Mahathi Gandhamaneni
Student Number: 1007008140
UTORid: gandhama

STA302: Final Project Report
Can we predict water temperature based on salinity in oceans?

Introduction

In this paper, I wish to explore the question of whether there is a linear relationship between water salinity & water temperature in oceans. In other words, can we predict the water temperature based on salinity in oceans? Being able to determine a linear relationship between water temperature and salinity could help us predict when water temperatures are getting warmer and thus, predict climate change and global warming (since increasing water temperatures are an indicator of climate change (Stefan & Sinokrot, 1993)).

In "Key Physical Variables in the Ocean: Temperature, Salinity, and Density", Pawlowicz notes a relationship between temperature, salinity and density of ocean water. A water sample with higher salinity will have greater mass, and will be more dense. The warmer the water, the more space it takes up, and the lower its density. This informs us that there is a relationship between temperature and salinity. I hope to incorporate this relationship into the model by considering density as well as salinity as predictor variables and carrying out a multivariable regression.

Methods

The dataset I chose was recorded by the California Cooperative Oceanic Fisheries Investigations (CalCOFI), a subset of which I found on Kaggle and decided to use for this analysis. The response variable we are going to use is Water temperature (T_degC). The predictor variables are Salinity (Salnty) and Depth in meters (Depthm). Since the dataset does not have a density variable, we will exploit the fact that density of water increases with depth (Webb, n.d.), and use the depth variable instead of density. The dataset was first randomly split into a training and testing set on a 70:30 split.

In order to arrive at our final model, two models built on the same training data were compared - a simple linear regression model with salinity as the predictor variable and temperature as the response, and a multiple linear regression model with salinity and depth as predictors and temperature as the response. After fitting the models, 95% confidence intervals were constructed for each coefficient. Hypothesis testing was performed to see if the values of the test statistics are plausible values of the coefficients or not. The ANOVA test was conducted to test the significance of the regression lines built by each model. In order to conduct diagnostic checking, the residuals were plotted against the fitted values and observed to determine any patterns.

Bad leverage points were determined using Cook's distance and residuals (which estimate the influence of a data point when performing regression). These points were then removed. The models were refit and the residuals vs fitted values plot was replotted. The normality and homoscedasticity of the errors was then assessed through normal Q-Q plots and standardized residual plots. Box-cox transformations were then performed on the response variable in order to obtain normality and remove homoscedasticity for each model.

In order to determine which model was the best to answer our question, the AIC, BIC, and adjusted R^2 values were calculated for each model and the one with the highest R^2 and lowest AIC and BIC values was selected. VIF values were calculated to check for multicollinearity. In order to check the accuracy of predictions on the selected model, the outcomes were predicted for the testing set and the mean squared error was assessed.

Results

Based on the boxplot for T_degC (Figure 1), its distribution appears to be skewed right. It also has a few outlier values after the right whisker ends. Since this data was recorded in California, these values are in accordance with its temperatures. However, the generalizability of observations drawn from this data to all oceans around the world may be a little low, since many oceans may have much lower temperatures than that recorded here. Based on the boxplot for Salnty (Figure 1), its distribution appears to be symmetric, maybe very slightly skewed left. It has a few outliers before the left whisker, and a few more after the right whisker. Based on the boxplot for Depthm (Figure 1), its distribution appears to be heavily skewed right. A large number of values were recorded at depths from 0 to around 500 meters since most organisms reside at these depths. Very few organisms reside in deep parts of the ocean, and it is also difficult to record such readings at such deep parts of the ocean, which explains the distribution of Depthm.

After fitting both models, hypothesis testing was conducted for each coefficient where the null hypothesis stated that the coefficient was equal to zero and the alternate hypothesis stated that the coefficient was not equal to zero. We noticed that their p-values were $<2e^{-16}$, implying that the null hypothesis is rejected. The estimate of each coefficient was well within their respective 95% CI's implying that the estimates were plausible values of the coefficients. On performing the ANOVA for both models, we saw that the p-value was $<2e^{-16}$ indicating that the variation in the response can be significantly explained by the regression line. From this point on, the models will be talked about separately due to their significantly different behavior.

Simple Linear Regression Model

A small curve was noticed in the residuals vs fitted values plot, implying that the relationship was not linear (Figure 2). This could have been caused by bad leverage points, which were identified and removed. After refitting, the pattern still persisted, so the process was repeated once more. Now,

the data seems to be a little more randomly distributed amongst 0, and the curve seems to have slightly improved, but it still exists. A normal Q-Q plot was also plotted where it was noticed, the the observations largely deviate from the straight line, implying that the normality assumption is also being violated (Figure 2). In order to correct the model so that these assumptions are followed, a Box-Cox transformation was applied to the response variable. We now see that the data points are much more evenly distributed across zero, largely improving from before. Looking at the normal Q-Q plot, we see that some observations still deviate from the straight line, but the amount that they deviate from it has reduced considerably (Figure 3). The adjusted R^2 value of the improved model is 0.5633 which is higher than that of the original model (0.2688), implying that the improved model is a better predictor than the original model.

Multiple Linear Regression Model

A curved pattern was immediately noticed in the residuals vs fitted values plot, with clusters of residuals that have obvious separation from the rest, indicating that the linearity assumption was violated and that the errors of the model were dependent (Figure 4). Since this could be the product of bad leverage points, these were identified and removed from the data. On replotting, the observations seem to be more randomly distributed around zero than they were before, but there is still an obvious curve in the residuals vs fitted values plot. The normal Q-Q plot shows us a deviation of observations from the straight line, implying that the normality assumption is also being violated (Figure 4). A Box-Cox transformation was applied with the optimal value of lambda (calculated using R) on the response variable and the model was then refitted. We now see that the data is much more randomly distributed across zero, with only a very small curve pattern. Looking at the normal Q-Q plot, we see that almost all the observations lie on the straight line, with very few values deviating from the line (Figure 5). This implies that the model meets linear regression assumptions and is a valid predictor. The adjusted R^2 value of the improved model is 0.9172, which is a large improvement from 0.4937 of the original model. Each predictor had aVIF value of 2.164, implying that the predictors are lowly correlated with each other.

Based on the AIC, BIC, and adjusted R^2 values, the final model that was selected was the improved multiple linear regression model mentioned above. The outcomes were predicted for the testing set and the mean squared error was approximately 0.823. Since this value is quite low, we can conclude that it is a good predictor for our research question based on our data.

Discussion

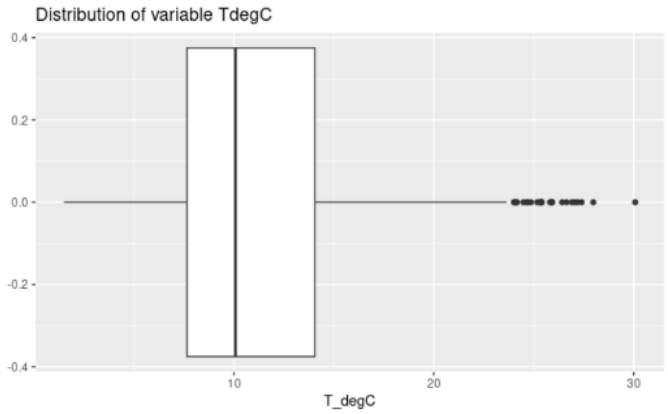
The final model is a multiple linear regression model where the predictors are salinity and depth, and the response is temperature. The slope is interpreted as the change in the water temperature for a unit change in salinity, keeping depth constant. Thus, we can establish a linear relationship between salinity and temperature, and predict temperature based on salinity at a particular depth with

small error. This relationship is important since we can now use the predictions generated by this model to anticipate climate change and global warming. However, there are potential limitations of the model that need to be explored.

This model has very high R^2 values on the training data and low mean squared error on the test data. This might indicate that the model has overfit on the training data, and there is not much variation between the training and testing data, leading to very misleading positive results on both. This possibility can be removed by performing bootstrapping, using a larger subset of data, or other similar methods in future iterations of the experiment.

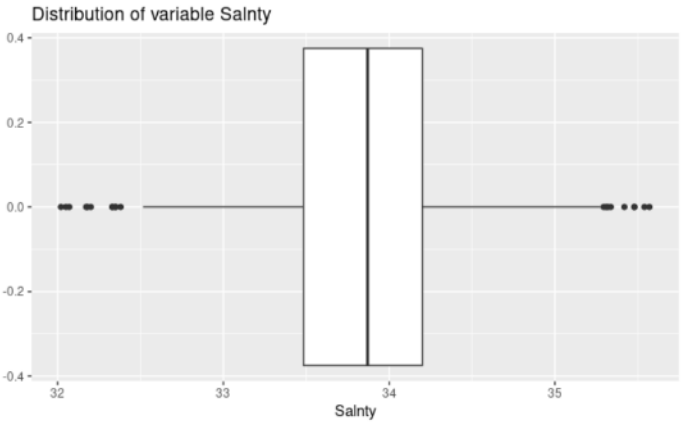
It is also possible that there are other confounding variables that affect this relationship that we have not accounted for in this analysis (leading to skewed results). From the beginning of this analysis, we were only concerned with our three main variables (or less in the case of the simple linear regression model), which we arrived at through reviewing relevant literature. In future iterations of this experiment, other impactful variables can be considered, such as biomass, etc.

Minimum	1.50
Maximum	30.08
Mean	10.86
Median	10.08
1st Quantile	7.65
3rd Quantile	14.06



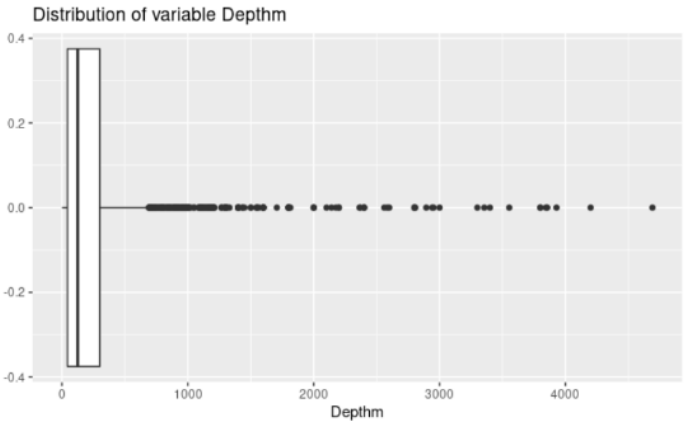
(a)

Minimum	32.02
Maximum	35.57
Mean	33.84
Median	33.87
1st Quantile	33.48
3rd Quantile	34.20



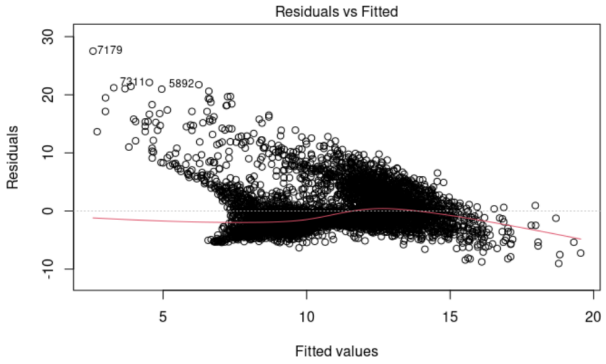
(b)

Minimum	0.0
Maximum	4691.0
Mean	223.98
Median	125.0
1st Quantile	43.25
3rd Quantile	300.0

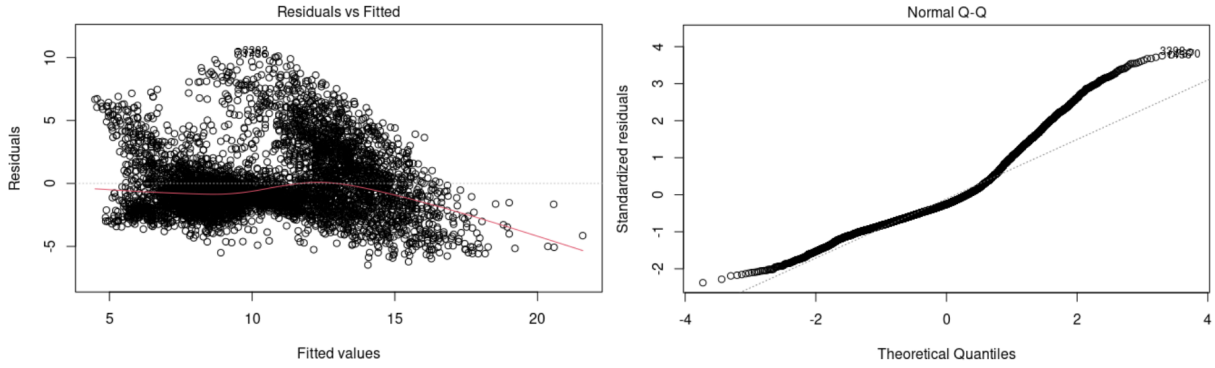


(c)

Figure 1: Numerical and visual summaries of variables (a) T_degC (b) Salinity & (c) Depthm

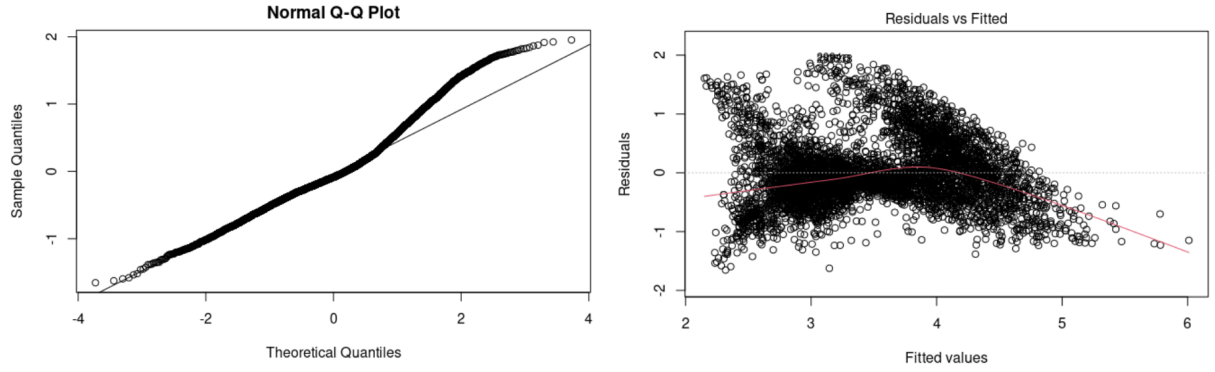


(a)



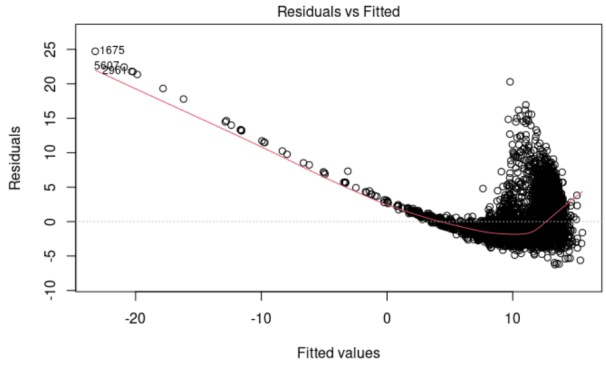
(b) (c)

Figure 2: Plots of the simple linear regression model (a) Residuals vs fitted plot of first version of the model (b) & (c) Residuals vs fitted plot and Normal Q-Q plot respectively after the second iteration of removal of bad leverage points (third version of the model)

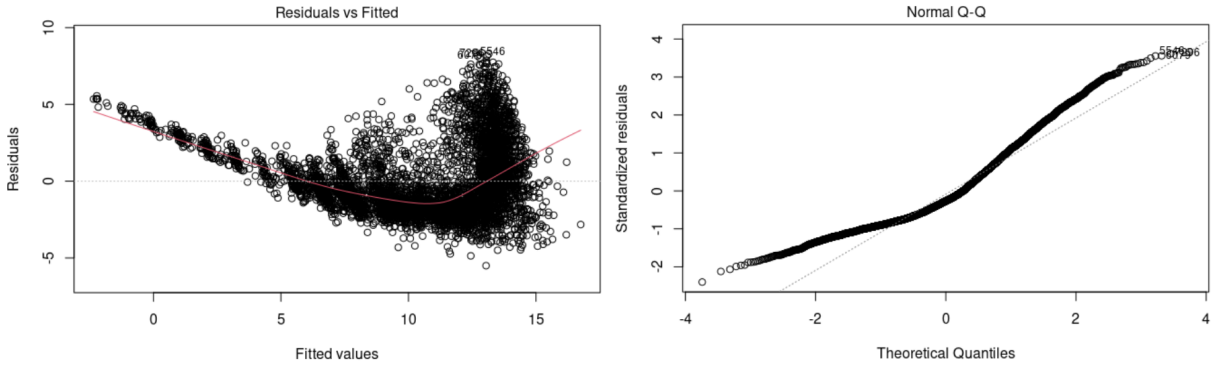


(a) (b)

Figure 3: (a) & (b) Normal Q-Q plot and Residuals vs fitted plot of the simple linear regression model after Box-Cox transformation.

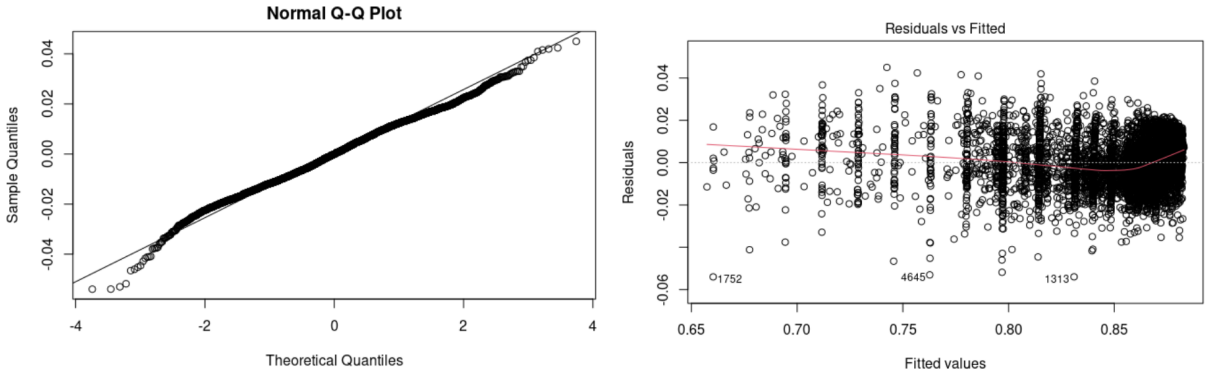


(a)



(b) (c)

Figure 4: Plots of the multivariate linear regression model (a) Residuals vs fitted plot of first version of the model (b) & (c) Residuals vs fitted plot and Normal Q-Q plot respectively after the removal of bad leverage points (first version of the model)



(a) (b)

Figure 5: (a) & (b) Normal Q-Q plot and Residuals vs fitted plot of the multiple linear regression model after Box-Cox transformation.

References

Dane, S. (2017). *CalCOFI* [Data set].

Datasets for regression analysis. (2017, December 1). Kaggle.com; Kaggle. https://www.kaggle.com/code/rtatman/datasets-for-regression-analysis/notebook

Pawlowicz, R. (2013). Key Physical Variables in the Ocean: Temperature, Salinity, and Density. *Nature Education Knowledge*, *4*(4).

Stefan, H. G., & Sinokrot, B. A. (1993). Projected global climate change impact on water temperatures in five north central U.S. streams. *Climatic Change*, *24*(4), 353–381. https://doi.org/10.1007/bf01091855

Webb, P. (n.d.). 6.3 density. In *Introduction to Oceanography*.