Data Mining Methods

Simple Linear Regression

Prof. Dr. Christina Andersson

High Integrity Systems
Frankfurt University of Applied Sciences

Booklet to sprint 4

1 Getting Started

In simple linear regression we model the relationship between one input variable and the target variable. Even if this situation with only one input variable isn't very realistic, we can still learn a lot about regression basics from this model, which can be used in the further sprints concerning regression analysis.

2 Some Useful Resources

- An introduction to the topic:
  https://online.stat.psu.edu/stat462/node/79/
- A good book, but it covers more than we need in this course:
  Rawlings, Pantula, Dichey: Applied Regression Analysis - A Research Tool

3 R

3.1 Packages

—

3.2 Some Useful Commands

lm	**plot**	**abline**	**summary**	**residuals**	**fitted**
influence.measures	**which**	**apply**	**read.csv**	**qqnorm**	**qqline**

4 Exercises

1. The data set in the table below shows the distance in kilometers traveled by 10 orienteering competitors, along with the elapsed time in hours. For example, the first competitor traveled 10 kilometers in 2 hours.

Time (in hours)	Distance (in km)
2	10
2	11
3	12
4	13
4	14
5	15
6	20
7	18
8	22
9	25

(a) Create a scatter plot for *Distance* and *Time*. Does the relationship seem to be linear?

(b) Use *Distance* as dependent variable and *Time* as independent (explanatory) variable and perform a simple linear regression anaysis. What is the estimated regression line? Produce a scatter plot including the regression line.

Hint: Use the functions *myvalues=lm(...), plot(...)* and *abline(...)*.

(c) Is the independent (explanatory) variable significant in the regression? How high is the coefficient of determination?

Hint: Use the function *summary(lm(...)*.

(d) The estimated prediction line can be used to make predictions about the distance traveled for a given number of hours. If a competitor travels 3 hours, what is then the predicted distance? If a competitor travels 30 hours, what is then the predicted distance? Comment on the usefulness of these predictions.

(e) Perform a residual analysis. Interpret the results.

Hint: Use the functions *residuals(lm(...))* and *fitted(lm(...))* to extract the residuals and the fitted values from the model. Construct appropriate residual plots using these values.

(f) We're now going to discuss outliers and influential observations. Suppose that there was a new observation, a real hard-core orienteering competitor, who hiked for 16 h and traveled 39 km. Update your original data with this observation and produce a scatter plot again.

(g) Estimate the regression line obtained with this hard-core competitor included.

(h) Is this hard-core competitor an influential observation? Use Cook's distance to answer this question.

Hint: Use the function *influence.measures(...)* and store the result in *myinfluence*. Then use *which(apply(myinfluence$is.inf,1,any))* and *summary(myinfluence)*. What are these functions doing? Interpret the results!

(i) Return to the original data set with the the first ten observations. Add an eleventh observation with time 5 hours and distance traveled 20 kilometers. Is this an influential observation?
(Also produce a scatter plot, regression line ...)

(j) Return to the original data set with the the first ten observations. Add an eleventh observation with time 10 hours and distance traveled 23 kilometers. Is this an influential observation?
(Also produce a scatter plot, regression line ...)

(k) Compare and interpret the results in h), i) and j)! Which observations are influential according to Cook's distance?

2. We are now going to investigate the use of transformations in regression analysis. The data set *baseball.txt* contains a collection of batting statistics for 331 baseball players in the American League. We're going to investigate whether there is a relationship between batting average and the number of home runs that a player hits.

   (a) First read the data set *baseball.txt* into R.
   Hint: You can use for example *read.csv("yourpath/baseball.txt", sep="")*

   (b) Baseball batting averages tend to be highly variable for low numbers of at bats, we restrict our data set in the following tasks to those players who had at least 100 at bats for this season.
   Hint: To extract only those observations with $x \geq 5$ from a data frame called z consisting of the variables x and y, you can use $z[z\$x >= 5,]$.

   (c) How many observations are there now in this restricted data set (with only those players with at least 100 at bats)?

   (d) Produce a scatter plot of *home runs* versus *batting average* (=bat_ave).

   (e) Based on the scatter plot, is a transformation to linearity called for? Why or why not?

   (f) First, try to perform a regression analysis without any transformation. Use batting average as explanatory variable and home runs as response variable. What is the estimated regression line? Significance?

   (g) Perform a residual analysis (also including a normal probability plot) for your regression model. Comments?

   (h) Now, use the natural logarithm to transform the variable home run. Produce a scatter plot and perform the regression analysis again with this transformed variable as response variable. What is the estimated regression line? Significance?

   (i) Perform a residual analysis for this new regression model. Comments?

   (j) Compare these two regression models. Comments? Which one would you choose? Why?

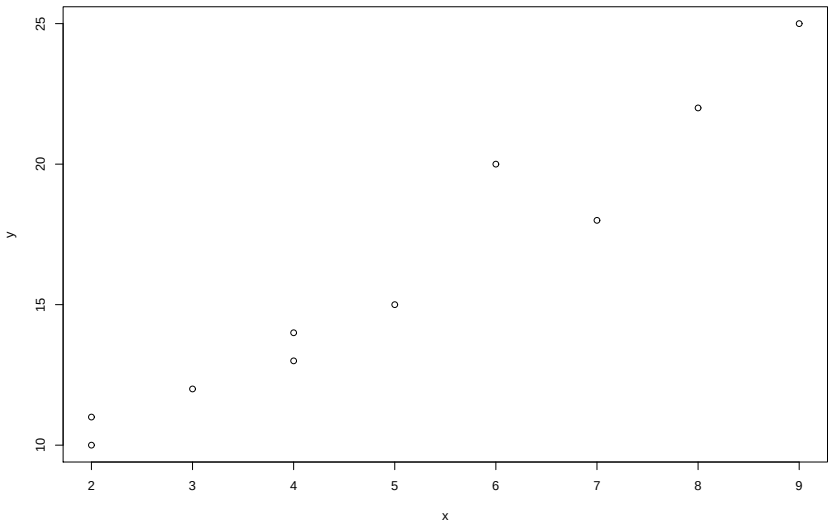
5 Solutions to Exercises

1. The data set in the table below shows the distance in kilometers traveled by 10 orienteering competitors, along with the elapsed time in hours. For example, the first competitor traveled 10 kilometers in 2 hours.

Time (in hours)	Distance (in km)
2	10
2	11
3	12
4	13
4	14
5	15
6	20
7	18
8	22
9	25

(a) Create a scatter plot for *Distance* and *Time*. Does the relationship seem to be linear?

```
> y=c(10,11,12,13,14,15,20,18,22,25)
> x=c(2,2,3,4,4,5,6,7,8,9)
> plot(x,y)
```

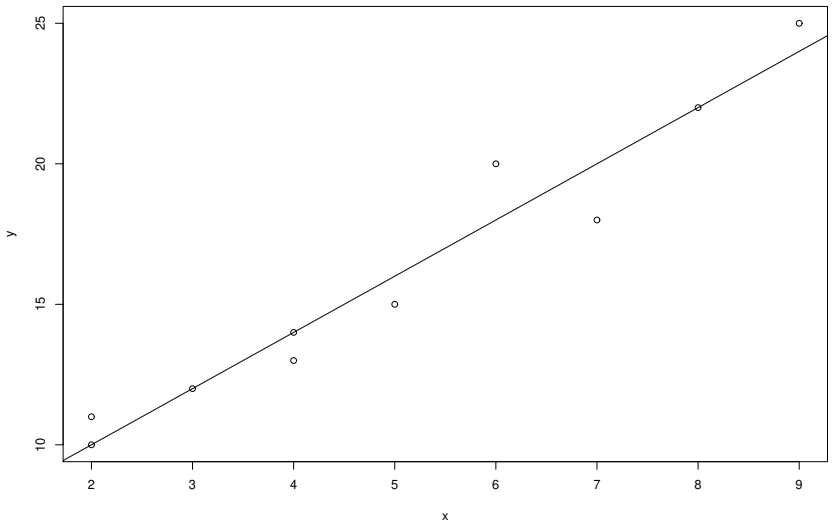


(b) Use *Distance* as dependent variable and *Time* as independent (explanatory) variable and perform a simple linear regression anaysis. What is the estimated regression line? Produce a scatter plot including the regression line.
Hint: Use the functions *myvalues=lm(...), plot(...)* and *abline(...)*.

Solution:

```
> myvalues=lm(y~x)
> myvalues

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
          6            2

> plot(y~x)
> abline(myvalues)
```

y=6+2*x



(c) Is the independent (explanatory) variable significant in the regression? How high is the coefficient of determination?
Hint: Use the function *summary(lm(...)*.

Solution:

```
> summary(myvalues)

Call:
lm(formula = y ~ x)

Residuals:
   Min     1Q Median     3Q    Max
```

```
 -2.00  -0.75   0.00   0.75   2.00

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.0000     0.9189   6.529 0.000182 ***
x             2.0000     0.1667  12.000 2.14e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.225 on 8 degrees of freedom
Multiple R-squared: 0.9474,Adjusted R-squared: 0.9408
F-statistic:   144 on 1 and 8 DF,  p-value: 2.144e-06
```

(d) The estimated prediction line can be used to make predictions about the distance traveled for a given number of hours. If a competitor travels 3 hours, what is then the predicted distance? If a competitor travels 30 hours, what is then the predicted distance? Comment on the usefulness of these predictions.

Solution:

```
y=6+2*3=12
y=6+2*30=66
```

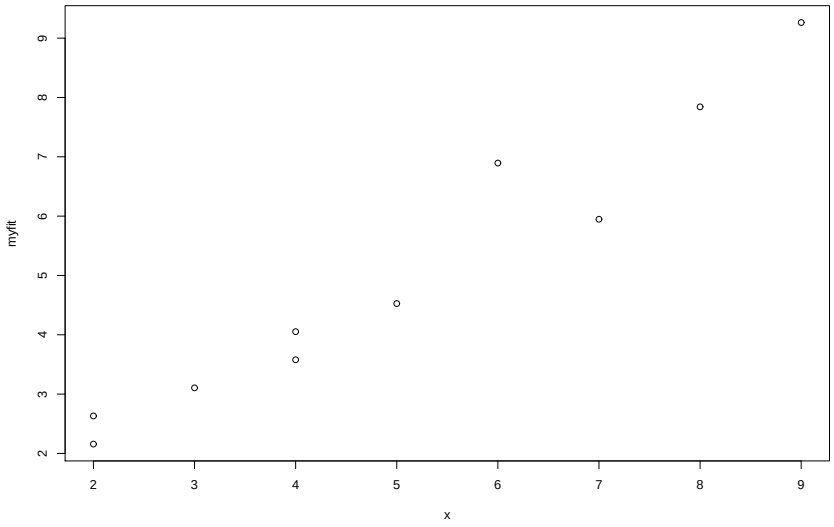
(e) Perform a residual analysis. Interpret the results.
Hint: Use the functions *residuals(lm(...))* and *fitted(lm(...))* to extract the residuals and the fitted values from the model. Construct appropriate residual plots using these values.

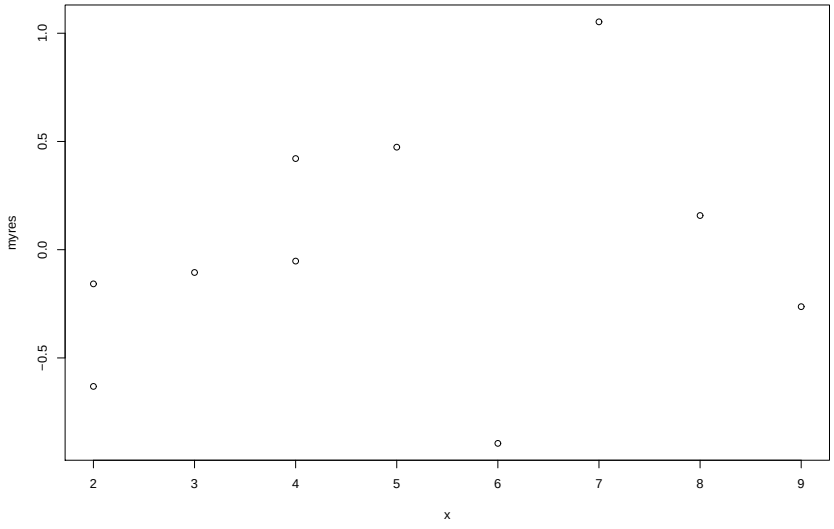
Solution:

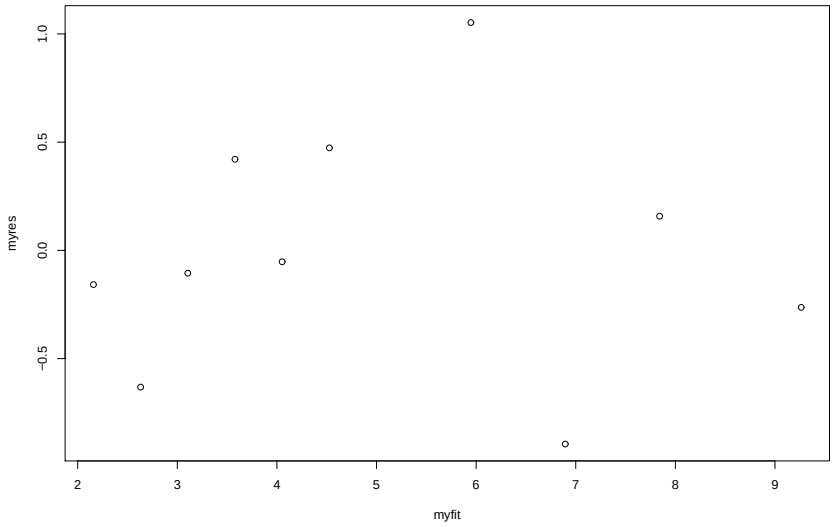
```
> myres=residuals(myvalues)
> myfit=fitted(myvalues)
> plot(x,myres)
> plot(x,myfit)
```

or

```
> plot(myfit,myres)
```


myfit
x
2
3
4
5
6
7
8
9

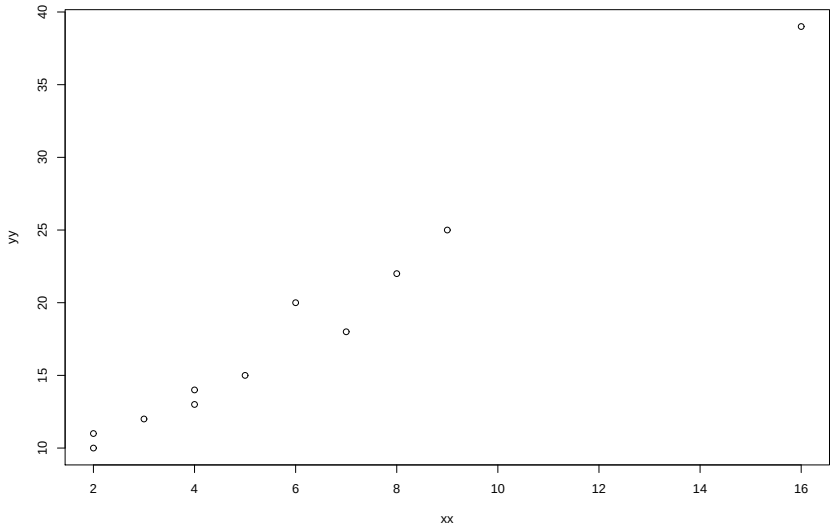

myres
x
−0.5
0.0
0.5
1.0
2
3
4
5
6
7
8
9



(f) We're now going to discuss outliers and influential observations. Suppose that there was a new observation, a real hard-core orienteering competitor, who hiked for 16 h and traveled 39 km. Update your original data with this observation and produce a scatter plot again.

Solution:

```
> xx =c(2 , 2,  3,  4,  4,  5,  6,  7,  8,  9, 16)
> yy =c(10, 11, 12, 13, 14, 15, 20, 18, 22, 25, 39)
> plot(xx,yy)
```



(g) Estimate the regression line obtained with this hard-core competitor included.

Solution:

```
> myvalues=lm(yy~xx)
> summary(myvalues)

Call:
lm(formula = yy ~ xx)

Residuals:
     Min       1Q   Median       3Q      Max
-2.15188 -0.59091  0.09202  0.51275  1.90909

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.72506    0.65132    8.79 1.04e-05 ***
xx           2.06098    0.09128   22.58 3.11e-09 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.169 on 9 degrees of freedom
Multiple R-squared: 0.9827,Adjusted R-squared: 0.9807
F-statistic: 509.7 on 1 and 9 DF,  p-value: 3.11e-09
```

(h) Is this hard-core competitor an influential observation? Use Cook's

distance to answer this question.

Hint: Use the function *influence.measures(...)* and store the result in *myinfluence*. Then use *which(apply(myinfluence$is.inf,1,any))* and *summary(myinfluence)*. What are these functions doing? Interpret the results!

Solution:

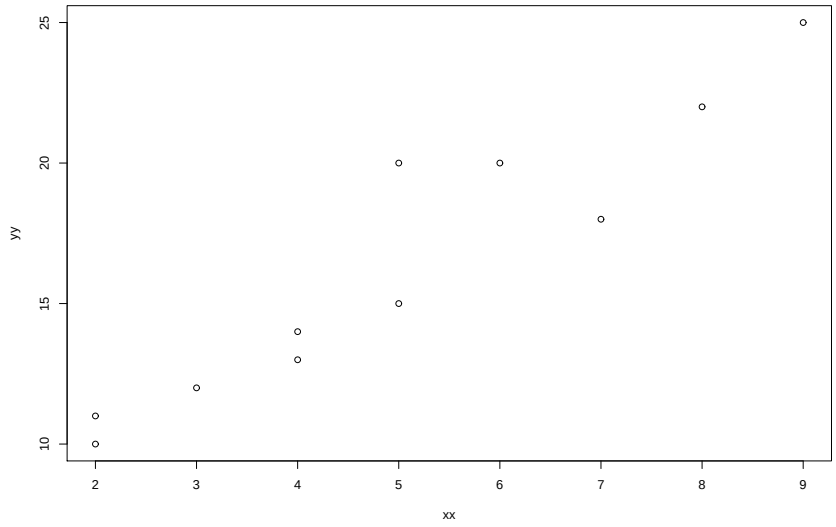
```
> influence.measures(myvalues)
Influence measures of
 lm(formula = yy ~ xx) :

     dfb.1_    dfb.xx    dffit cov.r   cook.d    hat inf
1   0.06482 -4.75e-02  0.06609 1.552 2.45e-03 0.1885
2   0.52408 -3.84e-01  0.53430 1.172 1.39e-01 0.1885
3   0.03130 -2.04e-02  0.03319 1.479 6.19e-04 0.1458
4  -0.27214  1.44e-01 -0.31378 1.194 5.06e-02 0.1153
5   0.00833 -4.42e-03  0.00961 1.430 5.19e-05 0.1153
6  -0.22133  7.55e-02 -0.30125 1.147 4.62e-02 0.0970
7   0.33660 -2.30e-17  0.62199 0.633 1.47e-01 0.0909
8  -0.24541 -1.97e-01 -0.78394 0.476 2.02e-01 0.0970
9  -0.00619 -3.04e-02 -0.06603 1.419 2.44e-03 0.1153
10 -0.02380  1.65e-01  0.26861 1.337 3.85e-02 0.1458
11 -0.40292  6.38e-01  0.68345 4.025 2.56e-01 0.7007   *

> myinfluence=influence.measures(myvalues)
> which(apply(myinfluence$is.inf,1,any))
11
11

> summary(myinfluence)
Potentially influential observations of
 lm(formula = yy ~ xx) :

   dfb.1_ dfb.xx dffit cov.r   cook.d hat
11 -0.40   0.64   0.68  4.02_*  0.26   0.70_*
```

(i) Return to the original data set with the the first ten observations. Add an eleventh observation with time 5 hours and distance traveled 20 kilometers. Is this an influential observation?
(Also produce a scatter plot, regression line ...)

Solution:

```
> yy=c(10,11,12,13,14,15,20,18,22,25,20)
> xx=c(2,2,3,4,4,5,6,7,8,9,5)
> plot(xx,yy)
```

```
> myvalues=lm(yy~xx)
> summary(myvalues)

Call:
lm(formula = yy ~ xx)

Residuals:
    Min      1Q  Median      3Q     Max
-2.3636 -0.8636 -0.3636  0.6364  3.6364

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.3636     1.2781   4.979 0.000761 ***
xx            2.0000     0.2337   8.558 1.29e-05 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.717 on 9 degrees of freedom
Multiple R-squared: 0.8906,Adjusted R-squared: 0.8784
F-statistic: 73.23 on 1 and 9 DF,  p-value: 1.287e-05

> myinfluence=influence.measures(myvalues)
> summary(myinfluence)
Potentially influential observations of
 lm(formula = yy ~ xx) :

   dfb.1_ dfb.xx dffit cov.r   cook.d hat
1  -0.13   0.11  -0.14  1.68_*  0.01   0.26
9   0.07  -0.11  -0.14  1.68_*  0.01   0.26
10 -0.22   0.31   0.36  1.96_*  0.07   0.39
11  0.40   0.00   0.98  0.28_*  0.25   0.09
```

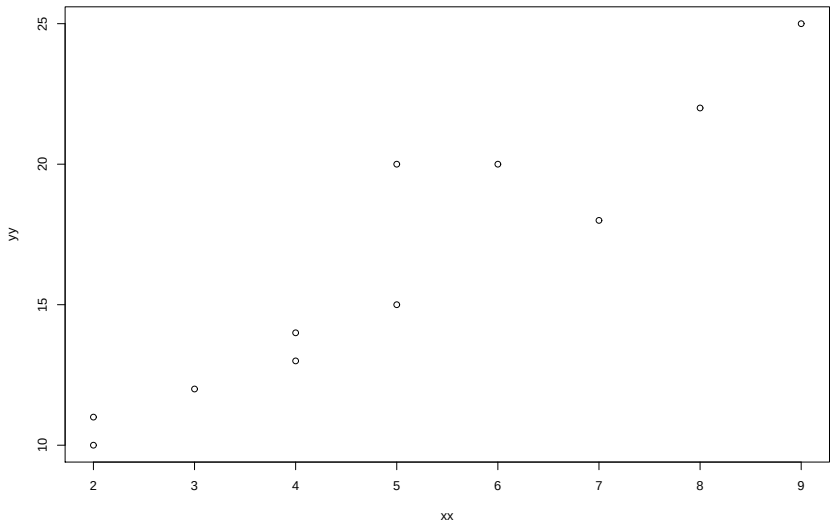


(j) Return to the original data set with the the first ten observations. Add an eleventh observation with time 10 hours and distance traveled 23 kilometers. Is this an influential observation?
(Also produce a scatter plot, regression line ...)

Solution:

```
> yy=c(10,11,12,13,14,15,20,18,22,25,23)
> xx=c(2,2,3,4,4,5,6,7,8,9,10)
> plot(xx,yy)
> myvalues=lm(yy~xx)
> summary(myvalues)

Call:
lm(formula = yy ~ xx)

Residuals:
    Min      1Q  Median      3Q     Max
-1.9194 -0.8969 -0.1635  0.6919  2.3697

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.6967     0.9718   6.891 7.14e-05 ***
xx            1.8223     0.1604  11.363 1.22e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

```
Residual standard error: 1.405 on 9 degrees of freedom
Multiple R-squared: 0.9348,Adjusted R-squared: 0.9276
F-statistic: 129.1 on 1 and 9 DF,  p-value: 1.223e-06

> myinfluence=influence.measures(myvalues)
> summary(myinfluence)
Potentially influential observations of
 lm(formula = yy ~ xx) :

   dfb.1_ dfb.xx  dffit   cov.r cook.d  hat
11  0.82  -1.27_* -1.47_*  0.90  0.82_*  0.36
```



(k) Compare and interpret the results in h), i) and j)! Which observations are influential according to Cook's distance?

Solution:
Only the last one influential according to Cook's distance.

2. We are now going to investigate the use of transformations in regression analysis. The data set *baseball.txt* contains a collection of batting statistics for 331 baseball players in the American League. We're going to investigate whether there is a relationship between batting average and the number of home runs that a player hits.

(a) First read the data set *baseball.txt* into R.
Hint: You can use for example

```
> baseball= read.csv("file://yourpath/baseball.txt", sep="",)
```

Solution:

```
> baseball= read.csv("file://yourpath/baseball.txt", sep="",)
```

(b) Baseball batting averages tend to be highly variable for low numbers of at bats, we restrict our data set in the following tasks to those players who had at least 100 at bats for this season.

Hint: To extract only those observations with $x \geq 5$ from a data frame called z consisting of the variables x and y, you can use $z[z\$x >= 5,]$.

Solution:

```
> baseball_new=baseball[baseball$at_bats>=100,]
> baseball_new$at_bats
  [1] 696 662 647 644 638 637 635 634 625 624 623 623 615 612 608 608 608 607
 [19] 590 587 585 582 577 577 573 570 569 569 566 564 561 560 560 557 556 554
 [37] 553 549 546 545 532 523 518 511 505 497 492 492 491 490 490 488 485 483
 [55] 482 482 480 475 475 474 474 471 467 467 466 466 465 463 461 458 455 454
 [73] 451 450 444 444 440 436 429 428 426 422 419 416 412 410 408 403 398 397
 [91] 397 383 378 374 374 367 366 359 358 357 353 352 351 347 345 344 342 337
[109] 330 329 328 326 325 321 320 316 314 312 304 302 300 289 288 284 282 281
[127] 277 277 275 273 265 265 264 263 263 260 259 258 252 251 245 241 240 237
[145] 229 228 222 222 219 212 204 202 201 200 199 197 196 194 193 192 182 181
[163] 179 179 171 168 167 163 161 159 158 157 157 156 155 153 151 150 149 148
[181] 147 144 137 135 134 133 132 132 132 130 130 128 127 127 125 117 113 112
[199] 112 112 112 112 111 109 108 107 105 104 103
```

(c) How many observations are there now in this restricted data set (with only those players with at least 100 at bats)?

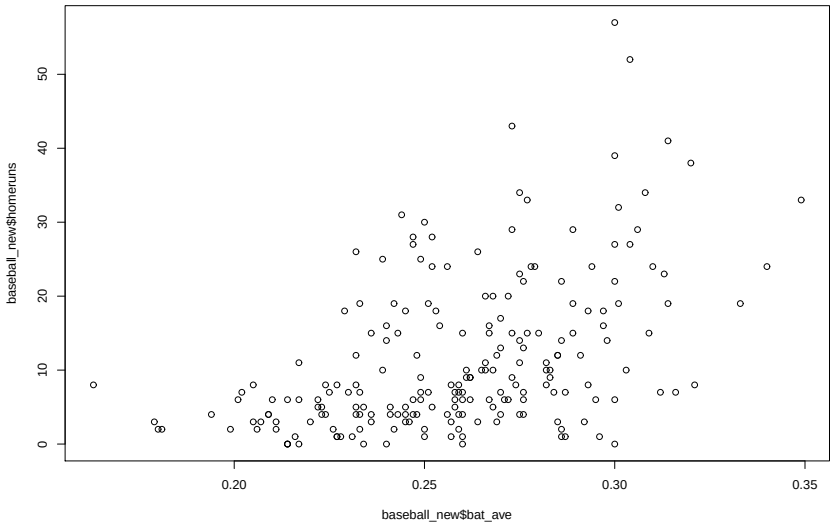
Solution:

```
> dim(baseball_new)
[1] 209  19
```

(d) Produce a scatter plot of *home runs* versus *batting average* (=bat_ave).

Solution:

```
> plot(baseball_new$bat_ave,baseball_new$homeruns)
```



(e) Based on the scatter plot, is a transformation to linearity called for? Why or why not?

Solution:
Not linear. Transformation required.

(f) First, try to perform a regression analysis without any transformation. Use batting average as explanatory variable and home runs as response variable. What is the estimated regression line? Significance?

Solution:

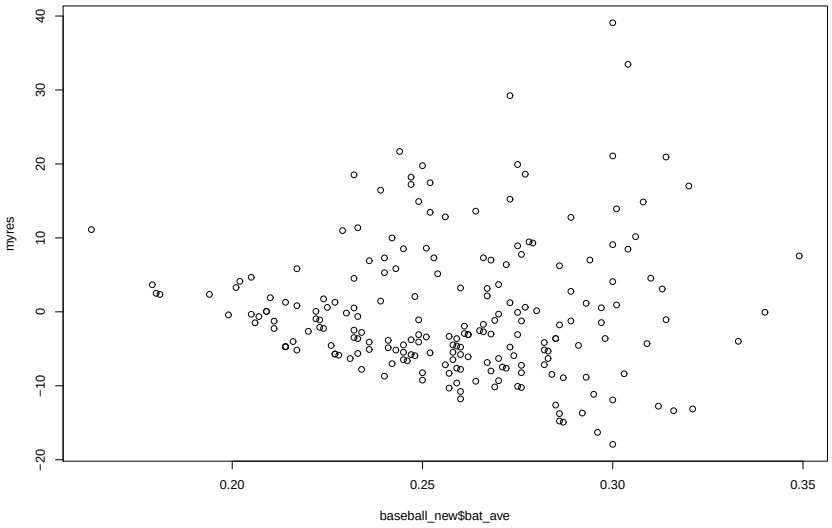
```
> myreg=lm(baseball_new$homeruns~baseball_new$bat_ave)
> summary(myreg)

Call:
lm(formula = baseball_new$homeruns ~ baseball_new$bat_ave)

Residuals:
    Min      1Q  Median      3Q     Max
-17.916  -5.774  -2.093   4.132  39.084

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          -28.149      5.083  -5.538 9.22e-08 ***
baseball_new$bat_ave  153.551     19.503   7.873 1.91e-13 ***
```

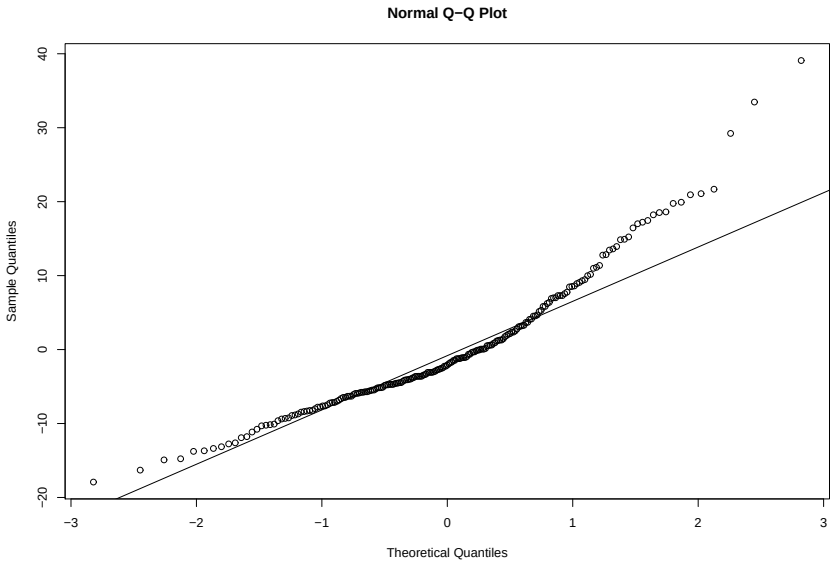
```
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 9.14 on 207 degrees of freedom
Multiple R-squared: 0.2305,Adjusted R-squared: 0.2267
F-statistic: 61.99 on 1 and 207 DF,  p-value: 1.909e-13
```

(g) Perform a residual analysis (also including a normal probability plot) for your regression model. Comments?

Solution:

```
> myres=residuals(myreg)
> plot(baseball_new$bat_ave,myres)

> qqnorm(myres)
> qqline(myres)
```



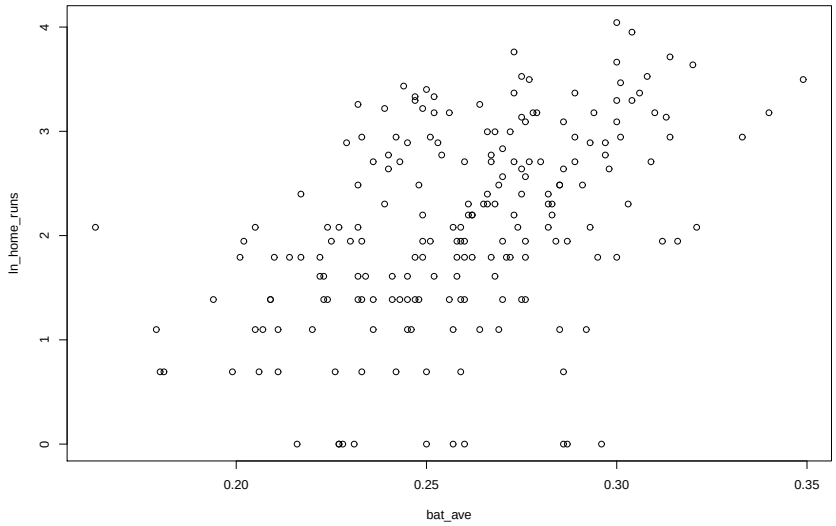


(h) Now, use the natural logarithm to transform the variable home run. Produce a scatter plot and perform the regression analysis again with this transformed variable as response variable. What is the estimated regression line? Significance?

Solution:

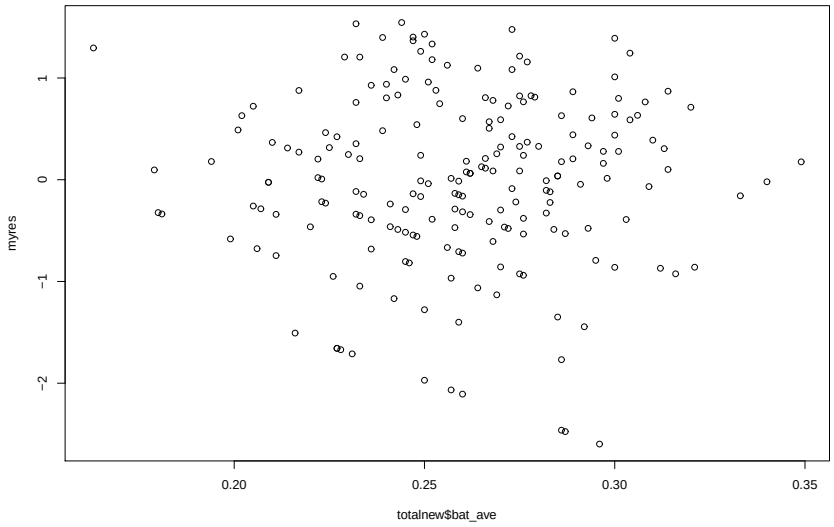
```
> ln_home_runs=log(baseball_new$homerun)
> bat_ave=baseball_new$bat_ave
> plot(bat_ave,ln_home_runs)

> total=data.frame(ln_home_runs,bat_ave)
> totalnew=total[total$ln_home_runs>=-10,]
> myr=lm(totalnew$ln_home_runs~totalnew$bat_ave)
> summary(myr)

Call:
lm(formula = totalnew$ln_home_runs ~ totalnew$bat_ave)

Residuals:
     Min       1Q   Median       3Q      Max
-2.59815 -0.46545  0.03677  0.60086  1.54499

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      -1.4386     0.4775  -3.013  0.00293 **
```

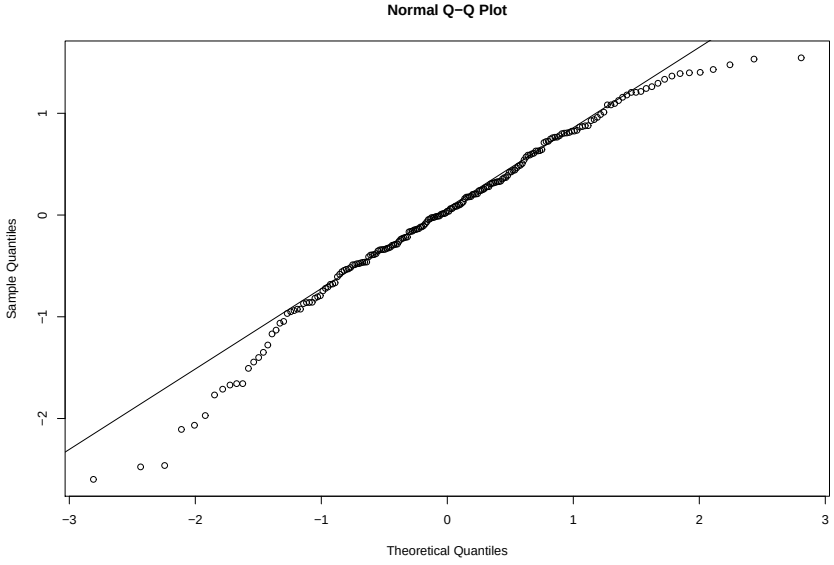
```
totalnew$bat_ave  13.6375     1.8263   7.467 2.51e-12 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.8352 on 199 degrees of freedom
Multiple R-squared: 0.2189,Adjusted R-squared: 0.215
F-statistic: 55.76 on 1 and 199 DF,  p-value: 2.511e-12
```



(i) Perform a residual analysis for this new regression model. Comments?

Solution:





(j) Compare these two regression models. Comments? Which one would you choose? Why?

Solution:

`Yes, which one would you choose?`