GLM

Emilio Madrazo

November 30, 2022

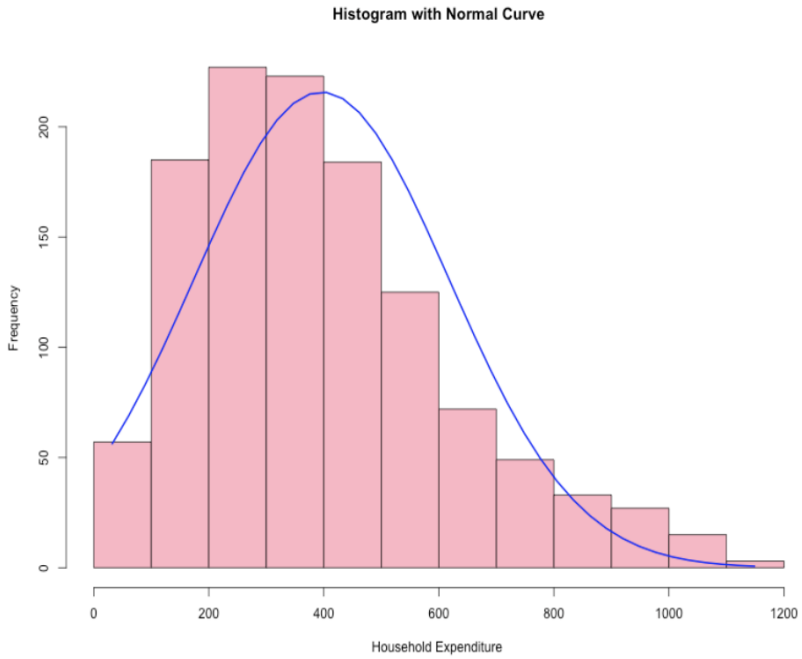
1 Task 1:

1.1 part 1

Graphs and relations;

Distribution of Expenditure:

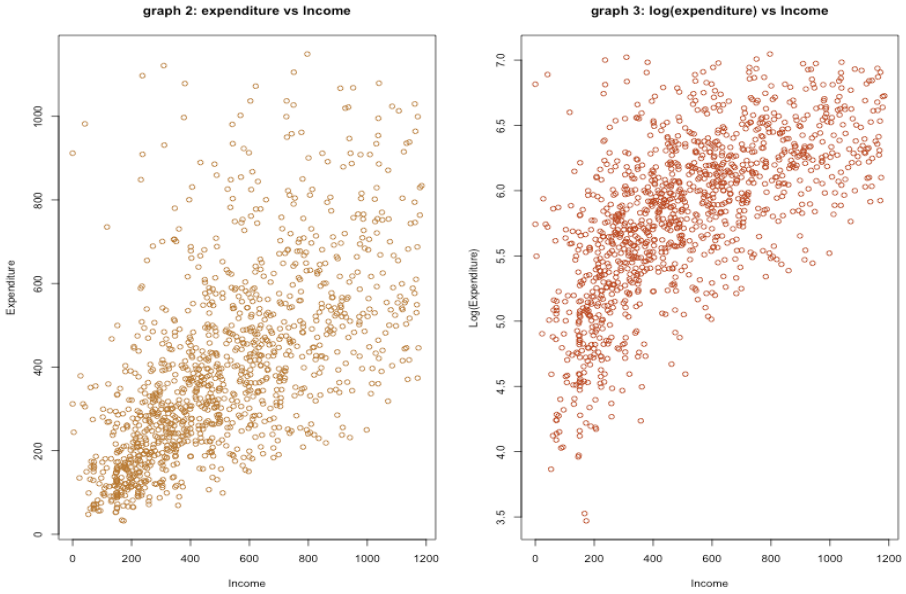
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
32.14	230.69	358.79	395.89	521.16	1149.01



From this graph we can see that the distribution of Expenditure is bell shaped, however does not follow a Normal distribution as shown by the blue line. The

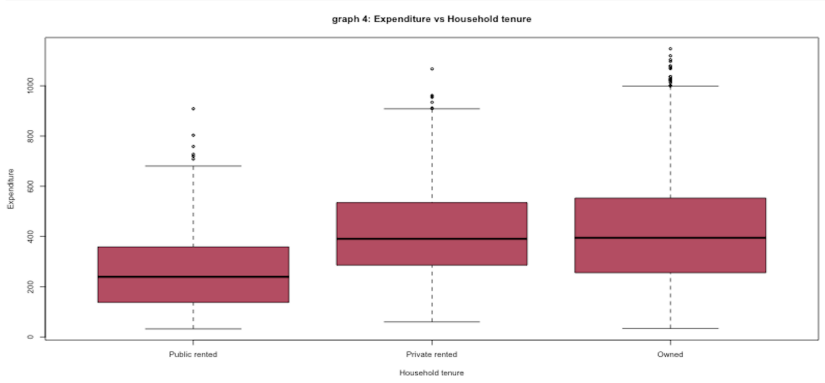
normal distribution and the histogram do not align, and hence we can say that the distribution of Expenditure is right-skewed.

Expenditure - Income:



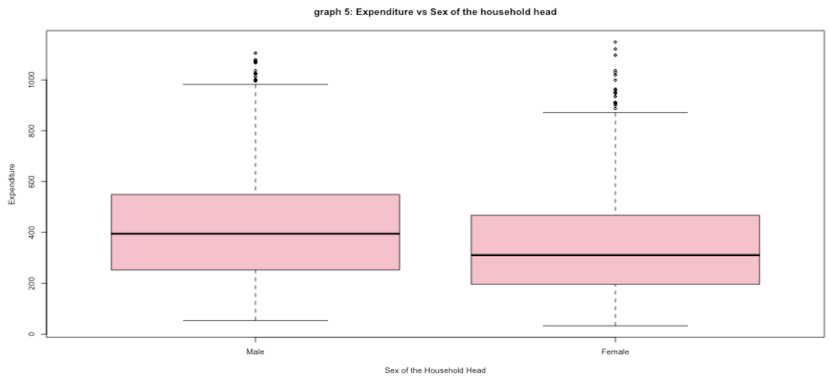
In these two graphs we can see that there is a positive relation ship between Income and Expenditure. In graph 2. we see that as income increases expenditure also increases but with bigger variance. Hence in graph 3. we can see that logarithmically the expenditure seems to reach a plateau as Income increases.

Expenditure - Household tenure:



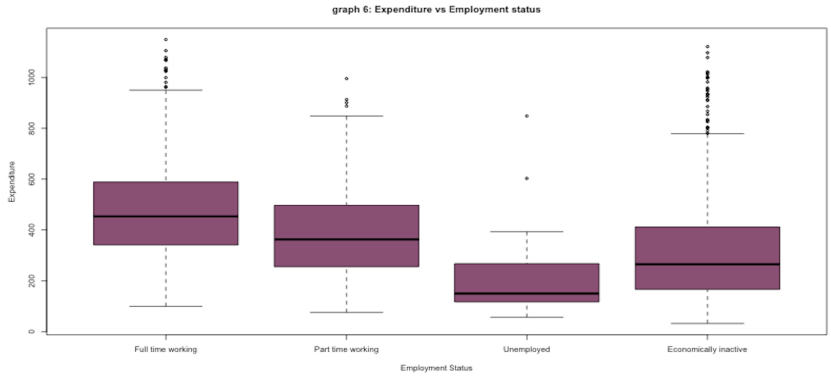
As seen in graph 4. There is positive relationship between public and private renting; those household that rent private have a higher 1st-quartile, median, and 3rd-quartile expenditure, than those who rent public. However owned households have a bigger variance in their expenditure than the previous two and have a higher median than those publicly rented.

Expenditure - Sex of the household head:



The is a clear negative relation between sex of household lead and expenditure. Those houses with men as their household lead have a higher distribution of expenditure than those with a woman.

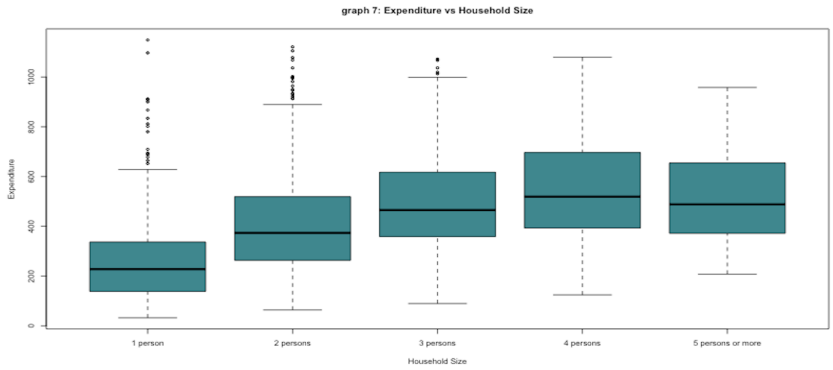
Expenditure - Employment status:



There is a positive relation between employment time and expenditure. Those households whose lead is unemployed have the lowest expenditure median and variance. While those who are part-time employed have a distribution of expenditure much closer to that of the full-time employed households, yet slightly lower. Finally, those who are economically unemployed have a much wider

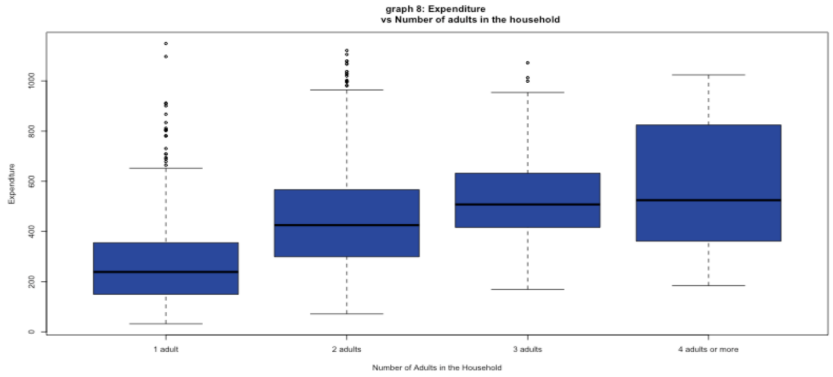
distribution of expenditure with the median being lower than those who are part-time and higher of those unemployed.

Expenditure - Household size:



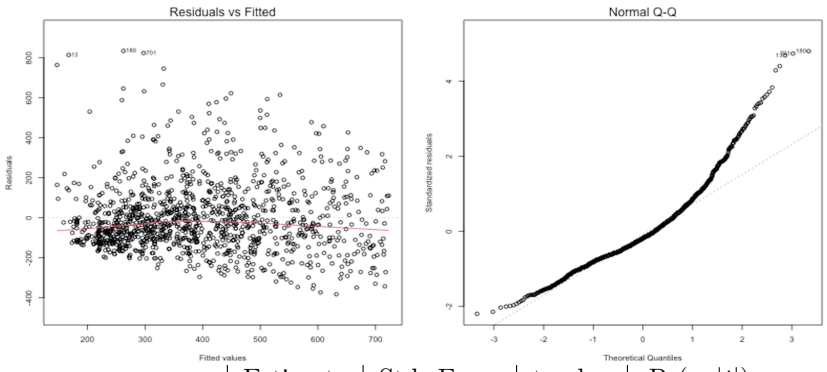
As it is to be expected in graph 7, there is a positive relation between number of people living in the house hold and expenditure. However the data shows that households with 5 or more people spend with lesser variance, and a lower median, than those with only 4 people.

Expenditure - Number of adults in the household:



Graph 8 shows a similar data than that in graph 7. There is a positive relation between expenditure and number of adults in the house hold. however what is interesting here is that house holds with 4 or more adults have a very wide distribution of expenditure even though the median spend than those house holds with only 3 adults.

1.2 part 2: Expenditure vs Income → Model 1

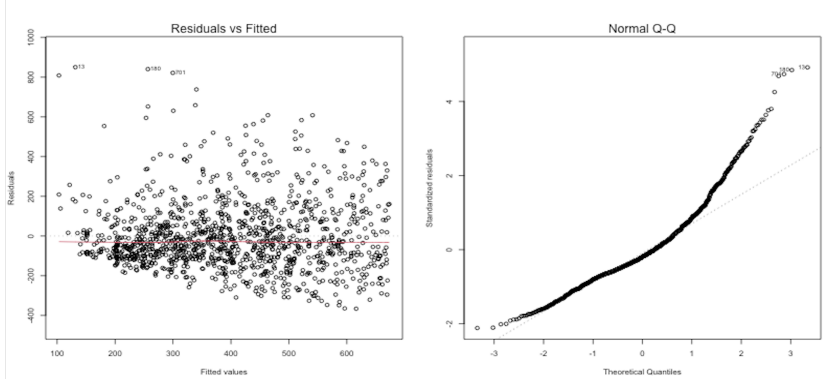


	Estimate	Std. Error	t value	Pr($>\|t\|$)
(Intercept)	147.22	10.39	14.18	$< 2e-16$
Income	0.486	0.0178	27.36	$< 2e-16$

Multiple R-squared: 0.3846, Adjusted R-squared: 0.3841

In model 1 we see that there there is a correlation between expenditure and Income which is significant to more than 99.9%. However as we can see in the Residuals vs Fitted line graph, the model does not follow a horizontal line of points, as the general data is tilted down. However the variance of the residuals appears to remain constant. Hence there is Homoscedasticity but no linearity. And as seen in the Q-Q Plot the distribution is right skewed. Hence no normality.

1.3 part 3: Expenditure vs Income + Income^2→ Model 2

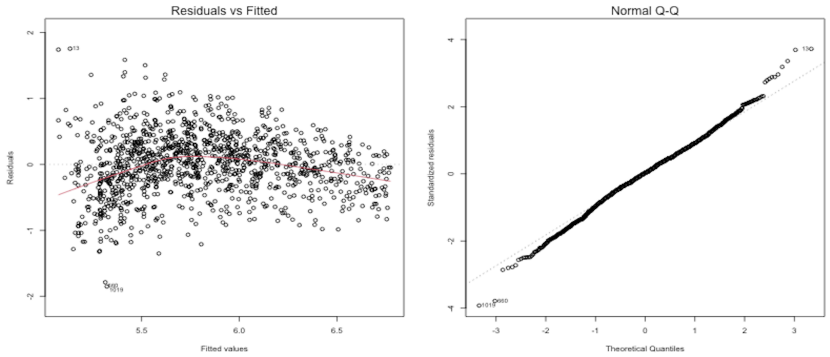


	Estimate	Std. Error	t value	Pr($>\|t\|$)
(Intercept)	1.032e+02	1.805e+01	5.716	1.38e-08
Income	0.690	0.0707	9.76	$< 2e-16$
Income^2	-1.76e-04	5.91e-05	-2.98	0.00295

Multiple R-squared: 0.3891, Adjusted R-squared: 0.3881

Similarly to Model 1, Model 2 shows a strong correlation between income, income^2 and Expenditure significant to 99.9% interval, and to 99% respectively. Similarly to Model 1, the Residuals vs Fitted plot shows a negative slope and hence there is no linearity, additionally the variance of residuals seems to expand, so there is also no Homoscedasticity. Finally in the QQ plot we still see the right side of the graph diverge from the normal line. Hence, again - no Normality.

1.4 part 4: log(expenditure) vs Income

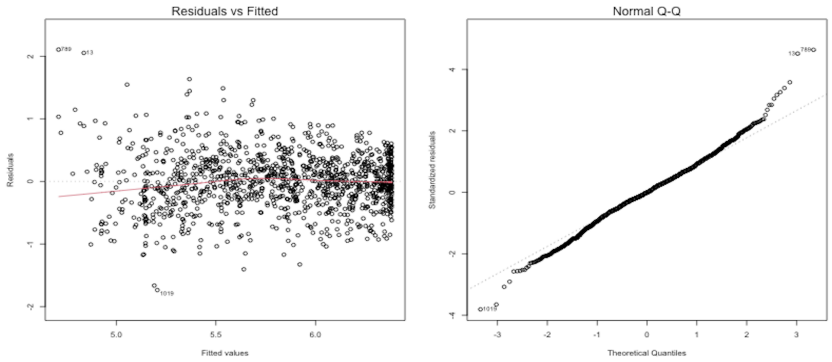


	Estimate	Std. Error	t value	Pr($>\|t\|$)
(Intercept)	5.074	2.82e-02	180.00	$< 2e-16$
Income	1.44e-03	4.82e-05	29.77	$< 2e-16$

Multiple R-squared: 0.4253, Adjusted R-squared: 0.4248

the Residuals vs Fitted line of Model 3 shows a n-shaped curve formed by the residuals, so no linearity. additionally the variance of the residuals shortens, and hence there is no Homoscedasticity. Finally the QQ plot perfectly mimics that of the normal curve, except on the outliers which is to be expected. However, as there is no Homoscedasticity and no Linearity - there is no Normality.

1.5 part 5: log(expenditure) vs Income + Income^2



	Estimate	Std. Error	t value	Pr($>\|t\|$)
(Intercept)	4.71	4.74e-02	99.22	$< 2e-16$
Income	3.12e-03	1.86e-04	16.77	$< 2e-16$
Income^2	-1.45e-06	1.55e-07	-9.349	$< 2e-16$

Multiple R-squared: 0.4644, Adjusted R-squared: 0.4635

Model 4 shows an relatively even variance of the residuals in a straight horizontal line. Hence there is both Homoscedasticity and Linearity. Additionally the QQ plot very closely follows the Normal line, except on the outliers which is to be expected. Hence with all three conditions checked - we can say model 4 follows Normality.

1.6 part 6: selecting a model

Ww can use the AIC estimator to measure the amount of error in each one of our models and compare them, and the Adjusted-R^2 (AR^2) to measure how well fitted the model is to the data normalizing by explanatory variables to see which is the best model. Therefore:

$$AIC(model4) < AIC(model2) < AIC(model1) < AIC(model3) \quad (1)$$

$$1528 < 1578 < 15792 < 1610 \quad (2)$$

and

$$AR^2(model4) > AR^2(model3) > AR^2(model2) > AR^2(model1) \quad (3)$$

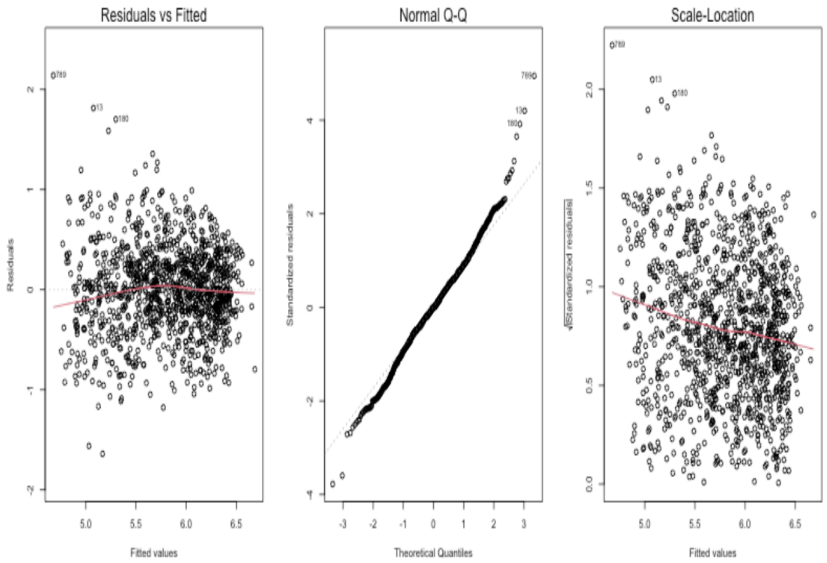
$$0.4635 > 0.4248 > 0.3881 > 0.3841 \quad (4)$$

In addition to the fact that is the only model that follows Normality, we can confidently select model 4 as the best model at predicting Expenditure with Income.

1.7 part 7: My model

The first, naive, thing I did was to run a linear model adding all the available explanatory variables. Naturally, as there is more data, the model improved from that of model 4, with a smaller AIC, and higher AR^2.
However, looking at the relation between expenditure and the different explanatory variables I realised there is a strong interaction between house hold size and number of adults in the house hold.
Similarly, looking at the data, I realised there was another strong interaction between income, and labour force. Even though about 50% of the data-set is economically inactive, the other half has a strong interaction between time worked and income.
Hence I made a second version of my model, multiplying the interactions already stated, and got the following:

	Estimate	Std. Error	t value	$\Pr(>\|t\|)$
(Intercept)	4.59	0.134	34.216	$< 2e-16$
income	1.488e-03	2.81e-04	5.301	1.37e-07
$income^2$	-7.11e-07	1.77e-07	-4.012	6.39e-05
lab.force	-0.121	2.43e-02	-4.964	7.90e-07
hh.size	1.656e-01	3.192e-02	5.189	2.49e-07
hh.adults	0.237	5.14e-02	4.61	4.63e-06
sex.hh	7.42e-02	2.691e-02	2.756	0.00593
house.ten	7.33e-02	1.72e-02	4.26	2.24e-05
income:lab.force	1.722e-04	4.235e-05	4.067	5.08e-05
hh.size:hh.adults	-5.490e-02	1.578e-02	-3.479	0.000522



Ajusted-R^2 : 0.512 AIC : 1423

Which is an improvement in very sense to that of model 4. The residuals follow a horizontal line and a constant variance across the Residuals vd Fitted graph, hence showing linearity and homoscedasticity. While the QQ plot, has a few more outliers, yet never skewes far from the Normal line. And hence this Model holds Normality. Finally the Scale-location plot shows the spread around the red line doesn't change much with the fitted values, hence the variance of magnitudes doesn't vary much as a function of the fitted values.

1.8 part 8: Explaining My model

Every explanatory variable and interaction between them - except the sex of the house hold lead - has a p-value higher than 99.9% significance. Where sex of the house hold lead has a p value of at least 99% significance.In other words, every variable and interaction is highly important in describing expenditure.

From the model results we can see that a 1 unit increase in income results in a 0.00149% increase in Expenditure. Similarly for house hold size and number of adults, a 1 unit increase in people represents a 0.165% and 0.237% increase in expenditure respectively. House tenure is also important in the analysis of expenditure as 1 unit increase in house hold tenure (moving from public-rented, to private-rented, to owned) represents a 0.0733% increase in expenditure. The last of the explanatory variables is that of sex, which says that households with men as the household lead spend 0.0742% than those with women as their household lead.

In terms of interactions of variables we can say that increase in labour force results in an increase in income represents an increase income, which is to be expected given the relation previously stated in part 7. However, what is interesting is that the relation between house hold size and number of adults is negative. Hence we can say that as the number of people in a house increase, the expenditure increases, however slower than that of the number of adults.

2 Task 2:

2.1 part 1

Since

$$f(y;\sigma) = \sigma \exp(-\sigma y), y > 0, \sigma > 0 \tag{5}$$

then we can take log(exp(f(y;σ))) = f(y;σ) = exp($log(\sigma)$-σy)

which is in the form of an EFD, where : a(y) = y, b(σ) = -σ, c(σ) = log(σ), and d(y) = 0

And since we know that:
E(Y_i) = $1/\sigma_i = \mu_i$
Var(Y_i) = $1/\sigma_i^2 = \mu_i^2$, since Y is distributed exponentially
log(μ_i) = $B_0 + B_1 x_i$, and hence $\mu_i = \exp(B_0 + B_1 x_i)$

Then we derive that,
systematic component: $\eta_i = B_0 + B_1 x_i$
link function: $\eta_i = \log(\mu_i)$

We know that in general the score vector is defined by

$$U(B_j) = \Sigma_i^n [(dl_i/d\sigma_i)(d\sigma_i/d\mu_i)(d\mu_i/d\eta_i)(d\eta_i/dB_j)] \tag{6}$$

By parts:

1. dl_i / dσ_i: where l_i is the log-likelihood of Y, and since Y is EFD we know that l_i = log(σ_i)-$\sigma_i y_i$ and hence:

dl_i / dσ_i = ($1/\sigma_i$)-y_i = -(y_i - μ_i)

2. $d\sigma_i/d\mu_i = (d\mu_i/d\sigma_i)^{-1} = (d(1/\sigma_i)/d\sigma_i)^{-1} = (\text{-}1/\sigma_i^2)^{-1} = \text{-}\sigma_i^2 = \text{-}1/\mu_i^2$

3. $d\mu_i/d\eta_i = (d\eta_i/d\mu_i)^{-1} = (dlog(\mu_i)/d\mu_i)^{-1} = (1/d\mu_i)^{-1} = \mu_i$

4. $d\eta_i/dB_j = x_{ij}$

hence,

$$u_j = \Sigma_i^n [\frac{-(y_i - \mu_i)(\mu_i)}{-\mu_i^2} x_{ij}] = \Sigma_i^n [\frac{(y_i - \mu_i)}{\mu_i} x_{ij}] \tag{7}$$

In vector notation:

$$U = X^T (y - \mu)/\mu \tag{8}$$

Similarly we know that:

$$I_{jk} = \Sigma_i^n [\frac{x_{ij} x_{ik}}{Var(y_i)} (d\mu_i/d\eta_i)^2] = \Sigma_i^n [\frac{x_{ij} x_{ik}}{\mu_i^2} \mu_i^2] = \Sigma_i^n x_{ij} x_{ik} \tag{9}$$

In vector notation:

$$I = X^T X \tag{10}$$

2.2 part 2

After the iterations in R we get that B = (B_0, B_1) = (-0.121, 0.00526)

2.3 part 3

The variance-covariance matrix is:

1.84e-02	-2.87e-05
-2.87e-05	4.72e-08

2.4 part 4

Since we want to measure what is the importance of B_1 we assume the mean of the distribution we are comparing to is 0, and hence:
$t_{B1} = 24.2$
hence for our 988 degrees of freedom, $\Pr(> |t_{B1}|) < 0.001$, and so we can say it is highly significant