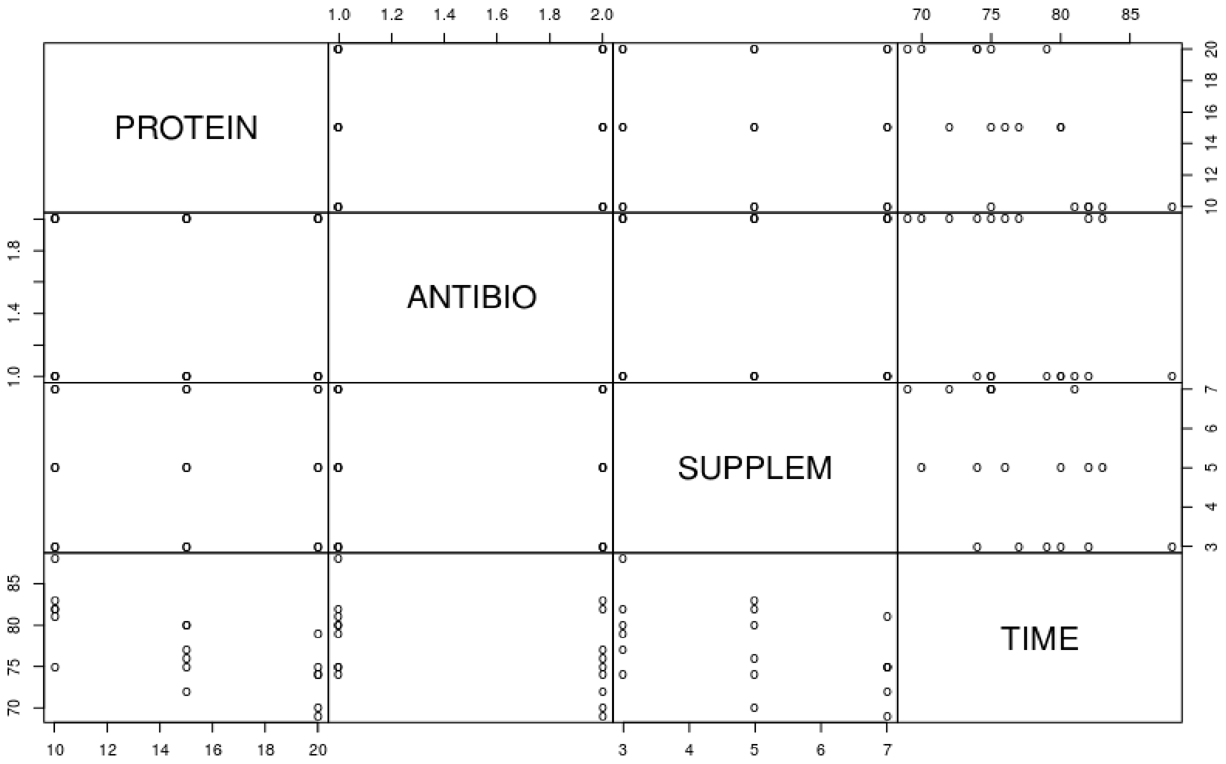
# STAT 512 – Assignment 1

Vignesh M. Pagadala
Vignesh.Pagadala@ColoState.Edu

Feb 11, 2019

**1. Show the pairwise scatterplots between all 4 variables (Y=Time, X1=Protein, X2=Antibio, X3=Supplem).**

**ANSWER:**



**2. Calculate pairwise (Pearson) correlations between all 4 variables.**

**ANSWER:**
The following table shows the Pearson correlation coefficient values calculated between all four variables.

| | PROTEIN | ANTIBIO | SUPPLEM | TIME |
|---|---|---|---|---|
| **PROTEIN** | 1.0000000 | 0.0000000 | 0.0000000 | -0.7111002 |
| **ANTIBIO** | 0.0000000 | 1.0000000 | 0.0000000 | -0.4180398 |
| **SUPPLEM** | 0.0000000 | 0.0000000 | 1.0000000 | -0.4693261 |
| **TIME** | -0.7111002 | -0.4180398 | -0.4693261 | 1.0000000 |

**3. Run the 3 simple linear regressions of Time vs each of the above three predictor variables. Show the parameter estimates ("Coefficients" table) and R 2 values. You can just copy/paste the relevant output from R. (6 pts)**

**ANSWER:**

A. Response of TIME with respect to PROTEIN

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  89.8333     3.2022  28.054 4.92e-15 ***
PROTEIN      -0.8333     0.2060  -4.046 0.000938 ***

Multiple R-squared:  0.5057
```

B. Response of TIME with respect to ANTIBIO

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   83.333      3.436  24.254  4.8e-14 ***
ANTIBIO       -4.000      2.173  -1.841   0.0843 .

Multiple R-squared:  0.1748
```

C. Response of TIME with respect to SUPPLEM

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  84.2083     3.4019  24.753 3.49e-14 ***
SUPPLEM      -1.3750     0.6468  -2.126   0.0494 *

Multiple R-squared:  0.2203
```

**4. Now run multiple regression of Time on all three predictor variables. Show the parameter estimates ("Coefficients" table) and R 2 value. We will use this the "full" model for the remaining questions.**

**ANSWER:**

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 102.7083     2.3104  44.455  < 2e-16 ***
PROTEIN      -0.8333     0.0987  -8.443 7.27e-07 ***
ANTIBIO      -4.0000     0.8059  -4.963 0.000208 ***
SUPPLEM      -1.3750     0.2467  -5.572 6.88e-05 ***

Multiple R-squared:  0.9007
```

**5. Note that (1) the slope estimates from the simple linear regressions are the same as the slope estimates from the "full" model and (2) the R 2 values from the simple linear regressions sum to the R 2 value from the "full" model. In general, this will not be the case (as we saw with the Rice Example). What is different about this data (as compared to the Rice Example)? Hint: Consider the result of question 2.**

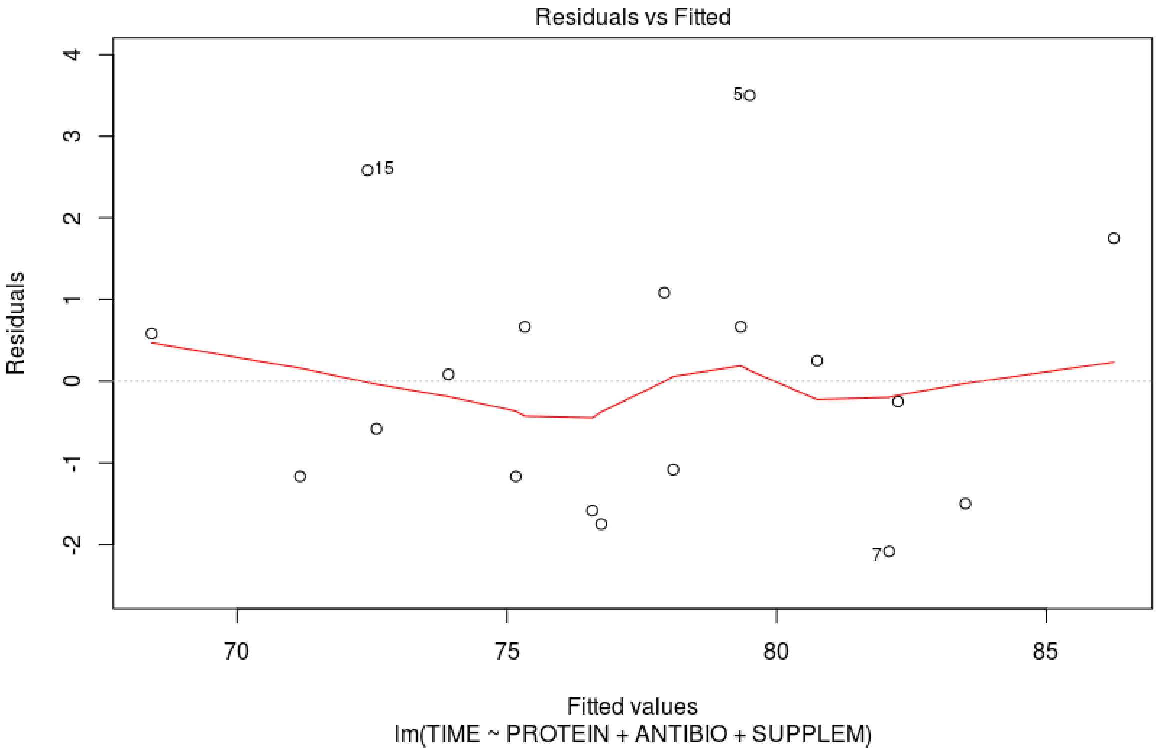
**ANSWER:**

In the rice example, the variables yeild, ht and tillers have some degree of correlation between them. However, in our current example, the predictor variables have absolutely no correlation between them, as observed in the pairwise correlation coefficient from the table in question 2. The $R^2$ (multiple regression) would be equal to the sum of the individual $r^2$ (linear regression) values only if there is no correlation amongst the predictor variables.

**6. Create plots of (A) Residuals vs Fitted values and (B) QQplot of residuals. Include these plots in your assignment. Thinking about model assumptions, what do we learn from each plot? (4 pts)**

**ANSWER:**

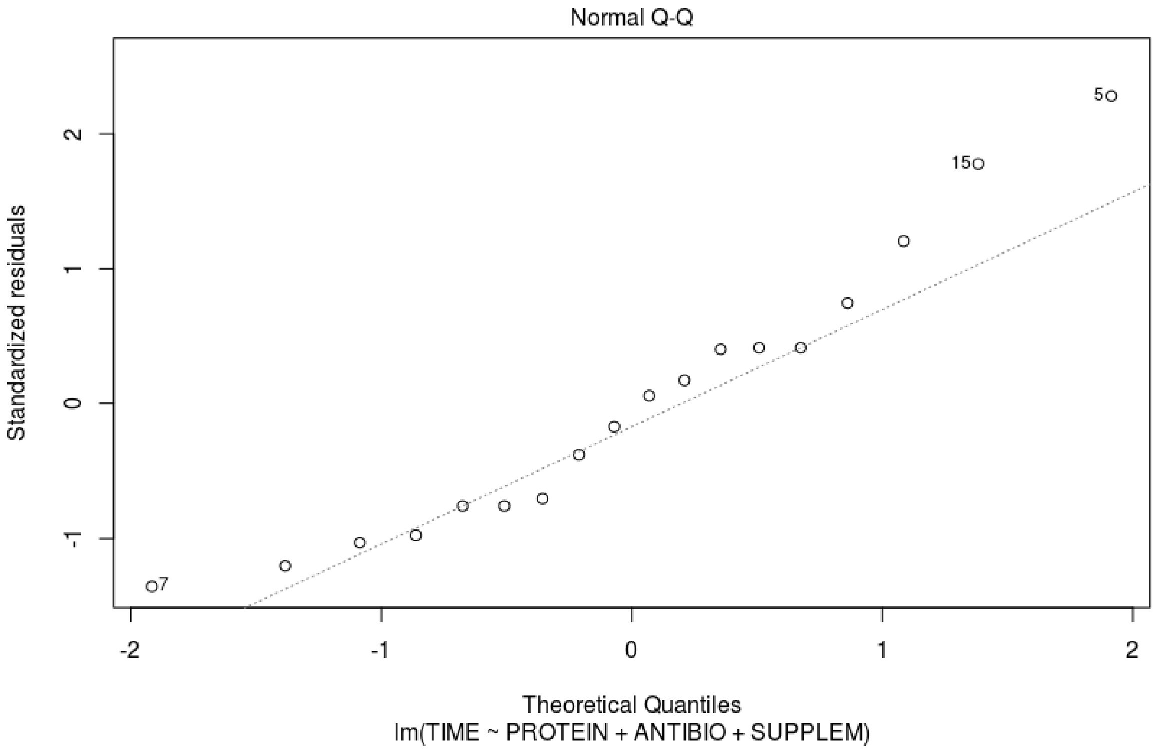
A.



From this plot, we can check for linearity i.e. if the assumption {**E($\varepsilon_i$ = 0) for all i**} is valid or not. If a noticeable curve or trend if observed, then the assumption does not hold. In this case, we can clearly notice an upside-down 'U' curve, hence this assumption does not hold.

We can also check for equal variance of the residuals. If the points are equally scattered, then the equal variance assumption can be validated. From the plot, there appear to be a few outliers (data items 5, 7 and 15). Therefore, this assumption is not validated either.

B.



This plot can tell us if the normality assumption is valid (if residuals are normally distributed). If the residuals line-up linearly, then this assumption holds. However, we can notice from the scatter, that there appears to be a slightly curved feature, curving upwards. Hence, this assumption does not hold.

**7. Interpret the $R^2$ value from "full" model.**

**ANSWER:**

The full model's $R^2$ value is observed to be:

```
Multiple R-squared:  0.9007
```

This value essentially indicates that 90.07 % of the variation in the number of days to bring beef cattle to market weight (TIME) is explained by the model. Since the $R^2$ value is high, predictions can be very precise. However, does this represent a good fit? In our case, no. This is because, the $R^2$ value alone is not sufficient to determine if the model is a good fit. Analyzing the residual plots, we can observe patterns in it, which indicates that our model is, in fact, biased.

**8. Give a one-sentence interpretation of estimated partial regression coefficient for AntiBio in the multiple regression.**

**ANSWER:**

The coefficient value indicates that the Response (TIME) is more sensitive to changes in antibiotic concentration than the other two predictor variables, assuming the model is true.

**9. Working from the “full” model, for each of the four b’s (intercept and three partial regression coefficients) give a p-value for the hypothesis that the true parameter value is zero vs a two-sided alternative. In other words, test $H_0 : \beta_i = 0$ versus $H_A : \beta_i \neq 0$.**

**ANSWER:**

**(A) $H_0 : \beta_i = 0$**

**i = 0**

```
Linear hypothesis test

Hypothesis:
(Intercept) = 0

Model 1: restricted model
Model 2: TIME ~ PROTEIN + ANTIBIO + SUPPLEM

  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     15 5816.8
2     14   40.9  1    5775.9 1976.3 < 2.2e-16 ***
```

**i = 1**

**Linear hypothesis test**

```
Hypothesis:
PROTEIN = 0

Model 1: restricted model
Model 2: TIME ~ PROTEIN + ANTIBIO + SUPPLEM

  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1     15 249.250
2     14  40.917  1    208.33 71.283 7.273e-07 ***
---
```

**i = 2**

**Linear hypothesis test**

```
Hypothesis:
ANTIBIO = 0

Model 1: restricted model
Model 2: TIME ~ PROTEIN + ANTIBIO + SUPPLEM

  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1     15 112.917
2     14  40.917  1        72 24.635 0.0002082 ***
```

**i = 3**

### Linear hypothesis test

```
Hypothesis:
SUPPLEM = 0

Model 1: restricted model
Model 2: TIME ~ PROTEIN + ANTIBIO + SUPPLEM

  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1     15 131.667
2     14  40.917  1     90.75 31.051 6.876e-05 ***
```

**10. Working from the "full" model, test the null hypothesis that the partial regression coefficient for Protein equals -3.0 versus a two-sided alternative. In other words, test $H_0 : \beta_1 = -3$ versus $H_A : \beta_1 \neq -3$. Give a test statistic, p-value and conclusion. (4 pts) Note: One approach to this question uses the car package. Remember you need to install a package the first time you use it and load the package every time you use it!**

### Linear hypothesis test

```
Hypothesis:
PROTEIN = - 3

Model 1: restricted model
Model 2: TIME ~ PROTEIN + ANTIBIO + SUPPLEM

  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1     15 1449.25
2     14   40.92  1    1408.3 481.87 3.033e-12 ***
```

$t = \sqrt{F} = 21.951$
$p = 3.033e^{-12}$
Since p value is lesser than α, we reject the above null hypothesis.

**ANSWER:**

**11. Working from the "full" model, give 95% confidence intervals for each of the four b's (intercept and three partial regression coefficients)**

**ANSWER:**
The 95 % confidence intervals for each of the coefficients are given below:
For Intercept:
$97.75268 \leq \beta \leq 107.66332$

For Protein:
$-1.04499 \leq \beta \leq -0.62161$

For Antibio
$-5.72848 \leq \beta \leq -2.27152$

For Supplem:
$-1.90412 \leq \beta \leq -0.84588$