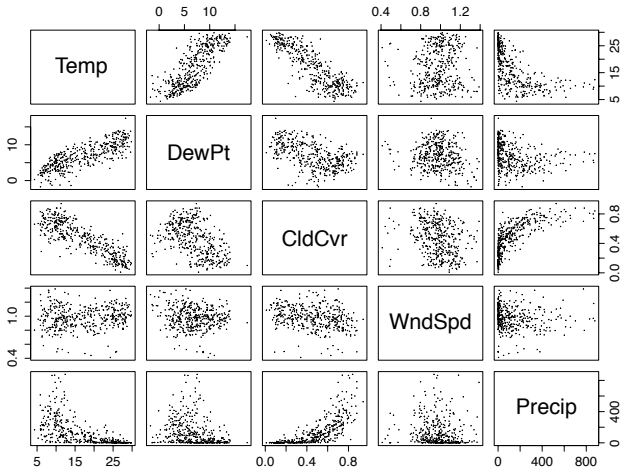
EESS 260 Final Project: Effect of the Climate Variables on the Log-transformed Inflow to Lake Shasta

[1] High Dimensional Regression ----------

[a] The condition number is **7.459**, using the log-transformed inflow values. Technically, having a condition number less than 10 falls into the "low" multicollinearity category, and having a condition number between 10 and 30 falls into the "moderate" multicollinearity category; however, since the condition number is rather close to the decision boundary, I would say that there exists some problems with multicollinearity, though not too severe. To briefly look into the potential sources of multicollinearity, we can look at the correlation matrix / plot of the predictors.

Looking at the pairwise correlation matrix and scatter plots between predictors (below), we see that a few predictors-pairs have relatively strong correlation (>|0.7|) with each other: 1) "Temp" and "DewPt", 2) "Temp" and "CldCvr", and 3) "CldCvr" and "Percip". The correlations also makes physical sense: 1) high temperature correlates with high dew point, 2) high temperature correlates with low cloud coverage, and 3) high cloud coverage correlates with high precipitation. We note these statistics / plots could not show collinearity between multiple predictors, which could very well exist.

	Temp	*DewPt*	*CldCvr*	*WndSpd*	*Precip*
Temp	*1.00000*	***0.8009***	***-0.8527***	*0.07986*	*-0.6083*
DewPt	***0.80091***	*1.0000*	*-0.5014*	*-0.14961*	*-0.2701*
CldCvr	***-0.85272***	*-0.5014*	*1.0000*	*-0.27639*	***0.7326***
WndSpd	*0.07986*	*-0.1496*	*-0.2764*	*1.00000*	*-0.1094*
Precip	*-0.60827*	*-0.2701*	***0.7326***	*-0.10942*	*1.0000*



[b] Looking at the p-values for the predictors, **Temperature**, **Dew Point**, and **Wind Speed** are the least significant; in fact, we would not reject the null hypothesis that their coefficients are zero, at the 5% level. This might be due to the fact that they are simply insignificant, or that they are redundant, such that their explanatory powers are captured by their collinear counterparts. For instance, "Temp" is insignificant and "CldCvr" is significant, but as part [a] has shown, " Temp" and "CldCvr " has a correlation of -0.85272. The coefficients' 95% confidence intervals are also as follows.

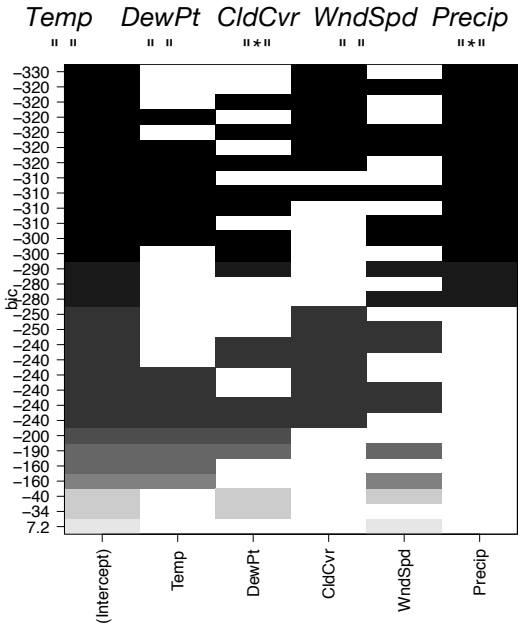
95% confidence interval:

	2.5 %	*97.5 %*
(Intercept)	*3.913378*	*5.033541*
Temp	*-0.018760*	*0.020262*
DewPt	*-0.030802*	*0.019648*
CldCvr	*0.509343*	*1.481815*
WndSpd	*-0.161658*	*0.426544*
Precip	*0.001361*	*0.002086*

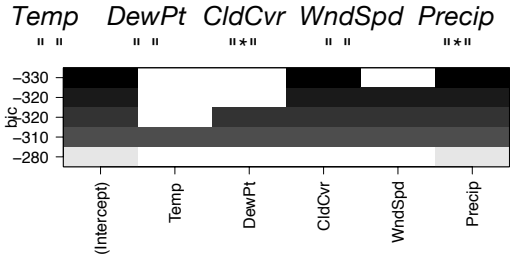
Coefficients:

	Estimate	*Std. Error*	*t value*	*Pr(>|t|)*
(Intercept)	*4.473460*	*0.284883*	*15.70*	*< 2e-16 ****
Temp	*0.000751*	*0.009924*	*0.08*	*0.94*
DewPt	*-0.005577*	*0.012831*	*-0.43*	*0.66*
CldCvr	*0.995579*	*0.247322*	*4.03*	*6.8e-05 ****
WndSpd	*0.132443*	*0.149593*	*0.89*	*0.38*
Precip	*0.001723*	*0.000184*	*9.34*	*< 2e-16 ****

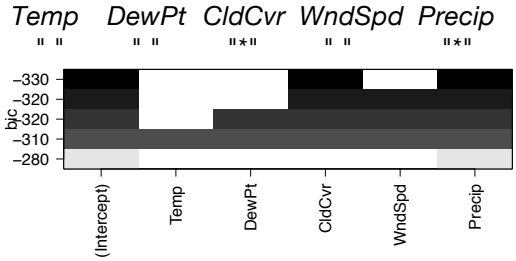
[c] From conducting an exhaustive search, the best subset of predictors with the BIC is of size 2, and includes "**CldCvr**" and "**Precip**".



[d] The forward search results in the same subset as the exhaustive search: the best subset of predictors is of size 2, and includes "**CldCvr**" and "**Precip**".



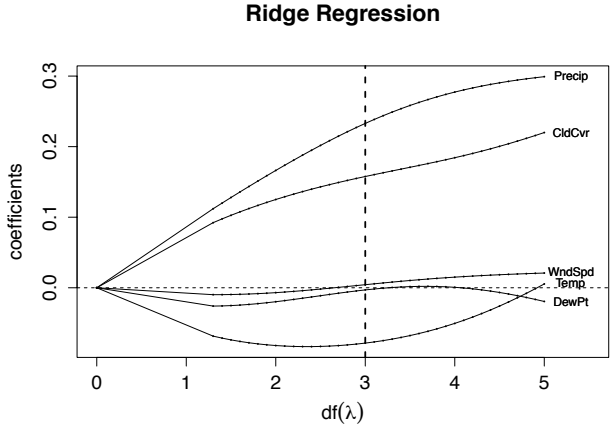
[e] The backward search results in the same subset as the forward search: the best subset of predictors is of size 2, and includes "**CldCvr**" and "**Precip**".



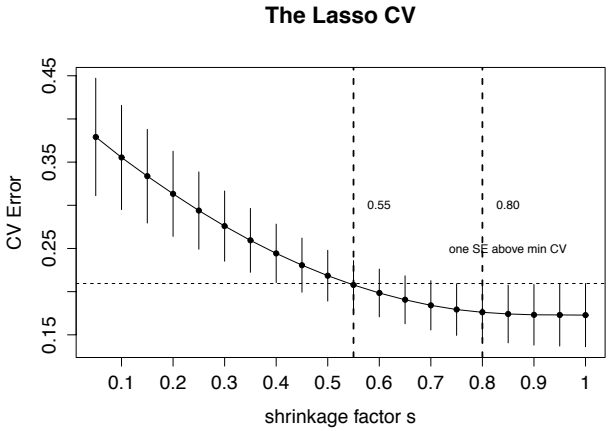
[f] The ridge regression coefficients for df(λ) = 3 is computed.

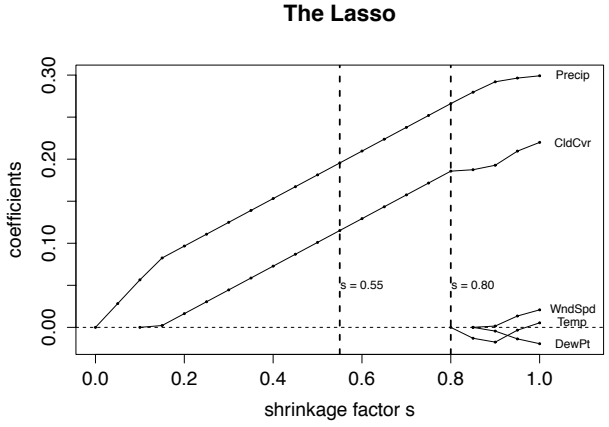
Coefficients (df(λ) = 3):

(Intercept)	*5.266*
Temp	*-0.078418*
DewPt	*-0.003045*
CldCvr	*0.157678*
WndSpd	*0.004351*
Precip	*0.232927*



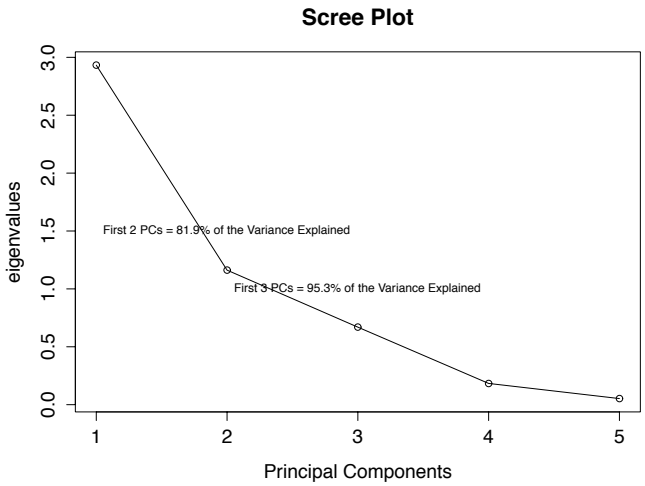
[g] The "one standard deviation rule" states once we find the $\hat{s}$ that minimizes the CV error, we move towards the direction of increasing regularization as much as we can such that CV error is still within one standard error of CV($\hat{s}$): $CV(s) \leq CV(\hat{s}) + \mathrm{SE}(\hat{s})$. Thus, **s = 0.55.** However, we note that between s = 0.8 and s = 0.55, the Lasso does not perform additional regularizations in terms of reducing the number of predictors. Further, s = 0.8 (0.2864) provides a lower mean square error than s = 0.55 (0.2888); thus we would use **s = 0.8** for part [k].





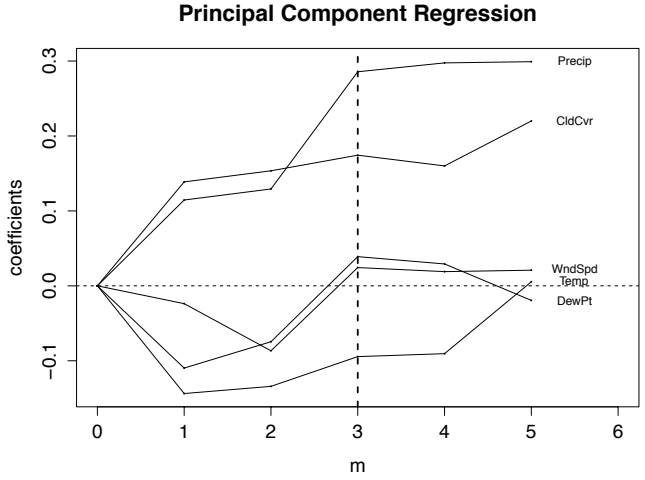
[h] As seen from part [f], the Lasso with **s = 0.55** and **s = 0.8** both choose a subset of predictors of size 2, which includes "**CldCvr**" and "**Precip**". This is the same subset as the ones found in the forward and the backward stepwise regressions.

[i] The first **3** principal components explain 95.3% of the variance, and is the least number of principal components that have >90% of the variance explained.



[j] Shown below are the coefficients from the principal component regression, using the first 3 principal components. We note that the intercept has the same value as that of all other methods used in this problem.

Coefficients (PCs=first 3):
(Intercept) 5.2661
Temp -0.09446
DewPt 0.03895
CldCvr 0.17438
WndSpd 0.02439
Precip 0.28570



[k] To measure the (out of sample) prediction error, we'll use the mean squared error (MSE) formula, and scaled the training and the test sets separately; the coefficients resulted from each method are shown below as well. The method with the lowest MSE is **the Lasso**.

$$MSE = \frac{1}{54}\sum_{i=1}^{n=54}(y_i - \hat{y}_i)^2$$

Method	[b] OLS	[c] Best Subset	[f] Ridge	[h] The Lasso	[i] PC
MSE	0.2959	0.2964	0.2963	**0.2864**	0.2968

Coefficients:

OLS *(the results below are different from those in part [b], since the data used here has been scaled):*

	Estimate	*Std. Error*	*t value*	*Pr(>|t|)*
(Intercept)	*5.2661*	*0.0207*	*254.31*	*< 2e-16 ****
Temp	*0.0054*	*0.0714*	*0.08*	*0.94*
DewPt	*-0.0194*	*0.0447*	*-0.43*	*0.66*
CldCvr	*0.2200*	*0.0546*	*4.03*	*6.8e-05 ****
WndSpd	*0.0209*	*0.0236*	*0.89*	*0.38*
Precip	*0.2992*	*0.0320*	*9.34*	*< 2e-16 ****

Best Subset*:*

	Estimate	*Std. Error*	*t value*	*Pr(>|t|)*
(Intercept)	*5.2661*	*0.0207*	*254.66*	*<2e-16 ****
CldCvr	*0.2191*	*0.0304*	*7.20*	*3e-12 ****
Precip	*0.2995*	*0.0304*	*9.84*	*<2e-16 ****

Ridge Regression*:*

please see part [f]

The Lasso *(MSE calculation uses s=0.8; we note that the Lasso does not have an analytical solution for the standard error; therefore, problem 2 exists):*

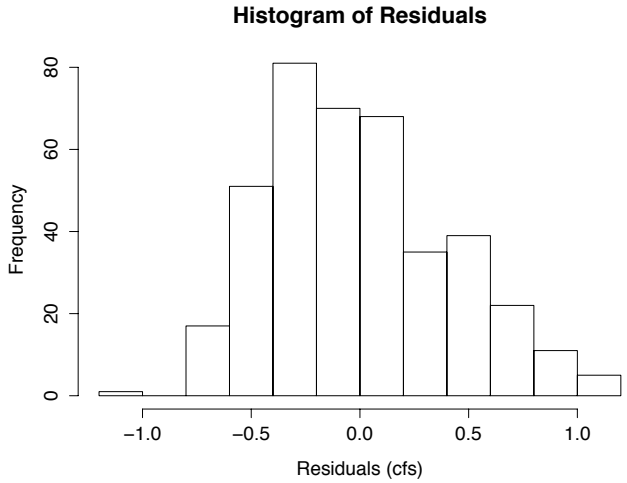
	(s=0.55)	*(s=0.80)*
(Intercept)	*5.2661*	*5.2661*
CldCvr	*0.1151*	*0.1857*
Precip	*0.1955*	*0.2661*

Principal Component Regression*:*

please see part [j]

[2] The Bootstrap ----------

[a] The histogram of the residuals resulted from running a regression with the Lasso and s = 0.8 looks slightly positively skewed than a normal distribution. For this problem, only the training set is considered (i.e. first 400 values).



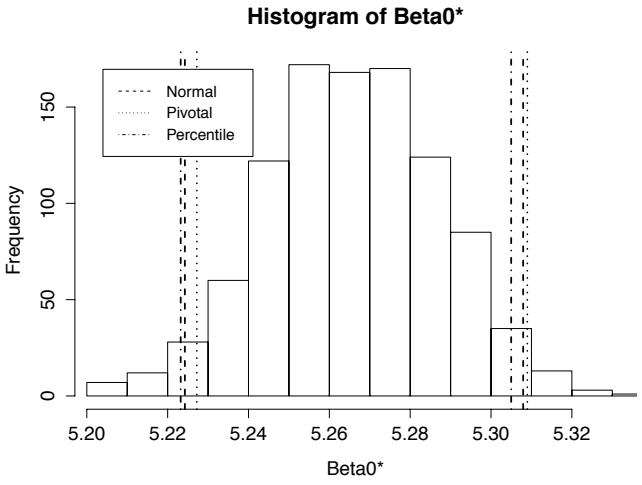
[b] Running a non-parametric bootstrap on the Lasso's residuals allows us to quantitatively say that three predictors are not significantly different from zero by use of any of the three confidence intervals: "**Temp**",

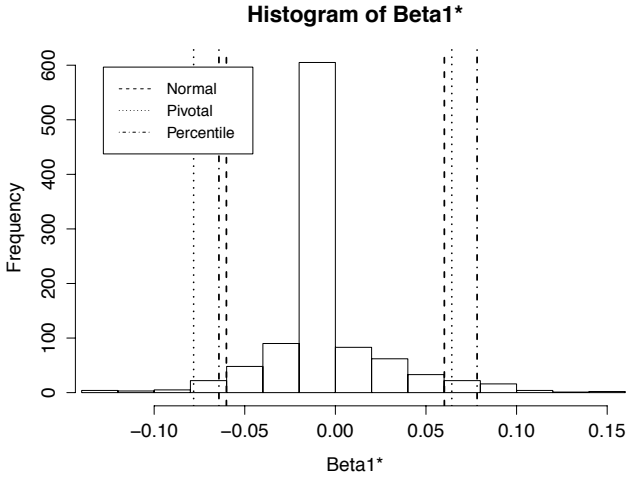
"**DewPt**", and "**WndSpd**". We note that the OLS model in problem 1 results in the same conclusion, and that the regularizations done by the Lasso (with s = 0.8 or s = 0.55) as well as that by the best subset selection rule the same predictors out.

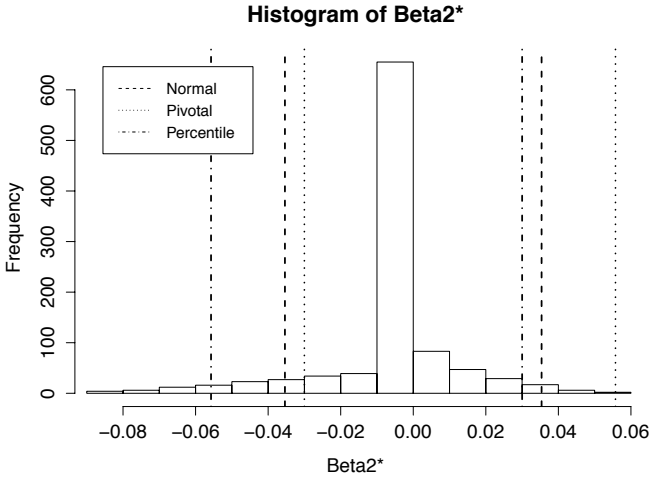
Algorithm for bootstrapping the residual:

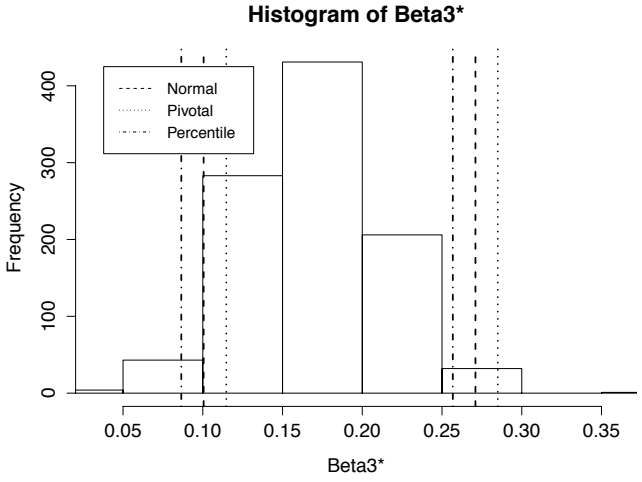
(1) $Run\ the\ Lasso\ to\ get\ \hat{\beta}\ for\ Y = X\beta + \varepsilon$
(2) $Get\ the\ residual\colon \varepsilon = Y - X\hat{\beta}$
(3) $get\ the\ bootstrap\ \varepsilon^*$
(4) $Get\ the\ bootstrap\ esimate\ on\ \hat{Y}^*\colon \hat{Y}^* = X\hat{\beta} + \varepsilon^*$
(5) $Run\ the\ Lasso\ to\ get\ \hat{\beta}^*\colon \hat{Y}^* = X\hat{\beta}^*$
(6) $Repeat\ steps\ (3)\ to\ (5)\ 1000\ times\ to\ get\colon \hat{\beta}_1^*, \hat{\beta}_2^*, \ldots, \hat{\beta}_{1000}^*$

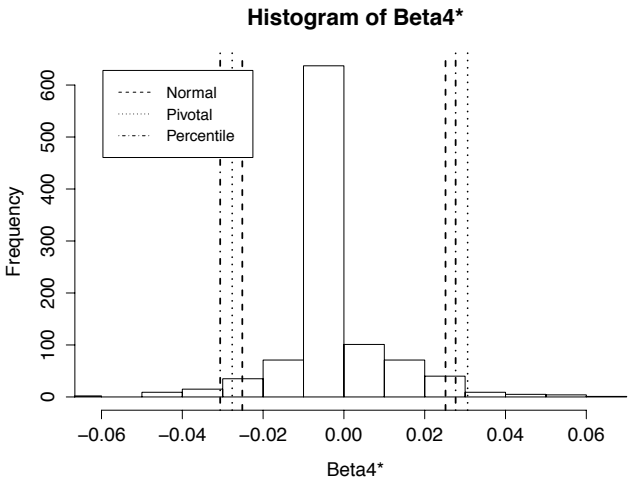
The Lasso with s = 0.8	β_0 **"intercept"**	β_1 **"Temp"**	β_2 **"DewPt "**	β_3 **"CldCvr"**	β_4 **"WndSpd"**	β_5 **"Precip"**
$\hat{\beta}_i$	5.2661	0.0000	0.0000	0.1857	0.0000	0.2661
Bootstrap Interval	β_0	β_1	β_2	β_3	β_4	β_5
Normal	(5.224, 5.307)	(-0.060, 0.060)	(-0.035, 0.035)	(0.101, 0.271)	(-0.025, 0.025)	(0.203, 0.329)
Pivotal	(5.227, 5.309)	(-0.078, 0.064)	(-0.030, 0.056)	(0.115, 0.285)	(-0.028, 0.030)	(0.222, 0.344)
Percentile	(5.223, 5.305)	(-0.064, 0.078)	(-0.056, 0.030)	(0.087, 0.257)	(-0.031, 0.028)	(0.188, 0.310)

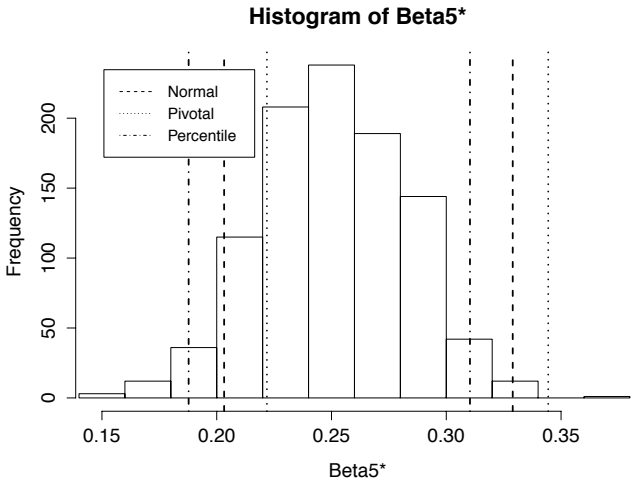






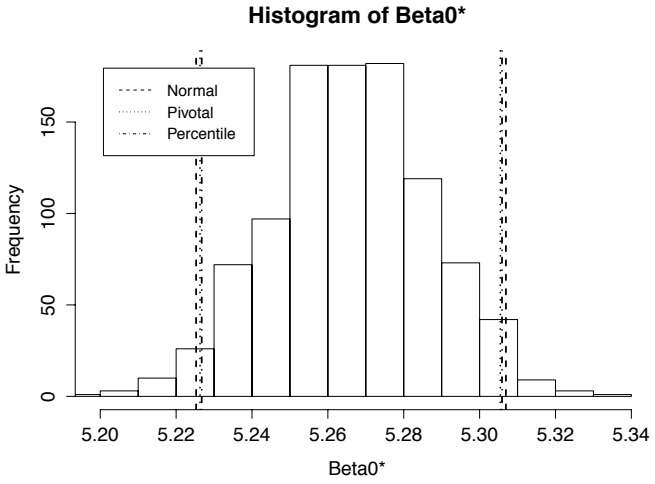


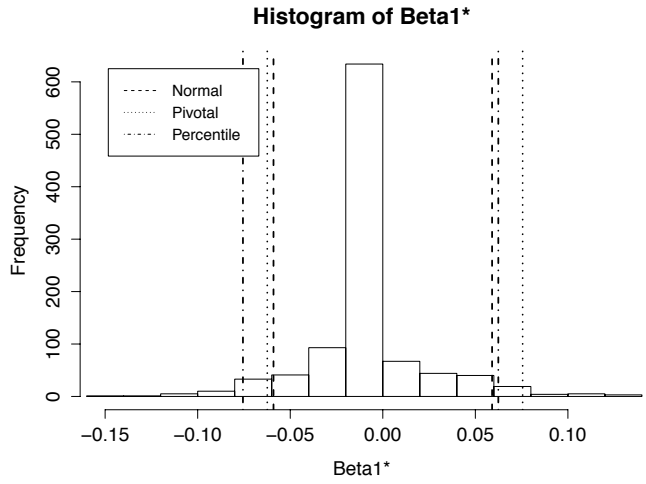


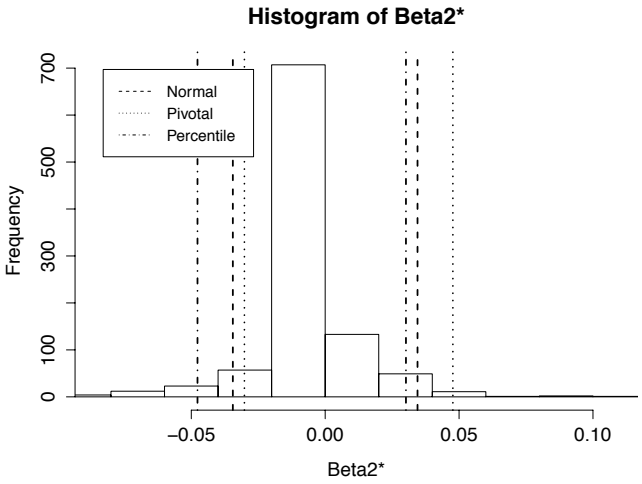


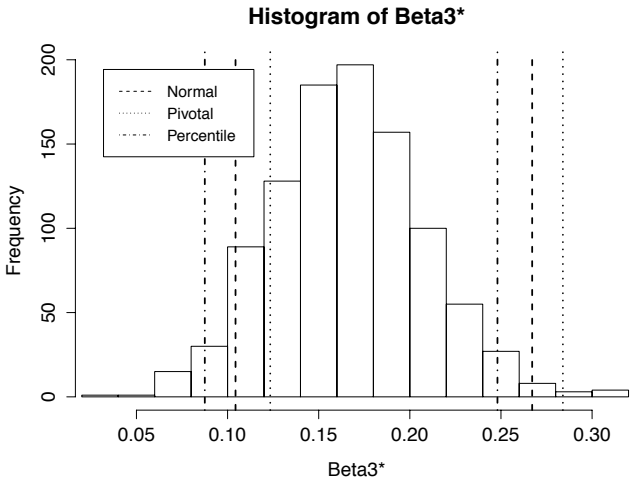
[c] Running a Gaussian parametric bootstrap allows us to make the same statement as in part [b]: the three predictors not significantly different from zero by use of any of the three confidence intervals are "**Temp**", "**DewPt**", and "**WndSpd**".

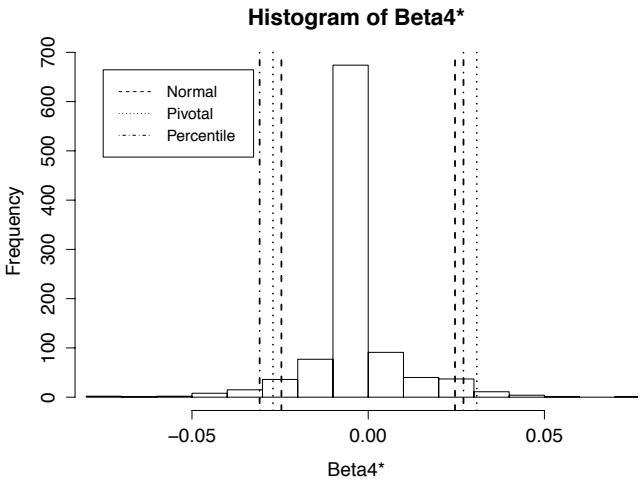
The Lasso with s = 0.8	β_0 **"intercept"**	β_1 **"Temp"**	β_2 **"DewPt "**	β_3 **"CldCvr"**	β_4 **"WndSpd"**	β_5 **"Precip"**
$\hat{\beta}_i$	5.2661	0.0000	0.0000	0.1857	0.0000	0.2661
Bootstrap Interval	β_0	β_1	β_2	β_3	β_4	β_5
Normal	(5.225, 5.307)	(-0.059, 0.059)	(-0.034, 0.034)	(0.104, 0.267)	(-0.025, 0.025)	(0.203, 0.330)
Pivotal	(5.226, 5.305)	(-0.062, 0.076)	(-0.030, 0.048)	(0.123, 0.284)	(-0.027, 0.031)	(0.215, 0.341)
Percentile	(5.226, 5.306)	(-0.076, 0.062)	(-0.048, 0.030)	(0.088, 0.248)	(-0.031, 0.027)	(0.191, 0.317)

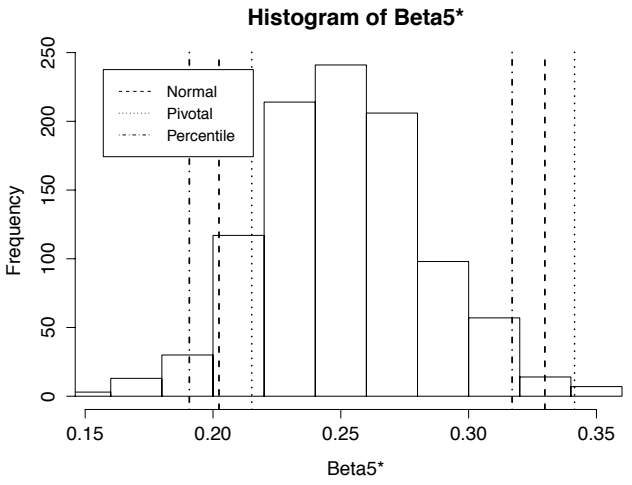




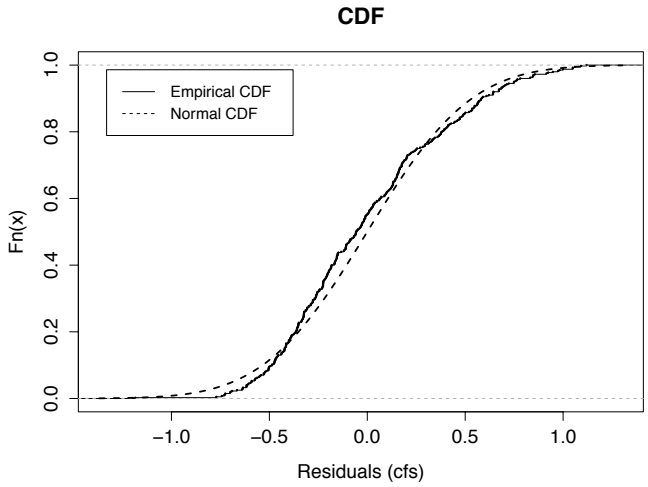






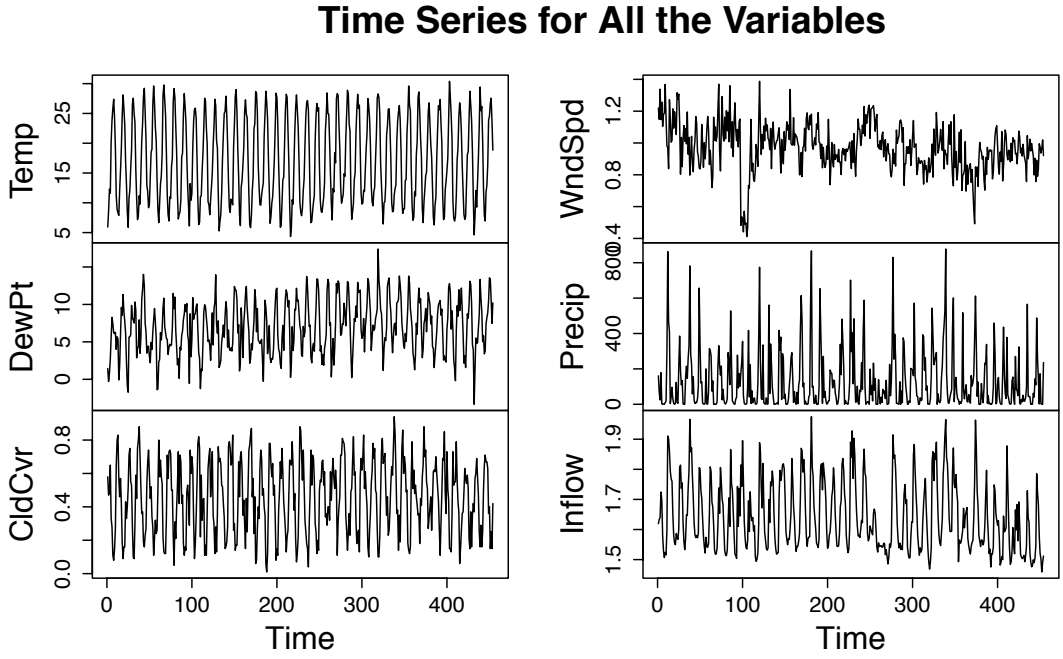


[d] Comparing part [b] and part [c], we note that the confidence intervals for all six betas are similar, and that the final conclusion is the same: not rejecting the null hypothesis that the coefficients of "Temp", "DewPt", and "WndSpd" are zero. In order words, if after assuming that the residuals are normally distributed, we get similar results, then the normal model is a decently valid assumption for the residuals. As a last comparison, we see that the normal CDF tracks the empirical CDF quite well. (We also note that the empirical data has indeed a slight positive skew, noted by the fact that the empirical CDF reaches the median before that of the normal).

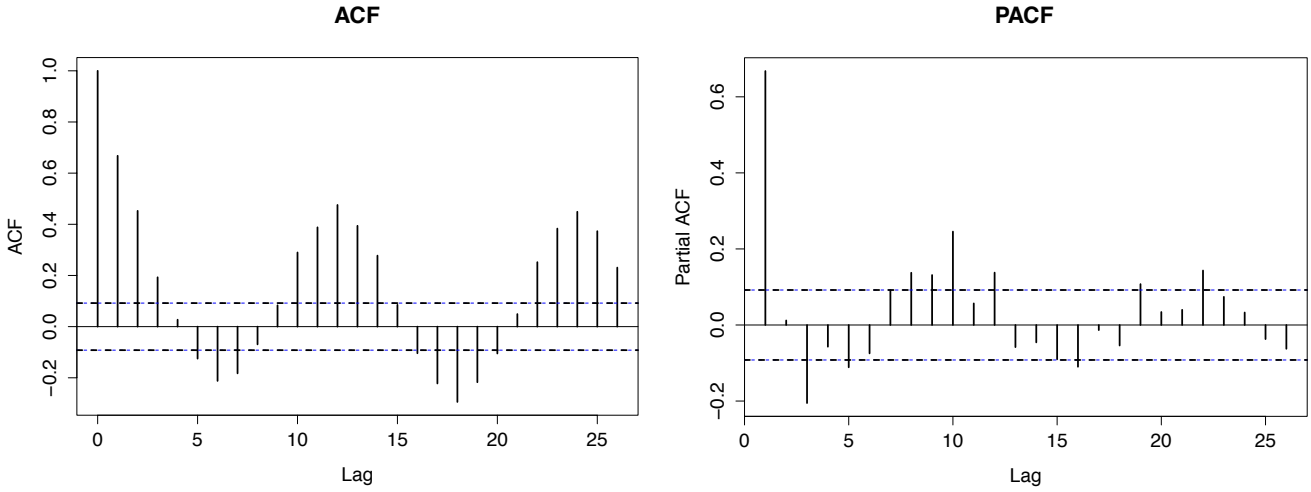


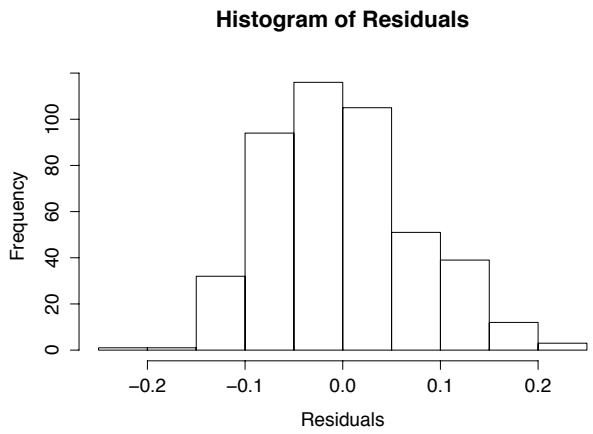
[3] Data with Correlated Errors ----------

[a] The time series plots indicate that "WndSpd", "Precip" and "Inflow" are likely non-stationary, as noted by the varying standard deviation size and (for "WndSpd") the slightly negatively sloping trend. "Infow" also appears to be non-stationary for similar reasons, though only slightly (and thus the characteristic is debatable). "Temp", "DewPt", and "CldCvr" appear to be non-stationary, as all seems to be following a periodic trend (that is relatively frequent in the given time domain); "DewPt" also seem to posses a slightly positively sloping trend. The strong periodic movement is discussed further in part [b] (though the discussion will be on the residuals from an OLS model).

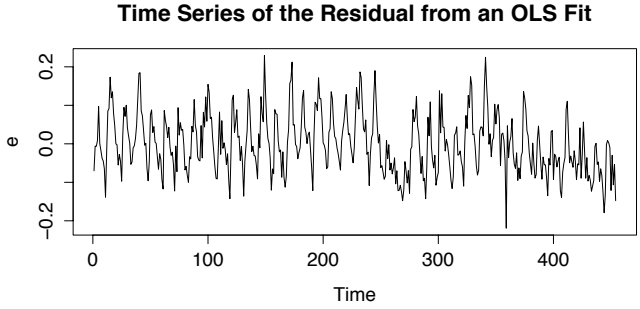


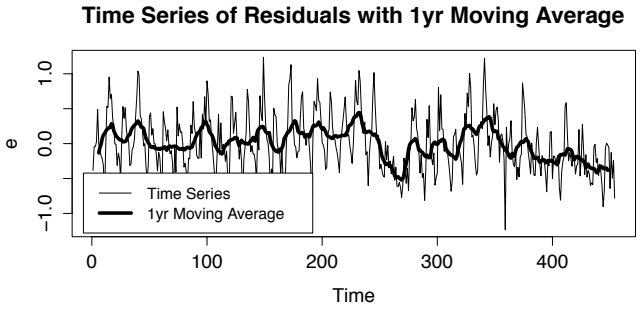
[b] The ACF tells us that the residuals from an OLS model have significant autocorrelations and a clear 12-month periodicity (and thus the residuals are non-stationary). As the PACF shows, even after removing the autocorrelations of the in-between lags, there still remains significant autocorrelations and a noticeable roughly 10-month periodicity. This autocorrelation is a phenomenon we would not see by only plotting a histogram of the residuals, which appears to be normally distributed. The patterns in both ACF and PACF also hint that there is likely no AR, MA, or ARMA model that can fit these residuals well; part [c] address this aspect further.

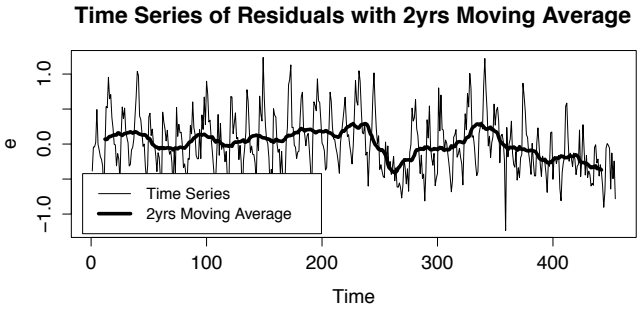




[c] Two lag sizes are tested in an attempt to fit a moving average model: 1 year and 2 years; both are multiples of the 12-month timeframe, which appears to be a period that could contribute to the non-stationarity of the residuals (part [b]). Unlike differencing, fitting a moving average model does not directly attempt to remove the non-stationarity, and is simply a smoothing technique; that said, it could provide insights into the potential trends and periodicity of the process at hand. For our residuals, there appears to be no clear trend, though both the dip around year 275 and the slightly downward trajectory after around year 350 become more pronounced.



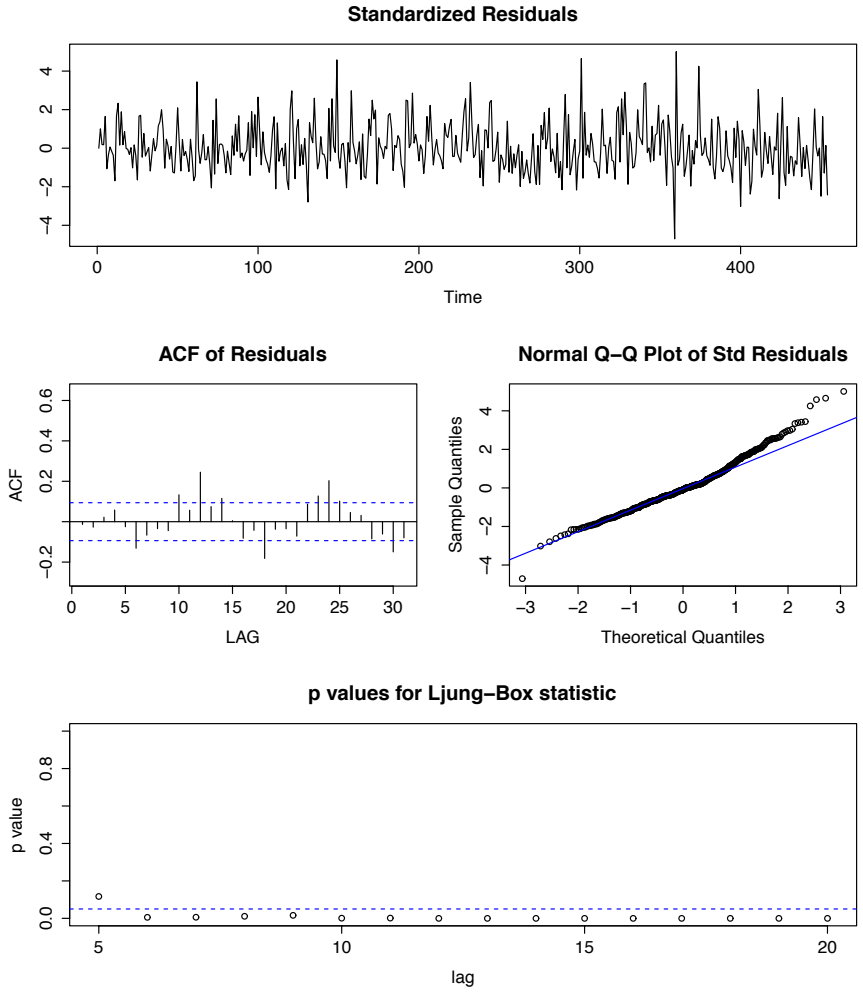




[d] *auto.arima()* suggest a model with parameters (p=3, d=1, q=1). That said, except for lag 5, all of the lags have p-values small enough to reject that null hypothesis that the autocorrelations up to and including that lag is zero; in other words, autocorrelation remains, and thus the ARIMA(3, 1, 1) does not seem to fully explain the underlying phenomenon. Other models (not shown here) were attempted on the full time length, as well as on a truncated version (namely before the "big dip": 0 to 250), yet none appears to be a decent fit of the time series so as to remove all autocorrelations. More advanced techniques could be performed, and perhaps more substantiated statements can be made on this time series after transferring it to the frequency domain.

Coefficients:

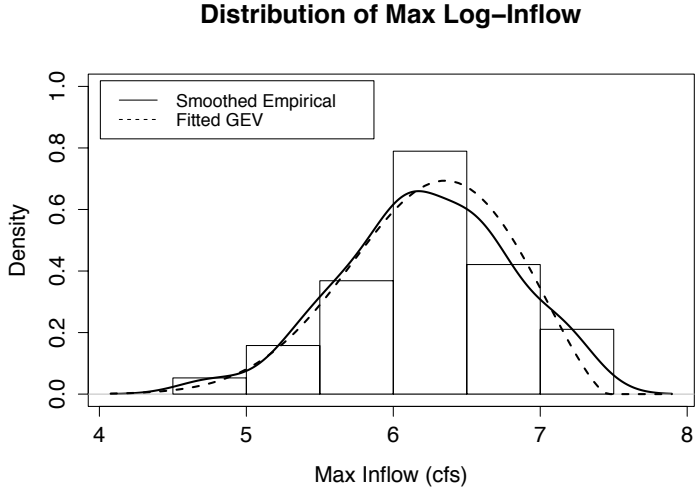
	ar1	*ar2*	*ar3*	*ma1*
	0.634	*0.125*	*-0.207*	*-0.978*
s.e.	*0.047*	*0.055*	*0.047*	*0.015*



[4] Extreme Value Analysis ----------

[a] The parameters of the fitted generalized extreme value (GEV) distribution are as follows:

Location, μ = 6.0624 (SE = 0.10636)
Scale, σ = 0.5896 (SE = 0.07934)
Shape, ξ = -0.4237 (SE = 0.12492)



[b] The fitted GEV and the empirical distribution result in values that are about 0.037 apart, and there are valid arguments for using either (or both). Both the CDF and the Q-Q plots shows the GEV model tracking the empirical values nicely. We also note that we only have 38 values in our dataset, suggesting that more data could potentially allow for a better fit of a GEV model. However, problem 1 to 3 tells us that the inflow has correlated errors and is non-stationary, thereby violating the i.i.d assumption in the GEV model; this observation could lean us towards using the empirical distribution instead. Finally, my take is that if having a max log-inflow above 7 is "dangerous" and if want to make a more cautious statement (say in designing for mitigation of this inflow), we would also use the empirical distribution, which gives a higher probability of seeing such phenomenon. Therefore, I would place more confidence in the empirical value.

P[Max Inflow ≥ 7 | GEV] = 0.06863
P[Max Inflow ≥ 7 | Empirical] = 0.1053