Residual_Discussion

Zachary Palmore

4/6/2021

Prompt

Using R, build a regression model for data that interests you. Conduct residual analysis. Was the linear model appropriate? Why or why not?

Data

This data was collected from a radiosonde during an upper air sounding at MPX in Chanhassen, MN on January 1. The data has been cleaned and reduced to display only the variables of interest: ambient air pressure in hectopascals (hPa), radiosonde height in meters (m), and the temperature at each of those points recorded in degrees Celsius (C). A sample of the data is shown below.

```
data <- data.frame(read.delim("https://raw.githubusercontent.com/palmorezm/msds/main/605/upperair_72747_
colnames(data) <- c("Pressure", "Height", "Temperature")
head(data)
```

```
##   Pressure Height Temperature
## 1     1000    193         0.0
## 2      988    287        -6.1
## 3      984    319        -5.9
## 4      948    610        -8.0
## 5      935    717        -8.7
## 6      925    801        -7.1
```

```
tbl <- rmarkdown::paged_table(head(data))
write.table(tbl, file="C:/data/tbl.png")
```

A summary of the data shows that minimum height is 193. However, we know that the elevation of the local area is about 287m. Notice that there are no other observations at 193. This implies we could simply exclude it from data set since it is not a true measurement but perhaps a value that appeared during the instrument setup. This summary and the exclusion of the underground observation is shown below through the removal of the first row.

```
summary(data)
```

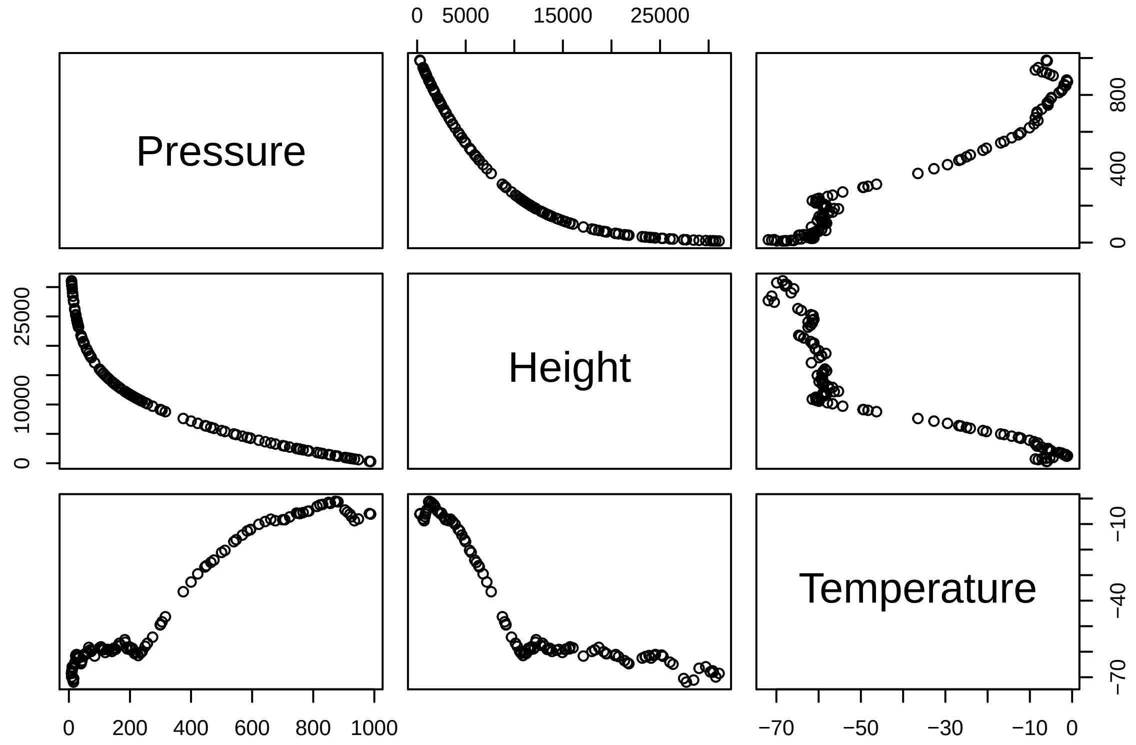
```
##     Pressure          Height       Temperature
##  Min.   :   8.6   Min.   :  193   Min.   :-71.90
##  1st Qu.:  59.3   1st Qu.: 4080   1st Qu.:-60.95
##  Median : 203.0   Median :11597   Median :-58.30
##  Mean   : 341.1   Mean   :12332   Mean   :-41.72
##  3rd Qu.: 608.5   3rd Qu.:19309   3rd Qu.:-11.10
##  Max.   :1000.0   Max.   :31073   Max.   :  0.00
```

```
data <- data[-1,]
```

Variable Selection

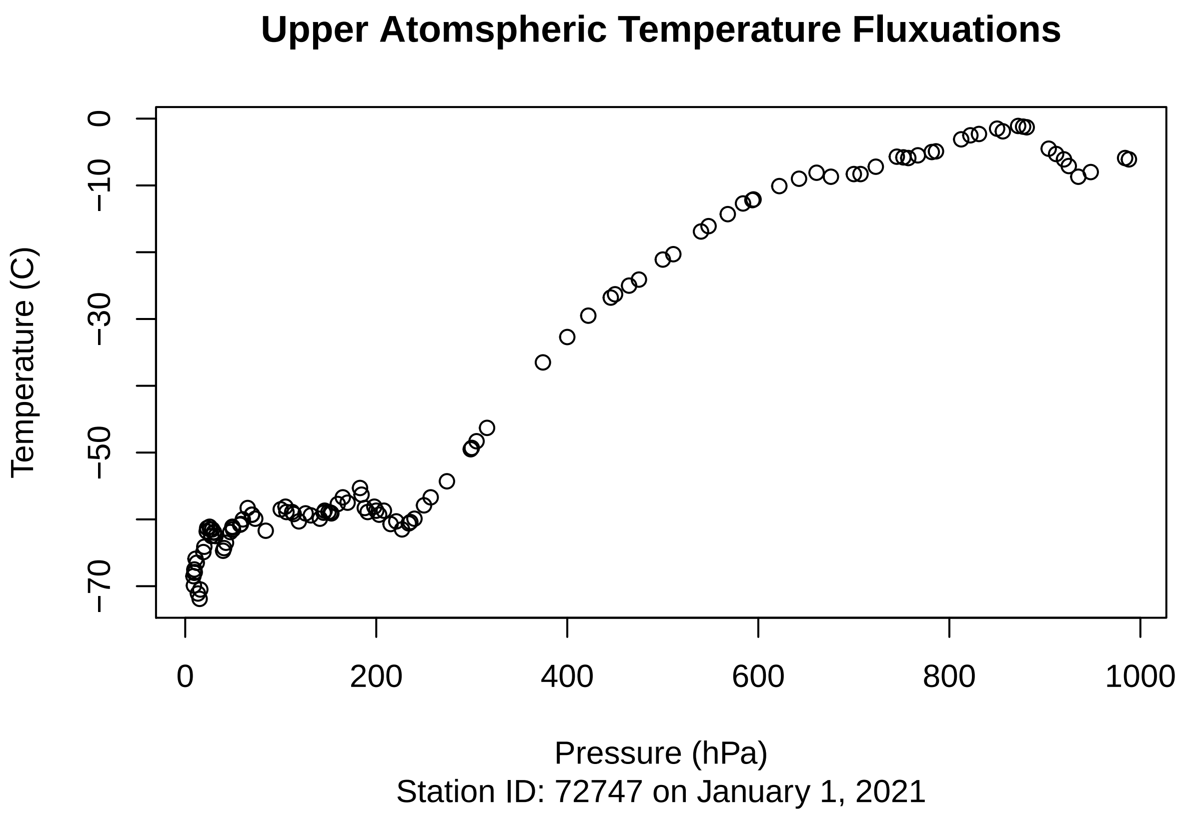
Now we plot the data to see which variables interest us most. The interaction of these three variables should demonstrate some physical constants. For example, there should be a decrease in pressure with height as the radiosonde move farther from earth and there should be a clear variation in the layering of temperature with height as well (to indicate layers of the atmosphere). Results are shown with all possible outcomes with the three variables plotted.

```
plot(data)
```



As expected, the results conform to the laws of physics, but we are not interested in the confirming their presence. From the chart, there are two variable combinations that are unique to this location around the twin cities of Minnesota. They are the plots of temperature and pressure which when plotted in either order describes the same atmospheric variation. For visual purposes, we will review what happens to temperature as pressure increases. Isolated from the matrix, that plot looks like:

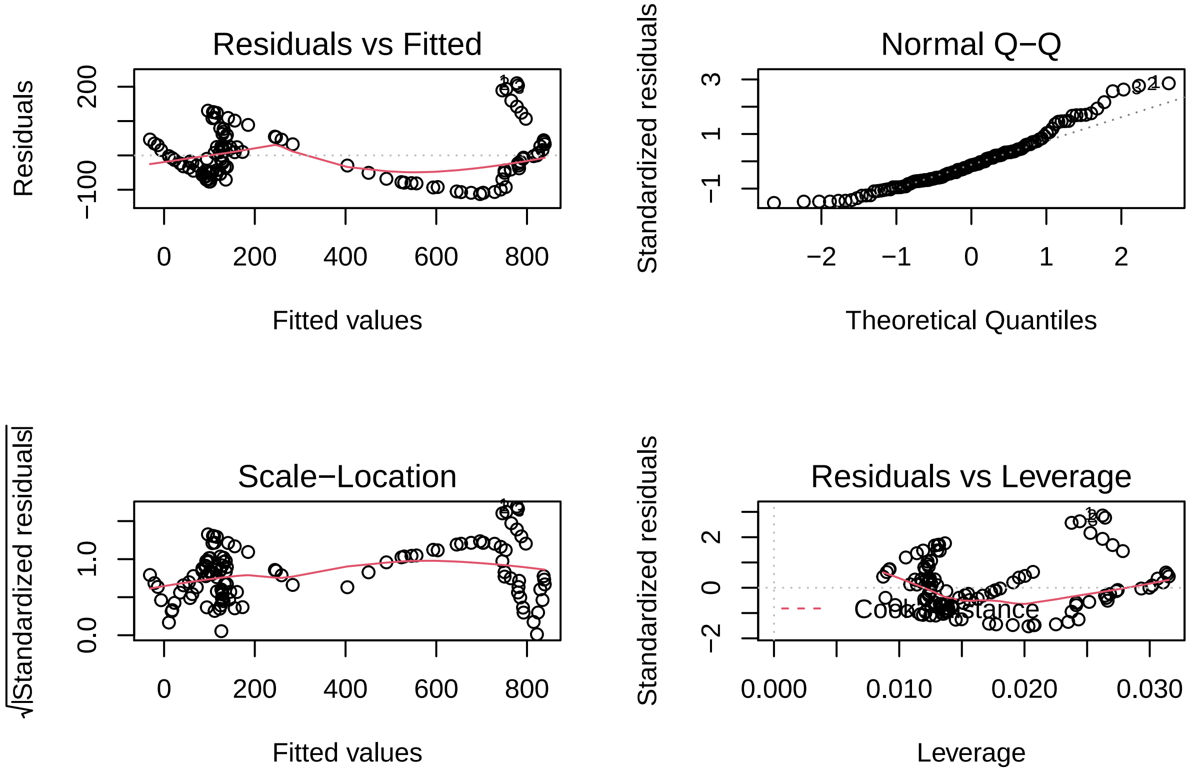
```
plot(data$Pressure, data$Temperature,
     xlab = "Pressure (hPa)", ylab = "Temperature (C)",
     main = "Upper Atomspheric Temperature Fluxuations",
     sub = "Station ID: 72747 on January 1, 2021")
```



Analysis

When we create a linear model using pressure modeled with temperature, we get the following diagnostic plots.

```
datalm <- lm(data$Pressure ~ data$Temperature)
par(mfrow=c(2,2))
plot(datalm)
```



In the Residuals vs Fitted we notice that there does not appear to be a clear pattern emerging from the plot. Although there are two distinct portions of the trend line in red. The first beginning with a steady increase at a constant slope until a point where the data changes into a wide 'u' shape. In my opinion, this is a close enough fit to a linear model that I would only transform it and repeat the analysis.

From the Normal Q-Q plot we see that in the distribution of standardized residuals over theoretical quantiles there is some deviation from a perfectly normal distribution which may be cause for some concern. Here again, a Box-Cox or similar transformation might have the desired side-effect of straightening this Normal Q-Q plot and normalizing the distribution. Importantly, the only major issue with the plot occurs at the higher theoretical quanitiles of this plot. The lower quantiles bear no issue.

For the scale-Location plot, the residuals are reasonably equally spread along the predictors. In other words, the trendline of standardized residuals are near flat against the fitted values. If they were not, there might appear to be a trend line extending diagonally across the plot as shown in the Normal Q-Q. This does not occur and the variation in these residuals is fairly random.

Lastly, we review the Residuals vs Leverage Plot. In it, there are no residuals with enough leverage to be considered extreme or outliers. The data is clustered randomly between the leverage values of about 0.05 and 0.35. Cook's distance lines which guide the analyst are not shown at all because the values all have such a small similar amount of low leverage. No points have substantially greater influence over the others.

Conclusion

In this case, the linear model was appropriate, but I do have some reservations. The data did not deviate much at all from expectations with a normal linear data set. Our residuals did not appear to show any clear patterns that might indicate a better model choice. Rather, those residuals were randomly distributed when compared to fitted values and in standardized residuals in our Scale-Location plot. There were no outliers or overly influential points, and the data is approximately normal. However, a Box-Cox or similar transformation would improve the normality assumption as well.

My only reservation in fitting this data using linear regression is that there may be multiple levels to the sounding that do not fit a normal Q-Q plot. If the data were grouped by pressure levels, it would be reasonable to reduce it into two groups, one from 0 – 200hPa and the other 201+hPa. This grouping would likely change the entire assessment to one that contains two parabolas. This of course, could result in a better fit with other models.

Source

University of Wyoming. Atmospheric Soundings. Retrieved April 06, 2021, http://weather.uwyo.edu/upperair/sounding.html