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(3359.82), we get the model with following variables (Table 1).

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **(Intercept)** | **gestation** | **parity** | **ht** | **drace** | **dwt** | **smoke** | **number** |
| -98.99311 | 0.454 | 0.74966 | 1.26968 | -0.56526 | 0.07689 | 2.15663 | -2.16762 |

Table.1 DataModel coefficents

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. (The name of the model is dataModel in r file, you could find all the r file in appendix modelselection-Su.r)

Fitting a First Order Interaction Model

In this model. We only looked at first order interactions between two variables. There is a possible that second order interaction between three or more variables maybe useful to analysis. We use all the variables and first order interaction between every two variables in the data except id and date, so we have roughly 200 to 300 variables to start with. Then we will proceed, choose AIC backwards selection approaches using step function. After this stage, we eliminate a high number of variables, being now reduce to some 50 – 60 variables. Then we examine the collinearity of the first order model. It is observed that there is a considerable number of variables which variance inflation factor is larger than 10, which prompt us to the following step.

1. We find the maximum number of GVIF; if it is larger than 10 then we remove it, and update our function;

2. Checking collinearity using vif function get the maximum - repeat the step 1;

After we remove all the variable which variance inflation number is larger than 10, we use AIC backwards selection again since we form a new model by deleting a lot of collinear variables. After model selection, we are left with 12 variables, the AIC scores of the model is 3358.58. We check the collinearity again and the variance inflation factor, now all VIF are less than 10. After we get the model below (Table.2), we check for the assumption.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **(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 |

Table.2 FinalModel coefficients

For normality, we use Shapiro Wilk normality test which null hypothesis is the population is normally distributed which our p value is 0.22, 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 0.16 which fail to reject the null hypothesis. So heteroskedasticity do not 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 0.08 which means that we fail to reject the null hypothesis. Now this model pass all the assumption. We also do a Anova test on the model, even though some of the p values are more than 0.05, because we are using AIC model selection, AIC and p value sometimes does not agree with each other. So, we could ignore the issue since we are using AIC model selection. (The name of the model is finalModel, you could find all the r file in appendix modelselection-Su.r)