# Motor Trend Analysis Report

*Arman Iskaliyev*

*10/03/2018*

## 1. Executive Summary

This report was developed for the Motor Trend Magazine to examine which type of automobile transmission has better result in mileage (miles per gallon or mpg). The report provides an analysis and answers for exploring questions. Methods of analysis include fitting regression models by using stepwise regression method, using T-test for hypothesis test. All R calculations and exploratory plots used in this report can be found in the appendices part. Results of analysis show that manual transmission type is better for mileage than automatic transmission type, and difference between automatic and manual transmissions is 1.8 MPG.

## 2. Questions of interest

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

## 3. Data Processing

### 3.1 Dataset description

The `mtcars` dataset from base R was used for this analysis. The data in dataset was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models).

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

### 3.2 Loading and exploring dataset

```
data("mtcars")
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 ...
```

```
summary(mtcars)
```

```
##       mpg             cyl             disp             hp
##  Min.   :10.40   Min.   :4.000   Min.   : 71.1   Min.   : 52.0
##  1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8   1st Qu.: 96.5
##  Median :19.20   Median :6.000   Median :196.3   Median :123.0
##  Mean   :20.09   Mean   :6.188   Mean   :230.7   Mean   :146.7
##  3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0   3rd Qu.:180.0
##  Max.   :33.90   Max.   :8.000   Max.   :472.0   Max.   :335.0
##       drat             wt             qsec             vs
##  Min.   :2.760   Min.   :1.513   Min.   :14.50   Min.   :0.0000
##  1st Qu.:3.080   1st Qu.:2.581   1st Qu.:16.89   1st Qu.:0.0000
##  Median :3.695   Median :3.325   Median :17.71   Median :0.0000
##  Mean   :3.597   Mean   :3.217   Mean   :17.85   Mean   :0.4375
##  3rd Qu.:3.920   3rd Qu.:3.610   3rd Qu.:18.90   3rd Qu.:1.0000
##  Max.   :4.930   Max.   :5.424   Max.   :22.90   Max.   :1.0000
##        am              gear            carb
##  Min.   :0.0000   Min.   :3.000   Min.   :1.000
##  1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:2.000
##  Median :0.0000   Median :4.000   Median :2.000
##  Mean   :0.4062   Mean   :3.688   Mean   :2.812
##  3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:4.000
##  Max.   :1.0000   Max.   :5.000   Max.   :8.000
```

### 3.3 Transformations

Some variables look like a categorical but they was saved as numeric. We converted them into factors manually.

```
vars_to_fctr <- c("am","cyl","carb","vs","gear")
#ds[vars_to_fctr] <- lapply(ds[vars_to_fctr], factor)
ds <- mtcars %>%
     tibble::rownames_to_column() %>%
     mutate_at(vars_to_fctr, funs(factor(.)))

levels(ds$am) <- c("automatic", "manual")
levels(ds$vs) <- c("V", "S")
str(ds)
```

```
## 'data.frame':    32 obs. of  12 variables:
##  $ rowname: chr  "Mazda RX4" "Mazda RX4 Wag" "Datsun 710" "Hornet 4 Drive" ...
##  $ mpg    : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl    : Factor w/ 3 levels "4","6","8": 2 2 1 2 3 2 3 1 1 2 ...
##  $ 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     : Factor w/ 2 levels "V","S": 1 1 2 2 1 2 1 2 2 2 ...
##  $ am     : Factor w/ 2 levels "automatic","manual": 2 2 2 1 1 1 1 1 1 1 ...
##  $ gear   : Factor w/ 3 levels "3","4","5": 2 2 2 1 1 1 1 2 2 2 ...
##  $ carb   : Factor w/ 6 levels "1","2","3","4",..: 4 4 1 1 2 1 4 2 2 4 ...
```

## 4. Exploratory Data Analysis

First we calculated correlation coefficients for our outcome variable, a `mpg`, vs. all other numeric variables.

```
ggcorr.out <- ds %>% select_if(is.numeric) %>%
     ggcorr(label = TRUE)
```

We found weak correlations in `qsec` and `drat` variables. So these variables will not be needed us in further analysis and they could be excluded from dataset.

```
ggcorr.out$data %>% filter(y == "mpg") %>% arrange(abs(coefficient)) %>% select(-label)
```

```
##      x   y coefficient
## 1 qsec mpg   0.4186840
## 2 drat mpg   0.6811719
## 3   hp mpg  -0.7761684
## 4 disp mpg  -0.8475514
## 5   wt mpg  -0.8676594
```

We also build boxplot pairs of `mpg` vs. all categorical variables. Here we saw robust relationship between `mpg ~ cyl`, `mpg ~ am` and `mpg ~ vs`, but there are some outliers here, in 8-cylinder observation (Cadillac Fleetwood, Lincoln Continental, Pontiac Firebird) and V type engine observation (Porsche 914-2). We decided to exclude `carb` and `gear` variables.

```
var_to_drop <- c("qsec", "drat", "gear", "carb")
ds %>% filter((vs == "V" & mpg > 25) | ((cyl == "8" & mpg < 12) | (cyl == "8" & mpg > 19)))
```

```
##               rowname  mpg cyl  disp  hp drat    wt  qsec vs        am
## 1  Cadillac Fleetwood 10.4   8 472.0 205 2.93 5.250 17.98  V automatic
## 2 Lincoln Continental 10.4   8 460.0 215 3.00 5.424 17.82  V automatic
## 3    Pontiac Firebird 19.2   8 400.0 175 3.08 3.845 17.05  V automatic
## 4       Porsche 914-2 26.0   4 120.3  91 4.43 2.140 16.70  V    manual
##   gear carb
## 1    3    4
## 2    3    4
## 3    3    2
## 4    5    2
```

```
ds <- ds %>% select(-one_of(var_to_drop), -rowname)
```

## 5. Question 1. Is an automatic or manual transmission better for MPG

As we just have seen in exploratory plot, there is strong relationship between mileage and transmission type, and manual transmission type looks better for MPG. To prove our assumption we conducted two-sided T-test.

```
(ttest.out <- t.test(mpg~am, data = ds))
```

```
##
##  Welch Two Sample t-test
##
## data:  mpg by am
## t = -3.7671, df = 18.332, p-value = 0.001374
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.280194  -3.209684
## sample estimates:
## mean in group automatic    mean in group manual
##                17.14737                24.39231
```

**Interpretation**

So calculated p-value 0.0013736 allow as to reject null hypothesis, and prove that manual transmission is better for MPG.

## 6. Question 2. Quantify the MPG difference between automatic and manual transmissions

### 6.1. Fit multiple models and Best model selection

Here we use stepwise regression method to fit multiple models and choose best model. We started with multivariate model including all variables.

Fit initial model:

```
mod.init <- lm(mpg ~ ., data = ds)
summary(mod.init)
```

```
##
## Call:
## lm(formula = mpg ~ ., data = ds)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.3385 -1.3302 -0.0046  1.1568  4.6601
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 31.341721   3.519033   8.906 4.49e-09 ***
## cyl6        -2.194219   1.691573  -1.297   0.2069
## cyl8        -0.255244   3.641927  -0.070   0.9447
## disp         0.003996   0.012710   0.314   0.7559
## hp          -0.035108   0.014121  -2.486   0.0203 *
## wt          -2.610044   1.176407  -2.219   0.0362 *
## vsS          1.986340   1.792830   1.108   0.2789
## ammanual     2.699155   1.628182   1.658   0.1104
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.442 on 24 degrees of freedom
## Multiple R-squared:  0.8729, Adjusted R-squared:  0.8359
```

```
## F-statistic: 23.55 on 7 and 24 DF,  p-value: 2.73e-09
```

Apply step function with backward direction to exclude variables one at time from formula, and to fit them. At the end of the function it give us best fitted model:

```
mod.best <- step(mod.init, direction = "backward", trace = 0)
summary(mod.best)
```

```
##
## Call:
## lm(formula = mpg ~ cyl + hp + wt + am, data = ds)
##
## 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 ***
## cyl6        -3.03134    1.40728  -2.154  0.04068 *
## cyl8        -2.16368    2.28425  -0.947  0.35225
## hp          -0.03211    0.01369  -2.345  0.02693 *
## wt          -2.49683    0.88559  -2.819  0.00908 **
## ammanual     1.80921    1.39630   1.296  0.20646
## ---
## 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
```

### Interpretation

According to the results of the stepwise method the best regression model for mpg includes all these variables:

- cyl
- hp
- wt
- am

This model has a quite high R-squared value equal to 0.8658799. Our model show that manual transmission type is 1.80 better than automatic.

## 7. Conclusion

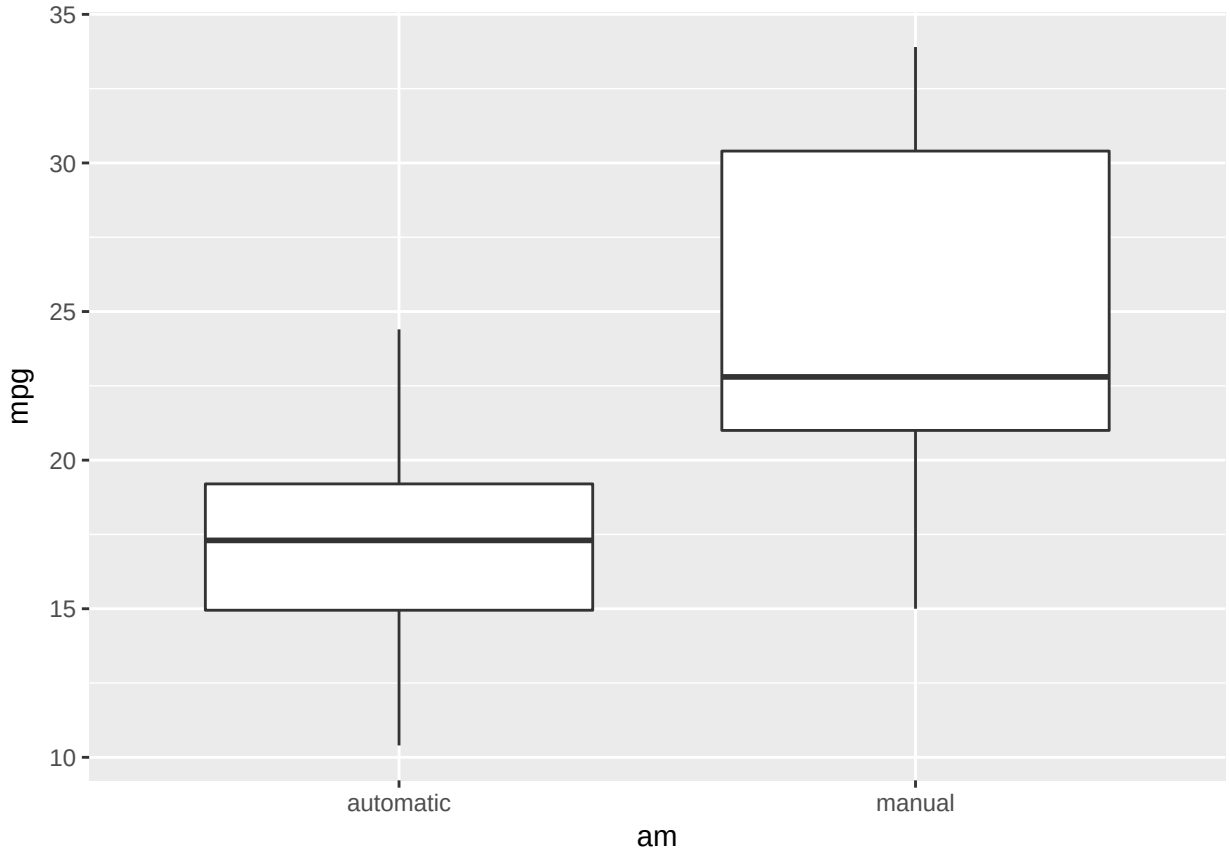
- Manual transmission type is better for mileage than automatic
- Manual transmission type is better for mileage by a factor of 1.8 than automatic

## Appendices

### Exploratory plots

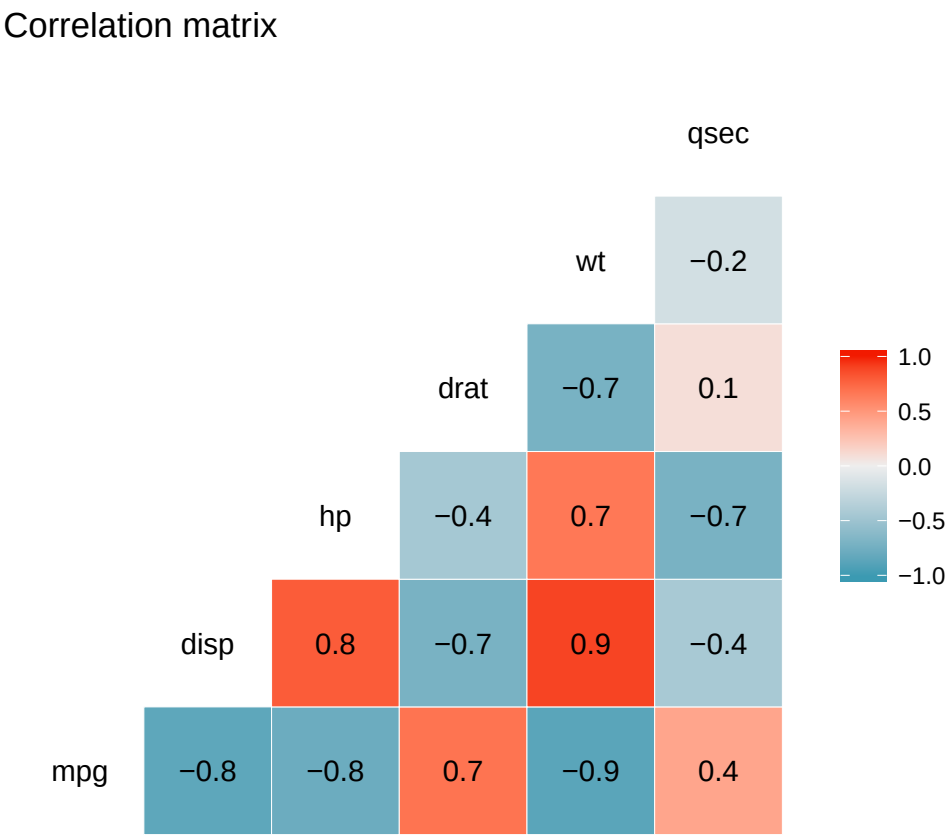
Relationship between mileage and transmission type.

```
qplot(x = am, y = mpg, data = ds, geom = "boxplot")
```




Correlation of mpg vs. all numerical variables.

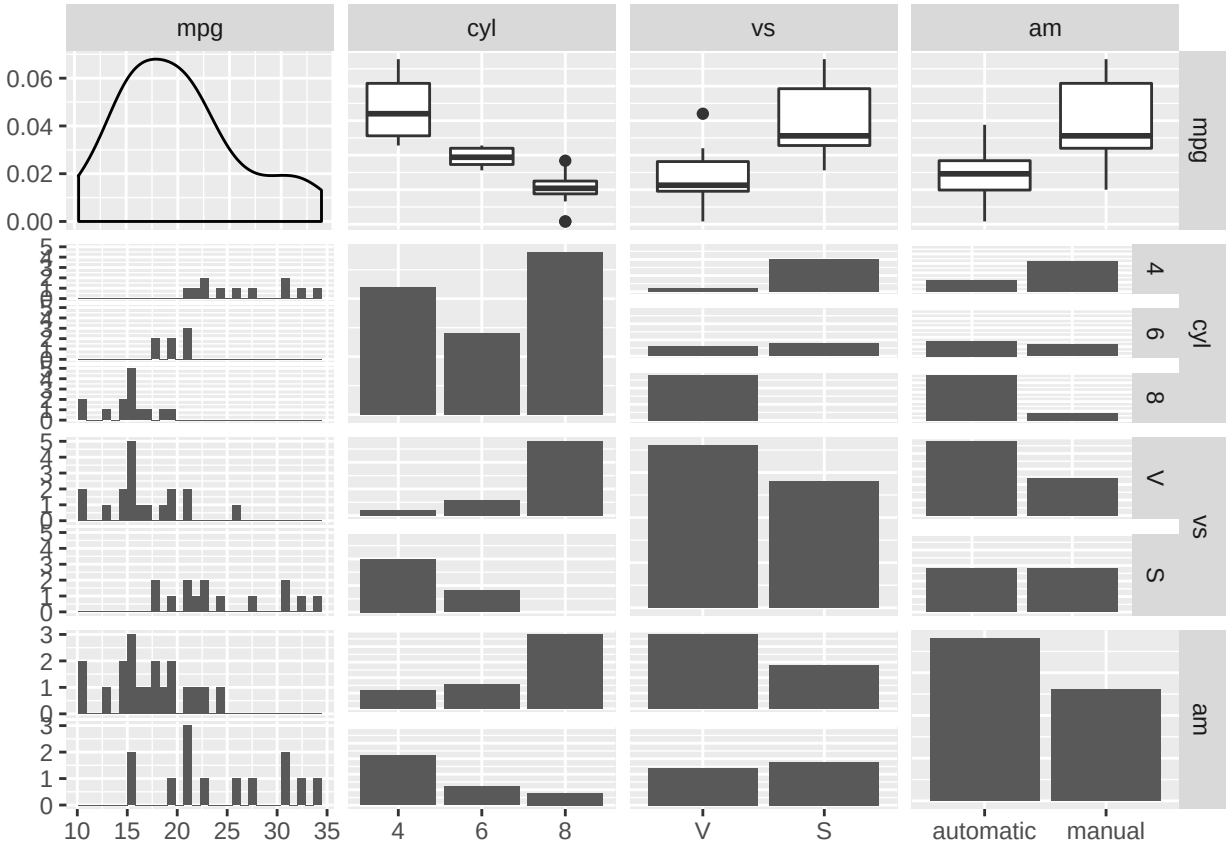
```
ggcorr.out + ggtitle("Correlation matrix")
```




Matrix of scatterplots mpg vs. all categorical variables.

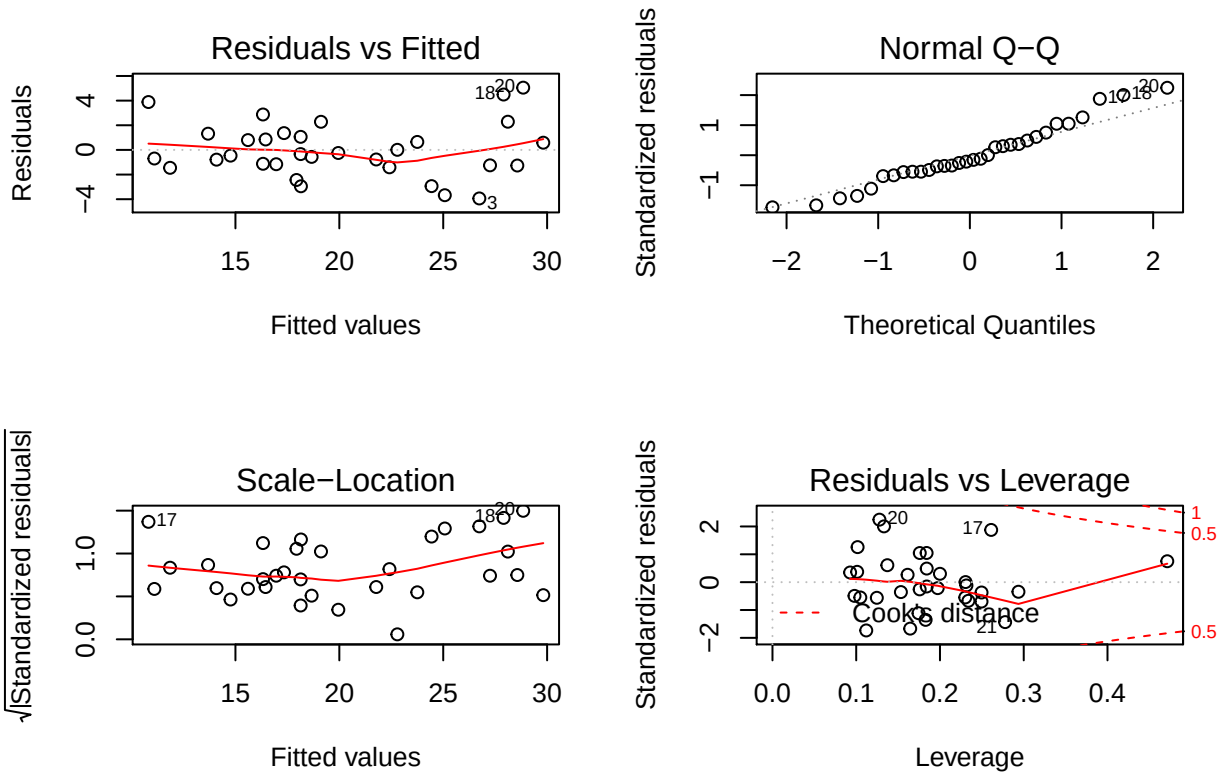
```
ds %>% select_if(is.factor) %>%
     bind_cols(select(ds, mpg)) %>%
     select(mpg, everything()) %>%
     ggpairs()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




Residual plots of best model

```
par(mfrow=c(2,2))
plot(mod.best)
```


Residuals vs Fitted
Residuals
Fitted values
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage