.869,  Adjusted R-squared:  0.8066
## F-statistic: 13.93 on 10 and 21 DF,  p-value: 3.793e-07
```

vif() bigger than 10 is cyl, disp and wt. ommit the disp for whose vif is bigger than 10 and is the highest one.

```
fit9<-lm(mpg~cyl+hp+drat+wt+qsec+vs+am+gear+carb, data=mtcars);vif(fit9);summary(fit9)
```

```
##       cyl        hp      drat        wt      qsec        vs        am      gear
## 14.284737  7.123361  3.329298  6.189050  6.914423  4.916053  4.645108  5.324402
##      carb
##  4.310597
```

```
##
## Call:
## lm(formula = mpg ~ cyl + hp + drat + wt + qsec + vs + am + gear +
##     carb, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7863 -1.4055 -0.2635  1.2029  4.4753
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.55052   18.52585   0.677   0.5052
## cyl          0.09627    0.99715   0.097   0.9240
## hp          -0.01295    0.01834  -0.706   0.4876
## drat         0.92864    1.60794   0.578   0.5694
## wt          -2.62694    1.19800  -2.193   0.0392 *
## qsec         0.66523    0.69335   0.959   0.3478
## vs           0.16035    2.07277   0.077   0.9390
## am1          2.47882    2.03513   1.218   0.2361
## gear         0.74300    1.47360   0.504   0.6191
## carb        -0.61686    0.60566  -1.018   0.3195
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.623 on 22 degrees of freedom
## Multiple R-squared:  0.8655, Adjusted R-squared:  0.8105
## F-statistic: 15.73 on 9 and 22 DF,  p-value: 1.183e-07
```

vif() bigger than 10 is cyl, ommit the cyl.

```
fit8<-lm(mpg~hp+drat+wt+qsec+vs+am+gear+carb, data=mtcars);vif(fit8)
```

```
##       hp     drat       wt     qsec       vs       am     gear     carb
## 6.015788 3.111501 6.051127 5.918682 4.270956 4.285815 4.690187 4.290468
```

vif() bigger than 5 is hp, ommit the hp.

```
fit7<-lm(mpg~drat+wt+qsec+vs+am+gear+carb, data=mtcars);vif(fit7)
```

```
##     drat       wt     qsec       vs       am     gear     carb
## 3.043073 5.104823 4.139107 4.191818 4.258479 4.688164 3.826243
```

vif() bigger than 5 is wt, ommit the wt.

```
fit6<-lm(mpg~drat+qsec+vs+am+gear+carb, data=mtcars);vif(fit6)
```

```
##     drat     qsec       vs       am     gear     carb
## 2.849229 3.753728 3.791734 3.901638 4.335401 2.456385
```

```
anova(fit1,fit6,fit7, fit8, fit9,fitall)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ drat + qsec + vs + am + gear + carb
## Model 3: mpg ~ drat + wt + qsec + vs + am + gear + carb
## Model 4: mpg ~ hp + drat + wt + qsec + vs + am + gear + carb
## Model 5: mpg ~ cyl + hp + drat + wt + qsec + vs + am + gear + carb
## Model 6: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     25 205.71  5    515.18 14.6702 3.032e-06 ***
## 3     24 155.11  1     50.60  7.2045   0.01389 *
## 4     23 151.48  1      3.64  0.5177   0.47976
## 5     22 151.41  1      0.06  0.0091   0.92477
## 6     21 147.49  1      3.92  0.5576   0.46349
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

So, wt shouldn't be omitted. omit gear.

```
fit6<-lm(mpg~drat+wt+qsec+vs+am+carb, data=mtcars);vif(fit6)
```

```
##     drat       wt     qsec       vs       am     carb
## 2.933371 4.720708 4.137804 4.042597 3.383725 2.663231
```

```
anova(fit1,fit6,fit7, fit8, fit9,fitall)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ drat + wt + qsec + vs + am + carb
## Model 3: mpg ~ drat + wt + qsec + vs + am + gear + carb
## Model 4: mpg ~ hp + drat + wt + qsec + vs + am + gear + carb
## Model 5: mpg ~ cyl + hp + drat + wt + qsec + vs + am + gear + carb
## Model 6: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     25 156.74  5    564.15 16.0647 1.472e-06 ***
## 3     24 155.11  1      1.63  0.2321    0.6349
## 4     23 151.48  1      3.64  0.5177    0.4798
## 5     22 151.41  1      0.06  0.0091    0.9248
## 6     21 147.49  1      3.92  0.5576    0.4635
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

omit qsec

```
fit5<-lm(mpg~drat+wt+vs+am+carb, data=mtcars);vif(fit5)
```

```
##     drat       wt       vs       am     carb
## 2.878459 4.260312 2.082614 3.257575 2.045591
```

```
anova(fit1,fit5,fit6,fit7, fit8, fit9,fitall)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ drat + wt + vs + am + carb
## Model 3: mpg ~ drat + wt + qsec + vs + am + carb
## Model 4: mpg ~ drat + wt + qsec + vs + am + gear + carb
## Model 5: mpg ~ hp + drat + wt + qsec + vs + am + gear + carb
## Model 6: mpg ~ cyl + hp + drat + wt + qsec + vs + am + gear + carb
## Model 7: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     26 176.13  4    544.77 19.3909 8.242e-07 ***
## 3     25 156.74  1     19.38  2.7598    0.1115
## 4     24 155.11  1      1.63  0.2321    0.6349
## 5     23 151.48  1      3.64  0.5177    0.4798
## 6     22 151.41  1      0.06  0.0091    0.9248
## 7     21 147.49  1      3.92  0.5576    0.4635
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

fit5 is selected. And fit5's R^2 is larger than 0.8 and it not bad.

Pr(>F) for model 2 is less than 0.05, and thus it fails to reject F-test. So the variables should be added into model and model fit8 is selected.

```
fitn<-lm(mpg~am+drat+wt+vs+carb, data=mtcars)
```

As the residuals plots about fitn in appendix fig-2, there is any of the patterned appearance. The residuals were independently and (almost) identically distributed with zero mean, and were uncorrelated with the fit and normality.

```
summary(fitn)$coefficients
```

```
##               Estimate Std. Error   t value    Pr(>|t|)
## (Intercept) 23.5400430  6.5641914  3.586130 0.001362001
## am1          2.4046645  1.6908215  1.422187 0.166858604
## drat         1.5451101  1.4833045  1.041668 0.307155619
## wt          -2.5470836  0.9861022 -2.582981 0.015774068
## vs           2.2689015  1.3384531  1.695167 0.101989311
## carb        -0.9889556  0.4139294 -2.389189 0.024430057
```

- For am is a two level factor, am==0 is reference, and the Intercept Estimatesv is the mean of the referenct group(am==0) and the other Estimates are the distances of the other groups' means from the reference mean. in average, mpg increase 2.4047(E(am==1)) , when changing am from 0 to 1 and holding all the other variables.

- At the same time, p-values are for tests of whether the groups are different than zero. i.e. H0: beta==0 vs. Ha: beta<>0. In Coefficients table, group(am==1): Pr(>|t|) =0.1669> 0.05, so we fail to reject H0. i.e. the mpg may be 0 when changing am from 0 to 1 and holding other variables.

Considering multivariable with interaction, change am from 0 to 1 while holding other variables constant:

E[mpgi|am=1,cyl=x2,disp=x3,...]-[mpgi|am=0,cyl=x2,disp=x3...]=beta_am+beta_hp*x2+beta_drat*x3+.

thus the expected change in per unit change in holding all else constant is not constant. Consider t test: **H0:bata_am=2.4046645 vs. Ha:btat_am<>2.4046645**, calculate the T-confidence interval:

```
m_am1<-summary(fitn)$coefficients[2]
SE<-summary(fitn)$coefficients[2,2]
t<-summary(fitn)$coefficients[2,3]
m_am1+c(-1,1)*SE*t
```

```
## [1] 0.000000 4.809329
```

The lower endpoint of T-confidence interval >= 0, so when am change from 0 to 1 and all other variables is held, mpg increase in the interval [0, 4.809]

Conclusion

1. Manual transmission is better for MPG than an automatic.
2. The MPG difference between automatic and manual transmissions is not a constant, which is in [0, 4.809] and relies other variables

Appendix

```
boxplot(mpg~am, data=mtcars,boxwex = 0.3,main="mpg vs. am")
```

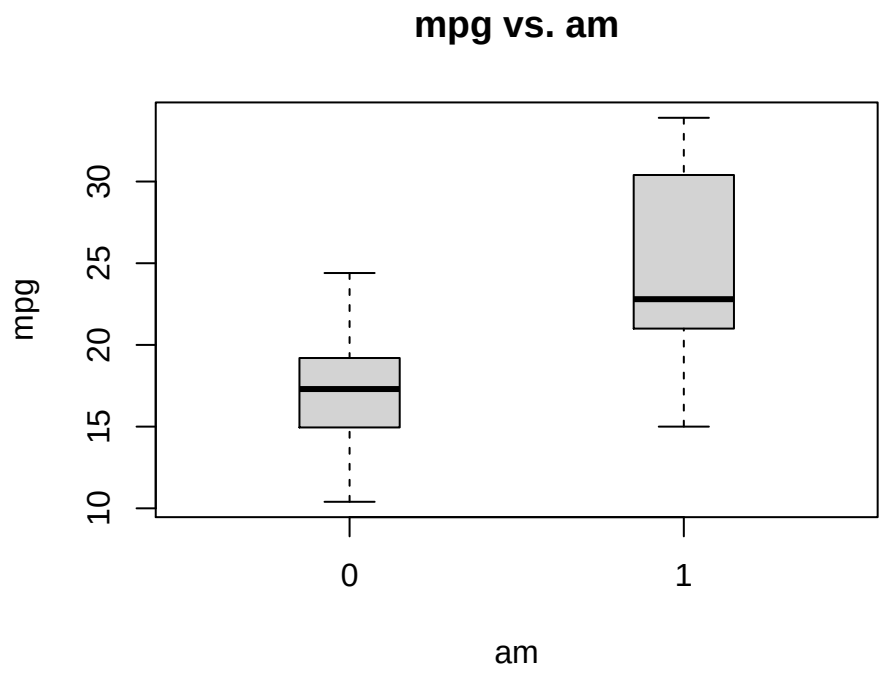
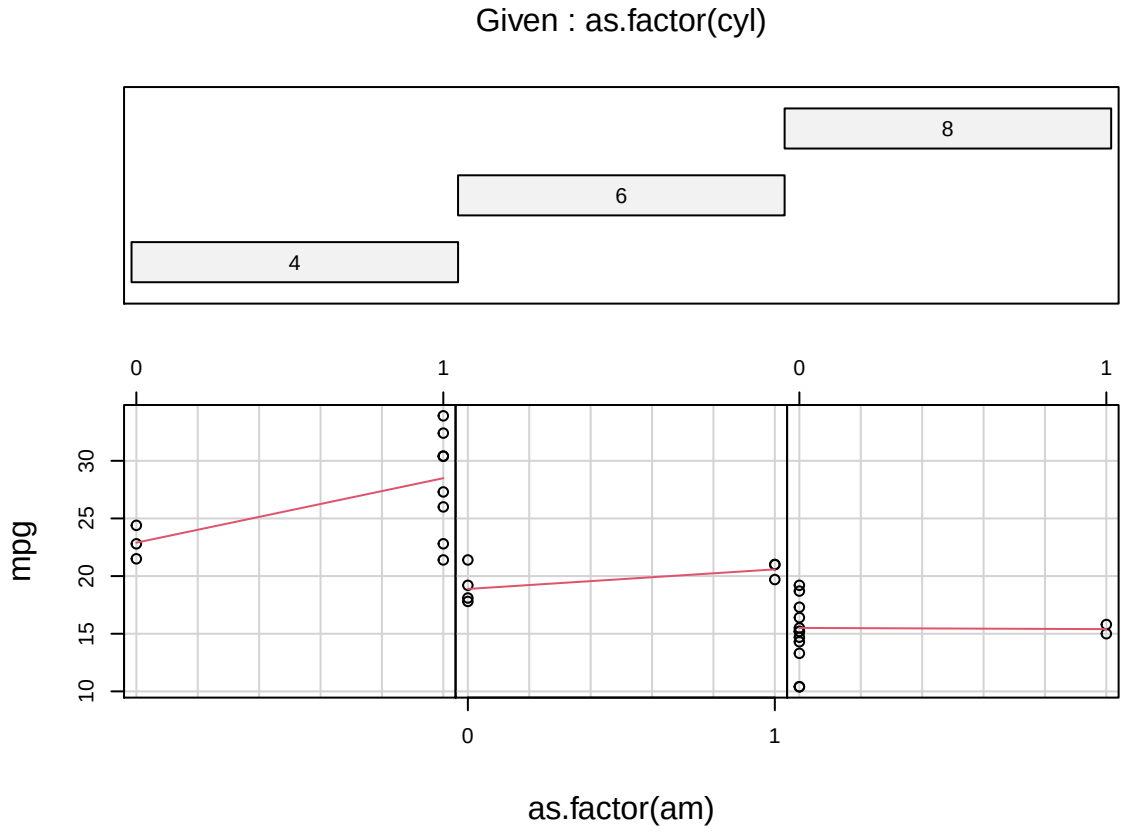


Fig-1: mpg ~ am with cyl as condition

```
coplot(mpg ~ as.factor(am) | as.factor(cyl), data = mtcars, panel = panel.smooth,row=1)
```



```
par(mfrow = c(1, 1))
rng<-round(c(-1,1)*max(abs(range(resid(fitn)))), 0)
plot(predict(fitn), resid(fitn), main="Resid vs. fit", ylim=rng)
abline(h=0, col="red", lty = 3)
abline(h=4, col="red", lty = 3)
abline(h=-4, col="red", lty = 3)
```

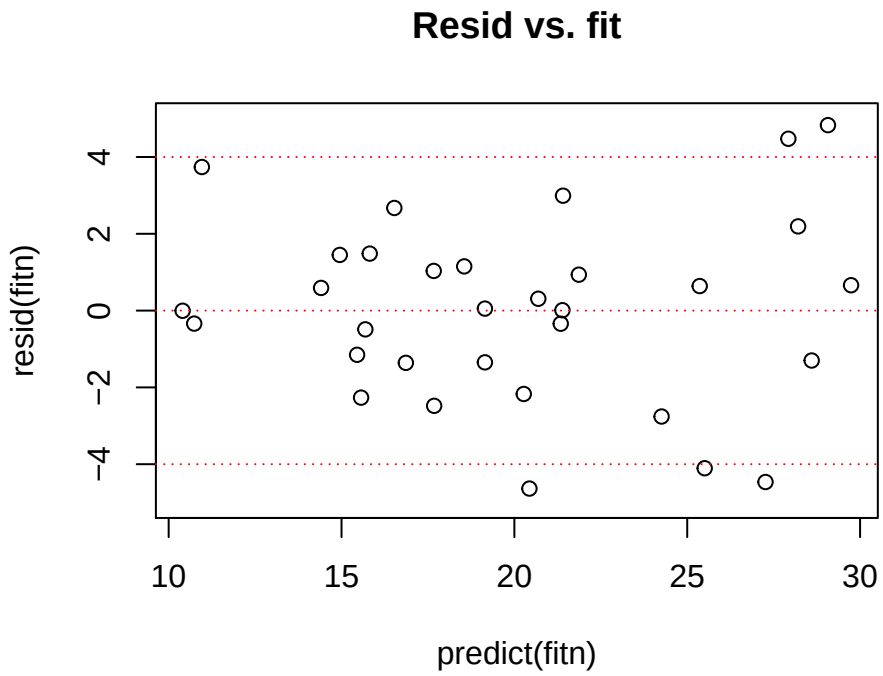
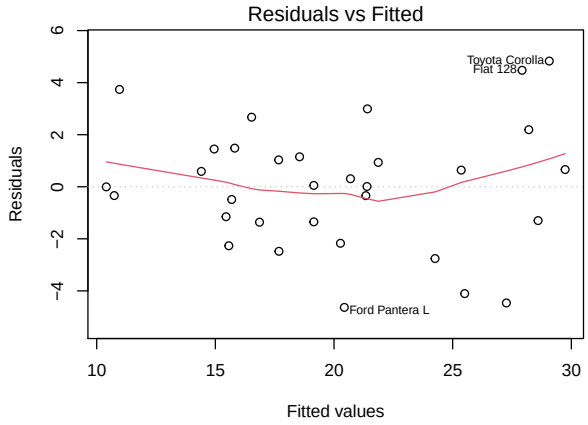
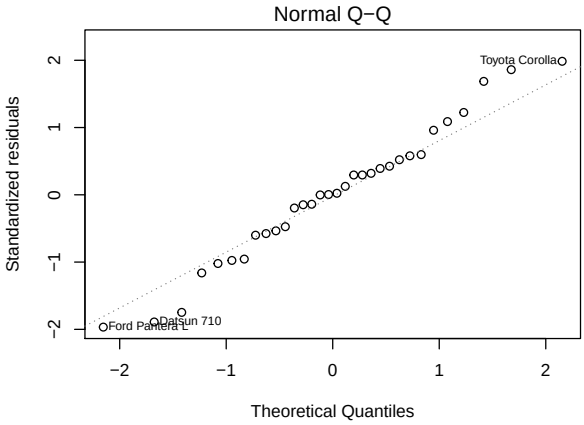
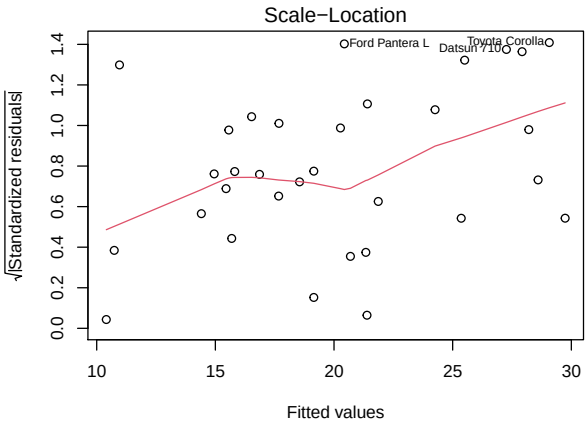


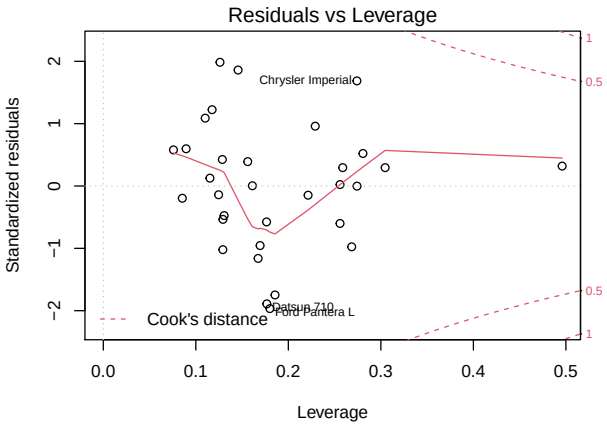
Fig-2: Residuals plot for fitn1

```
par(mfrow = c(2, 2)); plot(fitn)
```


Residuals vs Fitted
Residuals
Fitted values
Toyota Corolla
Fiat 128
Ford Pantera L


Normal Q−Q
Standardized residuals
Theoretical Quantiles
Toyota Corolla
Datsun 710
Ford Pantera L


Scale−Location
√|Standardized residuals|
Fitted values
Ford Pantera L
Datsun 710
Toyota Corolla


Residuals vs Leverage
Standardized residuals
Leverage
Chrysler Imperial
Datsun 710
Ford Pantera L
Cook's distance