Colin White
Jan 27, 2021
STAT 330 | INTRODUCTION TO REGRESSION
HOMEWORK ANALYSIS #1 – WINDMILLS

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.

We need to make sure a potential site for windmills will be profitable. To do this we need to know accurate information on the wind speed in candidate sites for windmills. Collecting this data would cost lots of time and money. However, we may be able to use already existing data from reference sites that are close to the candidate sites. If we can prove that the reference site's wind speed data is a good predictor of candidate sites' wind speed then we can make good business choices on where to invest in building windmills.

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

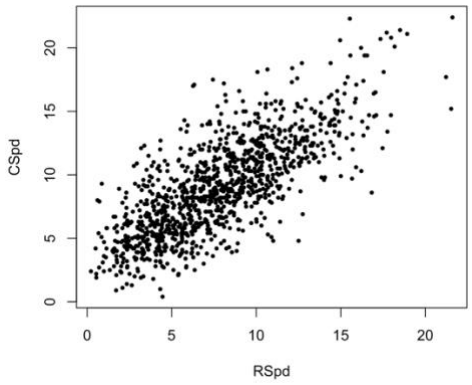


Figure 1

From the scatter plot in figure 1, it appears there is a medium to strong positive correlation between reference site wind speed (RSpd) and candidate site wind speed (RSpd.) The data seems to suggest that as RSpd increases, CSpd increases by nearly the same amount.

The correlation and covariance of RSpd to CSpd is 0.7556 and 10.6993 respectively. The Sigma is 2.4658.

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 $\_0$). 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.

$Y = ß_0 + ß_1 * x_1$
Y = The predicted value of CSpd
$ß_0$ = The y intercept, or the predicted value of CSpd if RSpd was zero.
$ß_1$ = The amount predicted wind speed (CSpd) increases, for every unit increase in reference wind speed (RSpd).

x_1 = The independent variable. Or observations of RSpd.

The SLR line will help us predict if the candidate site will have favorable wind conditions given the reference site wind conditions.

The 5 assumptions necessary to use SLR are:
1. Errors are Normally distributed.
2.Errors are independent of one another.
3.Errors are independent of covariates (x).
4.Covariates (x) are measured without error.
5.Relationship between covariate (x) and response (y) is linear.

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.

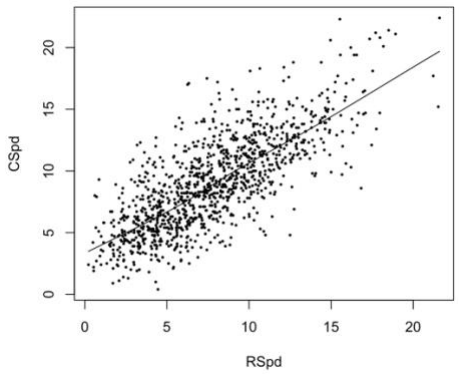


Fig 2

$Y = ß_0 + ß_1 * x_1$
CSpd = (3.1412) + (0.7557) * (RSpd)

This SLR equation implies that if RSpd is zero, CSpd would be predicted to be 3.1412. For each one unit RSpd increases, it is predicted that CSpd will increase by 0.7557.

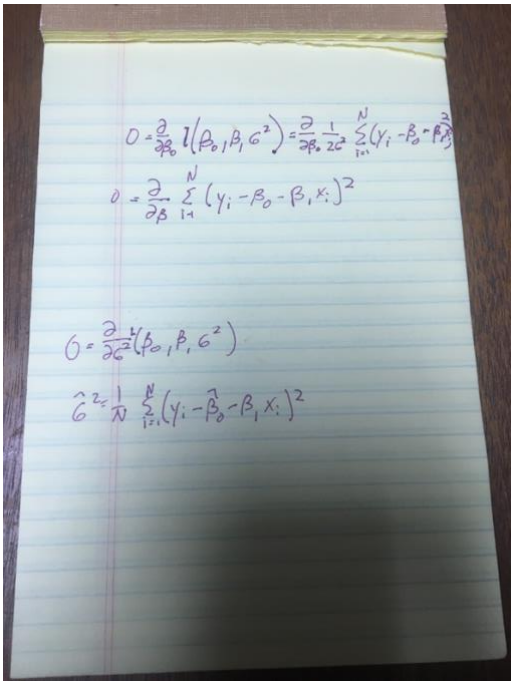
5. Explain in simple terms how you can use your fitted model to obtain predictions of windspeed at the candidate site given a windspeed at the reference site. As an example to illustrate your point, use your fitted model to obtain a prediction of the windspeed at the candidate site given the windspeed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

The model helps us predict the wind speed at a candidate site by taking the windspeed of the associated reference site and putting its value through the SLR equation, giving us the predicted candidate site wind speed. For instance, if a reference site has a wind speed of 12 m/s, we would predict that the wind speed at the associated candidate site would be 12.21 m/s (3.14123 + 0.75573 * 12 m/s.)

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

The SLR model has limits. It does not have enough data to reliably make predictions for RSpds above somewhere between 15 and 20 m/s. Trying to use the model to predict CSpds for RSpds greater than 15 or 20 m/s could yield an inaccurate prediction.

7. *This problem doesn't use the Windmill dataset.* For a dataset y_i for $i = 1, ..., n$, we assume an intercept-only model, $y_i \overset{iid}{\sim} \mathcal{N}(\beta_0, \sigma^2)$. Derive the maximum likelihood estimates of β_0 and σ^2.



CODE

```
#Code derived from class handouts
setwd("~/Desktop/1A School/1A Winter 2021/STAT330/HW1")

## Read in the Data
wind <- read.table("data.txt",header=TRUE)

summary(wind) ### gives general summary
head(wind)

## Plot the data, calculate correlation and covariance
plot(wind$RSpd,wind$CSpd,xlab="RSpd",pch=19,ylab="CSpd", cex = .3) # plot(x values, y values) and ($ means grab dat field )
scatter.smooth(wind$RSpd,wind$CSpd,xlab="RSpd",ylab="CSpd", pch=19, cex = .3)

cor(wind$RSpd,wind$CSpd) #correlation
```

```
cov(wind$RSpd,wind$CSpd) #covariance

## Fit an SLR Model
wind.lm <- lm(CSpd~RSpd,data=wind) #lm is linear model
summary(wind.lm)
summary(wind.lm)$sigma
```