Regression Techniques Project Report
CALCOFI DATA

Archi De

Diganta Bhattacharya

10 January 2021

List of Contents

1 Abstract **1**

2 The Data **1**

2.1 Introduction . . . 1

2.2 Collection of the Data . . . 1

2.3 Variables in the Data . . . 1

3 Methodology **2**

3.1 Cleaning the Data . . . 2

3.2 Model Selection . . . 2

3.3 Basic Linear Model and Assumptions . . . 3

3.4 Variable Selection . . . 3

3.4.1 Mallow's C_p . . . 3

3.4.2 General Methods for selection . . . 3

3.4.3 Penalized Regression . . . 4

3.4.4 Cases of penalised regression . . . 4

3.4.5 Result of Variable Selection . . . 5

3.5 Non-Parametric Regression . . . 5

3.5.1 Intuition for this approach . . . 5

3.5.2 Results . . . 5

3.6 Quantile Regression . . . 5

3.6.1 Intuition for this approach . . . 5

3.6.2 Fitting of the model . . . 5

3.6.3 What we found . . . 6

3.7 Checking Goodness of fit . . . 6

3.8 Measure of Influence . . . 7

3.8.1 Checking influence of the data points . . . 7

4 Data Analysis and Discussion **8**

5 Conclusion and Recommendations **9**

5.1 Conclusion . . . 9

6 References **10**

1 Abstract

We work with the CALCOFI data, which contains various details regarding the the study of the marine environment off the coast of California. Applying various techniques of data analysis we want to address the following questions from the data set such as how does salinity vary with change in the different variables measured and if it is possible to estimate salinity satisfactorily based on the different predictors taken into account. We also aim to find which predictors are the most important ones in the environment for predicting salinity and find an accurate model based on them for prediction. Lastly we check the goodness of fit of the models discussed.

2 The Data

2.1 Introduction

The California Cooperative Oceanic Fisheries Investigations (CalCOFI) conducts quarterly cruises off southern and central California, collecting a suite of hydrographic and biological data on station and underway. The organization was formed in 1949 to study the ecological aspects of the sardine population collapse off California. Today our focus has shifted to the study of the marine environment off the coast of California, the management of its living resources, and monitoring climate change. The physical, chemical, and biological data quickly became valuable for documenting climatic cycles in the California Current.

CalCOFI research drew world attention to the response to the dramatic Pacific-warming event in 1957-58 and introduced the term “El Niño” into the scientific literature.

2.2 Collection of the Data

The data collection systems, typically employ three classes of sensors:

1. Oceanographic sensors: sensors that measure properties of the ocean surface.
2. Meteorological sensors: sensors that measure properties of the air.
3. Navigational sensors: sensors that report what the ship is doing which collects the data.

These sensors report their information to a central computer which compiles the data and appends it to a daily log.

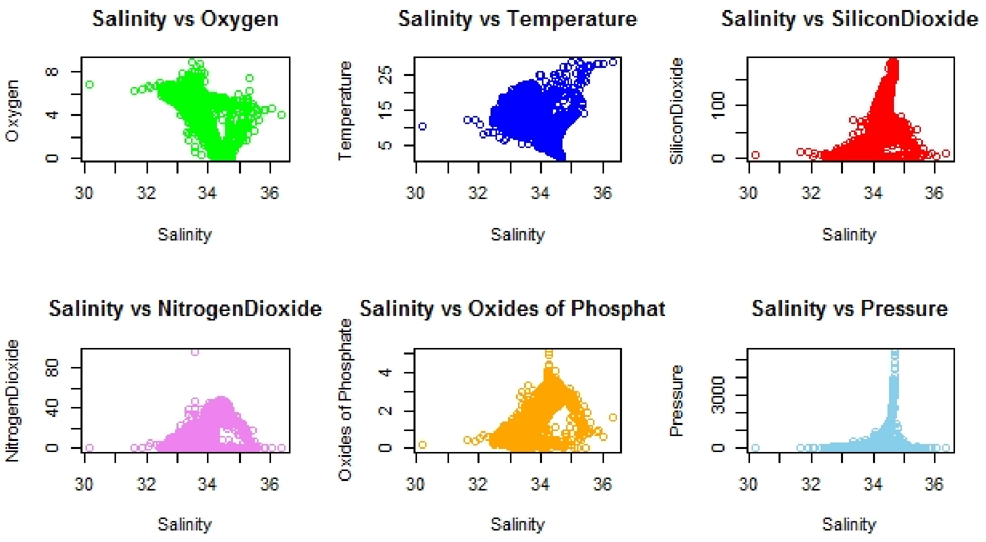
2.3 Variables in the Data

The original data had 306,533 values and 21 attributes. We have listed the attributes except the ones corresponding to the details of the ship and station on which the log was taken.

Categorical Attributes in Data	Name
Record indicator	recind
Temperature units of precision	tprec
Salinity units of precision	sprec

Attributes Collected in the Data	
Temperature of the water measured in degree Celcius	Reported Silicate Concentration
Reported depth in meters	Pressure measured in decibars
Reported potential density of water	Reported specific volume anomaly
Reported salinity	Reported potential temperature
Reported phosphate concentration	Reported dynamic height
Reported nitrite concentration	Reported oxygen saturation
Reported nitrate concentration	Potential density of water

Given below are the six plots of salinity against other significant continuous predictors ⟶



3 Methodology

3.1 Cleaning the Data

In the previous section, we have already mentioned the relevant variables of the data set. Some other columns in the original data correspond to the details of the names of stations and the ships from which the data was reported for ease of maintaining logs. Other columns indicated the precision of the instruments used for measurements. We omitted all those columns which could not be utilized in our analysis. We also changed all numeric character/string values to int to avoid any computational errors.

3.2 Model Selection

Salinity of water is an important factor for the survival of the flora and fauna in an aquatic environment. With climate change various attributes such as temperature and chemical concentrations are altering in these habitats. Therefore we want to model the impact observed on salinity of seawater. We want to answer:

What variables does Salinity depend on and how good is our Prediction?

We will use Basic Linear Models to answer this question. We will try to answer some more questions. We will use Non-parametric regression and Quantile regression to make better conclusions.

Quantile regression will help us in checking the fit of salinity of water. As we will see from the predicted linear model salinity of water, gives a pretty good linear fit except at top and bottom quantiles of some attributes . We want to check:

How does vary with the of the main attributes like oxygen saturation, sigma and other attributes?

3.3 Basic Linear Model and Assumptions

$Y = X\beta + \epsilon$ and, $\hat{\beta} = (X^TX)^{-1}X^TY \quad y \in \mathbb{R}^n, \beta \in \mathbb{R}^p, X \in \mathbb{R}^{nxp}$ With the standard model given above, several assumptions are made about the data and model that are not necessarily true for this data. The most common assumptions are:

- The errors are normally distributed.
- $E[\epsilon] = 0.$
- $V[\epsilon] = \sigma^2.$
- $Cov[\epsilon_i, \epsilon_j] = 0 \; \forall \; i \neq j.$
- Linearity of predictors.
- No multi-co-linearity.

So, in order to fit Linear model in the above data we checked the assumptions by the methods described:

1. Linearity of predictors can seen from plots in the following slides.
2. Q-Q Plot for residuals implied normality such that $E[\epsilon] = 0, V[\epsilon] = \sigma^2.$
3. VIF and Condition Index Matrix provided evidence for low multi-co-linearity.
4. We finally presented total correlation and partial correlations of the selected variables.

3.4 Variable Selection

Now we discuss the various methods used to choose the most important variables for predicting salinity.

3.4.1 Mallow's C_p

This method estimates the standardized total mean square of estimation for the partial model with the following formula and compares it's value with p:

$$C_p = \frac{SSE_p}{MSE_{all}} + 2(p+1) - n$$

3.4.2 General Methods for selection

Forward Selection
Backward Selection
AIC & BIC

3.4.3 Penalized Regression

Three types of penalized regression was fitted to extract the most important predictors.

1. For LASSO, the following is minimised.

$$(Y - X\beta)^T(Y - X\beta) + \lambda|\beta|_1 \quad \text{with,} \quad |\beta|_1 = \sum_{j=1}^{p} |\beta|_j$$

   The aim of LASSO is to add a penalty for each non zero coefficient kept, hence dropping the less essential attributes to extract the ones with most impact.

2. Ridge regression adds the L_2 penalty such that we have:

$$(Y - X\beta)^T(Y - X\beta) + \lambda\beta^T\beta \quad \text{where,} \quad \hat{\beta}_{ridge} = (X^TX + \lambda I)^{-1}X^TY$$

   As the value of λ increases, due to the penalty term, the coefficients shrink towards zero if they are corresponding to the less important attributes.

3. For elastic net, the following is minimised w.r.t β,

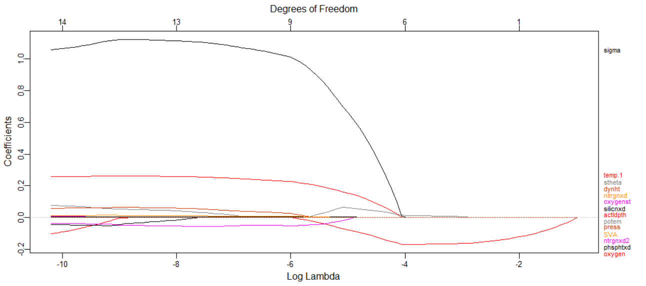
$$\sum i = 1^n\{y_i - \sum_{j=1}^{p} x_{ij}\beta_j\}^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2$$

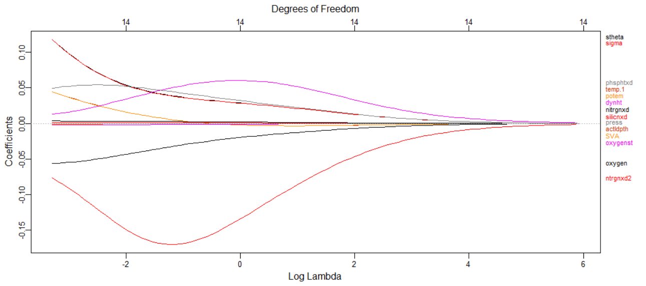
   This is a mixture of LASSO and ridge regression and can be used to extract the important predictors for the model.

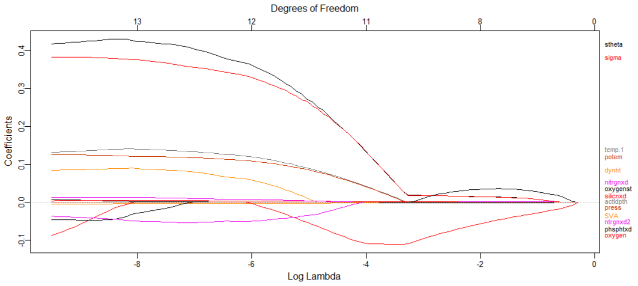
In all of the three described above, we chose the value of λ accordingly.

3.4.4 Cases of penalised regression

The following plot shows the values of coefficients of the predictors in the three types of penalised regression with increasing values of the penalty.







It can be noticed that LASSO works fastest for dropping the extra predictors followed by elastic net.

All these give the same 6 most important predictors, oxygen concentration being the most crucial one for predicting salinity.

3.4.5 Result of Variable Selection

The final model selected by the above methods is given by the following six predictors $\longrightarrow$
`Oxygen ,temperature ,dynamic height ,pressure ,density ,potem`
The model with above predictors give only a slight reduction in the R^2 value with $R^2 = 0.9757$ and dropped 12 unnecessary predictors.

3.5 Non-Parametric Regression

3.5.1 Intuition for this approach

Without assuming the distribution of the response variable which is salinity in our case, non parametric regression can be done. This is an effective method to detect if the relationship between the response and predictor variable is not linear.
It is not recommended to do non parametric regression for multiple predictors.
The Nadaraya Watson kernel estimate for a given kernel function $K_h()$ for bandwidth $h > 0$ is

$$\hat{m}_h = \frac{\sum K_h(x - x_i) y_i}{\sum K_h(x - x_i)}$$

3.5.2 Results

After viewing the predicted values by predicting with the six chosen predictors one by one, each show a linear trend except at the extreme quantiles for some of the fits.

3.6 Quantile Regression

3.6.1 Intuition for this approach

From the results of the non parametric regression, some deviation from the general trend was observed at the lower and upper quantiles. This motivated us to consider quantile regression to improve the prediction at the extreme quantiles.

3.6.2 Fitting of the model

Quantiles, such as the median $(p = 50\%)$, are robust to outliers. Quantile Regression Model Equation for the τ-th quantile is

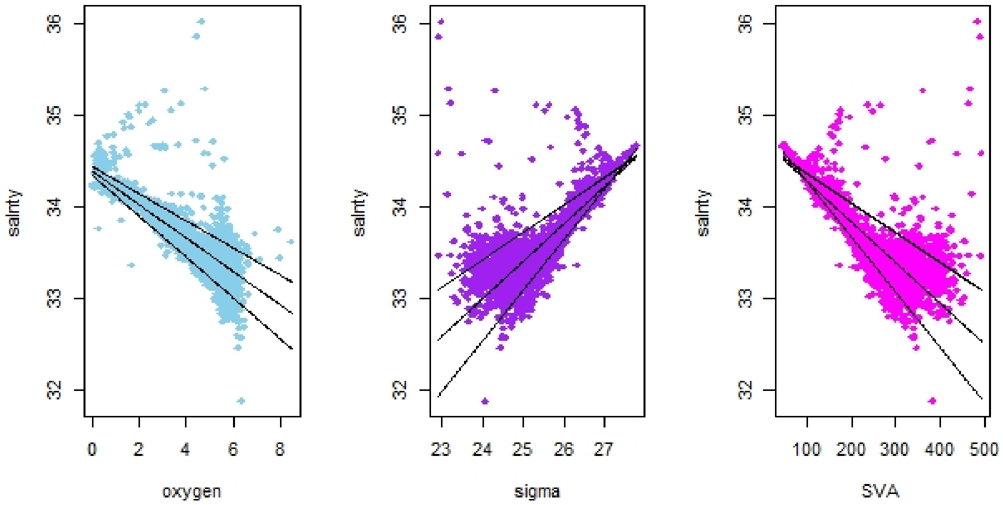
$$Q_\tau(y_i) = \beta_0(\tau) + \beta_1(\tau) x_{i1} + \cdots + \beta_p(\tau) x_{ip}$$

The estimates are found by the β for which we obtain

$$\min_{b \in \mathbb{R}^k} \sum_{i=1}^{n} \kappa_p \left(y_i - \mathbf{x}_i^\top \mathbf{b} \right)$$

where $\kappa_p(u) = u(p - I(u < 0))$, $0 < p < 1$.

3.6.3 What we found

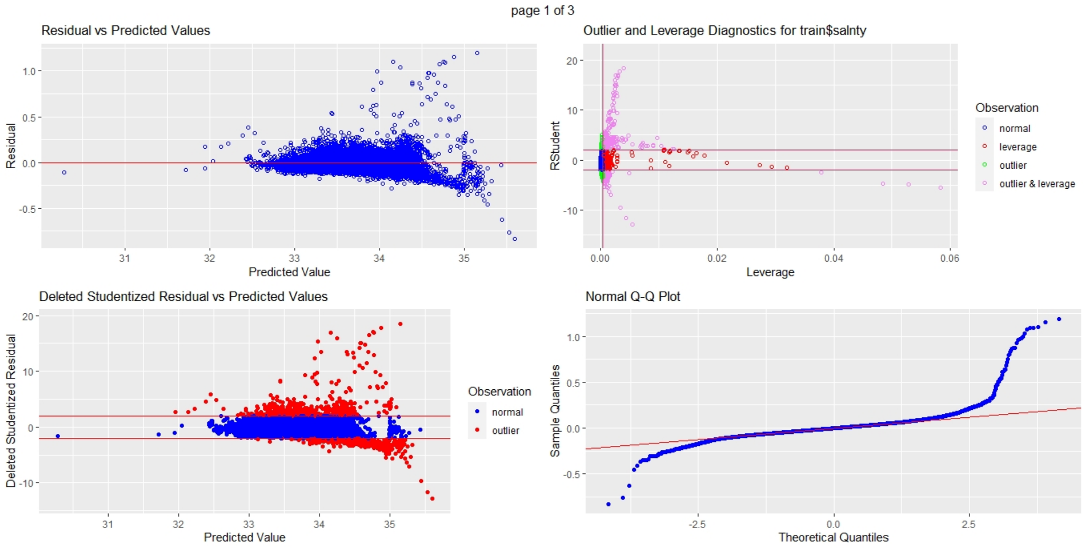


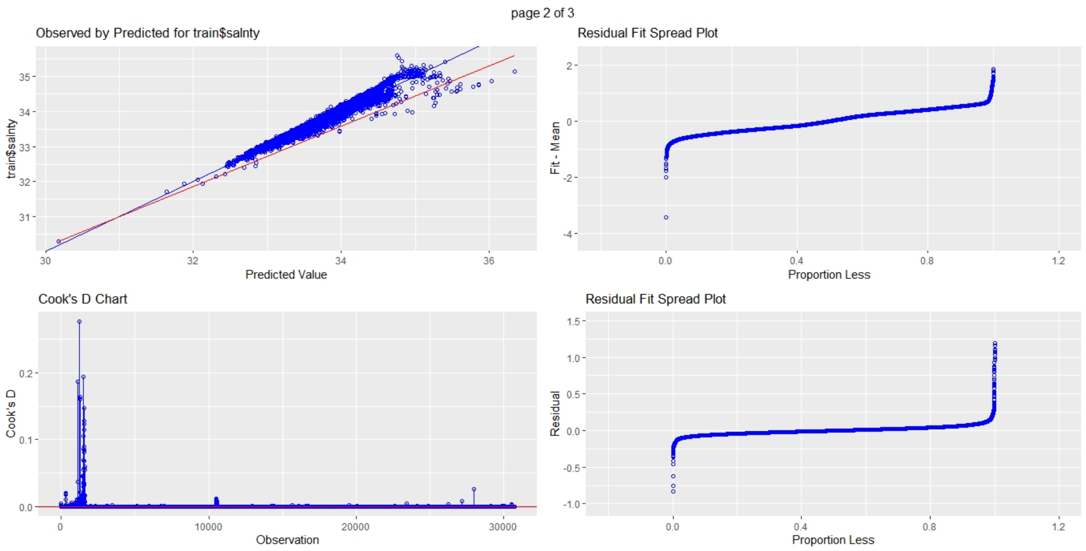
The fit for quantile regression is satisfactory when the single predictor oxygen concentration, sigma and sva is taken. This is in accordance with what we expected as these are the 3 most crucial predictors selected in variable selection.

3.7 Checking Goodness of fit

The various measures that we considered for checking goodness of fit of the chosen model are

- The multiple correlation coefficient which is given by $R^2 = 1 - \dfrac{SSE}{SST} = 0.9757.$
- The adjusted R^2 which is given by $R^2_{adj} = 1 - \dfrac{(n-1)}{(n-p-1)}\dfrac{SSE}{SST} = 0.9757$
- The PRESS R^2 statistic which is given by $R^2_{PRESS} = \sum_i \left(\dfrac{e_i}{1-h_{ii}}\right)^2 = 13.18$ where h_{ii} is the ith diagonal entry of $H = X(X^\top X)^{-1}X^\top$





Note that the Q-Q plot shows a straight line except at the extreme quantiles. This is also reflected in the previous sections which led to quantile regression giving better results.

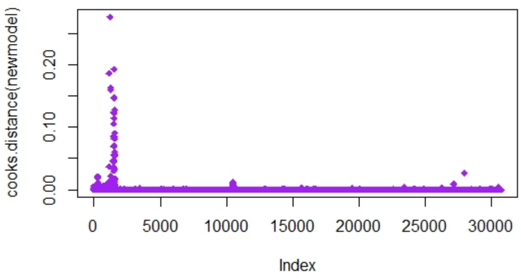
3.8 Measure of Influence

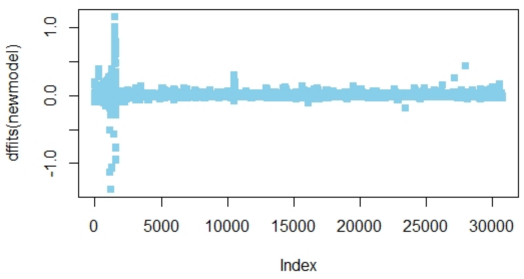
To check which of the data points from the data set have greater influence on the model, and can be possibly outliers, we calculated the following well known values for each of the data points.

1. **Cook's Distance:** $D_i = \frac{\sum_{j=1}^{n}(\hat{Y}_j - \hat{Y}_{j(i)})^2}{p \cdot MSE}$ where, $\hat{Y}_{j(i)}$ is the fitted value for the j observation without including the i-th observation in the data that will generate the model.
2. $DFFITS_i = \frac{\hat{y}_i - \hat{y}_{(i)}}{\sqrt{MSE_{(i)} h_{ii}}}$ and, an observation is deemed influential if the absolute value of its DFFITS value is greater than $2\sqrt{\frac{k+2}{n-k-2}}$.

3.8.1 Checking influence of the data points

The influence of the datapoints can be visualised from the given plots.





It can be seen that except for a few points there are not many outliers in the data set.

4 Data Analysis and Discussion

Why this Data?

This dataset collected in the aquatic habitat of the coast at Pacific was originally aimed to study the trend in Sardine population in the area. This data quickly became relevant in the study of how the different attributes in the data affect the aquatic habitat. Salinity of the habitat is a crucial factor in determining the flora and fauna in the ecosystem. This data can be used to study how salinity is affected by different attributes which can be measured by easier means.

What do we aim to answer?

It is to be seen that if we can indeed satisfactorily predict salinity with other attributes. If it is possible to predict we would be interested to know what of relation holds between the response and the predictors. If such a model fits well we can predict the trend of change in salinity if one of the predictors changes significantly. This is relevant in terms of environmental effects as temperature which of the crucial attributes which has increased in the recent years. It is also relevant to mention that the research on this data set introduced the term *"El Nino"* effect.

How did we proceed?

Our initial attempt was to first identify the crucial attributes among all the variables while the common variable selection methods like forward and backward selection and penalized regression. with the crucial attributes extracted we first fit a Linear Model to predict salinity which was sufficiently good. The residual analysis of this model showed little evidence of deviation from the standard Gauss-Markov setup except in case of some extreme quantiles. To verify if the assumption of linearity is justified Non-Parametric model was fit to view the general trend which was mostly Linear. Finally to address some inconsistency at the extreme quantiles we fit a quantile regression with the three most influential predictors. This fit was significantly accurate.

What did we find?

Among the 18 predictors initially present 6 were sufficient to predict salinity satisfactorily. The most important predictor was the Oxygen Concentration in water. Height, Temperature and Pressure were among the other important attributes affecting salinity. The relation between the response and the predictors was roughly linear.

How to interpret and use?

We can use the fitted model to estimate the salinity of a region in the aquatic habitat if we know the relevant predictors of the model. The effect of change in certain predictors on the aquatic habitat can be estimated from the predicted change in salinity.

5 Conclusion and Recommendations

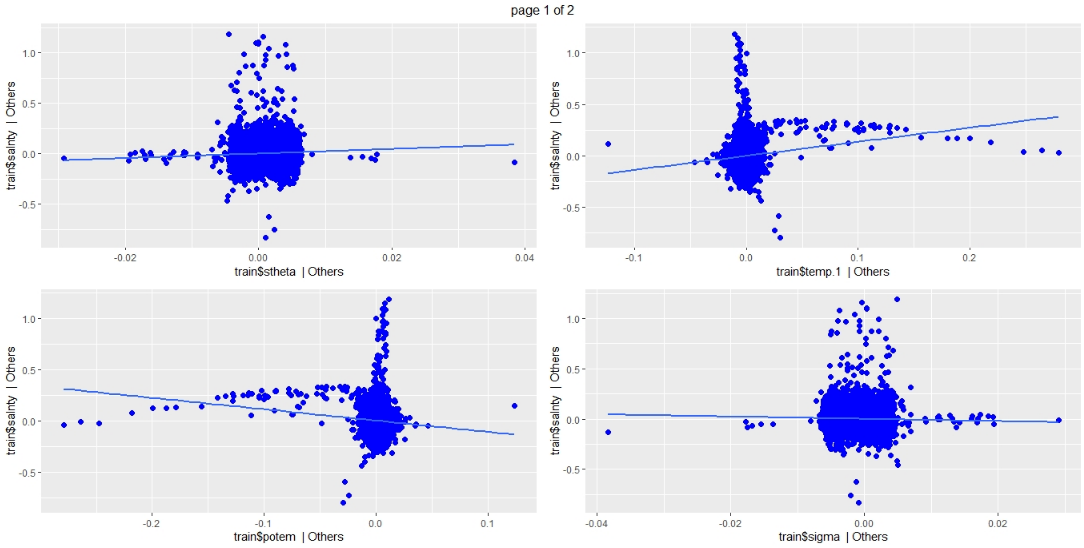
5.1 Conclusion

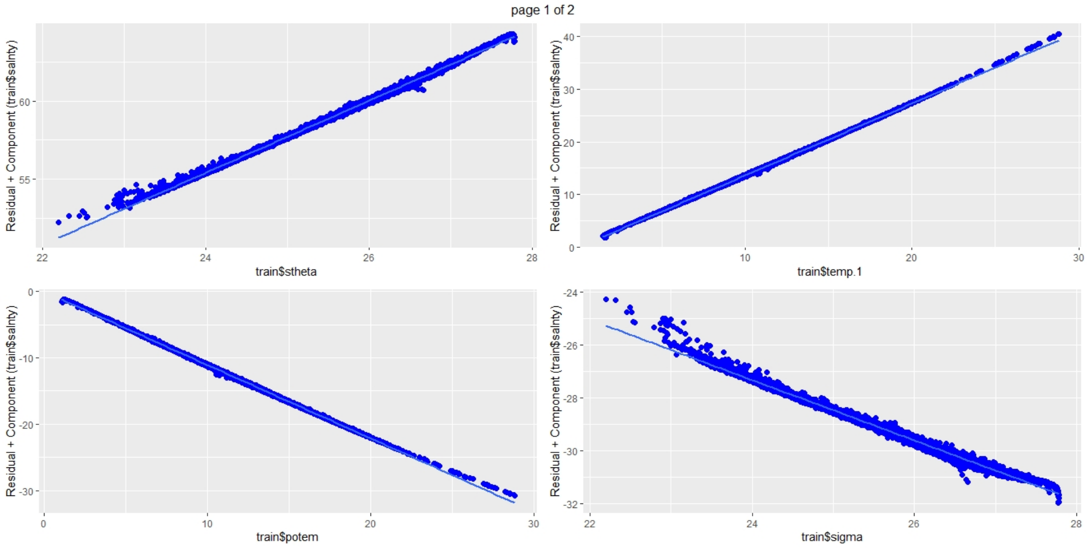
The final model was :

$$Y \sim \alpha + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \beta_3 \cdot X_3 + \beta_4 \cdot X_4 + \beta_5 \cdot X_5 + \beta_6 \cdot X_6$$

- Y is Salinity, the dependent variable.
- X_1 is Oxygen Saturation.
- X_2 is Potential Density of Water.
- X_3 is S_θ.
- X_4 is Dynamic Height.
- X_5 is Temperature.
- X_6 is Potential Temperature.

We have provided the Added Variable Plots and Residual Plus Component Plots for the final model.





6 References

- https://www.kaggle.com/sohier/calcofi

- Class Notes and Materials for R.

- https://bookdown.org/egarpor/PM-UC3M/glm-diagnostics.html

- https://www.r-bloggers.com/2019/01/quantile-regression-in-r-2/