Homework 1: Windmills Dataset

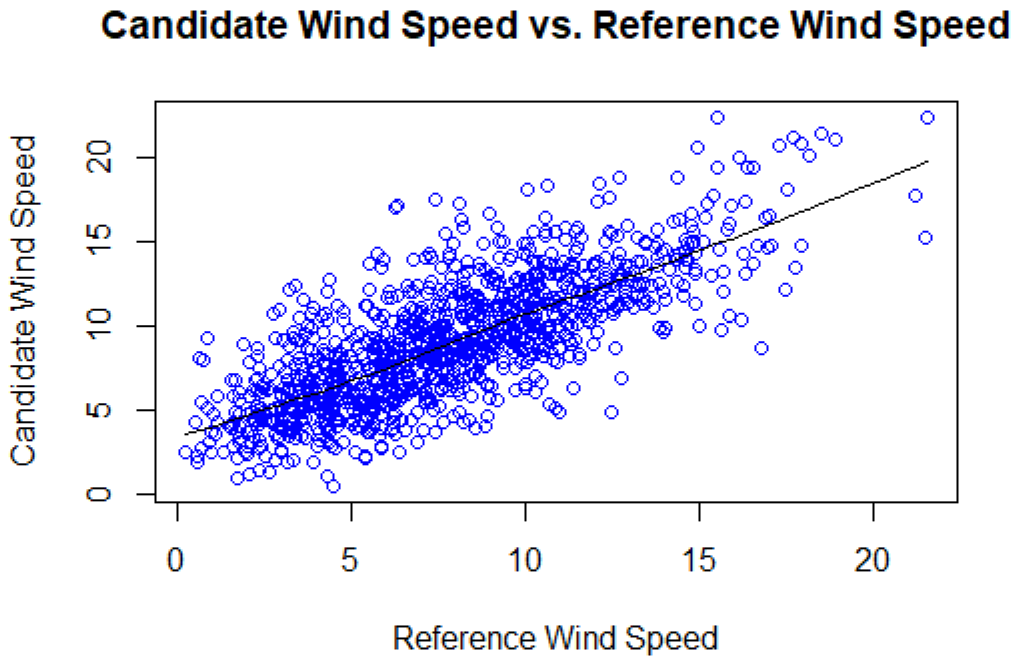
Kevin Toney

September 20, 2017

Question 1. Our main purpose is to predict wind speed of a candidate site for a wind farm, and thus determine if the farm should be built. Since the wind speed makes a large impact in the amount of wind energy produced, we will be careful to minimize errors as much as possible. In order to predict with the least amount of error and financial burden, we will use the wind speeds collected from a reference site that is similar, and close in distance, to the candidate site. We will also determine whether or not this reference site is a good predictor.

A good statistical model to answer our questions is simple linear regression. The purpose of linear regression is to determine the relationship between two variables, and use the relationships to predict a value of one of them. With this model, we will be able to study and predict what the wind speeds will be at the candidate site once the wind farm is built.

Question 2. Is simple linear regression appropriate for this situation? Let's start by seeing a scatter plot of the candidate site's wind speeds compared to the reference sites' wind speed.



The trend of the data suggests that the candidate site's wind speeds increase as the reference's wind speeds grow. The data seems to follow a linear pattern upward. To confirm our observations, we will use a metric called correlation. Correlation measures the strength and direction of a linear relationship. The correlation of the data, since it seems to follow a linear pattern, is 0.75. Since this metric is close to 1, we confirm that the candidate site's wind speed follows the same pattern as the reference site's wind speed.

Please be aware that there does not seem to be any outliers in the data.

Since the two sites' wind speeds have a linear relationship, we are justified in using simple linear regression.

Question 3. The simple linear regression model we will use to predict the candidate site's wind speed is:
$Candidate_i \sim NormalDistribution(\beta_0 + \beta_1 * Reference_i, \sigma^2)$

You may notice that the first parameter in the normal distribution is a linear equation.

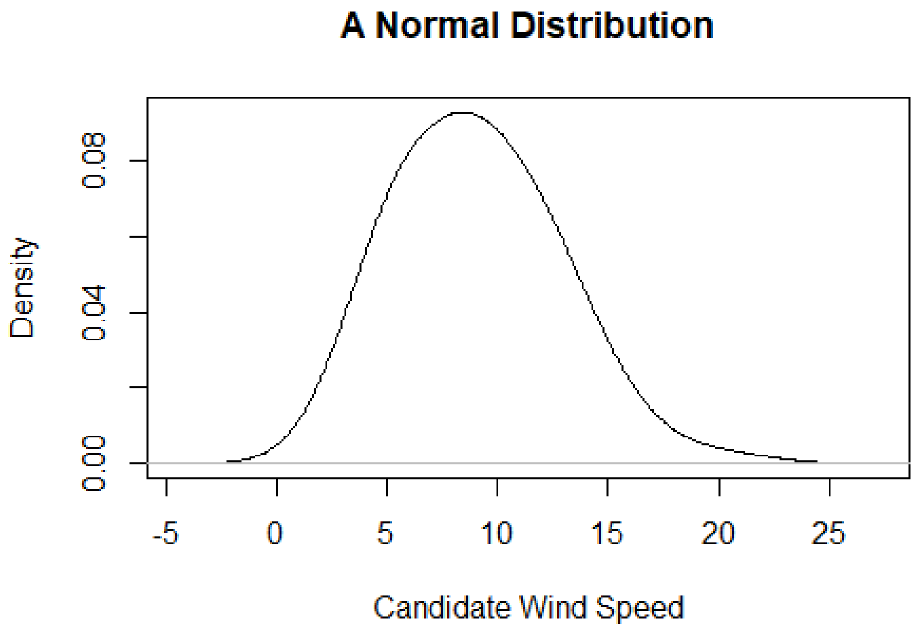
i = a certain wind speed from 1 to the number of recorded wind speeds (1,116).

β_0 = The candidate site's wind speed, on average, if the reference site's wind speed will equal 0 meters per second.

β_1 = The rate at which the candidate site's wind speed increases if the reference site's wind speed rises by 1 m/s.

σ^2 = The average distance of the candidate site's wind speeds from the mean line.

For your information, a normal distribution follows a bell curve with most of the data points being located close to the mean. This model will help us to predict wind speeds that are close to the mean of the normal distribution, i.e, the values that are most likely, given certain wind speeds from the reference site. An example of a normal distribution is shown below.



We are assuming each recorded wind speed is independent of each other. Moreover, we are assuming the prediction error follows a normal distribution. Also, we are assuming σ^2 of the wind speeds is equal all across the line created by the linear equation $\beta_0+\beta_1*Reference_i,$.

Question 4. The model, or the linear equation that provides the predicted mean line of the data with the least amount of error is below:
$Candidate_i \sim NormalDistribution(\hat{\beta}_0 + \hat{\beta}_1 * Reference_i, \hat{\sigma}^2)$

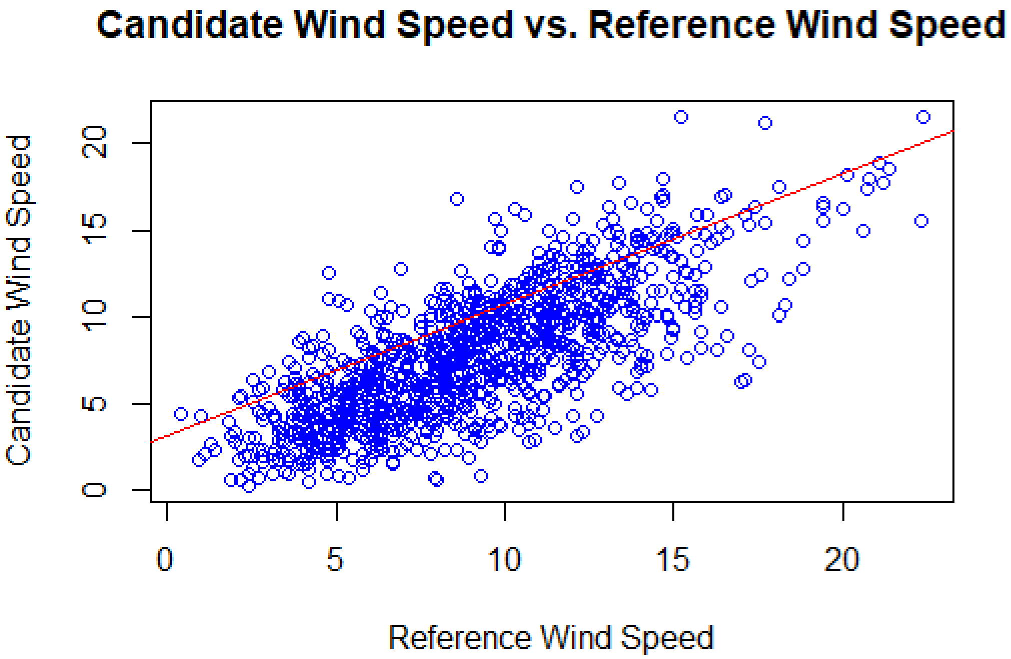
$i = 1$ to the number of recorded wind speeds (1,116).

$\hat{\beta}_0$ = The predicted candidate site's wind speed, given the reference site's wind speed equals 0 m/s.

β_1 = The predicted rate the candidate site's wind speed will increase if the reference site's wind speed rises by 1 m/s.

σ^2 = The predicted average distance of the candidate site's wind speeds from the mean line.

The data with the predicted mean line, which has the least amount of error, is below:



The red line shows the mean line of the candidate site's wind speed compared to the reference site's wind speed. This line has the least amount of prediction error. Therefore, we are confident the predicted wind speeds at the reference site will be close to this line.

Question 5. Now, we will be able to use the linear model $Candidate_i \sim NormalDistribution(\hat{\beta}_0 + \hat{\beta}_1 * Reference_i, \hat{\sigma}^2)$ to predict the wind speeds at the candidate site given a wind speed at the reference site. For example, when I input a reference site wind speed of 12 m/s, the model returns a predicted candidate site wind speed of 12.21 m/s.

Question 6. Nevertheless, there is a crucial limitation to predicting the candidate site's wind speed. We are only able to predict values that are in the range of the data. For example, the max of the recorded wind speeds 22.4 and 21.602. I will not be able to predict the candidate site's wind speed given the reference site's wind speed is 30 m/s. If we want to predict values that are more extreme than 22.4, we will need to collect more data.

A R Code:

```
rm(list=ls())

windmills.dat <− read.table(”Fall 2017/STAT 330/Homework/Homework1/
Windmill.txt”, header = T, sep = ””)

y <− windmills.dat$CSpd #this response variable is in
the 1st column of the dataset
x <− windmills.dat$RSpd
#did this data read in correctly? I will verify
my data with other classmates.

#according to the plot, I think my data read in correctly.
plot(y, x, xlab = ”Reference Wind Speed”,
ylab = ”Candidate Wind Speed”, col = ”blue”,
     main = ”Candidate Wind Speed vs. Reference Wind Speed”)
#replace my first scatter plot with a scatter plot that has a
#line running through the middle of the data points.
scatter.smooth(x, y, xlab = ”Reference Wind Speed”,
ylab = ”Candidate Wind Speed”, col = ”blue”,
               main = ”Candidate Wind Speed vs. Reference Wind Speed”)

cov(x, y) #the positive result tells me there is a
#positive linear relationship. As reference wind speed
#goes up, the candidate wind speed follows the same pattern.
cor(x, y) #the linear relationship is strong, since
#the correlation is above 0.5

plot(density(y), main=”An Example of a Histogram”,
xlab=”Candidate Wind Speed”)
plot(density(y, adjust = 2), main=”A Normal Distribution”,
xlab=”Candidate Wind Speed”)
#the response variable seems to follow a normal distribution.
#I used adjust=2 to make the graph look smoother.

linear_model <− lm(y ˜ x)
```

```
summary(linear_model)
coef(linear_model)
#the estimate for the intercept, the value of the
candidate site's wind
#speed, on average, when the reference windspeed is 0 comes out to be 3.141.
#the intercept for the average rate of change, or b1 is 0.756.
#the estimate for the average variability (sigma^2) of the data points
#around the mean line is 2.466.

#the r-squared score the summary gave me is 0.5709. I don't think the fit
may be as good as I was hoping.

#plot the least squares regression line on the data.
plot(y, x, xlab = ”Reference Wind Speed”,
ylab = ”Candidate Wind Speed”, col = ”blue”,
     main = ”Candidate Wind Speed vs. Reference Wind Speed”)
abline(reg = linear_model, col=”red”)

#I don't think I read my data in correctly. Let's try copying and
pasting the data in a text file.
#nope. I read in the data correctly.
Why doesn't the line fit as well as I want it to?

new_dframe <- data.frame(x=12)
predict.lm(linear_model, newdata = new_dframe)
#given a reference site's wind speed at 12 m/s,
#the model predicts the candidate site's wind speed to be
#12.21 m/s.

#test our assumptions real quick.
library(lmtest)
bptest(linear_model)
library(MASS)
std.res <- stdres(linear_model)
plot(std.res)
plot(density(std.res))
#the residuals seem to follow a
normal distribution too.
#simple linear regression seems to be the best model.

max(windmills.dat$CSpd)
max(windmills.dat$RSpd)
```