Data Science and Machine Learning
National Technical University of Athens

Statistical Modelling

1st Assignment
Statistic Models

Nanos Georgios

nanosgiwrgos1997@gmail.com

"All models are wrong, but some are useful"

– George E. P. Box

Aim

The aim of this assignment is to get acquainted with the simple linear model. We will export the linear regression function for data of two applications and we will analyze and interpret the results from the graphs that we will export.

Part A

It is required to show for the simple linear model $E(y_x) = \beta_0 + \beta_1 * x$ that the following are valid:

1. $R^2 = r_{xy}^2$ where R^2 is the o coefficient of determination and r_{xy} is the sample factor (Pearson) correlation of x and y observations.
2. $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i$
3. $cov(\overline{y}, \hat{\beta}_1) = 0$
4. $\sum_{i=1}^{n} y_i \hat{y}_i = \sum_{i=1}^{n} \hat{y}_i{}^2$
5. $\sum_{i=1}^{n} (y_i - \hat{y}_i)(\hat{y}_i - \overline{y_i}) = 0$
6. $\frac{\hat{\beta}_1}{se(\hat{\beta}_1)} = \frac{r_{xy}\sqrt{n-1}}{\sqrt{1-r_{xy}^2}}$

1. The sample correlation coefficient (Pearson) r_{xy} of the x and y observations is equal to: $r_{xy} = \frac{s_{xy}}{s_x s_y}$. The coefficient of determination (or good fit) is equal to $R^2 = \frac{SSR}{SST} = \frac{\sum_{i=1}^{n} (\hat{y_i} - \overline{y})^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2} = \hat{\beta}^2 * \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2} = (\frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2})^2 * \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2 (y_i - \overline{y})^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2 \sum_{i=1}^{n} (y_i - \overline{y})^2} = \frac{s_{xy}^2}{s_x^2 s_y^2} = r_{xy}^2$

2. The estimation of the value of the random variable Y is equal to $\hat{y}_i = \hat{a} + \hat{\beta} x_i \sum_{i=1}^{n} \hat{y}_i = \sum_{i=1}^{n} \hat{\alpha} + \hat{\beta} x_i = \sum_{i=1}^{n} (\overline{y} - \hat{\beta}\overline{x}) + \sum_{i=1}^{n} \hat{\beta} x_i = \sum_{i=1}^{n} \overline{y} - \sum_{i=1}^{n} \hat{\beta}\overline{x} + \sum_{i=1}^{n} \hat{\beta} x_i = \sum_{i=1}^{n} \frac{\sum_{i=1}^{n} y_i}{n} - \sum_{i=1}^{n} \hat{\beta} \frac{\sum_{i=1}^{n} x_i}{n} + \sum_{i=1}^{n} \hat{\beta} x_i = \sum_{i=1}^{n} yi - \sum_{i=1}^{n} \hat{\beta} x_i + - \sum_{i=1}^{n} \hat{\beta} x_i = \sum_{i=1}^{n} yi$

3. $cov(\overline{y}, \hat{\beta}) = \sum_{i=1}^{n} (\overline{y} - \overline{\overline{y}})(\overline{\hat{\beta}} - \overline{\overline{\hat{\beta}}}) = \sum_{i=1}^{n} (1/n * \sum_{i=1}^{n} y_i - 1/n * \sum_{i=1}^{n} (1/n * \sum_{i=1}^{n} y_i)(\hat{\beta} - 1/n * \sum_{i=1}^{n} \hat{\beta})) = 0$

where,

$1/n * \sum_{i=1}^{n} y_i - 1/n * \sum_{i=1}^{n} (1/n * \sum_{i=1}^{n} y_i) = 1/n * \sum_{i=1}^{n} y_i - 1/n^2 * n * \sum_{i=1}^{n} y_i = 0$

4. From 2. we have proved that $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i$. Therefore multiplying both members by the term $\sum_{i=1}^{n} \hat{y}_i$ the result is obtained.

5. It is true that : $\overline{y} = \hat{\alpha} + \hat{\beta}\overline{x}$. Therefore, $\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \overline{y_i}) = \sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta}x_i)(\hat{\alpha} + \hat{\beta}x_i - (\hat{\alpha} + \hat{\beta}\overline{x})) = \sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta}x_i)(x_i - \overline{x}) * \hat{\beta} = \sum_{i=1}^{n}(y_i - \overline{y} + \hat{\beta}\overline{x} - \hat{\beta}x_i)(x_i - \overline{x}) * \hat{\beta} = \sum_{i=1}^{n}(y_i - \overline{y})(x_i - \overline{x}) * \hat{\beta} - \sum_{i=1}^{n}\hat{\beta}^2(x_i - \overline{x})^2$

The estimate of β is equal to: $\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$. Therefore, $\sum_{i=1}^{n}(y_i - \overline{y})(x_i - \overline{x}) * \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} - \sum_{i=1}^{n}(\frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2})^2(x_i - \overline{x})^2 = \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) * \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} - \sum_{i=1}^{n}(x_i - \overline{x})^2 * \frac{1}{(\sum_{i=1}^{n}(x_i - \overline{x})^2)^2} * \sum_{i=1}^{n}(x_i - \overline{x})^2(y_i - \overline{y})^2 = 0$

6. It is true that $\hat{\beta}1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{S_{xy}}{S_{xx}}$. Also, for the standar error it is true that $\hat{\beta}1$ ότι: $se(\hat{\beta}1) = \sqrt{\frac{S^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}} = \frac{S}{\sqrt{S_{xx}}}$, where S^2 is the real dispersion estimator of σ^2 which is defined as $S^2 = \frac{1}{n-2} * \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{n-2}(\sum_{i=1}^{n}(y_i - \overline{y})^2 - \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2) = \frac{1}{n-2}(S_{yy} - \sum_{i=1}^{n}(\hat{\beta}_0 + \hat{\beta}_1 x_i - \overline{y})^2) = \frac{1}{n-2}(S_{yy} - \hat{\beta}_1{}^2 * \sum_{i=1}^{n}(x_i - \overline{x})^2) = \frac{1}{n-2}(S_{yy} - (\frac{S_{xy}}{S_{xx}})^2 * S_{xx}) = \frac{1}{n-2}(S_{yy} - \frac{S_{xy}^2}{S_{xx}})$

Therefore, $\frac{\hat{\beta}_1}{se(\hat{\beta}_1)} = \frac{\sqrt{n-2} * S_{xy}}{\sqrt{Sxx}} * \frac{\sqrt{S_{xx}}}{\sqrt{S_{yy}S_{xx} - S_{xy}^2}} = \frac{\sqrt{n-2} * S_{xy}}{\sqrt{S_{xx} * S_{yy}}} * \frac{1}{\sqrt{1 - \frac{S_{xy}^2}{S_{xx} * S_{yy}}}} = \frac{r_{xy}\sqrt{n-2}}{\sqrt{1 - r_{xy}^2}}$

Part B

The data from the cholesterol.txt file refer to total cholesterol levels (mg / ml) of 24 patients (y) and their age (x). (i) It is required to construct a scatter plot between the two variables y and x and adapt the model $E(y) = \beta_0 + \beta_1 x$ to the data.

Initially we read our data through the commands:

```
adat <- read.table('./chol.csv', header = TRUE)
attach(adat)
adat
```

and print our data:

	y	x
1	3.5	46
2	4.5	57
3	2.1	22
4	2.5	22
5	2.3	28
6	3.3	29
7	1.9	20
8	3.0	25
9	3.8	43
10	4.6	63
11	4.0	49
12	3.2	34
13	4.0	52
14	2.9	28
15	4.1	57
16	3.2	40
17	4.3	52
18	2.5	24
19	2.6	30
20	3.8	36
21	3.0	33
22	4.2	48
23	3.9	58
24	3.3	50

Figure 1: The data of the chol.csv file

Next we construct a scatter plot and apply the regression function of the form $E(y) = \beta_0 + \beta_1 x$

```
plot(y~x, pch=19)
abline(lm(y~x))
model <- lm(y~x)
summary(model)
```

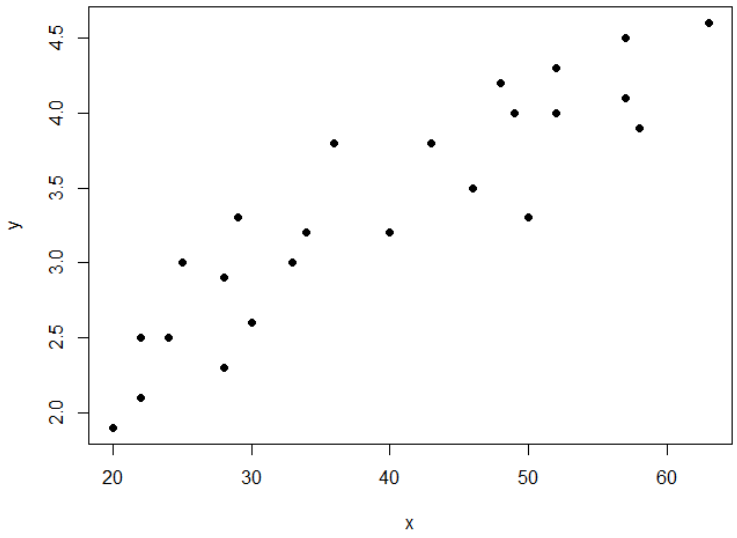

Figure 2: The scatter plot of the variables x and y

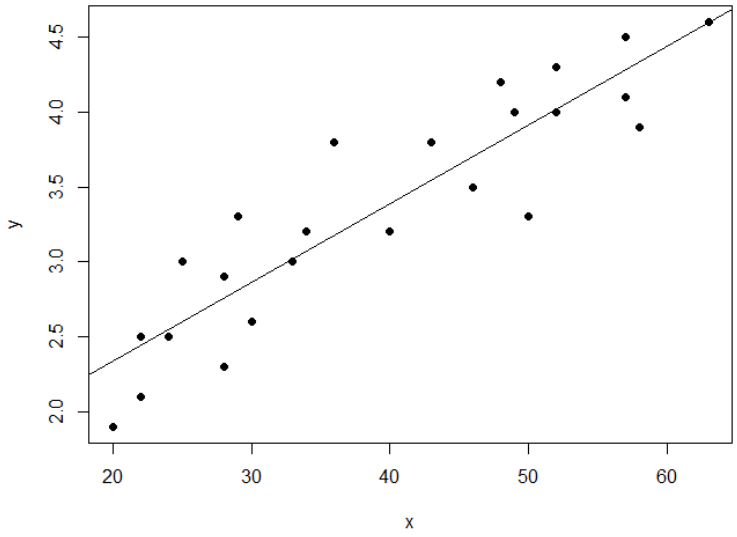

Figure 3: The regression function of the model

(ii) It is requested to check $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$ and in addition to specify a 0.95 confidence interval for the x factor in the adapted model.

```
Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-0.6111 -0.2151 -0.0058  0.2297  0.6256

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.279868   0.215699   5.934 5.69e-06 ***
x           0.052625   0.005192  10.136 9.43e-10 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.334 on 22 degrees of freedom
Multiple R-squared:  0.8236,    Adjusted R-squared:  0.8156
F-statistic: 102.7 on 1 and 22 DF,  p-value: 9.428e-10
```

Figure 4: Model estimations

As we see for the model $y_i = \hat{\beta}_0 + \hat{\beta}_1 x$ the estimates of the variables $\hat{\beta}_0 = 1.279868$ and $\hat{\beta}_1 = 0.052625$ have been made. We observe that the standard error of β_0 is equal to 0.215699 while that of β_1 is equal to 0.005192. We also see that we have 22 degrees of freedom which result from the total size of the observations subtracting the two variables estimated $\hat{\beta}_0$ and $\hat{\beta}_1$. We observe that t-value is quite far from the zero value (t-value = 10.136) and also differs greatly from the standard error (Std Error = 0.215699). At the same time, Pr (> | t |) equals something very small (= 9.43e-10), so we can rule out the null hypothesis H_0.

We specify a confidence interval of 0.95 through the command:

```
confint(model, level=0.95)
```

```
                2.5 %     97.5 %
(Intercept) 0.83253668 1.72720003
x           0.04185806 0.06339175
```

Figure 5: Application of 0.95 confidence interval

From the 0.95 confidence interval applied we see that it is 97.5 % probable β_0 and β_1 belong to the intervals (0.83253668, 1.72720003) and (0.04185806, 0.06339175) respectively.

Regression analysis is concerned with determining how changes in independent variables are related to changes in the dependent variable. B_1 refers to the slope of the regression function. If the X value increases by one unit, the cholesterol level is expected to increase by about 1.72 units.

(iii) Construct a 0.99-d.e. forecast for the cholesterol level y of a patient aged 35 years, as well as for its expected value, E (y).

```
predict(model, newdata=list(x=35), interval="prediction", level=.99)

predict(model, newdata=list(x=35), interval="confidence", level=.99)
```

```
       fit      lwr      upr
1 3.12174 2.918965 3.324515
```

Figure 6: Application of 0.99 confidence interval and forecast for age x = 35

```
       fit      lwr      upr
1 3.12174 2.158578 4.084902
```

Figure 7: Application 0.99 and forecast for age x = 35 for the expected value E (y)

(iv) Check the Normal distribution and the graph ei with $\hat{y}_i$, for the remaining ei.

```
qqnorm(model$residuals, pch=19)
qqline(model$residuals)
plot(model, which=1, pch = 19)
```

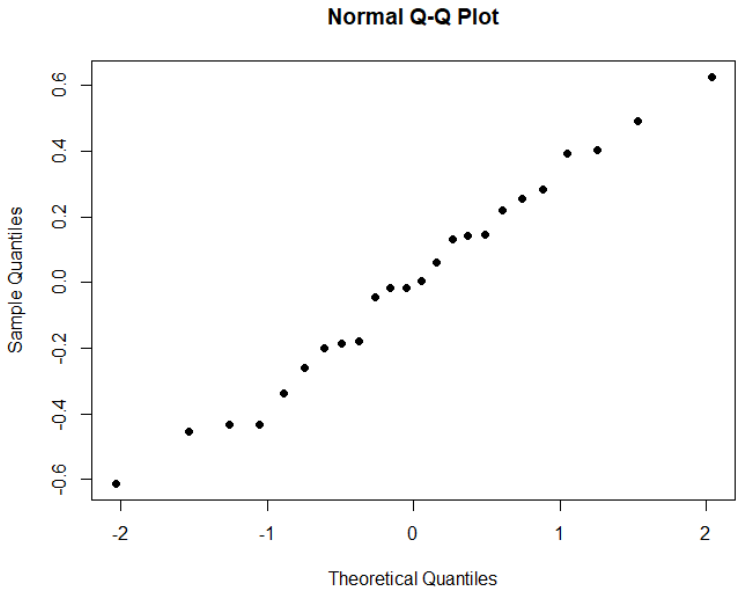

Figure 8: The graphical control test of the Normal distribution

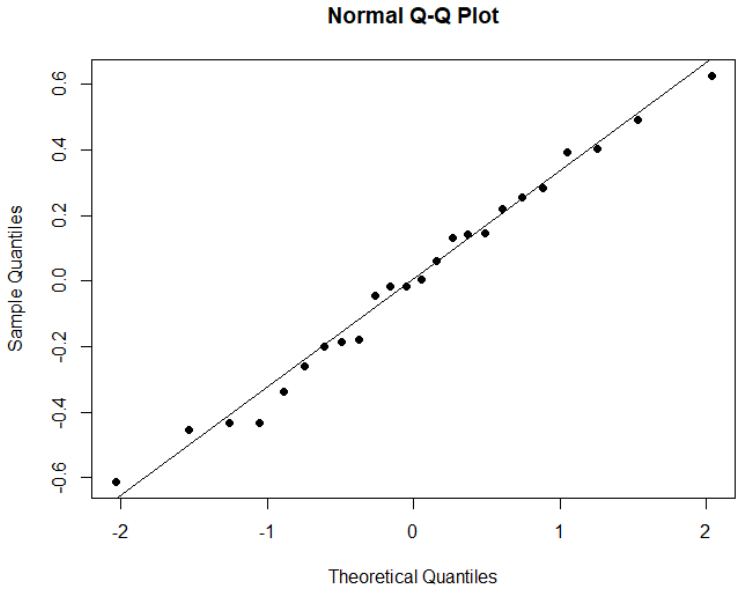

Figure 9: The graphical control test of the Normal distribution line

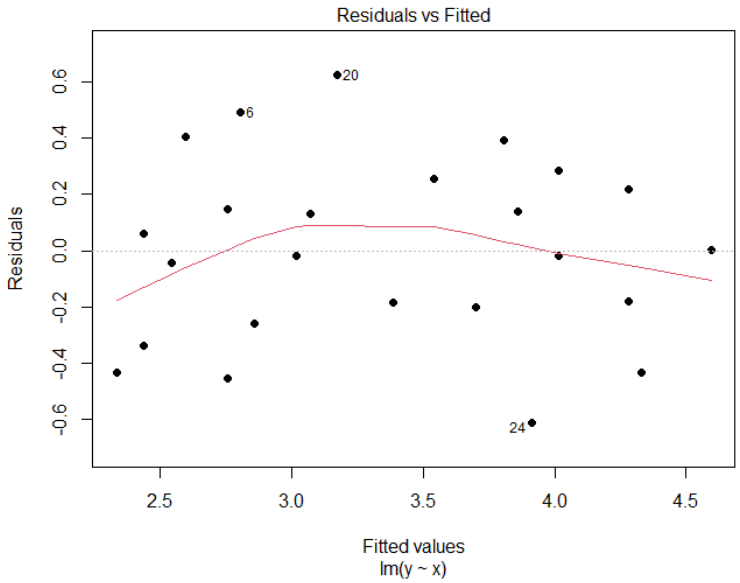

Figure 10: The graph ei with $\hat{y}_i$

As we can see from the normal distribution control diagram **??**, our data largely follows the normal distribution format. At the same time from the figure **??**. We conclude that the residuals do not follow any form or distribution, the element is absolutely desirable in our model, as it indicates the randomness of the errors of the model we are approaching.

Part C

Construct a scatter plot between the two variables y and x and adapt a model of the form $y = 3 - \alpha e^{\beta x}$. We observe that our model is non-linear therefore, we reduce it to linear through the metamarking $z = ln(3 - y) = c + \beta x$ where, $c = ln(\alpha)$.

x	2	4	6	12	18	24
y	1.07	1.88	2.26	2.78	2.97	2.99

Apply for the model $E(z) = c + \beta x$ the simple linear model and construct the scattering function.

```
#make a transformation
z = log(3-y)
plot(z~x, pch=19)
abline(lm(z~x))
model <- lm(z~x)
summary(model)
```

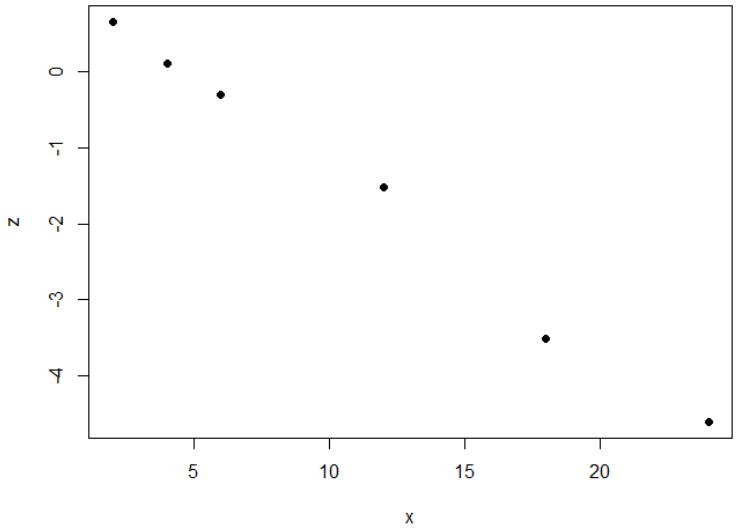

Figure 11: The scatter plot of the variables x and ln (3-y)

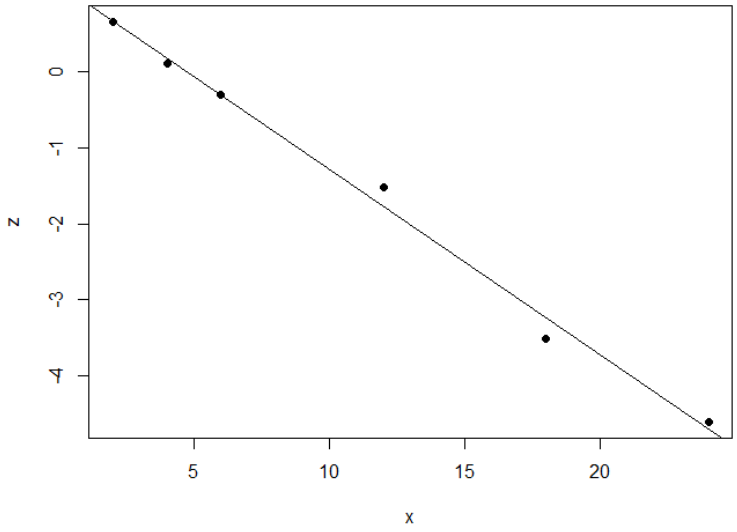

Figure 12: The regression function of the variables x and z = ln (3-y)

```
Call:
lm(formula = z ~ x)

Residuals:
       1        2        3        4        5        6
-0.009494 -0.066345  0.006562  0.255561 -0.274847  0.088563

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.15435    0.13703   8.424  0.00109 **
x           -0.24367    0.01012 -24.077 1.77e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1957 on 4 degrees of freedom
Multiple R-squared:  0.9931,    Adjusted R-squared:  0.9914
F-statistic: 579.7 on 1 and 4 DF,  p-value: 1.765e-05
```

Figure 13: The estimates of the transformed model

We observe that the estimator of c is equal to 1.15435, therefore $ln(\alpha) = e^{1.15435} => \alpha =$ 3.171961. The estimator of β is equal to -0.24367 with a standard error of c equal to 0.13703 and of β with 0.01012 respectively.

(ii) We construct a 0.95 confidence interval through the command:

```
confint(model, level=0.95)
```

```
                2.5 %    97.5 %
(Intercept)  0.7738880  1.534822
x           -0.2717697 -0.215571
```

Figure 14: Application of 0.95 confidence interval

We construct a 0.95 confidence interval to predict the observation y, as well as an approximate 95 % d.e. for its mean value, E (y), when x = 9 through the commands:

```
predict(model, newdata = list(x=9), interval = "prediction", level = 0.95
predict(model, newdata = list(x=9), interval = "confidence", level = 0.95
```

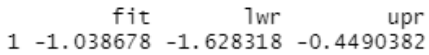

Figure 15: Apply 0.95 confidence interval and predict for x = 9

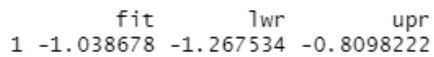

Figure 16: Application 0.95 and forecast for x = 9 for the expected value E (z)

The confidence interval and the forecast for x = 9 and also for the expected value were applied for the transformed model. Therefore, the predictions of the shapes 15 and 16 refer to approaches of the original model.