PSYCGR01: Statistics
GLM assumptions and diagnostics worksheet

Maarten Speekenbrink *UCL Experimental Psychology*
October, 2016

1 General Exercises

1. The parameters in a multiple regression model are called *partial regression coefficients* because they reflect the relation between a predictor and the dependent variable after the effect of all other predictors in the model has been accounted for (or "partialled out"). The *partial correlation coefficient* is related to this, reflecting the correlation between the predictor and dependent variable after "partialling out" the effect of all the other predictors in the model. To compute the partial correlation, you can take the square root of the Proportional Reduction in Error (PRE) of a MODEL A with the predictor included, compared to a MODEL C where the predictor has been removed (this PRE is referred to as the "coefficient of partial determination"). You then have to give it the sign (i.e., a "+" or "-", indicating that the the correlation is positive or negative respectively) of the partial regression coefficient of the predictor.

   Take for example the following models (from the diabetes and control data discussed in the next lecture):

$$\begin{aligned}
\text{MODEL C}: &\quad Y_i = 35.40 + e_i \\
\text{MODEL A1}: &\quad Y_i = 36.58 - 0.11X_{1i} + e_i \\
\text{MODEL A2}: &\quad Y_i = 47.84 - 0.65X_{2i} + e_i \\
\text{MODEL A3}: &\quad Y_i = 47.74 + 0.03X_{1i} - 0.66X_{2i} + e_i
\end{aligned}$$

   with respective Sums of Squared Errors of SSE(C) = 3530.28, SSE(A1) = 3487.37, SSE(A2) = 2678.66 and SSE(A3) = 2676.29.

   (a) What is the partial correlation between X_1 and Y in MODEL A3?
   Answer: $\sqrt{\frac{\text{SSE(A2)}-\text{SSE(A3)}}{\text{SSE(A2)}}} = \sqrt{\frac{2678.66-2676.29}{2678.66}} = 0.0297$

   (b) What is the partial correlation between X_2 and Y in MODEL A3?
   Answer: $-\sqrt{\frac{\text{SSE(A1)}-\text{SSE(A3)}}{\text{SSE(A1)}}} = -\sqrt{\frac{3487.37-2676.29}{3487.37}} = -0.4822$

   (c) The main objective in multiple regression analysis is to account for the variance of the dependent variable (e.g., we use the predictors to "explain" part of the variance of the dependent variable). If predictors are correlated, they are (partly) redundant in the sense that they account for (partly) the same variance of the

dependent variable Y. The coefficient of partial determination reflects the proportion of unique variance of Y (variance that the other predictors cannot account for) that a predictor can account for. You can compute this coefficient from R^2 values. For example, if we let R^2_{A2} denote the PRE of MODEL A2 compared to MODEL C, and R^2_{A3} denote the PRE of MODEL A3 compared to MODEL C, we can compute the coefficient of partial determination between X_1 and Y as

$$\frac{R^2_{A3} - R^2_{A2}}{1 - R^2_{A2}}$$

How would you describe the numerator and denominator in terms of "proportions of variance"?

Answer: The numerator $(R^2_{A3} - R^2_{A2})$ reflects the proportion of variance accounted for by X_1 that can not be accounted for by X_2. The denominator $(1 - R^2_{A2})$ is the proportion of variance *not* accounted for by X_2.

To check that this way of computing the coefficient of partial determination is correct, first note that

$$1 - R^2_{A2} = \frac{\text{SSE(C)}}{\text{SSE(C)}} - \frac{\text{SSE(C)} - \text{SSE(A2)}}{\text{SSE(C)}} = \frac{\text{SSE(A2)}}{\text{SSE(C)}}$$

Using this value and $R^2_{A3} - R^2_{A2} = \frac{\text{SSE(A2)} - \text{SSE(A3)}}{\text{SSE(C)}}$ gives

$$\begin{aligned} \frac{R^2_{A3} - R^2_{A2}}{1 - R^2_{A2}} &= \frac{\frac{\text{SSE(A2)} - \text{SSE(A3)}}{\text{SSE(C)}}}{\frac{\text{SSE(A2)}}{\text{SSE(C)}}} \\ &= \frac{\text{SSE(A2)} - \text{SSE(A3)}}{\text{SSE(A2)}} \end{aligned}$$

as required.

2 SPSS Exercises

1. Open the datafile `2016_Questionnaire_1.sav`. Like last week, we'll look at a model in which you predict weight from height and age. However, you will now analyse the full dataset.

   (a) To get a feel for the data, obtain descriptives (e.g., means, standard deviations, minima, maxima, etc.) for the three variables and create histograms and scatterplots. Do these variables look like what you would expect?

```
> # read the data
> library(foreign)
> dat <- as.data.frame(read.spss("2016_Questionnaire_1.sav"))
```

```
Warning in read.spss("2016_Questionnaire_1.sav"):  2016_Questionnaire_1.sav:
File-indicated value is different from internal value for at least one
of the three system values.  SYSMIS: indicated -1.79769e+308, expected
-1.79769e+308; HIGHEST: 1.79769e+308, 1.79769e+308; LOWEST: -1.79769e+308,
-1.79769e+308
Warning in read.spss("2016_Questionnaire_1.sav"):  2016_Questionnaire_1.sav:
Unrecognized record type 7, subtype 18 encountered in system file
```

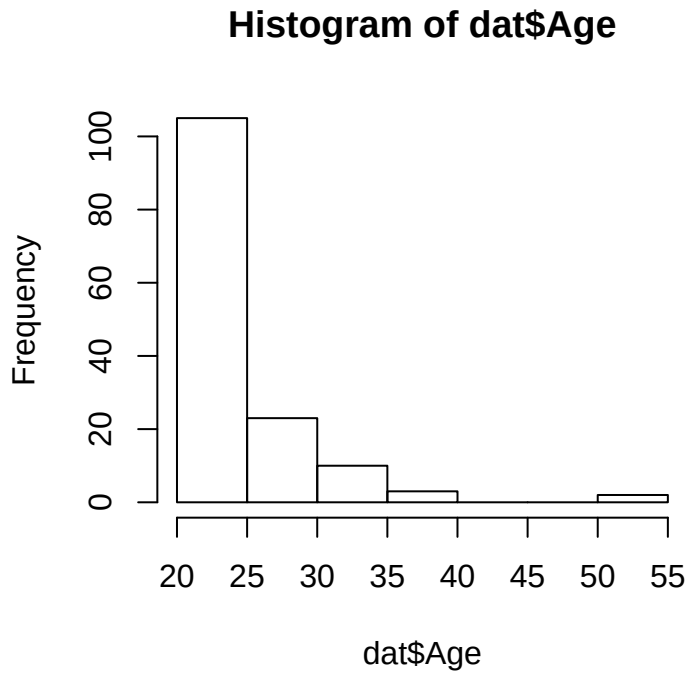
```
> # get means and ranges
> summary(dat[,c("Age","Height","Weight")])
```

```
      Age            Height          Weight
 Min.   :20.00   Min.   :  1.7   Min.   : 43.00
 1st Qu.:22.00   1st Qu.:163.0   1st Qu.: 54.50
 Median :23.00   Median :170.0   Median : 60.00
 Mean   :24.85   Mean   :167.8   Mean   : 64.49
 3rd Qu.:26.00   3rd Qu.:176.0   3rd Qu.: 69.50
 Max.   :52.00   Max.   :192.0   Max.   :181.00
```

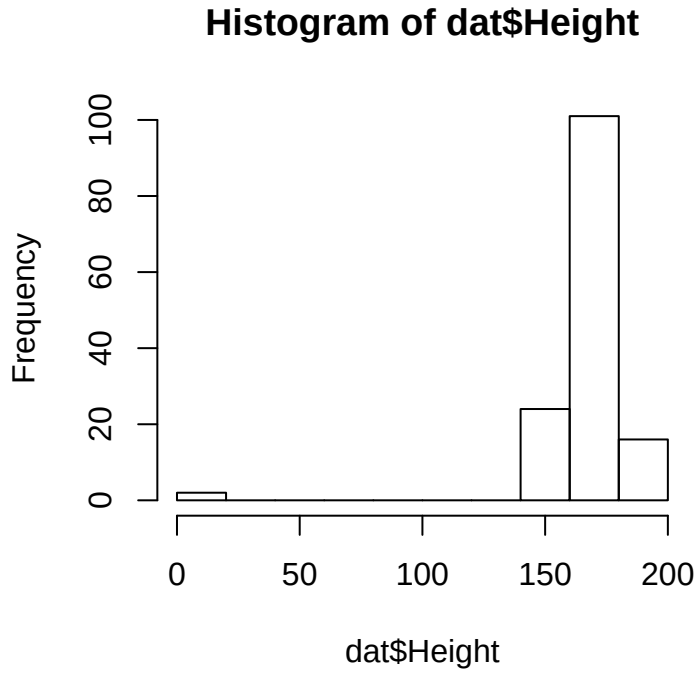
```
> # standard deviations
> apply(dat[,c("Age","Height","Weight")],2,sd)
```

```
      Age    Height    Weight
 4.815333 21.611478 17.194436
```

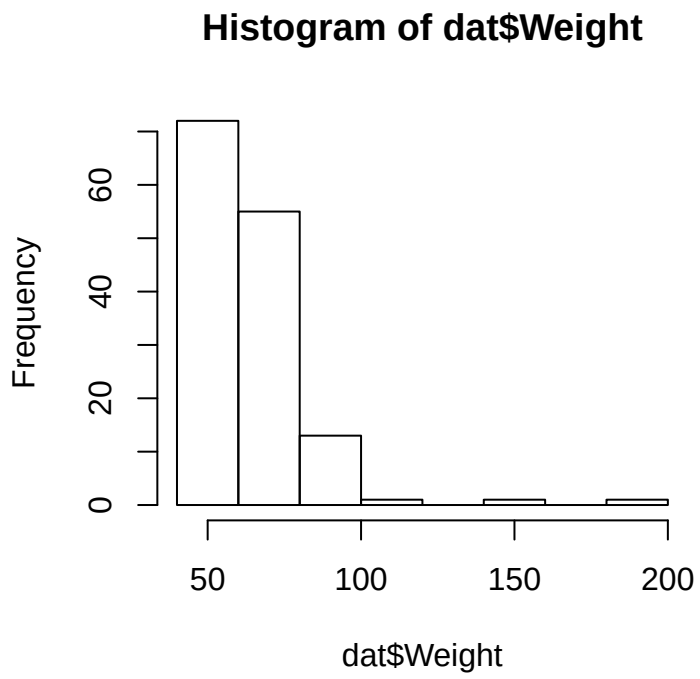
```
> # histograms
> hist(dat$Age)
```



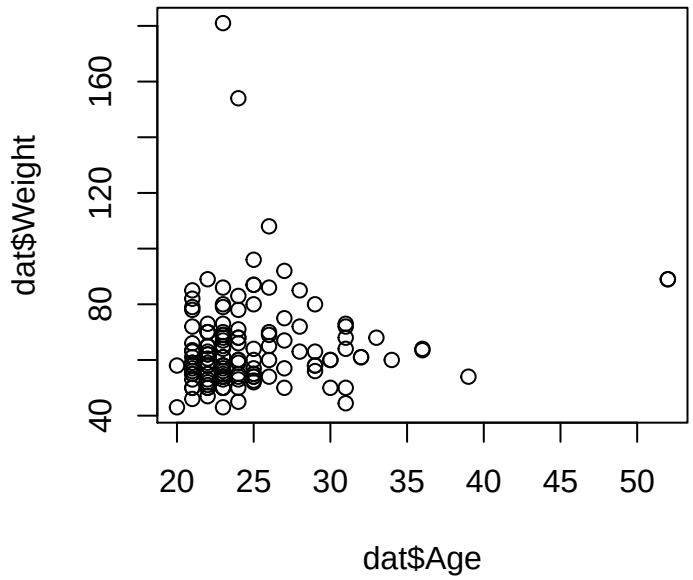

```
> hist(dat$Height)
```



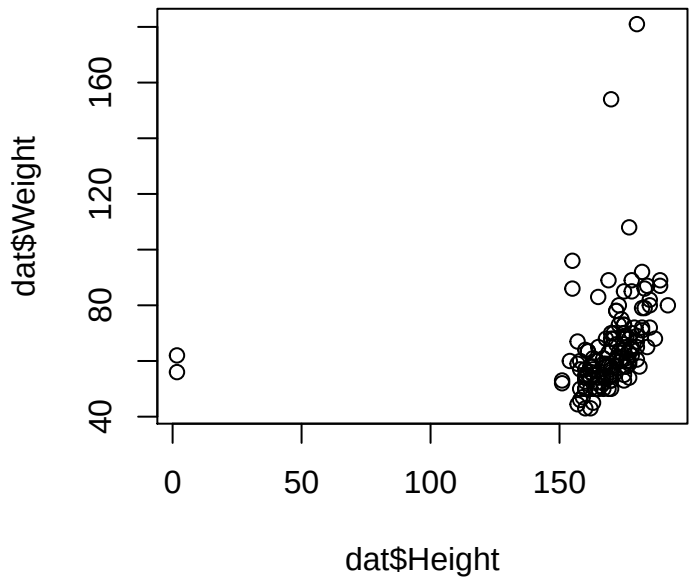
```
> hist(dat$Weight)
```



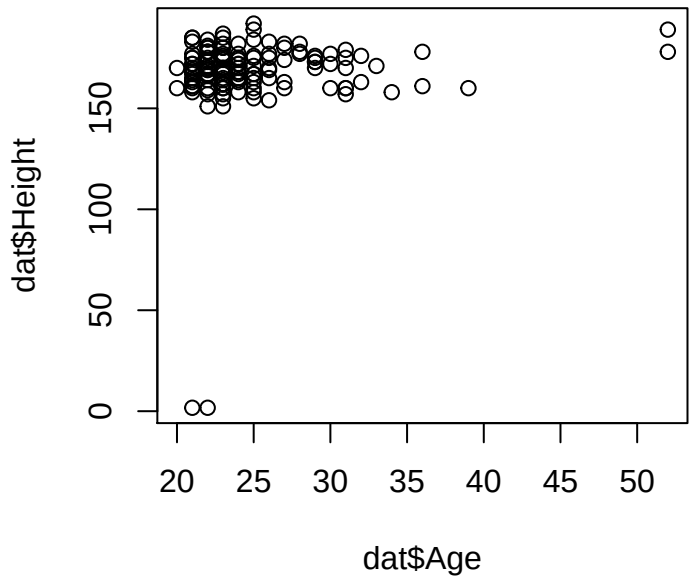

```
> # scatterplots
> plot(dat$Age,dat$Weight)
```




```
> plot(dat$Height,dat$Weight)
```



```
> plot(dat$Age, dat$Height)
```



Answer: There are some "strange" values in this data, such as a height of 1.7 cm (case 9 and case 33) and a weight of 181 kg (case 36). Especially the values of height seem to be clearly an error in data entry, or an untruthful response.

(b) Fit the parameters of the model

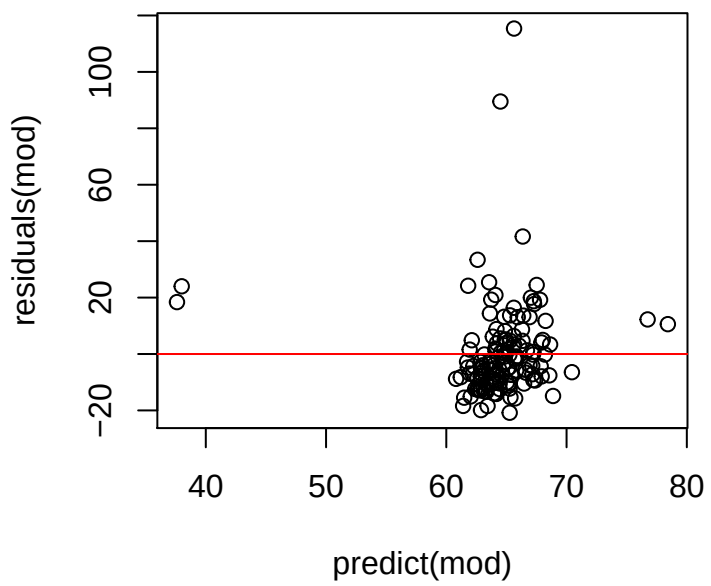
$$\texttt{Weight}_i = \beta_0 + \beta_1 \texttt{Height}_i + \beta_2 \texttt{Age}_i + \epsilon_i$$

and check the data for outliers and assess the assumptions underlying the model. If there are any outliers, re-estimate the model after removing them, and re-assess the assumptions of the model.

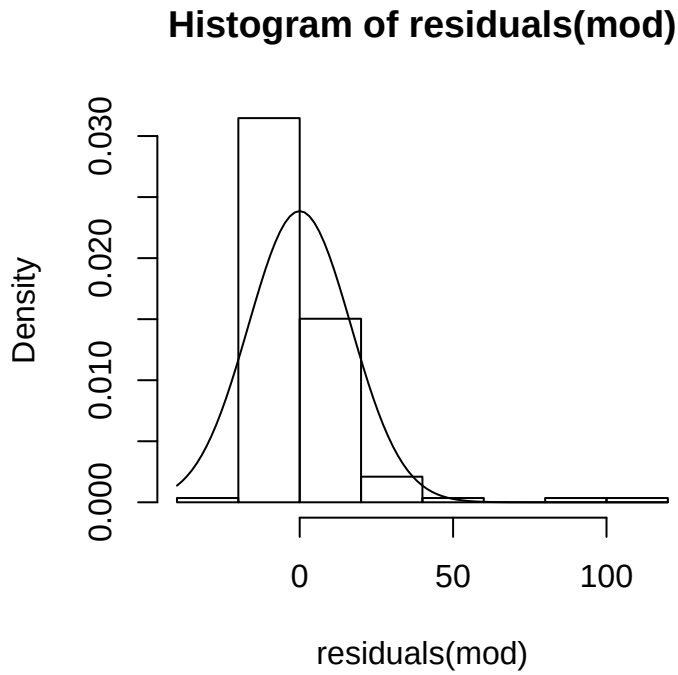
```
> mod <- lm(Weight ~ Height + Age,data=dat)
> summary(mod)


Call:
lm(formula = Weight ~ Height + Age, data = dat)

Residuals:
    Min      1Q  Median      3Q     Max
-20.819  -9.170  -4.251   4.689 115.365

Coefficients:
```

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 29.07969   12.60565   2.307   0.0225 *
Height       0.15280    0.06586   2.320   0.0218 *
Age          0.39355    0.29558   1.331   0.1852
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16.84 on 140 degrees of freedom
Multiple R-squared:  0.05401,Adjusted R-squared:  0.0405
F-statistic: 3.997 on 2 and 140 DF,  p-value: 0.02051
```

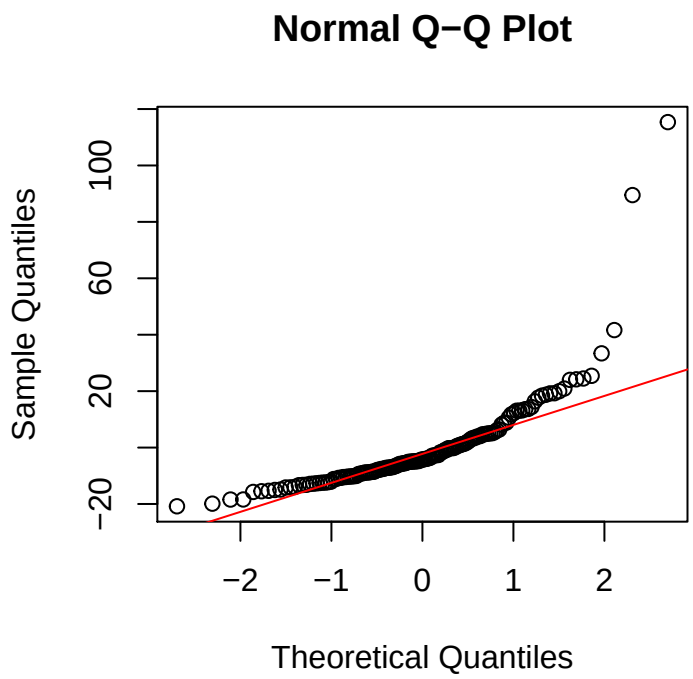
```
> # predicted vs residual plot
> plot(predict(mod),residuals(mod))
> abline(h=0,col="red")
```




```
> # histogram of residuals
> hist(residuals(mod),prob=TRUE)
> curve(dnorm(x,mean=0,sd=sd(residuals(mod))),add=TRUE)
```



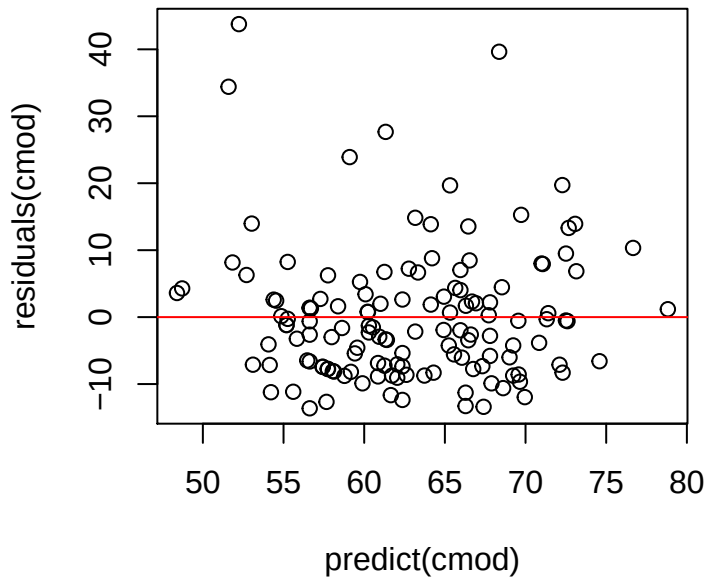

```
> # qqplot
> qqnorm(residuals(mod))
> qqline(residuals(mod), col = 2)
> ## normality tests
> shapiro.test(residuals(mod))


Shapiro-Wilk normality test

data:  residuals(mod)
W = 0.70179, p-value = 1.114e-15

> ks.test(residuals(mod),"pnorm",mean=0,sd=sd(residuals(mod)))

Warning in ks.test(residuals(mod), "pnorm", mean = 0, sd = sd(residuals(mod)
ties should not be present for the Kolmogorov-Smirnov test


One-sample Kolmogorov-Smirnov test

data:  residuals(mod)
D = 0.16276, p-value = 0.001025
alternative hypothesis: two-sided

> ## multicollinearity?
> library(car)
```



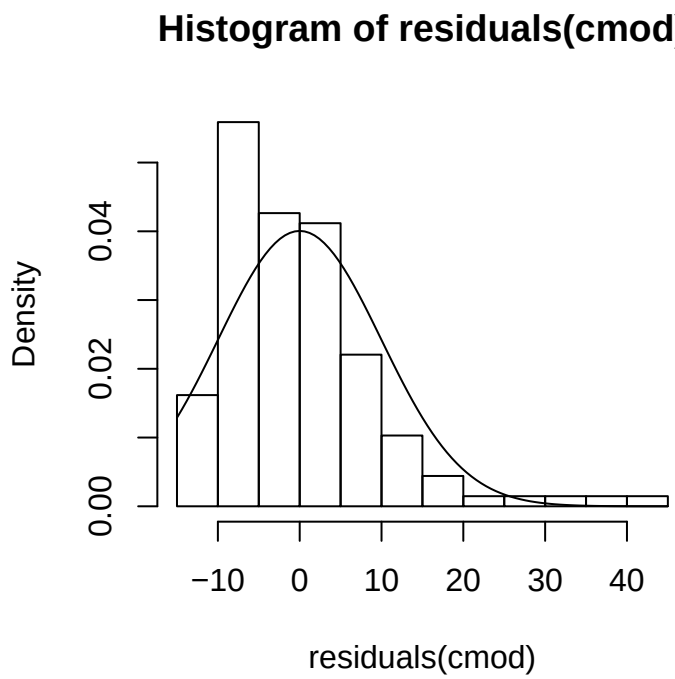

```
> vif(mod)

  Height      Age
1.014044 1.014044

> ## Outlier detection
> # mahalanobis distance
> preds <- as.matrix(dat[,c("Height","Age")])
> mahal <- mahalanobis(preds,colMeans(preds),cov(preds))
> which(mahal >= qchisq(1-.05/nrow(dat),df=2))

[1]  9 13 33 87

> # leverage
> lever <- hatvalues(mod)
> which(lever > (2*3)/nrow(dat))

  9  13  14  33  80  87 122
  9  13  14  33  80  87 122

> # studentized deleted residual
> sdr <- rstudent(mod)
> which(abs(sdr) >= qt(1-(.05/2)/nrow(dat),df=nrow(dat)-3-1))

36 53
36 53
```

```
> # cooks distance
> cooksd <- cooks.distance(mod)
> which(cooksd > 1)

named integer(0)

> ## let's have a look at the potential outliers
> cbind(dat[,c("Weight","Height","Age")],mahal,lever,sdr,cooksd)[
+   mahal >= qchisq(1-.05/nrow(dat),df=2) |
+     lever >= (2*3)/nrow(dat) |
+     abs(sdr) >= qt(1-(.05/2)/nrow(dat),df=nrow(dat)-3-1) |
+     cooksd >= 1,
+ ]

    Weight Height Age       mahal       lever         sdr      cooksd
9     62.0    1.7  22 59.15008850 0.423542926  1.8942506 0.862836636
13    89.0  178.0  52 31.83549778 0.231186653  0.8290644 0.069050700
14    64.0  178.0  36  5.40616628 0.045064601 -0.3904472 0.002412693
33    56.0    1.7  21 59.06187454 0.422921701  1.4433340 0.504998840
36   181.0  180.0  23  0.52546390 0.010693457  8.4382889 0.170867846
53   154.0  170.0  24  0.04641238 0.007319855  5.9534273 0.069916509
80    63.5  161.0  36  5.71349132 0.047228861 -0.2636096 0.001155889
87    89.0  189.0  52 31.90169675 0.231652843  0.7151450 0.051578095
122   54.0  160.0  39  9.14443080 0.071390407 -0.9160360 0.021528306

> ## for some reason, the leverage values computed with R
> ## don't correspond to those that SPSS gives...
> ## I'm going to check up on this, but for now
> ## and to give results consistent with those of SPSS
> ## I'm adding .01 to the criterion for the leverage
> ## This is not something that you should do normally!!
>
> ## remove the outliers
> cleandat <- subset(dat,
+     mahal < qchisq(1-.05/nrow(dat),df=2) &
+     lever < (2*3)/nrow(dat) + .01 &
+     abs(sdr) < qt(1-(.05/2)/nrow(dat),df=nrow(dat)-3-1) &
+     cooksd < 1)
> # re-estimate the model
> cmod <- lm(Weight ~ Height + Age,data=cleandat)
> summary(cmod)


Call:
lm(formula = Weight ~ Height + Age, data = cleandat)
```

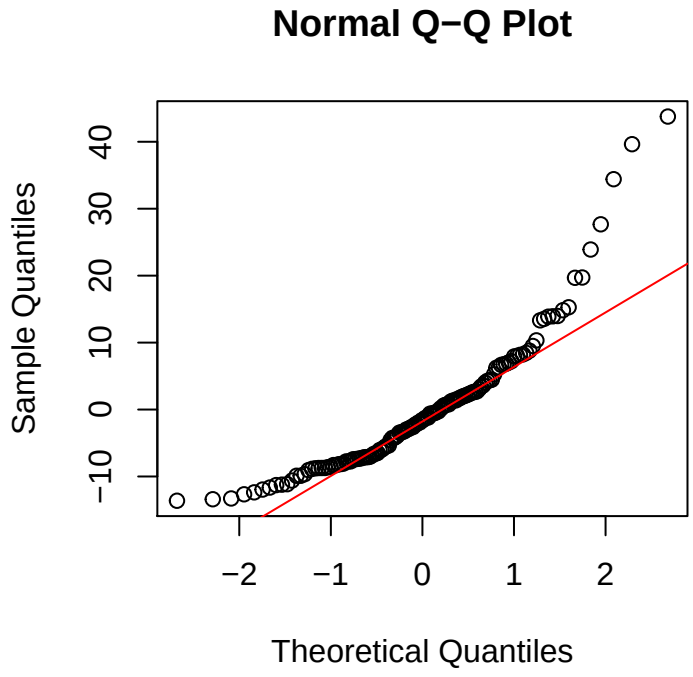
```
Residuals:
    Min      1Q  Median      3Q     Max
-13.621  -7.272  -1.581   3.713  43.766

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -67.1816    18.3071  -3.670  0.00035 ***
Height        0.7186     0.1016   7.073 7.77e-11 ***
Age           0.3215     0.2537   1.267  0.20719
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.03 on 133 degrees of freedom
Multiple R-squared:  0.2803,Adjusted R-squared:  0.2695
F-statistic:  25.9 on 2 and 133 DF,  p-value: 3.157e-10

> # check assumptions and potential problems
> # predicted vs residual plot
> plot(predict(cmod),residuals(cmod))
> abline(h=0,col="red")
```




```
> # histogram of residuals
> hist(residuals(cmod),prob=TRUE)
> curve(dnorm(x,mean=0,sd=sd(residuals(cmod))),add=TRUE)
```




```
> # qqplot
> qqnorm(residuals(cmod))
> qqline(residuals(cmod), col = 2)
```



```
> ## normality tests
> shapiro.test(residuals(cmod))


Shapiro-Wilk normality test

data:  residuals(cmod)
W = 0.86856, p-value = 1.241e-09

> ks.test(residuals(cmod),"pnorm",mean=0,sd=sd(residuals(cmod)))

Warning in ks.test(residuals(cmod), "pnorm", mean = 0, sd = sd(residuals(cmo
ties should not be present for the Kolmogorov-Smirnov test


One-sample Kolmogorov-Smirnov test

data:  residuals(cmod)
D = 0.12046, p-value = 0.03863
alternative hypothesis: two-sided

> # multicollinearity
> vif(cmod)

  Height      Age
1.000084 1.000084

> ## Outlier detection
> preds <- as.matrix(cleandat[,c("Height","Age")])
> cmahal <- mahalanobis(preds,colMeans(preds),cov(preds))
> which(cmahal >= qchisq(1-.05/nrow(cleandat),df=2))

named integer(0)

> clever <- hatvalues(cmod)
> which(clever > (2*3)/nrow(cleandat))

  8  14  35  38  51  75  80  92 104 121 125 132 142
  8  12  32  34  47  70  75  86  98 115 118 125 135

> csdr <- rstudent(cmod)
> which(abs(csdr) >= qt(1-(.05/2)/nrow(cleandat),df=nrow(cleandat)-3-1))

54 99
49 93

> ccooksd <- cooks.distance(cmod)
> which(ccooksd > 1)

named integer(0)
```

Answer: The estimated model is

$$\texttt{Weight}_i = 29.08 + 0.153 \times \texttt{Height}_i + 0.394 \times \texttt{Age}_i + e_i$$

and height is a significant predictor of weight. The predicted vs residual plot is difficult to interpret without first removing at least one outlier, but the histogram for the errors (residuals) shows that the distribution is somewhat skewed, which is also evident in the Q-Q plot, and which is furthermore confirmed by a significant Shapiro-Wilk test. Hence, the null-hypothesis of Normally distributed errors can be rejected. Checking the tolerance and/or VIF, Multicollinearity doesn't appear to be a problem. Before truly assessing the model assumptions, we should first remove outliers. For the leverage, we can take a heuristic that values larger than

$$\frac{2p}{n} = \frac{2 \times 3}{143} = 0.04196$$

indicate possible problems. For the Malahananobis distance, we can work out the critical value. Using a Bonferroni correction, the significance level becomes $\frac{.05}{143} = 0.0003497$, which gives a critical value of $\chi^2_{2;0.0003497} = 15.917$. For the Studentized deleted residual, we can use the same Bonferroni correction and work out the two-tailed critical values for df $= n - p - 1 = 143 - 3 - 1 = 139$ as -3.666 and 3.666. For the Cook's distance, we can use the heuristic that values larger than 1 indicate potential problems. There are 7 cases in the data for which at least one of these measures indicates a problem. Although it is questionable that all these 7 observations are "truly" outliers, as the number of potential outliers is relatively small compared to the total size of the sample, we can choose to remove all of these. After filtering out these nine potential outliers, the estimated model is

$$\texttt{Weight}_i = -67.182 + 0.719 \times \texttt{Height}_i + 0.322 \times \texttt{Age}_i + e_i$$

and again only the slope for height differs significantly from 0. Note that the slope is quite a bit larger than before. Re-checking for outliers shows that the Studentized deleted residual indicates two new outliers, and the leverage also indicates a number of observations with high weight. Cook's distance and Mahalanobis distance don't indicate potential problems. The predicted vs residual plot looks a lot better now. There is no clear reason to suspect the assumption of homoscedasticity is violated, although there is still some skew in the distribution of the errors, and a significant Shapiro-Wilks test indicates that the errors are not Normally distributed. You could try to remove more potential outliers, or transform the dependent variable (with e.g., a logarithmic or square-root transform) to reduce the skew. However, this is rather difficult. Personally, I don't think the problem with skew is too dramatic here and I doubt there will be any big changes to the results if it were diminished.

2. Open the dataset `reactionIQ.sav`. This data was simulated to replicate the results of Der & Deary (2003). The data contains two variables:

- `AH4`: Total number of correct answers in the Allice Heim 4 intelligence test, consisting of 65 items including series completion, mental arithmetic, vocabulary, and reasoning by analogy.
- `RT`: Average reaction time (in milliseconds) in a simple task in which participants pressed key 1, 2, 3, or 4 as soon as the corresponding digit was presented on screen

(a) Estimate the parameters of MODEL C:

$$\texttt{RT}_i = \beta_0 + \beta_1 \texttt{AH4}_i + \epsilon_i$$

and compute the SSE.

```
> dat <- as.data.frame(read.spss("reactionIQ.sav"))
> modC <- lm(RT~AH4,data=dat)
> summary(modC)


Call:
lm(formula = RT ~ AH4, data = dat)

Residuals:
    Min      1Q  Median      3Q     Max
-538.03  -65.56   -1.92   63.20  433.35

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 834.2973     8.0584  103.53   <2e-16 ***
AH4          -3.8537     0.2747  -14.03   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 100.5 on 898 degrees of freedom
Multiple R-squared:  0.1797,Adjusted R-squared:  0.1788
F-statistic: 196.7 on 1 and 898 DF,  p-value: < 2.2e-16

> SSEC <- sum(residuals(modC)^2)
> SSEC

[1] 9067692
```

Answer:

$$\texttt{RT}_i = 834.297 - 3.854 \times \texttt{AH4}_i + \epsilon_i$$

and $\text{SSE(C)} = 9,067,691.6$.

(b) Estimate the parameters of (polynomial regression) MODEL A:

$$\texttt{RT} = \beta_0 + \beta_1 \texttt{AH4}_i + \beta_2 \texttt{AH4}_i^2 + \epsilon_i$$

and compute the SSE and the PRE of this model compared to MODEL C. What is the null-hypothesis you test when comparing these two models? Test this null-hypothesis.

```
> dat$AH4sq <- dat$AH4^2
> modA <- lm(RT~AH4 + AH4sq,data=dat)
> summary(modA)


Call:
lm(formula = RT ~ AH4 + AH4sq, data = dat)

Residuals:
    Min      1Q  Median      3Q     Max
-561.83  -64.98    0.27   62.32  409.55

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 865.96919   13.57615  63.786  < 2e-16 ***
AH4          -6.62487    0.99638  -6.649 5.12e-11 ***
AH4sq         0.04911    0.01698   2.892  0.00392 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 100.1 on 897 degrees of freedom
Multiple R-squared:  0.1873,Adjusted R-squared:  0.1855
F-statistic: 103.4 on 2 and 897 DF,  p-value: < 2.2e-16

> SSEA <- sum(residuals(modA)^2)
> SSEA

[1] 8983902

> # PRE
> (SSEC - SSEA)/SSEC

[1] 0.009240466

> anova(modC,modA)

Analysis of Variance Table

Model 1: RT ~ AH4
Model 2: RT ~ AH4 + AH4sq
  Res.Df     RSS Df Sum of Sq     F   Pr(>F)
1    898 9067692
2    897 8983902  1     83790 8.366 0.003915 **
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Answer:

$$\texttt{RT} = 865.969 - 6.625 \times \texttt{AH4}_i + 0.049 \times \texttt{AH4}_i^2 + \epsilon_i$$

and SSE(A) $= 8,983,901.9$ and PRE $= \frac{9{,}067{,}691.6 - 8{,}983{,}901.9}{9{,}067{,}691.6} = 0.00924$. The null hypothesis tested is $H_0 : \beta_2 = 0$, and the test statistic is $F = 8.363$, with $P(F_{1,897} \geq 8.363) = .004$, so there is a significant quadratic effect of `AH4`.

(c) Screen the data for outliers and remove them if necessary. Base this analysis on MODEL A. Re-estimate the model after removal of any outliers and save the residuals as `RTres`.

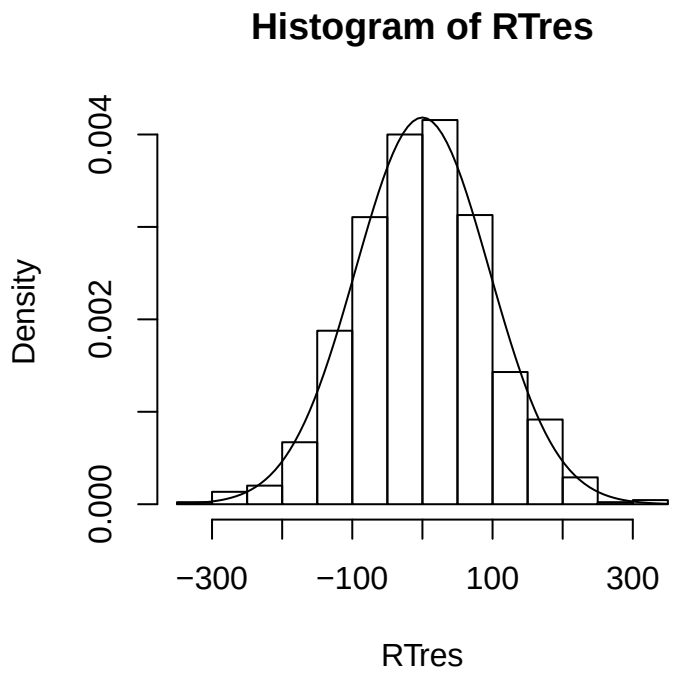
```
> sdr <- rstudent(modA)
> which(abs(sdr) >= qt(1-(.001/2),df=nrow(dat)-3-1))

34 40 42 43 87
34 40 42 43 87

> cooksd <- cooks.distance(modA)
> which(cooksd > 1)

named integer(0)

> cleandat <- dat[abs(sdr) < qt(1-(.001/2),df=nrow(dat)-3-1) & cooksd < 1, ]
> modA <- lm(RT~AH4 + AH4sq,data=cleandat)
> summary(modA)


Call:
lm(formula = RT ~ AH4 + AH4sq, data = cleandat)

Residuals:
     Min       1Q   Median       3Q      Max
-309.752  -65.181   -0.023   60.581  307.315

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 874.74741   13.27412  65.899  < 2e-16 ***
AH4          -7.20342    0.96844  -7.438  2.4e-13 ***
AH4sq         0.05756    0.01641   3.507 0.000476 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 95.46 on 892 degrees of freedom
Multiple R-squared:  0.2088,Adjusted R-squared:  0.207
F-statistic: 117.7 on 2 and 892 DF,  p-value: < 2.2e-16
```

```
> RTres <- residuals(modA)
```

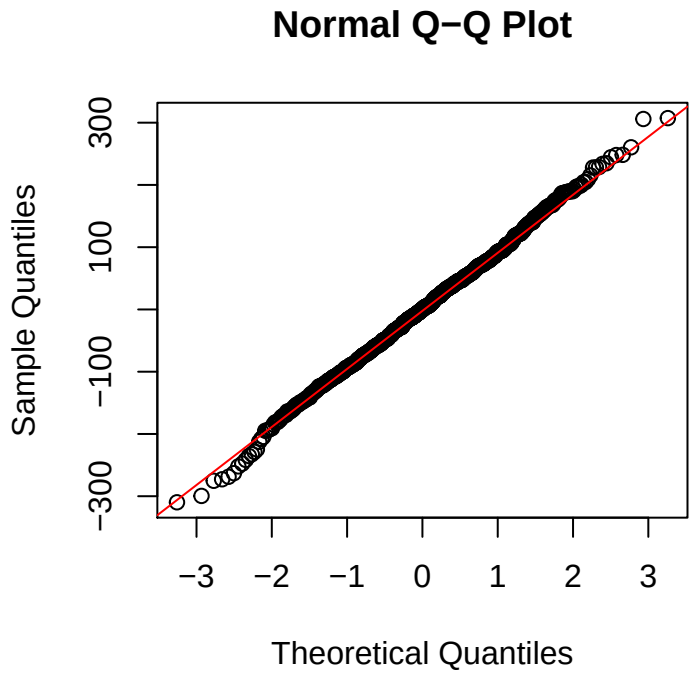
Answer: A full Bonferroni correction would result in $\frac{\alpha}{n} = \frac{.05}{900} = .0000056$ which is rather small and probably too conservative, so I'll use $\alpha = .001$ instead. The critical value for the Studentized deleted residual is then $t_{896;.0005} = 3.301$, or using a conservative value from a table, $t_{200;.0005} = 3.34$. I did not use the Mahalanobis distance here, because the second of the two predictors (`AH4`^{2}) is an artificial variable and completely dependent on `AH4`. Using the Studentized deleted residual, some outliers detected, although Cook's distance does not indicate any problems. While it is questionable that there really are outliers in this data, you could remove 5 "suspect" observations, which are mainly the more extreme `RT`'s when `AHA4` is low.

(d) Draw a histogram and Q-Q plot of `RTres`. Also draw a predicted by residual plot. Do these plots indicate potential problems?

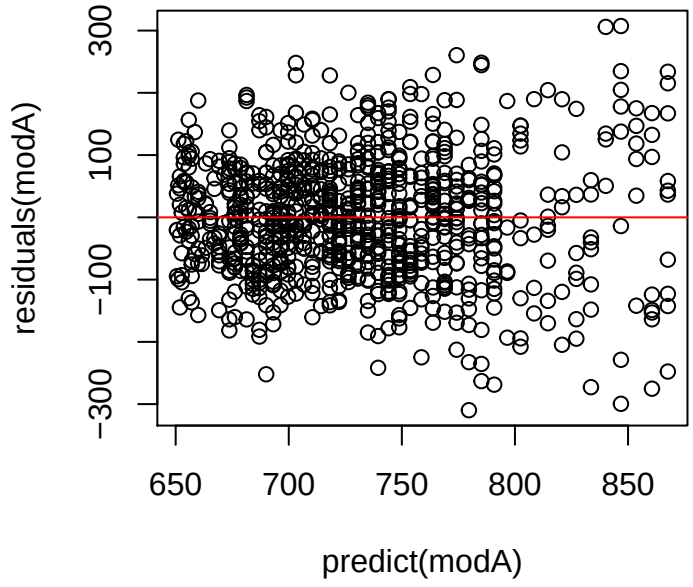
```
> hist(RTres,prob=TRUE)
> curve(dnorm(x,mean=0,sd=sd(RTres)),add=TRUE)
```




```
> qqnorm(RTres)
> qqline(RTres, col = 2)
```




```
> plot(predict(modA),residuals(modA))
> abline(h=0,col="red")
```



Answer: A histogram and QQ plot do not indicate major problems. However, the

predicted by residual plot indicates that there is heterogeneity of variance (higher RT's have larger variance)

(e) Use the Shapiro-Wilk test to assess whether the residuals (i.e., the model error) are normally distributed. Does this test indicate potential problems?

```
> shapiro.test(RTres)


Shapiro-Wilk normality test

data:  RTres
W = 0.99849, p-value = 0.6413
```

Answer: The Shapiro-Wilks test is not significant, so there is no reason to question that the data is Normally distributed.

(f) Use the Breusch-Pagan or Koenker test to test for homoscedasticity. You can find a description of these tests in the file `homoscedasticity.pdf` on Moodle. I would suggest you use the model

$$g_i = \beta_0 + \beta_1 \mathtt{AHA}_i + \beta_2 \mathtt{AHA}_i^2$$

for this, but you can try other possibilities as well.

```
> # compute SSR[Y]
> SSR_Y <- sum(residuals(lm(RT ~ 1,data=cleandat))^2) -
+   sum(residuals(modA)^2)
> # compute new variable g
> cleandat$g <- RTres^2/(SSR_Y/nrow(cleandat))
> gmod <- lm(g~AH4 + AH4sq,data=cleandat)
> summary(gmod)


Call:
lm(formula = g ~ AH4 + AH4sq, data = cleandat)

Residuals:
   Min     1Q Median     3Q    Max
-9.941 -2.717 -1.516  0.993 34.655

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.9582609  0.7328228  14.953  < 2e-16 ***
AH4         -0.4638690  0.0534644  -8.676  < 2e-16 ***
AH4sq        0.0060838  0.0009062   6.714 3.37e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 5.27 on 892 degrees of freedom
Multiple R-squared:  0.111,Adjusted R-squared:  0.109
F-statistic:  55.7 on 2 and 892 DF,  p-value: < 2.2e-16

> # Sum of Squares reduced for the Breusch-Pagan test
> SSR_g <- sum(residuals(lm(g ~ 1,data=cleandat))^2) -
+   sum(residuals(gmod)^2)
> # test statistic
> SSR_g/2

[1] 1547.14

> # p-value
> 1-pchisq(SSR_g/2,df=2)

[1] 0

> # Koenker test
> Rsq_g <- SSR_g/sum(residuals(lm(g ~ 1,data=cleandat))^2) # R-squared
> nrow(cleandat)*Rsq_g # Koenker test statistic

[1] 99.36931

> 1-pchisq(nrow(cleandat)*Rsq_g,df=2) # p-value

[1] 0
```

Answer: As the Shapiro-Wilks test is not significant, I'll use the Breusch-Pagan test here. First, I've computed a new variable

$$g_i = \frac{e_i^2}{\text{SSR}_Y/n} = \frac{\texttt{RTres}_i^2}{2145531.905/895}$$

and then estimated the following model

$$g_i = 10.958 - 0.464 \times \texttt{AHA}_i + 0.006 \times \texttt{AHA}_i^2 + e_i$$

with an $\text{SSR}_g = 3094.279$ and $R_g^2 = .111$. So the test statistic for the Breusch-Pagan test is

$$\text{LW}_B = \frac{\text{SSR}_g}{2} = \frac{3094.279}{2} = 1547.139$$

The critical value for a χ^2-distribution with 2 degrees of freedom and $\alpha = .05$ is 5.991, so the result is significant and the assumption of homoescedasticity should be rejected. For comparison, the Koenker test would be

$$\text{LW}_K = n \times R_g^2 = 895 \times .111 = 99.345$$

which is also significant.

(g) Reaction times are usually transformed before analysis. A common transform is the log-transform. Create the variable $\texttt{logRT} = \ln(\texttt{RT})$ (ln denotes the natural log, which is a logarithm with base e). Now estimate the parameters of a regression model with `logRT` as dependent variable and `AH4` as predictor. Use polynomial regression to check whether you should include powers of `AH4`. You can do this by starting with model

$$\texttt{logRT}_i = \beta_0 + \beta_1 \texttt{AHA}_i + \epsilon_i$$

and compare it to model

$$\texttt{logRT}_i = \beta_0 + \beta_1 \texttt{AHA}_i + \beta_2 \texttt{AHA}_i^2 + \epsilon_i$$

and if the comparison is significant, compare this model to

$$\texttt{logRT}_i = \beta_0 + \beta_1 \texttt{AHA}_i + \beta_2 \texttt{AHA}_i^2 + \beta_3 \texttt{AHA}_i^3 + \epsilon_i$$

etc.

```
> dat$logRT <- log(dat$RT)
> mod <- lm(logRT~AH4,data=dat)
> summary(mod)


Call:
lm(formula = logRT ~ AH4, data = dat)

Residuals:
     Min       1Q   Median       3Q      Max
-1.04960 -0.08591  0.00679  0.09312  0.43470

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.7159247  0.0111114  604.42   <2e-16 ***
AH4         -0.0049551  0.0003788  -13.08   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1386 on 898 degrees of freedom
Multiple R-squared:   0.16,Adjusted R-squared:  0.1591
F-statistic: 171.1 on 1 and 898 DF,  p-value: < 2.2e-16

> mod <- update(mod,.~.+AH4sq)
> summary(mod)


Call:
```

```
lm(formula = logRT ~ AH4 + AH4sq, data = dat)

Residuals:
     Min       1Q   Median       3Q      Max
-1.06644 -0.08495  0.00834  0.09303  0.41786

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.738e+00  1.878e-02 358.729  < 2e-16 ***
AH4         -6.915e-03  1.379e-03  -5.016 6.35e-07 ***
AH4sq        3.474e-05  2.349e-05   1.479     0.14
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1385 on 897 degrees of freedom
Multiple R-squared:  0.1621,Adjusted R-squared:  0.1602
F-statistic: 86.75 on 2 and 897 DF,  p-value: < 2.2e-16
```

Answer:

$$\texttt{logRT}_i = 6.716 - .005\texttt{AHA} + e_i$$

As the quadratic effect is not significant, there is no need to include it in the model.

(h) Screen the data for outliers and remove them if necessary. Base this analysis on the model you determined for `logRT`. Re-estimate the model after removal of outliers and save the residuals as `logRTres`.

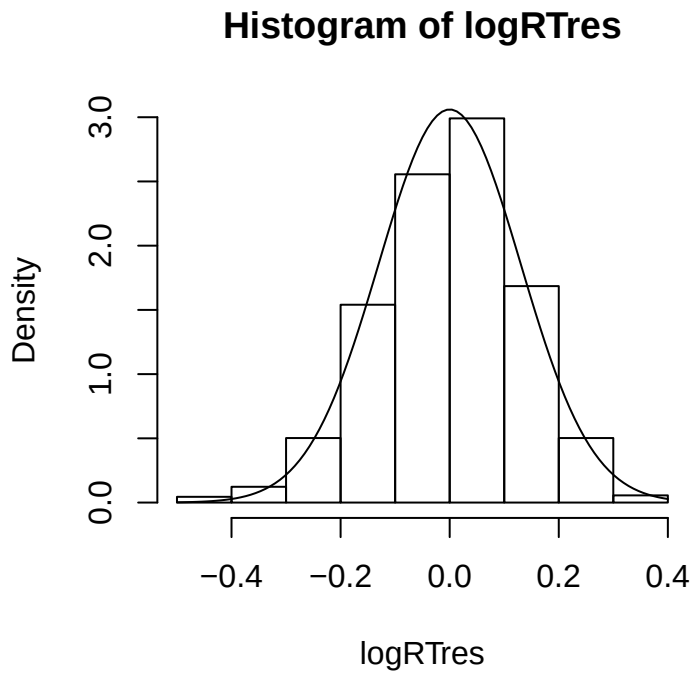
```
> sdr <- rstudent(mod)
> which(abs(sdr) > qt(1-.001/2,df=nrow(dat)-2-1))

40 42 87 98
40 42 87 98

> cooksd <- cooks.distance(mod)
> which(cooksd > 1)

named integer(0)

> cleandat <- dat[abs(sdr) < qt(1-(.001/2),df=nrow(dat)-2-1) & cooksd < 1, ]
> mod <- lm(logRT~AH4,data=cleandat)
> summary(mod)


Call:
lm(formula = logRT ~ AH4, data = cleandat)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.45451 -0.08812  0.00508  0.08919  0.42220

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.7296396  0.0105949  635.17   <2e-16 ***
AH4         -0.0053606  0.0003605  -14.87   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.131 on 894 degrees of freedom
Multiple R-squared:  0.1983,Adjusted R-squared:  0.1974
F-statistic: 221.1 on 1 and 894 DF,  p-value: < 2.2e-16

> summary(update(mod,.~.+AH4sq))


Call:
lm(formula = logRT ~ AH4 + AH4sq, data = cleandat)

Residuals:
     Min       1Q   Median       3Q      Max
-0.44874 -0.08598  0.00786  0.08938  0.39097

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.771e+00  1.799e-02 376.317  < 2e-16 ***
AH4         -8.960e-03  1.315e-03  -6.812 1.77e-11 ***
AH4sq        6.353e-05  2.233e-05   2.844  0.00455 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1305 on 893 degrees of freedom
Multiple R-squared:  0.2055,Adjusted R-squared:  0.2037
F-statistic: 115.5 on 2 and 893 DF,  p-value: < 2.2e-16

> mod <- lm(logRT~AH4+AH4sq,data=cleandat)
> logRTres <- residuals(mod)
```

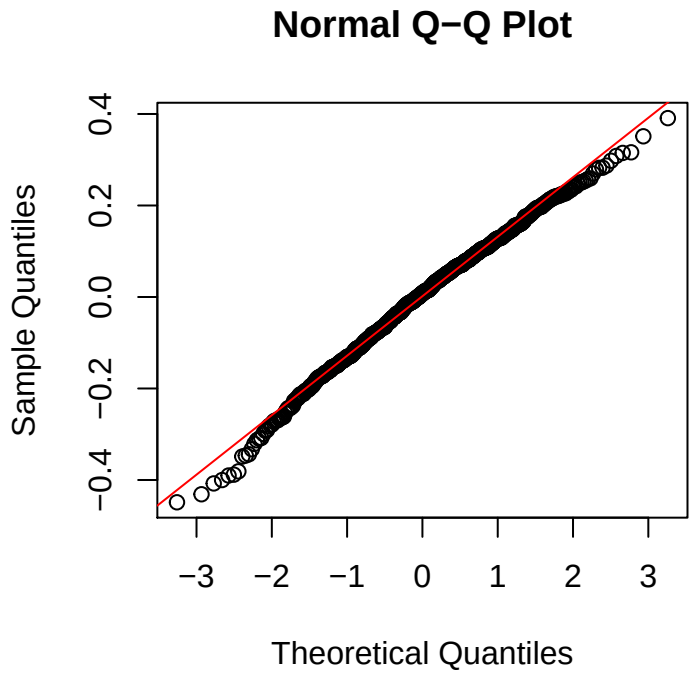
Answer: There are 4 outliers detected by the Studentized deleted residual. After removing outliers, you should re-estimate the model. It now also makes sense to check whether the quadratic effect is significant, which it now is.

(i) Draw a histogram and Q-Q plot of `logRTres`. Also draw a predicted by residual plot. Do these plots indicate potential problems?

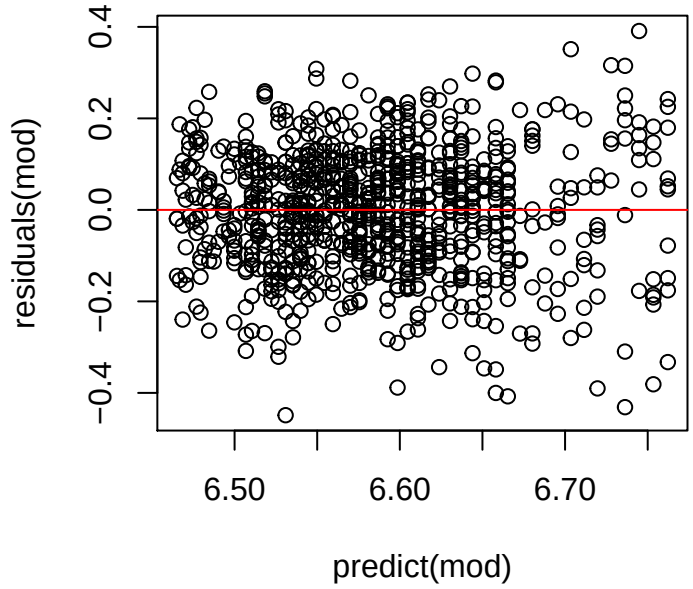
```
> hist(logRTres,prob=TRUE)
> curve(dnorm(x,mean=0,sd=sd(logRTres)),add=TRUE)
```




```
> qqnorm(logRTres)
> qqline(logRTres, col = 2)
```




```
> plot(predict(mod),residuals(mod))
> abline(h=0,col="red")
```



Answer: The histogram and Q-Q plot look quite good. The predicted by residual

plot looks improved as well.

(j) Use the Shapiro-Wilk test to assess whether the residuals (i.e., the model error) follow a Normal distribution. Does this test indicate potential problems?

```
> shapiro.test(logRTres)


Shapiro-Wilk normality test

data:  logRTres
W = 0.99453, p-value = 0.002508
```

Answer: The Shapiro-Wilks test is significant, so the null hypothesis that the errors are normally distributed is rejected.

(k) Use the Breusch-Pagan or Koenker test to test for homoscedasticity.

```
> # compute SSR[Y]
> SSR_Y <- sum(residuals(lm(logRT ~ 1,data=cleandat))^2) -
+ sum(residuals(mod)^2)
> # compute new variable g
> cleandat$g <- logRTres^2/(SSR_Y/nrow(cleandat))
> gmod <- lm(g~AH4 + AH4sq,data=cleandat)
> summary(gmod)


Call:
lm(formula = g ~ AH4 + AH4sq, data = cleandat)

Residuals:
   Min     1Q Median     3Q    Max
-8.536 -2.951 -1.772  0.980 43.197

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.3644600  0.7584016  12.348  < 2e-16 ***
AH4         -0.3746863  0.0554368  -6.759 2.51e-11 ***
AH4sq        0.0052447  0.0009414   5.571 3.35e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.5 on 893 degrees of freedom
Multiple R-squared:  0.0603,Adjusted R-squared:  0.0582
F-statistic: 28.65 on 2 and 893 DF,  p-value: 8.693e-13

> # Sum of Squares reduced for the Breusch-Pagan test
> SSR_g <- sum(residuals(lm(g ~ 1,data=cleandat))^2) -
```

```
+   sum(residuals(gmod)^2)
> # test statistic
> SSR_g/2

[1] 866.8516

> # p-value
> 1-pchisq(SSR_g/2,df=2)

[1] 0

> # Koenker test
> Rsq_g <- SSR_g/sum(residuals(lm(g ~ 1,data=cleandat))^2) # R-squared
> nrow(cleandat)*Rsq_g # Koenker test statistic

[1] 54.03115

> 1-pchisq(nrow(cleandat)*Rsq_g,df=2) # p-value

[1] 1.85052e-12
```

Answer: As the Shapiro-Wilks test is significant, We should use the Koenker test. First, I've computed a new variable

$$g_i = \frac{e_i^2}{\text{SSR}_Y/n} = \frac{\texttt{logRTres}_i^2}{3.933/896}$$

and then estimated the following model

$$g_i = \beta_0 + \beta_1 \times \texttt{AH4}_i + \beta_2 \times \texttt{AH4}_i^2 + e_i$$

with an $\text{SSR}_g = 1733.983$ and $R_g^2 = .060$. So the test statistic for the Koenker test is

$$\text{LW}_K = nR_g^2 = 896 \times 0.060 = 53.76$$

The critical value for a χ^2-distribution with 2 degrees of freedom is 5.991, so the result is significant and the assumption of homoscedasticity should again be rejected. For comparison, the test statistic for the Breusch-Pagan test is

$$\text{LW}_B = \frac{\text{SSR}_g}{2} = \frac{1733.983}{2} = 866.9915$$

which is also significant.

While we seem to have reduced the level of heteroscedasticity, the log transformation did not manage to eliminate it. It might be wise to check some other transformations to see whether that gives better results.

3. Open the dataset `mystery.sav`. This dataset contains three "mystery" variables, `y`, `x1` and `x2`. You will need to predict `y`.

(a) Estimate a model to predict `y` from `x1` and `x2`. How much variance of `y` do `x1` and `x2` "explain"?

```
> dat <- as.data.frame(read.spss("mystery.sav"))
> mod <- lm(y~x1+x2,data=dat)
> summary(mod)


Call:
lm(formula = y ~ x1 + x2, data = dat)

Residuals:
    Min      1Q  Median      3Q     Max
-0.6893 -0.4855 -0.3408  0.1382  2.2563

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.68142    0.65320   1.043    0.311
x1           0.49918    0.07789   6.409 6.48e-06 ***
x2           0.86442    0.03416  25.308 6.20e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8504 on 17 degrees of freedom
Multiple R-squared:  0.9742,Adjusted R-squared:  0.9711
F-statistic: 320.6 on 2 and 17 DF,  p-value: 3.179e-14
```

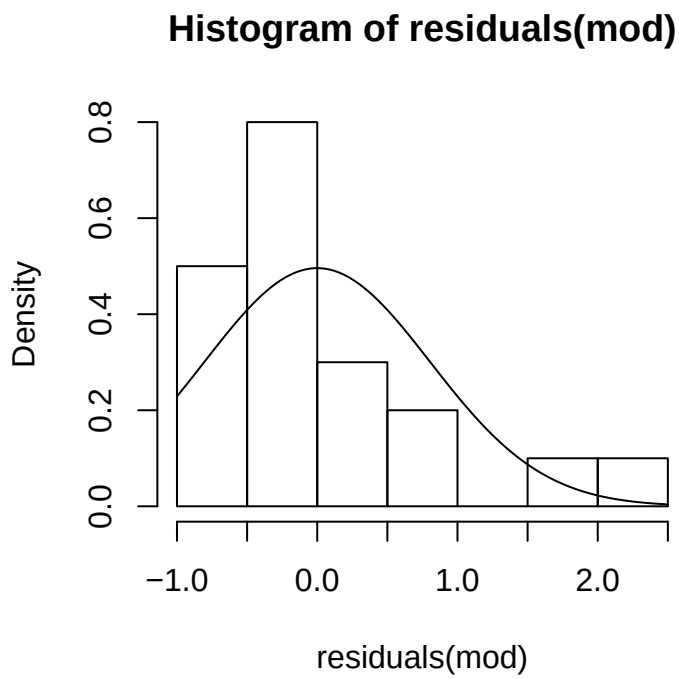
Answer: Using a multiple regression model, we get

$$Y_i = 0.681 + 0.499X_{1i} + 0.864X_{2i} + e_i$$

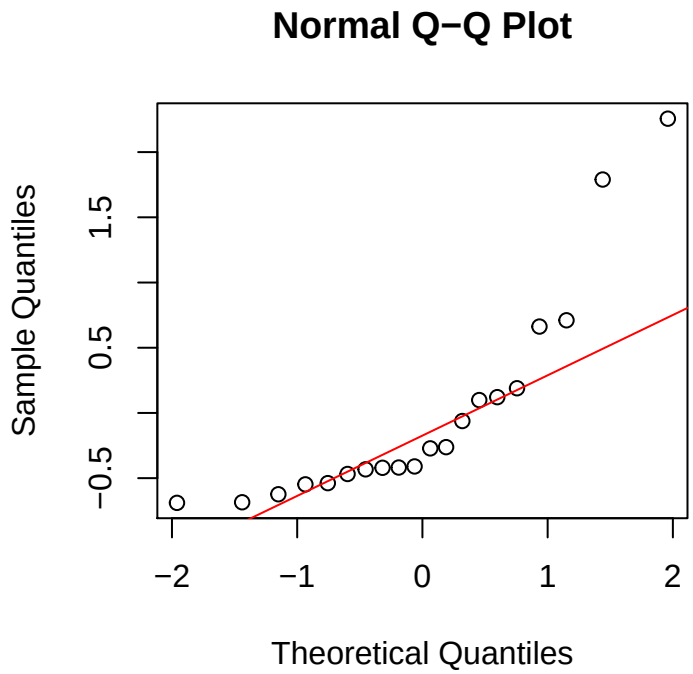
The "proportion of variance explained" $R^2 = .974$ is very high.

(b) Assess the assumptions underlying the model you specified. Can you improve upon the model?

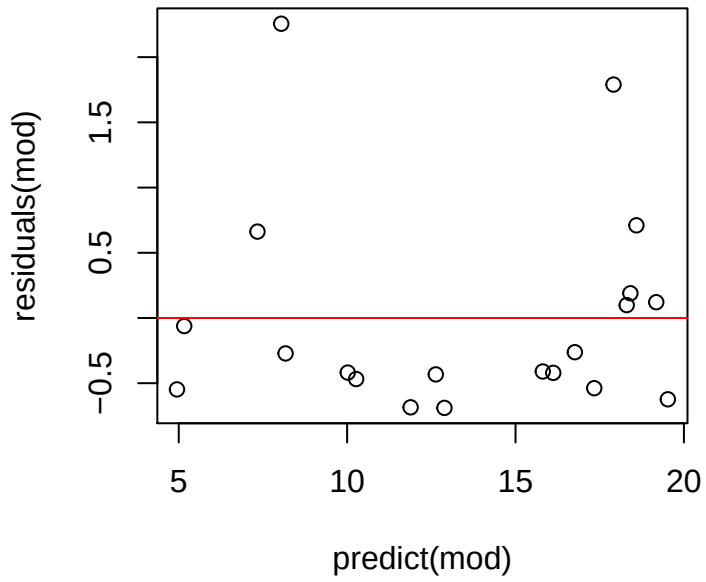
```
> hist(residuals(mod),prob=TRUE)
> curve(dnorm(x,mean=0,sd=sd(residuals(mod))),add=TRUE)
```



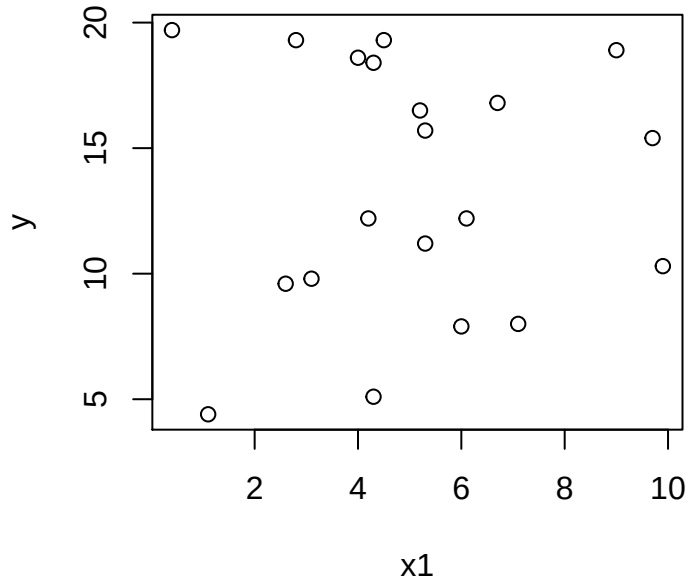

```
> qqnorm(residuals(mod))
> qqline(residuals(mod), col = 2)
```



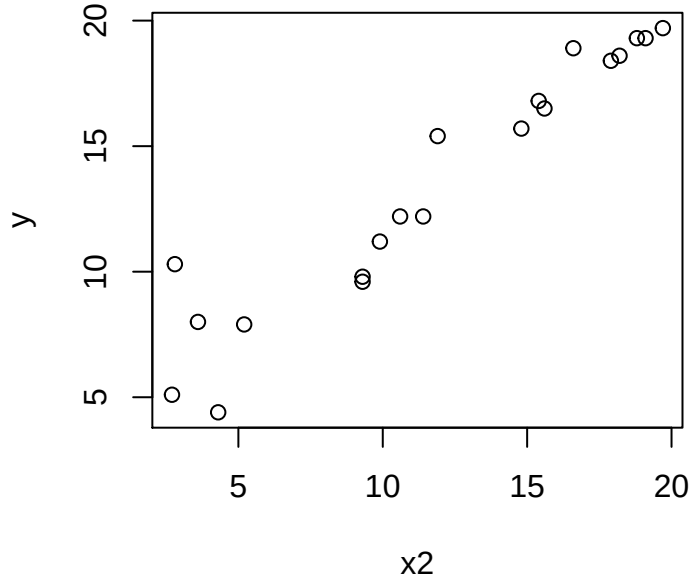
```
> plot(predict(mod),residuals(mod))
> abline(h=0,col="red")
```




```
> shapiro.test(residuals(mod))


Shapiro-Wilk normality test

data:  residuals(mod)
W = 0.76823, p-value = 0.0003012

> # evidence of nonlinearity?
> plot(y~x1,data=dat)
```




```
> plot(y~x2,data=dat)
```



Answer: The graphical methods to detect problems in the assumptions show some cause for concern. A histogram and QQ plot of the residuals indicate a deviation

from normality. A predicted by residual plot also shows a slightly strange pattern. From a scatterplot of the predictors and outcome, there is no immediate reason to suspect a nonlinear relation between Y and either X_1 or X_2, so a linear regression model looks reasonable for this data. But perhaps the error is not Normally distributed.

The problem is that you can't always detect the cause for violations of the model assumptions. In this case, the real model underlying the data was the "Pythagorean" model $Y_i^2 = X_{1i}^2 + X_{2i}^2$ (or $Y_i = \sqrt{X_{1i} + X_{2i}}$) Estimating this correct model (by squaring the predictors and dependent variable and then using a linear regression) gives

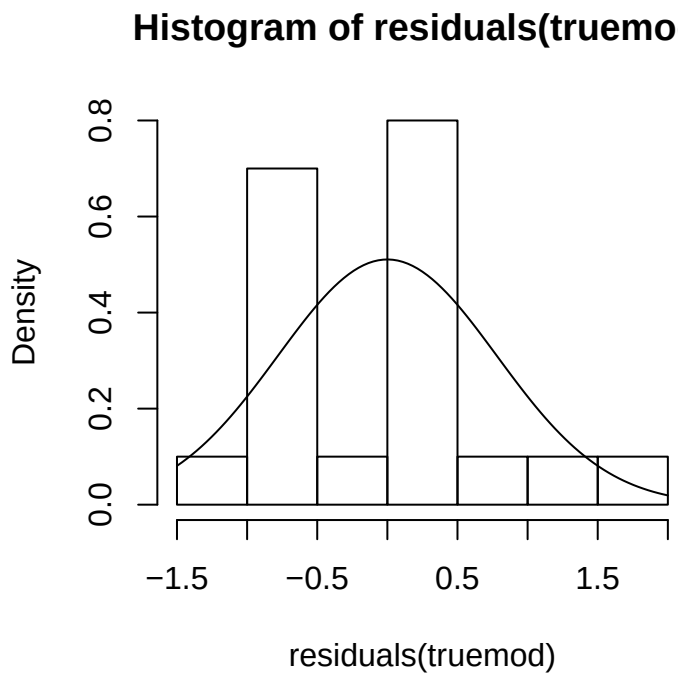
$$Y_i^2 = 1.013X_{1i}^2 + 1.000X_{2i}^2 + e_i$$

which can "explain" 100% of the variance of Y ($R^2 = 1.00$). Note also that the residuals look much better now.

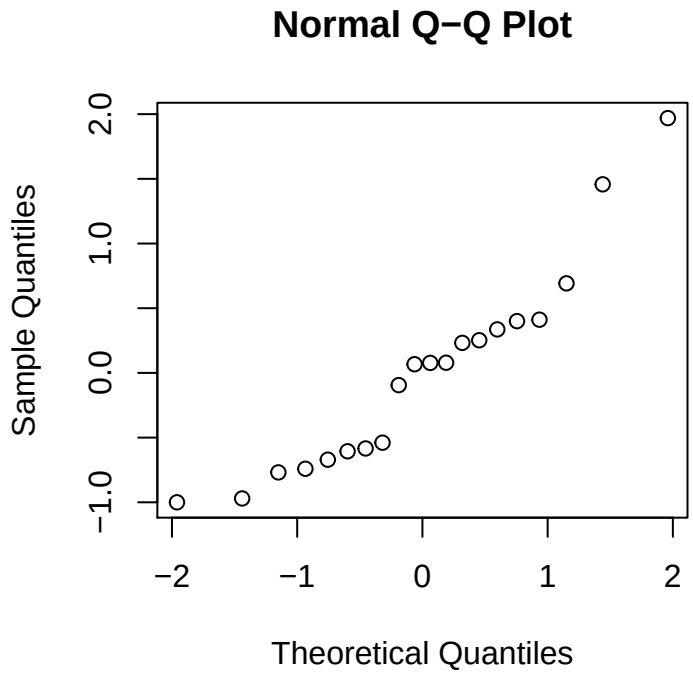
The idea for this exercise was not for you to come up with the correct model, although you may have tried to transform the dependent variable using e.g. a log or square-root transform. What this exercise was intended to show is that a linear model can be used quite successfully (in terms of R^2, the linear model does extremely well here) even though it is essentially incorrect. As a (first) approximation, a linear model often works quite well, especially with multiple predictors, in which case determining the true form of the relation can become exceedingly difficult.

```
> # the "true" model
> # within a formula, you can use the I()
> # function so that you can directly
> # transform variables without having to
> # store them first
> truemod <- lm(I(y^2) ~ I(x1^2) + I(x2^2),data=dat)
> summary(truemod)


Call:
lm(formula = I(y^2) ~ I(x1^2) + I(x2^2), data = dat)

Residuals:
     Min       1Q   Median       3Q      Max
-1.00005 -0.62210  0.07219  0.35223  1.96931

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.418815   0.414298  -1.011    0.326
I(x1^2)      1.012920   0.006753 149.990   <2e-16 ***
I(x2^2)      0.999795   0.001466 682.027   <2e-16 ***
```

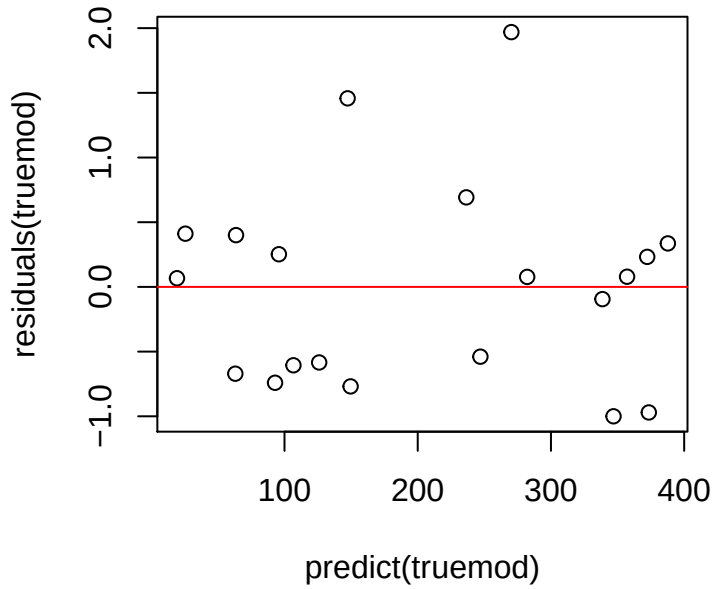
```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.826 on 17 degrees of freedom
Multiple R-squared:      1,Adjusted R-squared:      1
F-statistic: 2.327e+05 on 2 and 17 DF,  p-value: < 2.2e-16

> hist(residuals(truemod),prob=TRUE)
> curve(dnorm(x,mean=0,sd=sd(residuals(truemod))),add=TRUE)
```




```
> qqnorm(residuals(truemod))
```




```
> qqline(residuals(truemod), col = 2)

Error:  $ operator is invalid for atomic vectors

> plot(predict(truemod),residuals(truemod))
> abline(h=0,col="red")
```




```
> shapiro.test(residuals(truemod))


Shapiro-Wilk normality test

data:  residuals(truemod)
W = 0.91303, p-value = 0.07281
```