6010 Project

Johan Boer

12/5/2021

Description:

This project is an exploratoration of the predictive relationship of the temperature of one location to the temperature of another. This relationship will be analysed using linear regression. In particular this project is going to look at the effect of the temperature of several places around Utah on the temperature of the Salt Lake airport.

The Data:

The data used for this project all comes from the temperature records from NOAA around several locations in Utah. Each weather station records a maximum and minimum temperature each day. For this project, I have collected the data from the first 6 months in 2004 from 5 weather stations shown in Fig.1. There is one station near the Salt Lake airport, one near Provo, Logan, and Cedar City, and another weather station in the Uinta mountains near King's peak. Each of these locations have a unique feature relative to the Salt Lake airport. The station in Logan has a very similar elevation, while the Station in the Uintas is at a much higher elevation. Provo has the nearest station and Cedar City has the farthest.

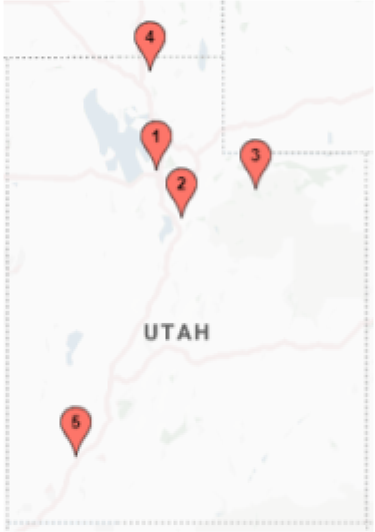


Fig.1

I will use the average temperature for each day in each place as the data used in my regression analysis. The average temperature for this project is defined as the mean of the minimum and maximum temperature of a particular day. There is no missing data at any of the chosen locations in the first six months of 2004, and as a result each location has 182 observations for average temperature. The particular stations used in this analysis each have a station code; the list of stations used is: USC00421918, USW00024127, USS0010J30S, USS0011J70S, and USW00093129.

Study Question:

The primary question is: how do the temperatures at the other stations relate to the temperature at the Salt Lake airport. Is a linear model a good representation of the relationship? If so, what is the best model? I also want to leverage the unique characteristics of each station to get a basic understanding of what factors, in general, make a good predictor of temperature.

The Model:

To try to answer the question of how temperatures relate I want to use a linear model. As a first step, I will plot a series of pairwise linear regressions to get a general idea of how the data looks. The simple linear regression plots are in Fig. 2. The pairwise data seems to indicate that the data is linear. The data does not seem to have any clear curvature, which is a good indication a linear model with only first order terms might be a good fit. Also, each term seems to have a linear

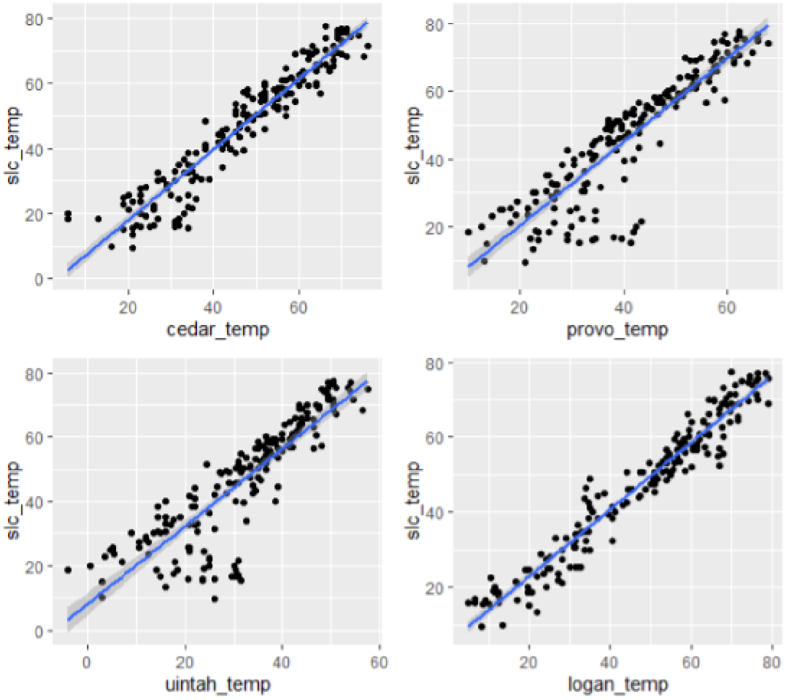


Fig. 2

relationship so there is some evidence that all the variables are informative.

I continue by creating a multivariate linear model of the form:

$$T_{SLC} = \beta_0 + \beta_1 T_{Cedar\ City} + \beta_2 T_{Provo} + \beta_3 T_{Uinta} + \beta_4 T_{Logan}$$

After running the regression I get the following:

Coefficients	Estimate	Std. Error	T Value	P(>|t|)
Intercept	-2.07076	1.04962	-1.973	0.050069
Cedar City	0.26516	0.06241	4.249	3.47e-05
Provo	0.32028	0.08930	3.586	0.000434
Uinta	-0.12364	0.08453	-1.463	0.145328
Logan	0.58202	0.03127	18.610	< 2e-16

The residual standard error is 3.291 with 178 degrees of freedom and the model has an adjusted R-squared of 0.9687.

By using the Bonferroni inequality I can take the joint 99% confidence interval for each individual variable to get a 95% confidence set. The 99% Confidence intervals are

	Intercept	Cedar City	Provo	Uinta	Logan
Lower	-4.8038661	0.1026585	0.0877460	-0.3437348	0.5005883
Upper	0.6623499	0.4276693	0.5528190	0.0964630	0.6634606

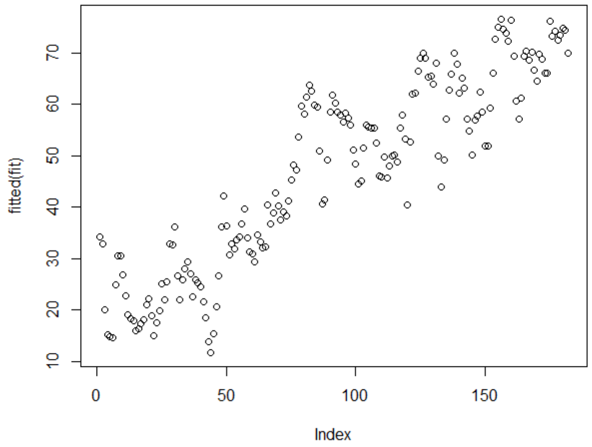
So, the joint confidence set is the union of the above.

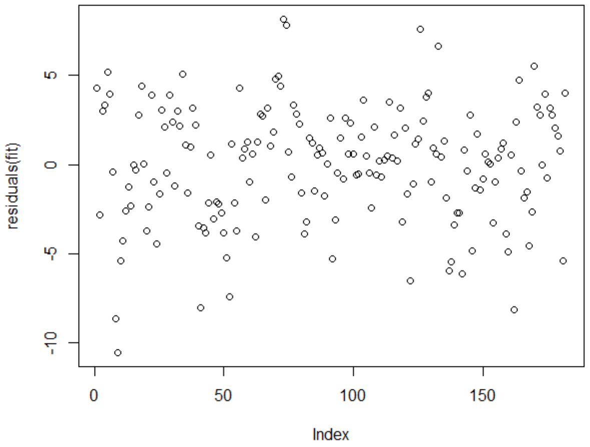
From both the P-value and the fact that zero is in the confidence set for both the intercept and the beta associated with Uinta, it is clear that neither the intercept or the predictor for Uinta is significant with a 95% total significance level.

If I run a modified model where I exclude the data from the Uinta station, the other parameters have very similar estimators and nearly the same R-squared value. Removing the temperature data from the Uinta station does not really affect the model, so choosing which of the two models is better is not obvious. While there is an argument to use the simpler model that excludes insignificant variables, there is also a valid argument to use the bigger model for a slightly better residual standard error and adjusted R-squared. I think there is utility in using the full model particularly in regards to more broad points about the data. For this project I will report the full model.

Plots and Comments:

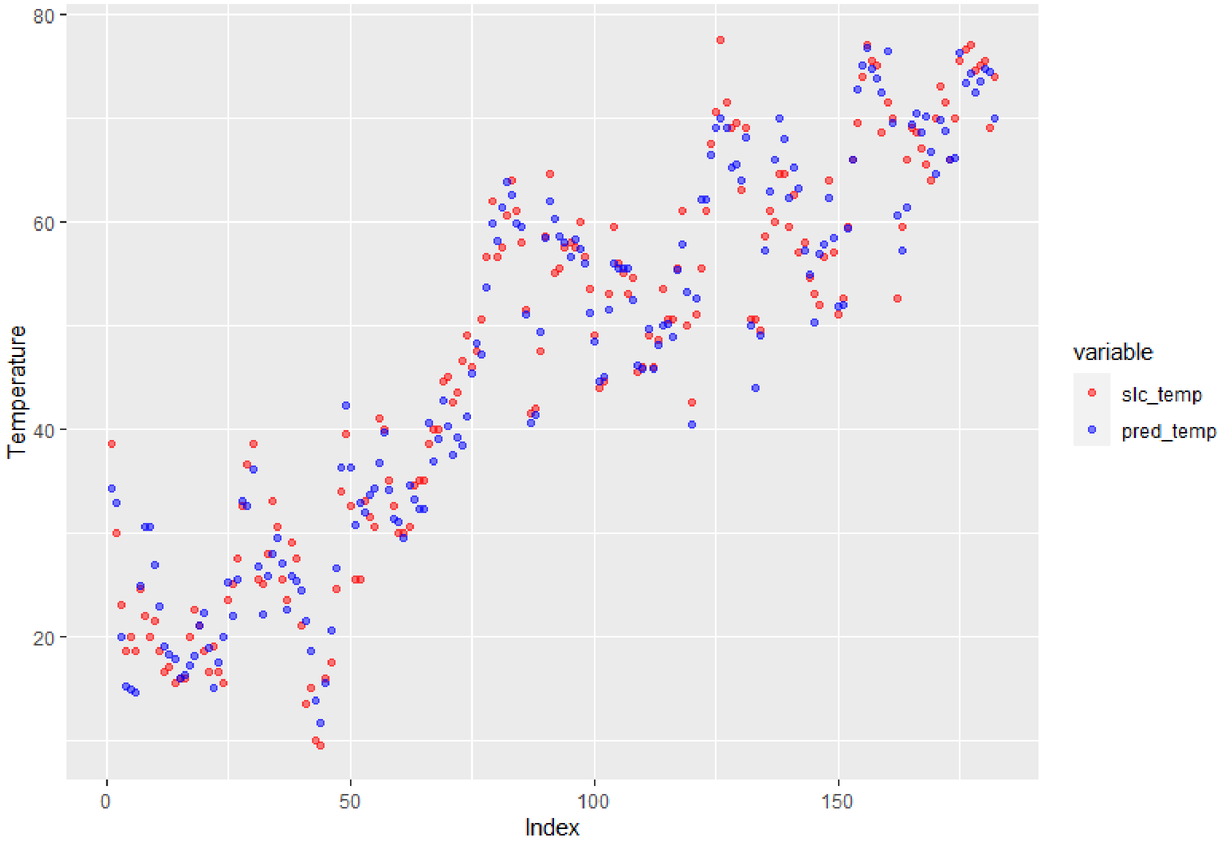
The first pair of plots I will look at is the plot of predicted values and the plot of the residuals index by date (1 corresponds to January 1st and 182 corresponds to June 30th).





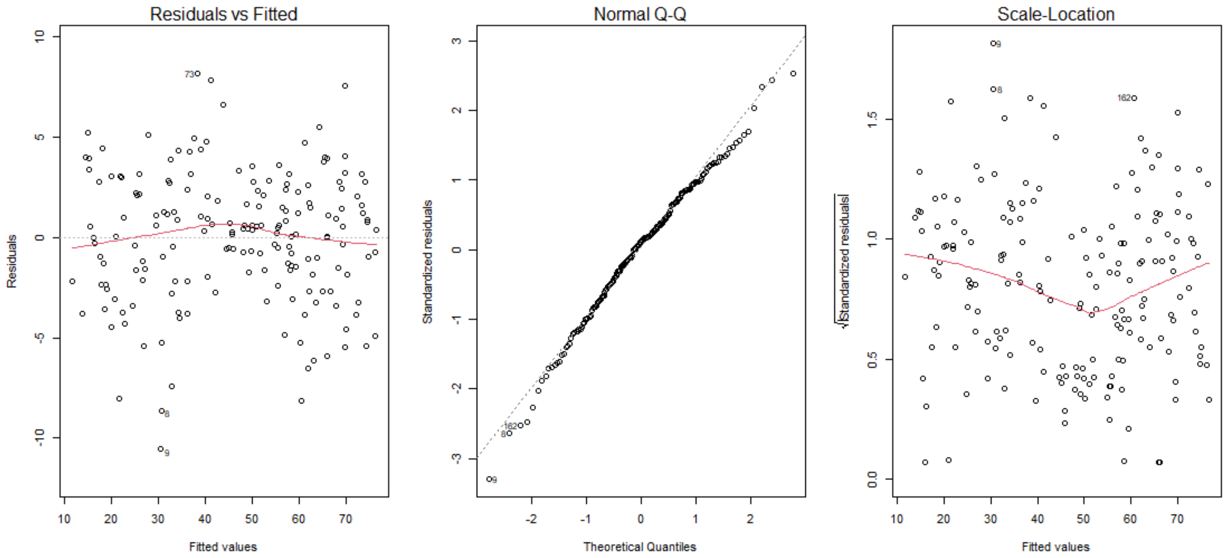
The predicted values behave mostly as expected, in January the temperature in SLC should be colder than in June and the temperature increases in an approximately linear fashion. The residuals look good. The mean is near zero and the variance looks constant everywhere. There also isn't any clear pattern, so the assumption that the errors are effectively random seems to hold.

The next plot is a representation of the observed data compared to the fitted values again indexed by date.



This plot shows that the fitted values do approximate the observed values, confirming that this model is a good fit.

Next I will show some diagnostic plots to make sure the assumptions of linear regression are met.



The first plot shows that the residuals and the fitted values are uncorrelated and indicates that there is a linear relationship. The Q-Q plot indicates that the distribution of the residuals is normal. The Salce-Location plot indicates that the variance of the residuals is lower around 50 degrees, but not by a lot. The assumption of homoscedasticity is supported by the mostly flat Scale-Location plot and the Residual vs Fitted plot, although the assumption is not perfect. Overall these plots support the assumptions of linear regression using the ordinary least squares and normally distributed errors.

Conclusion:

This analysis shows that the temperature of the Salt Lake airport can be modeled well by using a linear combination of the temperatures of other points around Utah. The pairwise linear models suggest that choosing an arbitrary point should give useful information about the temperature at the Salt Lake airport in a linear relationship, but as the multiple linear model showed, selecting many arbitrary points may not improve the model by a significant amount.

Looking at the coefficients of the multiple linear regression model, I notice some trends. Since I only have 4 other locations nothing conclusive can be said about the model, but we can get an inclination about what factors make a good predictor location. The highest correlation was between Salt Lake and Logan, and the lowest was between Salt Lake and the station in the Uintas. The elevation of Logan and Salt Lake is almost the same, and the Uintas has a much higher elevation, so it seems like differences in elevation have a large impact on the correlation of temperature. The next lowest correlation was between Salt Lake and Cedar City, the farthest location away. This would indicate that distance away may also be an important factor in the ability to predict the temperature.

The temperature of one location can be modeled well by the temperature of different locations, but factors like distance and change in elevation may affect how much predictive information you can get from a particular location.

Code:

This project made use of some R code given on the next page.

```
library(ggplot2)
library(gridExtra)
library(reshape2)

# reading and manipulating the data

weather = read.csv("C:\\Users\\Johan\\Documents\\6010Project\\2801622.csv")


# splitting the data into subsets based on location
weather.cedar = subset(weather,STATION=="USW00093129")
weather.provo = subset(weather,STATION=="USS0011J70S")
weather.uintah = subset(weather,STATION=="USS0010J30S")
weather.slc = subset(weather, STATION=="USW00024127")
weather.logan = subset(weather, STATION=="USC00421918")


# collect only the mean temperature into a single dataframe
cedar_temp = weather.cedar$TAVG
provo_temp = weather.provo$TAVG
uintah_temp = weather.uintah$TAVG
logan_temp = weather.logan$TAVG
slc_temp = weather.slc$TAVG

temperatures = data.frame(cedar_temp,provo_temp,uintah_temp,logan_temp,slc_temp)



# create exploratory plots

p1 = ggplot(temperatures,aes(cedar_temp,slc_temp))+
 geom_point()+
 geom_smooth(method='lm')

p2 = ggplot(temperatures,aes(provo_temp,slc_temp))+
 geom_point()+
 geom_smooth(method='lm')

p3 = ggplot(temperatures,aes(uintah_temp,slc_temp))+
 geom_point()+
 geom_smooth(method='lm')

p4 = ggplot(temperatures,aes(logan_temp,slc_temp))+
 geom_point()+
 geom_smooth(method='lm')
```

```
grid.arrange(p1,p2,p3,p4,nrow=2)



# create and analyze a linear model

fit = lm(slc_temp~cedar_temp+provo_temp+uintah_temp+logan_temp,
data=temperatures)
summary(fit)

confint(fit,level=.99)

# alternate model where all variables are significant
fit1 = lm(slc_temp~cedar_temp+provo_temp+logan_temp, data=temperatures)
summary(fit1)

confint(fit1,level=.99)


# plot data

pred_temp = fitted(fit) # the predicted values

layout(matrix(c(1,2),1,2))
plot(fitted(fit1))
plot(residuals(fit1))

# fitted and observed data
pf = data.frame(index = c(1:182),slc_temp,pred_temp)
pf = melt(pf,id.var=1)
ggplot(data=pf, aes(index,value,color=variable))+
  geom_point(alpha=.5)+
  scale_color_manual(values=c("red","blue"))+
  labs(x="Index",y="Temperature")


# diagnostic plots
layout(matrix(c(1,2,3),1,3))
plot(fit)
```