HOMEWORK ANALYSIS #1 - WINDMILLS

Jillian Maw

1/21/2022

Energy can be produced from wind using windmills. Choosing a site for a wind farm (i.e. the location of the windmills), however, can be a multimillion dollar gamble. If wind is inadequate at the site, then the energy produced over the lifetime of the wind farm can be much less than the cost of building the operation. Hence, accurate prediction of wind speed at a candidate site can be an important component in the decision to build or not to build. Since energy produced varies as the square of the wind speed, even small errors in prediction can have serious consequences.

One possible solution to help predict wind speed at a candidate site is to use wind speed at a nearby reference site. A reference site is a nearby location where the wind speed is already being monitored and should, theoretically, be similar to the candidate site. Using information from the reference site will allow windmill companies to know the wind speed at the candidate site without going through a costly data collection period if the reference site is a good predictor. The dataset `windmill.txt` on the course webpage contains measurements of wind speed (in meters per second, m/s) at a candidate site (`CSpd`) and at an accompanying reference site (`RSpd`).

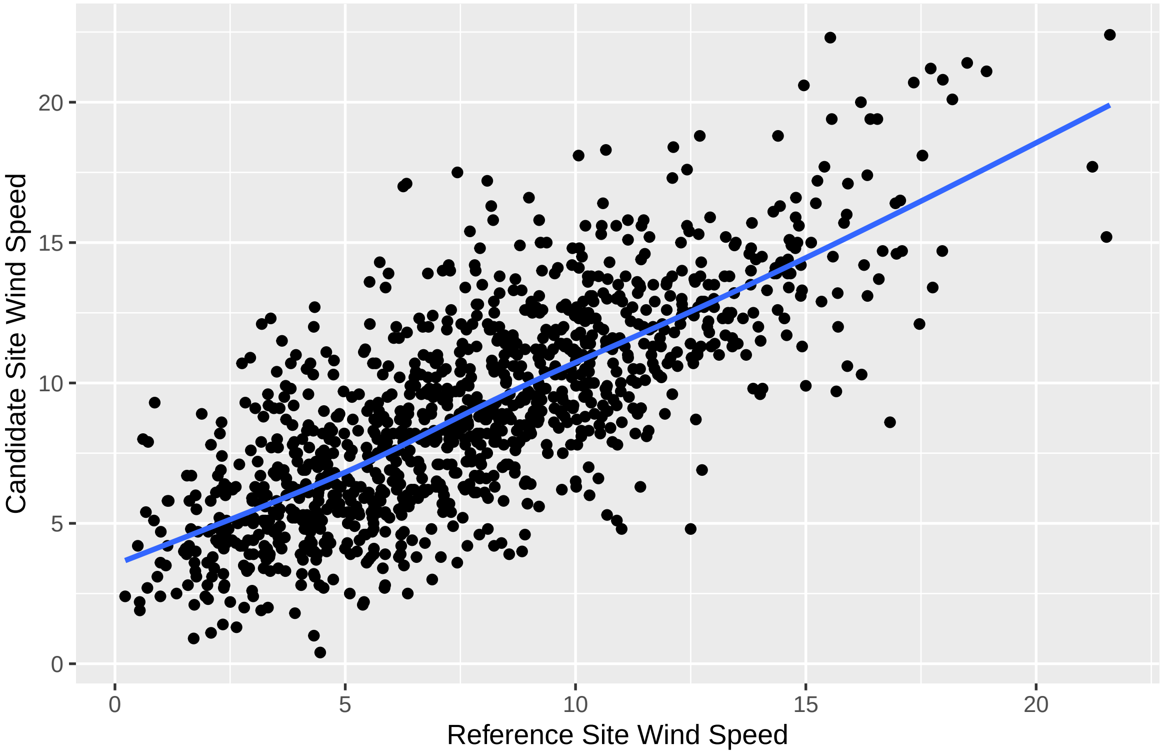
In each of the following questions, assume that your audience (the people you are writing your answer to) are civil engineers who have a strong mathematical background but a weak statistical background. Please attach your clearly commented code (R or Python) to the back of your answers as an appendix.

1. In your own words, summarize the overarching problem and specific questions that need to be answered using the windmill data. Discuss how statistical modeling will be able to answer the posed questions.

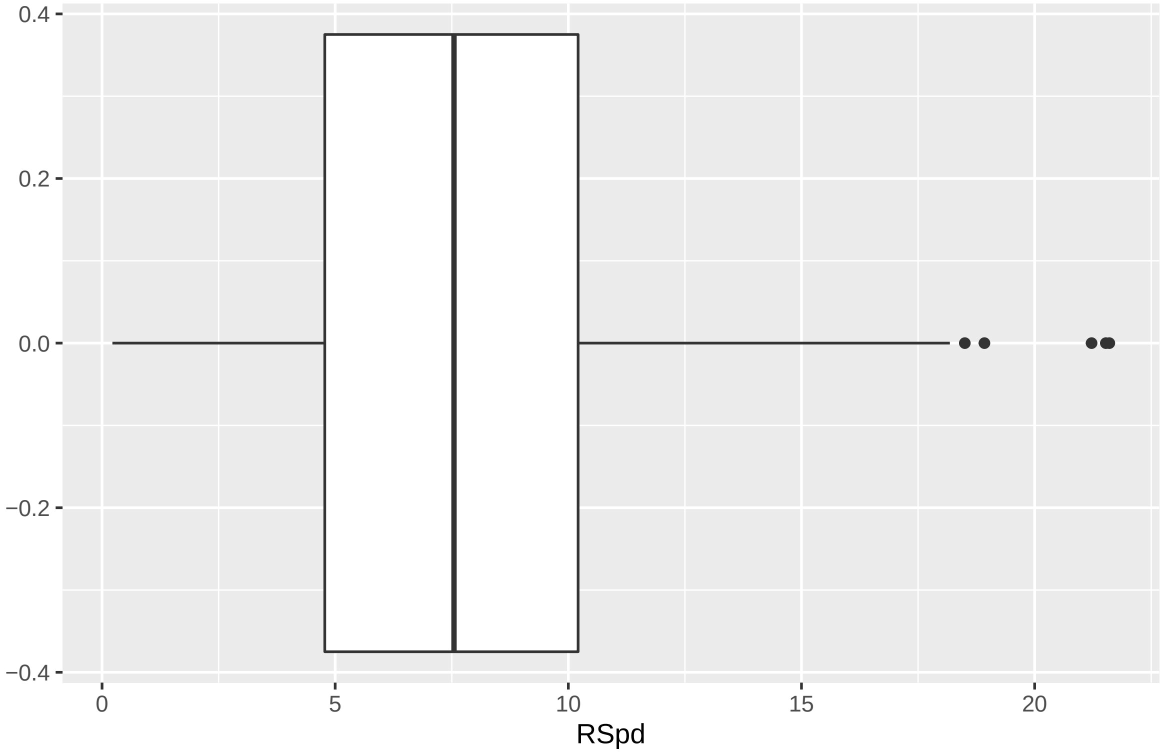
The overarching problem is to determine if the reference sites have similar wind speeds to the candidate sites, to prove whether or not they are a good predictor of wind speeds at the candidate sites. We will show that the reference site wind speeds can predict the candidate site wind speeds by quantifying the strength of the relationship between them. Since energy produced by a wind farm varies as the square of the wind speed, we will use least squares estimation to determine the quantitative relationship in wind speed, creating a formula that optimizes the distance between all points and the line of regression to be as small as possible, on average. This will tell us if the relationship between candidate and reference site wind speeds is strong and can be used to make predictions of candidate site wind speeds.

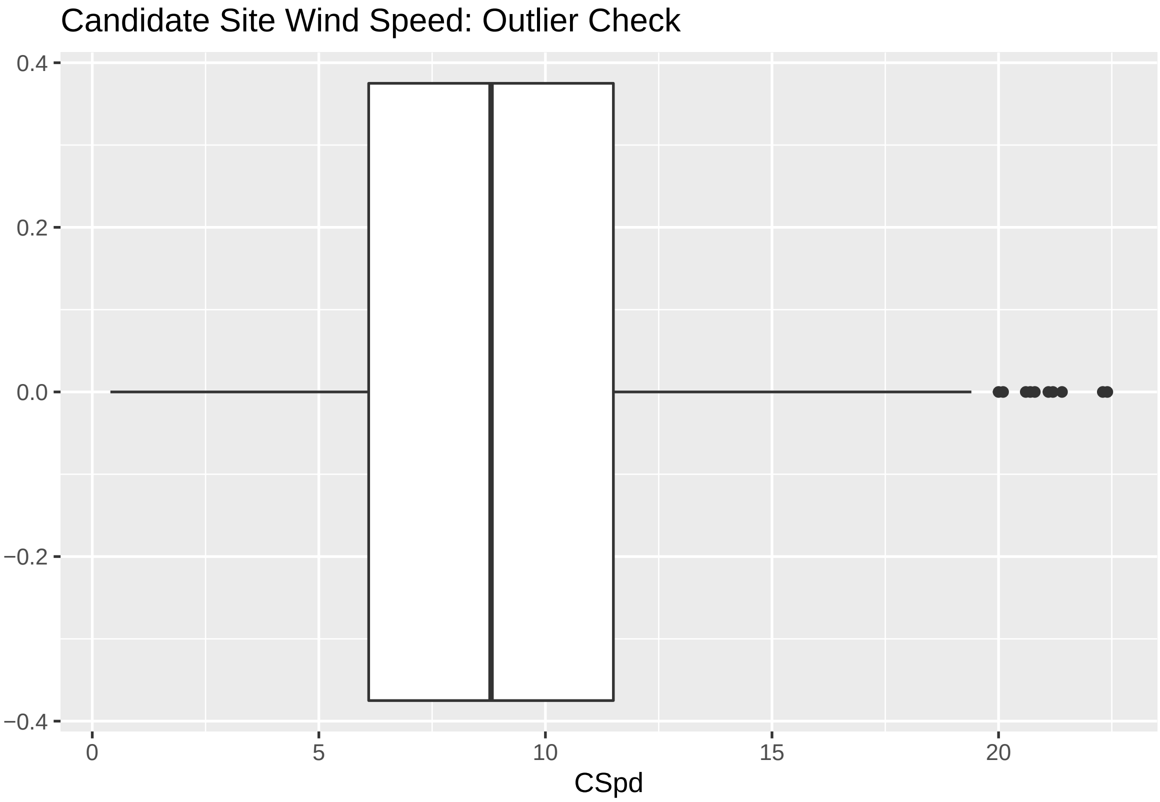
2. Explore the data using basic exploratory graphics and summary statistics. Comment on any potential relationships you see through this exploratory analysis.

Wind Speed Comparisons: Scatterplot and Trend Line



Reference Site Wind Speed: Outlier Check





Some of the summary statistics for this data is shown in the table below:

Wind Speed Location	Min. Wind Speed	Median Wind Speed	Mean Wind Speed	Max. Wind Speed
Reference Site	0.2221	7.5478	7.7773	21.6015
Candidate Site	0.4	8.8	9.0188	22.4

Also of note is the covariance between the reference site wind speed and candidate site wind speed, 10.6993, which tells us the relationship between the two is generally positive: on average, when the reference site wind speed increases, the candidate wind speed increases. For the final summary statistic, we note the correlation between the two as well, 0.7556, which quantifies the strength of the linear relationship as a positive, moderately strong relationship.

The scatterplot illustrates the same information of the summary statistics, especially the positive, moderately strong relationship between reference site wind speeds and candidate site wind speeds. It also shows the data is concentrated mostly between 0m/s and 15 m/s for both location types, with some data points greater than 15 m/s also showing on the scatterplot. A trend line add to the scatterplot demonstrates the linear direction and positive relationship between the two location types.

The boxplots show the range and spread of the data, and demonstrates there are a few outliers for both location types that could be affecting the strength of the relationship between the two location site types.

3. Regardless of your answer in #2, write out (in mathematical form with Greek letters) a SLR model that would help answer the questions in problem. Provide an interpretation of each mathematical term (variable or parameter) included in your model (e.g. interpret β_0). Using the mathematical form, discuss how your model, after fitting it to the data, will be able to answer the questions in this problem. List the four assumptions necessary to use SLR.

The model $y_i \overset{iid}{\sim} N(\beta_0 + \beta_1 x_i,\ \sigma^2)$ will serve as the simple linear regression model for our data. In this model, y_i represents the response variable at a given point i. The response variable of this report's data is candidate site wind speed, for example. The x_i represents the explanatory variable at a given point i, which we are using to explain the response variable (using statistical modeling). The symbol $\overset{iid}{\sim}$ means "independent and identically distributed", which means it meets two of the assumptions needed for simple linear regression. Those assumptions will be explained later, two paragraphs below. The N is short for Normal distribution, meaning the model's data follows that distribution shape and behaviors. β_0 represents the intercept coefficient, which says when x_i is 0, the mean y_i is the intercept coefficient. β_1 is the slope coefficient, which says as x_i increases, the mean y_i increases by the slope coefficient. The symbol σ^2 represents the variance of the data around the regression line fitted to the data by this model. Another way to think of the variance is that it is the square of the standard deviation. The standard deviation shows that for any x_i, 99.7% of the response variable will be within 3 standard deviations of the regression line made by $\beta_0 + \beta_1 x_i$, the intercept coefficient plus the product of the slope intercept and explanatory variable.

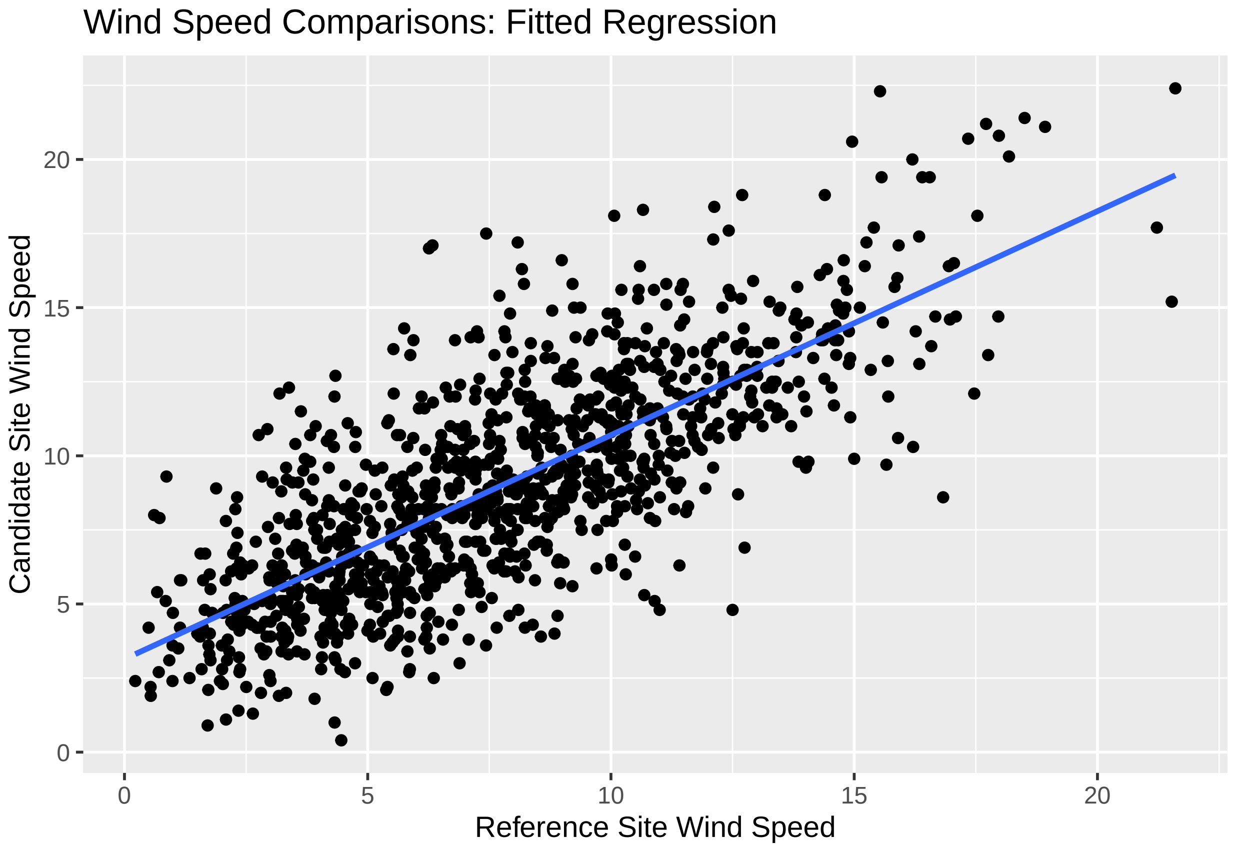
After fitting the above model to the windmill data, we will create a simple regression line that shows the most optimal linear relationship between all the reference site wind speeds and candidate site wind speeds. This will prove if the relationship between the two location types can be used to predict the wind speed of a candidate site for a windmill farm.

The four assumptions necessary to use simple linear regression are as follows:

- That the relationship between the data's variables is linear
- That the variables are independent of each other; that is, measuring one doesn't affect measuring the other variable
- That the population the data was sampled from follows a Normal distribution
- That the data is homoskedastic, or has an equal variance around the regression line fitted to it.

4. Fit your model in #3 to the windmill data and summarize the results by displaying the fitted model in equation form (do NOT just provide a screen shot of the R or Python summary output). Interpret each of the fitted parameters in the context of the problem. Provide a plot of the data with a fitted regression line.

Fitting our data to the aforementioned model, we get $y_i \overset{iid}{\sim} N(\ 3.1412 + 0.7557\ x_i,\ 6.0823\ )$. In this model, y_i represents the candidate site wind speed at location number i. The x_i represents the reference site wind speed. The symbol $\overset{iid}{\sim}$ means "independent and identically distributed", which means the windmill data meets two of the assumptions needed for simple linear regression. The N is short for Normal distribution, meaning the windmill data is distributed as Normal. The intercept coefficient has been replaced by 3.1412 m/s, which says when a reference site wind speed is 0 m/s, the mean candidate site wind speed is 3.1412 m/s. The slope coefficient has been replaced by 0.7557 m/s, which says as the reference site wind speed increases, the mean candidate site wind speed increases by 0.7557 m/s. The σ^2 has been updated to 6.0823, which represents the variance of the reference and candidate site wind speeds around the regression line fitted to the data by this model. Another way to think of that number is that for any reference site wind speed, 99.7% of the candidate site wind speed will be within 3 standard deviations (standard deviation being the square root of the variance σ^2) of the regression line made by $\beta_0 + \beta_1 x_i$, the intercept coefficient plus the product of the slope intercept and reference site wind speed.



5. Explain in simple terms how you can use your fitted model to obtain predictions of wind speed at the candidate site given a wind speed at the reference site. As an example to illustrate your point, use your fitted model to obtain a prediction of the wind speed at the candidate site given the wind speed at the reference site is 12 m/s.

We can use our fitted model to obtain predictions of wind speed at the candidate site, given the wind speed at the reference site, by substituting the reference site wind speed for x_i in the model. For example, if we want to predict the wind speed of a candidate site wind speed where the reference site wind speed is 12 m/s, we substitute 12 for x_i into the model's regression line, $3.1412 + 0.7557\ x_i$. We will get a mean estimate for the candidate site wind speed of 12.21 m/s.

6. Explain potential limitations of using your SLR model for prediction. As an example, use your fitted model to predict the wind speed at the candidate site given the wind speed at the reference site is 30 m/s.

One limitation of the model is that it cannot extrapolate a candidate site wind speed beyond the range of the collected data, as there would be so much variance that would be unknown and unaccounted for it could be utterly and stupidly wrong. For example, if we tried to predict the wind speed of a candidate site wind speed where the reference site wind speed is 30 m/s, far above the known maximum of collected data for reference site wind speeds, 21.6015 m/s, the calculated result is 25.8132 m/s, a difference of 4.2117 m/s and above the standard deviation of the rest of the model. The calculated result is also far above the maximum collected candidate site wind speed of 22.4 m/s. We have no way of knowing if this is even a valid measurement in for these sites in the first place.

Appendix of Code

```
library(knitr)
knitr::opts_chunk$set(echo = FALSE, include = FALSE)
library(tinytex)
library(ggplot2)
windmill <- read.table(file="~/R programming/STAT_330/Windmill.txt", header = TRUE)
#x=RSpd is explanatory variable; y=CSpd is response variable
windmill.scatter <- ggplot(data = windmill, mapping=aes(x=RSpd, y=CSpd)) +
  geom_point() +
  geom_smooth(se=FALSE) +
  xlab('Reference Site Wind Speed') +
  ylab('Candidate Site Wind Speed') +
  ggtitle('Wind Speed Comparisons: Scatterplot and Trend Line')
suppressMessages(print(windmill.scatter))
#Calculate the correlation and covariance between CSpd and RSpd
windmill.cov <- cov(windmill$CSpd, windmill$RSpd)
windmill.cor <- cor(windmill$CSpd, windmill$RSpd)
#Provide a summary of the main features of the data
windmill.RSpd.box <- ggplot(data = windmill, mapping=aes(RSpd)) +
  geom_boxplot() +
  ggtitle("Reference Site Wind Speed: Outlier Check")
windmill.CSpd.box <- ggplot(data = windmill, mapping=aes(CSpd)) +
  geom_boxplot() +
  ggtitle("Candidate Site Wind Speed: Outlier Check")
suppressMessages(print(windmill.RSpd.box))
suppressMessages(print(windmill.CSpd.box))
#clean up the summary statistics
R1 <- as.numeric(round(summary(windmill$RSpd)["Min."], digits = 4))
R2 <- as.numeric(round(summary(windmill$RSpd)["Median"], digits = 4))
R3 <- as.numeric(round(summary(windmill$RSpd)["Mean"], digits = 4))
R4 <- as.numeric(round(summary(windmill$RSpd)["Max."], digits = 4))
C1 <- as.numeric(round(summary(windmill$CSpd)["Min."], digits = 4))
C2 <- as.numeric(round(summary(windmill$CSpd)["Median"], digits = 4))
C3 <- as.numeric(round(summary(windmill$CSpd)["Mean"], digits = 4))
C4 <- as.numeric(round(summary(windmill$CSpd)["Max."], digits = 4))
#Fit a simple linear model to the windmill data where CSpd is the response variable and
#RSpd is the explanatory variable.
windmill.regress <- lm(formula=CSpd~RSpd, data=windmill)
#Identify the estimates beta sub zero, beta sub one, and sigma squared.
windmill.beta.0 <- round(as.numeric(coef(windmill.regress)["(Intercept)"]), digits = 4)
windmill.beta.1 <- round(as.numeric(coef(windmill.regress)["RSpd"]), digits = 4)
windmill.var <- round(sigma(windmill.regress)^2, digits = 4)
#Add your estimated regression line to the scatterplot you created above.
windmill.est.reg <- ggplot(windmill, aes(x=RSpd,y=CSpd)) +
  geom_point() +
  geom_smooth(method="lm",se=FALSE) +
  xlab('Reference Site Wind Speed') +
  ylab('Candidate Site Wind Speed') +
  ggtitle('Wind Speed Comparisons: Fitted Regression')
suppressMessages(print(windmill.est.reg))
#Generate predictions of CSpd for a RSpd of 12 m/s.
windmill.predict <- data.frame(RSpd = 12)
```

```r
windmill.est.1 <- round(as.numeric(predict.lm(windmill.regress,
                                              newdata=windmill.predict)), digits = 4)
#Generate predictions of CSpd for a RSpd of 30 m/s.
windmill.predict <- data.frame(RSpd = 30)
windmill.est.2 <- round(as.numeric(predict.lm(windmill.regress,
                                              newdata=windmill.predict)), digits = 4)
#End of homework's code
```