**AML 612 Homework #5**

**For all questions, please submit your R code, and a doc file with a copy of the screen output and plots (where applicable), and all other relevant files to** [**smtowers@asu.edu**](mailto:smtowers@asu.edu)**.**

**Due Friday March 23rd at noon (except for the in-class presentation, due Mon March 19th)**

**The code in your R files should exhibit all of the good coding practices mentioned in** [**http://sherrytowers.com/2012/12/14/good-programming-practices-in-any-language/**](http://sherrytowers.com/2012/12/14/good-programming-practices-in-any-language/)

**Plots should exhibit all of the good plotting practices mentioned in** [**http://sherrytowers.com/2013/01/04/good-practices-in-producing-plots/**](http://sherrytowers.com/2013/01/04/good-practices-in-producing-plots/)

**Please submit your files in a format hwk5\_<last name>.xxx where xxx is the format of the document (R, doc, tex, pdf, bib, etc)**

**Question 1)**

The project groups are:

Crime and student performance (prospectus #4):

   Rick

   Bechir

   Cesar

State alcohol laws and DUI fatalities (prospectus #7):

   Karen

   Juan

Bullying and academic achievement (prospectus #8):

   Mugdha

   Caleb

   Josean

Each project group will prepare a short presentation (15 minutes, plus 5 to 10 minutes for questions) on one or more analyses that were cited as background references in the original project prospectus (other analyses can be added as well, if they are relevant to the motivation and objective of the analysis). *You will be giving this presentation with the same detail as if you were the people who had actually done the analysis*. Thus, the presentation is expected to have an introduction to the topic describing the motivation and prior previous work, moving on to the objective, methods, detailed results, and discussion of how the work contributes to the body of literature on the topic. The project group will tag-team the presentation. The presentations will occur in class on Monday, March 19th.

**DO NOT GO OVERTIME IN YOUR PRESENTATIONS!**

All students in the audience will be expected to pose one question each to the presentors during the question period after each of the talks.

Tips on giving good presentations can be found at: <http://sherrytowers.com/?p=2486>

**Question 2)**

Publication of projects is encouraged, but not required in this course. However, what is required is that the projects be written up in publication quality format.

As the first step towards this, each project group needs to identify a journal that would be appropriate for publication of an analysis like theirs. Suggestions for potential journals can be obtained by looking at related work in the literature, and noting which journals those authors published in. I am also happy to give guidance… if this is needed, groups need to set up a meeting with me **at least a week in advance of the homework deadline of Friday March 23rd.**

Once a journal has been identified, go to the Instructions for Authors page on the journal website, and download the LaTex or Word template for the journal (if the journal offers both, pick LaTex). Also download any LaTex style files that the journal might require.

Using this template, add the title of the project, and the student author list and affiliation information. If the journal requires a running header title, fill that in too.

Then begin a structured Abstract that has the sections Background, Methods, Results, Conclusions. Fill in two or three sentences in the Background section that motivates the analysis.

Write the Introduction section, with the motivation thoroughly described, with relevant references. Also put a sentence or two stating the objective of the paper.

Leave all other sections (Methods and Materials, Results, Discussion, Conclusions) empty for the moment. You will be filling them in as we progress through the rest of the course.

Tips on writing papers can be found at <http://sherrytowers.com/?p=1876>

**Question 3)**

Three students gave feedback on the project prospectus documents. The feedback documents in the class Dropbox are linked to on the page: <http://sherrytowers.com/?p=3790>

Password: prospect

Go to the feedback pages in the Dropbox corresponding to your prospectus, and leave a comment on each of the pages indicating that you have read the comments.

**Question 4)**

In May, 2016, a professor obtained a Polar M400® fitness tracking watch. Since then, she has recorded the number of steps that she takes every day, with some missing data when she wasn’t wearing the watch.

The file <https://www.dropbox.com/s/c6to3zr3q929z2p/polar_m400_data.csv?dl=0> contains the daily data on the number of steps taken.

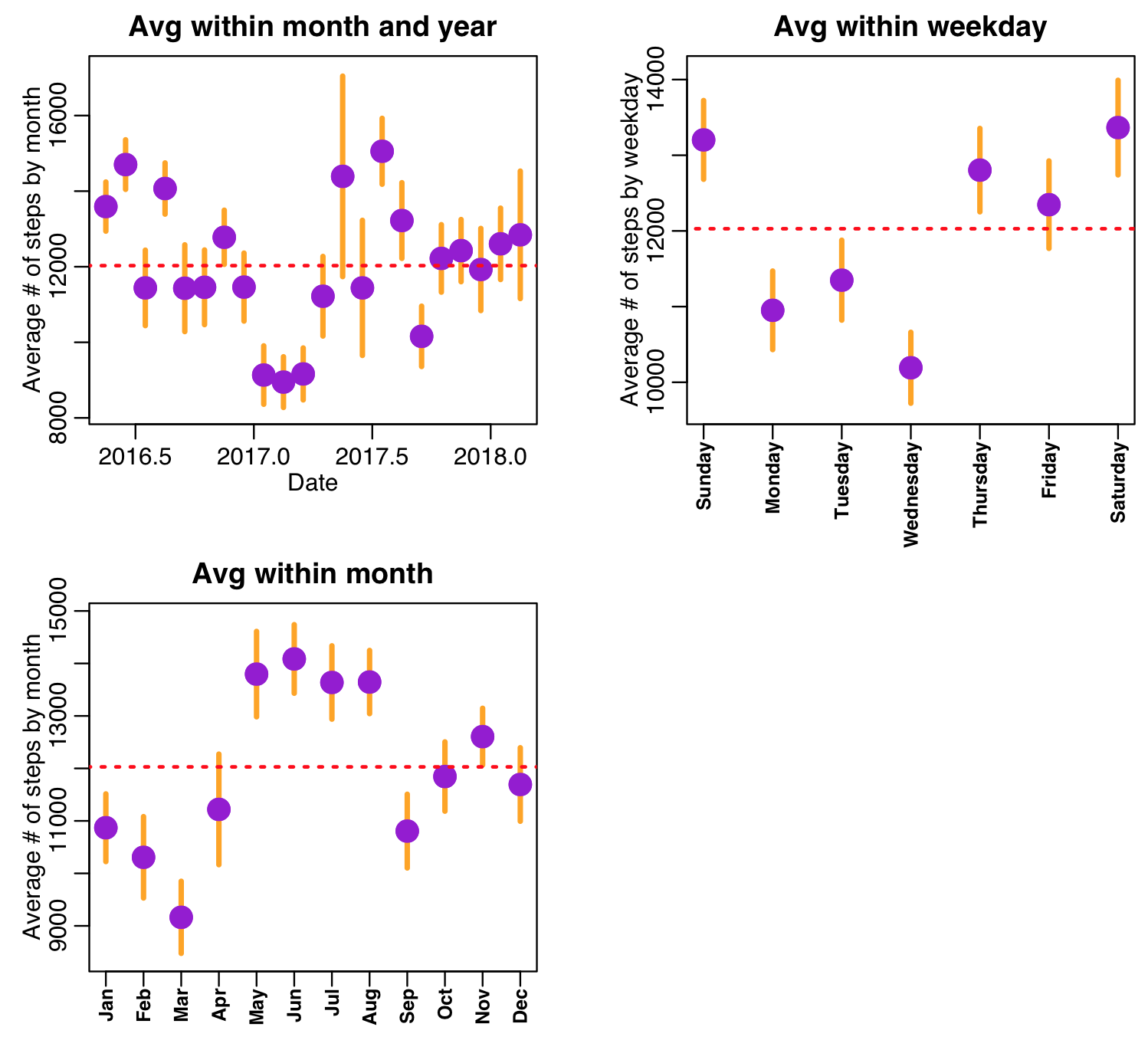
a) Average the data within date by year and month, where “date” is year+(month-0.5)/12. Also calculate the standard error on the average. Produce the plot below (using different colours than I used), where the error bars are the +/- one standard error on the mean, and the red dotted line represents the overall average. Make sure that the y-axis limits are sufficient that all error bars are fully shown.

Read the R help file for the day.of.week() function in the R chron library (install the chron library in R if you have not already done so). Using the day.of.week() function, determine the weekday of each data point. Average the data within weekday

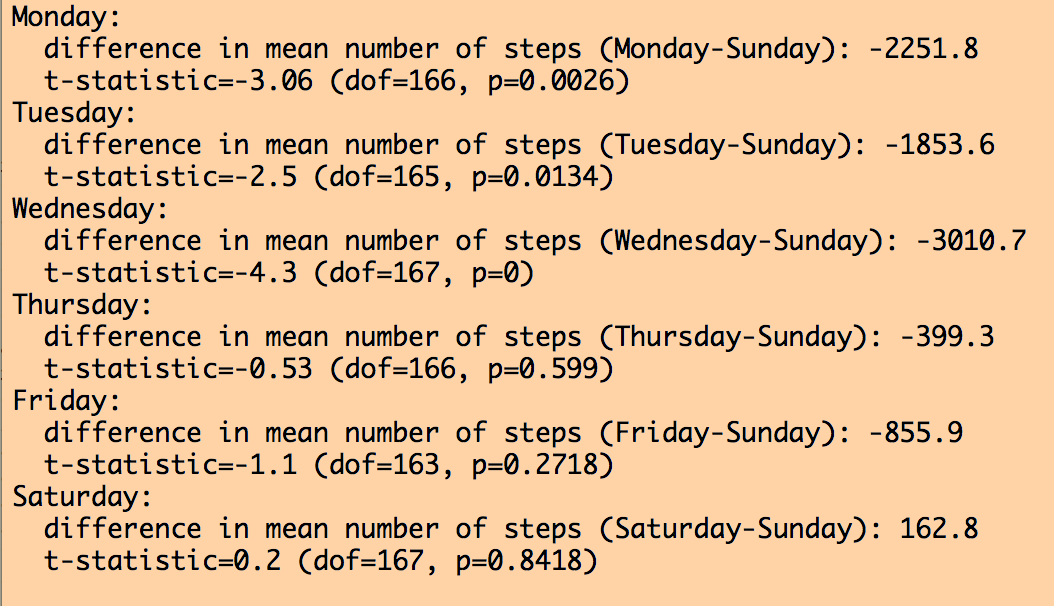
Repeat the above plot for averages within weekday.

Also repeat the above plot for averages within month.

Note that tips on specialized axis labels for plots can be found at http://sherrytowers.com/2012/12/11/the-basics-of-the-r-statistical-progamming-language/

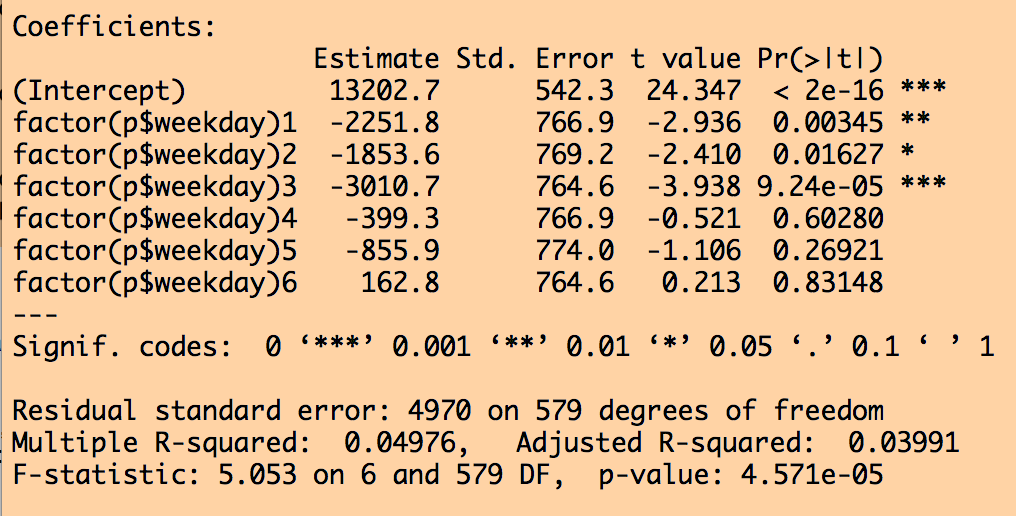


b) Using the two sample t-test under the assumption of equal variances, test whether or not the data averaged within each weekday is statistically consistent with the average on Sunday, and write the code to produce the following output:



Based on the output, which day appears to be the laziest day of the week for the professor?

c) Construct a linear regression model, regressing the number of steps on weekday as a factor. Re-produce the following output from the summary of your fit:



Is the Intercept term equal to the mean number of steps the professor took on Sundays?

Notice that the means by factor level are the same as in part(b), but p-values by factor level are slightly different than the p-values you obtained from the individual weekday t-test comparisons in part (b). Why is that?

Hint: what does the linear regression model assume about variances for ALL factor levels? What is the assumption of the variances used in the individual t-tests?

Under what conditions would the p-values from the individual comparisons be equal to the p-values from the overall linear regression fit?

Given that the p-values are nearly equal between the two methods, what does this indicate?

d) The number of steps the professor takes each day may be affected by the weather. The nearest airport to where the professor lives has airport code KLAF.

The Weather Underground website allows you to download the daily weather data for an airport a calendar year at a time. For example, the KLAF data for 2018 to present can be found at:

<https://www.wunderground.com/history/airport/KLAF/2018/1/1/CustomHistory.html?dayend=31&monthend=12&yearend=2018&req_city=&req_state=&req_statename=&reqdb.zip=&reqdb.magic=&reqdb.wmo>=

Copy and paste the weather history observations into a text file. You will need to delete the extra header lines between months (or comment them out), and add a column at the beginning of each line to indicate the numerical month.

While it is instructive for you to learn how to clean up data, making you do this for more than a few months of data is likely cruel and unusual punishment for not much gain in added insight. Thus, you can find the weather data for 2016 and 2017 in the files:

<https://www.dropbox.com/s/hi0po9o2wc6on8q/klaf_climate_2016.csv?dl=0>

<https://www.dropbox.com/s/d8nbl38imh0cwga/klaf_climate_2017.csv?dl=0>

Write the R code to read in the 2016 and 2017 data, and your 2018 data, and combine the data frames into one data frame using the R rbind() function. Hint: for ease of doing this, make the column names of your 2018 data match the column names in the 2016 and 2017 data sets.

For your climate data frame, calculate the number of julian days from Jan 1, 1970

Also, create a new variable in the data frame called levent, that is 0 if no weather event occurred (see the last column of the data frame for the various types of weather events that can occur), and 1 if some weather event occurred.

Now, for the original step count data frame, add columns temp\_avg and levent, and mesh in the data from your climate data frame using the R match() function.

e) The number of steps the professor takes each day may also be affected by the number of daylight hours in the day.

The R script <http://www.sherrytowers.com/AML_course_libs.R> contains a function

length\_of\_day(day\_of\_year,latitude)

that takes as its arguments the day-of-year (ie; Jan 1st is day\_of\_year=1, and December 31st is day\_of\_year=365), and the latitude, in degrees.

The professor lives at latitude 40.42 degrees

A failing of R is that it doesn’t have a nice, one-package-fits-all, set of simple functions for handling dates. In <http://www.sherrytowers.com/AML_course_libs.R> I have written a function for you

convert\_month\_day\_year\_to\_date\_information(month,day,year)

that converts month, day and year into a variety of date-related information, like the

date expressed as a fraction of the year (date), the weekday, the number of days since 1/1/1970 (jul), and the day\_of\_year.

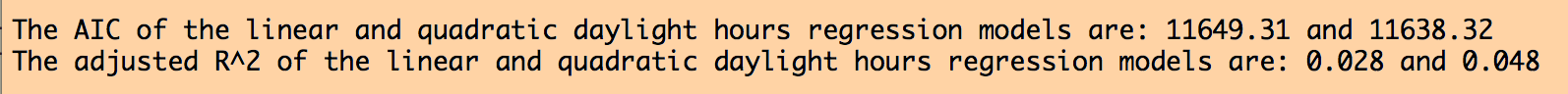
Using this function, calculate the day\_of\_year for each day of data for which there were step counts.

Then use the length\_of\_day() function to calculate the number of daylight hours. Put this information in the step count data frame as daylight\_hours

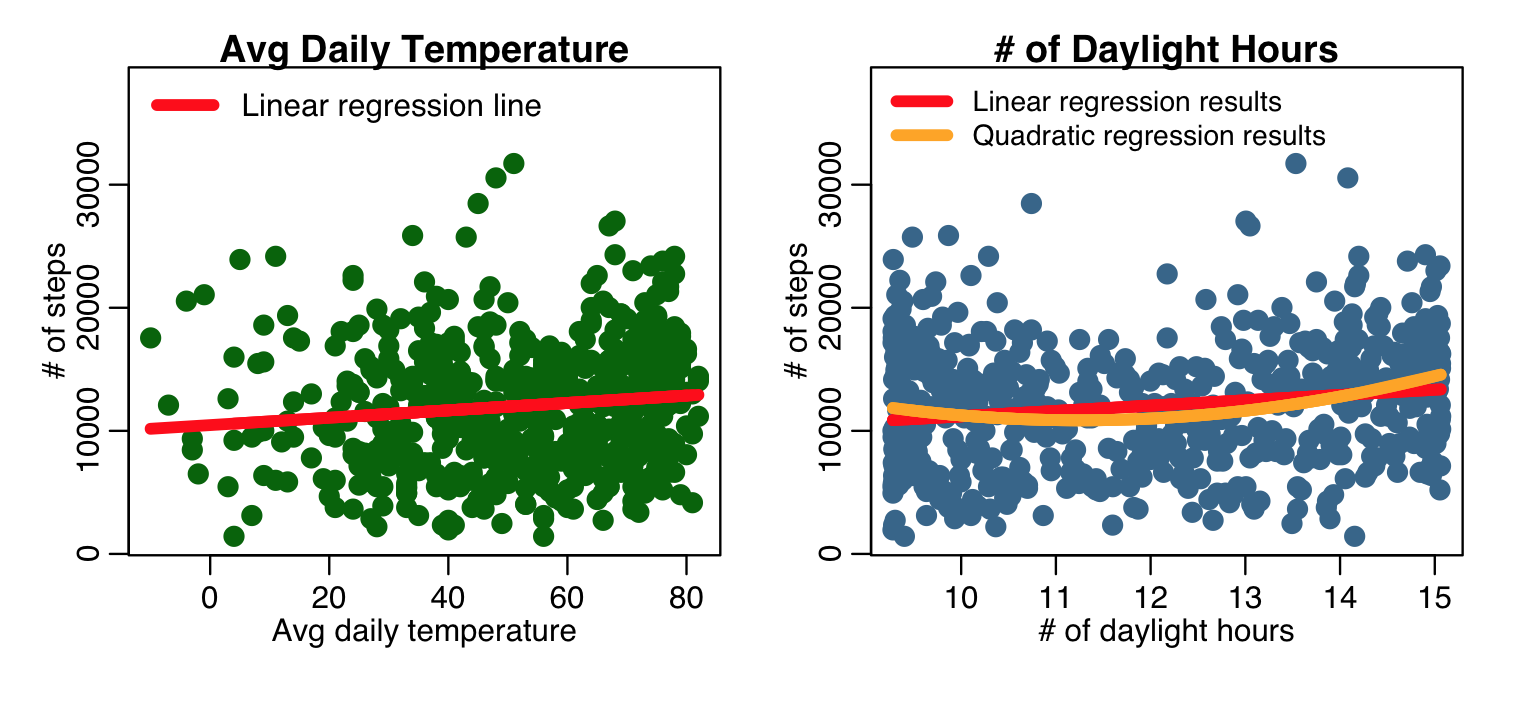
f) Now that you have the temperature, weather event, and daylight hours in the step count data frame, reproduce the plot below, using different colours than what I’ve used here (note that I adjusted my y axis range such that I could fit the legend over the plot without having points overlay it… this is something you should do for a publication-quality plot).

Overlay the linear regression line on the plot steps vs temperature and steps vs daylight hours. Note that there seems to be some funny business going on with the daylight hours plot… the residuals of the linear regression line are definitely not homoscedastic (they are spread wider for low and high daylight hours than they are for middle-values of daylight hours).

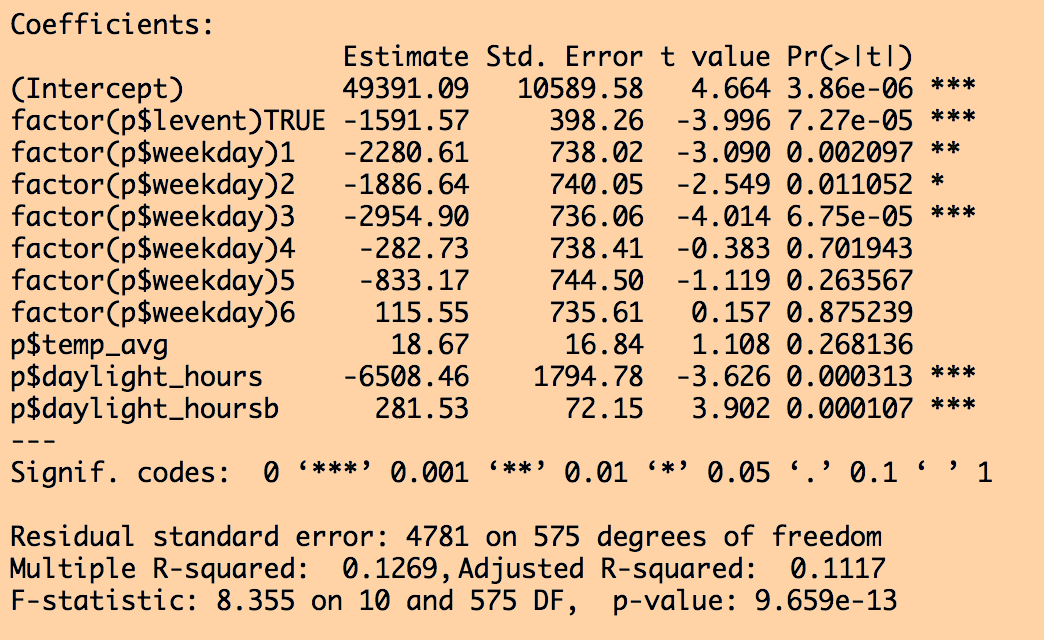
It looks like a linear model isn’t likely to be appropriate, and a quadratic would likely fit the data better. Add a quadratic term to your regression fit and repeat the fit, overlaying the new line on the plot. Also write the code to reproduce the following output:



Was the quadratic model a better fit?

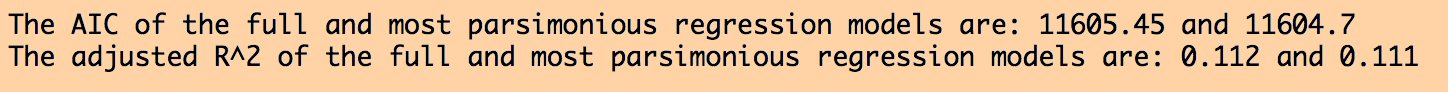


g) Now do a regression fit that includes temperature, daylight hours, daylight hours quadratic term, weekdays as a factor, and weather events as a factor. Reproduce the following output from the lm() method:



Which of the fit terms do not appear to be significant? (note, that when a multi-level factor is included, if one level is significant, the factor is considered significant in the fit)

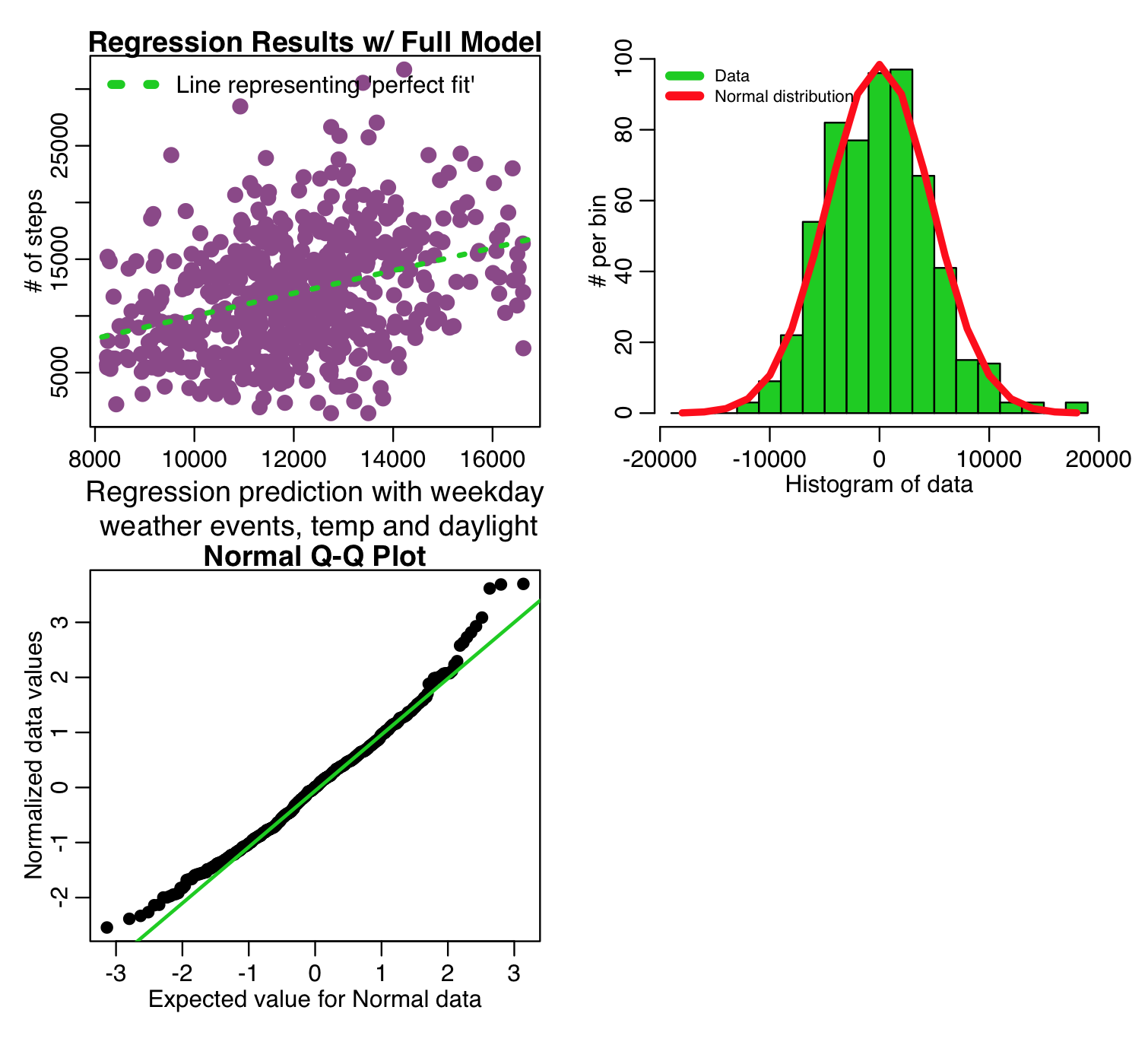
Using the R stepAIC() function, obtain the most parsimonious model, and reproduce the following output:



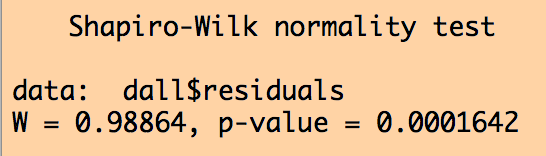
Which term went away in the full model? Did it make much of a difference either way?

h) For the most parsimonious model, reproduce the following set of plots (note that you can use the norm\_overlay\_and\_qq() function in the AML\_course\_libs.R script to produce the second and third plots). On the first plot, overlay the 45 degree line that represents a “perfect fit” (ie; a fit with perfect one-to-one correspondence between the fit results and the data).

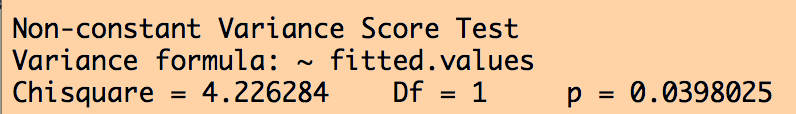
What pathology (fat tailed, narrow tailed, right skew, left skew, bimodal) does the QQ plot appear to represent?



i) Perform the Shapiro-Wilk test to test the Normality of the fit residuals, reproducing the following output:



j) In the R “car” library is the method ncvTest() which performs the Breusch-Pagan test for homoscedasticity of the residuals. Write the code to perform the test, and produce the following output:



k) Note that the QQ plot and the Shapiro-Wilk test both show some evidence of non-Normality of the residuals, and the Breusch-Pagan test indicates that there may be some evidence of heteroskedasticity.

This can reflect the presence of a confounding variable that we have not taken into account.

What assumption(s) did we make in this analysis that was almost certainly violated over the time period of the data (in fact, was violated several times)?

IMO, especially looking at the histogram of the residuals, they aren’t drastically non-Normal. If this were a publication, I’d likely proceed with the analysis because sometimes you just cannot account for all confounders.