STAT 500: HW5

Jasmine Mou

11/2/2017

1. Using the `stackloss` data, fit a model with `stack.loss` as the response and the other three variables as predictors using the following methods:

```
data(stackloss,package="datasets")
```

(a) Least squares

```
lm_stackloss <- lm(stack.loss~., data=stackloss)
summary(lm_stackloss)
```

```
##
## Call:
## lm(formula = stack.loss ~ ., data = stackloss)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.2377 -1.7117 -0.4551  2.3614  5.6978
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -39.9197    11.8960  -3.356  0.00375 **
## Air.Flow      0.7156     0.1349   5.307  5.8e-05 ***
## Water.Temp    1.2953     0.3680   3.520  0.00263 **
## Acid.Conc.   -0.1521     0.1563  -0.973  0.34405
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.243 on 17 degrees of freedom
## Multiple R-squared:  0.9136, Adjusted R-squared:  0.8983
## F-statistic:  59.9 on 3 and 17 DF,  p-value: 3.016e-09
```

(b) Least absolute deviations

```
library(quantreg)
rq_stackloss <- rq(stack.loss~., data=stackloss)
summary(rq_stackloss)
```

```
##
## Call: rq(formula = stack.loss ~ ., data = stackloss)
##
## tau: [1] 0.5
##
## Coefficients:
##             coefficients lower bd  upper bd
## (Intercept) -39.68986    -41.61973 -29.67754
## Air.Flow      0.83188      0.51278   1.14117
## Water.Temp    0.57391      0.32182   1.41090
## Acid.Conc.   -0.06087     -0.21348  -0.02891
```

(c) Huber method

```
library(MASS)
rlm_stackloss <- rlm(stack.loss~., data=stackloss)
summary(rlm_stackloss)
```

```
##
## Call: rlm(formula = stack.loss ~ ., data = stackloss)
## Residuals:
##      Min       1Q   Median       3Q      Max
## -8.91753 -1.73127  0.06187  1.54306  6.50163
##
## Coefficients:
##             Value    Std. Error t value
## (Intercept) -41.0265   9.8073    -4.1832
## Air.Flow      0.8294   0.1112     7.4597
## Water.Temp    0.9261   0.3034     3.0524
## Acid.Conc.   -0.1278   0.1289    -0.9922
##
## Residual standard error: 2.441 on 17 degrees of freedom
```

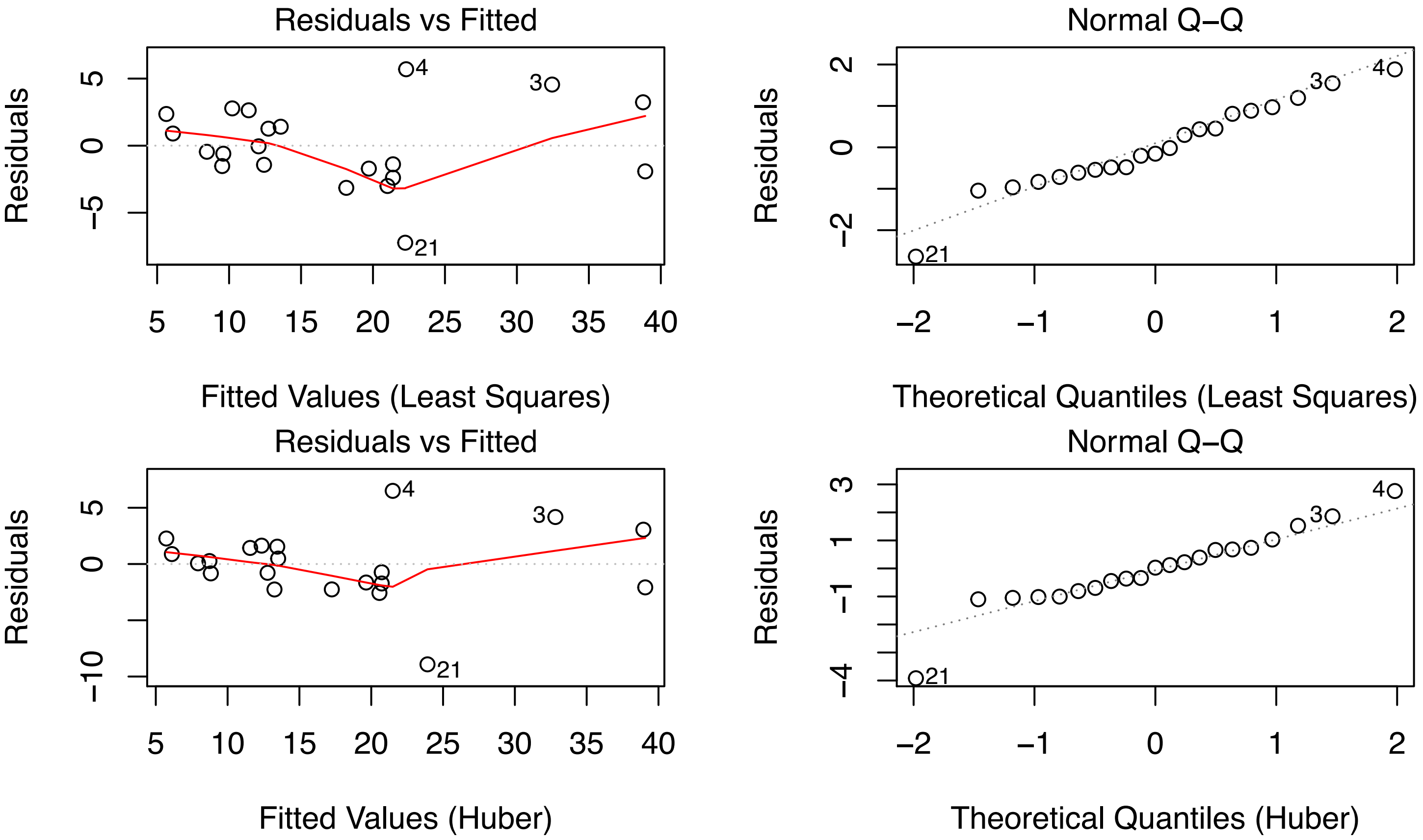
(d) Least trimmed squares

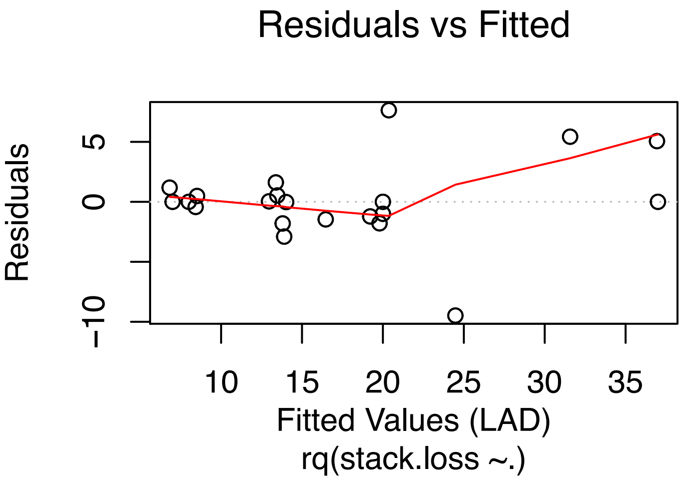
```
ltsreg_stackloss <- ltsreg(stack.loss~., data=stackloss, nsamp="exact")
ltsreg_stackloss
```

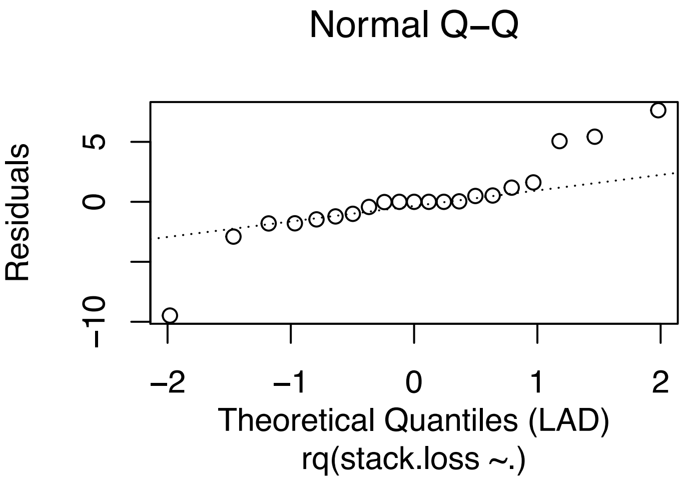
```
## Call:
## lqs.formula(formula = stack.loss ~ ., data = stackloss, nsamp = "exact",
##     method = "lts")
##
## Coefficients:
## (Intercept)     Air.Flow   Water.Temp   Acid.Conc.
##  -3.581e+01    7.500e-01    3.333e-01    3.489e-17
##
## Scale estimates 0.8482 0.8645
```

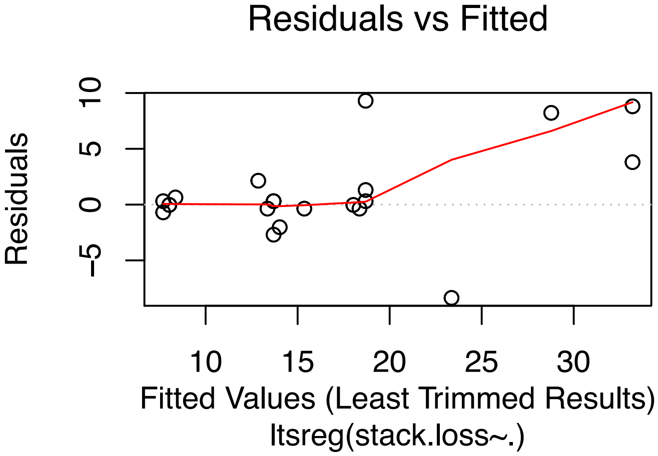
Compare the results.

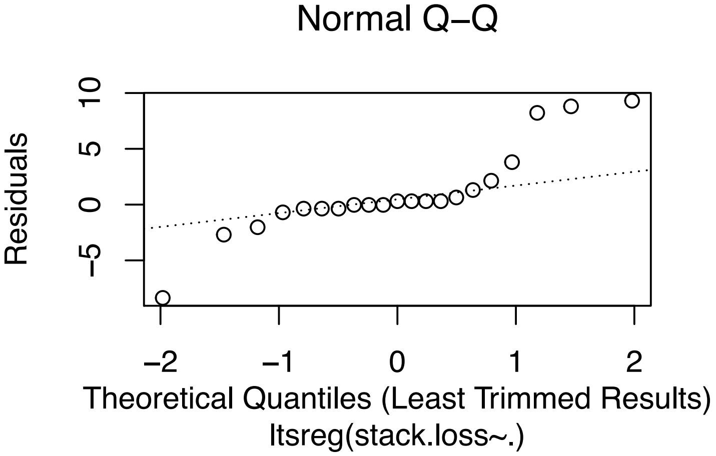
The variable `Acid.Conc.` has large p-value in Least Square, small confidence interval around 0 in LAD, small t-value in Huber method, and very small coefficient value in Least Trimmed Squares – all indicate that `Acid.Conc` is not significant in influencing the `stack.loss`. Let's compare different models by checking the constant variance assumption for the errors and the normality assumptions. From the plots we can see both Least Squares and Huber methods are significantly influenced by outlier.











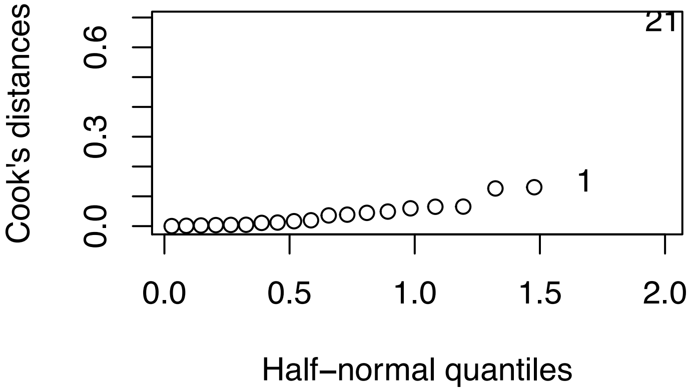
Now use diagnostic methods to detect any outliers or influential points.

To find the influential points, use Cook's distance to check. We can find that the observation #21 is an influential point/outlier.

```
n <- dim(stackloss)[1]
p <- dim(stackloss)[2]
cook <- cooks.distance(lm_stackloss)
cook[which(cook>4/(n-p-1))]
```

```
##        21
## 0.6919999
```

Plot the half-normal quantiles of the Cook Statistics to verify observation #21 is an outlier.



Remove these points and then use least squares. Compare the results.

Remove observation #21 and apply least squares. We can see the new Adjusted R^2 has rised from 0.8983 to 0.9392.

```
lm_stackloss_wo_outlier <- lm(formula(lm_stackloss), data=stackloss[-c(21),])
summary(lm_stackloss_wo_outlier)
```

```
##
```

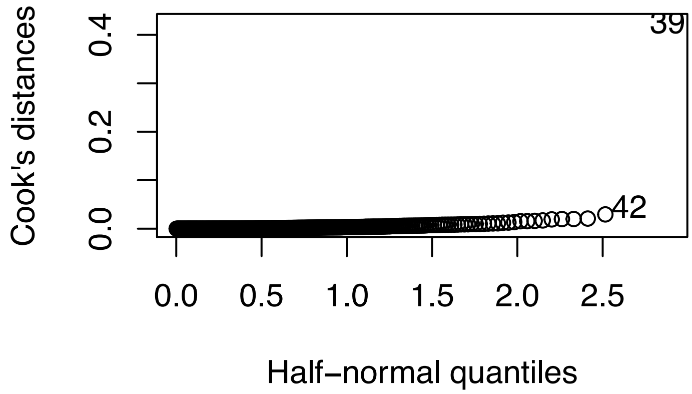
```
## Call:
## lm(formula = formula(lm_stackloss), data = stackloss[-c(21),
##     ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0449 -2.0578  0.1025  1.0709  6.3017
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -43.7040     9.4916  -4.605 0.000293 ***
## Air.Flow      0.8891     0.1188   7.481 1.31e-06 ***
## Water.Temp    0.8166     0.3250   2.512 0.023088 *
## Acid.Conc.   -0.1071     0.1245  -0.860 0.402338
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.569 on 16 degrees of freedom
## Multiple R-squared:  0.9488, Adjusted R-squared:  0.9392
## F-statistic: 98.82 on 3 and 16 DF,  p-value: 1.541e-10
```

2. For the `fat` data used in this chapter, a smaller model using only `age`, `weight`, `height` and `abdom` was proposed on the grounds that these predictors are either known by the individual or easily measured.

(a) Compare this model to the full thirteen-predictor model used earlier in the chapter. Is it justifiable to use the smaller model?
Conduct the ANOVA test. Suppose the null hypothesis is it's enough to use the smaller model, and the alternative hypothesis is that predictors should also include another 9 predictors other than the ones in small model. As the p-value for F-test is 0.002558 < 0.05, we have enough evidence to reject the null hypothesis. Thus it is not justifiable to use the smaller model, and the other 9 predictors in the full thirteen-predictor model are still useful.

```
## Analysis of Variance Table
##
## Model 1: brozek ~ age + weight + height + abdom
## Model 2: brozek ~ age + weight + height + neck + chest + abdom + hip +
##     thigh + knee + ankle + biceps + forearm + wrist
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1    247 4205.0
## 2    238 3785.1  9     419.9 2.9336 0.002558 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(c) For the smaller model, examine all the observations from case numbers 25 to 50. Which two observations seem particularly anomalous?
From the Cook's distance, we can see the observations #42 and #39 seem particularly anomalous.



3. Use the `fat` data, fitting the model described in Section 4.2.
See 2(a) for modeling fitting in `lm_fat_full`.

(a) Fit the same model but now using Huber's robust method. Comment on any substantial differences between this model and the least squares fit.

```
rlm_fat_full <- rlm(brozek ~ age + weight + height + neck + chest + abdom +
hip + thigh + knee + ankle + biceps + forearm + wrist, data=fat)
summary(lm_fat_full)
```

```
##
## Call:
## lm(formula = brozek ~ age + weight + height + neck + chest +
##     abdom + hip + thigh + knee + ankle + biceps + forearm + wrist,
##     data = fat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.264  -2.572  -0.097   2.898   9.327
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -15.29255   16.06992  -0.952  0.34225
## age           0.05679    0.02996   1.895  0.05929 .
## weight       -0.08031    0.04958  -1.620  0.10660
## height       -0.06460    0.08893  -0.726  0.46830
## neck         -0.43754    0.21533  -2.032  0.04327 *
## chest        -0.02360    0.09184  -0.257  0.79740
## abdom         0.88543    0.08008  11.057  < 2e-16 ***
## hip          -0.19842    0.13516  -1.468  0.14341
## thigh         0.23190    0.13372   1.734  0.08418 .
## knee         -0.01168    0.22414  -0.052  0.95850
## ankle         0.16354    0.20514   0.797  0.42614
## biceps        0.15280    0.15851   0.964  0.33605
## forearm       0.43049    0.18445   2.334  0.02044 *
## wrist        -1.47654    0.49552  -2.980  0.00318 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.988 on 238 degrees of freedom
## Multiple R-squared:  0.749,  Adjusted R-squared:  0.7353
## F-statistic: 54.63 on 13 and 238 DF,  p-value: < 2.2e-16
```

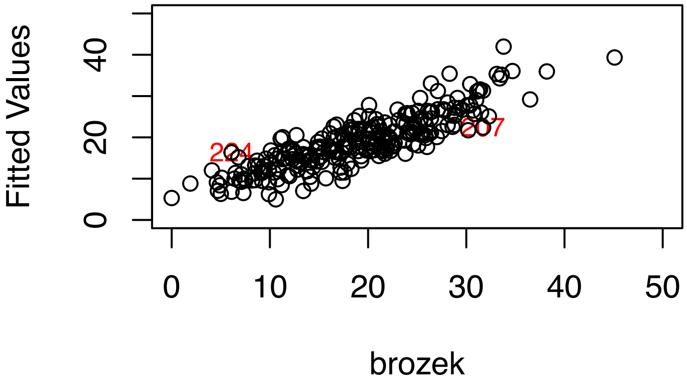
```
summary(rlm_fat_full)
```

```
##
## Call: rlm(formula = brozek ~ age + weight + height + neck + chest +
##     abdom + hip + thigh + knee + ankle + biceps + forearm + wrist,
##     data = fat)
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.3964  -2.7352  -0.1171   2.8008   9.4446
##
## Coefficients:
##             Value    Std. Error t value
## (Intercept) -11.3460  17.1216    -0.6627
## age           0.0650   0.0319     2.0368
## weight       -0.0643   0.0528    -1.2163
## height       -0.0625   0.0948    -0.6595
## neck         -0.4553   0.2294    -1.9846
## chest        -0.0256   0.0978    -0.2614
## abdom         0.8778   0.0853    10.2891
## hip          -0.2142   0.1440    -1.4872
## thigh         0.2632   0.1425     1.8473
## knee         -0.1076   0.2388    -0.4505
## ankle          0.1815   0.2186     0.8306
## biceps         0.1367   0.1689     0.8091
## forearm        0.4152   0.1965     2.1126
## wrist         -1.5739   0.5279    -2.9812
##
## Residual standard error: 4.073 on 238 degrees of freedom
```

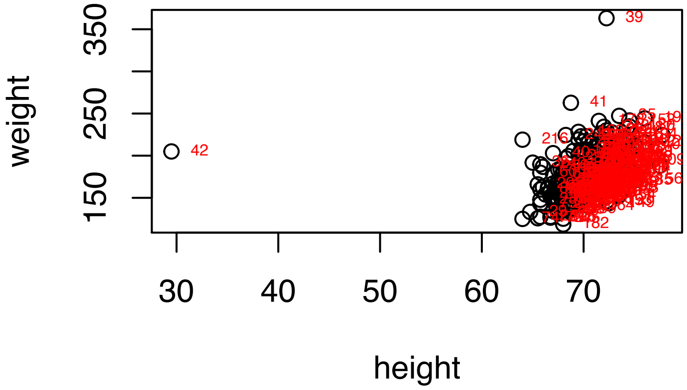
The predictors abdom, wrist, forearm, and age are top 4 significant predictors in Huber's robust method, while age predictor has a p-value > 0.05 in least square method, which making it not being significant. The residual standard error of least square method is 3.988 while that one in Huber's robust method is 4.073.

(b) Identify which two cases have the lowest weights in the Huber fit. What is unusual about these two points?
Observations #224 and #207 have the lowest weights in the Huber fit. These two points look deviate a little bit from the projected trend.

```
##       224       207
## 0.5269652 0.5800712
```



(c) Plot weight (of the man) against height. Identify the two outlying cases. Are these the same as those identified in the previous question? Discuss.
The two outlying cases are #42 and #39. They are different from the ones identified in the previous questions. As in previous questions, we are looking at the fitted values for variable brozek given 13 predictors, while in this question we are simply assuming a linear relationship between variable weight and height before looking for outliers that don't fit this relationship.



Attachment: RMarkdown Codes

~~~
---
title: 'STAT 500: HW5'
author: "Jasmine Mou"
date: "11/2/2017"
output: pdf_document

fontsize: 4pt
geometry: margin=0.5in
---
1. Using the `stackloss` data, fit a model with `stack.loss` as the response
and the other three variables as predictors using the following methods:
```{r}
data(stackloss,package="datasets")
```
(a) Least squares
```{r}
lm_stackloss <- lm(stack.loss~., data=stackloss)
summary(lm_stackloss)
```


(b) Least absolute deviations
```{r message=FALSE, warning=FALSE}
library(quantreg)
rq_stackloss <- rq(stack.loss~., data=stackloss)
summary(rq_stackloss)
```


(c) Huber method
```{r}
library(MASS)
rlm_stackloss <- rlm(stack.loss~., data=stackloss)
summary(rlm_stackloss)
```


(d) Least trimmed squares
```{r}
ltsreg_stackloss <- ltsreg(stack.loss~., data=stackloss, nsamp="exact")
ltsreg_stackloss
```


Compare the results.
*The variable `Acid.Conc.` has large p-value in Least Square, small
confidence interval around 0 in LAD, small t-value in Huber method, and very
small coefficient value in Least Trimmed Squares -- all indicate that
`Acid.Conc` is not significant in influencing the `stack.loss`. Let's compare
different models by checking the constant variance assumption for the errors
and the normality assumptions. From the plots we can see both Least Squares
and Huber methods are significantly influenced by outlier.*
```{r, fig.width=8, fig.height=3, echo=FALSE}
## define plotting functions
plot_residuals <- function(model, model_name){
~~~

~~~
  par(mfrow=c(1,2))
  ## Residuals vs Fitted
  par(col.lab="white")
  plot(model, which=1)
  par(col.lab="black")
  title(xlab=paste("Fitted Values (", model_name, ")", sep=""),
ylab="Residuals")

  ## Q-Q Plot
  par(col.lab="white")
  plot(model, which=2)
  par(col.lab="black")
  title(xlab=paste("Theoretical Quantiles (", model_name, ")", sep=""),
ylab="Residuals")
}

plot_residuals_customized <- function(model, model_name, model_func){
  par(mfrow=c(1,2))
  ## Residuals vs Fitted
  plot(fitted(model), residuals(model), xlab=paste("Fitted Values (",
model_name, ")\n", model_func, sep=""), ylab="Residuals", main="Residuals vs
Fitted", font.main=1)
  abline(h=0, lty=3, col="gray")
  panel.smooth(x=fitted(model), y=residuals(model))

  ## Q-Q Plot
  qqnorm(residuals(model), xlab=paste("Theoretical Quantiles (", model_name,
")\n", model_func, sep=""), ylab="Residuals", main="Normal Q-Q", font.main=1)
  qqline(residuals(model), lty=3)
}

## least squares
plot_residuals(lm_stackloss, "Least Squares")

## Huber
plot_residuals(rlm_stackloss, "Huber")

## LAD
plot_residuals_customized(rq_stackloss, "LAD", "rq(stack.loss ~.)")

## least trimmed results
plot_residuals_customized(ltsreg_stackloss, "Least Trimmed Results",
"ltsreg(stack.loss~.)")
```

Now use diagnostic methods to detect any outliers or influential points.
*To find the influential points, use Cook's distance to check. We can find
that the observation #21 is an influential point/outlier.  *

```{r}
n <- dim(stackloss)[1]
p <- dim(stackloss)[2]
cook <- cooks.distance(lm_stackloss)
~~~

~~~
cook[which(cook>4/(n-p-1))]
```
*Plot the half-normal quantiles of the Cook Statistics to verify observation
#21 is an outlier.*

```{r, fig.width=4, fig.height=3, echo=FALSE}
faraway::halfnorm(cook, 2, ylab="Cook's distances")
```


Remove these points and then use least squares. Compare the results.
*Remove observation #21 and apply least squares. We can see the new Adjusted
$R^2$ has rised from 0.8983 to 0.9392. *
```{r}
lm_stackloss_wo_outlier <- lm(formula(lm_stackloss), data=stackloss[-c(21),])
summary(lm_stackloss_wo_outlier)
```



2. For the `fat` data used in this chapter, a smaller model using only `age`,
`weight`, `height` and `abdom` was proposed on the grounds that these
predictors are either known by the individual or easily measured.
(a) Compare this model to the full thirteen-predictor model used earlier in
the chapter. Is it justifiable to use the smaller model?
*Conduct the ANOVA test. Suppose the null hypothesis is it's enough to use
the smaller model, and the alternative hypothesis is that predictors should
also include another 9 predictors other than the ones in small model. As the
p-value for F-test is 0.002558 < 0.05, we have enough evidence to reject the
null hypothesis. Thus it is not justifiable to use the smaller model, and the
other 9 predictors in the full thirteen-predictor model are still useful.*
```{r, echo=FALSE}
data(fat, package="faraway")
lm_fat_reduced <- lm(brozek ~ age + weight + height + abdom, data=fat)
lm_fat_full <- lm(brozek ~ age + weight + height + neck + chest + abdom +
hip + thigh + knee + ankle + biceps + forearm + wrist, data=fat)
anova(lm_fat_reduced, lm_fat_full)
```

(c) For the smaller model, examine all the observations from case numbers 25
to 50. Which two observations seem particularly anomalous?
*From the Cook's distance, we can see the observations #42 and #39 seem
particularly anomalous.*

```{r, fig.width=4, fig.height=3, echo=FALSE}
cook <- cooks.distance(lm_fat_reduced)
faraway::halfnorm(cook, 2, ylab="Cook's distances")
```



3. Use the `fat` data, fitting the model described in Section 4.2.
*See 2(a) for modeling fitting in `lm_fat_full`.*
(a) Fit the same model but now using Huber’s robust method. Comment on any
substantial differences between this model and the least squares fit.
```{r}
~~~

~~~
rlm_fat_full <- rlm(brozek ~ age + weight + height + neck + chest + abdom +
hip + thigh + knee + ankle + biceps + forearm + wrist, data=fat)
summary(lm_fat_full)
summary(rlm_fat_full)
```
*The predictors `abdom`, `wrist`, `forearm`, and `age` are top 4 significant
predictors in Huber's robust method, while `age` predictor has a p-value >
0.05 in least square method, which making it not being significant. The
residual standard error of least square method is 3.988 while that one in
Huber's robust method is 4.073. *

(b) Identify which two cases have the lowest weights in the Huber fit. What
is unusual about these two points?
*Observations #224 and #207 have the lowest weights in the Huber fit. These
two points look deviate a little bit from the projected trend. *
```{r, echo=FALSE}
wts <- rlm_fat_full$w
names(wts) <- row.names(fat)
head(sort(wts), 2)
```

```{r, fig.width=4, fig.height=3, echo=FALSE}
i <- c(207,224)
poi <- fat[i,]
plot(poi$brozek, lm_fat_full$fitted.values[i], xlim=c(0,50), ylim=c(0,50),
xlab="brozek", ylab="Fitted Values")
text(poi$brozek, lm_fat_full$fitted.values[i]+.1, i, col="red", cex=0.8)
points(fat$brozek, lm_fat_full$fitted.values)
```

(c) Plot weight (of the man) against height. Identify the two outlying cases.
Are these the same as those identified in the previous question? Discuss.
*The two outlying cases are #42 and #39. They are different from the ones
identified in the previous questions. As in previous questions, we are
looking at the fitted values for variable `brozek` given 13 predictors, while
in this question we are simply assuming a linear relationship between
variable `weight` and `height` before looking for outliers that don't fit
this relationship. *
```{r, fig.width=4, fig.height=3, echo=FALSE}
attach(fat)
plot(height, weight)
text(height, weight, row.names(fat), cex=0.5, pos=4, col="red")
```
~~~