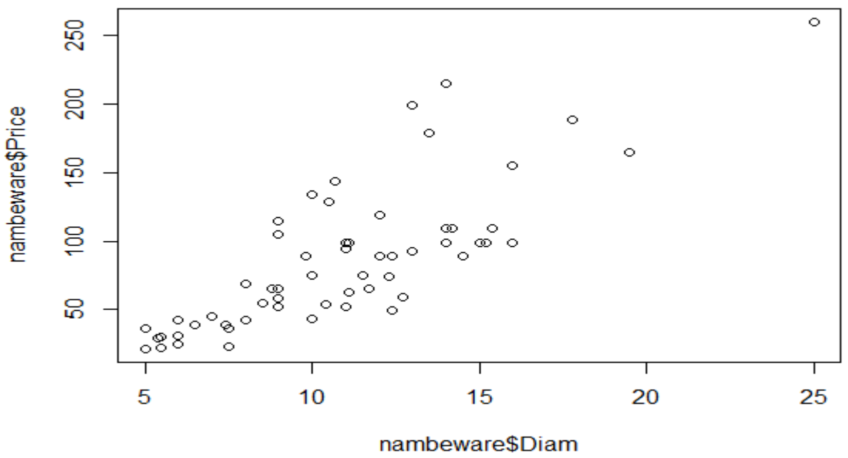
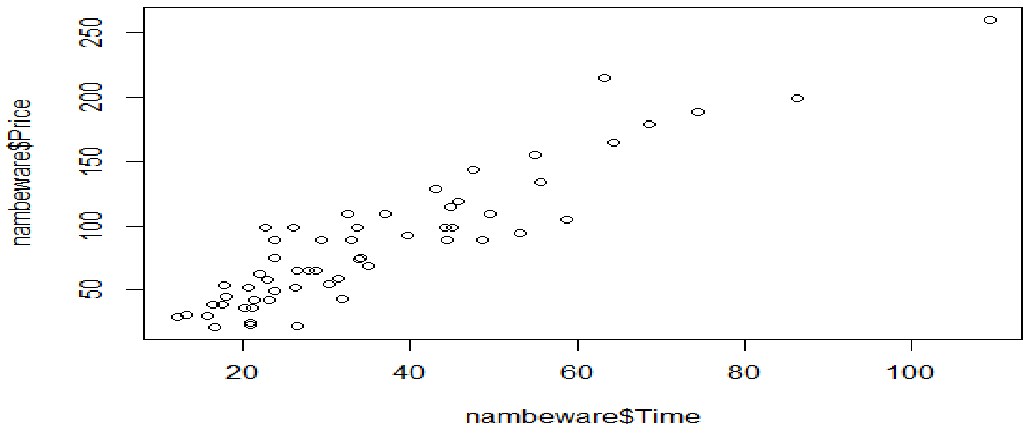
GLM ASSIGNMENT 3
TAMOGHNA DEY
19MSMS31

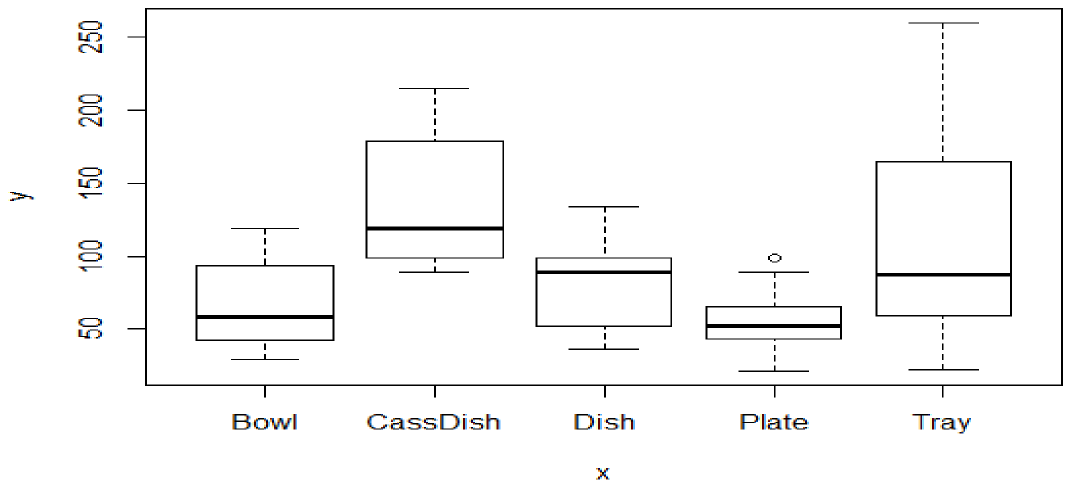
Data: nambeware
Package in R: GLMsData
Description: The data come from Nambe Mills, manufacturers of tableware made from sand casting a special alloy of several metals. The polishing times for the products are thought to be related to the size of the item, as indicated by the diameter. After casting, the pieces go through a series of shaping, grinding, buffing, and polishing steps. In 1989 the company began a program to rationalize its production schedule of some 100 items in its tableware line. The total grinding and polishing times listed here were a major output of this program.
Variables: The data has 59 observations on 4 variables
a) Type:
b) Diam
c) Time
d) Price
Objective: We suspect that Inverse Gaussian GLM with log as link function may be appropriate for this data. Here Price is our dependent variable of interest. We are to fit a model, assess it and draw inference on it.

EXPLORATORY DATA ANALYSIS



Plot of Diam against Price shows a positive relation, although it is probably not linear.





The plot of Time against Price shows positive trend which mean higher the polishing and grinding time for the product, higher its price.

The boxplot of the prices of different types of product show that Trays have the largest price range; Bowls, Dishes, Plates have low and small price range; however CassDish has a relatively high but small price range.

FITTING THE GLM

The model that we are fitting is

$$\eta = \log(\mu_y) = \beta_0 + \beta_1 I_{CassDish} + \beta_2 I_{Dish} + \beta_3 I_{Plate} + \beta_4 I_{Tray} + \beta_5(\text{Diam}) + \beta_6(\text{Time})$$

Where $\beta_0 = 2.523$, $\beta_1 = -0.013$, $\beta_2 = 0.081$, $\beta_3 = -0.331$, $\beta_4 = -0.494$, $\beta_5 = 0.119$, $\beta_6 = 0.018$

Point Estimation

We assume that Price follows an Inverse Gaussian Distribution with mean μ and shape parameter λ.
Estimate of μ= mean of Price= $\hat{\mu}$= 86.381
Estimate of λ= $\hat{\lambda}$= $(E[Y^{-1}] - \hat{\mu}^{-1})^{-1}$=203.783
The Pearson Estimate of the Dispersion Parameter is given by Φ= 0.0004

Interval Estimation

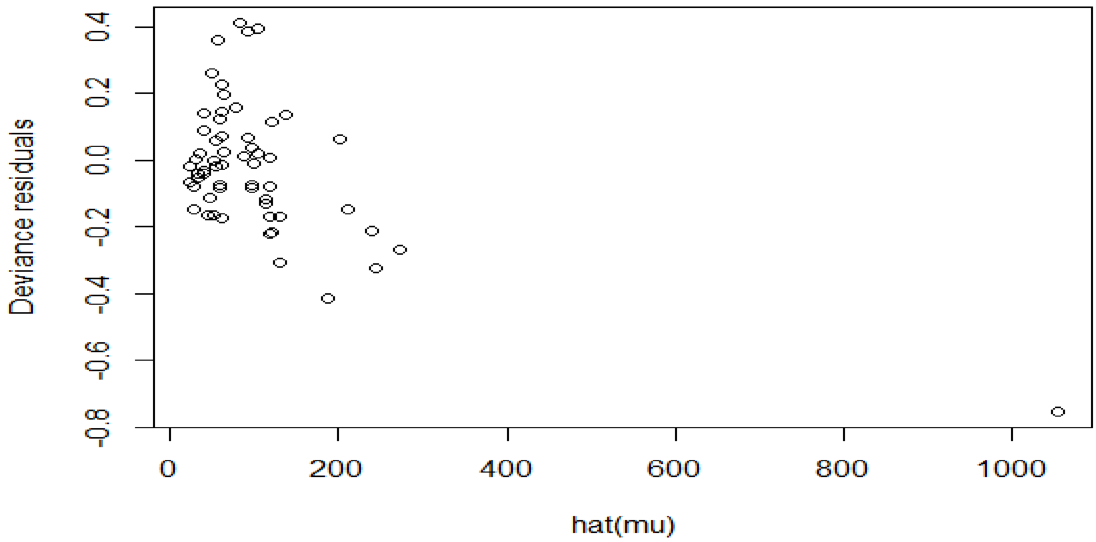
95% CI for β_0 = (2.384,2.664)
95% CI for β_1 = (-0.199,0.184)
95% CI for β_2 = (-0.054,0.226)
95% CI for β_3 = (-0.441,-0.218)
95% CI for β_4 = (-0.615,-0.367)
95% CI for β_5 = (0.101,0.136)
95% CI for β_6 = (0.012,0.023)

The confidence intervals for β_1 and β_2 contain 0; so the null hypotheses β_i=0,i=1,2 cannot be rejected.

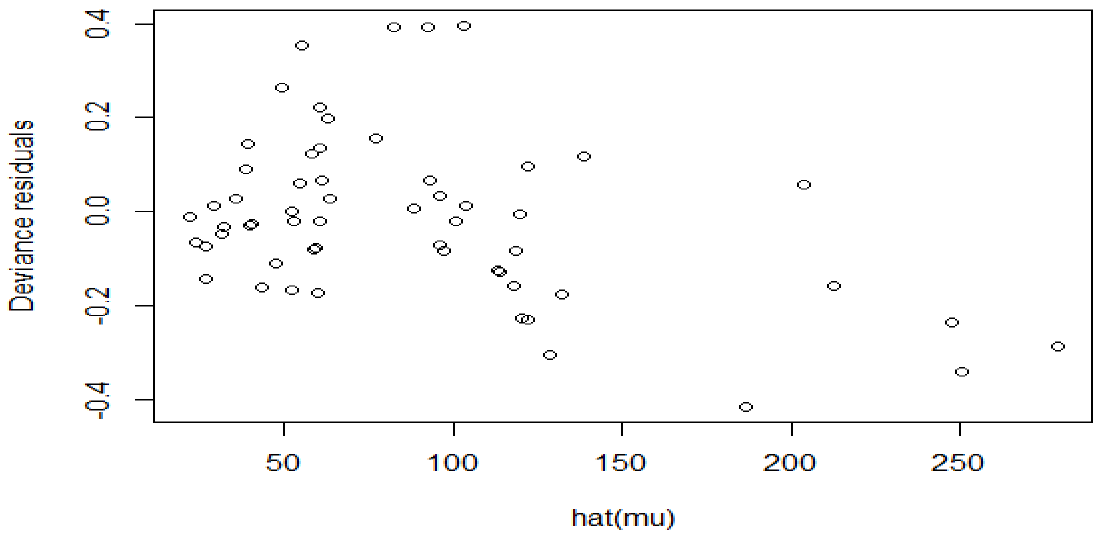
MODEL ASSESSMENT

Null Deviance : 0.289526 on 58 df
Residual Deviance : 0.022029 on 52 df
AIC : 476.55
The Residual Deviance is quite less than the Null Deviance, which means that most of the unexplained part of the null model can be explained by the fitted model. So we can assume that this is a fairly good fit.

We plot the fitted values against residuals to check if there are any irregularities in the model.



This is a very alarming residuals vs fitted plot. All the datapoints seem to cluster towards the top left corner. However on closer inspection we can see that 11th datapoint has an enormous fitted value and is affecting the entire plot. We now fit the model after omitting the 11th datapoint. Now we observe the residuals vs fitted plot for the new model.



This new plot, is a big improvement over the previous plot; although there is some signs of non-constant variance in this graph as well, indicated by a sort of funnel shape. We now check whether this new model has improved in terms of the other measures.

FITTING THE NEW MODEL

The model that we are fitting is

$\eta = \log(\mu_y) = \beta_0 + \beta_1 I_{CassDish} + \beta_2 I_{Dish} + \beta_3 I_{Plate} + \beta_4 I_{Tray} + \beta_5(\text{Diam}) + \beta_6(\text{Time})$

Where $\beta_0 = 2.507$, $\beta_1 = -0.028$, $\beta_2 = 0.077$, $\beta_3 = -0.331$, $\beta_4 = -0.495$, $\beta_5 = 0.118$, $\beta_6 = 0.019$

The Pearson Estimate of the Dispersion Parameter is given by $\Phi = 0.000396$

Estimate of μ= mean of Price= $\hat{\mu} = 83.39$

Estimate of $\lambda = \hat{\lambda} = (E[Y^{-1}] - \hat{\mu}^{-1})^{-1} = 212.76$

ASSESSING THE NEW MODEL

Null Deviance: 0.273153 on 57 df

Residual Deviance: 0.019806 on 51 df

AIC: 459.75

The difference between Null and Residual Deviance is larger than the previous model.

Also, the AIC value is less than the previous model.

So, we can conclude that the new model gives a better fit than the previous model.