STAT 210
Applied Statistics and Data Analysis:
Homework 7 - Solution

Due on November 16, 2025

Question 1

For this question use the data set `25Fhw7Q1.csv`.

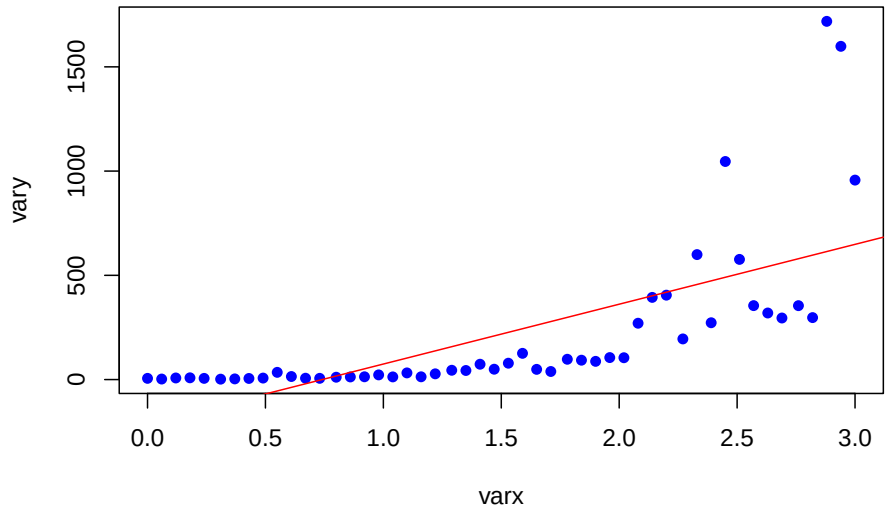
(i) Read the data and plot `vary` as a function of `varx`. Fit a simple linear regression for `vary` as a function of `varx` and add the regression line to the plot. Comment. Obtain a summary for the regression and interpret the result. Draw the diagnostic plots and use appropriate tests to check for normality and homogeneous variances. Comment on the results. What would be the predicted `vary` for a point with `varx` = 2.5 using this model? Include a confidence interval at the 98% confidence level.

```
dat1 <- read.csv('25Fhw7Q1.csv')
str(dat1)
```

```
## 'data.frame':    50 obs. of  2 variables:
##  $ varx: num  0 0.06 0.12 0.18 0.24 0.31 0.37 0.43 0.49 0.55 ...
##  $ vary: num  5.8 2.53 7.71 8.42 5.69 ...
```

The data set has only two variables, `varx` and `vary`. There are 50 observations of each.

```
plot(vary ~ varx, data = dat1, pch = 16, col = 'blue')
md1 <- lm(vary ~ varx, data = dat1)
abline(md1, col = 'red')
```



We observe that the fit is not good. There is an increasing trend for `vary` as `varx` increases, but the relation is not linear.

```
summary(md1)
```

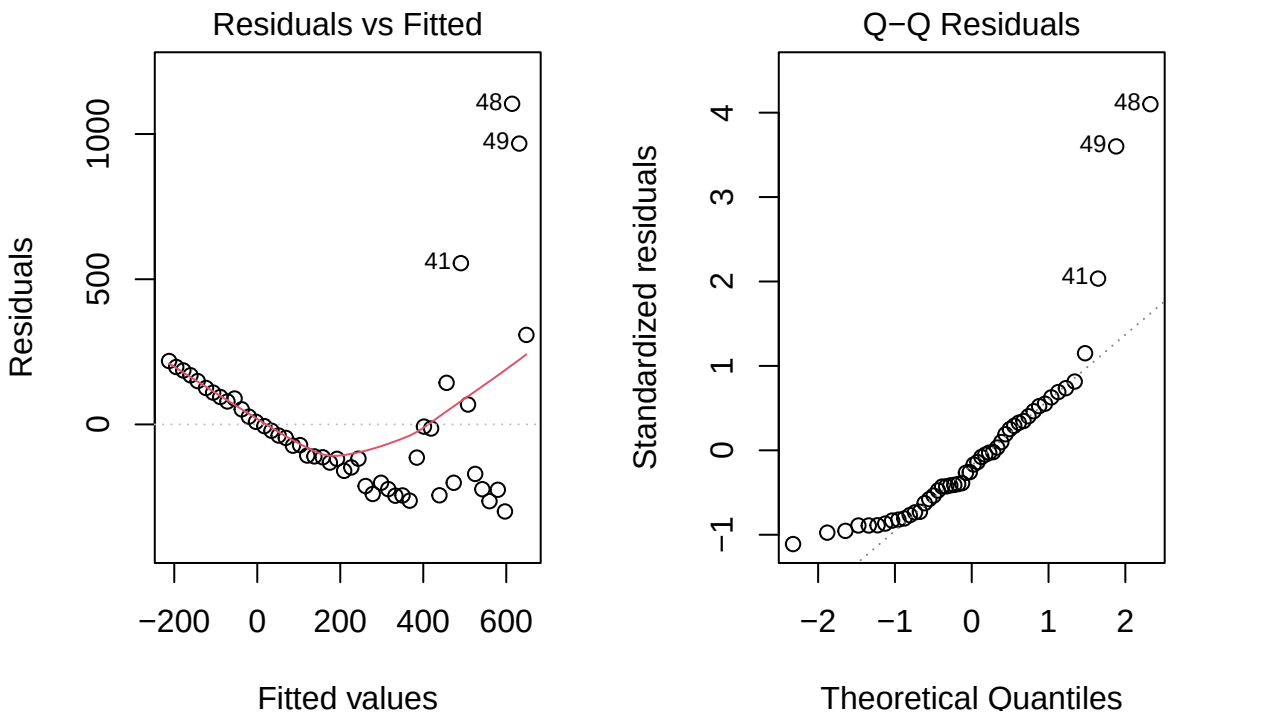
```
##
```

```
## Call:
## lm(formula = vary ~ varx, data = dat1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -299.66 -193.26  -58.74   93.20 1104.11
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -212.55      77.71  -2.735  0.00871 **
## varx          287.02      44.63   6.431 5.49e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 279 on 48 degrees of freedom
## Multiple R-squared:  0.4628, Adjusted R-squared:  0.4516
## F-statistic: 41.35 on 1 and 48 DF,  p-value: 5.494e-08
```

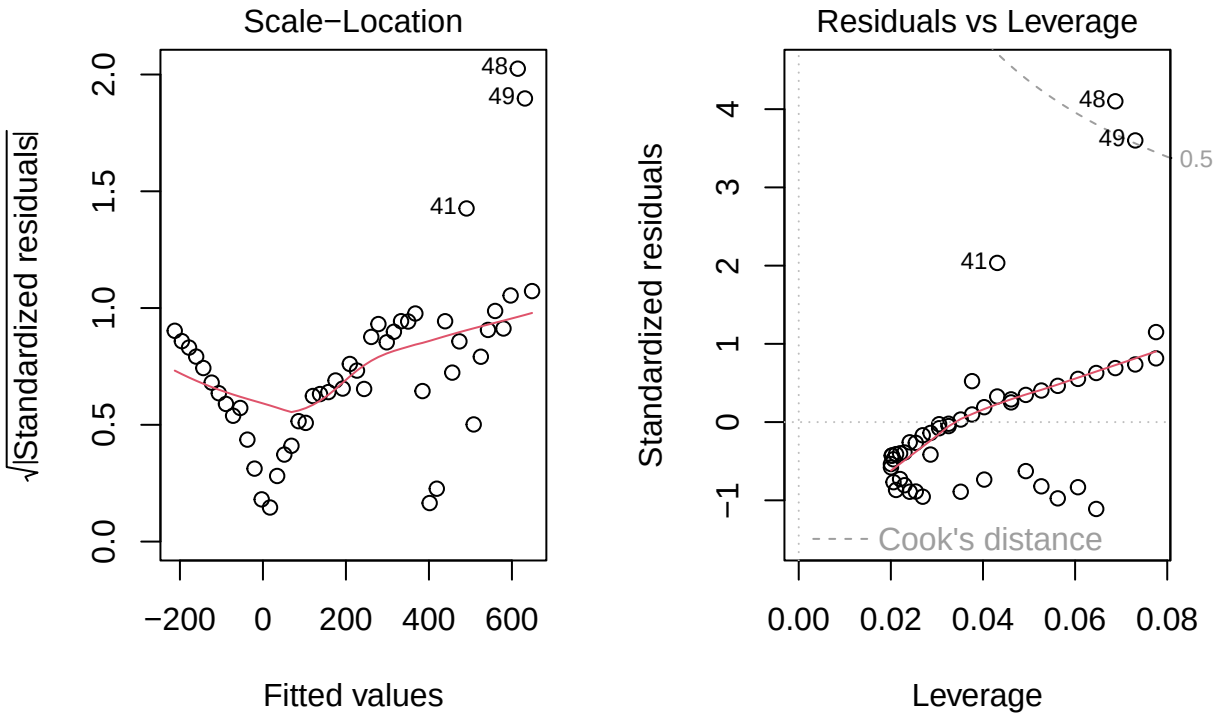
The summary for the residuals at the top of the table shows very asymmetric values, indicating that the assumptions for the model are probably not valid. Both estimated parameters have small p-values, indicating that they are significantly different from zero, and the estimated standard deviation for the errors is 279. The R^2 is about 46%.

The diagnostic plots are

```
par(mfrow = c(1,2))
plot(md1, which = c(1,2))
```




```
plot(md1, which = c(3,5))
```




```
par(mfrow = c(1,1))
```

All the plots show violations of the assumptions. The residuals vs fitted plot shows the residuals are not symmetrically distributed and the variance is not uniform. The local smoother shows a curvature that indicates that the model does not account for all the information in the data. The quantile plot shows clear deviations from the reference line. The scale-location shows the residuals do not have homogeneous variances and also that the size of residuals decreases and then increases, showing a pattern that one does not expect to see. Finally, the residuals vs leverage plot also shows an unexpected pattern, and some very large values for Cooks distance for two points. The conclusion is that this is not a good model. The Shapiro-Wilk and ncv tests below confirm this conclusion.

```
shapiro.test(rstandard(md1))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  rstandard(md1)
## W = 0.77134, p-value = 2.036e-07
```

```
library(car)
ncvTest(md1)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 32.0961, Df = 1, p = 1.4673e-08
```

Both p-values are small indicating that the assumptions of normality and homogeneous variances are not valid for the residuals.

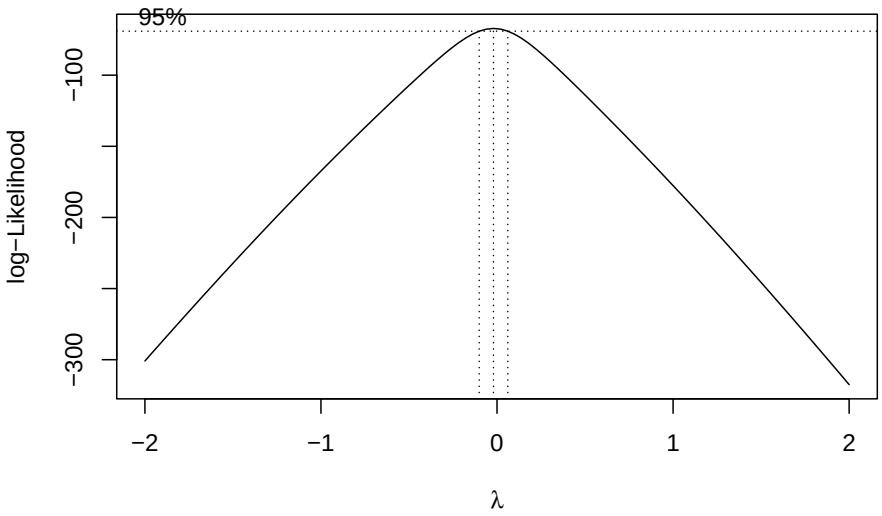
The predicted value with 98% confidence interval is

```
predict(md1, data.frame(varx = 2.5), interval = 'c', level = 0.98)
```

```
##        fit      lwr      upr
## 1 505.0074 361.6351 648.3796
```

(ii) In this part you have to use the Box-Cox transformations to improve the model. To simplify this problem, you have to choose between two transformations of the output variable `vary`, a square root or a logarithm. Use the function `boxcox` on the package `MASS` with the argument set to the model you fitted in (i). If the confidence interval in the graph includes zero, choose a logarithmic transformation for `vary`. If the confidence interval in the graph includes 0.5 then choose a square root transformation.

```
library(MASS)
boxcox(md1)
```



The confidence interval includes zero, so we proceed with a logarithmic transformation.

(iii) Fit a new model with the transformation that you choose in (ii). Obtain a summary for the new regression and compare with the previous one. Draw the diagnostic plots and compare with the previous results. Use appropriate tests to check for normality and homogeneous variances.

Fit the model:

```
md2  <- lm(log(vary) ~ varx, data = dat1)
summary(md2)
```

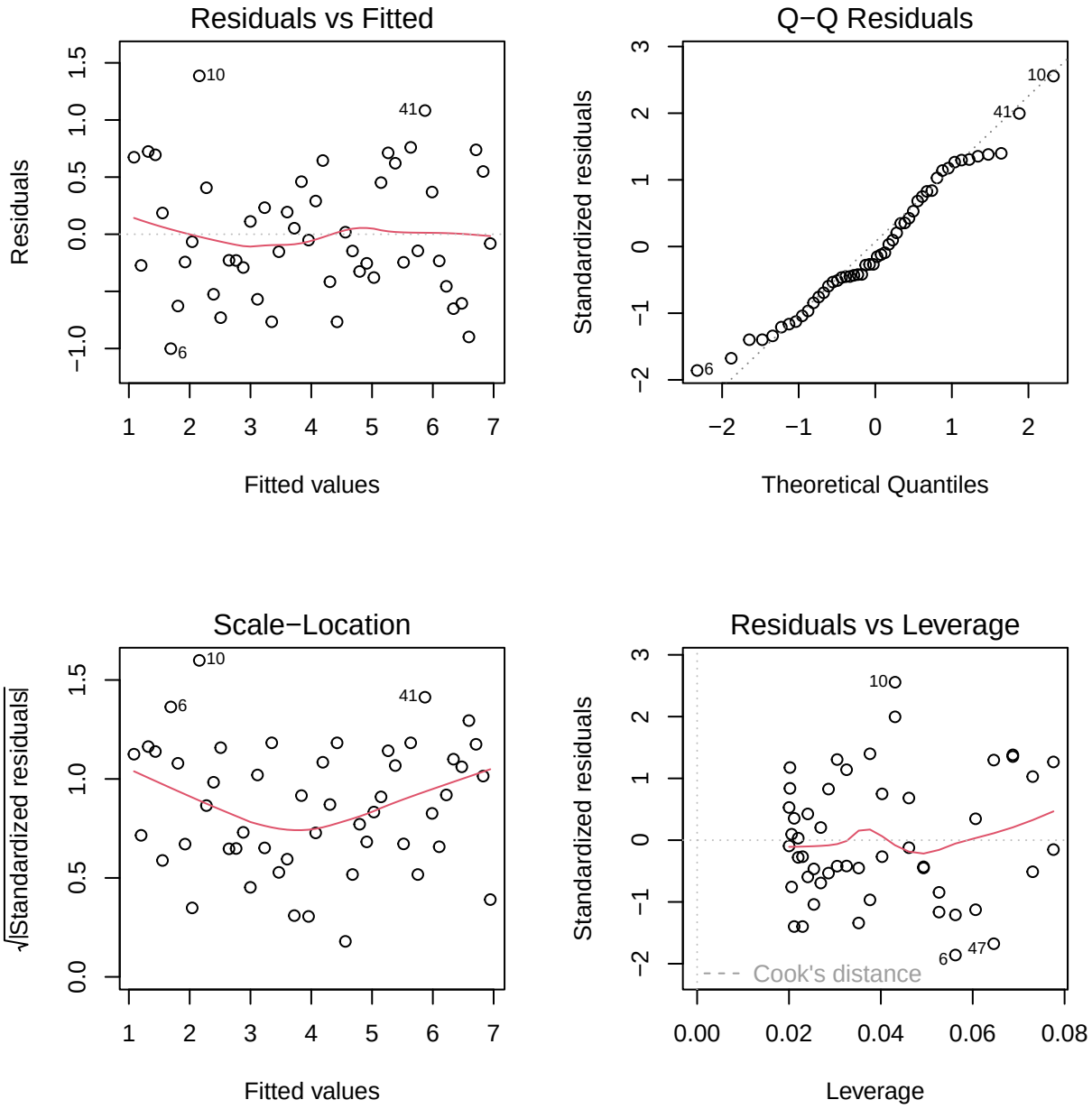
```
##
## Call:
## lm(formula = log(vary) ~ varx, data = dat1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.0014 -0.3658 -0.1132  0.4406  1.3863
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.08395    0.15440    7.02 6.84e-09 ***
## varx         1.95362    0.08868   22.03  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5543 on 48 degrees of freedom
## Multiple R-squared:   0.91,  Adjusted R-squared:  0.9081
## F-statistic: 485.4 on 1 and 48 DF,  p-value: < 2.2e-16
```

The summary for residuals shows approximately symmetric quartiles, which is an improvement on the previous model. The two estimated parameters have very small p-values, indicating that they are significantly different

from zero. The estimated standard deviation for the errors is 0.5543, down from 279. The R^2 is about 91%, up from 46%.

The diagnostic plots are

```
par(mfrow = c(2,2))
plot(md2)
```




```
par(mfrow = c(1,1))
```

The only plot that shows an inadequate pattern is the scale-location plot, where the local smoother is not horizontal, possibly indicating non-homogeneous variances in the residuals. We check this with the ncv test below.

```
shapiro.test(rstandard(md2))
```

```
##
```

```
##  Shapiro-Wilk normality test
##
## data:  rstandard(md2)
## W = 0.97334, p-value = 0.3147
```

```
library(car)
ncvTest(md2)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.06548595, Df = 1, p = 0.79803
```

The two tests have large p-values indicating that the null hypothesis of normality and homogeneous variances for the residuals cannot be rejected.

(iv) Write down and equation for the final model in terms of the original variables. What would be the predicted `vary` for a point with `varx` $= 2.5$? Include a confidence interval at the 98% confidence level (to do this, get a confidence interval for the model with the transformed `vary` (using the logarithm or the square root) and use the inverse transformation (exponential or square) on the extremes of the confidence interval). Draw a scatterplot of `vary` against `varx` and add the regression line for the first model and the curve you obtained with the second regression.

The equation for the model is

$$\log(\texttt{vary}) = 1.08395 + 1.95362 \cdot \texttt{varx}$$

or in terms of the original variables

$$\texttt{vary} = \exp(1.08395 + 1.95362 \cdot \texttt{varx}) = 2.956334 e^{1.95362 \cdot \texttt{varx}}$$

The predicted value for log(`vary`) with 98% confidence interval is

```
predict(md2, data.frame(varx = 2.5), interval = 'c', level = 0.98)
```

```
##        fit      lwr      upr
## 1 5.967997 5.683151 6.252844
```

The predicted value for `vary` is obtained exponentiating the previous values:

```
exp(predict(md2, data.frame(varx = 2.5), interval = 'c', level = 0.98))
```

```
##        fit     lwr      upr
## 1 390.7225 293.874 519.4881
```

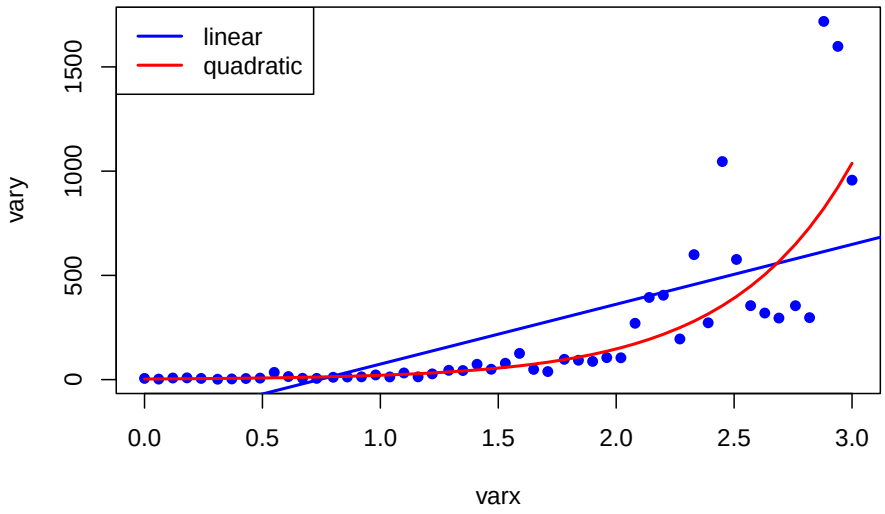
Compare with the prediction with the initial model:

```
predict(md1, data.frame(varx = 2.5), interval ='c', level = 0.98)
```

```
##        fit      lwr      upr
## 1 505.0074 361.6351 648.3796
```

Scatterplot including the two models:

```
plot(vary ~ varx, data = dat1, pch = 16, col = 'blue')
abline(md1, col = 'blue', lwd = 2)
pred_values <- predict(md2)
lines(dat1$varx, exp(pred_values), col = 'red', lwd = 2)
legend('topleft', c('linear', 'quadratic'), col = c('blue', 'red'), lwd = 2)
```


linear
quadratic
vary
varx
0
500
1000
1500
0.0
0.5
1.0
1.5
2.0
2.5
3.0

Question 2

The data for this question is in the file `25Fhw7Q2.csv`

Read the data:

```
dat2 <- read.csv('25Fhw7Q2.csv')
str(dat2)
```

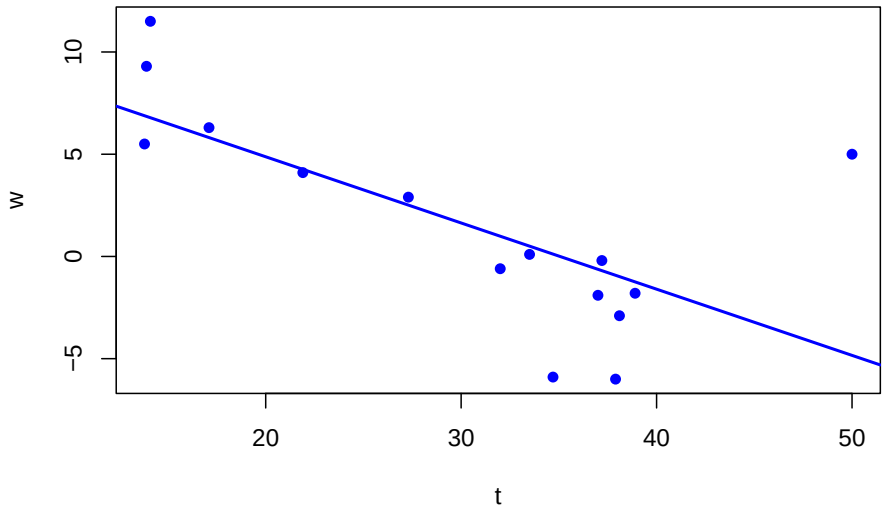
```
## 'data.frame':    15 obs. of  2 variables:
##  $ w: num  -1.9 4.1 5.5 -0.2 2.9 -0.6 9.3 -5.9 -2.9 -6 ...
##  $ t: num  37 21.9 13.8 37.2 27.3 32 13.9 34.7 38.1 37.9 ...
```

The data set has only two variables, `w` and `t`. There are 15 observations of each.

(i) Draw a scatterplot of `w` as a function of `t`. Fit a simple linear regression model and add the line to the plot. Comment. Obtain a summary of the regression and explain the different components of the output. Write down an equation for the model. What is the estimated standard deviation for the errors?

Plot and regression line

```
plot(w ~ t, data = dat2, pch = 16, col = 'blue')
md3 <- lm(w ~ t, data = dat2)
abline(md3, col = 'blue', lwd = 2)
```



The scatterplot shows a decreasing relation between the variables, which is reflected in the regression line. There is also a large amount of variability in the plot.

```
summary(md3)
```

```
##
## Call:
## lm(formula = w ~ t, data = dat2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.0153 -1.4863 -0.4039  0.4899  9.8390
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  11.3515     2.9511   3.847  0.00202 **
```

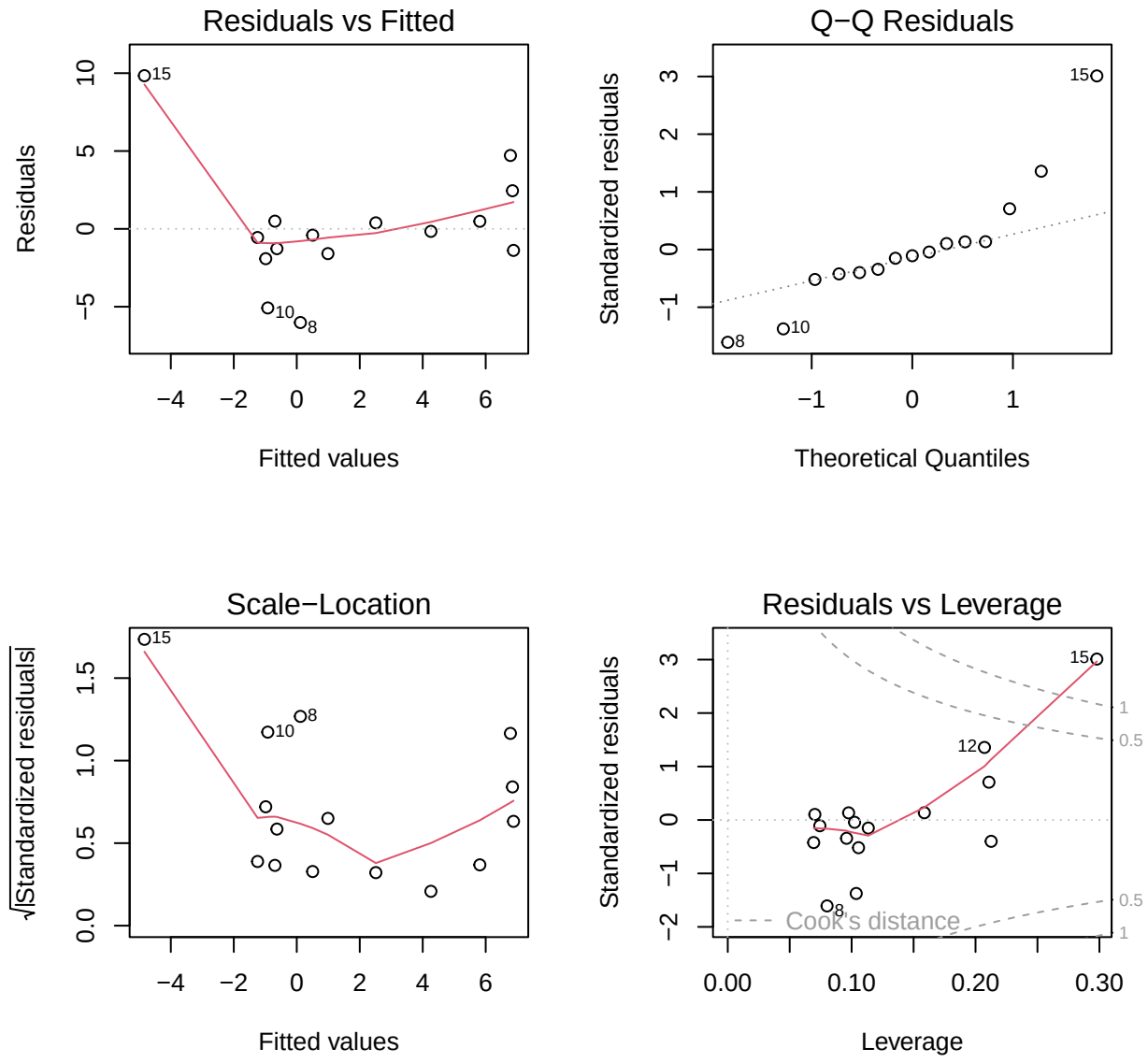
```
## t            -0.3238     0.0930  -3.482  0.00405 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.901 on 13 degrees of freedom
## Multiple R-squared:  0.4826, Adjusted R-squared:  0.4427
## F-statistic: 12.12 on 1 and 13 DF,  p-value: 0.004053
```

The residual summary at the top of the table shows important asymmetries. Both coefficients are significantly differnt from zero. The R^2 is around 48% and the standard deviation for the residuals is 3.9. The equation for this model is

$$\mathtt{w} = 11.3515 - 0.3238 \times \mathtt{t}$$

(ii) Draw the diagnostic plots and comment on what you observe. Do you identify one or more points as outliers? If you do, which points are they? Can you identify these points in the initial scatterplot?

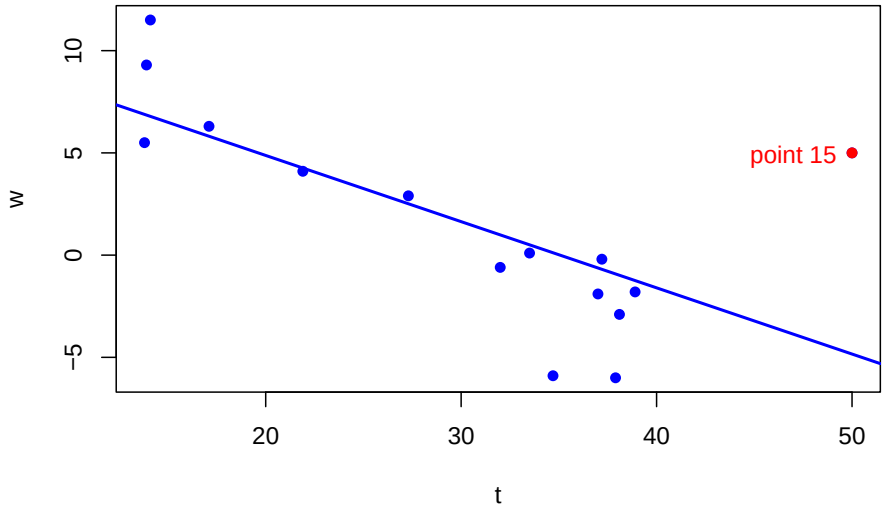
```
par(mfrow = c(2,2))
plot(md3)
```



```
par(mfrow = c(1,1))
```

The plots show that the model does not satisfy the assumptions of normality and homogeneous variance. Point 15 is flagged in all plots as having a more extreme behavior, different from the rest. For instance, in the quantile plot it has a very large distance from the reference line, and in the residuals vs leverage plot it has high leverage and a large residual, with a value for Cooks distance beyond the unit contour line. We conclude that point 15 is an outlier. We identify this point in the scatterplot:

```
plot(w ~ t, data = dat2, pch = 16, col = 'blue')
abline(md3, col = 'blue', lwd = 2)
points(dat2$t[15], dat2$w[15], pch = 16, col ='red')
text(47,4.8,c('point 15'), col = 'red')
```



We also check normality and homoscedasticity using tests (not requested in the statement of the question)

```
shapiro.test(rstandard(md3))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  rstandard(md3)
## W = 0.87004, p-value = 0.03379
```

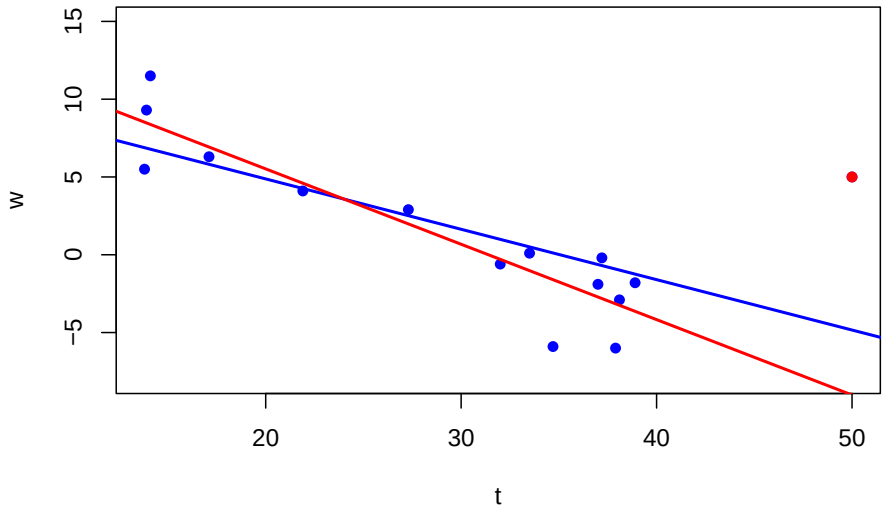
```
ncvTest(md3)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 5.962497, Df = 1, p = 0.014613
```

Both p-values are small, indicating that the null hypotheses are rejected.

(iii) Fit a new regression model excluding the outlier(s) that you identified in the previous section. Draw a scatterplot with both regression lines. Compare the summary tables. Write down an equation for the new model. What is the estimated standard deviation for the errors? Do you think the outliers are influential points?

```
md4 <- lm(w ~ t, data = dat2[-15,])
plot(w ~ t, data = dat2, pch = 16, col = 'blue', ylim = c(-8, 15))
abline(md3, col = 'blue', lwd = 2)
points(dat2$t[15], dat2$w[15], pch = 16, col ='red')
abline(md4, col = 'red', lwd = 2)
```



We see that excluding point 15 produces an important change in the regression line. The new slope is more pronounced. We print the summary table for the second model:

```
summary(md4)
```

```
##
## Call:
## lm(formula = w ~ t, data = dat2[-15, ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2981 -0.5933  0.5827  1.0696  3.1219
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 15.20911    1.84365   8.249 2.74e-06 ***
## t           -0.48447    0.06144  -7.885 4.36e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.236 on 12 degrees of freedom
## Multiple R-squared:  0.8382, Adjusted R-squared:  0.8247
## F-statistic: 62.17 on 1 and 12 DF,  p-value: 4.361e-06
```

The residual summary at the top of the table still shows some asymmetry, but it is less than for the initial model. Both parameters have small p-values indicating that they are significantly different from zero. The table below shows the changes in the coefficients, R^2 and standard deviation of the residuals

	Model 1	Model 2
Intercept	11.35	15.21
Slope	-0.324	-0.485
R^2	0.493	0.828
Standard deviation	3.9	2.24

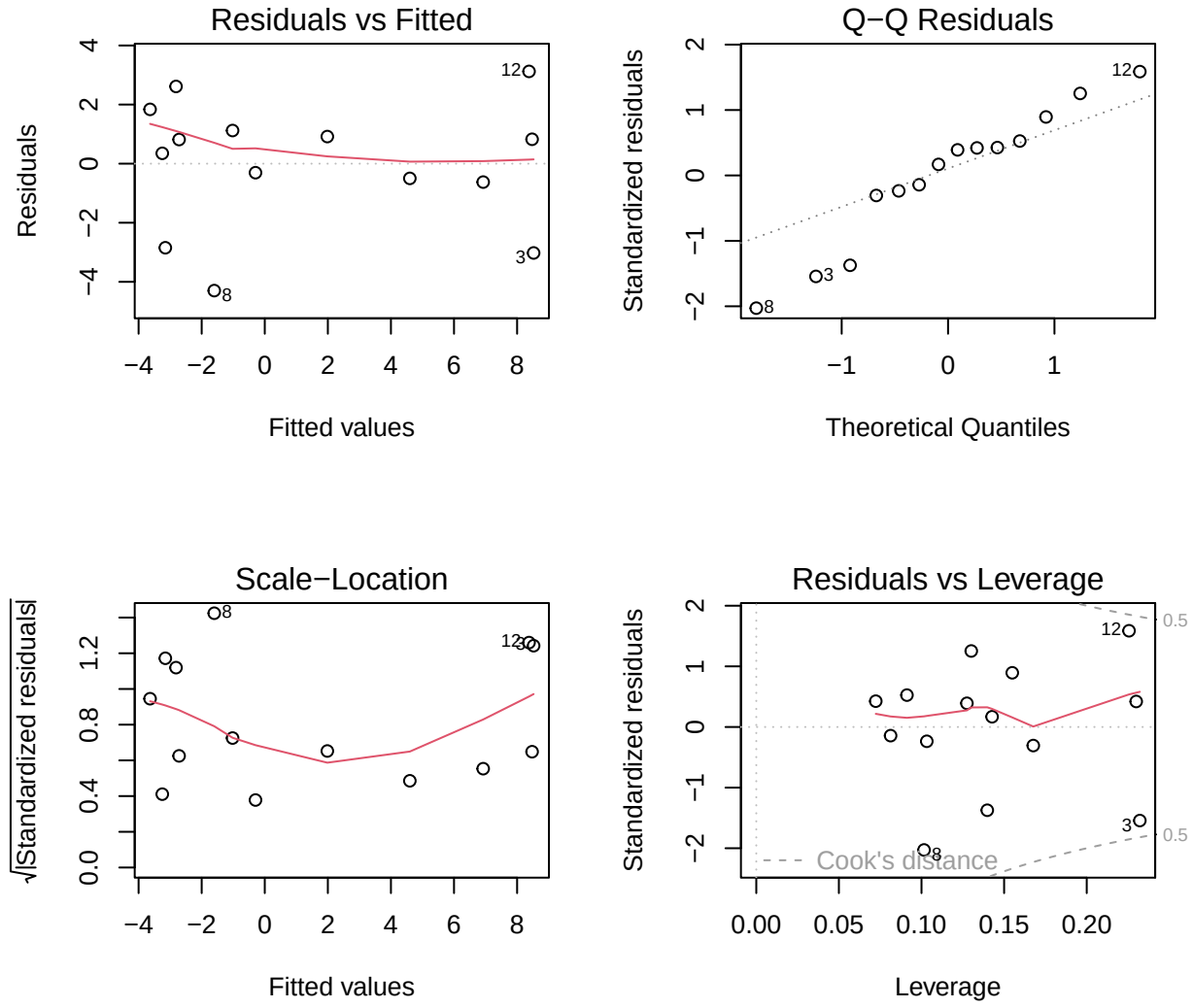
From the table we observe that there are important changes in all these parameters, and we conclude that point 15 is an influential point. The equation for the new model is

$$\texttt{w} = 15.21 - 0.485 \times \texttt{t}.$$

(iv) Draw the diagnostic plots for the new model and comment. Run the Shapiro-Wilk test on the standardized residuals for both models and compare the results. The standardized residuals for

`modelA` can be obtained with the command `rstandard(modelA)`. Check also whether the variances are homogeneous using the `ncvTest` in the `car` package.

```
par(mfrow = c(2,2))
plot(md4)
```




```
par(mfrow = c(1,1))
```

The plots are now much better. There is still some uncertainty regarding the quantile plot, as some points are far from the refernce line, and the scale-location plot also gives indications of non-homogeneous variances. We do the tests:

```
shapiro.test(rstandard(md4))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  rstandard(md4)
## W = 0.94062, p-value = 0.4264
```

```
ncvTest(md4)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
```

```
## Chisquare = 0.003913842, Df = 1, p = 0.95012
```

Both tests now have large p-values, indicating that we do not have evidence to reject the null hypotheses of normality and homoscedasticity.

(v) Describe the sampling distribution for the parameters in the model that excludes the outliers. Give confidence intervals at a confidence level of 99% for the parameters of the regression. Find the predicted value for t = 45 with a confidence interval at the 99% level using both models and compare the results.

The estimated parameters are $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1)$, which have a normal distribution:

$$\hat{\boldsymbol{\beta}} = N\big((\beta_0, \beta_1)', \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\big)$$

The estimated covariance matrix for $\hat{\boldsymbol{\beta}}$ can be obtained with

```
vcov(md4)
```

```
##             (Intercept)           t
## (Intercept)   3.3990338 -0.107168896
## t            -0.1071689  0.003775452
```

The confidence intervals at a confidence level of 99% for the parameters of the regression are

```
confint(md4, level = 0.99)
```

```
##                 0.5 %     99.5 %
## (Intercept)  9.577613 20.8405983
## t           -0.672153 -0.2967825
```

We see that both confidence intervals exclude zero, which agrees with the conclusion that both parameters are significantly different form zero.

The predicted values for $t = 45$ for both model are

```
# Model 1
(p1 <- predict(md3, data.frame(t = 45), interval = 'c', level = 0.99))
```

```
##         fit       lwr      upr
## 1 -3.219938 -8.442476 2.002599
```

```
# Model 2
(p2 <- predict(md4, data.frame(t = 45), interval = 'c', level = 0.99))
```

```
##         fit       lwr       upr
## 1 -6.591943 -10.20499 -2.978895
```

The predicted value with the second model is much smaller, due to the steeper slope. The width of the confidence interval in the second model is about 30% smaller:

```
# Width of CI with first model
p1[3]-p1[2]
```

```
## [1] 10.44507
```

```
# Width of CI with second model
p2[3]-p2[2]
```

```
## [1] 7.226096
```