Charles Valentine
Homework 9
11/6/2016

Note:
All code is in script hw08.R --- I have made use of cat and print statements to display information easily!

Problem 1

a)

OUTPUT

```
Problem 1a
==========
Correlation of GRO2 GRO2 oncogene and GRO3 GRO3 oncogene expression values is:
0.7966
```

b)

OUTPUT

```
Problem 1b
==========
Parametric 90 percent conf. interval for the correlation of GRO2 GRO2 oncogene and GRO
3 GRO3 oncogene expression values:
( 0.6703, 0.8781 )
```

c)

OUTPUT

```
Problem 1c
==========
Nonparametric 90 percent conf. interval for the correlation of GRO2 GRO2 oncogene and
GRO3 GRO3
oncogene expression values:
( 0.5954 0.8959 )
```

d) We have found the probability that the null hypothesis is supported which is far less than 0.05. We shall then reject the null hypothesis and accept that the correlation between the two gene expression values are significantly greater than 0.64 (one-tailed).

OUTPUT

```
Problem 1d
==========
Ho: The correlation is 0.64 or less
HA: The correlation is greater than 0.64
P-value:  2e-04
```

Problem 2

a)

OUTPUT

```
Problem 2a
==========
There are 85 genes highly negatively correlated with Zyxin
(correlation < -0.5)
```

b)

OUTPUT

```
Problem 2b
==========
Top five genes that are most negatively correlated with Zyxin

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
```

c) The following p-value threshold condition was satisfied for all 142 genes that were negatively correlated with Zyxin.

OUTPUT

```
Problem 2c
==========
There are 142 genes significantly negatively correlated with Zyxin
(FDR adjusted p-value < 0.05)
```

Problem 3

a) The p-values for the linear fit coefficients (intercept and slope) are both far below the level of 0.05 so we can conclude that the fit coefficients are statistically significant. The R-squared value gives an indication of how much variation in “GRO2 GRO2 oncogene” expression’s variation can be explained by the “GRO3 GRO3 oncogene” expression’s variation.

OUTPUT

```
Problem 3a
==========
P-values for the linear model of the GRO2 GRO2 oncogene predicting the GRO3 GRO3
oncogene

         (Intercept)         golub[gene2.loc, ]
           3.400e-05                  2.201e-09

The proportion of the GRO2 GRO2 oncogene expression's variation can be explained by
the GRO3 GRO3 oncogene expression's variation is:
0.6346
```

b) We shall accept the null hypothesis that the slope parameter is 0.5 or greater since the p-value is far above the level of 0.05.

OUTPUT

```
Problem 3b
==========
Ho: The slope parameter is 0.5 or greater
HA: The slope parameter is less than 0.5
P-value:  1
```

c) .

OUTPUT

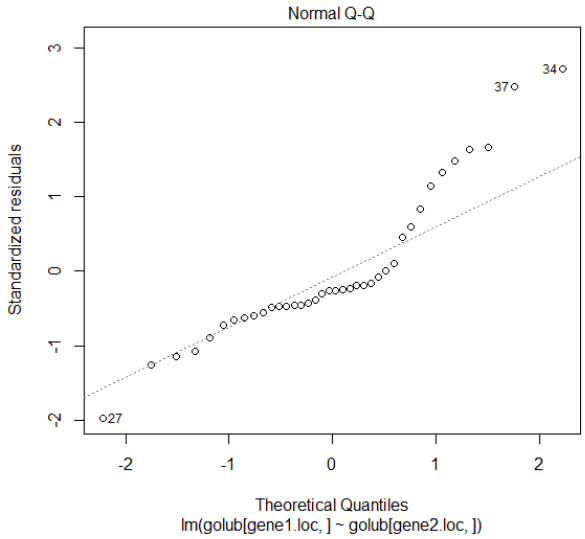
```
Problem 3c
==========
The 80 percent prediction interval for the GRO3 GRO3 oncogene expression when GRO2 GRO
2
oncogene is not expressed is:
( 0.2677, 2.2521 )
```

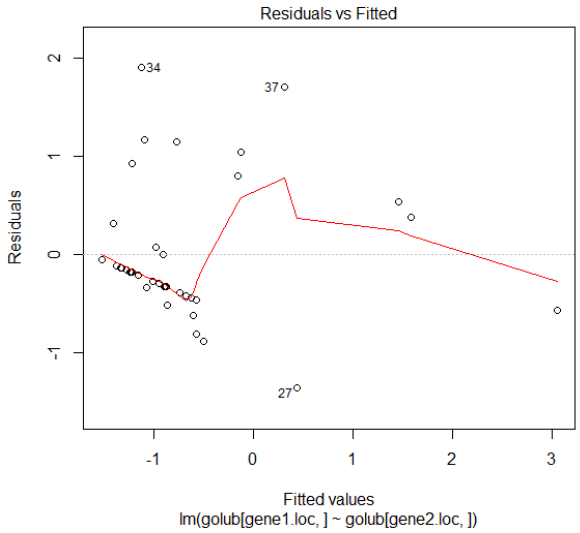
d) First we check the normality of the residuals of the linear fit. We do this with a Shapiro-Wilks test and find that our p-value is far less than 0.05 which indicates we should accept the alternate hypothesis that our data comes from a non-normal distribution. This is a violation of the linear fit model assumption. The two figures below test the model assumptions. The figure on the left, a QQ-plot also shows a non-linear relationship between the standardized residuals and the theoretical quantiles which is in congruence with our Shapiro-Wilks test that these data do not come from a normal distribution. Finally, the variance over the fitted values appears to change in the figure on the right which invalidates the assumption that we must have equal variance. Instead we do not homoscedasticity.

We should conclude that we cannot trust the linear fit on this dataset.

OUTPUT

```
Problem 3d
==========
Ho: The residuals of the linear fit follow a normal distribution
HA: The residuals of the linear fit do not follow a normal distribution
P-value:  0.0014611
```





Problem 4

a) The fitted regression equation can be formulated from the estimates from the multiple linear regression. The fitted regression equation is

$$\text{stack.loss} = -39.9197 + 1.2953 \cdot \text{water.temp} + 0.7156 \cdot \text{air.flow} - 0.1521 \cdot \text{acid.conc}$$

OUTPUT

```
Problem 4
==========

Call:
lm(formula = stack.loss ~ Air.Flow + Water.Temp + Acid.Conc.,
    data = stackloss)

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
```

b) Only air flow and water temperature have statistical significance at a level of 0.05. As determined by the p-value in the above summary.

c)

OUTPUT

```
The 90 percent confidence interval for stackloss
when Air Flow is 60, Water Temp is 20, and Acid Concentration is 90 is:
( 9.3312, 21.1357 )

The 90 percent prediction interval for stackloss
when Air Flow is 60, Water Temp is 20, and Acid Concentration is 90 is:
( 0.2677, 2.2521 )
```