Exercises 4 solutions

1. Model assumptions are violated for datasets 6 (t_3-distributed error), 7 (y is dependent on x_3^3 and x_4^3) and 12 (exponentially distributed error). The code that produced these datasets is available from the course Moodle page.

2. We have a response variable y and the model formalises an idea how this y depends on some explanatory variables $x_1, \ldots, x_p$. There are several assumptions. The first assumption is that there is some linear relationship, i.e., that, apart from some unsystematic random variation, y is a linear function of $x_1, \ldots, x_p$.

   Second, the random variation is assumed to work in the same way everywhere, regardless of what values the x-variables take or when observations have been made, i.e. the linear function that relates y to $x_1, \ldots, x_p$ contains the full information that the x-variables can give us about y.

   Third, the random variation for different observations is assumed to be unrelated. This assumption (as well as the previous one) is problematic if there are some systematic effects such as developments over time (if this is not included as x-variable) or some connection between "neighbouring" observations may be going on.

   Fourth, the random variation is assumed to have a "normal" distributional shape. This means that in most cases the random deviation from the linear function is expected to be rather small. Also, random variation is expected to yield observations below the assumed true regression function in the same way (as often, and of the same size on average) as above it, the larger the random deviation from the function is, the less likely, and extreme outliers (observations far away from the assumed model function) are assumed to happen very, very rarely (virtually never, unless the dataset is very, very large).

   (This question is a difficult one, even for me; probably most non-statisticians would find my explanation still too complicated.)

3. I have produced a **matrix plot** first, which indicates a nonlinear relationship between mpg and displacement, horsepower, weight, acceleration. Thinking about its functional shape, I decided to invert mpg to "gallons per mile" gpm=1/mpg (by the way, it is European standard to measure fuel consumption as consumption per distance while the American standard is distance per consumption). This produces a much nicer matrix plot.

   Furthermore, there is seemingly some strong linear dependence between displacement, horsepower, weight (though I won't do too much about this) and some potential outliers and leverage points.

   My **first model** fit is a LS regression for

   $$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6} + \beta_7 x_{i7} + e_i$$

   where $e_i \sim \mathcal{N}(0, \sigma^2)$ independently, $i = 1, \ldots, 390$. The variables should be defined but it suffices to make reference to the table defining the variables but $Y_i = 1/mpg_i$. Note that when asked to write down the model, it is wrong to replace the parameters by their estimated values (because it's part of the model that we do not know the true values of the parameters precisely).

   The R^2 for this model is 88.45%, which is not bad. (Note that the model with untransformed mpg - not shown - only yields 81%.) Diagnostic plots for this model reveal heteroscedasticity along the fitted values and some of the x-variables, obvious non-normality in the tails of the residual distribution, a dodgy leverage point no. 28, some out-of-line behaviour for the 3-cylinder cars. and a slightly and irregularly nonlinear behaviour along model year (which probably cannot be repaired). Note that the residuals vs. order plot, though shown, is useless because cars are ordered by model year (you have to look up the data file to find this).

In order to tackle the heteroscedasticity, I do something radical and transform everything to logarithms (most things have looked roughly linear before and therefore transforming only to Y to logarithms would have introduced nonlinearity). Note that when logarithmising, it does not make a difference whether mpg or gpm is used as the response, because log(mpg) is -log(gpm).

The matrix plot for this looks even better. So I now fit as **Model 2**

$$\log Y_i = \beta_0 + \beta_1 \log x_{i1} + \beta_2 \log x_{i2} + \beta_3 \log x_{i3} + \beta_4 \log x_{i4} + \beta_5 \log x_{i5} + \beta_6 \log x_{i6} + \beta_7 \log x_{i7} + e_i$$

under the same assumptions as before (note that this means that e_i is no longer the same as above).

This improves the R^2 a bit and while there is still some worry about the tails of the residual distribution (qq-plot), the heteroscedasticity is pretty much gone and the Cook's distances look harmless now. A clear improvement. Some slight nonlinearity could be seen in disreplacement and model year.

I still try out an **MM-estimator** as my **Model 3** (so the same model is fitted again, but more robustly). This is justified because the error distribution seems to have heavier tails than the normal.

Of course, this doesn't make the qq-plot appear any more normal and rather shown more observations as belonging to the heavy tails than before, but an interesting feature is that disreplacement looks more linear now (except of an outlier) and also the pattern along model year has become clearer as strange behaviour for a group of cars with model year 1980 (and some other lower outliers for the later years). The robustness weights show that there is no clear distinction between extreme outliers and "good data" but many cars cause some heavier deviation from the regression than would be expected under the normal. The 3-cylinder group still is not in line but generally this is my **recommended fitted model** because it rather looks like "linear plus some irregular features" which can even be quite well described (for which robust estimators are suitable), while model 2 suggested some nonlinearity. Also the estimated residual standard error is lower than before though this is expected under even slight deviations from normality.

Report for institute

It has proven suitable to model the fuel consumption in gallons per mile as a product of the other variables. Generally, the predictive power of the model is quite strong. This has led to the following estimations:

- Multiplying cylinders by 2 means that, on average, gallons by mile has to be multiplied by $1.064872 = 2^{0.09068}$. Note that this effect is weak and could be due to random variation.
- Multiplying displacement by 2 means that, on average, gallons by mile has to be multiplied by 1.061813. Note that this effect is weak and could be due to random variation.
- Multiplying horsepower by 2 means that, on average, gallons by mile has to be multiplied by 1.172136. This is a strongly significant effect.
- Multiplying weight by 2 means that, on average, gallons by mile has to be multiplied by 1.448942. This is a strongly significant effect.
- Multiplying acceleration by 2 means that, on average, gallons by mile has to be multiplied by 1.181508. This is a strongly significant effect.
- Multiplying modelyear by 2 (which doesn't make real sense but may give you an idea of the strength of the trend) means that, on average, gallons by mile has to be multiplied by 0.2379631. This is a strongly significant effect.

All these effects are to be understood under the assumptions that all other variables are held fixed. This may be possible but difficult to achieve for new cars. In the sample, it has to be noted, that there is a clear positive relationship between cylinders, displacement, horsepower and weight, and a negative relationship with acceleration, which may confound to some extent some of the above estimates. Of course, the model year cannot be controlled anyway and was just included for informative reasons. It is a positive finding, at least, that fuel consumption seems to go down on average, so awareness for this necessity rises.

Note that the cars with 3 cylinders are not in line with these findings and have a higher fuel consumption (though there are only four of them in the sample) than would be expected under the model. Also, fuel consumption for a group of cars from 1981 was already lower than expected from the model, so hopefully the positive trend goes on.

Some general comments (again from the 2008 ICA and therefore partly with reference to robust estimation)

- It is important in such a situation to report and interpret the estimated parameter values and to comment on whether they are significant ("there is no clear evidence that the effect is real but it could have been caused by random variation alone" etc.) and not just their signs. Your clients, though not statisticians, still have a technical, somewhat scientific interest in the precise sizes of the effects.
- It makes sense to consider the robustness weights because they give the best information about "outlyingness". However, it's not so clear to decide against which variable they should be plotted. It is good to know the case numbers of potential outliers. Sometimes the people who collected the data can check what went on with these observations, and even if not, they can be identified in other plots. But it may also be informative to know whether observations with some specific values on another variable (for example cars with 3 cylinders) form groups of outliers, for which one would then not apply the model when it comes to prediction, for example.

```
autoc <- read.table("auto.dat",header=TRUE)

pairs(autoc)

# Transform mpg to gpm

autoc$gpm <- 1/autoc$mpg
pairs(autoc[,c(8,2:7)])

# Model 1

autolm <- lm(gpm~cylinders+displacement+horsepower+weight+acceleration+
             modelyear,data=autoc)

summary(autolm)

Call:
lm(formula = gpm ~ cylinders + displacement + horsepower + weight +
    acceleration + modelyear, data = autoc)

Residuals:
       Min         1Q     Median         3Q        Max
-0.0152885 -0.0033503 -0.0001614  0.0027882  0.0241131
```

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   8.956e-02  7.921e-03  11.306  < 2e-16 ***
cylinders     1.466e-03  5.540e-04   2.646  0.00848 **
displacement -2.024e-05  1.265e-05  -1.600  0.11045
horsepower    1.067e-04  2.381e-05   4.480 9.87e-06 ***
weight        1.156e-05  1.206e-06   9.586  < 2e-16 ***
acceleration  2.850e-04  1.739e-04   1.639  0.10204
modelyear    -1.262e-03  8.750e-05 -14.420  < 2e-16 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 0.005698 on 383 degrees of freedom
Multiple R-squared: 0.8845,     Adjusted R-squared: 0.8827
F-statistic: 488.8 on 6 and 383 DF,  p-value: < 2.2e-16


par(mfrow=c(2,2))
plot(autolm,which=1:4,ask=FALSE)

plot(autoc$cylinders,residuals(autolm))
plot(autoc$displacement,residuals(autolm))
plot(autoc$horsepower,residuals(autolm))
plot(autoc$weight,residuals(autolm))

plot(autoc$acceleration,residuals(autolm))
plot(autoc$modelyear,residuals(autolm))
plot(1:nrow(autoc),residuals(autolm),type="l",xlab="Order")

# log transform

lautoc <- log(autoc)

pairs(lautoc[,c(8,2:7)])

# Model 2

lautolm <- lm(gpm˜cylinders+displacement+horsepower+weight+acceleration+
             modelyear,data=lautoc)

summary(lautolm)

Call:
lm(formula = gpm ˜ cylinders + displacement + horsepower + weight +
    acceleration + modelyear, data = lautoc)

Residuals:
      Min        1Q    Median        3Q       Max
-0.406026 -0.063443 -0.001284  0.063282  0.412744

Coefficients:
```

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.03279    0.65308  -0.050  0.95998
cylinders     0.05438    0.06012   0.905  0.36623
displacement  0.05205    0.05127   1.015  0.31062
horsepower    0.26676    0.05919   4.507 8.75e-06 ***
weight        0.56815    0.08690   6.538 2.00e-10 ***
acceleration  0.17551    0.06072   2.891  0.00406 **
modelyear    -2.22844    0.13226 -16.849  < 2e-16 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 0.114 on 383 degrees of freedom
Multiple R-squared: 0.8892,     Adjusted R-squared: 0.8875
F-statistic: 512.5 on 6 and 383 DF,  p-value: < 2.2e-16


par(mfrow=c(2,2))
plot(lautolm,which=1:4,ask=FALSE)

plot(autoc$cylinders,residuals(lautolm))
plot(autoc$displacement,residuals(lautolm))
plot(autoc$acceleration,residuals(autolm))
plot(autoc$modelyear,residuals(autolm))

# Model 3

library(robustbase)
rlautolm <- lmrob(gpm~cylinders+displacement+horsepower+weight+acceleration+
             modelyear,data=lautoc)
summary(rlautolm)

Call:
lmrob(formula = gpm ~ cylinders + displacement + horsepower +
    weight + acceleration + modelyear, data = lautoc)

Weighted Residuals:
      Min        1Q    Median        3Q       Max
-0.450505 -0.063154 -0.001194  0.060842  0.462763

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.69007    0.67671  -1.020 0.308499
cylinders     0.09068    0.06118   1.482 0.139068
displacement  0.08653    0.06655   1.300 0.194278
horsepower    0.22914    0.07665   2.989 0.002976 **
weight        0.53500    0.11461   4.668 4.21e-06 ***
acceleration  0.24063    0.07028   3.424 0.000685 ***
modelyear    -2.07119    0.12609 -16.426  < 2e-16 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Robust residual standard error: 0.09353
```

```
Convergence in 20 IRWLS iterations

Robustness weights:
 2 observations c(110,380) are outliers with |weight| = 0 ( < 0.00026);
 32 weights are ~= 1. The remaining 356 ones are summarized as
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.1044  0.8635  0.9546  0.8799  0.9843  0.9990
Algorithmic parameters:
tuning.chi         bb tuning.psi refine.tol    rel.tol
 1.5476400  0.5000000  4.6850610  0.0000001  0.0000001
 nResample     max.it     groups    n.group   best.r.s   k.fast.s      k.max
       500         50          5        400          2          1        200
 trace.lev compute.rd
         0          0
seed : int(0)

plot(rlautolm,which=c(1,2,4,5),ask=FALSE)

plot(lautoc$cylinders,residuals(rlautolm))
plot(lautoc$displacement,residuals(rlautolm))
plot(lautoc$modelyear,residuals(rlautolm))
plot(1:nrow(autoc),rlautolm$weights)
```

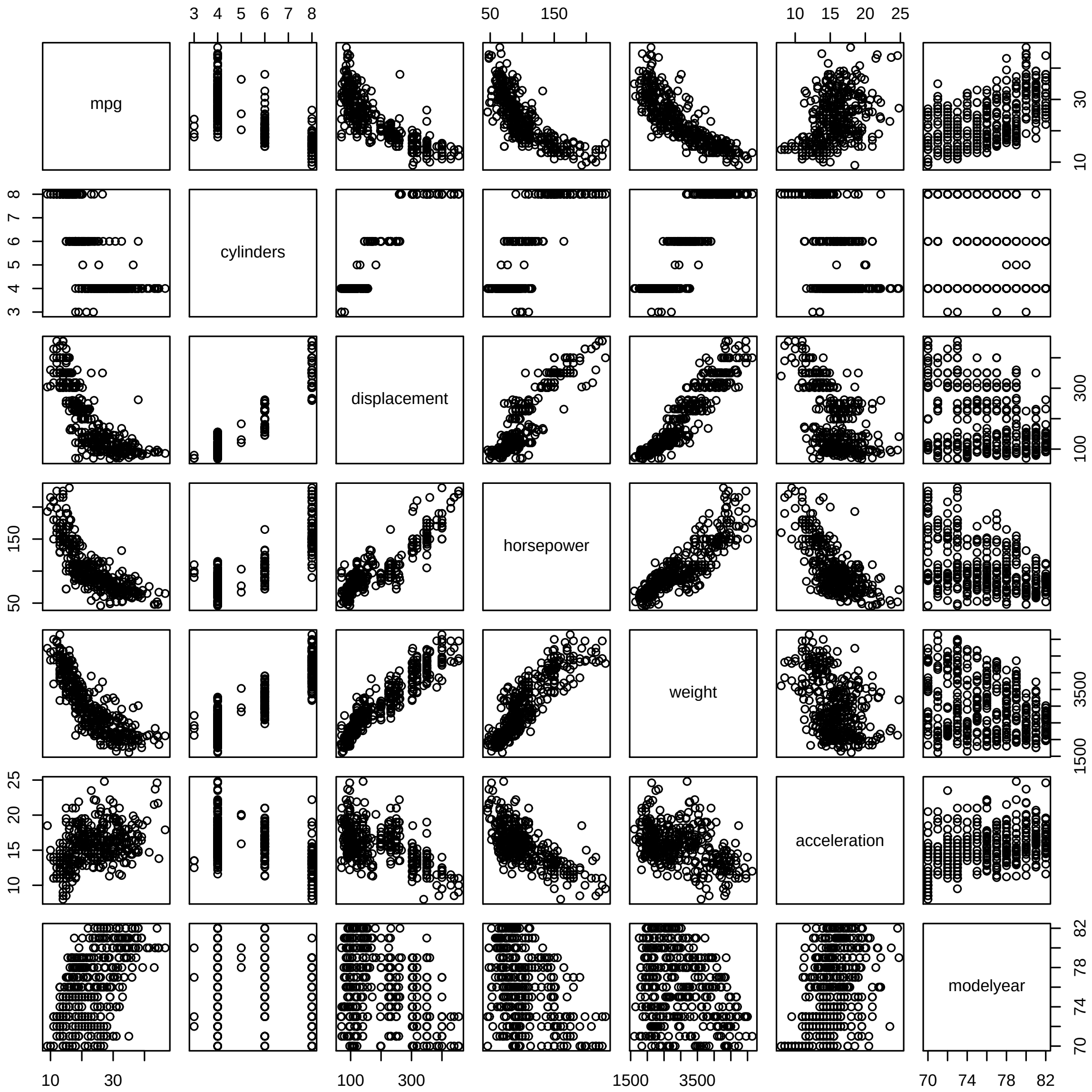

Figure 1: Matrix plot for raw data.

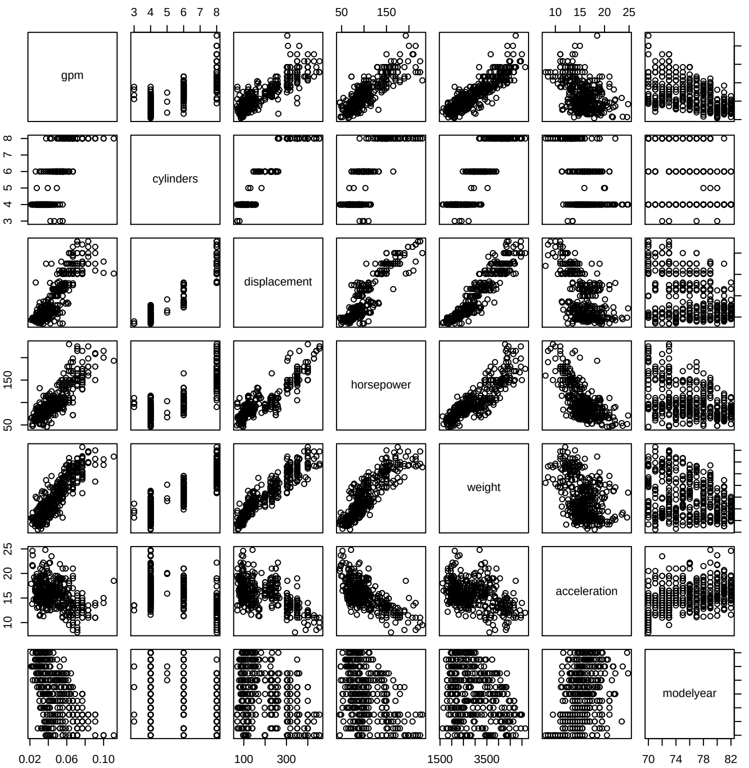

Figure 2: Matrix plot for data with gpm=1/mpg.

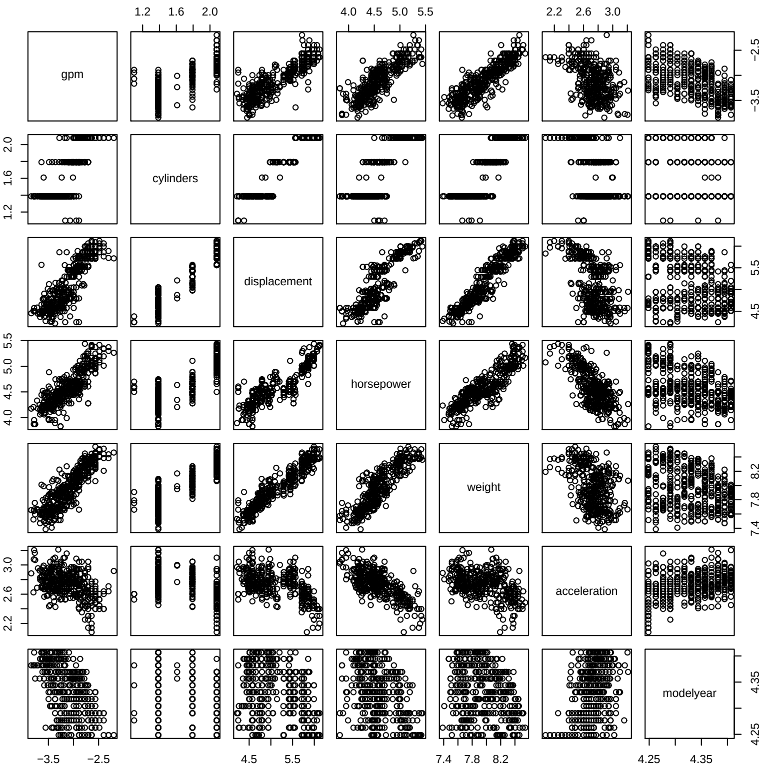

Figure 3: Matrix plot for logarithmised data with gpm=1/mpg.

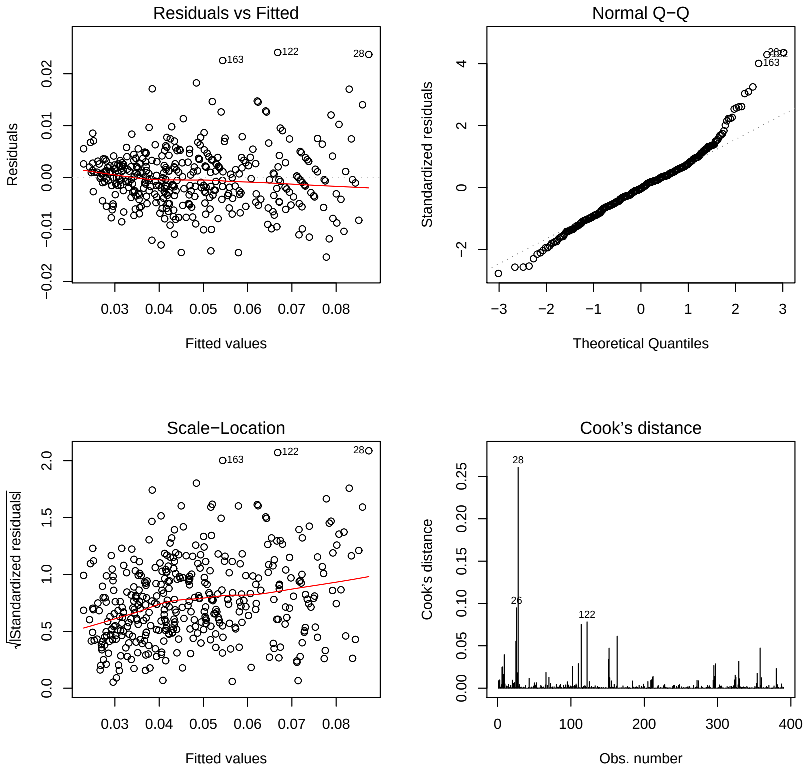

Figure 4: Standard residual plots for model 1 / LS.

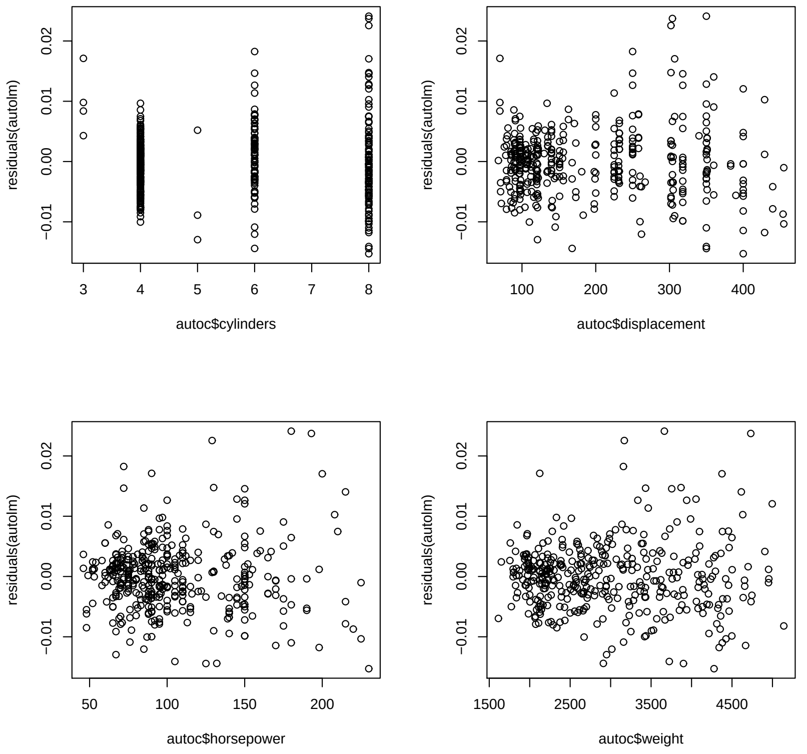

Figure 5: Residuals vs. x-variables plots for model 1 / LS.

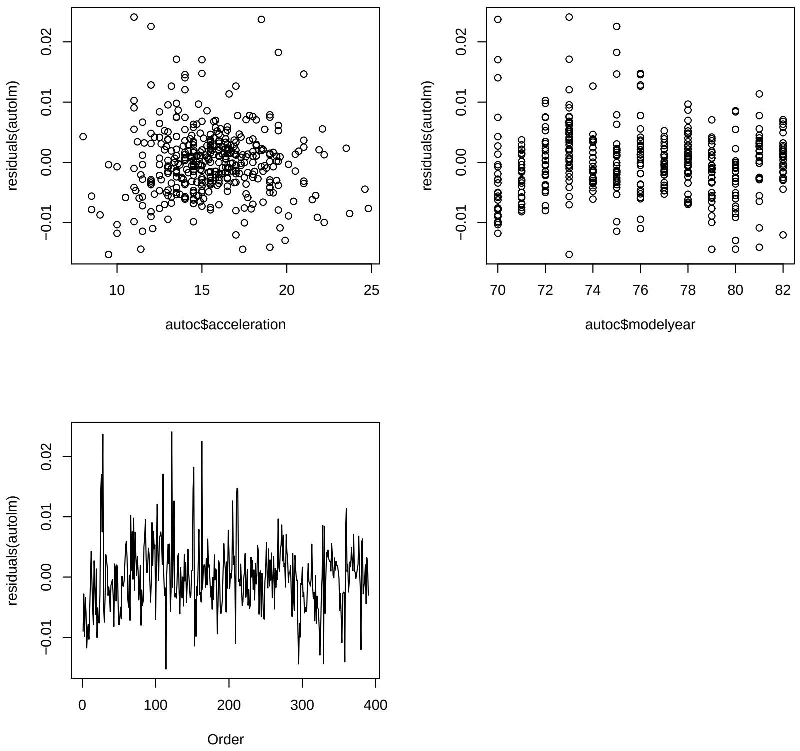

Figure 6: Residuals vs. x-variables and order for model 1 / LS.

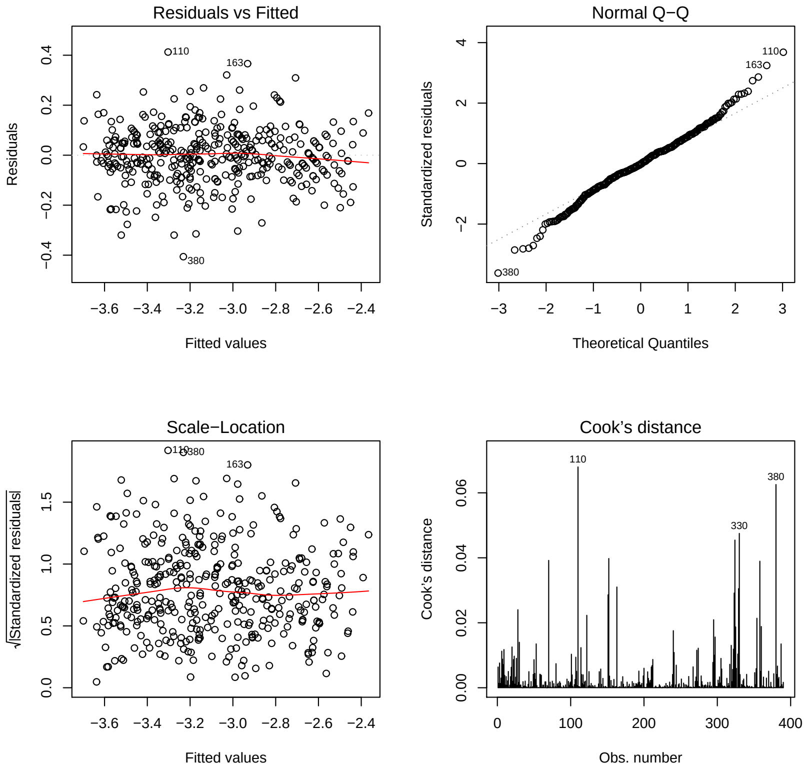

Figure 7: Standard residual plots for model 2 / LS with logs.

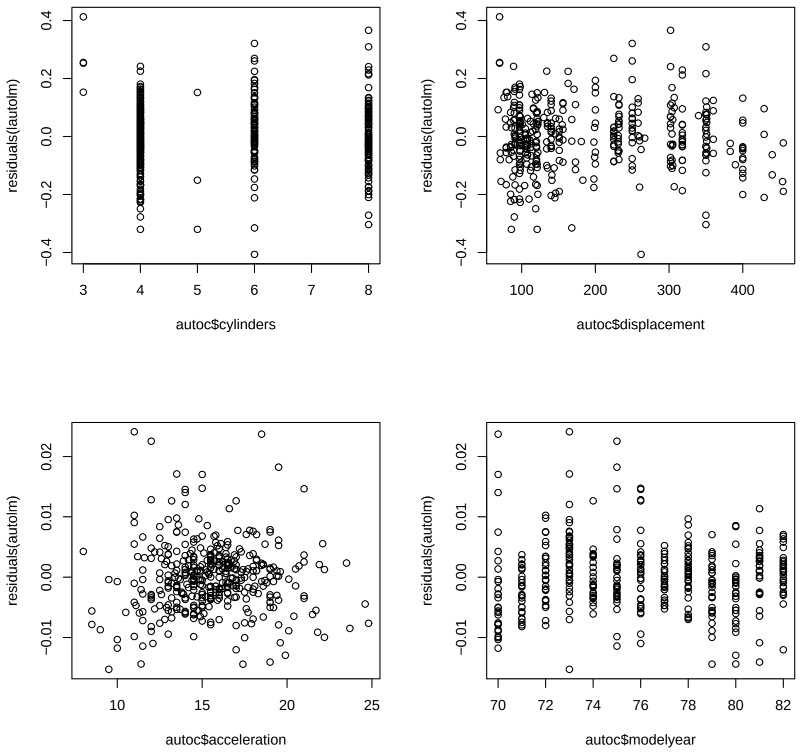

Figure 8: Some residual vs. x plots for model 2 / LS with logs.

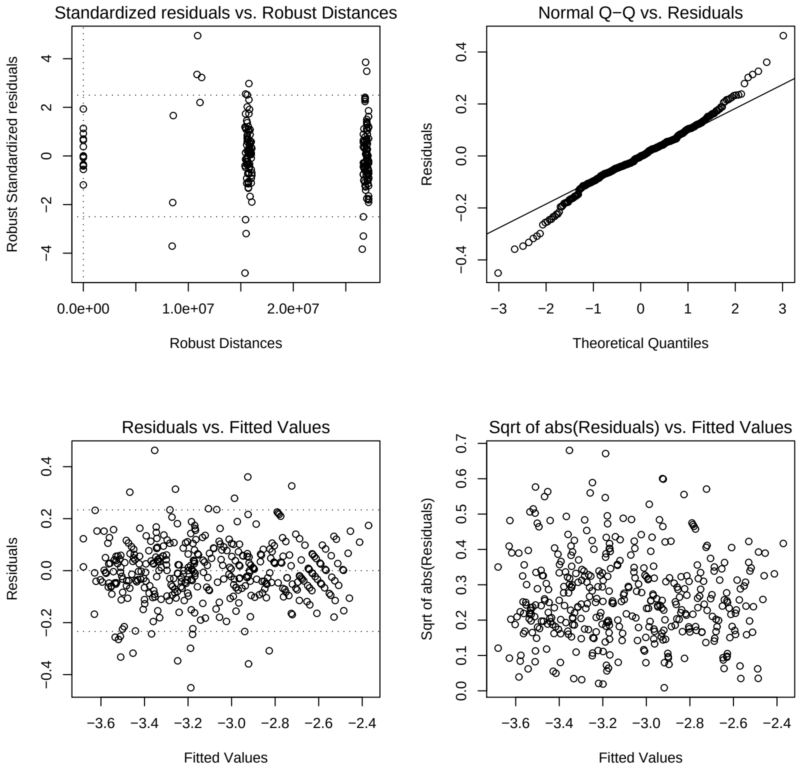

Figure 9: Standard residual plots for model 3 / MM with logs.

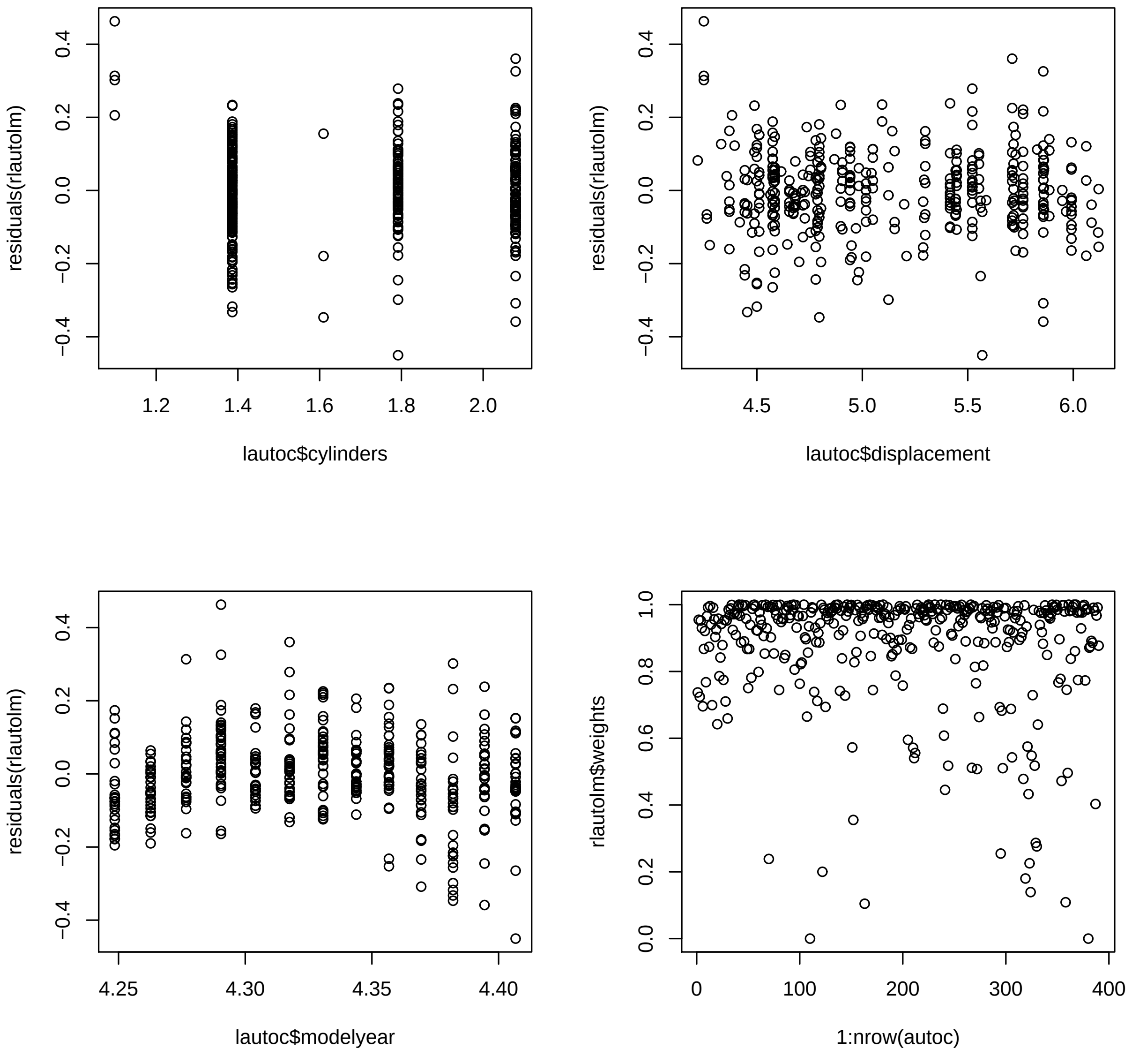

Figure 10: Some residual vs x plots and robustness weights for model 3 / MM with logs.