TF: Alexis

# Homework 3 Problem Set

## *Statistics E-109*

## *Due March 30, 2022 at 5:50 pm EST*

*Homework policies. Please provide concise, clear answers for each question. Note that only writing the result of a calculation (e.g., "SD = 3.3") without explanation is not sufficient. For problems involving R, include the code in your solution, along with any plots.*

*Please submit your homework assignment via Canvas as a PDF file.*

*We encourage you to discuss problems with other students (and, of course, with the course head and the TFs), but you must write your final answer in your own words. Solutions prepared "in committee" are not acceptable. If you do collaborate with classmates on a problem, please list your collaborators on your solution.*

*Max points: 100*

### PART 1 (25 points)

Perform the following commands in R:

```
#R Code here
set.seed(1)
x1 <- runif(100)
x2 <- 0.5*x1+rnorm(100)/10
y <- 2+2*x1+0.3*x2+rnorm(100)
```

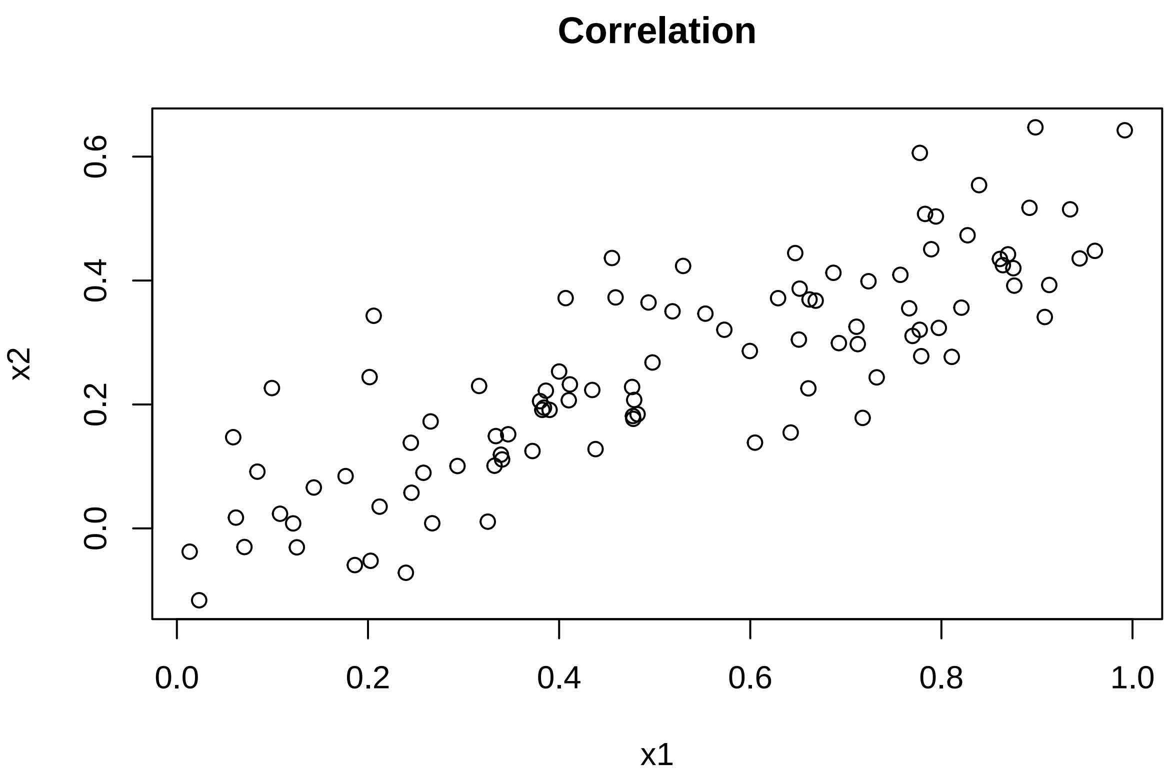
The last line corresponds to creating a linear model in which y is a function of x1 and x2.

a) What is the correlation between x1 and x2? Create a scatterplot displaying the relationship between the variables.

```
#R Code here
cor(x1, x2)
```

```
## [1] 0.8351212
```

```
plot(x1, x2, main = 'Correlation')
```




**Comment** The correlation between x1 and x2 is 0.8351212.

b) Using this data, fit a least squares regression to predict y using x1 and x2. Describe the results obtained. What are b0, b1, and b2? How do these relate to the true $\beta_0$, $\beta_1$, and $\beta_2$? Can you reject the null hypothesis $H0 : \beta_1 = 0$? How about the null hypothesis $H_0 : \beta_2 = 0$?

```
#R Code here

m <- lm(y ~ (x1 + x2))
summary(m)
```

```
##
## Call:
## lm(formula = y ~ (x1 + x2))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8311 -0.7273 -0.0537  0.6338  2.3359
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.1305     0.2319   9.188 7.61e-15 ***
## x1            1.4396     0.7212   1.996   0.0487 *
## x2            1.0097     1.1337   0.891   0.3754
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.056 on 97 degrees of freedom
## Multiple R-squared:  0.2088, Adjusted R-squared:  0.1925
## F-statistic:  12.8 on 2 and 97 DF,  p-value: 1.164e-05
```

```r
dif0 <- 2 - 2.1305
dif1 <- 2 - 1.4396
dif2 <- 0.3 - 1.0097
dif0; dif1; dif2
```

```
## [1] -0.1305
## [1] 0.5604
## [1] -0.7097
```

**Comment** b0 = 2.1305; b0 - ^b0 = -0.1305 b1 = 1.4396; b1 - ^b1 = 0.5604 b2 = 1.0097; b2 - ^b2 = -0.7097

The coefficient estimates for b0, b1, and b2 are 2.1305, 1.4396, and 1.0097 respectively. These values differ from the true coefficients by -0.1305, 0.5604, and -0.7097 respectively.

We can reject the null hypothesis at a 99% confidence value for b1, and we cannot reject the null hypothesis for b2.

c) Now fit a least squares regression to predict y using only x1. Comment on your results. Can you reject the null hypothesis $H_0 : \beta_1 = 0$?

```r
#R Code here
m2 <- lm(y ~ x1)
summary(m2)
```

```
##
## Call:
## lm(formula = y ~ x1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.89495 -0.66874 -0.07785  0.59221  2.45560
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.1124     0.2307   9.155 8.27e-15 ***
## x1            1.9759     0.3963   4.986 2.66e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.055 on 98 degrees of freedom
## Multiple R-squared:  0.2024, Adjusted R-squared:  0.1942
## F-statistic: 24.86 on 1 and 98 DF,  p-value: 2.661e-06
```

**Comment** The estimated coefficients for b0 and b1 are 2.1124 and 1.9759 respectively. The null hypothesis H0: b1 = 0 can be confidently rejected due to the very low p-value.

d) Now fit a least squares regression to predict y using only x2. Comment on your results. Can you reject the null hypothesis $H_0 : \beta_1 = 0$?

```r
#R Code here
m3 <- lm(y ~ x2)
summary(m3)
```

```
##
## Call:
## lm(formula = y ~ x2)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.62687 -0.75156 -0.03598  0.72383  2.44890
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.3899     0.1949   12.26  < 2e-16 ***
## x2            2.8996     0.6330    4.58 1.37e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.072 on 98 degrees of freedom
## Multiple R-squared:  0.1763, Adjusted R-squared:  0.1679
## F-statistic: 20.98 on 1 and 98 DF,  p-value: 1.366e-05
```

**Comment** The coefficient estimates are b0 = 2.3899 and 2.8996. The null hypothesis H0: b1 = 0 can be confidently rejected due to the very low p-value.

e) Do the results obtained in (b)-(d) contradict each other? Explain your answer.

**Comment** The results obtained using both x1 and x2 showed significance at a 99% confidence level for b1 and no significance for b2. The results obtained using either x1 or x2 were highly significant and confidently rejected the null hypothesis b1 = 0. These results do contradict; however, it can be explained by collinearity. In part A, the correlation between x1 and x2 was high at 0.835. Because of the similarity between x1 and x2, their combined estimates had additive effects in the prediction model, resulting in higher significance/confidence of the results.

f) Now suppose we obtain one additional observation, which was unfortunately mismeasured.

```
#R Code here
x1 <- c(x1 , 0.1)
x2 <- c(x2 , 0.8)
y <- c(y,6)
```

Re-fit the linear models from (b) to (d) using this new data. What effect does this new observation have on the each of the models? In each model, is this observation an outlier? A high-leverage point? Both? Explain your answers.

```
#R Code here
m <- lm(y ~ (x1 + x2))
m1 <- lm(y ~ x1)
m2 <- lm(y~ x2)

summary(m); summary(m1); summary(m2)
```

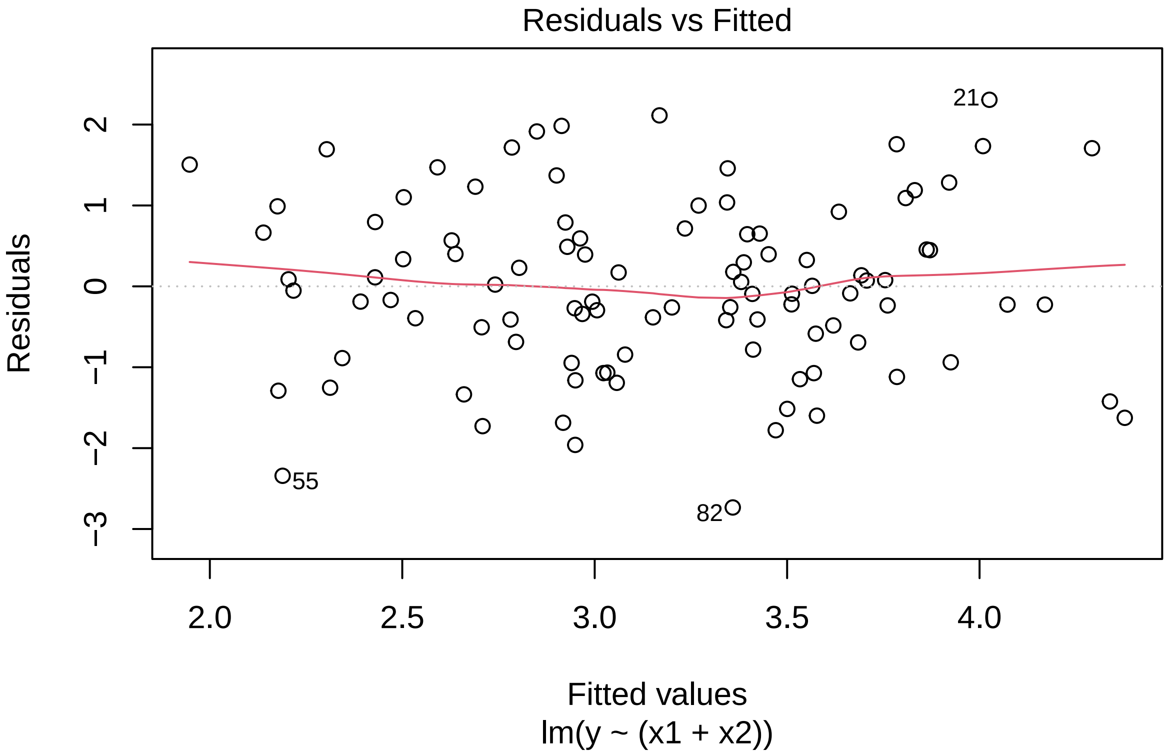
```
##
## Call:
## lm(formula = y ~ (x1 + x2))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.73348 -0.69318 -0.05263  0.66385  2.30619
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

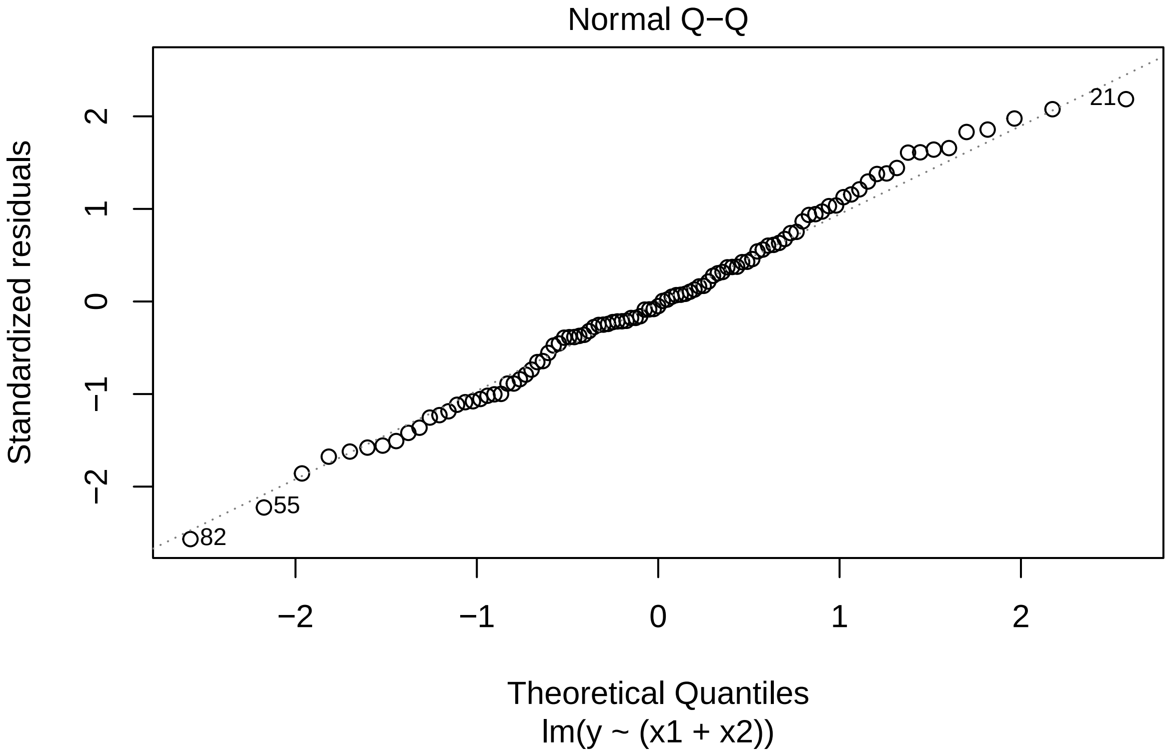
```
## (Intercept)   2.2267     0.2314   9.624 7.91e-16 ***
## x1            0.5394     0.5922   0.911  0.36458
## x2            2.5146     0.8977   2.801  0.00614 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.075 on 98 degrees of freedom
## Multiple R-squared:  0.2188, Adjusted R-squared:  0.2029
## F-statistic: 13.72 on 2 and 98 DF,  p-value: 5.564e-06
```

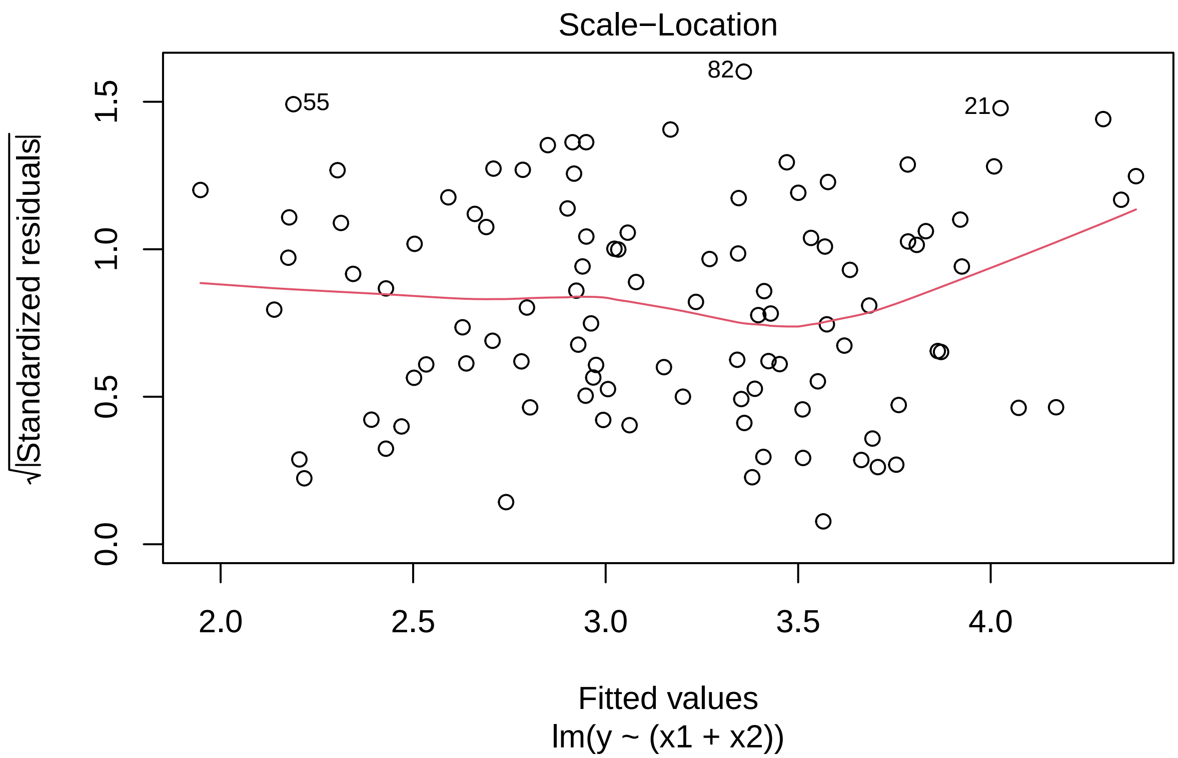
```
##
## Call:
## lm(formula = y ~ x1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8897 -0.6556 -0.0909  0.5682  3.5665
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.2569     0.2390   9.445 1.78e-15 ***
## x1            1.7657     0.4124   4.282 4.29e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.111 on 99 degrees of freedom
## Multiple R-squared:  0.1562, Adjusted R-squared:  0.1477
## F-statistic: 18.33 on 1 and 99 DF,  p-value: 4.295e-05
```

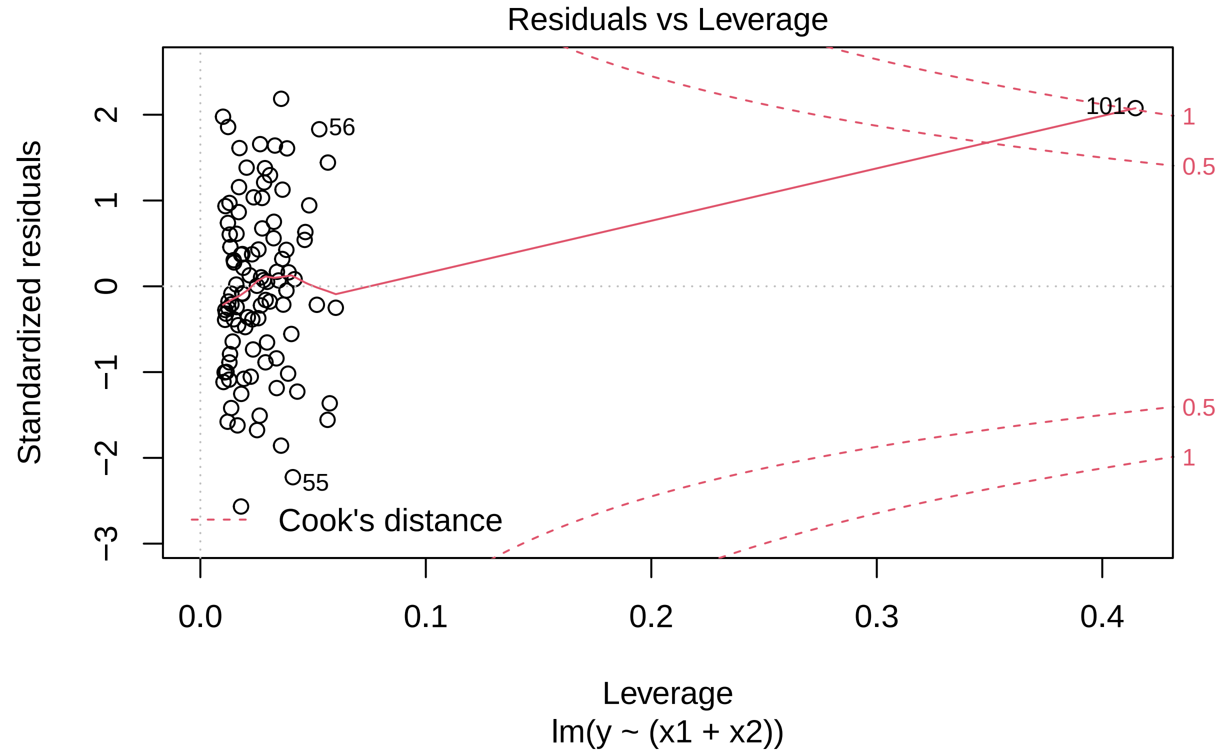
```
##
## Call:
## lm(formula = y ~ x2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.64729 -0.71021 -0.06899  0.72699  2.38074
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.3451     0.1912  12.264  < 2e-16 ***
## x2            3.1190     0.6040   5.164 1.25e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.074 on 99 degrees of freedom
## Multiple R-squared:  0.2122, Adjusted R-squared:  0.2042
## F-statistic: 26.66 on 1 and 99 DF,  p-value: 1.253e-06
```

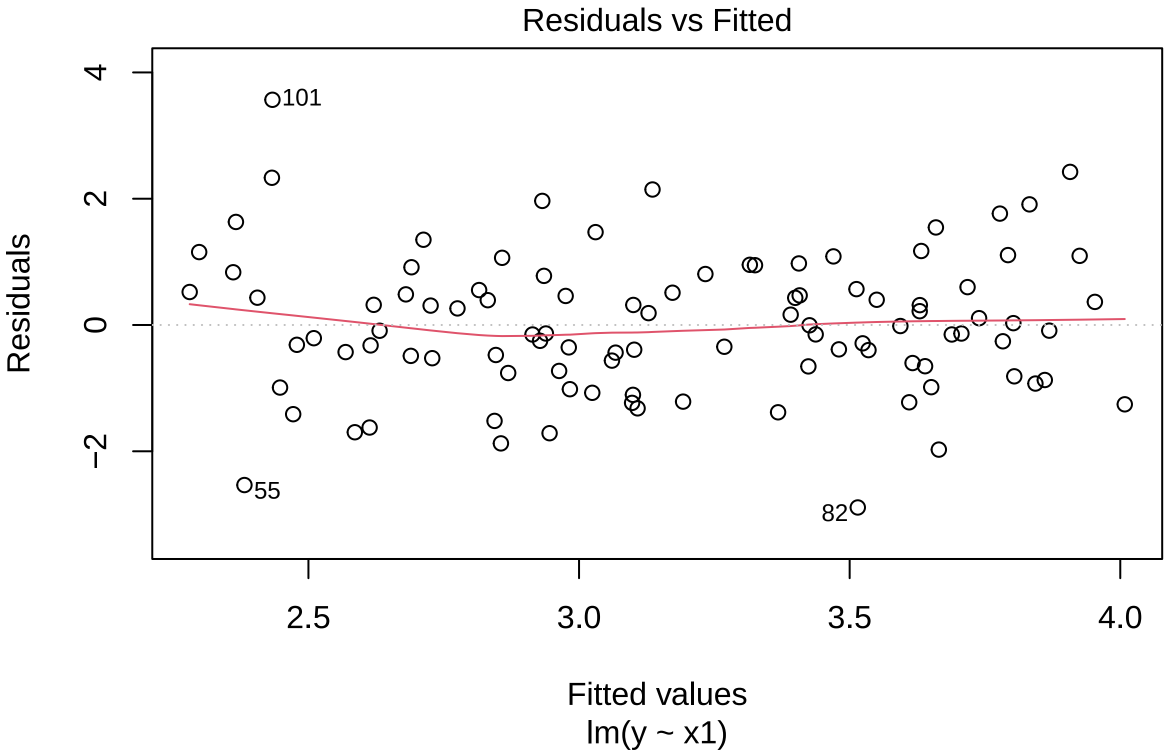
```
plot(m); plot(m1); plot(m2)
```


Residuals vs Fitted
Residuals
Fitted values
lm(y ~ (x1 + x2))
21
55
82
2.0
2.5
3.0
3.5
4.0
−3
−2
−1
0
1
2

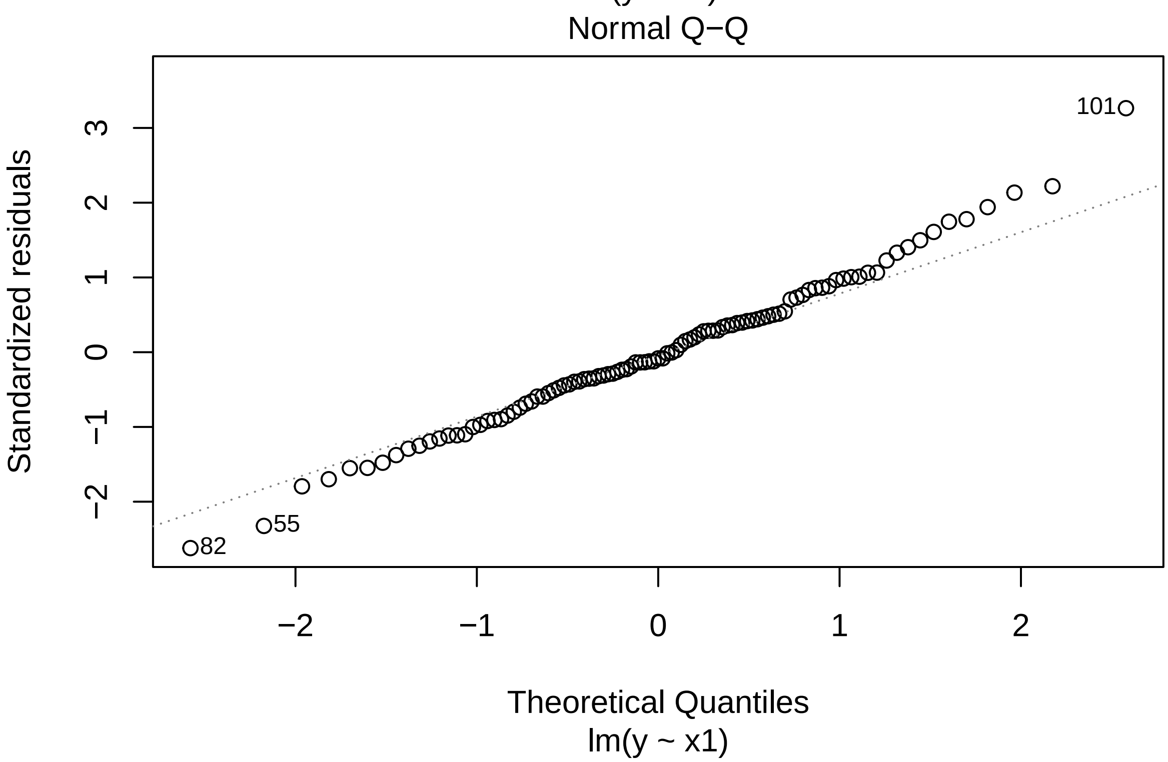

Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(y ~ (x1 + x2))
21
55
82
−2
−1
0
1
2


Scale−Location
√|Standardized residuals|
Fitted values
lm(y ~ (x1 + x2))
82
55
21
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0

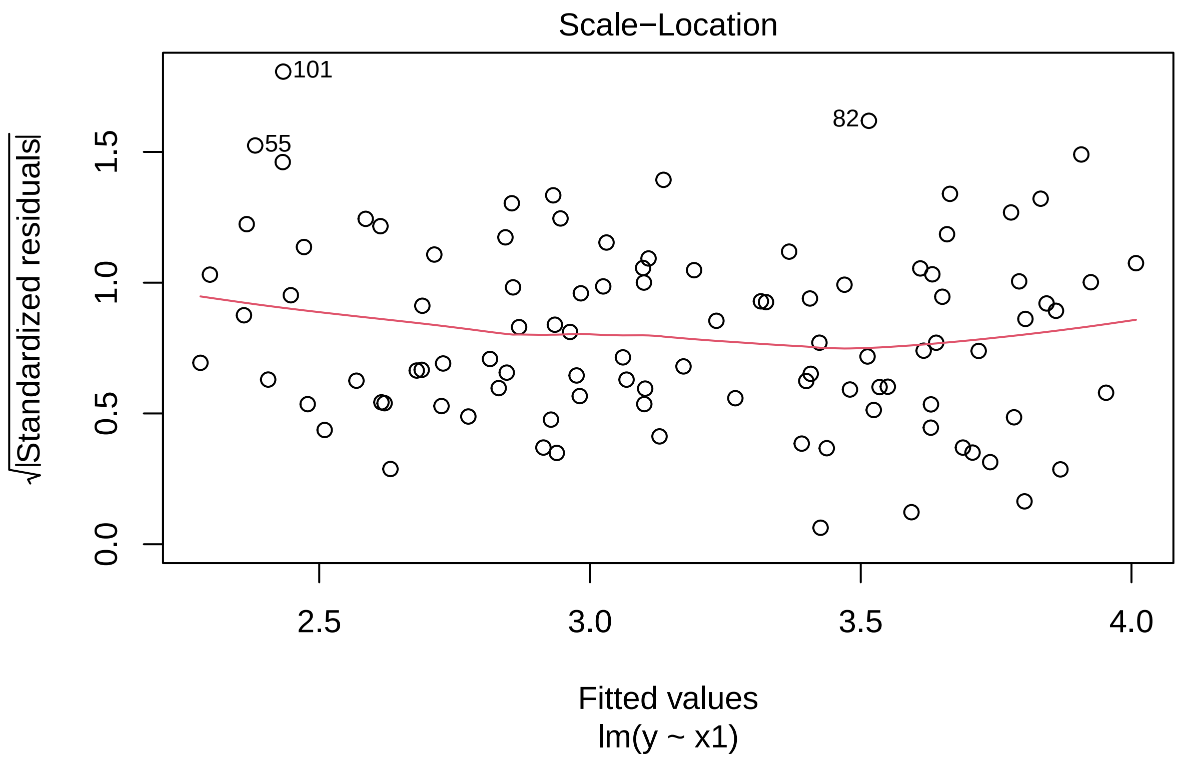

Residuals vs Leverage
Standardized residuals
Leverage
lm(y ~ (x1 + x2))
101
56
55
Cook's distance
1
0.5
0.5
1
−3
−2
−1
0
1
2
0.0
0.1
0.2
0.3
0.4

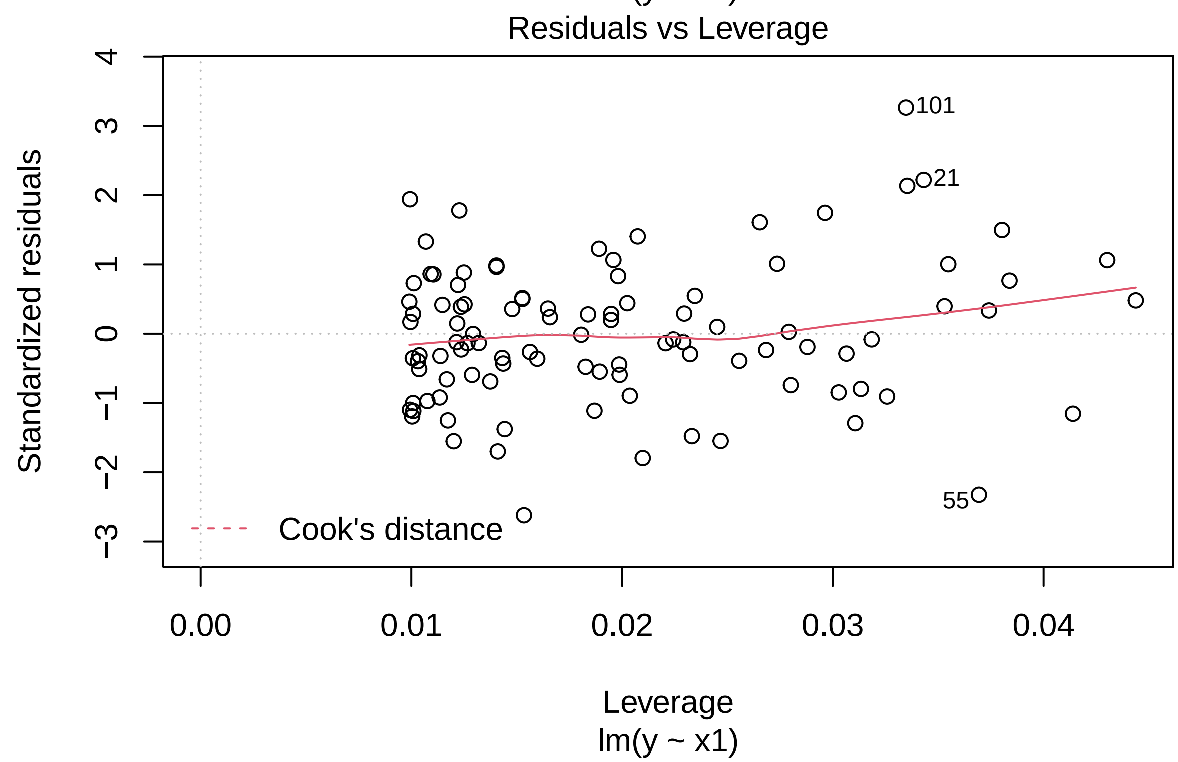


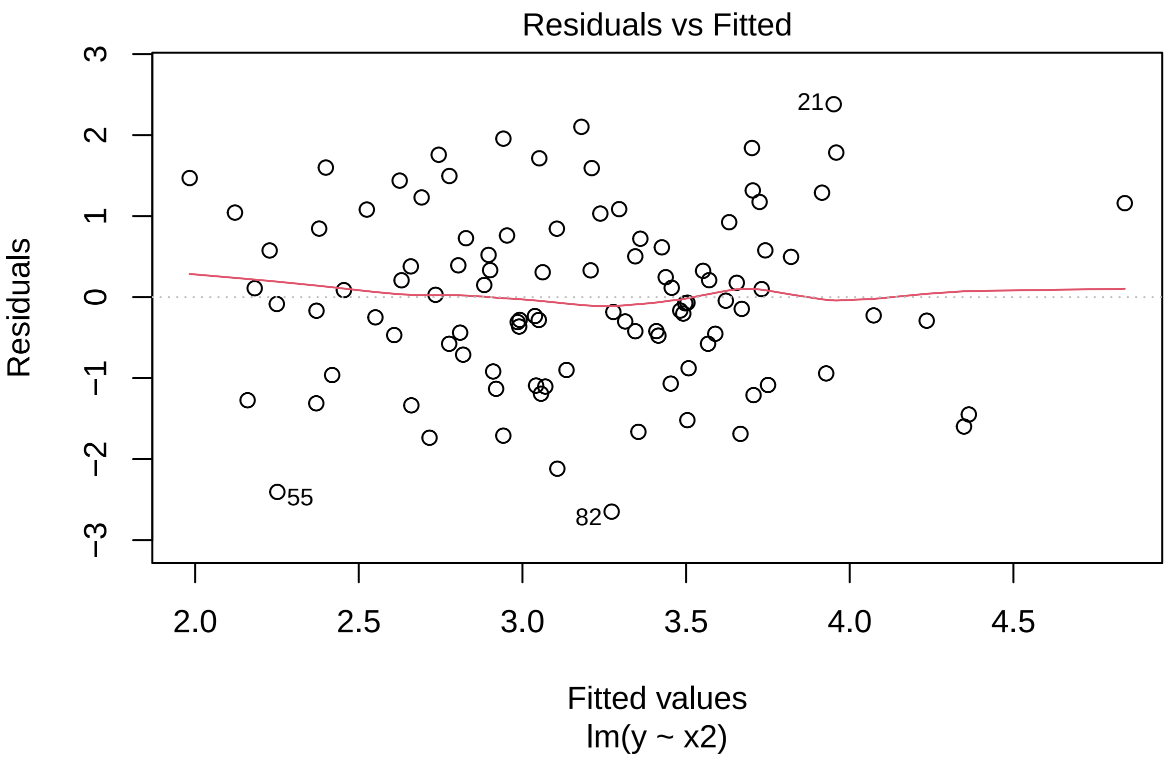
Residuals vs Fitted
Residuals
4
2
0
−2
101
55
82
2.5
3.0
3.5
4.0
Fitted values
lm(y ~ x1)

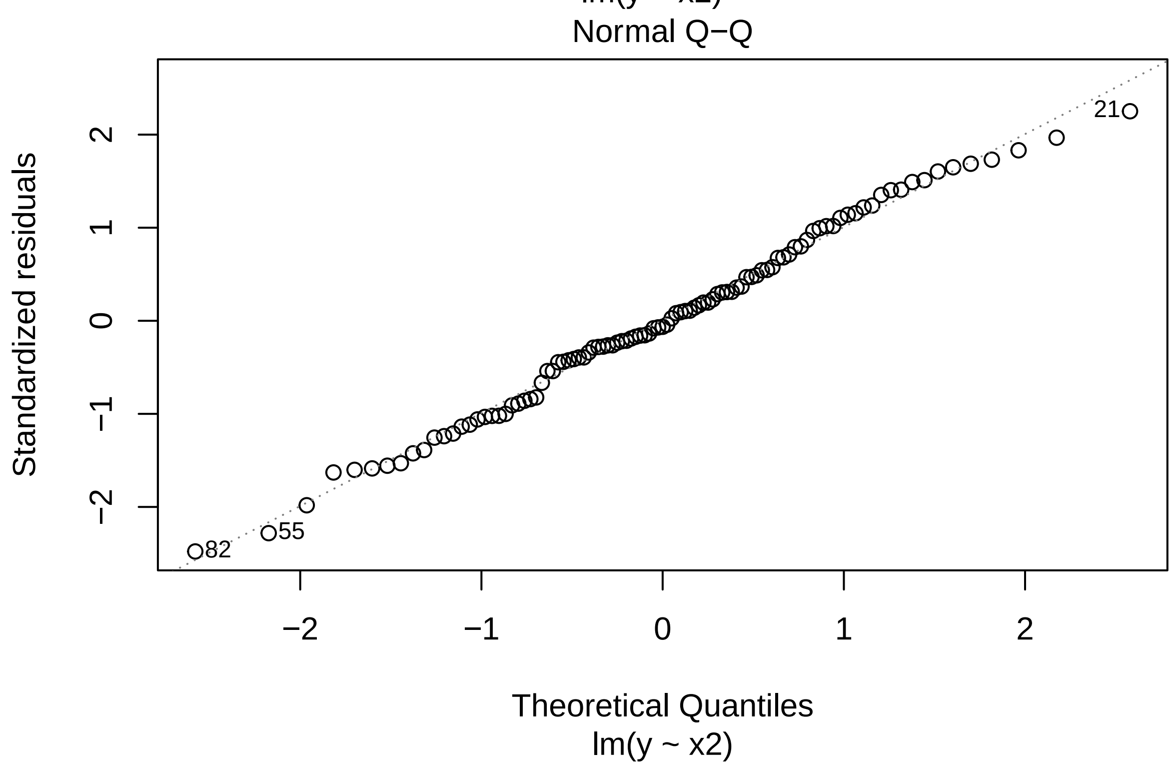



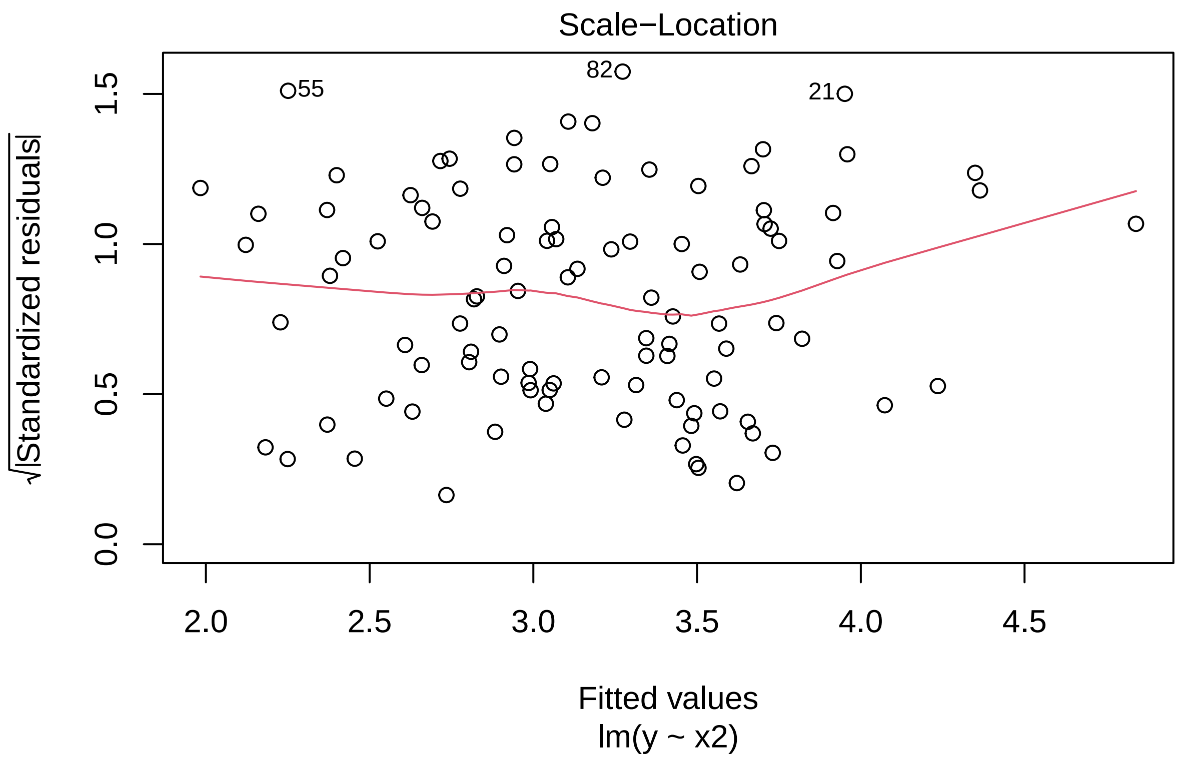
Normal Q−Q
Standardized residuals
3
2
1
0
−1
−2
101
55
82
−2
−1
0
1
2
Theoretical Quantiles
lm(y ~ x1)

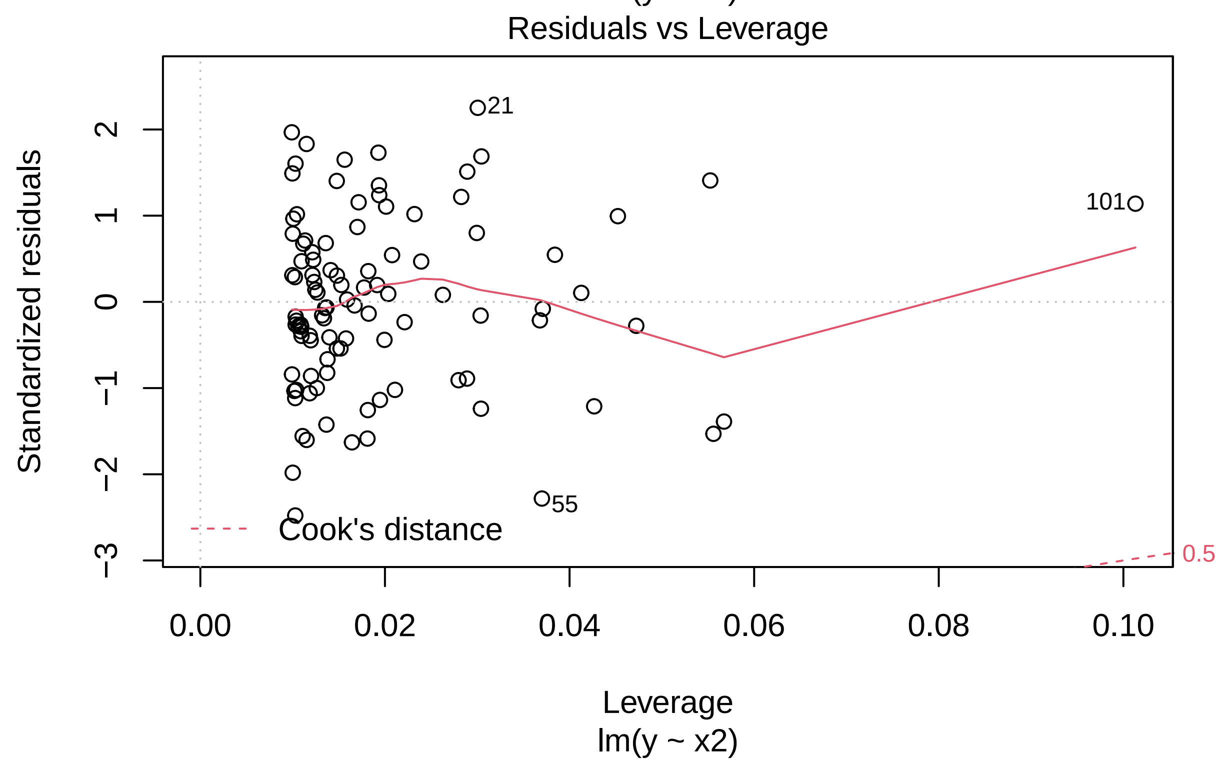

Scale−Location
√|Standardized residuals|
Fitted values
lm(y ~ x1)
101
55
82


Residuals vs Leverage
Standardized residuals
Leverage
lm(y ~ x1)
Cook's distance
101
21
55


Residuals vs Fitted
Residuals
3
2
1
0
−1
−2
−3
21
55
82
2.0
2.5
3.0
3.5
4.0
4.5
Fitted values
lm(y ~ x2)


Normal Q−Q
Standardized residuals
2
1
0
−1
−2
21
55
82
−2
−1
0
1
2
Theoretical Quantiles
lm(y ~ x2)




**Comment** Individually, using x1 or x2 as a predictor of y gives highly significant results and confidently rejects the null hypothesis. Using both x1 and x2, however, reduces the significant of both x1 and x2. For x1, there is now no statistical significance, and for x2, the null hypothesis can only be rejected at 99% confidence level, which is the opposite of the previous models. Based on cook's distance, the added point is a high leverage point, but not an outlier, for the x1 + x2 model. For the x1 model, the added point is an outlier, but not a high leverage point. For the x2 model, the added point is neither outlier nor high-levarage point based on cook's distance.

**PART 2 (25 points)**

(Use TermLife.csv data file) Term Life Insurance: Here we examine the 2004 Survey of Consumer Finances (SCF), a nationally representative sample that contains extensive information on assets, liabilities, income, and demographic characteristics of those sampled (potential U.S. customers). We study a random sample of 500 families with positive incomes. From the sample of 500, we initially consider a subsample of n = 275 families that purchased term life insurance. Note: For n = 275, we want you to subset the data so that you are only looking at rows where FACE > 0. Also, variable LNFACE = log of the face variable and LNINCOME = log of the income variable.

a) Fit a linear regression model of LNINCOME, EDUCATION, NUMHH, MARSTAT, AGE, and GENDER on LNFACE.

```r
#R Code here
TermLife <- read.csv('TermLife.csv')
subTerm <- subset(TermLife, subset = FACE > 0)
subTerm$LNFACE <- with(subTerm, log(FACE))
subTerm$LNINCOME <- with(subTerm, log(INCOME))

m <- lm(LNFACE ~ (LNINCOME + EDUCATION + NUMHH + MARSTAT + AGE + GENDER), data = subTerm)
summary(m)
```

```
##
## Call:
## lm(formula = LNFACE ~ (LNINCOME + EDUCATION + NUMHH + MARSTAT +
##     AGE + GENDER), data = subTerm)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.7719 -0.8990  0.0914  0.8407  4.6519
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.527493   0.907585   2.785  0.00574 **
## LNINCOME     0.450635   0.078786   5.720 2.84e-08 ***
## EDUCATION    0.218198   0.038729   5.634 4.45e-08 ***
## NUMHH        0.263542   0.072279   3.646  0.00032 ***
## MARSTAT     -0.163366   0.268683  -0.608  0.54369
## AGE         -0.003529   0.007928  -0.445  0.65656
## GENDER       0.917195   0.352626   2.601  0.00981 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.511 on 268 degrees of freedom
## Multiple R-squared:  0.3616, Adjusted R-squared:  0.3474
## F-statistic:  25.3 on 6 and 268 DF,  p-value: < 2.2e-16
```

b) Check if multicollinearity is present.

```r
#R Code here
library(car)
```

```
## Loading required package: carData
```

```r
vif(m)
```

```
##  LNINCOME EDUCATION     NUMHH   MARSTAT       AGE    GENDER
##  1.248831  1.169355  1.396531  1.863770  1.140301  1.743312
```

**Comment** Each variable shows a low VIF score, between 1.14 and 1.86, which suggests that multicollinearity is not present.

c) Briefly explain the idea of collinearity and a variance inflation factor. What constitutes a large variance inflation factor?

**Comment** Collinearity is when multiple variables in a multiple linear regression model are highly correlated to each other, making it difficult to independently predict the dependant variable and affecting the statistical significance measurement. Variance inflation factor measures the amount of multicollinearity in multiple regression variables by taking the ratio between the overall model variance and the variance of the single independent variable. A VIF score above 10 suggests very high correlation between the variables and multicolliniarity.

d) Supplement the variance inflation factor statistics with a table of correlations of explanatory variables. Given these statistics, is collinearity an issue with this fitted model? Why or why not?

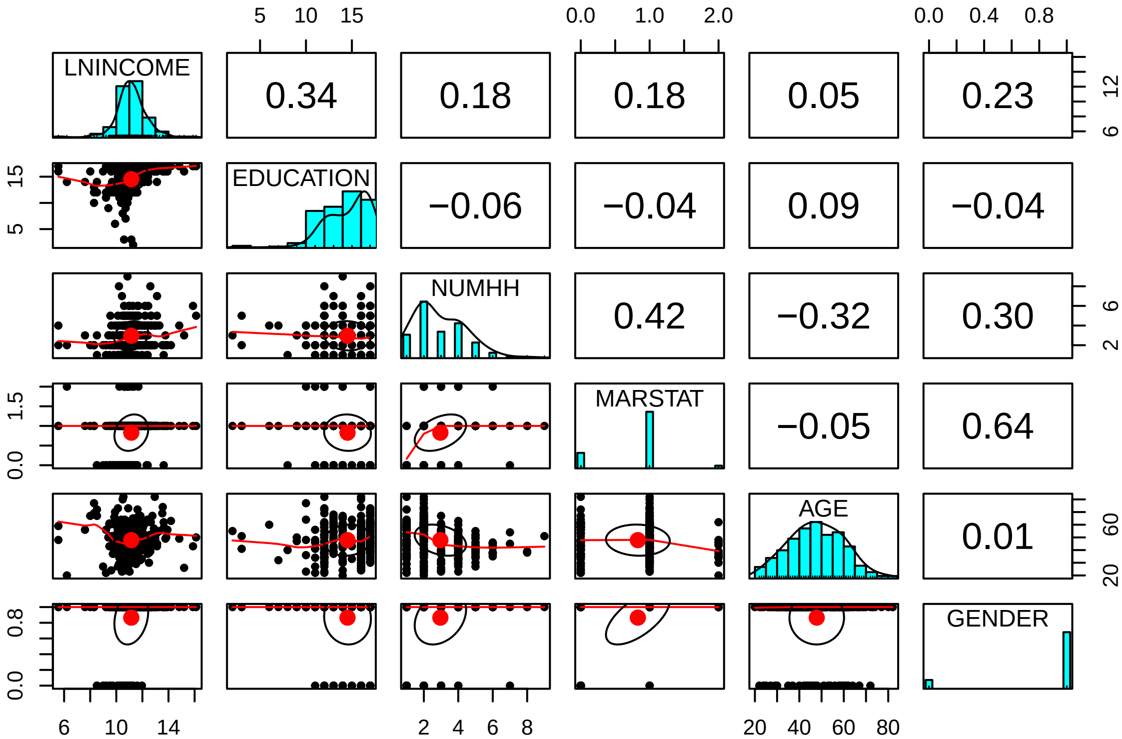
```r
#R Code here
library(foreign)
library(psych)
```

```
##
## Attaching package: 'psych'
```

```
## The following object is masked from 'package:car':
##
##     logit
```

```r
library(car)

y <- subTerm[, c('LNINCOME', 'EDUCATION', 'NUMHH', 'MARSTAT', 'AGE', 'GENDER')]
pairs.panels(y)
```




**Comment** The greatest correlation between variables in this model is 0.64 with MARSTAT and GENDER. Because the pairwise correlation is greater than 0.5, collinearity may be an issue between these variables.

### PART 3 (25 points)

(Use condo.csv data file) A real estate agent wishes to determine the selling price of residences using the size (square feet) and whether the residence is a condominium or a single- family home.

a) Fit a regression model to predict the selling price for residences and provide the regression equation.

```r
#R Code here
condo <- read.csv('condo.csv')
n <- lm(condo$price ~ (condo$sqfeet + condo$condo))
summary(n)
```

```
##
## Call:
## lm(formula = condo$price ~ (condo$sqfeet + condo$condo))
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -42627 -15636  -7005  10441  98485
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  66001.33   48476.33   1.362  0.19112
## condo$sqfeet    90.37      29.53   3.060  0.00708 **
## condo$condo   3629.50   15891.23   0.228  0.82206
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 32190 on 17 degrees of freedom
## Multiple R-squared:  0.4065, Adjusted R-squared:  0.3367
## F-statistic: 5.822 on 2 and 17 DF,  p-value: 0.01186
```

**Comment** b0 = 66001.33 b1 = 90.37 b2 = 3629.50 condo = 1 (condo) or 0 (single family home)

The regression equation is Selling Price = 66001.33 + 90.37sqft + 3629.50condo

b) Interpret the parameters $\beta_1$ and $\beta_2$ in the model given in part (a).

**Comment** As square footage increases, selling price will also increase. For b1, selling price is predicted to increase by $90.37 for each square foot. For b2, a condominim is expected to sell for $3629.50 more than a single-family home, assuming that condo = 1 is a condo and condo = 0 is a single-family home.

c) Fit a new regression model now including the interaction term x1* x2 and provide the regression equation.

```
#R Code here
a <- lm(condo$price ~ (condo$sqfeet * condo$condo))
summary(a)
```

```
##
## Call:
## lm(formula = condo$price ~ (condo$sqfeet * condo$condo))
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -24320 -14900  -7791   9482 101302
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               148863.88   64449.58   2.310   0.0346 *
## condo$sqfeet                  38.45      39.83   0.965   0.3488
## condo$condo              -167044.18   95189.90  -1.755   0.0984 .
## condo$sqfeet:condo$condo     100.69      55.46   1.815   0.0882 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 30210 on 16 degrees of freedom
## Multiple R-squared:  0.5079, Adjusted R-squared:  0.4156
## F-statistic: 5.504 on 3 and 16 DF,  p-value: 0.008608
```

**Comment** Selling Price = 148863.88 + 38.45sqft - 167044.18condo + 100.69sqft*condo

d) Describe what including this interaction term accomplishes. This interaction term makes sqft no longer significant, and it makes condo and sqft:condo significant.

**Comment**

e) Conduct a test of hypothesis to determine if the relationship between the selling price and the square footage is different between condominiums and single-family homes.

```
#R Code here
c = read.csv('condo.csv')
cm = lm(price ~ .*., data=c)
summary(cm)
```

```
##
## Call:
## lm(formula = price ~ . * ., data = c)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -24320 -14900  -7791   9482 101302
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   148863.88   64449.58   2.310   0.0346 *
## condo        -167044.18   95189.90  -1.755   0.0984 .
## sqfeet            38.45      39.83   0.965   0.3488
## condo:sqfeet     100.69      55.46   1.815   0.0882 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 30210 on 16 degrees of freedom
## Multiple R-squared:  0.5079, Adjusted R-squared:  0.4156
## F-statistic: 5.504 on 3 and 16 DF,  p-value: 0.008608
```

**Comment** The F value is 5.504 and the p-value is 0.008608. There is a significant difference between the price and square footage of condos and single-family homes. The null hypothesis can be rejected.

### PART 4 (25 points)

The data set fat (Library: UsingR) contains several body measurements that can be done using a scale and a tape measure. These can be used to predict the body-fat percentage (body.fat). Measuring body fat requires a special apparatus; if our resulting model fits well, we have a low-cost alternative.

a) Partition the data into 60% for training and 40% for testing. Use set.seed(25) before data partition.

```r
#R Code here
library(UsingR)
```

```
## Loading required package: MASS

## Loading required package: HistData

## Loading required package: Hmisc

## Loading required package: lattice

## Loading required package: survival

## Loading required package: Formula

## Loading required package: ggplot2

##
## Attaching package: 'ggplot2'

## The following objects are masked from 'package:psych':
##
##     %+%, alpha

##
## Attaching package: 'Hmisc'

## The following object is masked from 'package:psych':
```

```
##
##     describe
## The following objects are masked from 'package:base':
##
##     format.pval, units
##
## Attaching package: 'UsingR'
## The following object is masked from 'package:survival':
##
##     cancer
## The following object is masked from 'package:psych':
##
##     headtail
```

```
set.seed(25)

index.fat <- sample(x=nrow(fat), size=0.60*nrow(fat))
training.fat <- fat[index.fat,]
testing.fat <- fat[-index.fat,]
```

b) Use training data to develop a multiple linear regression model with body.fat as response variable and age, weight, height, BMI, neck, chest, abdomen, hip, thigh, knee, ankle, bicep, forearm, and wrist as independent variables.

```
#R Code here
m <- lm(body.fat ~ age + weight + height + BMI + neck + chest + abdomen + hip + thigh + knee + ankle + b
summary(m)
```

```
##
## Call:
## lm(formula = body.fat ~ age + weight + height + BMI + neck +
##     chest + abdomen + hip + thigh + knee + ankle + bicep + forearm +
##     wrist, data = training.fat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.7542 -2.8408 -0.0655  2.8135  8.2465
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.86609   45.57172   0.238   0.8119
## age          0.05581    0.04123   1.354   0.1781
## weight       0.01852    0.12728   0.146   0.8845
## height      -0.46390    0.62662  -0.740   0.4604
## BMI         -0.16532    0.90847  -0.182   0.8559
## neck        -0.56039    0.31467  -1.781   0.0772 .
## chest       -0.07661    0.13574  -0.564   0.5734
## abdomen      0.83618    0.11671   7.165 4.47e-11 ***
## hip         -0.28763    0.18055  -1.593   0.1135
## thigh        0.17800    0.19240   0.925   0.3565
## knee         0.22955    0.31363   0.732   0.4655
## ankle        0.29466    0.24351   1.210   0.2284
## bicep        0.02938    0.23813   0.123   0.9020
```

```
## forearm      0.23173    0.32809   0.706   0.4812
## wrist       -0.78185    0.70017  -1.117   0.2661
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.095 on 136 degrees of freedom
## Multiple R-squared:  0.747,  Adjusted R-squared:  0.7209
## F-statistic: 28.68 on 14 and 136 DF,  p-value: < 2.2e-16
```

c) Use the stepAIC function to select a model. Report model summary and provide equation for this model.

```r
#R Code here
s <- stepAIC(m, trace = F)
summary(s)
```

```
##
## Call:
## lm(formula = body.fat ~ height + neck + abdomen + hip, data = training.fat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.1132 -2.5769 -0.3226  2.7584  8.4243
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.47347    9.40857   0.688   0.4925
## height      -0.33918    0.14977  -2.265   0.0250 *
## neck        -0.57668    0.23220  -2.484   0.0141 *
## abdomen      0.84737    0.07226  11.726   <2e-16 ***
## hip         -0.20157    0.11571  -1.742   0.0836 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.037 on 146 degrees of freedom
## Multiple R-squared:  0.736,  Adjusted R-squared:  0.7287
## F-statistic: 101.7 on 4 and 146 DF,  p-value: < 2.2e-16
```

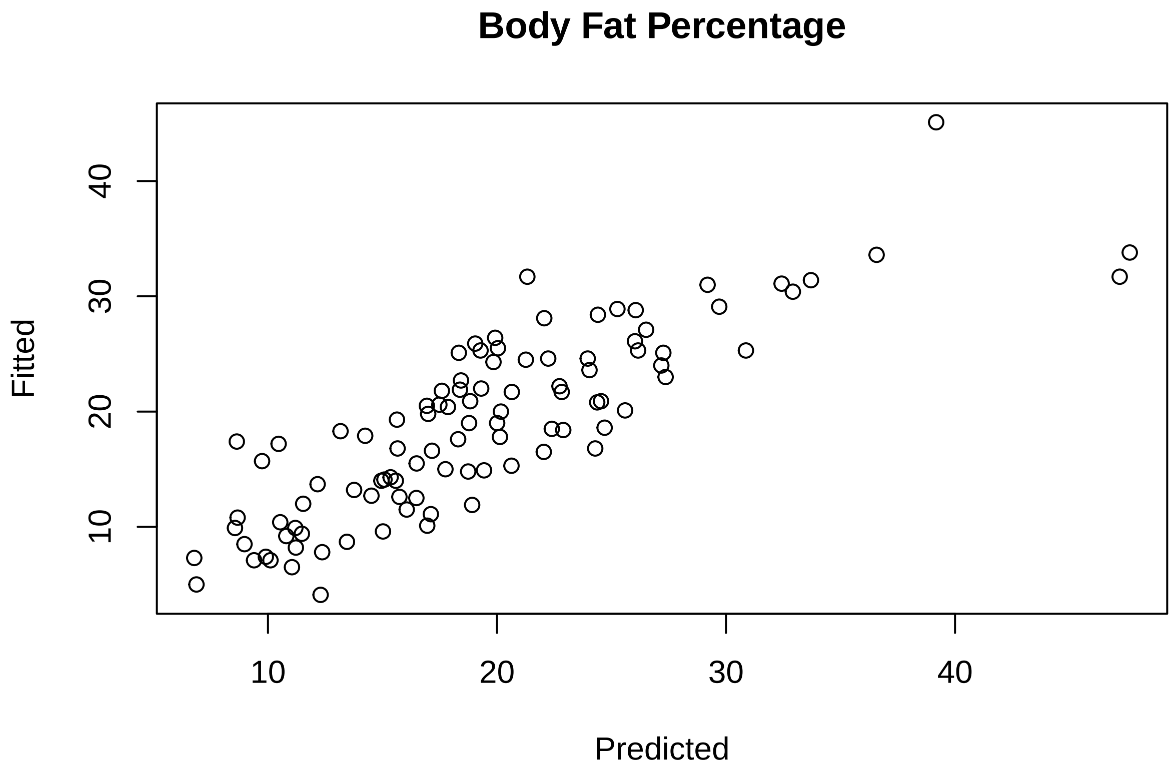
**Comment** The regression equation for this model is body fat = 6.47347 - 0.33918*height -0.57668*neck + 0.84737*abdomen - 0.20157*hip

d) What are the top three contributors to the body-fat percentage? Provide an interpretation for these three coefficients.

**Comment** The top 3 contributors to the body-fat percentage are abdomen, neck, and height. Higher abdomen values correspond to higher body-fat percentage. Lower height and neck values correspond to lower body-fat percentage.

e) Develop a scatter plot for predicted and fitted response values using the testing data. Obtain R2 using testing data based on predicted and fitted response values?

```r
#R Code here
p <- predict(m, testing.fat)
f <- (testing.fat$body.fat)
plot(p, f, xlab = 'Predicted', ylab = 'Fitted', main = 'Body Fat Percentage')
```




```
l <- lm(f ~ p)
summary(l)
```

```
##
## Call:
## lm(formula = f ~ p)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.4936  -3.1883  -0.3361   3.3522  11.2266
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.56634    1.14594    2.24   0.0274 *
## p            0.83971    0.05475   15.34   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.261 on 99 degrees of freedom
## Multiple R-squared:  0.7038, Adjusted R-squared:  0.7008
## F-statistic: 235.2 on 1 and 99 DF,  p-value: < 2.2e-16
```

**Comment** The R squared value for predicted and fitted values using the testing set is 0.70.