

Solutions:

1a)

> cor.test(GRO2.data,GRO3.data)

Pearson's product-moment correlation

data: GRO2.data and GRO3.data

t = 7.9074, df = 36, p-value = 2.201e-09

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

0.6399101 0.8897262

sample estimates:

cor

0.7966283

The Correlation between expression values of the 2 genes is 0.7966283

1b)

> cor.test(GRO2.data, GRO3.data, conf.level = 0.90)

Pearson's product-moment correlation

data: GRO2.data and GRO3.data

t = 7.9074, df = 36, p-value = 2.201e-09

alternative hypothesis: true correlation is not equal to 0

90 percent confidence interval:

0.6702984 0.8780861

sample estimates:

cor

0.7966283

the parametric 90% confident interval for the correlation with cor.test() is 0.6702984 0.8780861

1c)

> quantile(boot.cor[,1],c(0.05,0.95))

5% 95%

0.5971801 0.8928861

the bootstrap 90% confident interval for the correlation is 0.5971801 0.8928861

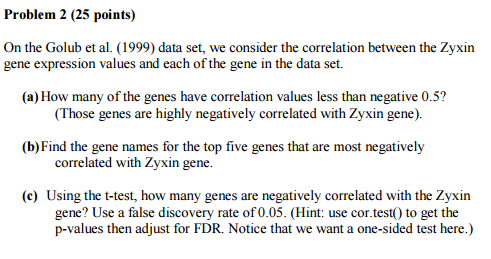
1d)

> quantile(boot.cor[,1],c(0.025,0.975))

2.5% 97.5%

0.5273127 0.9120214

As 0.64 comes within the CI (alpha = 0.05), we accept the null hypothesis that correlation = 0.64



Solutions:

2a)

[1] "no of genes that have correlation values less than negative 0.5"

> print(sum(cor.data < -0.5))

[1] 85

The no of genes have correlation values less that negative 0.5 is 85

2b) The gene names for the top five genes that are most negatively correlated with Zyxin gene:

[1] "gene names for the top five genes that are most negatively correlated with Zyxin gene."

> golub.gnames[order.cor[1:5],2]

[1] "Macmarcks"

[2] "Inducible protein mRNA"

[3] "C-myb gene extracted from Human (c-myb) gene, complete primary cds, and five complete alternatively spliced cds"

[4] "Oncoprotein 18 (Op18) gene"

[5] "54 kDa protein mRNA"

2c)

[1] "genes are negatively correlated with the Zyxin gene."

> sum(cor.ttest < 0.05)

[1] 572

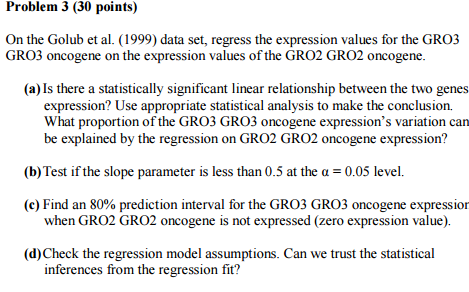
Using t-test, the no of genes that are negatively correlated with the Zyxin gene are 572

[1] "After FDR adjustment"

> sum(cor.fdr < 0.05)

[1] 142

After FDR adjustment, no of genes that are negatively correlated with the Zyxin gene are 142



Solutions:

3a)

> reg.fit <- lm(GRO3.data ~ GRO2.data)

> reg.fit

Call:

lm(formula = GRO3.data ~ GRO2.data)

Coefficients:

(Intercept) GRO2.data

-0.8426 0.3582

> summary(reg.fit)

Call:

lm(formula = GRO3.data ~ GRO2.data)

Residuals:

Min 1Q Median 3Q Max

-0.78038 -0.10639 -0.00553 0.14225 0.96298

Coefficients:

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

(Intercept) -0.84256 0.05941 -14.182 2.62e-16 \*\*\*

GRO2.data 0.35820 0.04530 7.907 2.20e-09 \*\*\*

---

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

Residual standard error: 0.3201 on 36 degrees of freedom

Multiple R-squared: 0.6346, Adjusted R-squared: 0.6245

F-statistic: 62.53 on 1 and 36 DF, p-value: 2.201e-09

We conclude that there is a statistically significant relationship between the two gene expression, as both the p-values are < 0.05 and we reject the Null Hypothesis that beta = 0

The proportion of the GRO3 GRO3 oncogene expression’s variation can be explained by the regression on GRO2 GRO2 oncogene expression through Multiple R-squared = 0.6346

3b)

> confint(reg.fit, level = 0.95)

2.5 % 97.5 %

(Intercept) -0.9630448 -0.7220732

GRO2.data 0.2663291 0.4500727

The interval for slope parameter is 0.2663291 0.4500727, so yes it is less than 0.5

3c)

> predict(reg.fit, newdata, interval="prediction", level = 0.80)

fit lwr upr

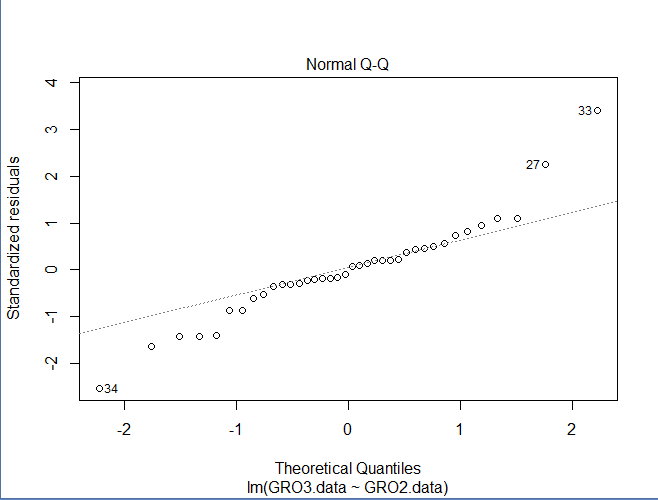
1 -0.842559 -1.267563 -0.4175553

The 80% prediction interval for the GRO3 GRO3 oncogene expression when GRO2 GRO2 oncogene is not expressed

lwr upr

-1.267563 -0.4175553

3d)



The Q-Q line seems to be normal, mainly for the central data

To confirm, we use Shapiro test:

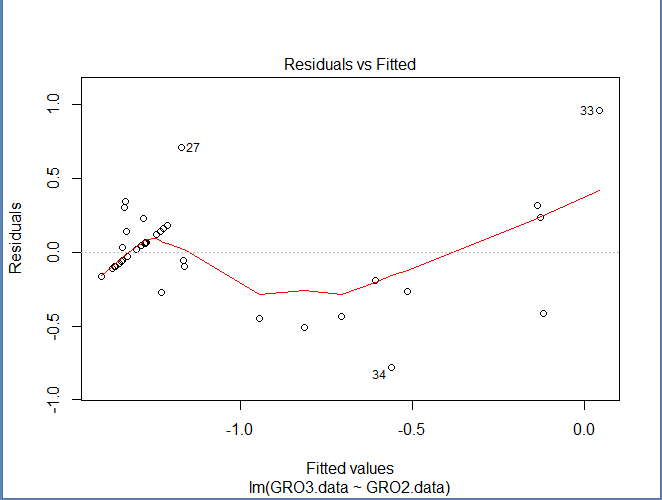
Shapiro-Wilk normality test

data: resid(reg.fit)

W = 0.9478, p-value = 0.07532

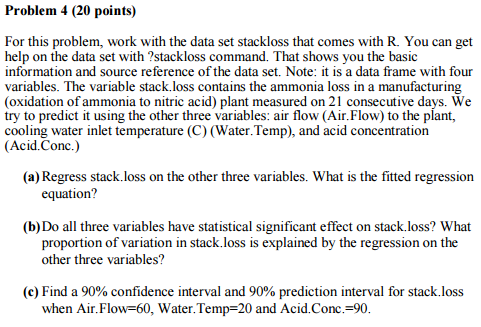
The p-value is 0.07532. As p > 0.5 we Accept the null Hypothesis

But we should not trust the statistical inferences from regression fit as all the assumptions are not true



Non-linear mean patterns are observed.

Variance seems to be different at different points. It is violating homoscedasticity.



Solutions:

4a)

Call:

lm(formula = stack.loss ~ Air.Flow + Water.Temp + Acid.Conc.,

data = stack.loss)

Residuals:

Min 1Q Median 3Q Max

-7.2377 -1.7117 -0.4551 2.3614 5.6978

Coefficients:

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

(Intercept) -39.9197 11.8960 -3.356 0.00375 \*\*

Air.Flow 0.7156 0.1349 5.307 5.8e-05 \*\*\*

Water.Temp 1.2953 0.3680 3.520 0.00263 \*\*

Acid.Conc. -0.1521 0.1563 -0.973 0.34405

---

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

Residual standard error: 3.243 on 17 degrees of freedom

Multiple R-squared: 0.9136, Adjusted R-squared: 0.8983

F-statistic: 59.9 on 3 and 17 DF, p-value: 3.016e-09

The fitted regression equation is :

-39.9197+ 0.715 Air.Flow + 1.295 Water.Temp – 0.15 Acid.Conc

4b)

As the p-value for Air.Flow(5.8e-05) and Water.Temp(0.00263) are both < 0.05, it has no statistical significant effect on stack.loss

As p-value for Acid.Conc(0.34405) is >0.05, it has no statistically significant effect on stack.loss

Proportion of stack.loss that can be explained by the regression on other three variables = R squared = 0.9136

4c)

90% CI = 13.50069 16.96617

> predict(lin.reg, given.data, interval="confidence", level = 0.90)

fit lwr upr

1 15.23343 13.50069 16.96617

90% prediction interval = 9.331184 21.13568

> predict(lin.reg, given.data, interval="prediction", level = 0.90)

fit lwr upr

1 15.23343 9.331184 21.13568