MSG500 Final 2016-08-25

Examiner: Rebecka Jörnsten, 0760-491949

Remember: To pass this course you also have to hand in a final project to the examiner.
Open book, open notes but no calculators or computers allowed

Make sure to give detailed and specific answers. Avoid yes/no answers. You should also provide a motivation. Good Luck!

Question 1(35=5+5+5+5+5+5+5)

The NHANES data set is a very large public health data set. Here you will look at a small subset of subjects and variables recorded. Specifically, you will look at the cholesterol level of 1000 subjects. Other variables recorded are; gender, age, HDL (the ”good” cholesterol), DBP, SBP (dpb is a more reliable marker for cardiovascular risk than systolic bp), weight (wt), height (ht) and body mass index (bmi).
a) Below I summarize the fit of a regression model including all variables. Comment on the appropriateness of the fit, the meaning of the model, if you think variables have to be transformed etc. Make sure you state how you come to your finding. Discuss other information you might want to have access to to say more about the fit and model.

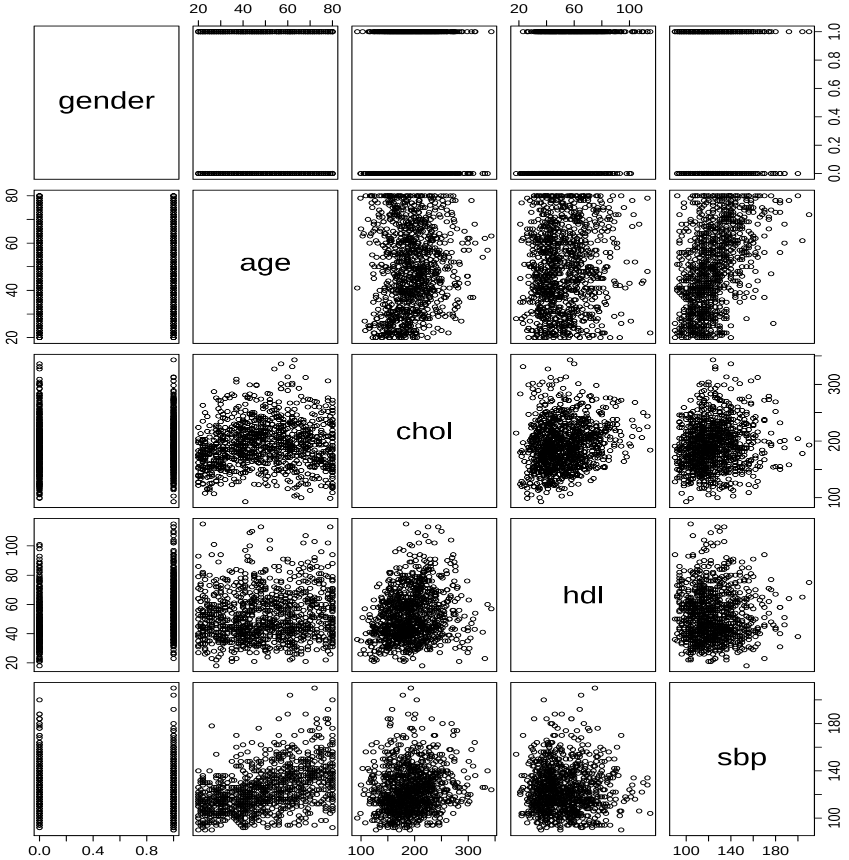

Figure 1: Scatter plots

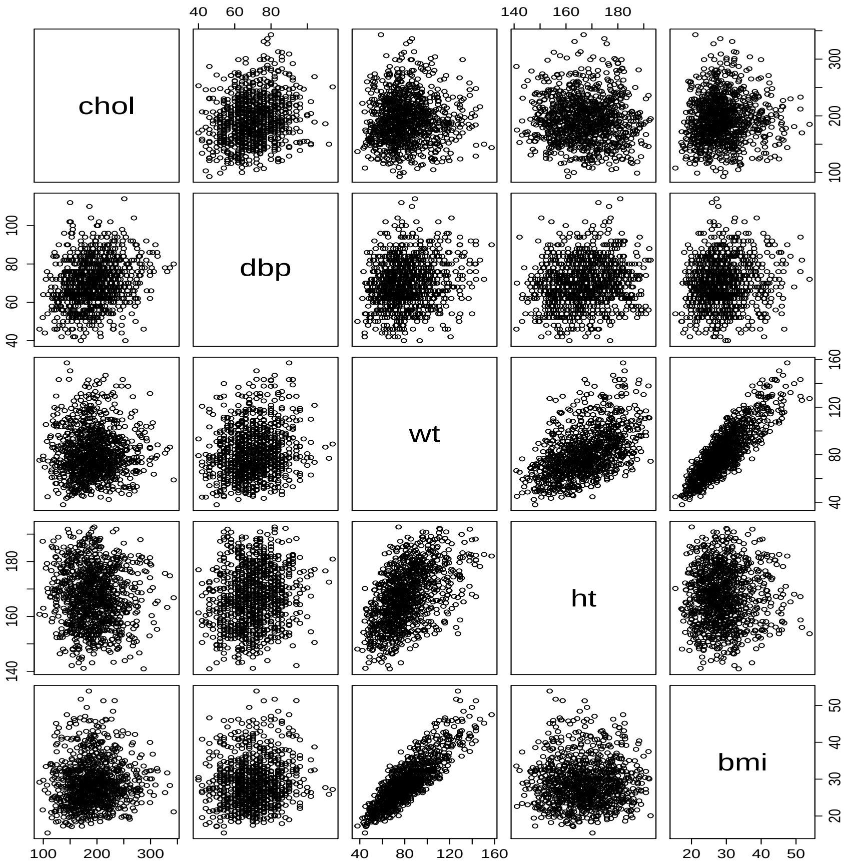

Figure 2: Scatter plots

```
Residuals:
    Min      1Q  Median      3Q     Max
-91.503 -25.689  -2.733  21.014 145.386

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 197.01226   98.64927   1.997   0.0461 *
gender        0.02981    3.61927   0.008   0.9934
age           0.13571    0.08057   1.684   0.0924 .
hdl           0.54652    0.08364   6.534 1.02e-10 ***
sbp           0.02790    0.08128   0.343   0.7315
dbp           0.72565    0.11594   6.259 5.77e-10 ***
wt            0.61421    0.57596   1.066   0.2865
ht           -0.67870    0.58532  -1.160   0.2465
bmi          -1.01345    1.61637  -0.627   0.5308
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 38.4 on 991 degrees of freedom
Multiple R-squared:  0.1069,Adjusted R-squared:  0.09964
F-statistic: 14.82 on 8 and 991 DF,  p-value: < 2.2e-16
```

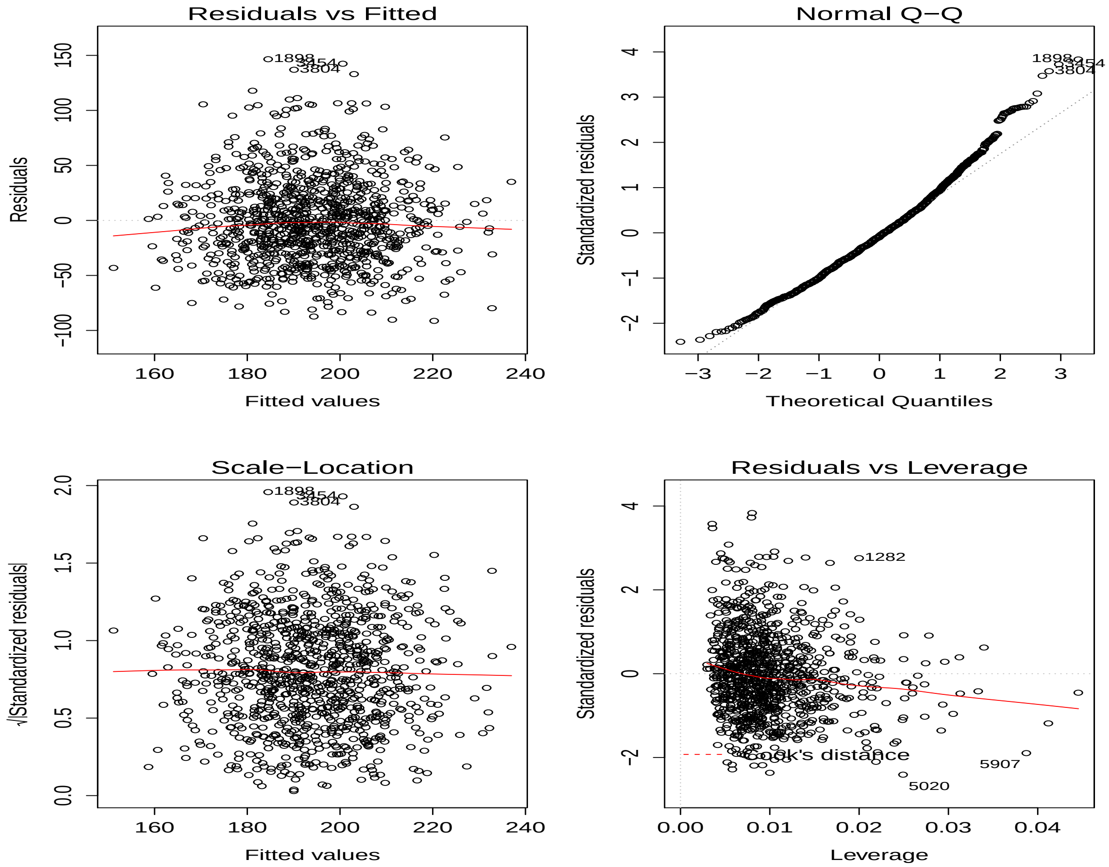

Figure 3: Residual diagnostics

b) Backward model-selection via AIC results in `age + hdl + dbp + wt + ht` as the final model. Comment and discuss.

c) We randomly split the data in 50% for training and 50% for testing. We repeat this 100 times, each time recording the model size, the prediction error and the selected model. Below are the results. Comment and discuss.
Average PE Cp: 1505.298, PE AIC: 1505.298, PE BIC: 1510.798
Average Modsize Cp: 4.92, Modsize AIC: 4.92, Modsize BIC: 3.72

```
              modselcp modselaic modselbic
[1,] "gender" "8"      "8"       "0"
[2,] "age"    "58"     "58"      "9"
[3,] "hdl"    "100"    "100"     "100"
[4,] "sbp"    "14"     "14"      "2"
[5,] "dbp"    "100"    "100"     "99"
[6,] "wt"     "21"     "21"      "5"
[7,] "ht"     "13"     "13"      "1"
[8,] "bmi"    "78"     "78"      "56"
```

d) I repeat the above using only 25% for training and get the following selection results. Comment and discuss.
Average Modsize Cp: 4.44, Modsize AIC: 4.46, Modsize BIC: 3.08
Average PE Cp: 1525.443, PE AIC: 1525.449, PE BIC: 1536.828

```
                modselcp modselaic modselbic
[1,] "gender" "11"     "11"      "3"
[2,] "age"    "32"     "33"      "5"
[3,] "hdl"    "96"     "96"      "82"
[4,] "sbp"    "12"     "12"      "2"
[5,] "dbp"    "99"     "99"      "93"
[6,] "wt"     "31"     "32"      "5"
[7,] "ht"     "16"     "17"      "4"
[8,] "bmi"    "47"     "46"      "14"
```

e) What would an interaction between gender and age mean? (Come up with a concrete example what such a model might look like and what the coefficients would tell you about subjects in the NHANES data).

f) A variable ”marital status” is also available for analysis. It has 5 levels (single, married, widowed, divorced, separated). Discuss how you would include such a variable in your model. Comment on what kind of results such a model might produce.

g) Are there any interactions that you would be particularly interesting in estimating? (including with marital status). Discuss and motivate.

Question 2(15=5+5+5)

a) Below you see 4 different nonlinear model descriptions. In each case, identify which parameters are linear and nonlinear. Also identify if there is a data transform that can bring this model into a linear form.
i) $y = a + bx + cx^2 + dx^3$ (parameters a,b,c,d)
ii) $y = a + c * exp(-exp(-b(x-d)))$ parameters(a,b,c,d)
iii) $y = a + b * exp(-cx) + dx$ parameters(a,b,c,d)

b) A nonlinear model $y = a + \frac{c}{1+exp(-b(x-d))}$ is fit to a data set of sample size 44 using nonlinear least squares. The regression mean SS is 15.06 and the MSE is 0.13. The F-statistics is thus 114.23. Is this significant? (Say how you conclude this). What does this mean?

c) The parameter estimates are as follows (as outputted from R):

```
parameter   estimate    s.e.
b           0.472       0.116
d           27.415      0.551
c           2.617       0.242
d           8.839       0.175
```

Explain how you would go about setting up CI and assessing significance. What are the assumption you make in order to set up such CI (or compute these p-values)? How would you go about checking those assumptions?

Question 3(20=5+5+5+5)

Data are collected on average race times for 35 different races. Information about the races include the total distance and the overall climb (increase in elevation) during the race.

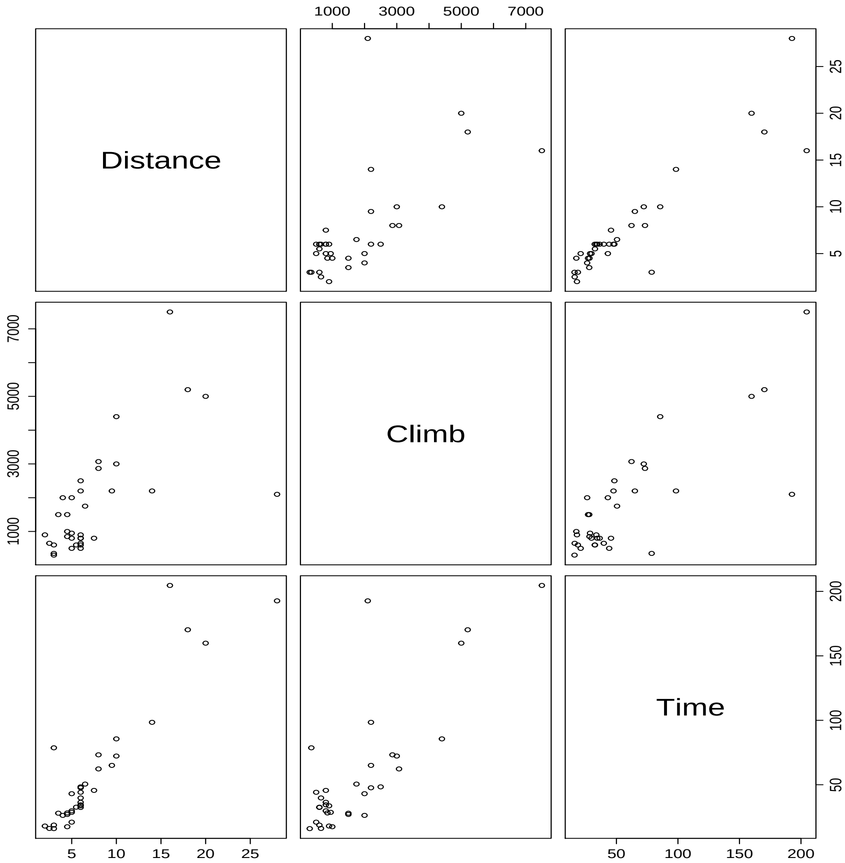

Figure 4: Scatter plots

a) A linear model is fit to the data to predict race times from distance and climb. The results are provided below. Comment on the fit and interpret the model. Any need for transformations? Any outliers? What's the impact on the fit of these/this if any outliers are present? Explain.

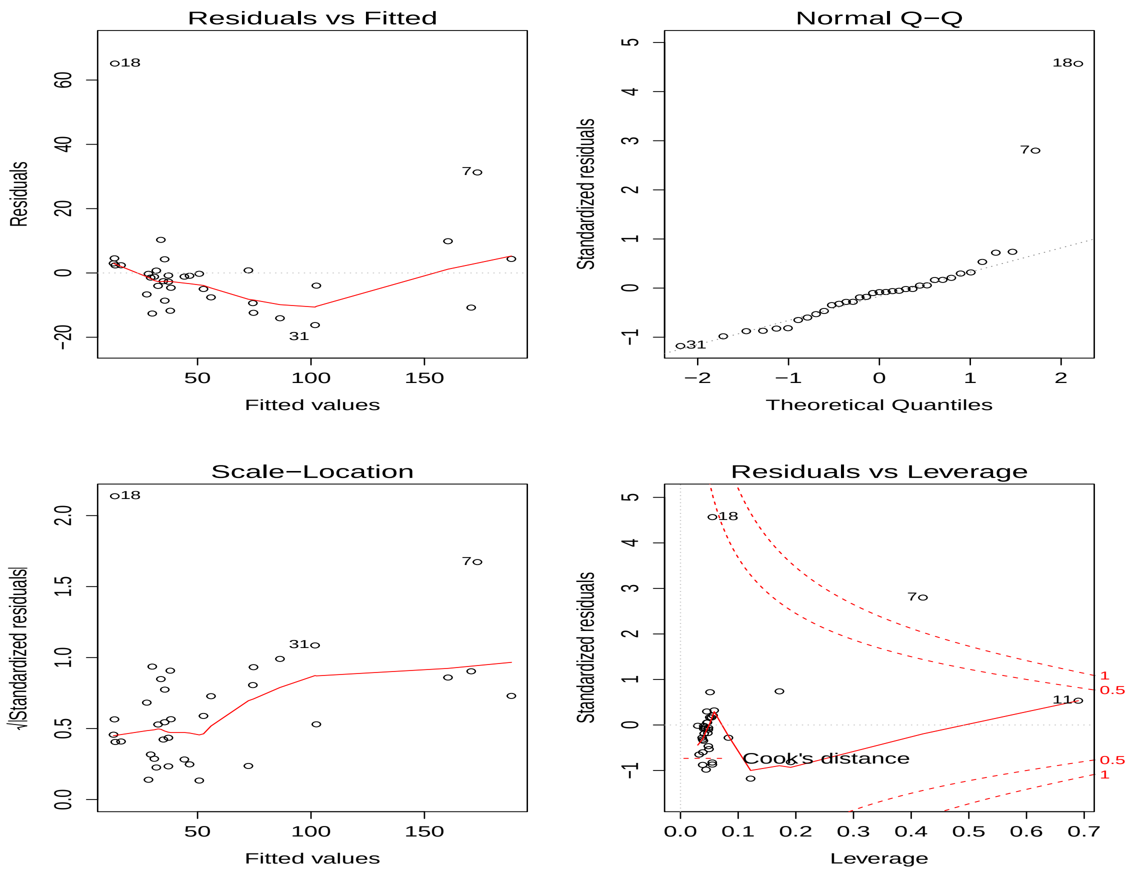

Figure 5: Residual diagnostics

```
Residuals:
    Min      1Q  Median      3Q     Max 
-16.215  -7.129  -1.186   2.371  65.121 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -8.992039   4.302734  -2.090   0.0447 *  
Distance     6.217956   0.601148  10.343 9.86e-12 ***
Climb        0.011048   0.002051   5.387 6.45e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 14.68 on 32 degrees of freedom
Multiple R-squared:  0.9191,Adjusted R-squared:  0.914 
F-statistic: 181.7 on 2 and 32 DF,  p-value: < 2.2e-16
```

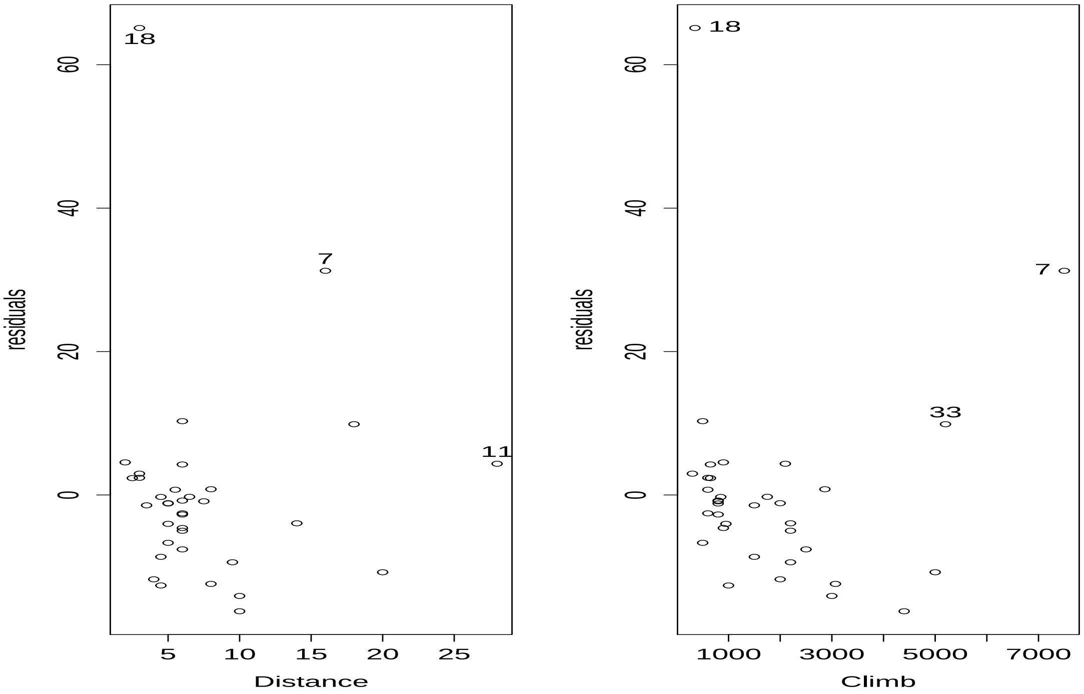

Figure 6: Residuals vs covariates

b) What is the meaning of an interaction term between Climb and Distance? Explain in concrete terms (i.e. not just the statistical definition but what it would mean in this data setting)

c) Below is the result when an interaction term is included. Comment on the fit. Climb is no longer significant. Comment, interpret and explain.

```
Residuals:
    Min      1Q  Median      3Q     Max
-25.994  -4.968  -2.220   2.381  56.115

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     9.3954374  6.8790233   1.366  0.18183
Distance        4.1489201  0.8352489   4.967 2.36e-05 ***
Climb          -0.0009710  0.0041648  -0.233  0.81718
Distance:Climb  0.0009831  0.0003070   3.203  0.00314 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 12.92 on 31 degrees of freedom
Multiple R-squared:  0.9392,Adjusted R-squared:  0.9333
F-statistic: 159.6 on 3 and 31 DF,  p-value: < 2.2e-16
```

d) Let's say more data are collected for races at high elevation locations. These appear as a group in the data with longer race times overall and especially for races with larger Climbs. How would you go about modeling this kind of data? (Hint: Draw such a data set in the scatter plots and go from there.)

Question 4(15=5+5+5)

Consider a simple regression model, y on x.
a) When would you consider transforming y? Explain what graphical and numerical tools you would use to determine the need for transformation. Same regarding x.
b) One of the additional assumptions we make in modeling is that of normality. We can use the QQ plot to check this. QQ plots of which of i-iv are relevant for modeling assessment here?
i) QQ of y
ii) QQ of x
iii) QQ of residuals
iv) QQ of standardized residuals
Say which you would use and why.
c) State an example where a variable transformation can help against collinearity.

Question 5(15=5+5+5)

Below is a scatter plot for a data set.

a) Identify outliers in the data set. Say which measure (leverage, Cook's distance,... that could be used to identify each of the outliers).

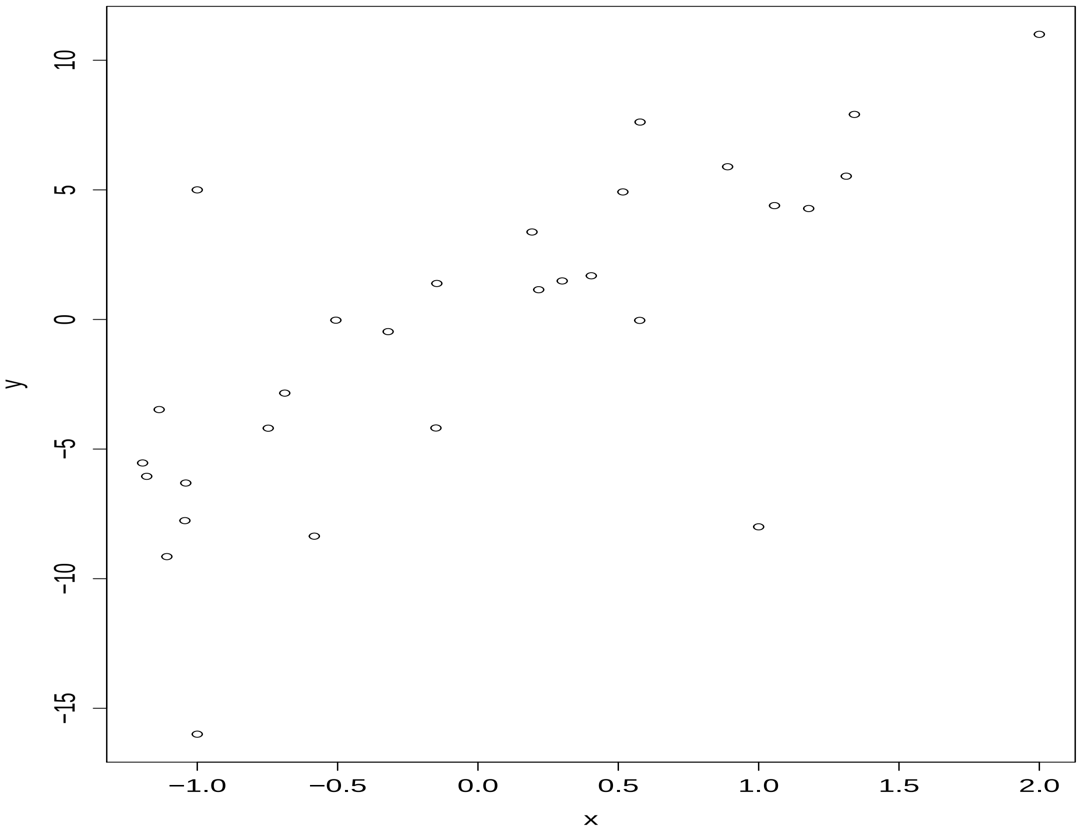

Figure 7: Scatter plot

b) Which of R^2, $\hat{\sigma}$, $se(\hat{\beta}_1)$ and the t-value for testing $\beta_1 = 0$ would change if you dropped these outliers (all of them)?

c) Same as b) but you drop only one of the outliers (do this for each of the outliers you identified).