**STAT 40001/STAT 50001 Statistical Computing Fall 2024**

**Lab -15**

**Q.N. 1)** The more beer you drink, the more your blood alcohol level (BAL) rises. Table below contains a data set on beer consumption.

Beers 5 2 9 8 3 7 3 5 3 5

BAL 0.10 0.03 0.19 0.12 0.04 0.095 0.07 0.06 0.02 0.05

1. Make a scatterplot with a regression line
2. Calculate 95% confidence interval for the model parameters
3. State the estimated linear regression model.

> # Q1

> beers <- scan()

1: 5 2 9 8 3 7 3 5 3 5

11:

Read 10 items

> bal <- scan()

1: 0.10 0.03 0.19 0.12 0.04 0.095 0.07 0.06 0.02 0.05

11:

Read 10 items

> beers; bal

[1] 5 2 9 8 3 7 3 5 3 5

[1] 0.100 0.030 0.190 0.120 0.040 0.095 0.070 0.060 0.020 0.050

> plot(beers, bal, pch = 17, col = 2, main = "Beers VS BAL")

> model1 <- lm(bal~beers)

> model1

Call:

lm(formula = bal ~ beers)

Coefficients:

(Intercept) beers

-0.0185 0.0192

> abline(model1, lwd = 2, col = "blue")

> cat("Fitter Model: BAL = -0.0185 + 0.0912 \* Beers")

Fitter Model: BAL = -0.0185 + 0.0912 \* Beers

> confint(model1, level = 0.9)

5 % 90 %

(Intercept) -0.05425885 0.01725885

beers 0.01267136 0.02572864

> summary(model1)

Call:

lm(formula = bal ~ beers)

Residuals:

Min 1Q Median 3Q Max

-0.0275 -0.0187 -0.0071 0.0194 0.0357

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -0.018500 0.019230 -0.962 0.364200

beers 0.019200 0.003511 5.469 0.000595 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

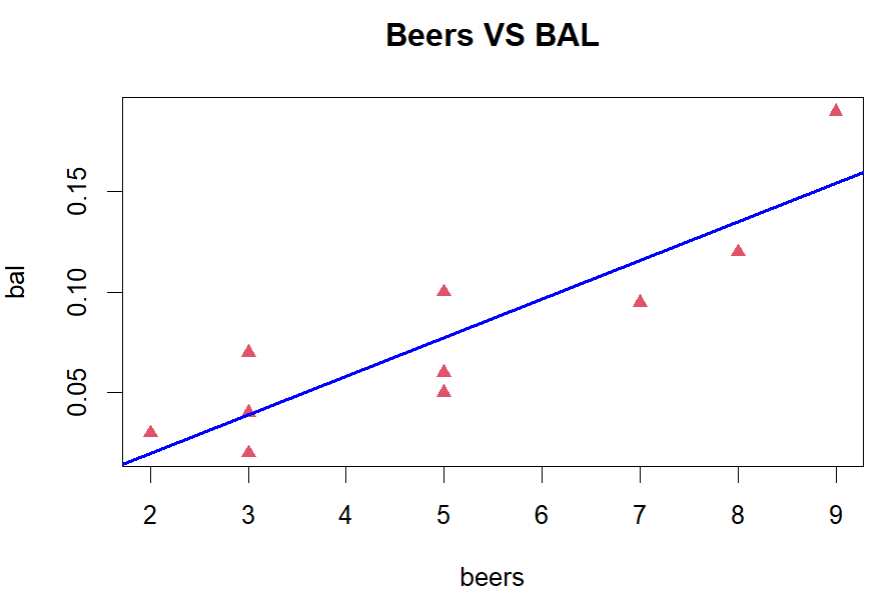
Residual standard error: 0.02483 on 8 degrees of freedom

Multiple R-squared: 0.789, Adjusted R-squared: 0.7626

F-statistic: 29.91 on 1 and 8 DF, p-value: 0.0005953

> cat("Beta < alpha -> reject the null hypothesis, we conclude that BAL value is dependent on the beers variable")

Beta < alpha -> reject the null hypothesis, we conclude that BAL value is dependent on the beers variable



**Q.N. 2)** A marketing researcher studied annual sales of a product that had been introduced 10 years ago. The data are as follows, where x is the year coded and y is the sales in thousands of units:

x: 0 1 2 3 4 5 6 7 8 9

y: 98 135 162 178 221 232 283 300 374 395

a) Prepare a scatter plot of the data

c) State the estimated regression line for the data and add it to the scatter plot.

d) Use the model to predict the sales in the 10th year (i.e. For x=10). Also provide the 95% and 90% confidence interval for the predicted value.

> #Q2

> x <- scan()

1: 0 1 2 3 4 5 6 7 8 9

11:

Read 10 items

> y <- scan()

1: 98 135 162 178 221 232 283 300 374 395

11:

Read 10 items

> x;y

[1] 0 1 2 3 4 5 6 7 8 9

[1] 98 135 162 178 221 232 283 300 374 395

> plot(x,y,pch = 17, col = 2, main = "X vs Y")

> model2 <- lm(y~x)

> model2

Call:

lm(formula = y ~ x)

Coefficients:

(Intercept) x

91.56 32.50

> cat("Fitted Model: y = 91.56 + 32.50 \* x")

Fitted Model: y = 91.56 + 32.50 \* x

> abline(model2, lwd = 2, col = "green")

> # ALSO: abline(91.56, 32.50 , lwd = 2, col = "green")

> predict(model2, data.frame(x = 10))

1

416.5333

> cat("For x = 10, y -> 416.5333")

For x = 10, y -> 416.5333

> predict(model2, data.frame(x = 10),interval = "conf", level = 0.95) # 95%

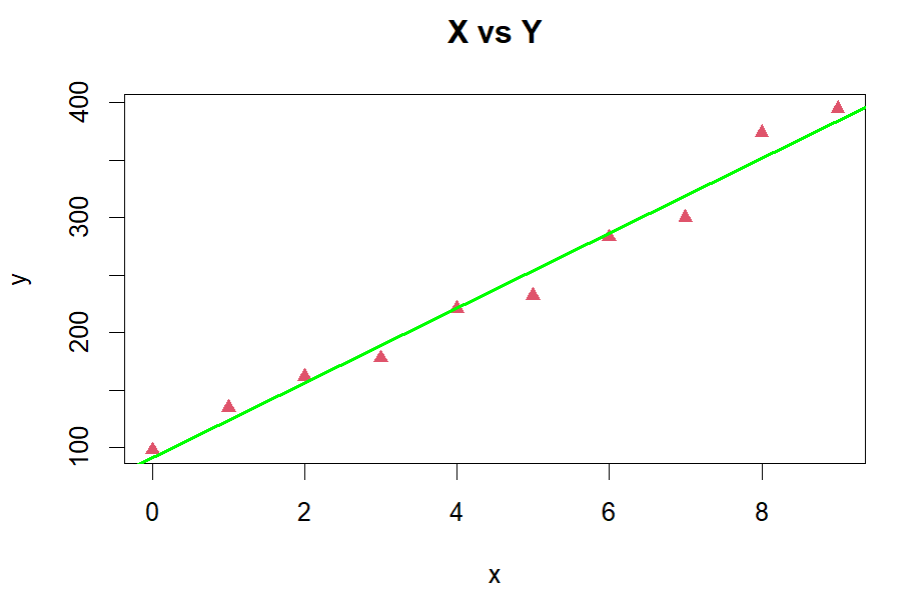
fit lwr upr

1 416.5333 392.9089 440.1578

> predict(model2, data.frame(x = 10),interval = "conf", level = 0.9) # 90%

fit lwr upr

1 416.5333 397.4827 435.5839



**Q.N. 3)** A simple random sample of apparently healthy children between the ages of 6 months and 15 years yielded the following data on age, X, and liver volume per unit of body weight (ml/kg), Y

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| X | 0.5 | 0.7 | 2.5 | 4.1 | 5.9 | 6.1 | 7 | 8.2 | 10 | 10.1 | 10.9 | 11.5 | 12.1 | 14.1 | 15 |
| Y | 41 | 55 | 41 | 39 | 50 | 32 | 41 | 42 | 26 | 35 | 25 | 31 | 31 | 29 | 23 |

a) Prepare a scatter plot of the data

c) State the estimated regression line for the data and add it to the scatter plot.

d) Use the model to predict the liver volume of 8 years old child.

e) Construct a 90% Confidence interval for the predicted value of the liver volume of 8 years old child

f) Construct a 90% prediction interval for the predicted value of the liver volume of 8 years old child

> #Q3

> X <- scan()

1: 0.5 0.7 2.5 4.1 5.9 6.1 7 8.2 10 10.1 10.9 11.5 12.1 14.1 15

16:

Read 15 items

> Y <- scan()

1: 41 55 41 39 50 32 41 42 26 35 25 31 31 29 23

16:

Read 15 items

> X;Y

[1] 0.5 0.7 2.5 4.1 5.9 6.1 7.0 8.2 10.0 10.1 10.9 11.5 12.1 14.1 15.0

[1] 41 55 41 39 50 32 41 42 26 35 25 31 31 29 23

> plot(X,Y,main = "X against Y", pch = 17, col = 2)

> model3 <- lm(Y~X)

> model3

Call:

lm(formula = Y ~ X)

Coefficients:

(Intercept) X

48.540 -1.576

> abline(model3, lwd = 2, col = "green")

> cat("Linear Fitted Line: Y = 48.540 - 1.576\*X")

Linear Fitted Line: Y = 48.540 - 1.576\*X

> predict(model3, data.frame(X=8))

1

35.93006

> cat("For X=8 -> Y=35.93006")

For X=8 -> Y=35.93006

> predict(model3, data.frame(X=8),interval = "conf",level = 0.95)

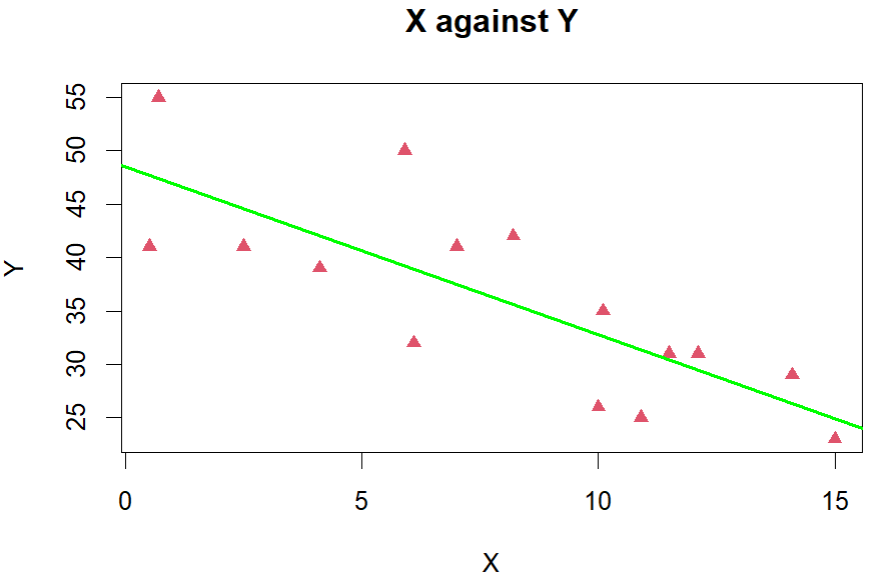
fit lwr upr

1 35.93006 32.65689 39.20324

> predict(model3, data.frame(X=8),interval = "conf",level = 0.90)

fit lwr upr

1 35.93006 33.24692 38.61321



**Q.N. 4)** The gross domestic product (GDP) of the United States in trillions of dollars from 1950- 2013 are provided in the link below

<http://media.pearsoncmg.com/aw/aw_sharpe_business_3/datasets/txt/GDP_2013.txt>

a) Display the data using a scatterplot.

b) Fit a simple linear regression model.

c) Add the fitted line to the scatter plot.

d) Determine the coefficient of determination.

d) Analyze the residual plots. Is your model questionable?

e) Use Box-Cox Transformation to see whether this model can be improved

> #Q4

> Q4 <- read.table("http://media.pearsoncmg.com/aw/aw\_sharpe\_business\_3/datasets/txt/GDP\_2013.txt",sep = "\t", header = T)

> head(Q4,5)

Year GDP...T.

1 2013 13.75

2 2012 13.67

3 2011 13.44

4 2010 13.18

5 2009 12.87

> dim(Q4)

[1] 64 2

> names(Q4) = c("Year","GDP")

> head(Q4,5)

Year GDP

1 2013 13.75

2 2012 13.67

3 2011 13.44

4 2010 13.18

5 2009 12.87

> attach(Q4)

> plot(Q4,pch = 17, col = 2)

> model4 <- lm(GDP~Year)

> model4

Call:

lm(formula = GDP ~ Year)

Coefficients:

(Intercept) Year

-387.8433 0.1993

> cat("Fitted Model: GDP = -387.8433 + 0.1993\*Year")

Fitted Model: GDP = -387.8433 + 0.1993\*Year

> abline(model4, lwd = 2, col = "green")

> summary(model4)

Call:

lm(formula = GDP ~ Year)

Residuals:

Min 1Q Median 3Q Max

-1.23604 -0.63427 -0.07458 0.51860 1.35054

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -3.878e+02 9.217e+00 -42.08 <2e-16 \*\*\*

Year 1.993e-01 4.651e-03 42.84 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6874 on 62 degrees of freedom

Multiple R-squared: 0.9673, Adjusted R-squared: 0.9668

F-statistic: 1835 on 1 and 62 DF, p-value: < 2.2e-16

> cat("Multiple R-squared: 96.73%")

Multiple R-squared: 96.73%

> plot(model4,1)

> cat("need transformation -> as it is expected to be scattered")

need transformation -> as it is expected to be scattered

> install.packages("MASS")

> library(MASS)

> b = boxcox(model4)

> y1 = GDP^(0.25)

> model\_new = lm(y1~Year)

> model\_new

Call:

lm(formula = y1 ~ Year)

Coefficients:

(Intercept) Year

-22.66422 0.01224

> cat("New Fitted Model: y1^0.25 = -22.66422 + 0.01224 \* Year")

New Fitted Model: y1^0.25 = -22.66422 + 0.01224 \* Year

> summary(model\_new)

Call:

lm(formula = y1 ~ Year)

Residuals:

Min 1Q Median 3Q Max

-0.039953 -0.010020 -0.000066 0.010955 0.028132

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -2.266e+01 2.015e-01 -112.5 <2e-16 \*\*\*

Year 1.224e-02 1.017e-04 120.3 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.01503 on 62 degrees of freedom

Multiple R-squared: 0.9957, Adjusted R-squared: 0.9957

F-statistic: 1.448e+04 on 1 and 62 DF, p-value: < 2.2e-16

> cat("Increased R square value -> Good!!")

Increased R square value -> Good!!

> plot(model\_new,1)

> cat("Residual Plot is way better than the previous one!!")

Residual Plot is way better than the previous one!!

