**STAT 40001/MA59800 Statistical Computing Fall 2017**

**Lab -19**

1. An article in the Journal of the Environmental Engineering Division [Least Squares Estimates of BOD Parameters (1980, Vol. 106, pp. 1197–1202)] took a sample from the Holston River below Kingport, Tennessee, during August 1977. The biochemical oxygen demand (BOD) test is conducted over a period of time in days. The resulting data follow:

Time (days): 1 2 4 6 8 10 12 14 16 18 20

BOD(mg/liter): 0.6 0.7 1.5 1.9 2.1 2.6 2.9 3.7 3.5 3.7 3.8

1. Assuming that a linear regression model is appropriate, fit the regression model relating BOD (y) to the time (x).

> Time = scan()

1: 1 2 4 6 8 10 12 14 16 18 20

12:

Read 11 items

> Time

[1] 1 2 4 6 8 10 12 14 16 18 20

> BOD = scan()

1: 0.6 0.7 1.5 1.9 2.1 2.6 2.9 3.7 3.5 3.7 3.8

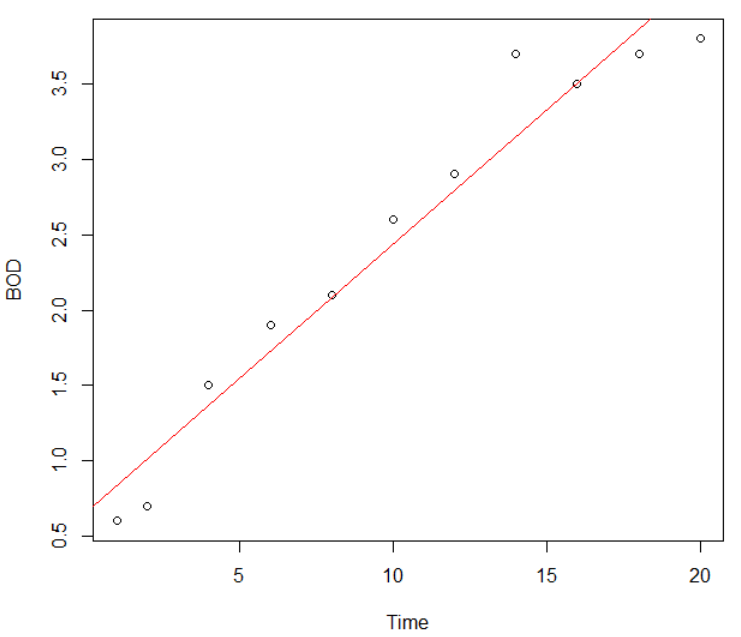
12:

Read 11 items

> plot(Time,BOD)

> model = lm(BOD~Time)

> abline(model,col=2)



1. What is the estimated value of?

> summary(model)

Call:

lm(formula = BOD ~ Time)

Residuals:

Min 1Q Median 3Q Max

-0.41892 -0.19933 0.01775 0.14581 0.54942

Coefficients:

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

(Intercept) 0.6578 0.1657 3.969 0.00326 \*\*

Time 0.1781 0.0140 12.716 4.69e-07 \*\*\*

---

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

Residual standard error: 0.2873 on 9 degrees of freedom

Multiple R-squared: 0.9473, Adjusted R-squared: 0.9414

F-statistic: 161.7 on 1 and 9 DF, p-value: 4.694e-07

> model

Call:

lm(formula = BOD ~ Time)

Coefficients:

(Intercept) Time

0.6578 0.1781

> 0.2873^2

[1] 0.08254129

σ^2 = 0.2873^2 = 0.08254129

1. What is the estimate of expected BOD level when the time is 15 days? Provide a 90% confidence interval and prediction interval

> predict(model,data.frame(Time=15),interval = "conf",level = 0.9)

fit lwr upr

1 3.328639 3.125933 3.531346

> predict(model,data.frame(Time=15),interval = "pred",level = 0.9)

fit lwr upr

1 3.328639 2.764355 3.892924

1. What change in mean BOD is expected when the time changes by three days?  
   0.6578\*3 = 1.9734
2. Air Pollution is currently one of the most serious public health worries worldwide. Many epidemiological studies have proved that some chemical compounds such as sulphur dioxide (SO2), nitrogen dioxide (NO2), ozone (O3) or other air-borne dust particles can have on our health. Link below contains 112 observations recorded during summer 2001 in Rennes (France).

Measurements for many variables are recorded

<http://www.agrocampus-ouest.fr/igagrocampus-ouest.fr/math/RforStat/ozone.txt>

1. Import the dataset in R.

> data = read.table('http://www.agrocampus-ouest.fr/igagrocampus-ouest.fr/math/RforStat/ozone.txt')

> head(data)

maxO3 T9 T12 T15 Ne9 Ne12 Ne15 Wx9 Wx12 Wx15 maxO3y wind rain

20010601 87 15.6 18.5 18.4 4 4 8 0.6946 -1.7101 -0.6946 84 North Dry

20010602 82 17.0 18.4 17.7 5 5 7 -4.3301 -4.0000 -3.0000 87 North Dry

20010603 92 15.3 17.6 19.5 2 5 4 2.9544 1.8794 0.5209 82 East Dry

20010604 114 16.2 19.7 22.5 1 1 0 0.9848 0.3473 -0.1736 92 North Dry

20010605 94 17.4 20.5 20.4 8 8 7 -0.5000 -2.9544 -4.3301 114 West Dry

20010606 80 17.7 19.8 18.3 6 6 7 -5.6382 -5.0000 -6.0000 94 West Rainy

1. Generate the list of variables included in the data using R.

> names(data)

[1] "maxO3" "T9" "T12" "T15" "Ne9" "Ne12" "Ne15" "Wx9" "Wx12" "Wx15" "maxO3y"

[12] "wind" "rain"

1. Create the subset of the data containing only the first 11 variables.

> data1 = data[,-c(12,13)]

1. Fit a multiple linear regression model for maxO3 as a response variable and all the reaming 10 variables as regressor variables. Use summary option to identify which of these variables appear to be significant?

> model = lm(maxO3~T9+T12+T15+Ne9+Ne12+Ne15+Wx9+Wx15+Wx15+maxO3y)

> summary(model)

Call:

lm(formula = maxO3 ~ T9 + T12 + T15 + Ne9 + Ne12 + Ne15 + Wx9 +

Wx15 + Wx15 + maxO3y)

Residuals:

Min 1Q Median 3Q Max

-53.576 -8.740 -0.390 7.578 39.468

Coefficients:

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

(Intercept) 12.30831 13.23214 0.930 0.354

T9 -0.01167 1.09196 -0.011 0.991

T12 2.21258 1.39642 1.584 0.116

T15 0.55990 1.13808 0.492 0.624

Ne9 -2.18837 0.93331 -2.345 0.021 \*

Ne12 -0.42910 1.33344 -0.322 0.748

Ne15 0.18345 0.99782 0.184 0.854

Wx9 0.95990 0.81317 1.180 0.241

Wx15 0.43591 0.70031 0.622 0.535

maxO3y 0.35183 0.06238 5.640 1.53e-07 \*\*\*

---

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

Residual standard error: 14.29 on 102 degrees of freedom

Multiple R-squared: 0.7638, Adjusted R-squared: 0.743

F-statistic: 36.66 on 9 and 102 DF, p-value: < 2.2e-16

maxO3y is the most significant variable.