STAT 210
Applied Statistics and Data Analysis:
Homework 6 - Solution

Due on November 09/2025

You cannot use artificial intelligence tools to solve this homework.

Show complete solutions to get full credit. Writing code is not enough to answer a question. Your comments are more important than the code. Do not write comments in chunks. Label your graphs appropriately

For all tests in this HW use a significance level of $\alpha = 0.02$.

Question 1 (50 pts)

The data for this question is stored in the file `CHFLS` in the library `HSAUR3` and comes from a survey of 60 villages and urban neighborhoods in China published in 2003. It has 1534 observations of 10 variables, but we will focus on `R_age` (age), `R_happy` (self-reportet happiness), and `R_region` (region).

(a) Creat a new data frame called `df1` that only includes `R_age`, `R_happy` and `R_region`. Check whether the new data frame has missing data. Explore the distribution of `R_age` for the different regions. Do boxplots of age as a function of region and comment on what you observe. Calculate mean, standard deviation, median and interquartile range for `R_age` for each of the six regions and comment.

Solution:

Load the data

```
library(HSAUR3)
str(CHFLS)
```

```
## 'data.frame':    1534 obs. of  10 variables:
##  $ R_region: Factor w/ 6 levels "Coastal South",..: 5 5 5 5 5 5 5 5 5 5 ...
##  $ R_age   : num  54 46 48 46 45 36 48 36 20 30 ...
##  $ R_edu   : Ord.factor w/ 6 levels "Never attended school"<..: 4 4 4 3 3 4 3 3 3 4 ...
##  $ R_income: num  900 500 800 300 300 500 0 100 200 400 ...
##  $ R_health: Ord.factor w/ 5 levels "Poor"<"Not good"<..: 4 3 4 3 3 5 2 4 3 4 ...
##  $ R_height: num  165 156 163 164 162 161 167 156 158 160 ...
##  $ R_happy : Ord.factor w/ 4 levels "Very unhappy"<..: 3 3 3 3 3 3 4 2 2 3 ...
##  $ A_height: num  172 170 172 174 172 180 168 173 178 176 ...
##  $ A_edu   : Ord.factor w/ 6 levels "Never attended school"<..: 4 4 3 2 3 5 3 4 5 4 ...
##  $ A_income: num  500 800 700 700 400 900 300 800 200 600 ...
```

Create the new data frame and check for missing values.

```
df1 <- subset(CHFLS, select = c(R_age, R_happy, R_region))
str(df1)
```

```
## 'data.frame':    1534 obs. of  3 variables:
```

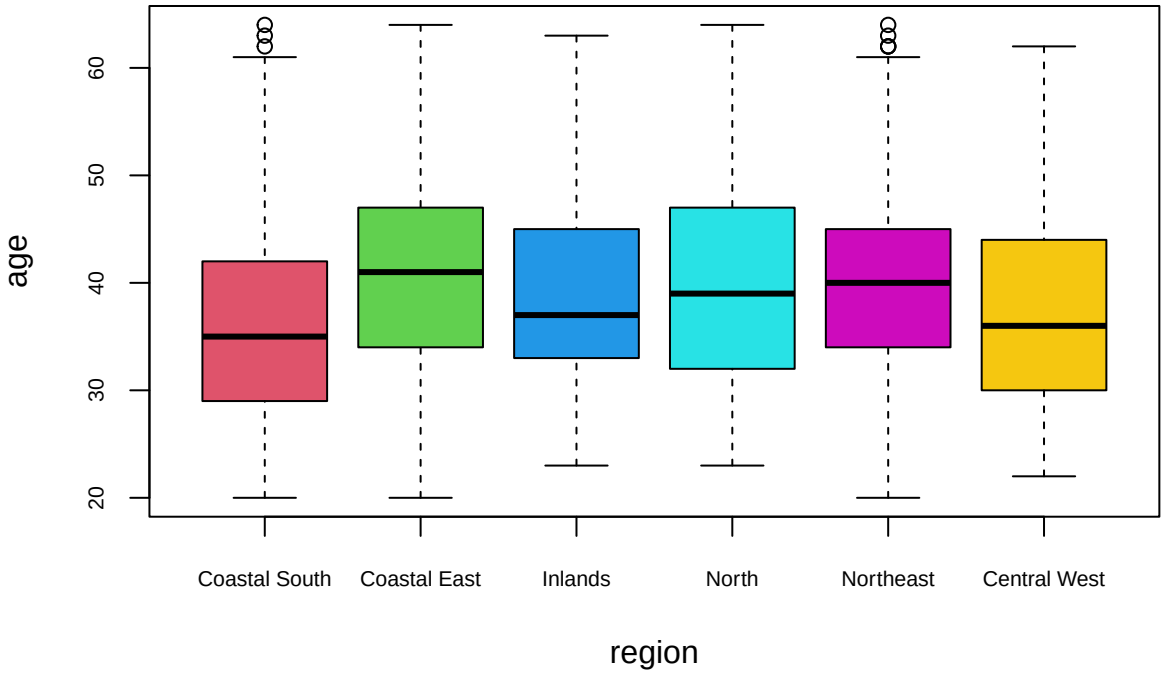
```
##  $ R_age   : num  54 46 48 46 45 36 48 36 20 30 ...
##  $ R_happy : Ord.factor w/ 4 levels "Very unhappy"<..: 3 3 3 3 3 3 4 2 2 3 ...
##  $ R_region: Factor w/ 6 levels "Coastal South",..: 5 5 5 5 5 5 5 5 5 5 ...
```

```
sum(is.na(df1))
```

```
## [1] 0
```

There are no missing values. Boxplots:

```
plot(R_age ~ R_region, data = df1, col = 2:7, cex.axis = 0.7,
     xlab = 'region', ylab = 'age')
```



The differences we observe are moderate. Perhaps the largest difference is between Coastal South and Coastal East. The first one has the youngest population among the six regions, with the lowest median and quartiles, while the second has the largest values.

Empirical parameters for the regions:

```
options(width = 100)
cat('Means','\n')
tapply(df1$R_age, df1$R_region, mean)
cat('Standard Deviations','\n')
tapply(df1$R_age, df1$R_region, sd)
cat('Medians','\n')
tapply(df1$R_age, df1$R_region, median)
cat('IQR','\n')
tapply(df1$R_age, df1$R_region, IQR)
```

```
## Means
## Coastal South  Coastal East       Inlands         North     Northeast  Central West
##      36.04702      41.05740      39.64103      39.93361      40.02867      37.26923
## Standard Deviations
## Coastal South  Coastal East       Inlands         North     Northeast  Central West
##      9.255339      9.743982      9.266623      9.528234      9.100512      9.423202
## Medians
## Coastal South  Coastal East       Inlands         North     Northeast  Central West
```

```
##            35            41            37            39            40            36
## IQR
## Coastal South  Coastal East       Inlands         North     Northeast  Central West
##            13            13            12            15            11            14
```

The parameters confirm our observations, with Coastal South having the lowest values for mean and median, while Coastal East has the largest values. The standard deviations are similar for all regions.

(b) Create a new ordered factor in `df1` named `R_a` by dividing `R_age` into five groups having approximately the same number of subjects. Name the levels `a1, a2, a3, a4,` and `a5`.

Solution:

New factor

```
df1$R_a <- cut(df1$R_age, quantile(df1$R_age,c(0,.2, .4, .6, .8, 1) ),
               labels = c('a1', 'a2', 'a3', 'a4', 'a5'),
               include.lowest = T,ordered_result = T)
str(df1)
```

```
## 'data.frame':    1534 obs. of  4 variables:
##  $ R_age   : num  54 46 48 46 45 36 48 36 20 30 ...
##  $ R_happy : Ord.factor w/ 4 levels "Very unhappy"<..: 3 3 3 3 3 3 4 2 2 3 ...
##  $ R_region: Factor w/ 6 levels "Coastal South",..: 5 5 5 5 5 5 5 5 5 5 ...
##  $ R_a     : Ord.factor w/ 5 levels "a1"<"a2"<"a3"<..: 5 4 5 4 4 2 5 2 1 1 ...
```

(c) Produce a table of `R_region` against `R_a` (age should be in the columns of the table). Graph a mosaic plot of this table using different colors for the rectangles. Comment on what you observe. Produce a second table with proportions relative to the regions. Comment.

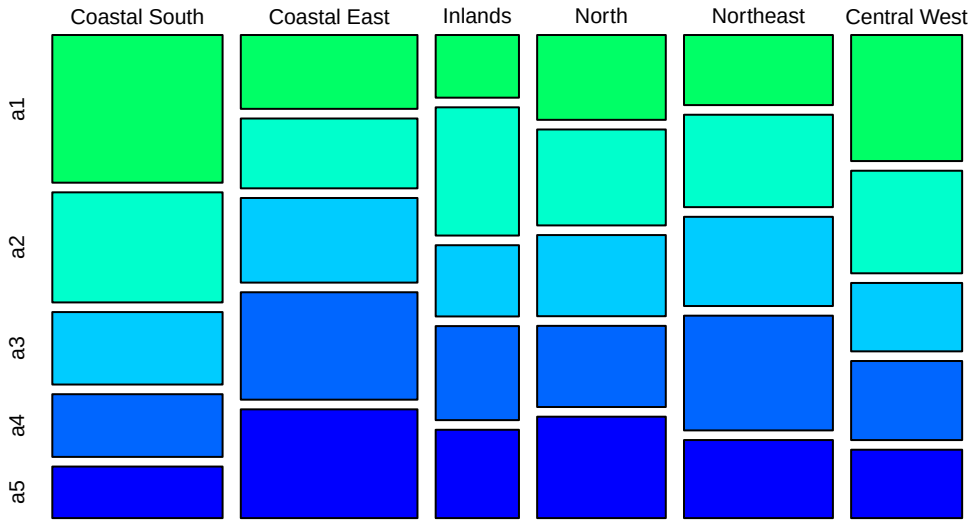
Solution:

Table

```
(tab1 <- table(df1$R_region, df1$R_a))
```

```
##
##                  a1  a2  a3  a4  a5
##   Coastal South 106  79  52  45  37
##   Coastal East   55  52  63  80  81
##   Inlands        22  45  25  33  31
##   North          46  52  44  44  55
##   Northeast      44  58  56  72  49
##   Central West   59  48  32  37  32
```

```
mosaicplot(tab1, color = rainbow(30)[11+2*1:5])
```



Observe that in the plot, the youngest group (`a1`) is represented on top. Coastal South has the largest proportion of people in this group, followed by Central West and North. Coastal East and Northeast have similar proportions and Inlands comes last. On the other hand, for the oldest group, which corresponds to class `a5`, Coastal East has the largest proportion, followed by North, Inlands, Northeast, central West and Coastal South.

In general, the distribution of the different age groups seems to have important differences in the regions represented in the study.

Second table

```
tab2 = prop.table(tab1, margin = 1)
print(round(tab2,3))
```

```
##
##                    a1    a2    a3    a4    a5
##   Coastal South 0.332 0.248 0.163 0.141 0.116
##   Coastal East  0.166 0.157 0.190 0.242 0.245
##   Inlands       0.141 0.288 0.160 0.212 0.199
##   North         0.191 0.216 0.183 0.183 0.228
##   Northeast     0.158 0.208 0.201 0.258 0.176
##   Central West  0.284 0.231 0.154 0.178 0.154
```

```
rowSums(tab2)
```

```
## Coastal South  Coastal East       Inlands         North     Northeast  Central West
##             1             1             1             1             1             1
```

We confirm the conclusions from the mosaic plot, and observed that there are important differences in the proportions of the different age groups in the six regions considered. For instance, the proportion of subjects in age group `a1` in Coastal South, is more than twice that of Inlands and Norteast, while the proportion of subjects in the oldest group, `a5`, in Coastal East is more than twice that of Coastal South.

(d) You want to determine whether the age groups have a homogeneous distribution across regions. Which test (or tests) do you know that can be used for this? What are the underlying assumptions? Are they satisfied in this case? Carry out all the tests that apply and discuss the results. What are your conclusions?

Solution:

We can use the Chi-square test to check for homogeneous distributions across regions. The condition is that the entries in the matrix of expected values have to be all above 5. We check this after doing the test. The test is

```
(csq1 <- chisq.test(tab1))
```

```
##
##  Pearson's Chi-squared test
##
## data:  tab1
## X-squared = 82.7, df = 20, p-value = 1.36e-09
```

The p-value is small and we reject the null hypothesis of homogeneous distributions across regions. To check the condition for the validity of the test, we print the table of expected values:

```
round(csq1$expected,2)
```

```
##
##                    a1    a2    a3    a4    a5
##   Coastal South 69.04 69.46 56.56 64.67 59.27
##   Coastal East  71.64 72.07 58.69 67.11 61.50
##   Inlands       33.76 33.97 27.66 31.63 28.98
##   North         52.16 52.47 42.73 48.86 44.78
##   Northeast     60.38 60.75 49.47 56.56 51.84
##   Central West  45.02 45.29 36.88 42.17 38.64
```

We see that all the values are large and the condition is satisfied.

(e) To explore the relation between age and happiness, build a contingency table for `R_happy` against `R_a`. Use the Chi-square test on this table. Are the conditions for the test satisfied? Why or why not?

Solution:

Print the table:

```
(tab3 <- table(df1$R_happy, df1$R_a))
```

```
##
##                   a1  a2  a3  a4  a5
##   Very unhappy     2   1   1   5   5
##   Not too happy   33  45  35  40  32
##   Somewhat happy 231 228 184 213 199
##   Very happy      66  60  52  53  49
```

In this table we observe that very few people are in the `Very unhappy` category. We have to print the table of expected values to check if the test is applicable to this table. The test is

```
(csq2 <- chisq.test(tab3))
```

```
## Warning in chisq.test(tab3): Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  tab3
## X-squared = 9.9143, df = 12, p-value = 0.6235
```

The warining in the output says that the condition is not verified. We check by printing the table of expected values:

```
round(csq2$expected,2)
```

```
##
##                       a1     a2     a3     a4     a5
##   Very unhappy      3.03   3.05   2.48   2.84   2.60
##   Not too happy    40.04  40.28  32.80  37.51  34.37
##   Somewhat happy 228.33 229.71 187.07 213.89 196.01
##   Very happy       60.60  60.96  49.65  56.77  52.02
```

All values in the first row are below five, so the condition is not satisfied.

(f) Create a new ordered factor called `R_h` in `df1` by joining the two lower levels of `R_happy`, i.e., `R_h` will have three level named `unhappy`, `Somewhat happy`, and `very happy`. The values for `unhappy` come from re-naming the levels `Very unhappy` and `Not too happy` as `unhappy`. One easy way to do this is to use the `labels` argument. Look at the help of the `factor` function to see how this is done.

Solution:

We create the new factor using the `labels` option in the function call:

```
df1$R_h <- factor(df1$R_happy,labels =  c('unhappy','unhappy','happy','very happy'))
str(df1)
```

```
## 'data.frame':    1534 obs. of  5 variables:
##  $ R_age   : num  54 46 48 46 45 36 48 36 20 30 ...
##  $ R_happy : Ord.factor w/ 4 levels "Very unhappy"<..: 3 3 3 3 3 3 4 2 2 3 ...
##  $ R_region: Factor w/ 6 levels "Coastal South",..: 5 5 5 5 5 5 5 5 5 5 ...
##  $ R_a     : Ord.factor w/ 5 levels "a1"<"a2"<"a3"<..: 5 4 5 4 4 2 5 2 1 1 ...
##  $ R_h     : Ord.factor w/ 3 levels "unhappy"<"happy"<..: 2 2 2 2 2 2 3 1 1 2 ...
```

(g) Build a contingency table for `R_h` against `R_a`. Graph a mosaic plot of this table using different colors for the rectangles. Comment on what you observe. Use the Chi-square test on this table. Are the conditions for the test satisfied? Why or why not? What is your conclusion after using the test?

Solution:

```
(tab4 <- table(df1$R_h, df1$R_a))
```

```
##
##               a1  a2  a3  a4  a5
##   unhappy     35  46  36  45  37
##   happy      231 228 184 213 199
##   very happy  66  60  52  53  49
```

```
mosaicplot(tab4, color = rainbow(30)[11+2*1:5])
```

There are no significant differences in the distributions accross ages. We use the Chi-square test:

```
(csq3 <- chisq.test(tab4))
```

```
##
##  Pearson's Chi-squared test
##
## data:  tab4
## X-squared = 3.4049, df = 8, p-value = 0.9064
```

```
round(csq3$expected,3)
```

```
##
##                    a1      a2      a3      a4      a5
##   unhappy      43.069  43.329  35.286  40.345  36.972
##   happy       228.331 229.707 187.066 213.889 196.007
##   very happy   60.600  60.965  49.648  56.767  52.021
```

The conditions for the test are satisfied. The p-value is large and we do not reject the null hypothesis of homogeneous distributions.

Question 2 (50 pts)

For this problem use the data set **women** which is available in R, and has 'average heights and weights for American women aged 30-39', according to the help file for the data set. Height is measured in inches while weight is measured in pounds.

Load the data:

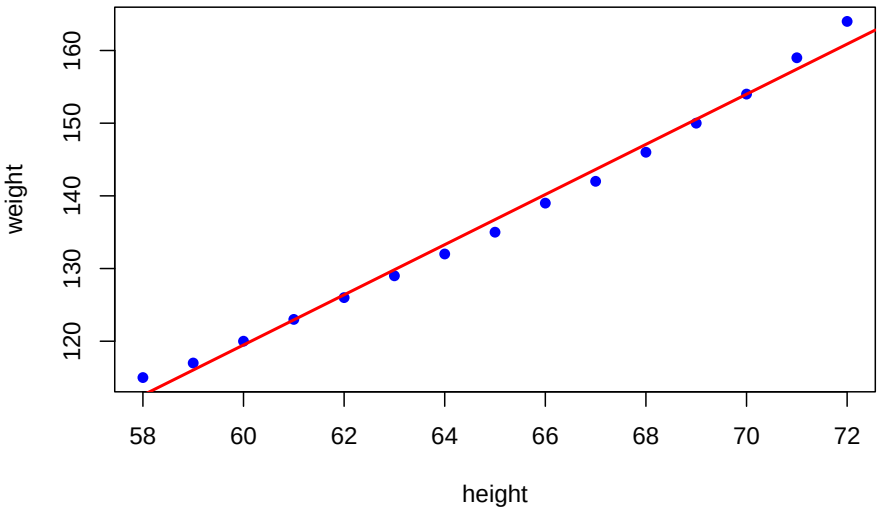
```
data(women)
str(women)
```

```
## 'data.frame':    15 obs. of  2 variables:
##  $ height: num  58 59 60 61 62 63 64 65 66 67 ...
##  $ weight: num  115 117 120 123 126 129 132 135 139 142 ...
```

(a) Fit a simple linear regression model for **weight** as a function of **height**. Produce a scatterplot and add the regression line.

Fit the model:

```
mod1 <- lm(weight ~ height, data = women)
plot(weight ~ height, data = women, pch = 16, col = 'blue')
abline(mod1, col = 'red', lwd = 2)
```



Although the regression line follows the trend in the scatterplot, the fit is not good. The points in the central part of the plot are all below the regression line, while the points at the extremes are above, indicating that there is a non-linear relation between the variables that is not captured by the model.

(b) Print a summary table for the model and interpret the results. Write an equation for the model and interpret the coefficients. What is the R^2 for this model?

Summary table:

```
summary(mod1)
```

```
##
## Call:
## lm(formula = weight ~ height, data = women)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7333 -1.1333 -0.3833  0.7417  3.1167
##
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -87.51667    5.93694  -14.74 1.71e-09 ***
## height        3.45000    0.09114   37.85 1.09e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.525 on 13 degrees of freedom
## Multiple R-squared:  0.991,  Adjusted R-squared:  0.9903
## F-statistic:  1433 on 1 and 13 DF,  p-value: 1.091e-14
```

We see some assymmetry in the residual summary at the top. Both coefficients have very small p-values, indicating that they are significantly different from zero. The standard deviation for the errors is 1.525 while the R^2 is 0.991, a very high value, indicating that the model explains most of the variability in the data.

(c) Predict the weight for a woman of 65 inches including a confidence interval.

We predict the `weight` including a CI:

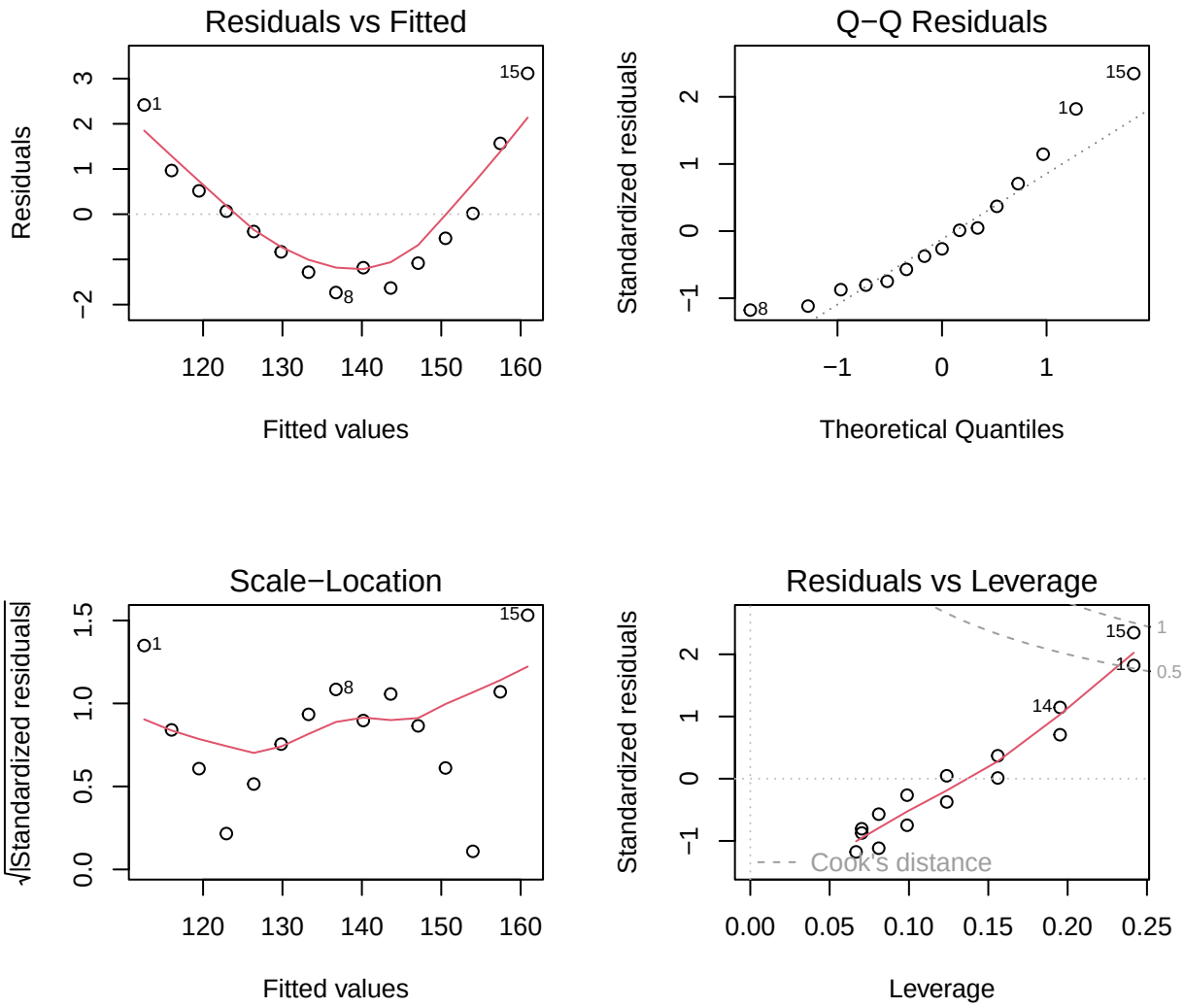
```
predict(mod1, data.frame(height = 65), interval = 'c', level = 0.98)
```

```
##        fit      lwr      upr
## 1 136.7333 135.6898 137.7769
```

(d) State explicitly the assumptions on which the model is based and using plots and tests verify whether they are satisfied.

The model is based on the assumptions that the errors are independent and have a common normal distribution with mean zero and unknown variance σ^2. We start by ploting the diagnostic graphs

```
par(mfrow = c(2,2))
plot(mod1)
```




```
par(mfrow = c(2,2))
```

The plots show that the assumptions are not satisfied. The quantile plot for normality does not show an adequate fit, while the residuals vs fitted values plot shows a U pattern that indicates that the model is not adequate. The residual vs leverage plot shows an increasing pattern and two of the values are beyond the 0.5 countour line for Cook's distance.

We also use ths Shapiro-Wilk and ncv tests:

```
shapiro.test(rstandard(mod1))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  rstandard(mod1)
## W = 0.90662, p-value = 0.1202
```

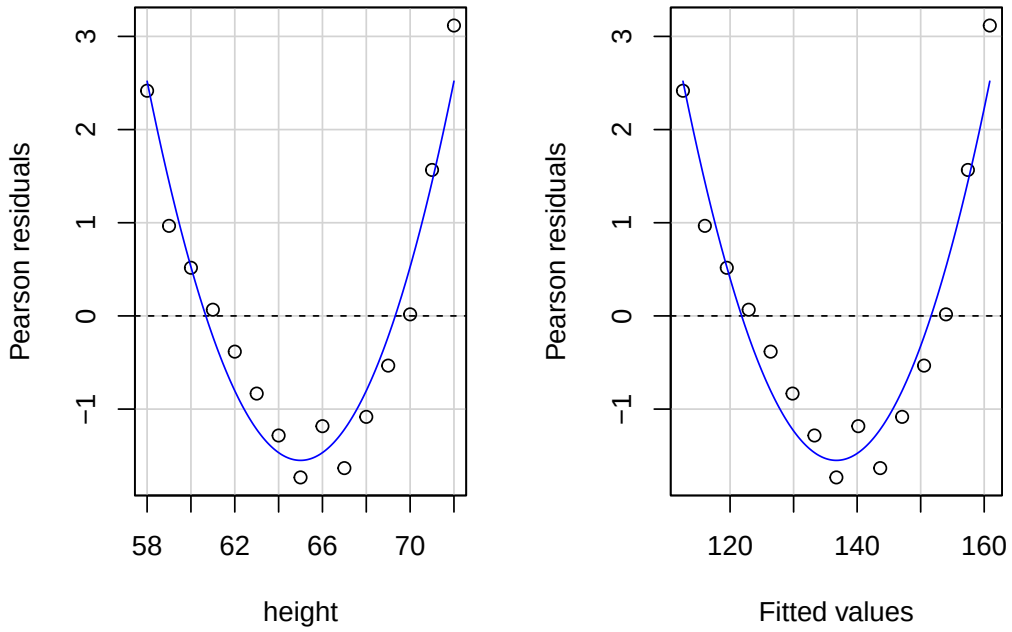
```
library(car)
ncvTest(mod1)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.8052115, Df = 1, p = 0.36954
```

Both p-values are above the significance threshold, and we do not reject the hypotheses of normality and constant variance.

(e) Use the function `residualPlots` in the `car` library. The argument of the function is the name of your model and the output is a couple of residual plots plus some summary information about two tests. The first plot is residuals vs. the regressor (`height` in this case) and the second is residuals against fitted values. In both cases, the blue line shown in the plot is not a local smoother but a quadratic term added to the model. If the line is flat, it indicates that the term will not improve the model. The results of tests of curvature are also shown. We will consider only the first one, corresponding to `height`. This is a test on the coefficient for a quadratic term in `height` added to the model. The null hypothesis is that coefficient corresponding to the quadratic term is zero. Interpret the output you get when using this function on your model.

```
residualPlots(mod1)
```




```
##             Test stat Pr(>|Test stat|)
## height         13.891        9.322e-09 ***
## Tukey test     13.891        < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The interpretation is clear, we should include a quadratic term in the model.

(f) Add a quadratic term in `height` to your model (you have to use the expression `I(height^2)` in the equation to do this). Print the summary table and interpret the results. What is the R^2 for this model and how does it compare with the previous model?

```
mod2 <- lm(weight ~ height + I(height^2), data = women)
summary(mod2)
```

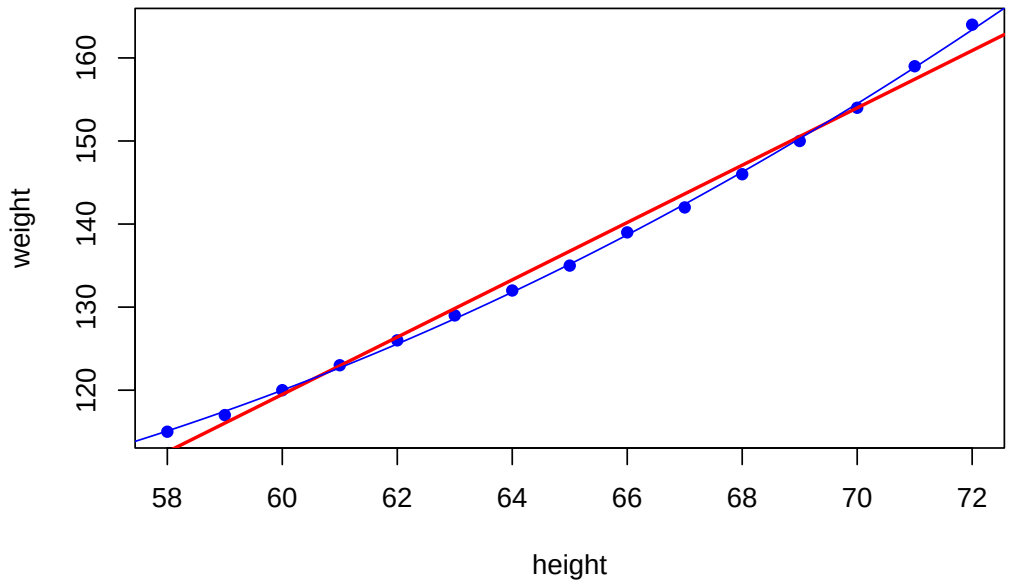
```
##
## Call:
## lm(formula = weight ~ height + I(height^2), data = women)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.50941 -0.29611 -0.00941  0.28615  0.59706
```

```
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 261.87818   25.19677  10.393 2.36e-07 ***
## height       -7.34832    0.77769  -9.449 6.58e-07 ***
## I(height^2)   0.08306    0.00598  13.891 9.32e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3841 on 12 degrees of freedom
## Multiple R-squared:  0.9995, Adjusted R-squared:  0.9994
## F-statistic: 1.139e+04 on 2 and 12 DF,  p-value: < 2.2e-16
```

The residual summary at the top of the table shows that the residuals seem to be symmetric. All the coefficients have small p-values, indicating that they are significantly different from zero. The standard deviation for the errors is 0.3841, down from 1.525 while the R^2 is 0.9995, up from 0.991.

(g) Plot the data and add the lines corresponding to the two model you fitted.

```
plot(weight ~ height, data = women, pch = 16, col = 'blue')
abline(mod1, col = 'red', lwd = 2)
curve(261.87818 -7.34832*x +0.08306*x^2, 57, 73, col = 'blue', add = T)
```



We see that the quadratic model has a much better fit to the data.

(h) Write down an equation for the final model. Predict the weight for a woman of 65 inches including a confidence interval using the quadratic model and compare with your previous prediction.

The equation for the model is

$$\texttt{weight} = 261.87818 - 7.34832 \cdot \texttt{height} + 0.08306 \cdot \texttt{height}^2.$$

The predicted value is

```
predict(mod2, data.frame(height = 65), interval = 'c', level = 0.98)
```

```
##        fit      lwr      upr
## 1 135.1828 134.7825 135.5831
```

Compare with previous prediction:

```r
predict(mod1, data.frame(height = 65), interval = 'c', level = 0.98)
```

```
##        fit      lwr      upr
## 1 136.7333 135.6898 137.7769
```

The new predicted value is smaller and the two confidence intervals are disjoint.