Homework 1 | STAT 330

Simple Linear Regression

AUTHOR
Zeb Sorenson

Data and Description

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 multi-million 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 estimate the wind speed at the candidate site without going through a costly data collection period, if the reference site is a good predictor.

The Windmill data set contains measurements of wind speed (in meters per second m/s) at a **candidate site (CSpd) (column 1)** and at an accompanying **reference site (RSpd) (column 2)** for 1,116 areas. Download the Windmill.txt file from Canvas, and put it in the same folder as this quarto file.

0. Replace the text "< PUT YOUR NAME HERE >" (above next to "author:") with your full name.

1. Briefly explain why simple linear regression could be a useful tool for this problem.

Linear regression could offer valuable insights into whether or not this candidate site will produce the wind needed to generate enough energy to offset costs and offer a viable energy source.

If we are able to correctly fit our model using the given data, the model can give us a concrete prediction of wind tendencies in the area which we can use to decide whether or not the location serves for our purposes.

2. Read in the data set, and call the tibble "wind". Print a summary of the data and make sure the data makes sense.

```
# <your code here>

wind <- read.csv("~/Desktop/Stat 330/Windmill.txt", sep="")

View(wind)

summary(wind)
```

```
      CSpd             RSpd        
 Min.   : 0.400   Min.   : 0.2221  
 1st Qu.: 6.100   1st Qu.: 4.7769  
 Median : 8.800   Median : 7.5477  
 Mean   : 9.019   Mean   : 7.7773  
 3rd Qu.:11.500   3rd Qu.:10.2096  
 Max.   :22.400   Max.   :21.6015  
```

3. What is the outcome variable in this situation? (Think about which variable makes the most sense to be the response.)

Wind Speed of our candidate site is our outcome variable. We are looking at wind speed data based on our reference location. So CSpd will the our Y axis on our graph.

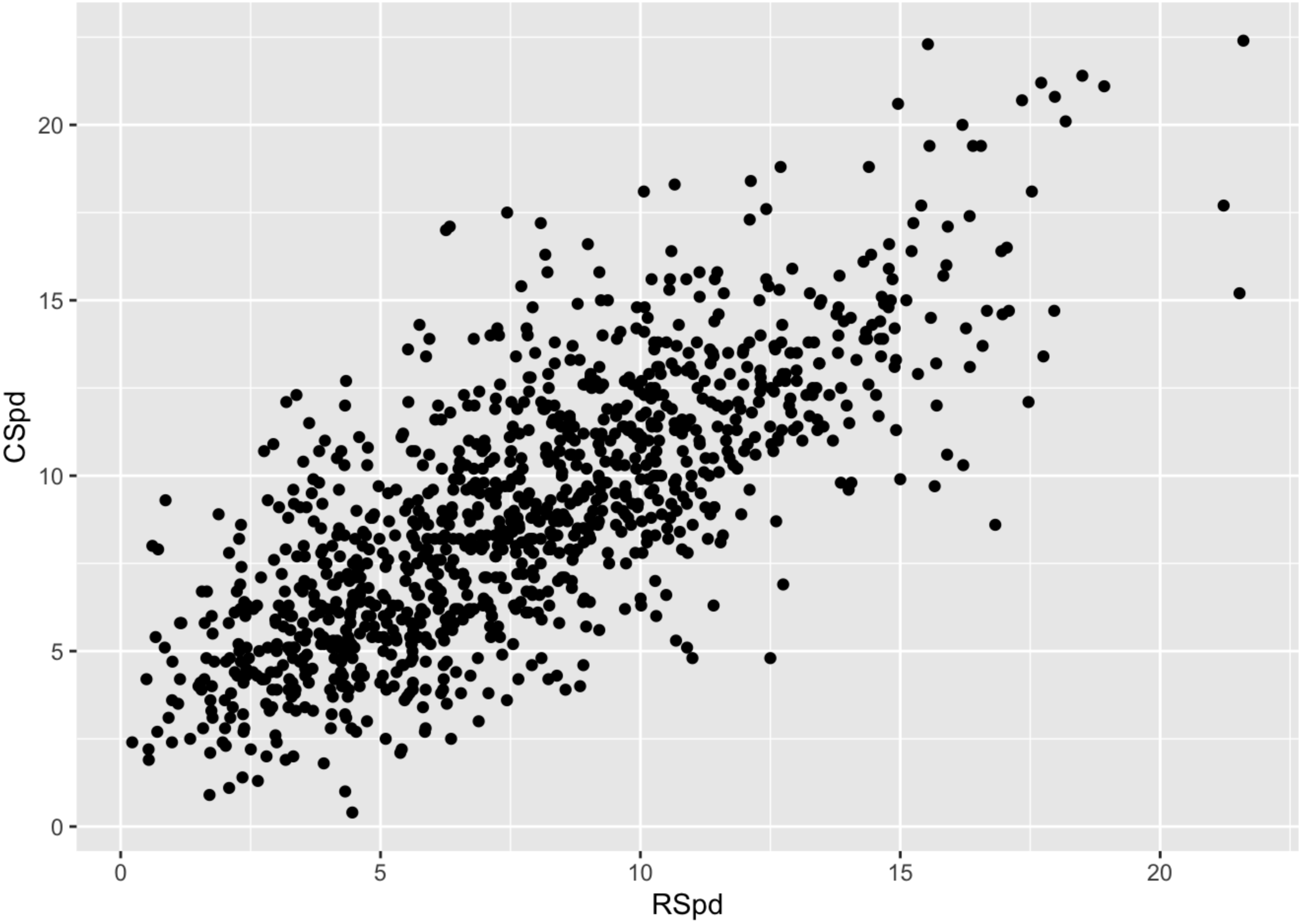
4. What is the explanatory variable in this situation?

The wind speed of our reference site is our explanatory variable. Meaning it will be our X axis when we create our graph.

5. Create a scatterplot of the data with variables on the appropriate axes. Add descriptive axis labels with appropriate units. Save the plot to a variable and print the plot.

```r
wind_scatter_plot <- ggplot(data = wind,
                            mapping = aes(x = RSpd, y = CSpd)) +
  geom_point()

wind_scatter_plot
```



```
#Plot is created by modifying in class coding activity 1
```

6. Briefly describe the relationship between RSpd and CSpd. (Hint: you should use 3 key words in a complete sentence that includes referencing the variables.)

The relationship between our reference site and candidate site is a moderate to strong positive linear relationship. Meaning, where our reference site is low, our candidate site is also low and the same for when the reference site is high. Going in a positive linear direction.

7. Calculate the correlation coefficient for the two variables (you may use a built-in R function). Print the result.

```
CS_PD_Cor <- cor(wind$CSpd, wind$RSpd) #save to a variable to possible use later and print

print(CS_PD_Cor)
```

```
[1] 0.7555948
```

8. Briefly interpret the number you calculated for the correlation coefficient (what is the direction and strength of the correlation?).

Our correlation coefficient is both positive and close to 1. Meaning that the strength relationship between these two variables is strong and positive. Zero would indicate no relationship and a negative number would indicate a negative relation.

9. Mathematically write out the simple linear regression model for this data set (using parameters (βs), not estimates, and not using matrix notation). Clearly explain which part of the model is deterministic and which part is random. Do not use “x” and “y” in your model - use variable names that are fairly descriptive.

candidate_wind_speed = β_0 + $\beta_i \times$ref_site_speed +ϵ_i

β_0 = our intercept. Expected wind speed of our candidate site when the wind speed at the reference site is 0.

$\beta_i \times$ref_site_speed = Slope. The change in our candidate site speed for every unit of change in speed at the ref site.

ϵ_i = Our residuals. The difference between our predicted value and actual value.

Where

ϵ_i is $\epsilon \sim \mathcal{N}(0, \sigma^2)$ independent and identically distributed

Both our intercept and our slope are deterministic because we can determine “hard” actual values for these variables, while our residuals will be random since they cannot be predicted.

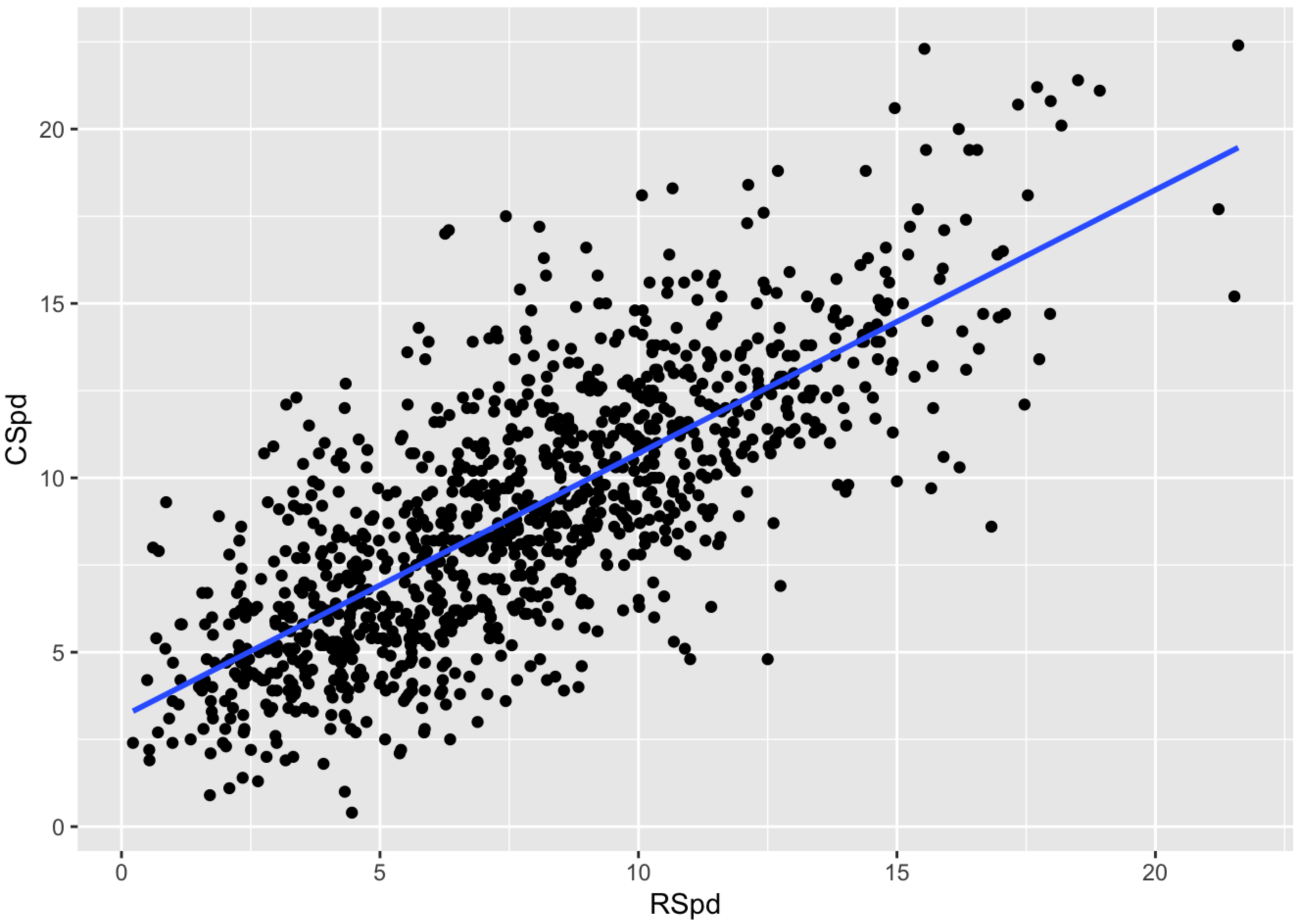
10. Add the OLS regression line to the scatterplot you created in 4. Print the result. You can remove the standard error line with the option `se = FALSE`.

```
wind_scatter_plot +

  geom_smooth(mapping = aes(x = RSpd, y = CSpd),

               method = "lm",
              se = FALSE)
```

```
`geom_smooth()` using formula = 'y ~ x'
```



```
#modified from code from in class #1
```

11. (a) Apply linear regression to the data. (b) Print out a summary of the results from the `lm` function. (c) Save the residuals and fitted values to the `wind` tibble. (d) Print the first few rows of the `wind` tibble.

```
#A

wind_speed_LM <- lm(CSpd ~ RSpd, data = wind)

#Candidate Speed Explained by Reference Site Speed

#B

summary(wind_speed_LM )
```

```
Call:
lm(formula = CSpd ~ RSpd, data = wind)

Residuals:
    Min      1Q  Median      3Q     Max 
-7.7877 -1.5864 -0.1994  1.4403  9.1738 
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3.14123    0.16958   18.52   <2e-16 ***
RSpd         0.75573    0.01963   38.50   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.466 on 1114 degrees of freedom
Multiple R-squared:  0.5709,	Adjusted R-squared:  0.5705 
F-statistic:  1482 on 1 and 1114 DF,  p-value: < 2.2e-16
```

```
#C

wind$residuals <- wind_speed_LM$residuals
wind$fits <- wind_speed_LM$fitted.values

#D

head(wind)
```

```
  CSpd   RSpd  residuals     fits
1  6.9 5.9666 -0.7503908 7.650391
2  7.1 7.2176 -1.4958132 8.595813
3  7.8 7.9405 -1.3421328 9.142133
4  6.9 6.0174 -0.7887821 7.688782
5  5.5 6.1646 -2.3000260 7.800026
6  3.1 1.7687 -1.3778979 4.477898
```

12. Briefly explain the rationale behind minimizing squared error loss. How does OLS choose the parameter estimates?

Squared error loss allows us to minimize the difference between our observed and predicted values giving us an unbiased estimate. It is also much easier to compute vs absolute value error. Absolute value error will give us an over fitted line, squaring gives us a better fit, minimizing squared residuals. OLS chooses the parameter estimates by giving us our estimated Y value GIVEN x. We use our x value which has already been given to us and allows us to create the signal as a straight line. Giving us our Beta Zero and Beta*X. It also assumes the variance is constant.

13. Mathematically write out the fitted simple linear regression model for this data set using the coefficients you found above (do not use parameters/βs and do not use matrix notation). Do not use “x” and “y” in your model - use variable names that are fairly descriptive.

Estimated_Y_GivenX = 3.14123 + 0.75573$\times$ref_site_speed

14. Interpret the coefficient for the slope.

This means that for every change in 1 unit of measurement in our reference site wind speed, we will have a change of 0.75573 unit of change in our candidate site.

15. Interpret the coefficient for the intercept.

This means that when our references site's wind speed is 0, our model is predicting that our candidate site's wind speed will be 3.14123

16. What is the estimated average wind speed at the candidate site (CSpd) when the wind speed at the reference site (RSpd) is 12 m/s? Show your code, and print the result.

```
Estimated_Candidate_Speed <-function(speed){
  speed<-3.14123+(0.75573*speed)
}

candidate <- Estimated_Candidate_Speed(12)

print(candidate)
```

```
[1] 12.20999
```

17. Briefly explain why it would be risky to answer this question: What is the estimated average wind speed at the candidate site (CSpd) when the wind speed at the reference site (RSpd) is 25 m/s?

We can briefly look at our scatter plot and see there is no data for our reference site at 25 m/s. Yes, we have been able to create a linear model that will give us a predicted number for our candidate site, however, looking specifically at our situation and the data, it may be unreliable to predict a number that high.

18. Calculate the (unbiased) estimate of σ^2, the average squared variability of the residuals around the line. Show your code, and print the result.

```
MSE <- sum(wind_speed_LM$residuals^2)/(wind_speed_LM$df.residual)

print(MSE)
```

```
[1] 6.082312
```

19. Create the design matrix and store it in a variable. Print the first few rows of the design matrix.

```
#we will multiply by reference site speed

#Use the cbind function, not the matrix function

wind_design_matrix  <- cbind(rep(1, length(wind$RSpd)), wind$RSpd)

head(wind_design_matrix)
```

```
     [,1]   [,2]
[1,]    1 5.9666
[2,]    1 7.2176
[3,]    1 7.9405
```

```
[4,]    1 6.0174
[5,]    1 6.1646
[6,]    1 1.7687
```

20. Obtain, and print, the parameter estimates for this data set (found above using `lm`) using matrix multiplication. You should use the following in your computations: t() [tranpose], solve() [inverse], %*% [matrix multiplicaiton].

```
#Equation on page 10 of lecture notes mod 1 lecture 4

#Y = CSpd values

#Do the computation from equation on page 10.

Y <- wind$CSpd

#Take the transpose first because giving error trying to do on one line of code

transposed_design_matrix <- t(wind_design_matrix)

Beta_Hat <- solve(transposed_design_matrix %*% wind_design_matrix) %*% (transposed_design_mat

print(Beta_Hat)
```

```
          [,1]
[1,] 3.1412324
[2,] 0.7557333
```

21. Briefly summarize what you learned, personally, from this analysis about the statistics, model fitting process, etc.

This assignment was intimidating at first but I learned a lot! It was very educational to be given data and then need to apply regression to that data and interpret our outputs rather than just reading about it and answering a few questions. I now feel like if I were given data in the workplace, I have a foundation to build on to do regression.This assignment helped me to learn to first explore the data before jumping into model fitting and then making sure our models make sense before proceeding. I also think it was important to see that although we were about to fit our model, that doesn’t mean we can predict any value, given the context of our situation. And although I don’t love R, it was great to improve my R skills and see how it can help us when working with a large amount of data.

22. Briefly summarize what you learned from this analysis *to a non-statistician*. Write a few sentences about (1) the purpose of this data set and analysis and (2) what you learned about this data set from your analysis. Write your response as if you were addressing a business manager (avoid using statistics jargon) and just provide the main take-aways.

The purpose of these data sets is to see if we can use the data we have from our reference site to be able to predict what wind speeds will be at out candidate site. If they are similar enough, we can then make predictions for wind speeds to help us determine if choosing the candidate site is a good choice. After performing our analysis, we saw that the reference site is indeed a good predictor tool for our candidate site in such a way that we can make reasonable predictions on how effective our candidate site will be within a reasonable range of numbers. Based on this anaylsis, the candidate site is a viable option!