Fitting a Model Using Provided Variables

We use all the variable provided in the data except id and data in our analysis to find a linear model that have the best fit. We choose AIC backwards selection method by using step function. After we get the lowest AIC scores, we get the model with following variables (table 1).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Coefficients(Intercept)** | **gestation** | **ht** | **ded** | **dwt** | **inc** | **time** |
| -83.21759 | 0.44988 | 1.04438 | -1.33282 | 0.07541 | -0.50323 | 1.99309 |
|  | **number** | **inc:parity** | **wt.1:ed** | **ed:drace** | **time:ed** | **ded:smoke** |
|  | -2.02635 | 0.15018 | 0.01692 | -0.20793 | -0.64961 | 0.61855 |

We also use Anova function to check whether the variable have contributed to the predictive ability of the model, all the p value is less than 0.05, which means all the variables that we selection contribute to the model’s predictive ability. We trace down the extreme residuals in the model. Then we want to check the assumptions about the model. For normality, we use Shapiro Wilk normality test which null hypothesis is the population is normally distributed which our p value is 0.09, we fail to reject the hypothesis. From QQ plot of residuals of the model below and Shapiro Wilk normality test, we could conclude that the model fit normality. For linearity, we have plotted a residual against fitted values graph, even though our graph is not perfect, it is a well-behaved graph that show the linearity of the model. For heteroskedasticity, we use Breusche-Pagan test (ncvTest) which null hypothesis is constant error variance, p value is 0.05 which reject the null hypothesis. So heteroskedasticity do exist and we could also see from the graph of residual against fitted data. For autocorrelation, we use Durbin Watson statistic which null hypothesis is linear regression residuals are uncorrelated, p value is 0.54 which means that we fail to reject the null hypothesis. For collinearity, we use variance inflation factors to find collinearity in our model. Since all of our variance inflation factor is less than 10, we don’t need to worry about collinearity in our model.