# Sean Kennedy: Final Exam Statistics 6371

# a):

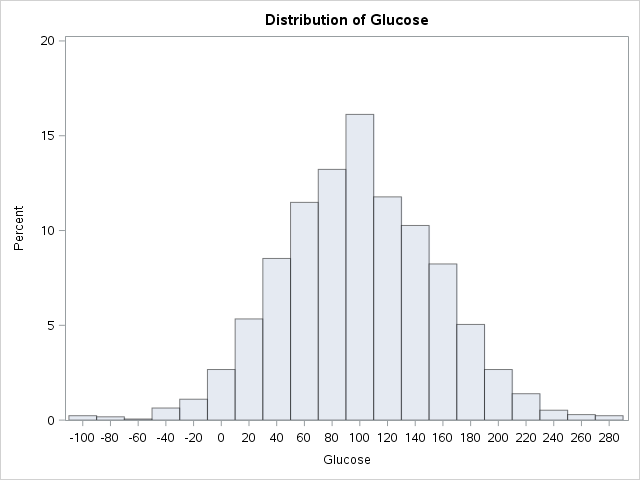
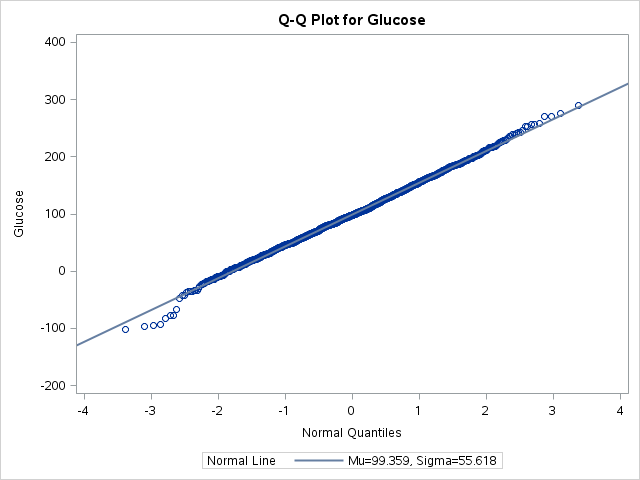
Fit a model for {Glucose|X10}: Glucose =  +  X10

## Model assumptions:

### Normality:

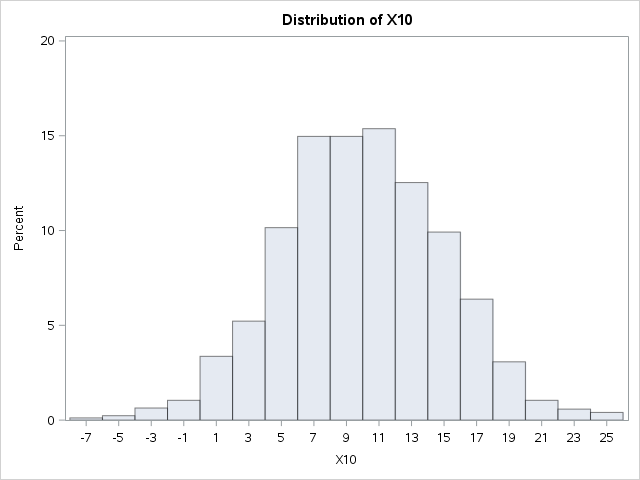
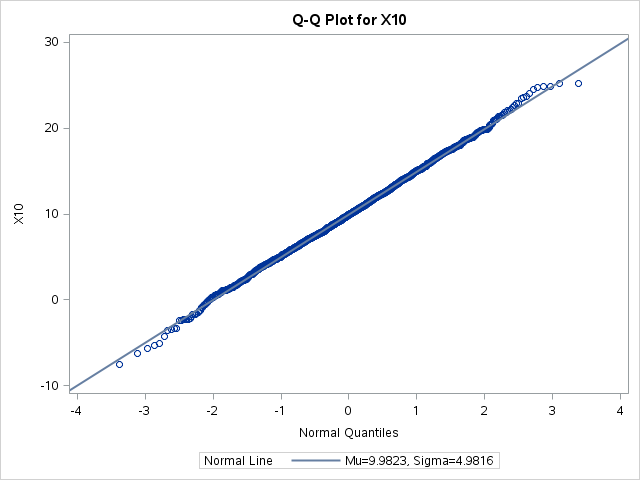
Glucose:

No apparent deviations from normality

X10

Somewhat fat tailed (on both sides) but sample size is large enough to invoke CLT

### Linearity:

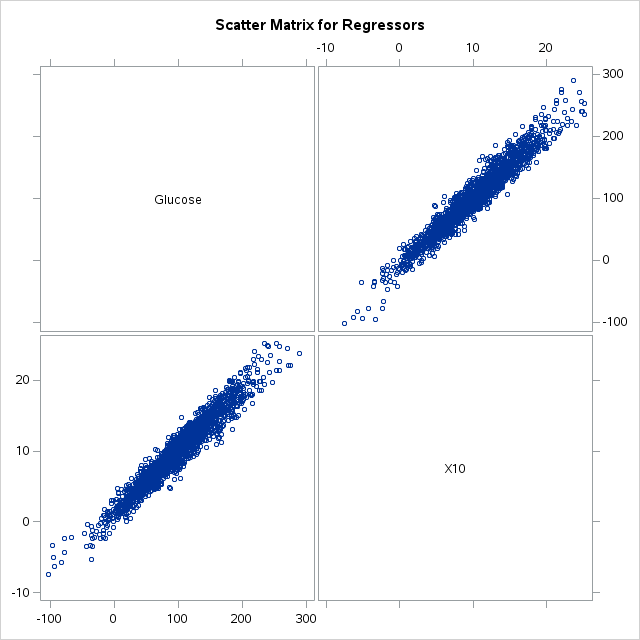
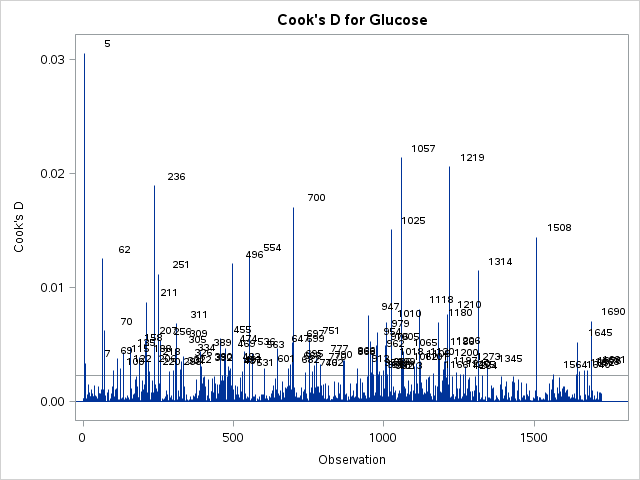
Hard to show more evidence of linearity than this 😊

### Constant Standard Deviation:

No evidence of heteroscedasticity – distribution of glucose at each level of X10 seems to be constant.

### Independence:

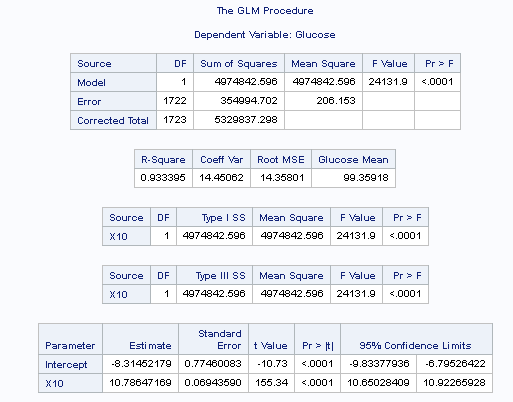
Not applicable here

### Outliers:

Observation 5 is slightly higher than out cooks d cutoff of 0.023 (4/n) – but overall should not be a problem.

## Results:



## Interpretation:

According to our model – we can expect an increase of 10.786 in the glucose production rate for every one unit increase in the expression level of gene X10. A 95% CL for this effect is [10.650, 10.923].

The effective regression line is as follows:

{Glucose|X10} : Glucose = -8.314 + 10.786 X10

# b)

Fit a model for (6 parameters with interactions)

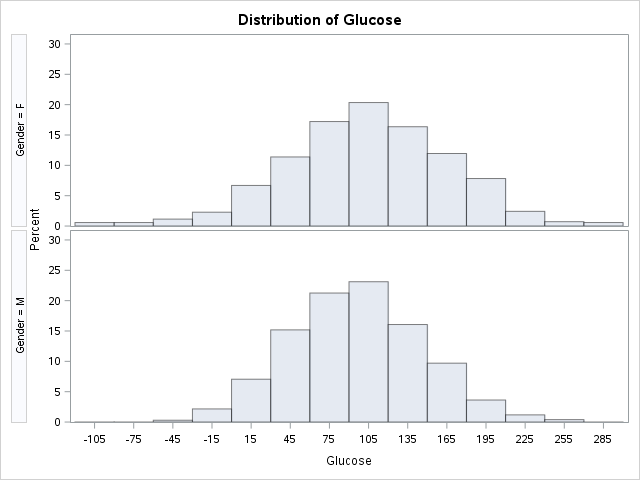
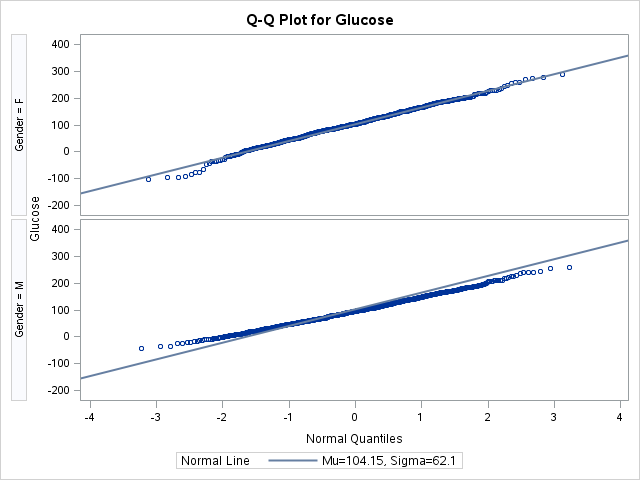
{Glucose|X10 Gender X10(Gender)}: Glucose =  +  X10 +  GenderM +  GenderF +  GenderM X10 +  GenderF X10

## Model assumptions:

### Normality:

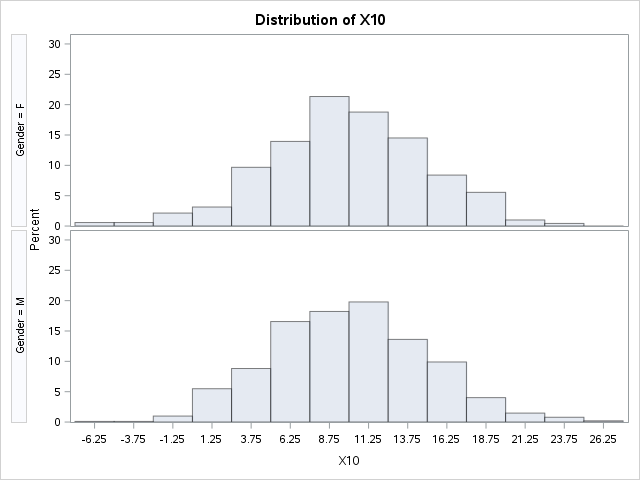
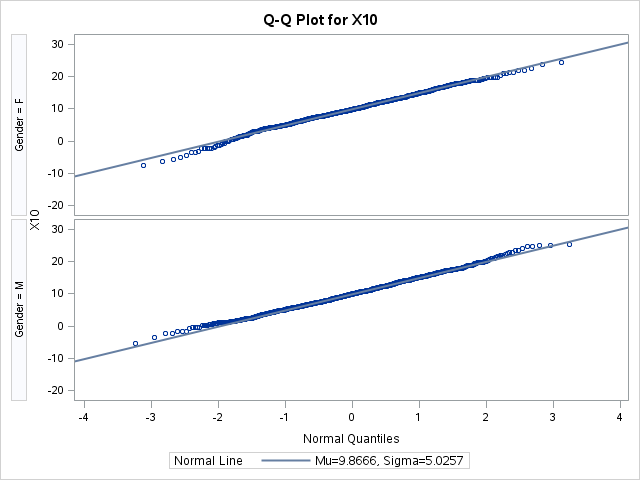
Glucose:

No apparent deviations from normality for either gender

X10

Somewhat fat tailed (on both sides) but sample size is large enough to invoke CLT

### Linearity:

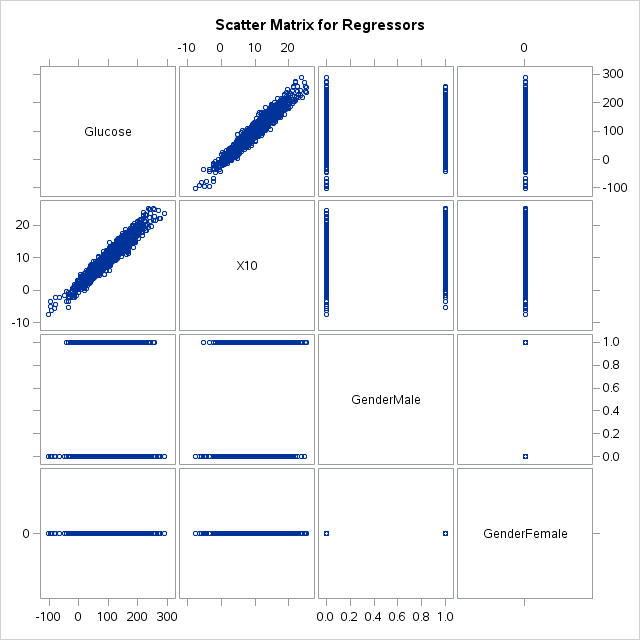
Still hard to show more evidence of linearity than this 😊

### Constant Standard Deviation:

No evidence of heteroscedasticity – distribution of glucose at each level of X10 seems to be constant.

### Independence:

Can only be male or female



## Results:



## Interpretation:

Our general equation for regression is (female is ref level):

{Glucose|X10 Gender X10(Gender)}: Glucose =  +  X10 +  GenderM +  GenderF +  GenderM X10 +  GenderF X10

Or

Glucose = -14.158 + 11.990 \* X10 + 9.942 GenderM -2.024 GenderM X10

For females (our reference level, GenderM = 0):

Glucose = -14.158 + 11.990 \* X10

An increase in the level of X10 gene corresponds to an increase of 11.990 in the rate of glucose production. A 95% CL for this effect is [11.806, 12.174].

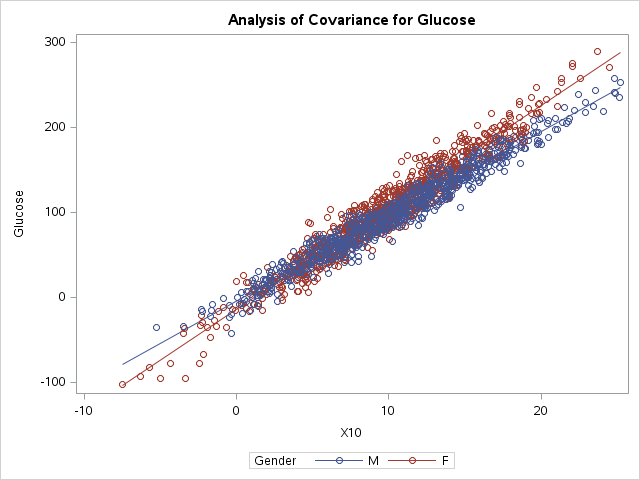
For males:

Glucose = -4.216 + 9.966 X10

An increase in the level of X10 gene corresponds to an increase of 9.966 in the rate of glucose production.

This analysis shows that the relationship between glucose and X10 gene expression differs rather significantly amongst male and female populations (11.990 vs 9.966 – a difference of almost 20%).

The following chart illustrates this difference in slope:



# C):

Fit a model for (12 parameters with interactions)

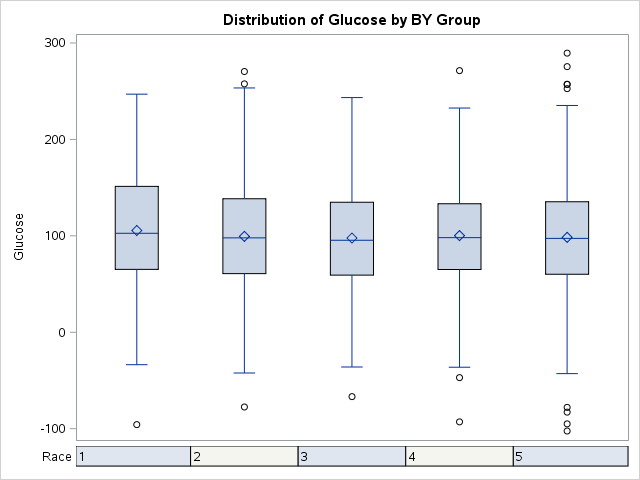
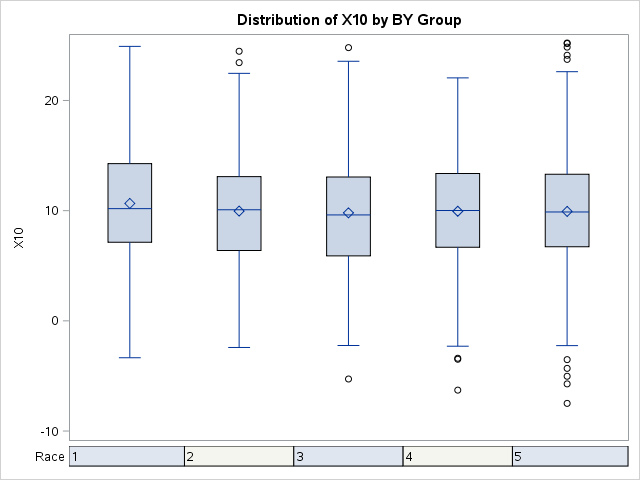
{Glucose|X10 Race X10(Race)}: Glucose =  +  X10 +  Race1 +  Race2 + Race3 +  Race4 +  Race5 +  Race1\_X10 +  Race2\_X10 +  Race3\_X10 +  Race4\_X10 +  Race5\_X10

## Model assumptions:

### Normality:

Glucose:

No apparent deviations from normality for any race – some evidence of positive skew (race 3) but not enough to violate normality given the sample size. Race 5 also appears most effected by outliers/spread.



X10

Somewhat fat tailed (on both sides) but sample size is large enough to invoke CLT

### Linearity:

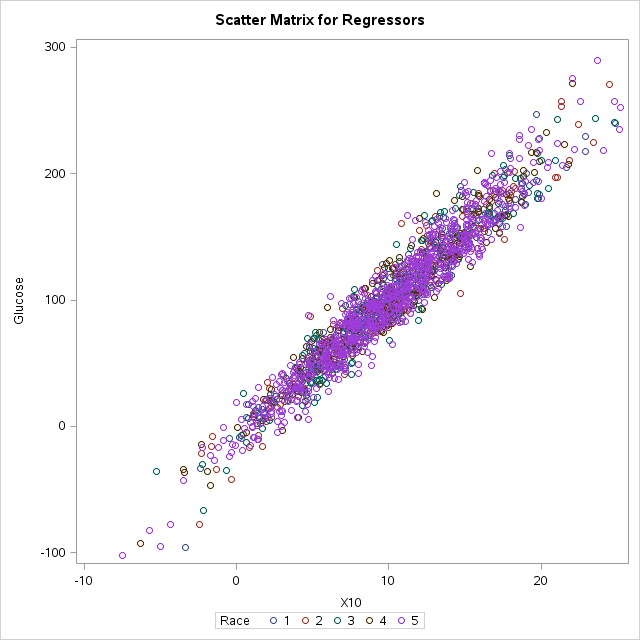
Still hard to show more evidence of linearity than this – starting to see a trend here 😊

### Constant Standard Deviation:

No evidence of heteroscedasticity – distribution of glucose at each level of X10 seems to be constant.

### Independence:

Assuming you can only be in one racial category (big assumption these days)



## Results:

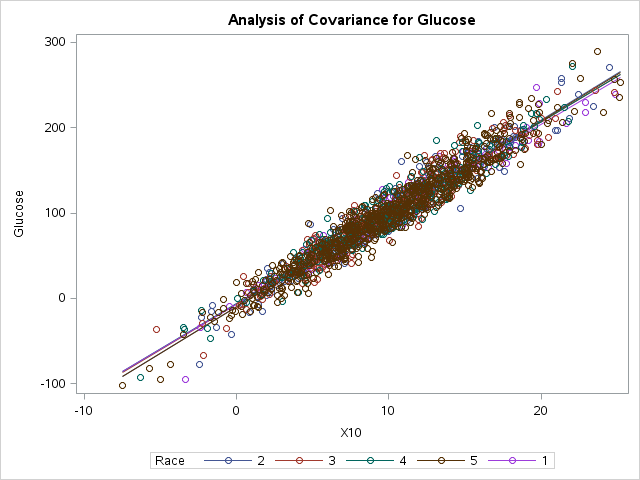


## Interpretation:

Due to the complex nature of these interactions it is difficult to assess the relationship numerically - one could follow the approach for males/females carried out in b) – but it appears from these results (with and without interactions) – that the effect is only significant at the reference level Race = 1.

None of the parameter estimates other than the intercept show statistical significance – whether used as interaction terms or as individual features.

This analysis shows that the relationship between glucose and X10 gene expression does not seem to depend on race. This can also be clearly seen by the almost negligible increase in explanatory power of our model (Adj R2) when adding Race as a factor. This can be clearly seen from the chart below – nearly all lines are identical indicating a constant effect across Racial groups.



# d)

Prediction at for a Male with X10 = 17

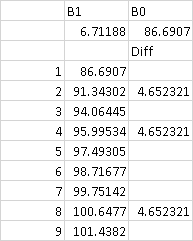
Running a 99% prediction interval in R yields the following estimate:

171.0336 [136.2921, 205.775]

# e)

None of the above – the equation given has a doubling differential = \* LN(2) = 4.652

The table below shows this relationship clearly:

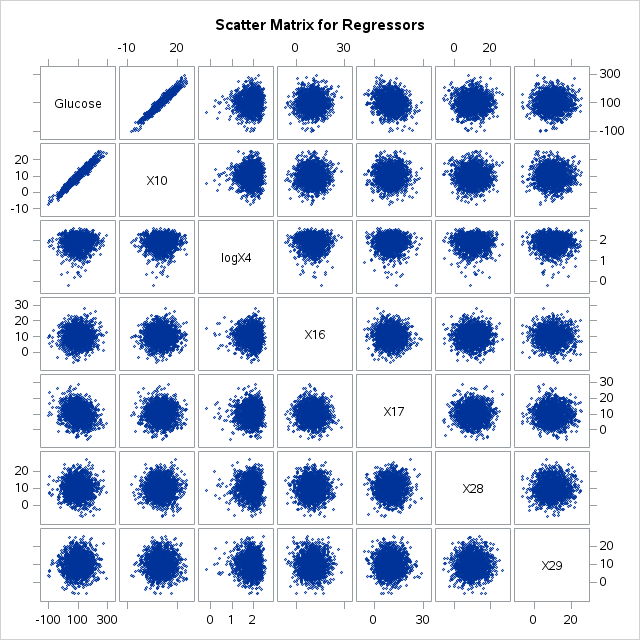


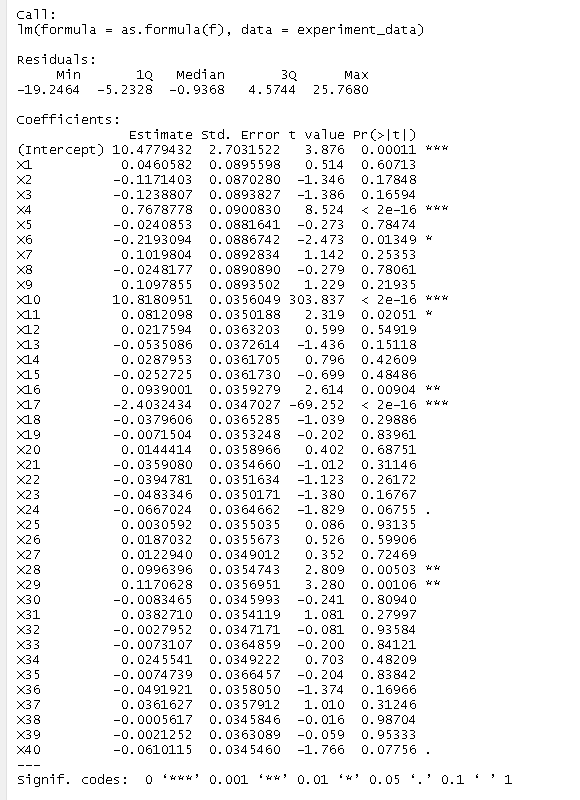
# f)

Failing to reject a false null hypothesis qualifies as a type II error

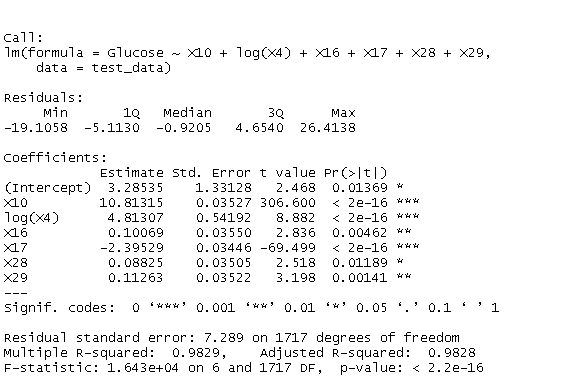
# g)

Running a full model in R across all genes (no interactions considered) shows statistical significance across genes X4, X10, X16, X17, X28 and X29. Running a scatter matrix in SAS shows the following:





Running a model with only those parameters inlcuded proves to be very predictive (making the log transform on X4 as suggested) – accounts for almost 98% of variance.



Further analysis could be done to see if relationships amongst gender and race exist amongst the new continuous regression variables.

# h)

Because this was study was done by volunteers it is only applicable to those involved in the study. Had it been participants been chosen at random – we could extend the analysis to a larger population.