Regression Analysis in R

2/24/2020

Background of the data set

The yacht data set gives six variables that could potentially be used to predict residuary resistant (ResidResis) of sailing yachts. Knowledge of a ship's residuary resistance can be useful to determine the ship's performance and to estimate the required propulsive power. The six variables are longitudinal position of the center of buoyancy (LongiPos), prismatic coefficient (PrisCoef), length-displacement ratio (LDRatio), beam-draught ratio (BDRatio), length-bean ratio (LBRatio), and Froude number (Froude), all adimensional.

Initial Interpretations

We re-code LongiPos as a categorical variable with five levels: position1=-5.0, position2=-2.4, position3=-2.3, position4=-2.2, and position5=0.0. The proportion of observations with position 1,2,3,4, and 5 are 0.182, 0.137, 0.453 ,0.046, and 0.182, respectively. The sample mean for the remaining five variables and ResidResis are 0.564, 4.790, 3.936, 3.208, 0.275, and 10.378, respectively. Their corresponding sample standard deviations are 0.023, 0.252, 0.549, 0.247, 0.101, and 15.043. We note that the last observation has a missing value for ResidResis. Ideally, we could attempt to find the missing value. Perhaps the value was mistakenly left out. We could delete the missing observation, impute the missing value by using the rest of the data to predict this value, use the means and standard deviations of pairs of predictor and response variables to estimate the missing value, or use maximum likelihood methods (such as the EM algorithm). As only one observation contains a missing value, we will remove it from the data set and perform the following analysis using the remaining observations. Taking LongiPosT (transformed) as our last listed variable in the data set, our initial model is:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \beta_6 X_{i6} + \beta_7 X_{i7} + \beta_8 X_{i8} + \beta_9 X_{i9} + \epsilon_i$$

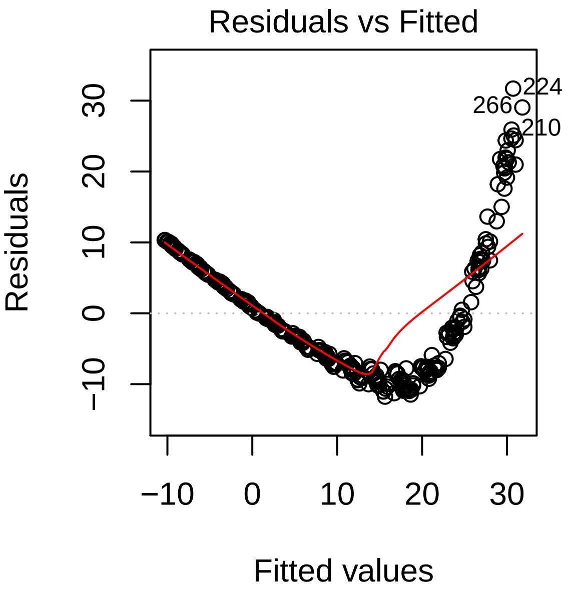
where X_1 through X_5 =PrisCoef, LDRatio, BDRatio, LBRatio, and Froude, respectively and

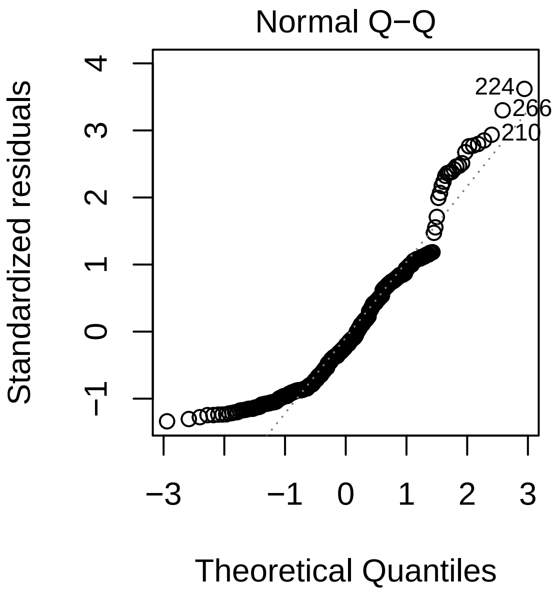
X_6=1 if position2, else 0 X_7=1 if position3, else 0 X_8=1 if position4, else 0 X_9=1 if position5, else 0

where $\beta_0, \dots, \beta_{p-1}$ are the parameters, $X_{i1}, \dots, X_{i,p-1}$ are known constants, ϵ_i are independent $N(0, \sigma^2)$ for $i = 1, \dots, n$.

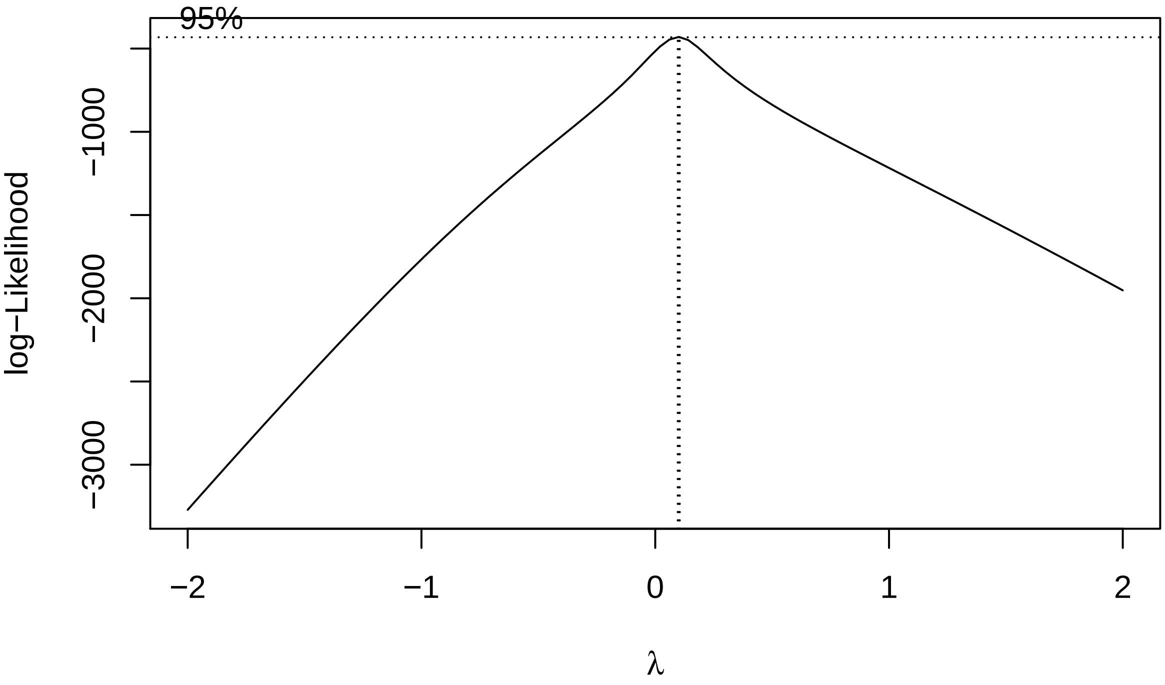
Initial model selection

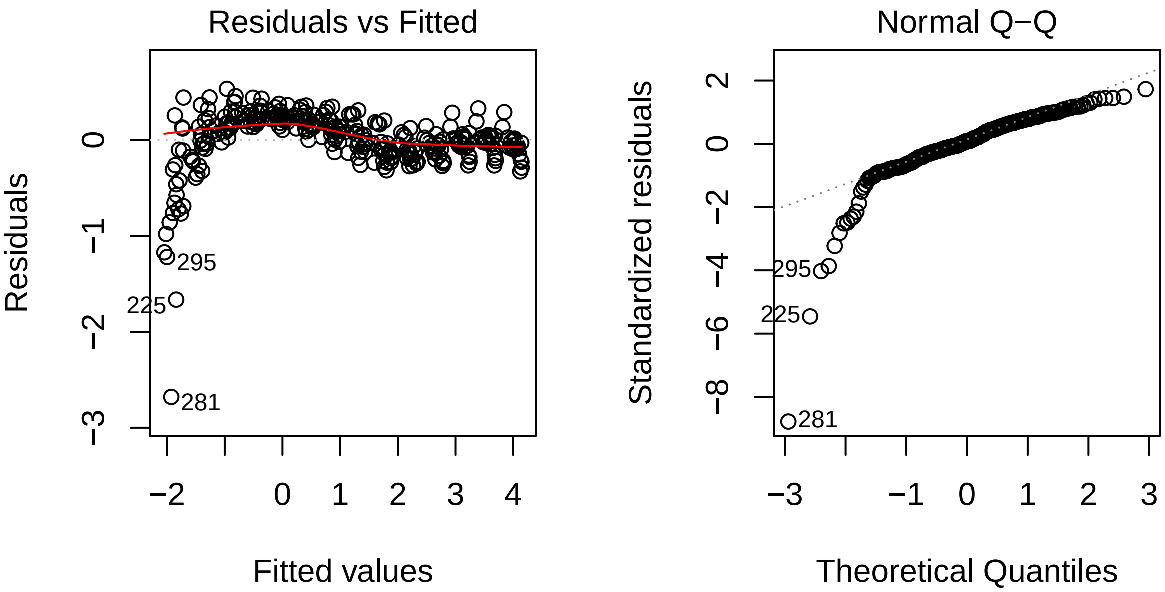
We start with a first-order regression model with all predictors as a starting point and determine if a variable transformation in X, Y, or both is necessary. From the scatter plot matrix (not shown), none of the predictors appear to have a linear regression with ResidResis.





The residual vs fitted values plot above shows that curvature and nonconstancy of error variance are apparent. The residuals are not randomly scattered about the horizontal and they do not roughly form a horizontal band around the zero line. The normal probability plot of residuals also indicates that the errors are not normally distributed as they do not form a line. Therefore we should perform a transformation on Y. As the distribution of the error terms are not close to a normal distribution and do not have a somewhat constant variance, a transformation on X is may not needed. Since the residual vs fitted values plot does not clearly match a prototype regression pattern in the literature, we will perform a box-cox transformation on ResidResis and note that all of the values for the response are strictly positive.





The boxcox procedure indicates that $\lambda = 0.1010101$ is the appropriate parameter. Since the likelihood estimate of λ is subject to sampling variability and SSE is usually stable in the neighborhood around the estimate, λ is often rounded to the nearest half. After the transformation using $\lambda = 0$ (or $Y' = log_e Y$), the residuals still are not randomly scattered around the line zero. In fact, they appear to follow a pattern, suggesting that a linear relationship does not exist between the predictor and response variables. The residuals do almost form a constant horizontal band around the zero line. In the normal probability plot, the residuals appear to be somewhat normally distributed with the presence of a few outliers on the left side of the graph and have a heavy left tail.

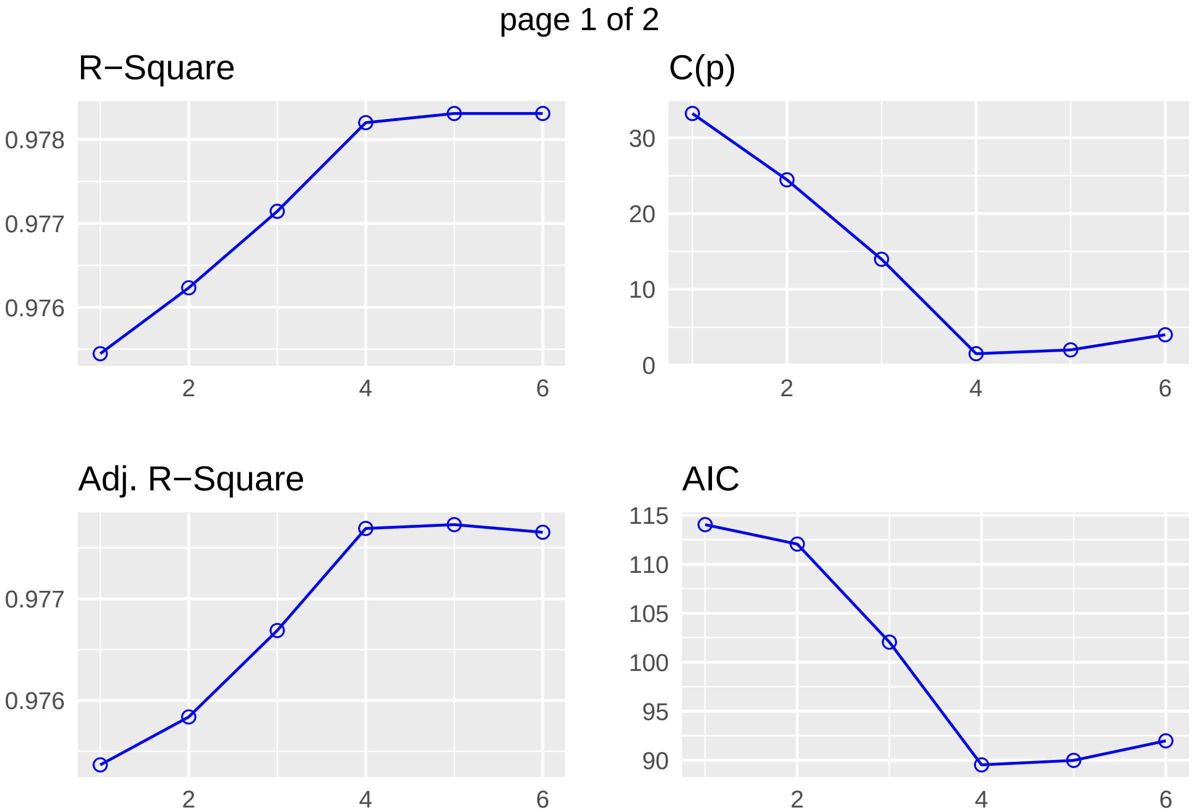
Now we consider a transformation on X. We use the Box-Tidwell power transformations to determine which λ to use for each predictor and note that all values for the predictors are strictly positive.

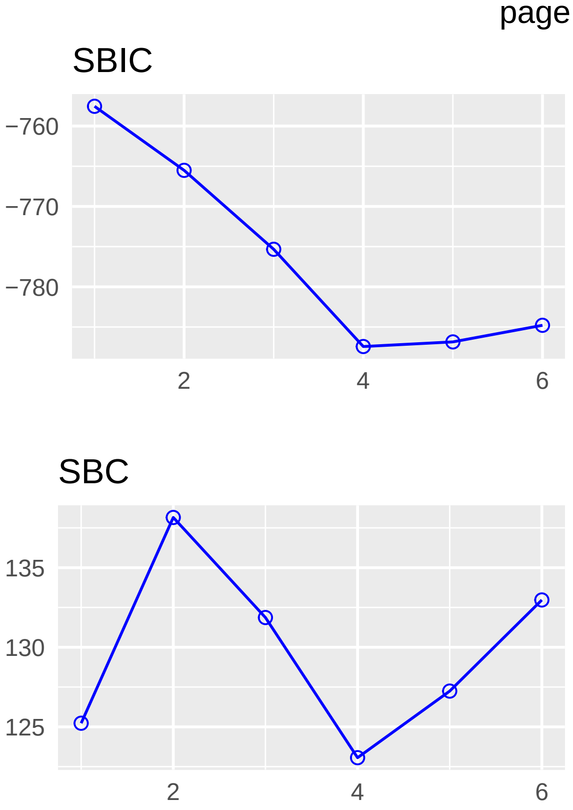
```
##           MLE of lambda Score Statistic (z)  Pr(>|z|)
## LDRatio      -0.48443             -0.5143    0.6070
## BDRatio      -0.38700              0.8112    0.4173
## LBRatio       1.38298             -0.2329    0.8159
## Froude        0.51784             -8.2056 2.294e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## iterations =  3
```

Note that the variable PrisCoef was not included in this analysis as R returned an error when it was included in the analysis with other variables. When each variable was analyzed by itself, only PrisCoef presented an error message. It appears that the only difference between PrisCoef and the other numeric variables is that the values of PrisCoef have a small range (from 0.530 to 0.600) while the other variables have larger ranges. Therefore the error code could be due to a bug in the boxTidwell itself. We then only transform the four predictors listed above, rounding each λ to the nearest half and get our final, transformed model.

Now we determine the best regression model to predict ResidResis from the six variables, using different model-building criteria and automatic selection procedures. As the number of predictors is small, we use a best subsets algorithm and note that there are $2^6 = 64$ possible models. To pick the best model, we want to maximize the $R_p^2 = 1 - \frac{SSE_p}{SSTO}$ and $R_{a,p}^2 = 1 - (\frac{n-1}{n-p})\frac{SSE_p}{SSTO}$ criteria, identify subsets for which Mallow's $C_p = \frac{SSE_p}{MSE(X_1,\ldots,X_{P-1})} - (n - 2p)$ is small and near p, minimize the $AIC_p = n \ln SSE_p - n \ln n + 2p$ and $SBC_p = n \ln SSE_p - n \ln n + [\ln n]p$ criteria, and minimize the $PRESS_p = \sum_{i=1}^{n}(Y_i - \hat{Y}_{i(i)})^2$. The package olsrr uses a variety criteria to identify the best model but does not include the $PRESS_p$ criterion.

```
##                        Best Subsets Regression
## --------------------------------------------------------------------
## Model Index    Predictors
## --------------------------------------------------------------------
##      1         FroudeT
##      2         FroudeT LongiPos2
##      3         BDRatioT FroudeT LongiPos2
##      4         PrisCoef BDRatioT FroudeT LongiPos2
##      5         PrisCoef BDRatioT LBRatioT FroudeT LongiPos2
##      6         PrisCoef LDRatioT BDRatioT LBRatioT FroudeT LongiPos2
## --------------------------------------------------------------------
##
##                                                    Subsets Regression Summary
## ----------------------------------------------------------------------------------------------------
##                        Adj.        Pred
## Model    R-Square    R-Square    R-Square     C(p)        AIC         SBIC         SBC         MSEP
## ----------------------------------------------------------------------------------------------------
##   1        0.9754      0.9754      0.9749    33.2127    114.0468    -757.5441    125.2274    0.0843
##   2        0.9762      0.9758      0.9752    24.4560    112.0641    -765.5108    138.1521    0.0830
##   3        0.9771      0.9767       0.976    13.9672    102.0516    -775.3299    131.8664    0.0803
##   4        0.9782      0.9777      0.9769     1.4988     89.5205    -787.4246    123.0621    0.0771
##   5        0.9783      0.9777      0.9768     2.0016     89.9768    -786.8498    127.2452    0.0773
##   6        0.9783      0.9777      0.9766     4.0000     91.9752    -784.7840    132.9705    0.0778
## ----------------------------------------------------------------------------------------------------
## AIC: Akaike Information Criteria
##  SBIC: Sawa's Bayesian Information Criteria
##  SBC: Schwarz Bayesian Criteria
##  MSEP: Estimated error of prediction, assuming multivariate normality
##  FPE: Final Prediction Error
##  HSP: Hocking's Sp
##  APC: Amemiya Prediction Criteria
```





The graphs indicate that the model with PrisCoef, BDRatio (transformed), Froude (transformed), and LongiPos2 (categorical) is the best regression model. We fit the data to this model (with ResidResis transformed) and conduct further analysis improve this model. Our transformed model is now:

$Y_i' = \beta_0 + \beta_1 X_{i1} + \beta_3 X_{i3}' + \beta_5 X_{i5}' + \beta_6 X_{i6}' + \beta_7 X_{i7}' + \beta_8 X_{i8}' + \beta_9 X_{i9}' + \epsilon_i$

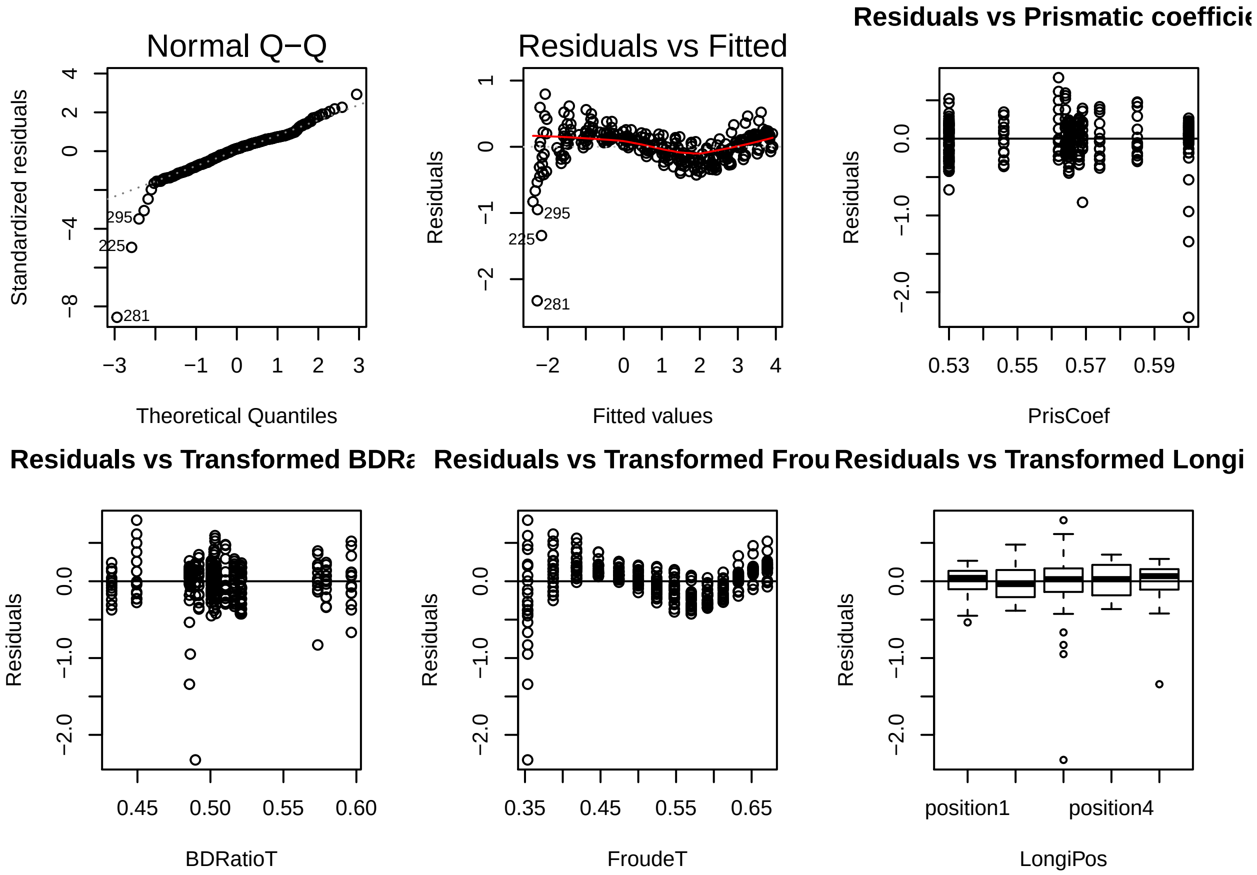
Model Improvement

We conduct a correlation test for normality. The coefficient of correlation between the ordered residuals and their expected values under normality is 0.9138386. For $\alpha = 0.05$, we find from Table B.6 that the critical value for n=307 is greater than 0.987. As the observed value is smaller, this implies that the distribution of the error terms may not be normally distributed. As we cannot entirely confirm that the the error terms are normally distributed, we use the Brown-Forsythe Test to test for constancy of error variance for $\alpha = 0.05$ instead of the Breusch-Pagan test as the Brown-Forsythe test is a non-parametric test that is robust against serious departures from normality while the Breusch-Pagan test assumes that the error terms are normally distributed. Group 1 consists of the first 153 observations and the remaining cases are placed into group 2. We have:

H_o: the variance is constant vs H_a: the variance is not constant

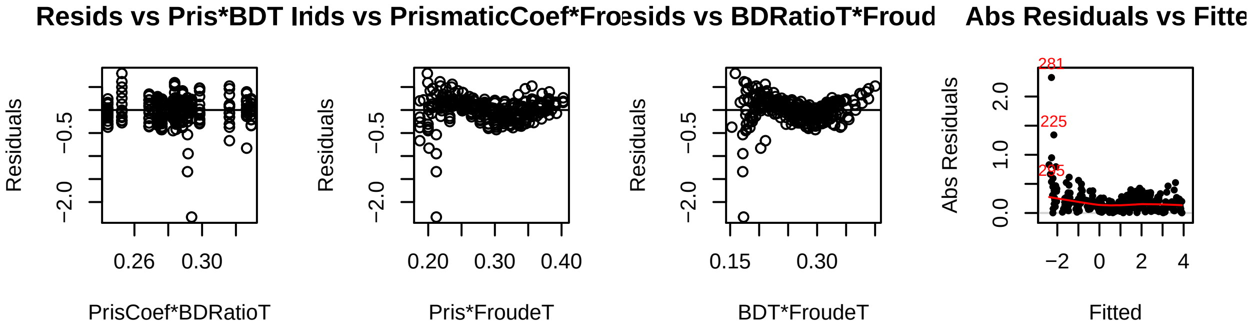
If $|t^*| \leq t(1 - \alpha/2, n1 + n2 - 2)$, conclude the error variance is constant, else conclude H_a

As $t^* = 1.1100$ is less than 1.9678, we conclude that the error variance is constant.



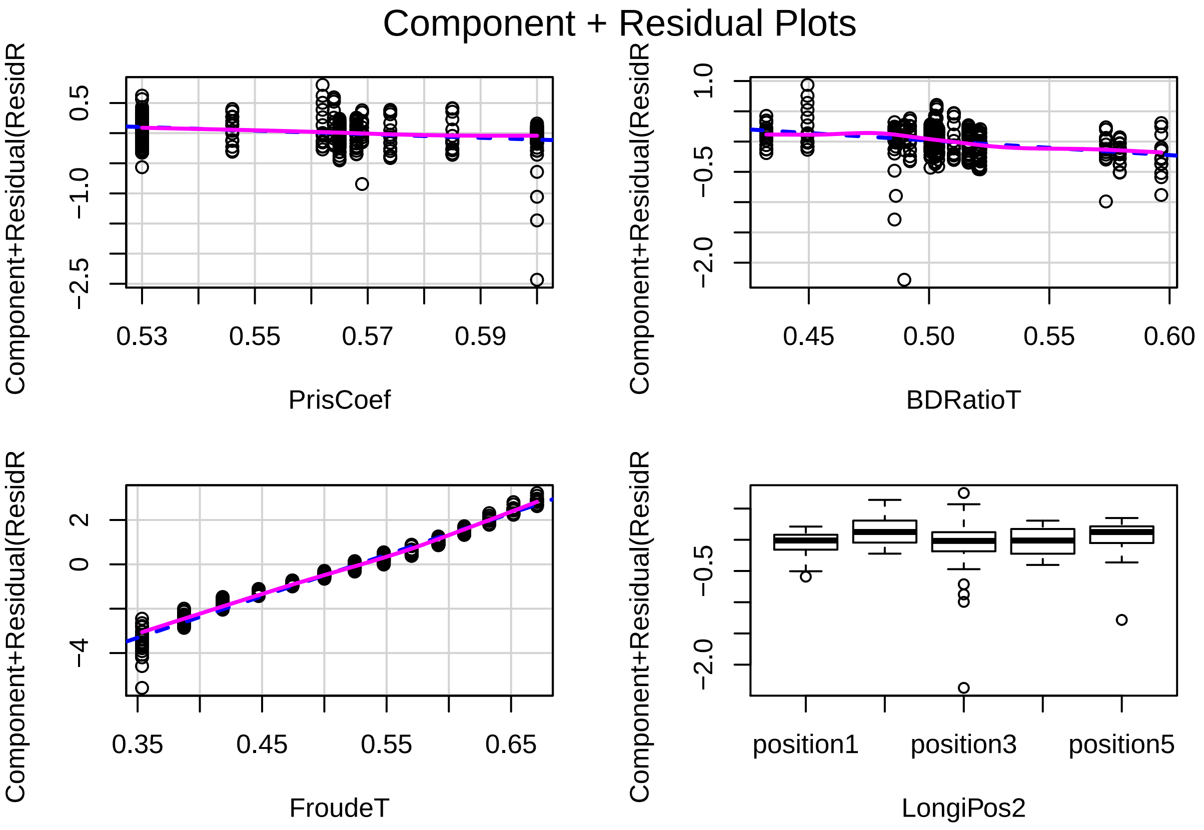
The normal probability plot above suggests that the errors may be normally distributed but note a heavy left tail with a few outliers. In the residuals vs fitted plot, the residuals do not appear to be randomly scattered around the zero line, implying that the relationship between ResidResis and the predictors is not linear. However, they do from a slight horizontal band. There are a few outliers on the left side of the graph as well. From the residuals vs predictor plots, only the residual vs FroudeT plot suggests that there may be a linear relationship with ResidResis. However, we note a somewhat curvature relationship suggested by the graph.

The PrisCoef and BDRatioT graphs in general form a constant band but are not randomly scattered about the horizontal.



Note that in the residuals vs interaction plots above, only PrisCoef(BDRatioT) appears to be symmetric about the horizontal with the exception of a few outliers while the other graphs fail to have a clear, symmetric pattern; interaction effects for PrisCoef(BDRatioT) are not present. The plot of absolute residuals against ResidResis has a few outliers but suggests that the errors have constant variance.

In the added variable (partial regression) plot for FroudeT when the other predictors are already in the model, implies that a linear term in FroudeT may be useful. As the remaining plots form almost a horizontal band about the line zero, the corresponding predictors contain no additional information useful for predicting ResiResis when the other predictors are already in the model.



Last, we conduct global and partial F-tests. For testing: $H_0 : \beta_1 = \cdots = \beta_7$ vs H_a : not all β_k equal zero for $k = 1, 2 \ldots, 7$, the test statistic is $F^* = \frac{MSR}{MSE}$, and if $F^* \leq F(1-\alpha; p-1, n-p)$, conclude H_0. At 5% significance level, we conclude H_a as $1917 > 2.0403$, that there is a regression relation between at least one of the predictors and ResidResis.

For testing $H_0 : \beta_k = 0$ vs $H_a : \beta_k \neq 0$, the test statistic is $t^* = \frac{b_k}{s(b_k)}$, and if $|t^*| \leq t(1-\alpha/2; n-p)$, conclude H_0. At 5% significance level, we conclude that all of the predictors besides LongiPost for position 3 and 4 have a linear relation with ResidResis.

The residual vs predictors plot indicated that FroudeT a linear relationship may exist with ResidResis but the graph had a slight curvature effect. The interaction plots implied that only the interaction term PrisCoef(BDRatioT) is likely useful. The added variable plots were very clear in indicating that adding FroudeT to the model when other predictors are already present would provide helpful information and that a linear relationship likely exists with ResidResis. Although the partial F-tests suggested that other predictors as well as FroudeT may have a linear relation with the response, FroudeT had the greatest test statistic. Therefore we include (PrisCoef*BDRatioT) as an interaction term, their lower terms, and FroudeT in our model: $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \beta_4 X_{i4} \epsilon_i$.

Further Model Diagnostics

We now identify outlying Y observations using studentized deleted residuals. Outlying Y observations are those with large absolute studentized deleted residuals. We use the Bonferroni simultaneous test procedure with a family significance level of $\alpha = .10$. We require $t(1-\alpha/2n; n-p-1) = t(.999919; 307-5-1) = 3.820$ and conclude that an observation is not an outlier if $|t_i| \leq 3.820$. We find that observations 225 and 281 are outlying Y observations.

We now identify outlying X observations using leverage values. Note that leverage values greater than $\frac{2p}{n} = 0.0325$ are considered outlying cases. We note that observations 15, 71, 72 84, and 211 through 224 have outlying observations in terms of their X values.

Now we determine if these outlying cases in both Y and X are influential using DFFITS, Cook's Distance, and DFBETA measures. A case is influential if the absolute value of DFFITS exceeds $2\sqrt{p/n} = 0.2552$, if the absolute value of DFBETAS is greater than $2/\sqrt{n} = 0.1141$, or if Cook's distance is greater than $\frac{4}{n-k-1} = 0.0133$. The observations that appear to be influential using all three measures are 15, 211, 223, 224, 225, and 281. We now determine if multicolinearity is present in the data. The largest variance inflation factor is 923.17 (for the interaction term), which is well above 1, indicating multicolinearity may be influencing the least squares estimates. We also note that adding or removing predictors to our model changes the regression coefficients. When the PrisCoef(BDRatioT) interaction term is removed from the model, all VIF are near zero. However, our previous analysis indicated that PrisCoef and BDRatioT were not good predictors for ResidResis. Therefore if we remove the interaction term wel also remove its lower orders, leaving only FourdeT. Our final model is $Y' = \beta_0 + \beta_1 X' + \epsilon_i$.

Final Model

Our fitted regression model is $\hat{Y}' = -8.794 + 18.734X'$ For every one unitless increase in the transformed Froude number, ResidResis increases by 18.734. This is reasonable as the Froude number is a ratio of two positive values. A Froude number greater than 1 indicates a fast rapid flow, which would cause a greater residuary resistance per unit weight of displacement. As the scope of the model does not cover X=0, β_0 has no particular meaning as a separate term in the model.

Association of residuary resistance and Froude number

Let's test to see if there is association between ResidResis and FroudeT was the same across levels of LongiPos2. To do this, we will conduct a partial F test. The full model is $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + + \beta_4 X_{i4} + \beta_5 X_{i5} + \beta_6 X_{i2} X_{i1} + \beta_7 X_{i3} X_{i1} + \beta_8 X_{i4} X_{i1} + \beta_9 X_{i5} X_{i1} + \epsilon_i$, where the interaction terms are the interactions between FroudeT and each level of LongiPos2. The reduced model is our final model in part E.

We test $H_0 : \beta_6 = \beta_7 = \beta_8 = \beta_9$ vs H_a: at least one is different. If $F^* \leq F(1 - .05; 1, n - 4)$, conclude H_0. At 5% significance level, we have $3.9095 > 3.8723$ and conclude that there is an interaction effect and so the association between ResidResis and Froude number was not the same across every level of LongiPos2.

Conclusion

From our initial model, we transformed the data both in Y and X using Box-Cox and Box-Tidwell transformations. A best subsets algorithm determined that a model with PrisCoef, BDRatioT, FroudeT, and LongiPos2 predictors was the best. We further improved on the model and determined that the transformed Froude numbers was the best variable to predict residuary resistance. We also concluded that the association between residuary resistance and Froude number was not the same across levels of LongiPos.