

# 02 Regression Homework

Megha Pandit

February 1, 2019

1. Describe the null hypotheses to which the p-values given in Table 3.4 correspond. Explain what conclusions you can draw based on these p-values. Your explanation should be phrased in terms of sales, TV, radio, and newspaper, rather than in terms of the coefficients of the linear model.

The null hypotheses to which the p-values given in Table 3.4 correspond to are that TV, radio and newspaper advertising have no relationship with sales. More specifically,  $H_0^{(1)} : \beta_1 = 0$ ,  $H_0^{(2)} : \beta_2 = 0$ ,  $H_0^{(3)} : \beta_3 = 0$ , where,  $\beta_1$  is the coefficient for TV advertising budget,  $\beta_2$  is the coefficient for radio advertising budget and  $\beta_3$  is the coefficient for newspaper advertising budget. For TV and radio advertising, the p-values are very small indicating that we can reject their corresponding null hypotheses. Therefore, we can reject  $H_0^{(1)}$  and  $H_0^{(2)}$ . However, the p-value for newspaper advertising is large indicating that we cannot reject the corresponding null hypothesis,  $H_0^{(3)}$ . Therefore, from the p-value, newspaper advertising does not have an effect on sales.

2. Carefully explain the differences between the KNN classifier and KNN regression methods.

KNN classifier is used to solve classification problems, usually where the response is qualitative with more than one or two levels. KNN regression method is used to solve regression problems where the response is either quantitative or qualitative in some cases. The major difference is in the mechanism of the two methods. KNN classifier identifies the neighborhood of  $x_0$  and then estimates  $f(x_0)$  as the average of all the responses in the neighborhood.

5. Consider the fitted values that result from performing linear regression without an intercept. In this setting, the  $i$ th fitted value takes the form  $\hat{y}_i = x_i \hat{\beta}$  where,

$$\hat{\beta} = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i'=1}^n x_{i'}^2}$$

Show that we can write

$$\hat{y}_i = \sum_{i'=1}^n a_{i'} y_{i'}$$

what is  $a_{i'}$  ?

We have  $\hat{y}_i = x_i \hat{\beta}$  and  $\hat{\beta} = (\sum_{i=1}^n x_i y_i) / (\sum_{i'=1}^n x_{i'}^2)$  therefore,

$$\begin{aligned} \hat{y}_i &= x_i \frac{\sum_{i=1}^n x_i y_i}{\sum_{i'=1}^n x_{i'}^2} = \frac{\sum_{i=1}^n x_i \frac{x_i}{n} y_i}{\sum_{i'=1}^n x_{i'}^2} \\ \hat{y}_i &= \sum_{i'=1}^n \left( \frac{\frac{x_{i'}^2 y_i}{n}}{x_{i'}^2} \right) = \sum_{i'=1}^n \frac{1}{n} y_{i'} = \sum_{i'=1}^n a_{i'} y_{i'} \end{aligned}$$

Therefore,

$$a_{i'} = \frac{1}{n}$$

6. Using (3.4), argue that in the case of simple linear regression, the least squares line always passes through the point  $(\bar{x}, \bar{y})$ .

Substituting  $\bar{x}$  for  $x$  and  $\bar{y}$  for  $y$  in the least squares equation, we get:

$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$$

From 3.4, we have  $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ . Therefore,  $\bar{y} = \bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 \bar{x} = \bar{y}$ . The above implies that the least squares line always passes through the point  $(\bar{x}, \bar{y})$ .

11. In this problem we will investigate the t-statistic for the null hypothesis  $H_0 : \beta = 0$  in simple linear regression without an intercept. To begin, we generate a predictor  $x$  and a response  $y$  as follows.

```
set.seed(1)
x=rnorm(100)
y=2*x+rnorm(100)
```

- (a) Perform a simple linear regression of  $y$  onto  $x$ , without an intercept. Report the coefficient estimate  $\hat{\beta}$ , the standard error of this coefficient estimate, and the t-statistic and p-value associated with the null hypothesis  $H_0 : \beta = 0$ . Comment on these results. (You can perform regression without an intercept using the command `lm(y~x+0)`.)

```
fit1 <- lm(y~x + 0)
summary(fit1)

##
## Call:
## lm(formula = y ~ x + 0)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.9154 -0.6472 -0.1771  0.5056  2.3109
##
## Coefficients:
##      Estimate Std. Error t value Pr(>|t|)
## x      1.9939      0.1065   18.73  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9586 on 99 degrees of freedom
## Multiple R-squared:  0.7798, Adjusted R-squared:  0.7776
## F-statistic: 350.7 on 1 and 99 DF,  p-value: < 2.2e-16
```

The coefficient estimate  $\hat{\beta}$  is 1.9939, the standard error of  $\hat{\beta}$  is 0.1065, the t-statistic is 18.73, and the p-value is  $<2e-16$ . The null hypothesis in this case is  $H_0 : \beta = 0$ . But, the large t-statistic and the small p-value ( $<2e-16$ ) allows us to reject the null hypothesis. Therefore, there is a significant relationship between  $x$  and  $y$ .

- (b) Now perform a simple linear regression of  $x$  onto  $y$  without an intercept, and report the coefficient estimate, its standard error, and the corresponding t-statistic and p-values associated with the null hypothesis  $H_0 : \beta = 0$ . Comment on these results.

```
fit2 <- lm(x~y + 0)
summary(fit2)

##
## Call:
## lm(formula = x ~ y + 0)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.8699 -0.2368  0.1030  0.2858  0.8938
##
## Coefficients:
##      Estimate Std. Error t value Pr(>|t|)
## y   0.39111     0.02089   18.73  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4246 on 99 degrees of freedom
## Multiple R-squared:  0.7798, Adjusted R-squared:  0.7776
## F-statistic: 350.7 on 1 and 99 DF,  p-value: < 2.2e-16
```

The coefficient estimate  $\hat{\beta}$  is 0.39111, the standard error of  $\hat{\beta}$  is 0.02089, the t-statistic is 18.73, and the p-value is <2e-16. The null hypothesis in this case is  $H_0 : \beta = 0$ . But, the large t-statistic and the small p-value (<2e-16) allows us to reject the null hypothesis. Therefore, there is a significant relationship between y and x.

- (c) What is the relationship between the results obtained in (a) and (b)? In both (a) and (b), the values for the t-statistic and the p-values are the same. This implies that both of them reflect the same line, i.e.,  $y = 2x + \epsilon$  can also be written as  $x = 0.5(y - \epsilon)$ .

- (d) For the regression of Y onto X without an intercept, the t-statistic for  $H_0 : \beta = 0$  takes the form  $\hat{\beta}/SE(\hat{\beta})$ , where  $\hat{\beta}$  is given by (3.38), and where

$$SE(\hat{\beta}) = \sqrt{\frac{\sum_{i=1}^n (y_i - x_i \hat{\beta})^2}{(n-1) \sum_{i=1}^n x_i^2}}$$

(These formulas are slightly different from those given in Sections 3.1.1 and 3.1.2, since here we are performing regression without an intercept.) Show algebraically, and confirm numerically in R, that the t-statistic can be written as:

$$\frac{(\sqrt{n-1}) \sum_{i=1}^n x_i y_i}{\sqrt{(\sum_{i=1}^n x_i^2)(\sum_{i=1}^n y_i^2) - (\sum_{i=1}^n x_i y_i)^2}}$$

We know

$$t = \frac{\hat{\beta}}{SE(\hat{\beta})} = \frac{\frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}}{\sqrt{\frac{\sum_{i=1}^n (y_i - x_i \hat{\beta})^2}{(n-1) \sum_{i=1}^n x_i^2}}} = \frac{(\sqrt{n-1}) \sum_{i=1}^n x_i y_i}{\sqrt{(\sum_{i=1}^n x_i^2)(\sum_{i=1}^n (y_i^2 - 2y_i x_i \hat{\beta} + x_i^2 \hat{\beta}^2))}} = \frac{(\sqrt{n-1}) \sum_{i=1}^n x_i y_i}{\sqrt{\sum_{i=1}^n x_i^2 \sum_{i=1}^n y_i^2 - \sum_{i=1}^n x_i^2 \hat{\beta} (2 \sum_{i=1}^n x_i y_i)}}$$

$$t = \frac{(\sqrt{n-1}) \sum_{i=1}^n x_i y_i}{\sqrt{\sum_{i=1}^n x_i^2 \sum_{i=1}^n y_i^2 - \sum_{i=1}^n x_i y_i (2 \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i y_i)}} = \frac{(\sqrt{n-1}) \sum_{i=1}^n x_i y_i}{\sqrt{\sum_{i=1}^n x_i^2 \sum_{i=1}^n y_i^2 - (\sum_{i=1}^n x_i y_i)^2}}$$

```
n <- length(x)
t <- sqrt(n - 1)*(x %*% y)/sqrt(sum(x^2) * sum(y^2) - (x %*% y)^2)
as.numeric(t)
```

```
## [1] 18.72593
```

The t-statistic above is the same as the one in part (a) and (b).

- (e) Using the results from (d), argue that the t-statistic for the regression of y onto x is the same as the t-statistic for the regression of x onto y.

If we replace  $x_i$  with  $y_i$  in the above equations, we would get the same result.

- (f) In R, show that when regression is performed with an intercept, the t-statistic for  $H_0 : \beta_1 = 0$  is the same for the regression of y onto x as it is for the regression of x onto y.

```
fit3 <- lm(y ~ x)
summary(fit3)

##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.8768 -0.6138 -0.1395  0.5394  2.3462
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.03769    0.09699  -0.389   0.698
## x            1.99894    0.10773  18.556 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9628 on 98 degrees of freedom
## Multiple R-squared:  0.7784, Adjusted R-squared:  0.7762
## F-statistic: 344.3 on 1 and 98 DF,  p-value: < 2.2e-16

fit4 <- lm(x ~ y)
summary(fit4)

##
## Call:
## lm(formula = x ~ y)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.90848 -0.28101  0.06274  0.24570  0.85736
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.03880    0.04266   0.91   0.365
## y            0.38942    0.02099  18.56 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4249 on 98 degrees of freedom
## Multiple R-squared:  0.7784, Adjusted R-squared:  0.7762
## F-statistic: 344.3 on 1 and 98 DF,  p-value: < 2.2e-16
```

As we can see from the above two regressions, the t-statistic for  $\beta_1$  for both the regressions is the same.

12. This problem involves simple linear regression without an intercept.

- (a) Recall that the coefficient estimate  $\hat{\beta}$  for the linear regression of Y onto X without an intercept is given by (3.38). Under what circumstance is the coefficient estimate for the regression of X onto Y the same as the coefficient estimate for the regression of Y onto X?

The coefficient estimate for the regression of Y onto X is given by

$$\hat{\beta} = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}$$

The coefficient estimate for the regression of X onto Y is given by

$$\hat{\beta} = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n y_i^2}$$

Therefore, the coefficient estimate for the regression of X onto Y is the same as the coefficient estimate of Y onto X when  $\sum_{i=1}^n x_i^2 = \sum_{i=1}^n y_i^2$

- (b) Generate an example in R with  $n = 100$  observations in which the coefficient estimate for the regression of X onto Y is different from the coefficient estimate for the regression of Y onto X.

```
x <- 1:100
sum(x^2)

## [1] 338350

y <- 2*x+rnorm(100)
sum(y^2)

## [1] 1354360

lm1 <- lm(y ~ x)
summary(lm1)

##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.90447 -0.49129 -0.00858  0.67418  2.60155
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.008886   0.209427  -0.042   0.966
## x            2.000764   0.003600 555.708 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.039 on 98 degrees of freedom
## Multiple R-squared:  0.9997, Adjusted R-squared:  0.9997
## F-statistic: 3.088e+05 on 1 and 98 DF, p-value: < 2.2e-16

lm2 <- lm(x ~ y)
summary(lm2)

##
## Call:
```

```
## lm(formula = x ~ y)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.29241 -0.33546 -0.00682  0.25149  1.44535
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.0204610  0.1046374   0.196   0.845
## y           0.4996506  0.0008991 555.708 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5194 on 98 degrees of freedom
## Multiple R-squared:  0.9997, Adjusted R-squared:  0.9997
## F-statistic: 3.088e+05 on 1 and 98 DF,  p-value: < 2.2e-16
```

- (c) Generate an example in R with  $n = 100$  observations in which the coefficient estimate for the regression of  $X$  onto  $Y$  is the same as the coefficient estimate for the regression of  $Y$  onto  $X$ .

```
x <- 1:100
sum(x^2)

## [1] 338350

y <- 100:1
sum(y^2)

## [1] 338350

lm1 <- lm(y ~ x)
summary(lm1)

##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.575e-14 -5.302e-15 -2.850e-15  4.300e-16  2.680e-13
##
## Coefficients:
##              Estimate Std. Error    t value Pr(>|t|)
## (Intercept)  1.010e+02  5.598e-15  1.804e+16 <2e-16 ***
## x           -1.000e+00  9.624e-17 -1.039e+16 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.778e-14 on 98 degrees of freedom
## Multiple R-squared:  1, Adjusted R-squared:  1
## F-statistic: 1.08e+32 on 1 and 98 DF,  p-value: < 2.2e-16

lm2 <- lm(x ~ y)
summary(lm2)

##
```

```
## Call:
## lm(formula = x ~ y)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.680e-13 -4.300e-16  2.850e-15  5.302e-15  3.575e-14
##
## Coefficients:
##              Estimate Std. Error    t value Pr(>|t|)
## (Intercept)  1.010e+02  5.598e-15  1.804e+16  <2e-16 ***
## y           -1.000e+00  9.624e-17 -1.039e+16  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.778e-14 on 98 degrees of freedom
## Multiple R-squared:  1, Adjusted R-squared:  1
## F-statistic: 1.08e+32 on 1 and 98 DF, p-value: < 2.2e-16
```

13. In this exercise you will create some simulated data and will fit simple linear regression models to it. Make sure to use `set.seed(1)` prior to starting part (a) to ensure consistent results.

- (a) Using the `rnorm()` function, create a vector, `x`, containing 100 observations drawn from a  $N(0, 1)$  distribution. This represents a feature,  $X$ .

```
set.seed(100)
x <- rnorm(100, mean = 0, sd = 1)
```

- (b) Using the `rnorm()` function, create a vector, `eps`, containing 100 observations drawn from a  $N(0, 0.25)$  distribution i.e. a normal distribution with mean zero and variance 0.25.

```
eps <- rnorm(100, mean = 0, sd = 0.25)
```

- (c) Using `x` and `eps`, generate a vector `y` according to the model  $Y = \beta_0 + 0.5X + \epsilon$ . (3.39) What is the length of the vector `y`? What are the values of  $\beta_0$  and  $\beta_1$  in this linear model?

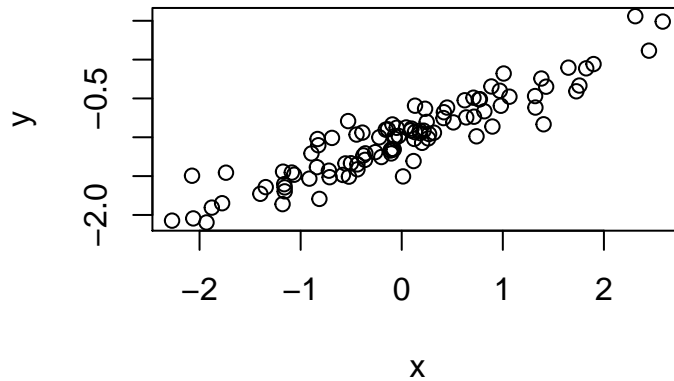
```
y <- -1 + 0.5*x + eps
length(y)
```

```
## [1] 100
```

The values for the coefficient estimates are:  $\beta_0 = -1$  and  $\beta_1 = 0.5$ .

- (d) Create a scatterplot displaying the relationship between `x` and `y`. Comment on what you observe.

```
plot(x,y)
```



From the scatterplot, x and y have a linear relationship.

- (e) Fit a least squares linear model to predict y using x. Comment on the model obtained. How do  $\hat{\beta}_0$  and  $\hat{\beta}_1$  compare to  $\beta_0$  and  $\beta_1$ ?

```
fit5 <- lm(y ~ x)
summary(fit5)
```

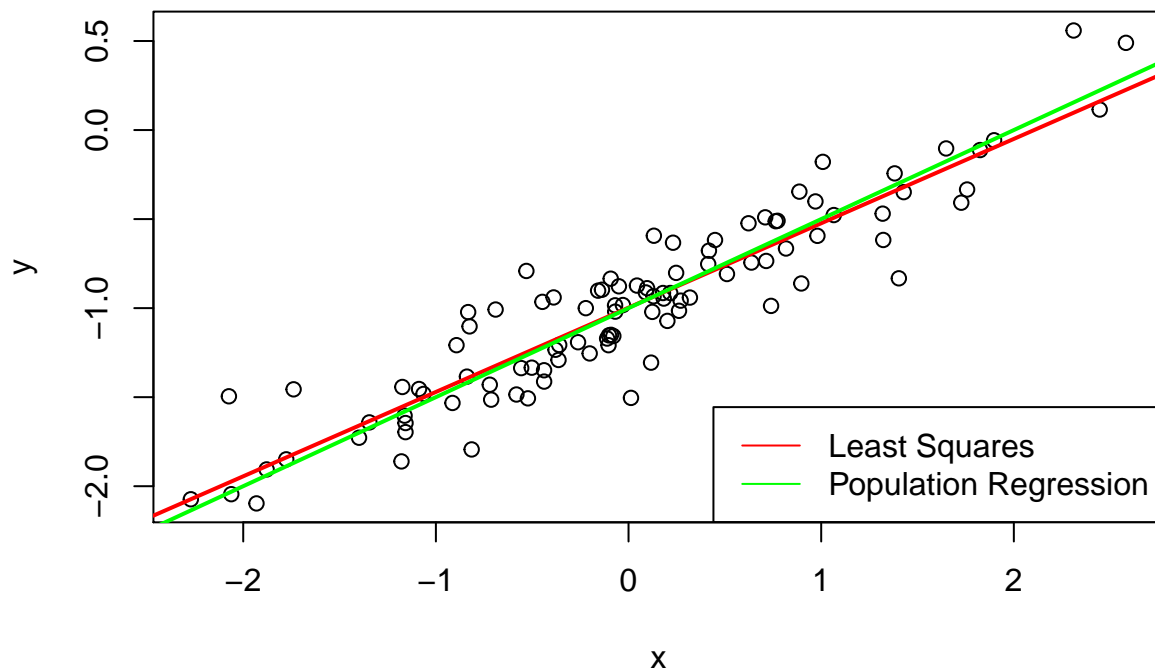
```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.51299 -0.10816 -0.01964  0.12146  0.48464
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.99714    0.01982  -50.30  <2e-16 ***
## x           0.47366    0.01952   24.27  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1982 on 98 degrees of freedom
## Multiple R-squared:  0.8573, Adjusted R-squared:  0.8559
## F-statistic: 588.9 on 1 and 98 DF,  p-value: < 2.2e-16
```

From the above summary,  $\hat{\beta}_0$  and  $\hat{\beta}_1$  are close to the values for  $\beta_0$  and  $\beta_1$ .

- (f) Display the least squares line on the scatterplot obtained in (d). Draw the population regression line on the plot, in a different color. Use the legend() command to create an appropriate legend.



```
plot(x,y)
abline(fit5, col = "red", lwd = 2)
abline(-1, 0.5, col = "green", lwd = 2)
legend("bottomright", c("Least Squares", "Population Regression"), col = c("red", "green"), lty = c(1,1))
```



(g) Now fit a polynomial regression model that predicts  $y$  using  $x$  and  $x^2$ . Is there evidence that the quadratic term improves the model fit? Explain your answer.

```
fit6 <- lm(y ~ x + I(x^2))
summary(fit6)
```

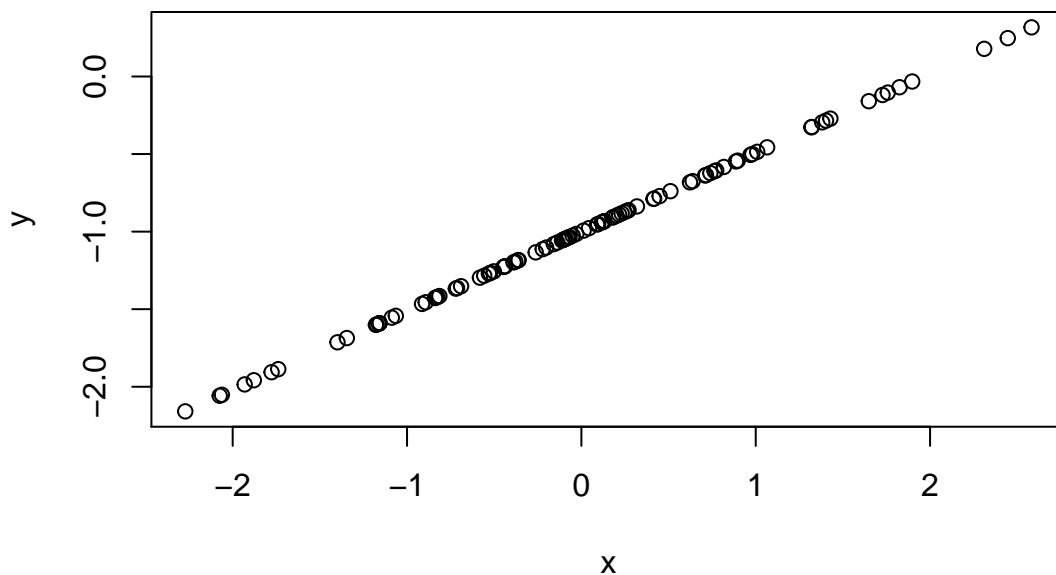
```
##
## Call:
## lm(formula = y ~ x + I(x^2))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.5128 -0.1061 -0.0308  0.1323  0.4691
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.01611    0.02429  -41.838  <2e-16 ***
## x             0.47044    0.01959   24.018  <2e-16 ***
## I(x^2)        0.01840    0.01372    1.341    0.183
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1974 on 97 degrees of freedom
## Multiple R-squared:  0.8599, Adjusted R-squared:  0.857
## F-statistic: 297.8 on 2 and 97 DF,  p-value: < 2.2e-16
```

The inclusion of a quadratic term does not improve the model much. The adjusted R-squared changed from 0.8559 to 0.857, which is a very small improvement. The RSE also shows little improvement from 0.1982 to 0.1974. This can be because, as seen previously,  $x$  and  $y$  share a linear relationship.

- (h) Repeat (a)-(f) after modifying the data generation process in such a way that there is less noise in the data. The model (3.39) should remain the same. You can do this by decreasing the variance of the normal distribution used to generate the error term  $\epsilon$  in (b). Describe your results.

```
set.seed(100)
eps <- rnorm(100, sd = 0.01)
y <- -1 + 0.5*x + eps
plot(x,y)
```



```
fit7 <- lm(y ~ x)
summary(fit7)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.875e-16 -1.038e-16 -5.120e-17 -1.100e-18  3.545e-15
##
```

```
## Coefficients:
##           Estimate Std. Error    t value Pr(>|t|)
## (Intercept) -1.000e+00  4.211e-17 -2.375e+16  <2e-16 ***
## x           5.100e-01  4.146e-17  1.230e+16  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.211e-16 on 98 degrees of freedom
## Multiple R-squared:      1, Adjusted R-squared:      1
## F-statistic: 1.513e+32 on 1 and 98 DF,  p-value: < 2.2e-16
```

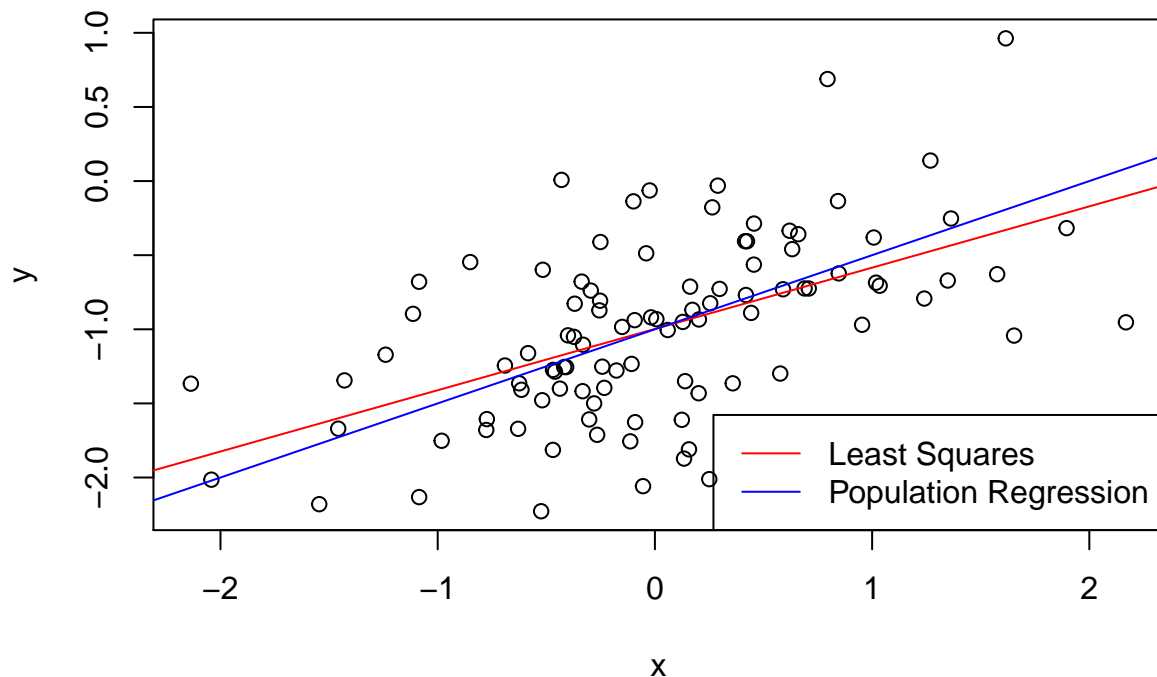
The t-statistic and the p-value both show that the coefficient estimate of  $x$  is significant. As we reduced the noise, the R-squared and RSE values imply a perfect fit, a perfect linear relationship.

- (i) Repeat (a)-(f) after modifying the data generation process in such a way that there is more noise in the data. The model (3.39) should remain the same. You can do this by increasing the variance of the normal distribution used to generate the error term  $\epsilon$  in (b). Describe your results.

```
set.seed(100)
eps <- rnorm(100, sd = 0.5)
x <- rnorm(100)
y <- -1 + 0.5*x + eps
fit8 <- lm(y ~ x)
summary(fit8)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.11672 -0.30279 -0.01578  0.30175  1.35737
##
## Coefficients:
##           Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.99758    0.05083 -19.626  < 2e-16 ***
## x           0.41341    0.06416   6.444 4.38e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5083 on 98 degrees of freedom
## Multiple R-squared:  0.2976, Adjusted R-squared:  0.2904
## F-statistic: 41.52 on 1 and 98 DF,  p-value: 4.378e-09
```

```
plot(x,y)
abline(fit8, col = "red")
abline(-1, 0.5, col = "blue")
legend("bottomright", c("Least Squares", "Population Regression"), col = c("red", "blue"), lty = c(1,1))
```



Increasing the variance of the normal distribution led to an increase in the RSE value and a drastic decrease in the R-Squared value. The two regression lines are still quite close given the large dataset we have.

- (j) What are the confidence intervals for  $\beta_0$  and  $\beta_1$  based on the original data set, the noisier data set, and the less noisy data set? Comment on your results.

```
original <- c(confint(fit5))
print(paste0("Confidence interval for  $\beta_0$  in the original data set: ", "[",original[1],",", original[3]

## [1] "Confidence interval for  $\beta_0$  in the original data set: [-1.03647564193144,-0.957800495099955]"
print(paste0("Confidence interval for  $\beta_1$  in the original data set: ", "[",original[2],",", original[4]

## [1] "Confidence interval for  $\beta_1$  in the original data set: [0.434924735825272,0.512391542209512]"
noisier <- c(confint(fit8))
print(paste0("Confidence interval for  $\beta_0$  in the noisier data set: ", "[",noisier[1],",", noisier[3],")")

## [1] "Confidence interval for  $\beta_0$  in the noisier data set: [-1.09844980910578,-0.896708339737883]"
print(paste0("Confidence interval for  $\beta_1$  in the noisier data set: ", "[",noisier[2],",", noisier[4],")")

## [1] "Confidence interval for  $\beta_1$  in the noisier data set: [0.286096018887651,0.540731301583647]"
less_noisy <- c(confint(fit7))

## Warning in summary.lm(object): essentially perfect fit: summary may be
## unreliable
```

```
print(paste0("Confidence interval for ??0 in the less noisy data set: ", "[",less_noisy[1],",", less_no
## [1] "Confidence interval for ??0 in the less noisy data set: [-1,-0.999999999999999]"
print(paste0("Confidence interval for ??1 in the less noisy data set: ", "[",less_noisy[2],",", less_no
## [1] "Confidence interval for ??1 in the less noisy data set: [0.51,0.51]"
```

The intervals seem to be centered around 0.5. With more noise, the confidence intervals become wider and with lesser noise, narrower. The confidence intervals for the less noisy data set are as seen because the model is a perfect fit for the true linear relationship between  $x$  and  $y$ . Also, the R-squared value = 1 and the extremely small RSE suggest that the model is a perfect fit and that the coefficient estimates are almost equal to the true parameter values.

14. This problem focuses on the collinearity problem.

(a) Perform the following commands in R:

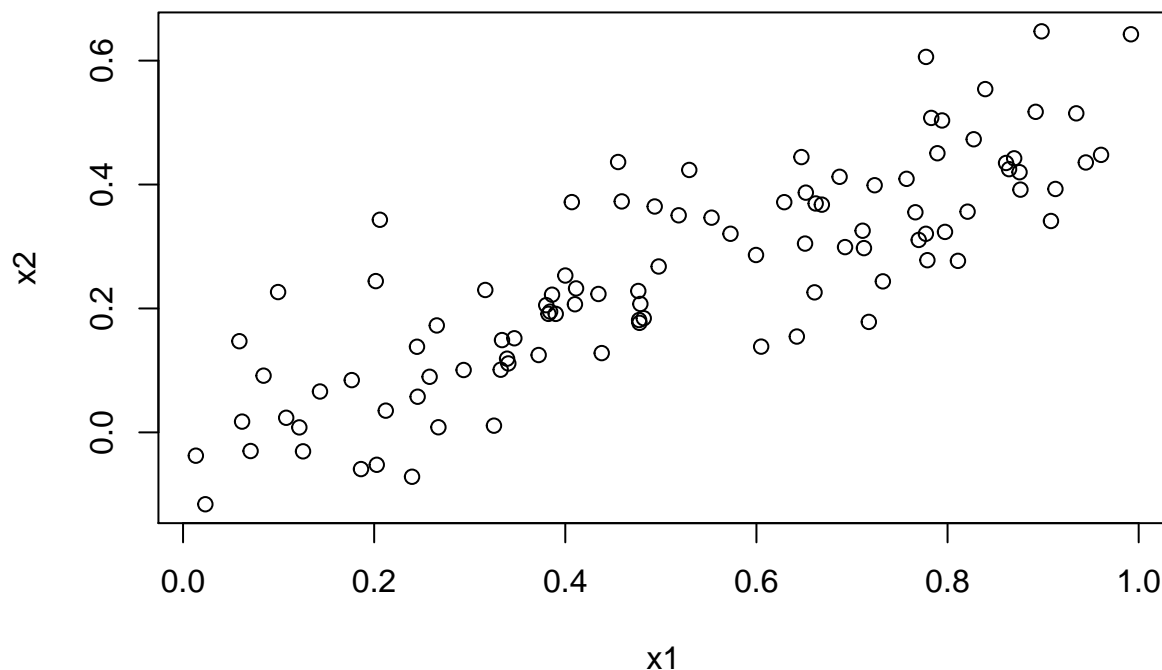
```
set.seed (1)
x1=runif (100)
x2 =0.5* x1+rnorm (100) /10
y=2+2* x1 +0.3* x2+rnorm (100)
```

The last line corresponds to creating a linear model in which  $y$  is a function of  $x_1$  and  $x_2$ . Write out the form of the linear model. What are the regression coefficients?

The linear model is of the form:  $Y = 2 + 2X_1 + 0.3X_2 + \epsilon$ , where  $\epsilon$  is a  $N(0,1)$  random variable. The regression coefficients are  $\beta_0 = 2, \beta_1 = 2$  and  $\beta_2 = 0.3$ .

(b) What is the correlation between  $x_1$  and  $x_2$ ? Create a scatterplot displaying the relationship between the variables.

```
print(paste0("Correlation between x1 and x2: ",cor(x1, x2)))
## [1] "Correlation between x1 and x2: 0.835121242463113"
plot(x1, x2)
```



x1 and x2 seem to be highly correlated.

- (c) Using this data, fit a least squares regression to predict y using x1 and x2. Describe the results obtained. What are  $\hat{\beta}_0$ ,  $\hat{\beta}_1$ , and  $\hat{\beta}_2$ ? How do these relate to the true  $\beta_0$ ,  $\beta_1$ , and  $\beta_2$ ? Can you reject the null hypothesis  $H_0 : \beta_1 = 0$ ? How about the null hypothesis  $H_0 : \beta_2 = 0$ ?

```
fit9 <- lm(y ~ x1 + x2)
summary(fit9)
```

```
##
## 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
```

The coefficient estimate for  $\beta_0$  is significant. Though the estimate for  $\beta_1$  is not completely two standard errors away from the mean, the corresponding p-value is less than 0.05 indicating that the coefficient is significant, and hence, we can reject the null hypothesis that  $\beta_1 = 0$ . As for the estimate for  $\beta_2$ , the p-value, much greater than 0.05, suggests that the coefficient is not statistically significant.

- (d) Now fit a least squares regression to predict  $y$  using only  $x_1$ . Comment on your results. Can you reject the null hypothesis  $H_0 : \beta_1 = 0$ ?

```
fit10 <- lm(y ~ x1)
summary(fit10)
```

```
##
## 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
```

The t-statistic for the estimate  $\beta_1$  is more than 2 and the p-value is much lower than 0.05. Hence, we can reject the null hypothesis  $H_0 : \beta_1 = 0$ . The coefficient is statistically significant.

- (e) Now fit a least squares regression to predict  $y$  using only  $x_2$ . Comment on your results. Can you reject the null hypothesis  $H_0 : \beta_1 = 0$ ?

```
fit11 <- lm(y ~ x2)
summary(fit11)
```

```
##
## 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
```

The t-statistic for the estimate  $\beta_1$  is more than 2 and the p-value is much lower than 0.05. Hence, we can reject the null hypothesis  $H_0 : \beta_1 = 0$ . The coefficient is statistically significant.

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

The results are not contradictory because the model in (c) has the effect of  $x_1$  and  $x_2$  together and the models in (d) and (e) see the effects of  $x_1$  and  $x_2$  on  $y$  individually. Since there is a correlation between  $x_1$  and  $x_2$ , the standard error of the coefficient estimate for  $\beta_1$  becomes larger than it should be when both  $x_1$  and  $x_2$  are included in the model. Also, the importance of  $x_2$  for  $y$  in the (c) model may have been masked due to the presence of correlation.

(g) Now suppose we obtain one additional observation, which was unfortunately mis-measured.

```
x1=c(x1 , 0.1)
x2=c(x2 , 0.8)
y=c(y,6)
```

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

```
fit12 <- lm(y ~ x1 + x2)
summary(fit12)
```

```
##
## 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
```

```
plot(fit12)
```

```
fit13 <- lm(y ~ x1)
summary(fit13)
```

```
##
## Call:
```



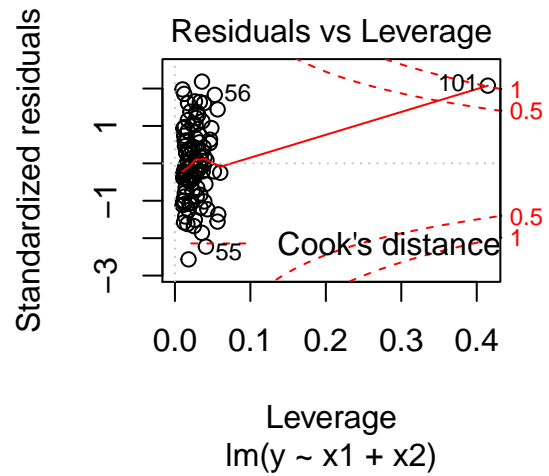
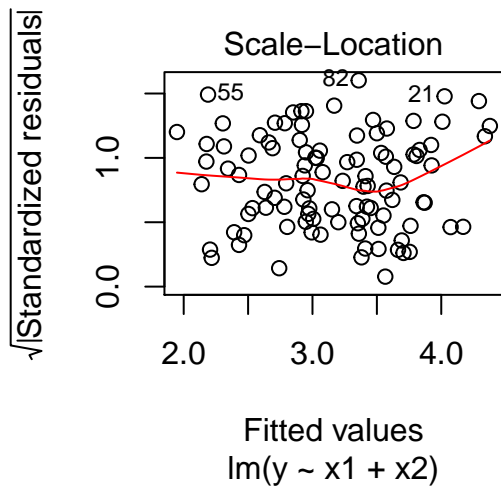
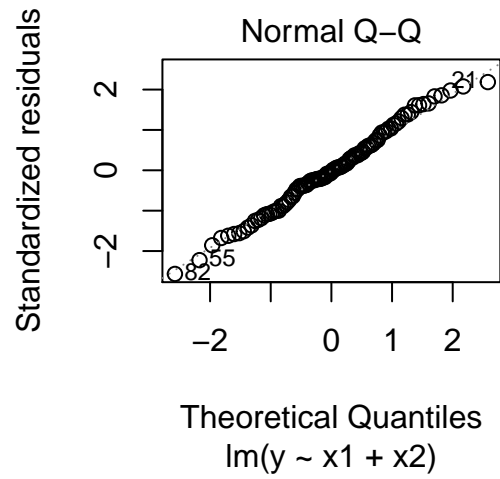
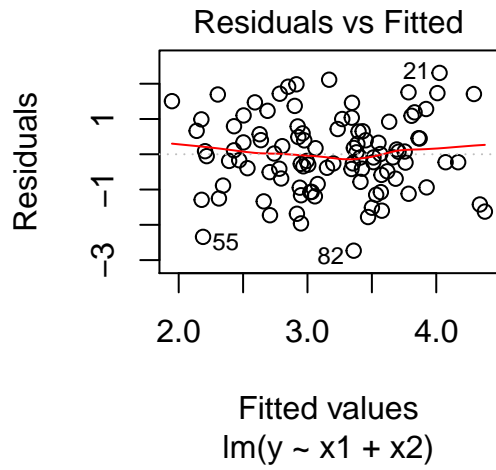
```
## lm(formula = y ~ x1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.8897 -0.6556 -0.0909  0.5682  3.5665
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.2569     0.2390   9.445 1.78e-15 ***
## x1             1.7657     0.4124   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
```

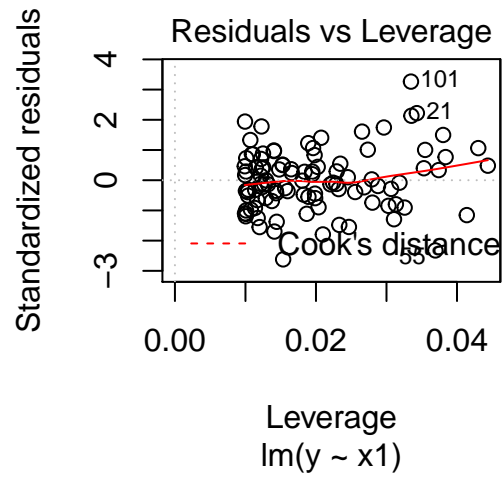
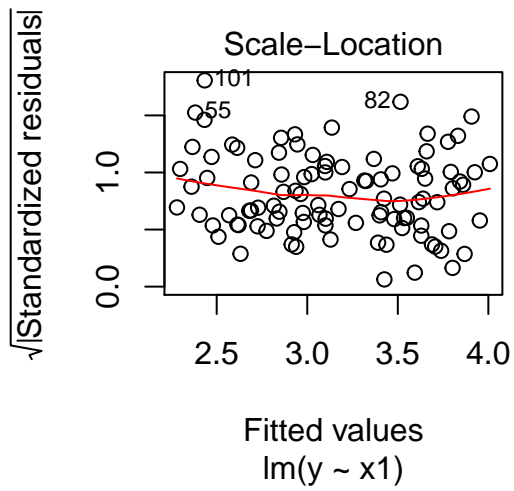
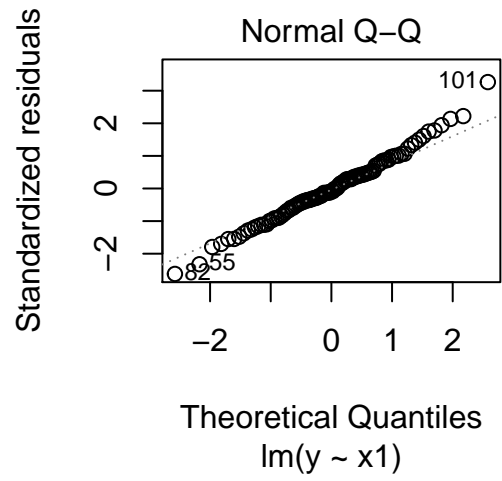
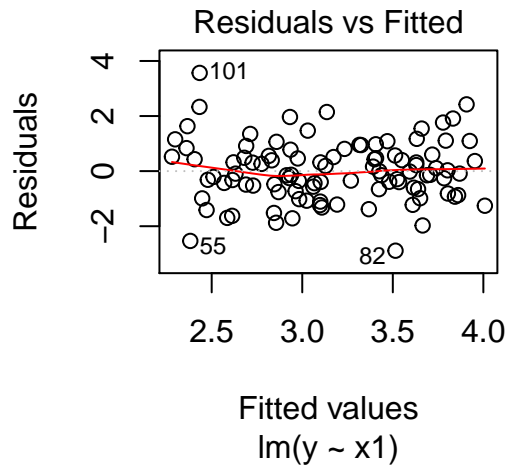
```
plot(fit13)
```

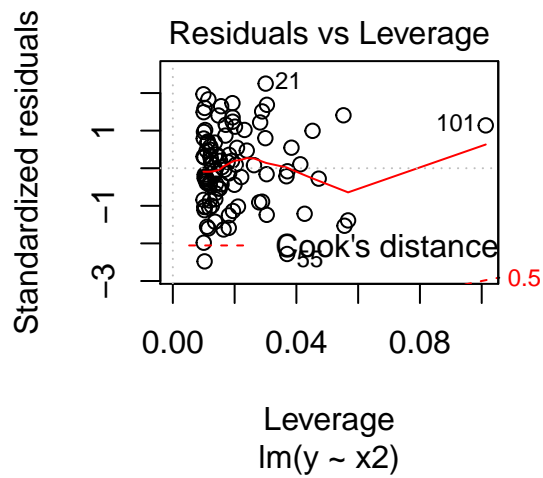
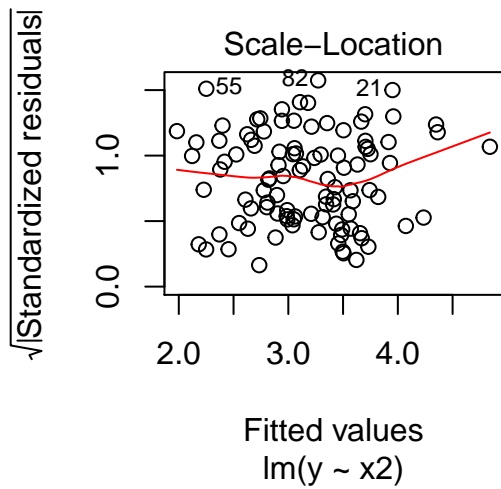
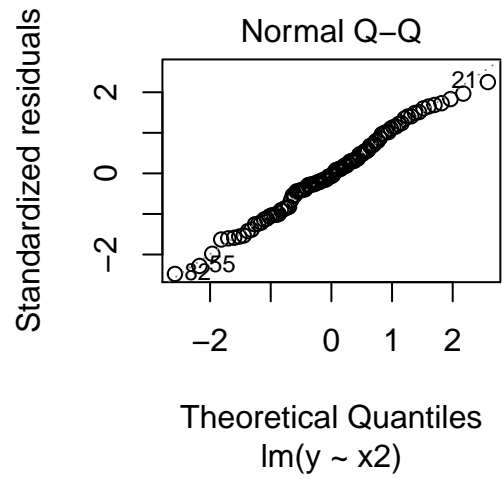
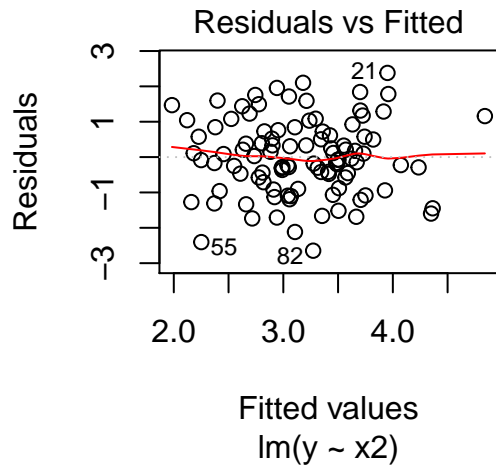
```
fit14 <- lm(y ~ x2)
summary(fit14)
```

```
##
## 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(fit14)
```







In the model with both  $x1$  and  $x2$  as predictors, the last point seems to be a high leverage point, from the residuals vs leverage plot. In the model with  $x1$  as the sole predictor, the last point is an outlier. In the model with  $x2$  as the sole predictor, the last point is a high leverage point.