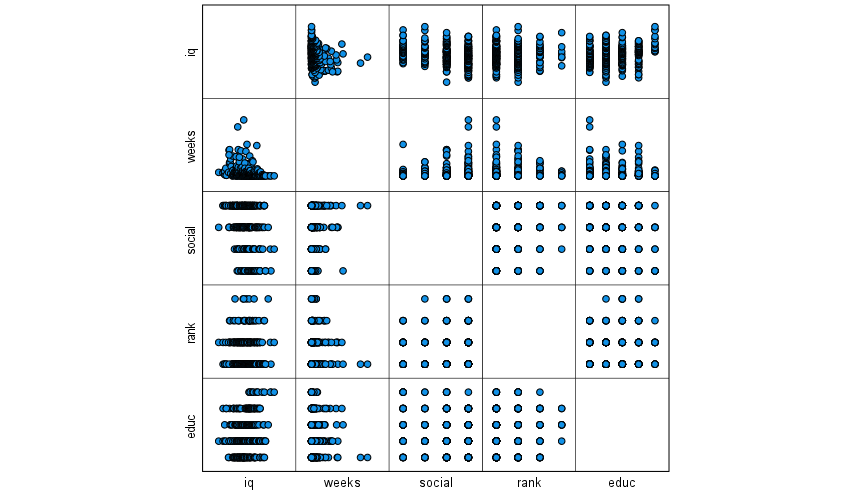
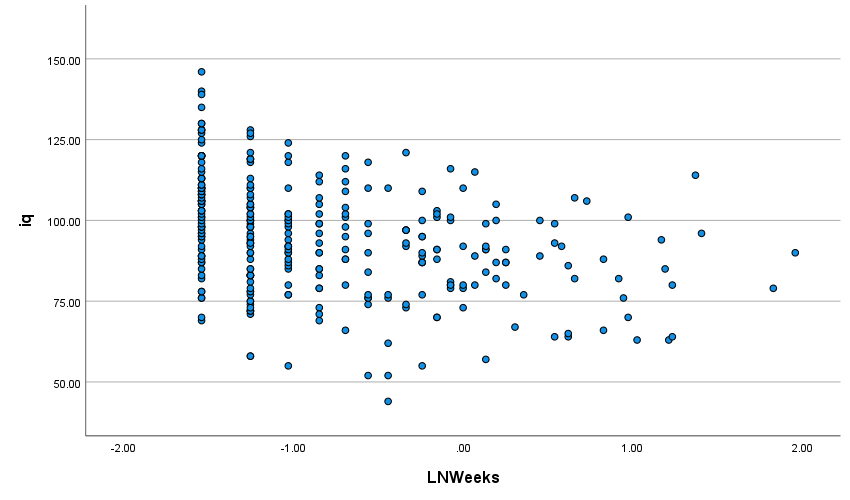
1. This study is an observational study, with the subjects being 300 babies who were admitted to a special-care unit, as we cannot assign treatments (breastfed or not) to the babies due to ethics and as such, we just measure other factors as they appear in nature. We can not make causal nor population inferences in this case as our samples were not randomly chosen and as such can’t be generalized to the population. As for why we are unable to make causal inferences, it is due to this study being an observational study.
2. It shouldn’t affect the conclusion too much in the end as the babies are still receiving mother’s milk in the end (same source) but it just involves a different feeding method. However, we can use a tube to ensure that the babies are receiving milk as it could potentially spill during feeding whereas it is less likely from a tube which means that the babies could effectively get more milk.
3. The benefit of treating them as numerical variables is that we can perform more statistical process and calculation on them than with categorical variables due to the numerical nature. However, one assumption we have is that each predictor has an equal interval between each value.



* 1. IQ and Social appears to have a weak negative linear relationship so one could use a linear model for this. IQ and education level appears to have an outlier in the data with rank 5 on average having high IQ levels compared to the non-existent trend for prior levels. Rank and IQ does not seem to indicate a linear relationship at all with Rank 5 having really wide data range with a small amount of values. When comparing IQ and weeks, we can clearly see a non-linear relationship with a few outliers so using a linear model(as with rank and education) would not be effective.



* 1. There seems to be a weak negative linear relationship between the two which has improved when comparing weeks to IQ which indicated no linear relationship at all.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Correlations** | | | | | | | | |
|  | | iq | social | rank | educ | LNWeeks | milk | fem |
| iq | Pearson Correlation | 1 | -.282\*\* | -.012 | .217\*\* | -.354\*\* | .340\*\* | .142\* |
| Sig. (2-tailed) |  | <.001 | .838 | <.001 | <.001 | <.001 | .014 |
| N | 300 | 300 | 300 | 300 | 300 | 300 | 300 |
| social | Pearson Correlation | -.282\*\* | 1 | .010 | -.141\* | .125\* | -.224\*\* | -.034 |
| Sig. (2-tailed) | <.001 |  | .862 | .015 | .030 | <.001 | .555 |
| N | 300 | 300 | 300 | 300 | 300 | 300 | 300 |
| rank | Pearson Correlation | -.012 | .010 | 1 | .090 | .055 | -.019 | -.021 |
| Sig. (2-tailed) | .838 | .862 |  | .121 | .346 | .739 | .717 |
| N | 300 | 300 | 300 | 300 | 300 | 300 | 300 |
| educ | Pearson Correlation | .217\*\* | -.141\* | .090 | 1 | -.123\* | .194\*\* | .036 |
| Sig. (2-tailed) | <.001 | .015 | .121 |  | .033 | <.001 | .540 |
| N | 300 | 300 | 300 | 300 | 300 | 300 | 300 |
| LNWeeks | Pearson Correlation | -.354\*\* | .125\* | .055 | -.123\* | 1 | -.458\*\* | -.044 |
| Sig. (2-tailed) | <.001 | .030 | .346 | .033 |  | <.001 | .449 |
| N | 300 | 300 | 300 | 300 | 300 | 300 | 300 |
| milk | Pearson Correlation | .340\*\* | -.224\*\* | -.019 | .194\*\* | -.458\*\* | 1 | .010 |
| Sig. (2-tailed) | <.001 | <.001 | .739 | <.001 | <.001 |  | .858 |
| N | 300 | 300 | 300 | 300 | 300 | 300 | 300 |
| fem | Pearson Correlation | .142\* | -.034 | -.021 | .036 | -.044 | .010 | 1 |
| Sig. (2-tailed) | .014 | .555 | .717 | .540 | .449 | .858 |  |
| N | 300 | 300 | 300 | 300 | 300 | 300 | 300 |
| \*\*. Correlation is significant at the 0.01 level (2-tailed). | | | | | | | | |
| \*. Correlation is significant at the 0.05 level (2-tailed). | | | | | | | | |

* 1. From the correlation matrix we can see that the results match the results I mention above yet it appears we have an even weaker negative linear relationship than expected from 4a. From this we can see the strongest linear relationship with IQ is LNWeeks(weeks log transformed) so I would use this variable to predict IQ at age 8.
  2. We can see the strongest correlation coefficient value from the above table is between milk and LNWeeks with a value of -.458 which indicates there is a slightly weak correlation. However, this should cause an issue with multicollinearity as the relationship is not very strong.

1. IQ=β0+ β1\*Rank+ β2\*LNWeeks+ β3\*Social+ β4\*Female+ β5\*Education+ β6\*Milk

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model Summary** | | | | |
| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
| 1 | .487a | .237 | .222 | 15.31614 |
| 2 | .487b | .237 | .224 | 15.29014 |
| a. Predictors: (Constant), milk, fem, rank, educ, social, LNWeeks | | | | |
| b. Predictors: (Constant), milk, fem, educ, social, LNWeeks | | | | |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **ANOVAa** | | | | | | |
| Model | | Sum of Squares | df | Mean Square | F | Sig. |
| 1 | Regression | 21387.761 | 6 | 3564.627 | 15.196 | <.001b |
| Residual | 68733.156 | 293 | 234.584 |  |  |
| Total | 90120.917 | 299 |  |  |  |
| 2 | Regression | 21387.125 | 5 | 4277.425 | 18.296 | <.001c |
| Residual | 68733.791 | 294 | 233.788 |  |  |
| Total | 90120.917 | 299 |  |  |  |
| a. Dependent Variable: iq | | | | | | |
| b. Predictors: (Constant), milk, fem, rank, educ, social, LNWeeks | | | | | | |
| c. Predictors: (Constant), milk, fem, educ, social, LNWeeks | | | | | | |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Coefficientsa** | | | | | | |
| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. |
| B | Std. Error | Beta |
| 1 | (Constant) | 89.500 | 4.404 |  | 20.321 | <.001 |
| LNWeeks | -5.262 | 1.294 | -.234 | -4.068 | <.001 |
| educ | 1.845 | .774 | .125 | 2.382 | .018 |
| rank | -.061 | 1.181 | -.003 | -.052 | .959 |
| social | -3.293 | .894 | -.194 | -3.685 | <.001 |
| fem | 4.138 | 1.773 | .119 | 2.334 | .020 |
| milk | 6.156 | 2.225 | .163 | 2.767 | .006 |
| 2 | (Constant) | 89.410 | 4.039 |  | 22.139 | <.001 |
| LNWeeks | -5.266 | 1.290 | -.234 | -4.083 | <.001 |
| educ | 1.841 | .769 | .125 | 2.392 | .017 |
| social | -3.294 | .892 | -.194 | -3.692 | <.001 |
| fem | 4.140 | 1.770 | .119 | 2.340 | .020 |
| milk | 6.156 | 2.221 | .163 | 2.772 | .006 |
| a. Dependent Variable: iq | | | | | | |

From the above tables, we can see that only rank was removed from the model. That means the new model would be IQ=β0 + β1\*LNWeeks+ β2\*Social+ β3\*Female+ β4\*Education+ β5\*Milk

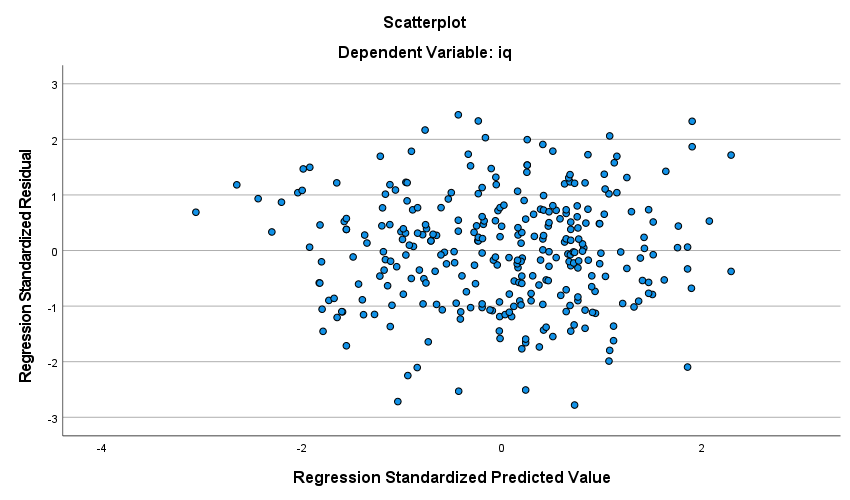
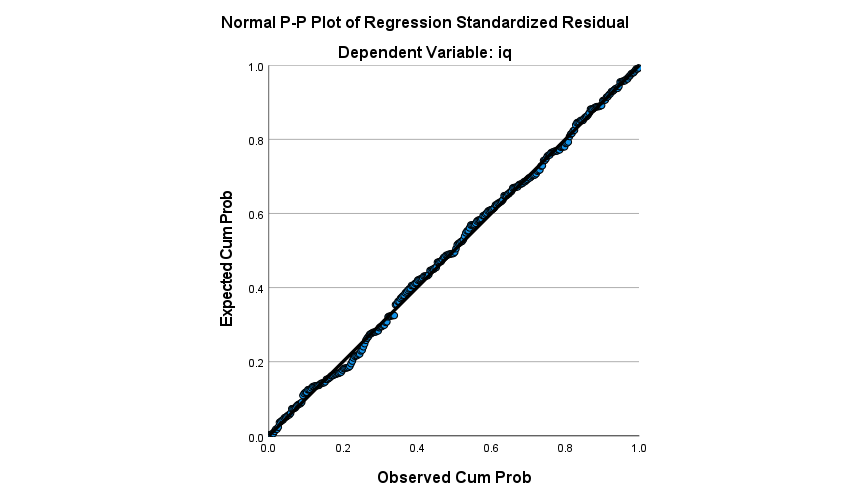
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model Summary** | | | | |
| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
| 1 | .487a | .237 | .224 | 15.29014 |
| a. Predictors: (Constant), milk, fem, educ, social, LNWeeks | | | | |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **ANOVAa** | | | | | | |
| Model | | Sum of Squares | df | Mean Square | F | Sig. |
| 1 | Regression | 21387.125 | 5 | 4277.425 | 18.296 | <.001b |
| Residual | 68733.791 | 294 | 233.788 |  |  |
| Total | 90120.917 | 299 |  |  |  |
| a. Dependent Variable: iq | | | | | | |
| b. Predictors: (Constant), milk, fem, educ, social, LNWeeks | | | | | | |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Coefficientsa** | | | | | | |
| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. |
| B | Std. Error | Beta |
| 1 | (Constant) | 89.410 | 4.039 |  | 22.139 | <.001 |
| LNWeeks | -5.266 | 1.290 | -.234 | -4.083 | <.001 |
| educ | 1.841 | .769 | .125 | 2.392 | .017 |
| social | -3.294 | .892 | -.194 | -3.692 | <.001 |
| fem | 4.140 | 1.770 | .119 | 2.340 | .020 |
| milk | 6.156 | 2.221 | .163 | 2.772 | .006 |
| a. Dependent Variable: iq | | | | | | |

From, the above tables, we can see our regression equation would become IQ=89.410 - 5.266\*LNWeeks - 3.294\*Social+ 4.140 \*Female+ 1.841 \*Education+ 6.156\*Milk.

* 1. Our Null hypothesis is the above model is not significant whereas our alternate statistic is the model is significant. We can see that that sum of squares residuals is 68733.791 with a df of 294. We have a test statistic of 18.296 and a P-value less than 0.001 which means we reject the null hypothesis and say our model is significant and useful.
  2. R2=.237. This means 23.7% of the variance is explained by the model

* 1. From the above graph, we can see homoscedasticity is not being violated as there is no pattern being found in the above graph. In other words, as the points are randomly scattered throughout the graph so variance of the residuals does not change with increasing fitted values 
  2. We can see that the points lie close to the lie with seldom few points lying far away from the line. This means that normality is not being violated.

1. There appears to be no problematic values within the data if by problematic, we mean that it as long as the value is student’s residual’s magnitude is above 2 and Cooks distance is close to/greater than one and Leverage value is greater than 10/300 (2p/n). None of the data values has Cooks value of greater than or even close to 1. For Leverage, we appear to have 12 values which are greater than 10/300 with our largest value being .05628940093691 and our smallest value above 10/300 is .03350940600287. For all the 12 values in this range however, none of the absolute value of the student’s residual is above 2(The largest absolute value is 1.84203610266671. Looking at Student Residuals, we can see 15 values which have an absolute value over 2 (From the negative perspective, the largest negative value is -2.80455124764296 while the largest positive value is 2.47064534853459) and none of those values have cooks value over 1 and one of the Leverage values being over 10/300.
   1. H0:Slope =0 whereas our Ha : Slope is different from 0. Our test statistic for milk is 2.772 which leads to a P-Value of .006. As this t statistic(Its null distribution is under the t distribution with a df of 294) gives a p-value which is less than our alpha value of 0.05, we reject the null hypothesis which means there is sufficient evidence to say that our slope is different from 0.
   2. Our Confidence interval is ( 1.785 , 10.527 )
2. H0:Slope =0 whereas our Ha : Slope is different from 0. Our test statistic for gender is 2.34 which leads to a P-Value of .02. As this t statistic (Its null distribution is under the t distribution with a df of 294) gives a p-value which is less than our alpha value of 0.05, we reject the null hypothesis which means there is sufficient evidence to say that our slope is different from 0. From the table, we can see that our average IQ difference between males and females is estimated to be about 4.140.
3. Our model from 7b. is IQ=89.410 - 5.266\*LNWeeks - 3.294\*Social+ 4.140 \*Female+ 1.841 \*Education+ 6.156\*Milk. If we are to use Ln(2Weeks), we get an equation of Q=89.410 - 5.266\*LN2Weeks - 3.294\*Social+ 4.140 \*Female+ 1.841 \*Education+ 6.156\*Milk. If we subtract the two we get- 5.266\*LN(2Weeks) + 5.266\*LNWeeks. We can further simplify this to - 5.266\*ln(2). That means the effect of spending double the time is - 5.266\*ln(2) which is about -3.650
4. H0:Social = Education=0 whereas our Ha : At least one of social or education is not equal to 0. F=(73727.826-68733.791/296-294)/ 68733.791/294 which is equal to 88.86599377 with a df of [2,294]. This means our null distribution of our test statistic is under the f-distribution with a df of [2,294]. This means we get a p-value of less than 0.001 so we reject our null hypothesis at the 5% significance and say with sufficient evidence that At least one of social or education is not equal to 0.
5. IQ=89.410 - 5.266\*LN(2.5) - 3.294\*4+ 4.140 \*0+ 1.841 \*4+ 6.156\*1. That gives us 84.929. Our prediction interval is (78.584, 91.270) while our confidence interval is (54.174, 115.681). We can see that out confidence interval is much wider than our prediction interval in this case.