Linear Regression Analysis

TEMPERATURE RECONSTRUCTION WITH PROXY DATA

December 13, 2016

Leonard Strnad
University of Colorado at Denver
Department of Mathematics and Statistics

Introduction

Temperature reconstruction is a huge part of Climatology. Environmental scientists and statisticians are using statistical methods to analyze data in order to gain an awareness of the state of our planet and our influence on it. Paleoclimatologists gather natural resources of climate change like tree ring, ice core, fossil pollen, ocean sediment, coral and historical information to study climate change. They study imprints weather has left behind on these resources or proxies. The NOAA's Center for Environmental Information hosts many rich data sets that are used in this field. Temperature reconstruction has been a topic of interest for some time and a lot of research has already taken place [1, 2, 3]. Jones 2004 specifically showed that there has been a global increase in the average northern hemisphere temperatures and claims it is a result of an "anthropogenic forcing" of climate. The data used in that study only has temperature measurements back to 1856, but there is proxy data that dates back to 1001. The focus of this paper uses the same dataset.

The data set contains 1000 years worth of proxy data of which the last 145 years have an associated average northern hemisphere temperature variable which expressed the difference from the global average temperature. The variable is expressed as nhtemp and ranges from about -.6 to .6. These differences are in Celsius. The predictors include tree ring proxy data: wusa, jasper, tornetask, urals, mongolia, tasman; ice core proxy data: westgreen; and sea shell proxy data: chesapeake. Each of the variable names reflect the geographical origin of the proxy data. The proxy data and the nhtemp variable have been smoothed by some gaussian smoothing kernel. There is a total of eight variables with one response. Again, complete data is only available for the last 145 years of the total 1000 years.

The paper by Jones 2004 mentioned above claims to be able to reconstruct atmospheric temperatures well enough to strongly suggest humans are responsible for the abnormal increase in temperature. We have not had the chance to cover time series regression analysis or been exposed to methods of extrapolation, but Jones 2004 claims to be properly reconstructing the atmospheric temperature. This paper simply aims at building the best fitting model that fits the historical data. The model we will consider will regress the proxy data on the historical data. A successful model will allow temperatures from 1856 to 2000 to be modeled by these proxies.

We first take a look at the data summary, collinearity, correlation and densities of the predictors. Then, a linear model is constructed to discuss model selection, and model diagnostics. Finally, we interpret the model and discuss prediction outside the extremes of the historical data. The R code that corresponds to this project can be found at `https://github.com/ljstrnadiii/LRAproject`

Data Summary

The variables wusa, jasper, tornetask, urals, mongolia, and tasman are all tree ring proxy data. The variable westgreen is ice core proxy data and chesapeake is sea shell proxy data. They all have a similar range of values. The minimum, maximum, mean, and quartiles of the predictors are shown in the table 1.

The correlation matrix in table 2 shows that mongolia has high correlation with wusa, jasper, tornetrask, urals, and year. Addionally, year has high correlation with a few variable but is the most correlated with the response variable. The scatter plot matrix shows the densities and bivariate densities between the variables. The urals and tornetrask proxies seem to potentially be bimodal and seem the furthest from a normal distribution. Also, most of the densities show that the distributions may be skewed. Next, we consider collinearity in the predictor space.

We check the variance decomposition displayed in table 3 and see that the condition index is high, however, there is only one significant proportion associated with urals. The proportions are included in the table below. One significantly large proportion does not give enough information

	nhtemp	wusa	jasper	westgreen	chesapeake	tornetrask	urals	mongolia	tasman
1	Min. :-0.6200	Min. :-1.67000	Min. :-0.9400	Min. :-1.18000	Min. :-2.2200	Min. :-1.7800	Min. :-1.4900	Min. :-1.7800	Min. :-1.41000
2	1st Qu.:-0.2800	1st Qu.:-0.45000	1st Qu.:-0.0800	1st Qu.:-0.26000	1st Qu.:-0.9300	1st Qu.:-0.7600	1st Qu.:-0.3800	1st Qu.:-0.2200	1st Qu.:-0.20000
3	Median :-0.1500	Median :-0.03000	Median : 0.2400	Median : 0.03000	Median :-0.4800	Median :-0.1200	Median :-0.1200	Median : 0.2700	Median : 0.09000
4	Mean :-0.1203	Mean : 0.04358	Mean : 0.1959	Mean : 0.02878	Mean :-0.4526	Mean :-0.1282	Mean :-0.1041	Mean : 0.2441	Mean : 0.08479
5	3rd Qu.: 0.0300	3rd Qu.: 0.49000	3rd Qu.: 0.4900	3rd Qu.: 0.33000	3rd Qu.: 0.0200	3rd Qu.: 0.3500	3rd Qu.: 0.2700	3rd Qu.: 0.7700	3rd Qu.: 0.37000
6	Max. : 0.6600	Max. : 2.07000	Max. : 1.5300	Max. : 1.22000	Max. : 1.9000	Max. : 1.7000	Max. : 0.7300	Max. : 1.7000	Max. : 1.21000
7	NA's :856								

Table 1: 5 number summary

	nhtemp	wusa	jasper	westgreen	chesapeake	tornetrask	urals	mongolia	tasman	year
nhtemp	1.00	0.50	0.43	0.29	-0.22	0.53	0.48	0.53	0.38	0.73
wusa	0.50	1.00	0.73	0.53	-0.33	0.44	0.66	0.77	0.18	0.84
jasper	0.43	0.73	1.00	0.30	-0.46	0.69	0.85	0.89	0.02	0.73
westgreen	0.29	0.53	0.30	1.00	-0.05	0.33	0.25	0.28	0.13	0.34
chesapeake	-0.22	-0.33	-0.46	-0.05	1.00	-0.12	-0.38	-0.54	-0.01	-0.45
tornetrask	0.53	0.44	0.69	0.33	-0.12	1.00	0.73	0.70	0.22	0.61
urals	0.48	0.66	0.85	0.25	-0.38	0.73	1.00	0.87	-0.11	0.73
mongolia	0.53	0.77	0.89	0.28	-0.54	0.70	0.87	1.00	0.05	0.87
tasman	0.38	0.18	0.02	0.13	-0.01	0.22	-0.11	0.05	1.00	0.32
year	0.73	0.84	0.73	0.34	-0.45	0.61	0.73	0.87	0.32	1.00

Table 2: Correlation matrix

	wusa	jasper	westgreen	chesapeake	tornetrask	urals	mongolia	tasman	year
1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.28
2	0.01	0.00	0.00	0.00	0.01	0.00	0.02	0.00	0.14
3	0.01	0.00	0.01	0.07	0.02	0.00	0.01	0.32	0.06
4	0.05	0.00	0.00	0.15	0.11	0.00	0.00	0.10	0.03
5	0.12	0.00	0.09	0.13	0.02	0.00	0.01	0.14	0.02
6	0.05	0.00	0.47	0.26	0.06	0.00	0.03	0.02	0.03
7	0.30	0.12	0.21	0.22	0.19	0.03	0.90	0.03	0.35
8	0.42	0.72	0.17	0.16	0.43	0.01	0.00	0.17	0.07
9	0.04	0.16	0.05	0.00	0.16	0.96	0.03	0.21	0.01

Table 3: Variance Decomposition

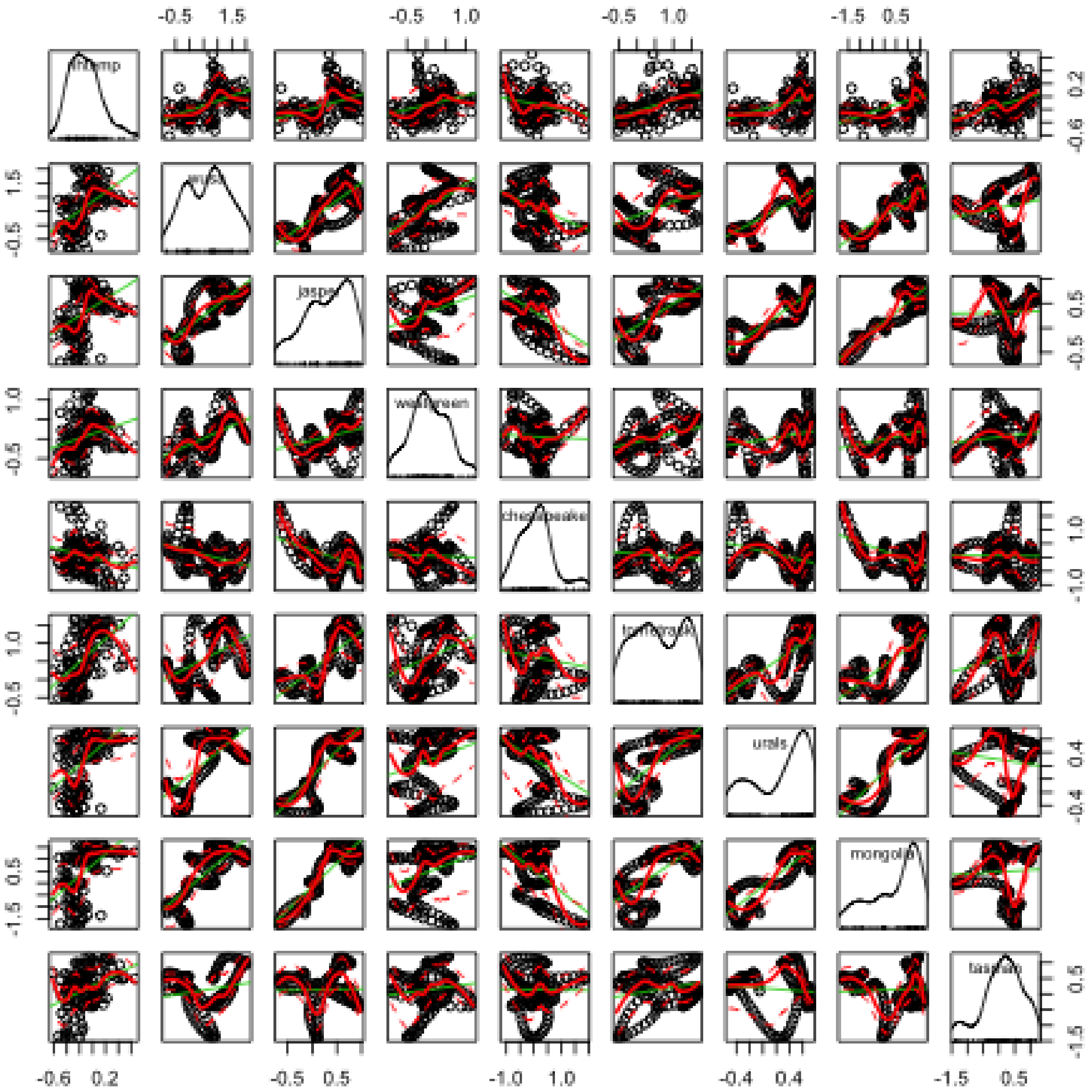

Figure 1: Scatterplot Matrix of nhtemp, wusa, jasper, westgreen, chesapeake, tornetrask, urals, mongolia, and tasman (in order left to right)

so we consider the variance inflation factor.

We fit the full model using all the predictors. Table 4 displays the variance inflation factors in the model. We see that the majority of the factors are tolerable except for Mongolia. So, we drop Mongolia, refit the model and consider the variance inflation factor again. Table 4 shows removing mongolia reduces the VIF, but we get pairs of significant variance decomposition proportions with a high condition index. However, performing method selection without Mongolia leads to a subset of predictors with reasonable variance inflation factors, but high variance decomposition proportions. In the model selection section we argue that perturbing the response does not lead to instability of the β coefficients which allows us to build the model regardless of the high variance proportions. Next, we use all predictors but mongolia in a model selection process.

	vif
wusa	7.82
jasper	7.10
westgreen	1.86
chesapeake	2.03
tornetrask	4.78
urals	6.89
mongolia	13.86
tasman	2.10
year	10.48

	vif
wusa	7.79
jasper	6.12
westgreen	1.83
chesapeake	1.78
tornetrask	4.43
urals	6.69
tasman	1.89
year	7.70

Table 4: Variance Inflation Factor before and after removing Mongolia

Model Selection

The section on data summary argues that Mongolia has a high variance inflation factor and removing it reduces the variance inflation factors of other variables as well. This is shown in Table 4. We perform model selection on the subset of predictors: wusa, jasper, westgreen, chesapeake, tornetrask, urals, tasman, and year. The model selection process considers the best subset methods using AIC, BIC, Cp-statistic, and the Adjusted R^2 criterion. The full model has an R^2=.61 with coefficients in table below.

The first method considered is best subset minimizing the AIC criterion. The AIC plot, figure 2, suggests five or six predictors which includes the intercept. The AIC criterion suggests that a more parsimonious model occurs at five predictors. A plot of the BIC criterion, figure 2, is to the right of the AIC. The BIC criterion also suggests that a model with five predictors including the

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-12.1097	1.5317	-7.91	0.0000
wusa	-0.1421	0.0438	-3.24	0.0015
jasper	-0.0784	0.0662	-1.18	0.2381
westgreen	0.0829	0.0355	2.34	0.0209
chesapeake	0.0393	0.0270	1.46	0.1478
tornetrask	0.0169	0.0366	0.46	0.6439
urals	0.0427	0.0762	0.56	0.5760
tasman	0.0334	0.0270	1.24	0.2184
year	0.0063	0.0008	7.74	0.0000

Full Model

intercept is the best model.

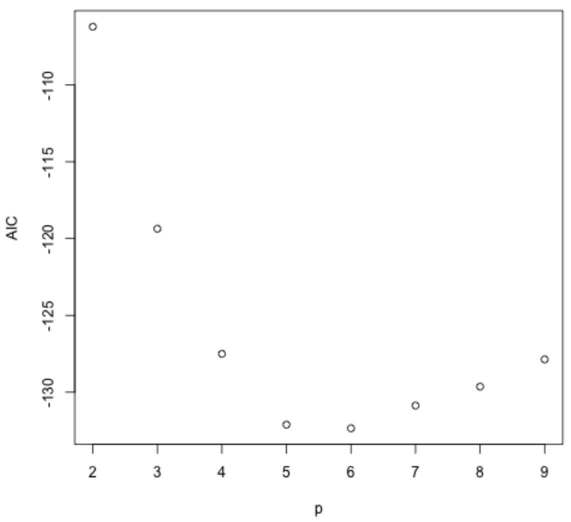

Figure 2: AIC plot vs # predictors

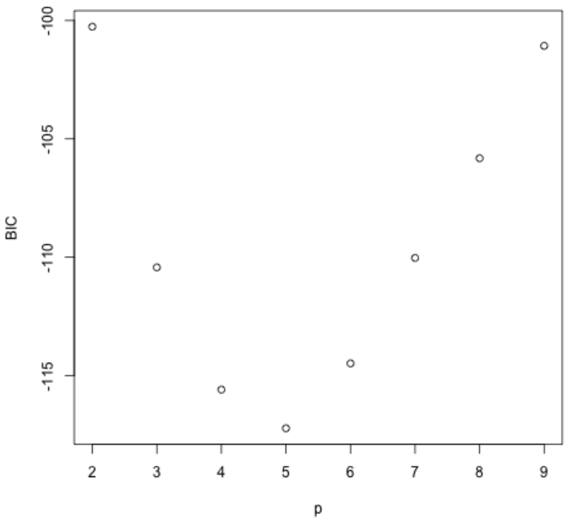

Figure 3: BIC plot vs # predictors

Next, we consider the adjusted R^2 and the C_p-statistic. The plots in figure 4 and 5 below display each criterion. The intent is to maximize the adjusted R^2. Figure 4 suggests that the model includes five or six predictors. Again, the difference between using five or six predictors is small and thus we use five predictors including the intercept in order to construct a more parsimonious model. Lastly, we consider the C_p-statistic, figure 5. We choose the number of predictors that is closest to and below the plotted line. This method also suggests that we choose 5 predictors including the intercept.

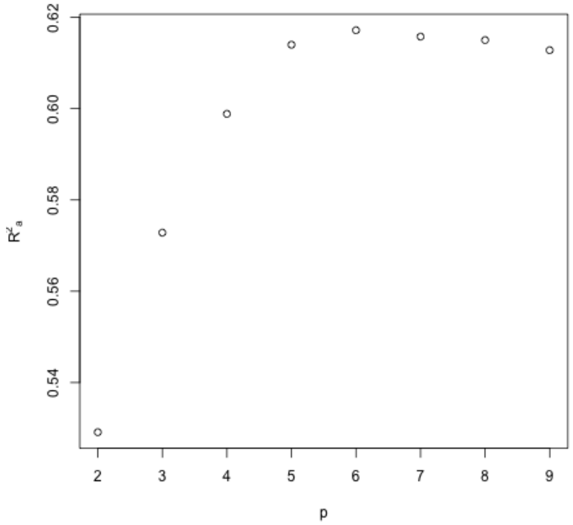

Figure 4: Adjusted R^2 plot vs # predictors

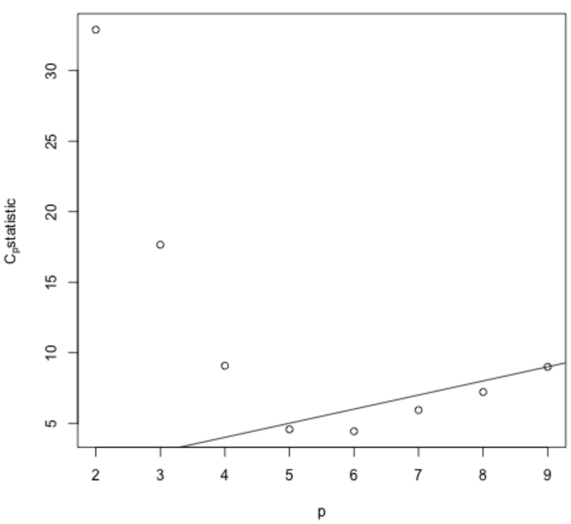

Figure 5: C_p-statistic plot vs # predictors

The majority of the methods agree to use five predictors including the intercept in the model. Using five predictors leads to a model that includes the intercept, wusa, westgreen, chesapeake, and year i.e. one tree ring proxy, one ice core proxy, one sea shell proxy and time. Using a model that has six predictors adds the tasman tree proxy variable. We can perform a hypothesis test to test if the larger model is preferred to the smaller model. This hypothesis test can be tested using the F statistic that the anova procedure calculates. The hypothesis test is as follows:

$$H_O : \beta_{tasman} = 0 \mid \beta_{wusa,westgreen,chesapeake,year} \quad \text{included}$$

$$H_A : \beta_{tasman} \neq 0 \mid \beta_{wusa,westgreen,chesapeake,year} \quad \text{included}$$

$$F = 2.16 \quad p = .14$$

There is not enough evidence to claim there is a difference between the two models. Therefore, we choose the simplest model which does not include tasman. Before continuing towards model diagnostics, we revisit the collinearity assessment that was addressed in the data summary section. Performing the variance decomposition with these four predictors leads to a high condition index, but low variance inflation factors displayed in the table 5 to the right. We argue that although there are more than two high variance decomposition proportions using these variable in the final model, perturbing the response does not lead to significant instability.

	x
(Intercept)	-0.01
wusa	-0.03
westgreen	-0.10
chesapeake	0.10
year	-0.01

	x
wusa	4.53
. westgreen	1.48
chesapeake	1.28
year	4.08

Table 5: percent change in coefficients (left). variance inflation factor(right)

Table 5 above displays the percent change in the coefficients after adding gaussian noise with variance that is 10% the standard deviation of the response to the response. We see that the coefficient associated with westgreen and chesapeake is approximately 10% lower and 10% higher respectively. These percent differences are not terribly significant which suggests that the model does not suffer from collinearity enough to have unstable coefficient estimates. Therefore, we continue with the model chosen by the four methods above which includes wusa, westgreen, chesapeake, and year.

Model Diagnostics

Model diagnostics are an important component of linear regression analysis. In order of most important to least, model selection includes checking the structure or link function of the model, dependence of errors to prevent overconfidence, non-constant variance and normality in the errors. Lastly, we will take a close look at any unusual observation and assess if there are any observation that are significantly affecting the structure of the model.

Structure and Constant Variance

First, we examine the structure of the model. It is most important to assess whether the underlying link function is appropriate. We plot the residuals vs. the fitted values. We also use this plot to assess if the model has constant variance. Figure 6 does not suggest any significant structure.

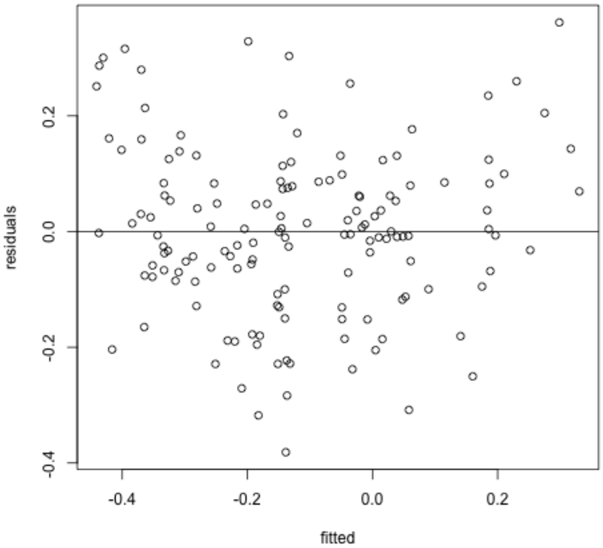

Figure 6: residuals vs. fitted

We also see that there is approximately equal thickness about $y = 0$ which suggests we have constant variance. There may be some concern with the constant variance assumption, but most importantly there does not seem to be any issue with the underlying structure. This can be verified by regressing the fitted values onto the residuals. One may suspect there is a polynomial structure in this plot above. We fit multiple models using a second, third and fourth degree polynomial basis. The model summary does not provide significant evidence of these polynomial structures. It seems reasonable to continue assuming the structure and constant variance assumptions are met. Next, we assess the dependence or the serial correlation.

Dependence of Errors

The second most important assumption is the independence of errors. This subsection addresses this issue by plotting the serial correlation of the model. The serial correlation is assessed by plotting the residuals vs. time or space. The final model chosen in the model selection section includes a year variable. So, we plot the residuals vs year in figure 7. The plot in figure 7 suggests that there is not significant serial correlation. In order to examine this more carefully we consider the Durbin-Watson hypothesis test:

$$H_O : \rho = 0$$

$$H_A : \rho \neq 0$$

$$DW = 1.1345 \quad p \approx 0.$$

The Durbin-Watson test says there is statistical evidence to claim there is positive serial correlation. Literature online suggests that a DW-statistic of 1.13 is only a minor case of positive serial correlation. The plot, however, does not seriously suggest positive correlation. The result of this hypothesis test is important to keep in mind because positive correlation leads to β estimates that do not achieve minimum variance and as a result there will be overconfidence in prediction. Further research could include a generalized least squares regression model which deals well with correlation in the residuals.

Normality

Lastly, the least important assumption in the linear model is that the errors are normal. Since we do not know the actual values of the errors we assess the normality of the residuals which are estimates of the errors. In order to do this we plot the theoretical normal quantiles against the actual quantiles. If we get a linear structure along $y = x$, then we assume that the residuals which

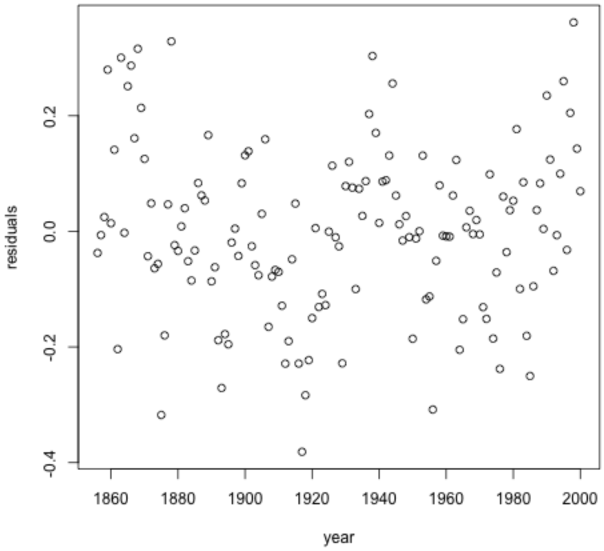

Figure 7: residuals vs. year

estimate the errors are normal. The QQ-plot in figure 8 displays a strong $y = x$ relationship. This suggests that the residuals are normal. Therefore, we can continue to assume the error distribution is normal.

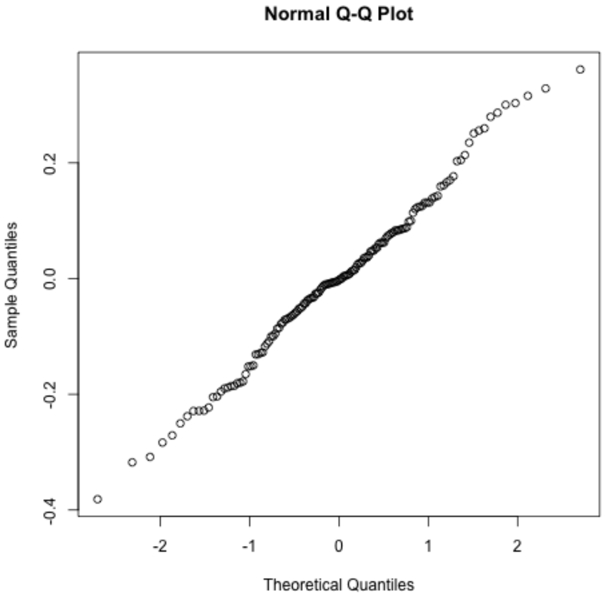

Figure 8: QQ-plot of residuals

The model assumptions are not entirely valid. We may have potential issues with constant variance which may suggest that the underlying link function is not correct. Box Cox transformations provide a way of transforming the response, but the suggested transformation does not lead to a better model. The link function is the most important structure of the linear model. Also, the violation of independent errors shown by the Durbin-Watson test is another violation to keep in mind. Even though the test statistic provides evidence of only a minor violation, we can no longer assume the estimates for β have the minimum variance which leads to overconfidence in prediction. Next, unusual observations are considered.

Unusual Observations

This section considers unusual observations which includes outliers, influential points, and leverage points. Detecting outliers provides insight of unusual observations in the response space. Detecting leverage points provides insight of unusual observation in the predictor space. Recall that the data has been smoothed by some kernel smoother before being used. This will likely remove leverage points or outliers. Influential points are points that significantly affect the fit of the model by causing a change in the coefficient estimation.

First, we consider the leverage points. In order to assess which observations may be leverage points we plot the halfnormal quantile-quantile plot of the hat values. All of the observations seem to be following the same trend in the plot of figure 9. There do not seem to be any obvious leverage points.

Next, we consider outlier points which are unusual observations in the response space–average northern hemisphere temperatures. In order to assess which values may be outliers we simply run the outlier test in R. The result states there are no studentized residuals with a significant Bonferonni p-value. This means that there are no significant outliers.

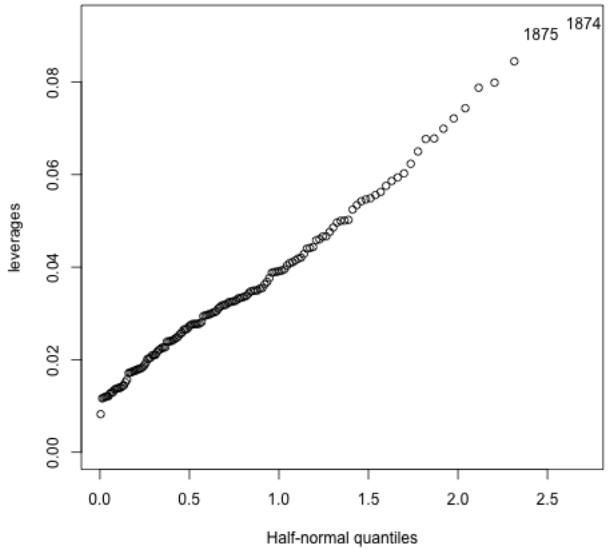

Figure 9: qq-halfnormal plot of hat-values

Lastly, we consider influential points. The cook's distances express the difference between the model fitted with and without a particular point. Cook's distance follows a halfnormal distribution. Therefore, we consider a qq-halfnormal plot of cook's distances. The QQ-plot of figure 10 suggests there are three obvious observations that do not follow the trend of the qq-halfnormal plot: 1875, 1878, and 1998. In order to assess how influential these points are we consider the percent change in the coefficients fitted with and without them. The tables in Table 6 show that the observation that corresponds to 1998 has a max coefficient difference of 8%, the observation corresponding to 1878 has a max coefficient difference of 11% and the observation corresponding to 1975 has a max coefficient difference of -19%. The observation corresponding to 1875 is clearly the most influential.

	1875
(Intercept)	-0.01
wusa	-0.04
westgreen	-0.12
chesapeake	-0.19
year	-0.01

	1878
(Intercept)	0.01
wusa	0.04
westgreen	0.08
chesapeake	0.11
year	0.01

	1998
(Intercept)	0.05
wusa	0.08
westgreen	-0.02
chesapeake	0.07
year	0.05

Table 6: Percent change in coefficients after removing the potential influential points

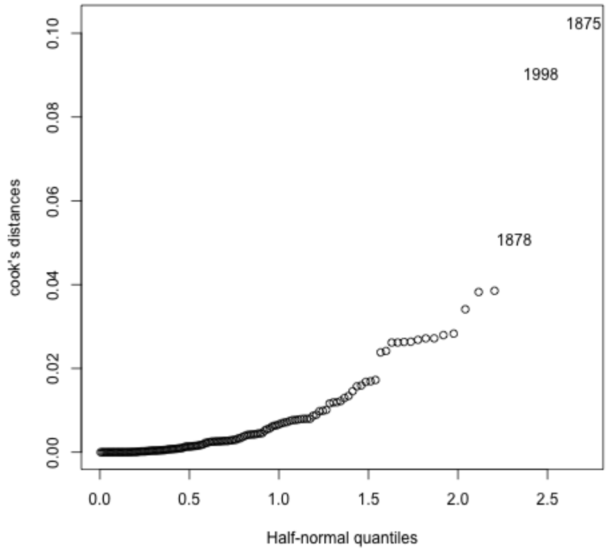

Figure 10: QQ halfnormal plot of cook's distances

This section on model diagnostics has shown that the we may have an issue with the constant variance assumption, the dependence of error assumption, and that we have a pretty significant influential point we may consider removing. In regards to the intent of this paper, it seems reasonable to claim that this model can roughly approximate the average northern hemisphere temperatures using only the geological and biological proxies that are the predictors. The final model is given in table 7 with an associated R^2=.62 and a mean squared error of .02.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-13.3437	1.1173	-11.94	0.0000
wusa	-0.1782	0.0334	-5.34	0.0000
westgreen	0.0965	0.0319	3.03	0.0029
chesapeake	0.0584	0.0229	2.56	0.0117
year	0.0069	0.0006	11.76	0.0000

Table 7: Final Model

Interpretation and Prediction

The final model regresses wusa, westgreen, chesapeake and year onto the average northern hemisphere change in temperature. The coefficient with largest magnitude corresponds to tree ring proxy data from western USA. Holding all other variables constant, one unit increase in tree ring proxy data leads to a decrease in the average northern hemisphere temperature of .18 degrees Celsius. The second largest coefficient corresponds to the ice core proxy data. Similarly, holding all other variables constant, one unit increase in the ice core proxy variable leads to an increase in the average northern hemisphere temperature by .10 degrees Celsius. The final model includes the best single tree ring proxy data source, the only ice core proxy data, the only sea shell proxy data and time. The mean squared error on the complete data is .02. Even though the model diagnostics suggest the model does not satisfy an assumption the mean squared error is low and we have a model with a significant R^2. It seems that this model does a decent job interpolating and reducing the number of measurements needed to build a model, but what happens when we try to extrapolate?

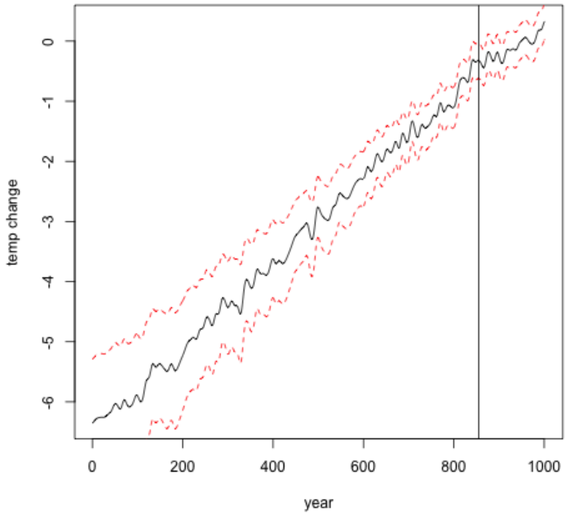

Figure 11: Final Model

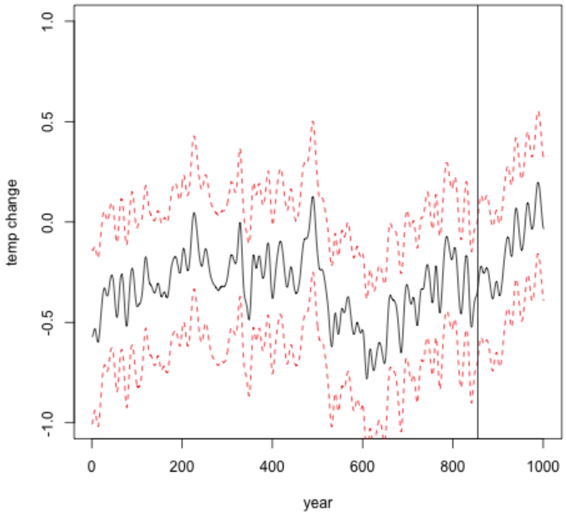

Figure 12: without year

Extrapolating the model to data points that lie outside the models training hull is dangerous because the underlying link function may only be acceptable within the range of the training data. Figure eleven and figure twelve display plots of two different models and their prediction intervals on the data that predates any historical record: 1001-2000. The left plot of figure eleven displays the prediction interval which correspond to the final model. The right plot in figure twelve is the result of best subset using AIC to find a model when ignoring year (time). This model uses wusa, jasper, chesapeake, tornetrask, urals, and tasman and it has an associated $R^2 = .45$, a mean squared error of .03 with one insignificant coefficient. The interesting observation is that the plot in figure twelve is much more like the reconstruction of Jones 2004 [1]. The corresponding plots of that study can be found here at `https://www.ncdc.noaa.gov/paleo/globalwarming/images/last2000-large.jpg`. The final model does an absolutely terrible job at extrapolating.

The final model predicts that at year 1000 the average northern hemisphere temperature was six degrees Celsius less than the global average. This is absolutely absurd. The five predictions on the top plot in the link are five different research paper's reconsructions of the temperature and they are have a maximum decrease predicted of about .2. The quick model just derived and mentioned above matches these reconstructions much, much better than the final model chosen with the model selection process and model diagnostics discussed above. This just simply shows that even if the R^2 value is the highest of any subset of predictors and the mean squared error is lower than the alternative model of figure twelve, the model may adopt a significant link function assumption error. Additionally, adding time enforces a stronger linear relationship because the data the final model trained on is completely inside a local increasing trend in temperature indicated by the plot right of the vertical line in figure twelve. The time variable deviates from its mean significantly when extrapolating compared to the other variables. This is a large part of the problem.

The conclusion of this regression analysis project is rich. Serious care must be taken while performing model selection and assessing the validity of the link function. Even though the residual plot vs the fitted values in figure six did not seem to indicate any serious structural error, there is clearly a more serious structural issue than the quick model (figure twelve) fitting wusa, jasper, chesapeake, tornetrask, urals, and tasman. This obvious structural difference only became apparent when considering performance of extrapolation. If the reconstructions of the papers discussed in the introduction were not available, then cross referencing would of never been able to occur in order to recognize just how poorly the final model extrapolates. Sacrificing model performance such as R^2, AIC, BIC, C_p, and MSE may have lead to a model with a more appropriate structure. The takeaway from this project is that if the plan is to extrapolate, one must make sure to pay particularly close attention to assumptions about the underlying structure.

References

[1] P. D. Jones and M. E. Mann, “Climate over past millennia,” *Reviews of Geophysics*, vol. 42, no. 2, 2004.

[2] A. Moberg, D. M. Sonechkin, K. Holmgren, N. M. Datsenko, and W. Karlén, “Highly variable northern hemisphere temperatures reconstructed from low-and high-resolution proxy data,” *Nature*, vol. 433, no. 7026, pp. 613–617, 2005.

[3] P. Jones, D. Parker, T. Osborn, and K. Briffa, “Global and hemispheric temperature anomalies–land and marine instrumental records,” *Trends: a compendium of data on global change*, 2006.