Windmill (Assignment 1)

Cody Frisby

January 19, 2016

#1

- We need to pick a site for a wind farm we have high confidence that we will get a return on our investment. Here, prediction is of the utmost importance. We need to be able to predict wind speeds at the canidate site with a high level of confidence. Statistical modeling can help help us account for a lot of the variability in the variable of interest if the other variable(s) are somewhat correlated with it.

#2

- Is an SLR model ok?

```
# first read the data into R.
W <- read.table("~/Documents/MATH3710/problem1/Windmill.txt", header = TRUE)
# look at a summary of the data.
summary(W)
```

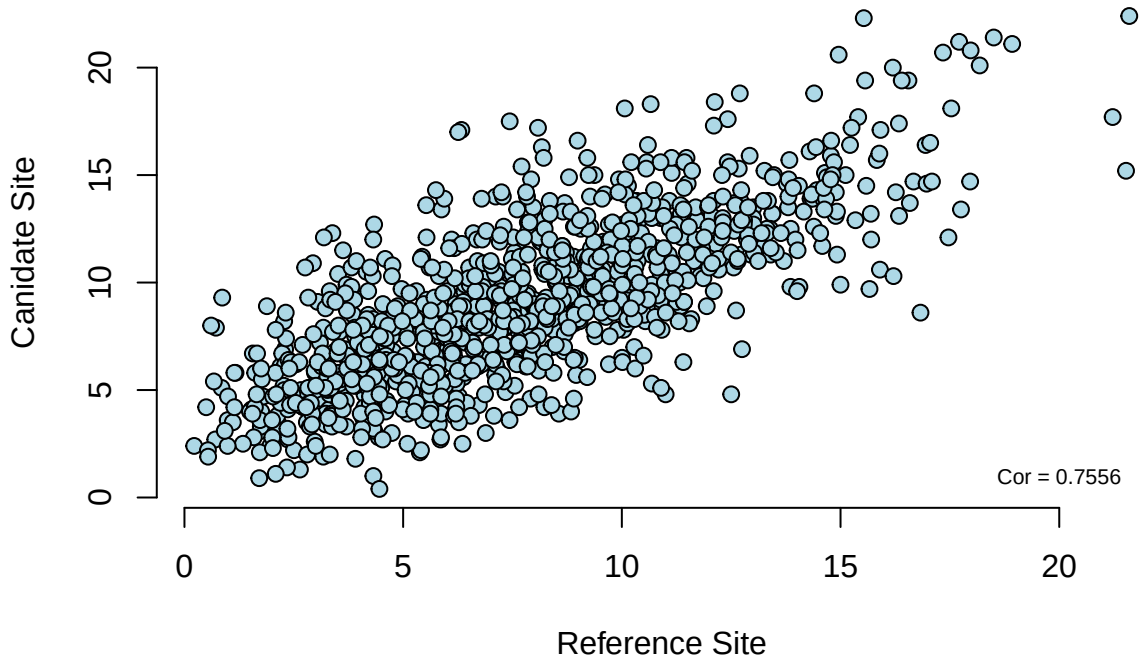
```
##       CSpd             RSpd
##  Min.   : 0.400   Min.   : 0.2221
##  1st Qu.: 6.100   1st Qu.: 4.7769
##  Median : 8.800   Median : 7.5477
##  Mean   : 9.019   Mean   : 7.7773
##  3rd Qu.:11.500   3rd Qu.:10.2096
##  Max.   :22.400   Max.   :21.6015
```

```
x <- W$RSpd
y <- W$CSpd
c(sd(x), sd(y))
```

```
## [1] 3.762639 3.763328
```

The two variables appear to be linearly related. They have similar variances but a little differnt means. A simple linear model may be ok. A scatter plot is here.

```
plot(x, y, bg = "lightblue", col = "black", cex = 1.1, pch = 21,
     frame = FALSE, xlab = "Reference Site", ylab = "Canidate Site")
#add correlation value to plot
text(x=20, y=1, paste0("Cor = ",round(cor(W)[2], 4)), cex = 0.7)
```

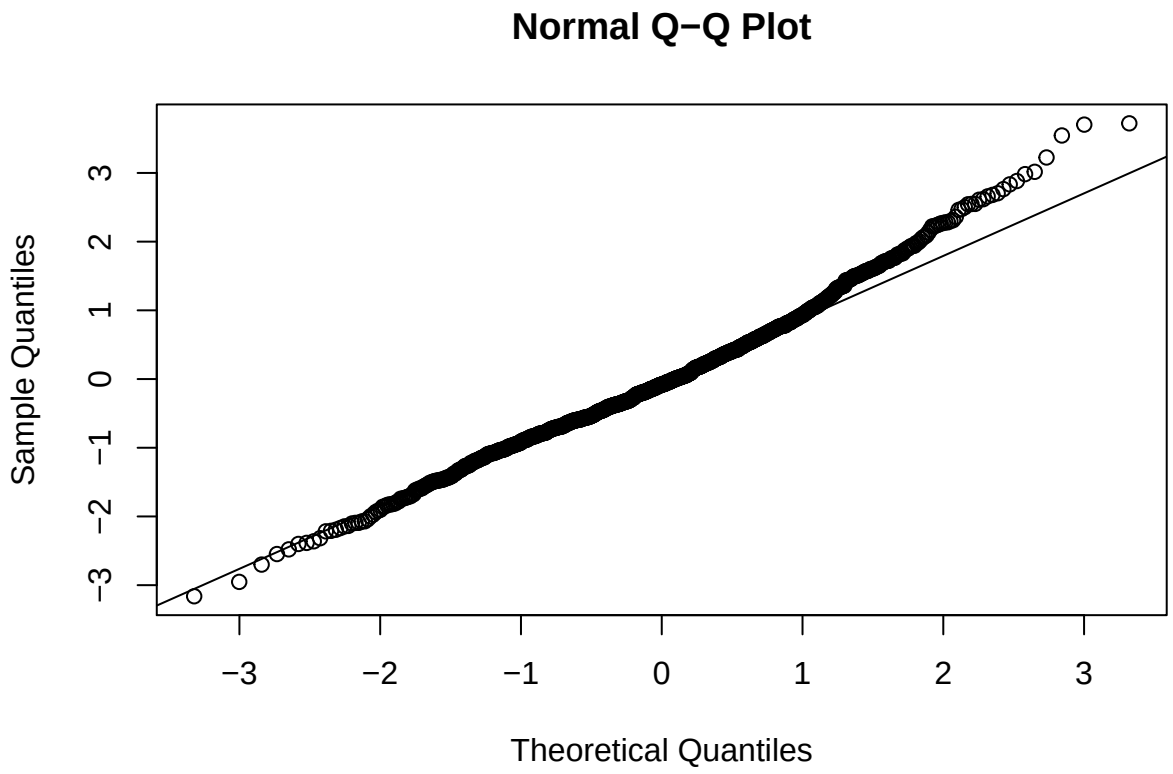


And now to look at some diagnostics of the standardized residuals from our model. This residual plot doesn't appear to violate any linearity assumptions. There are a few outliers beyond $\pm$ 3 standard deviations.

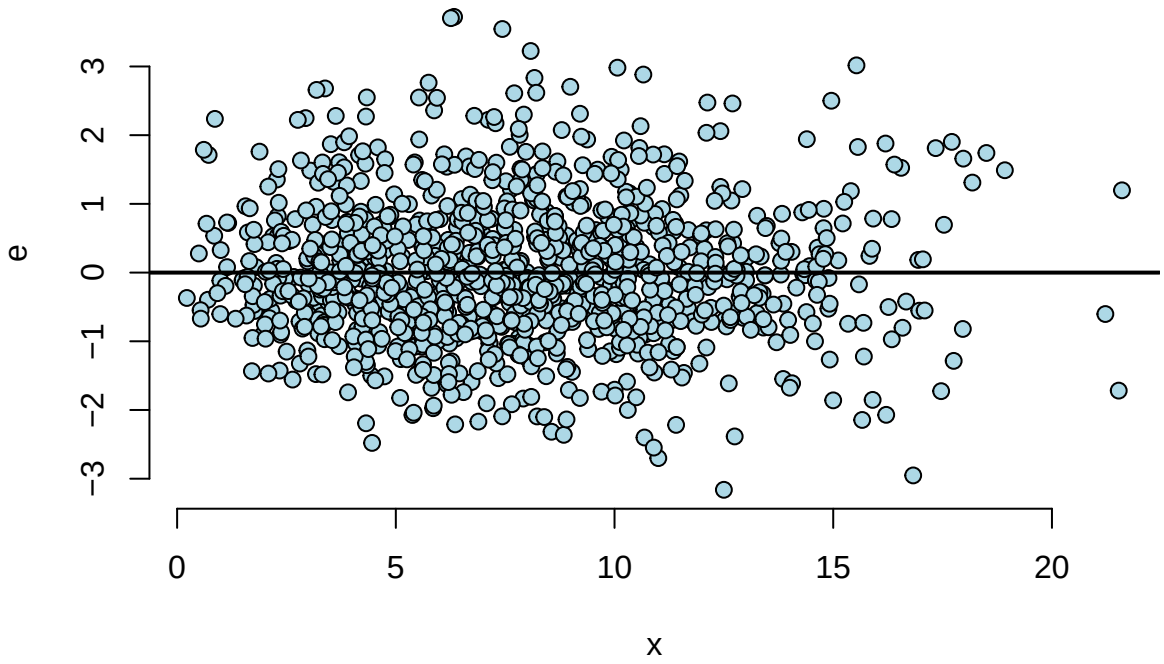
```
e <- rstandard(lm(y ~ x))
summary(e)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -3.161000 -0.644100 -0.080930  0.000012  0.584700  3.722000
```

```
qqnorm(e)
qqline(e)
```




```
# Residuals vs. x
plot(x, e, bg = "lightblue",
     col = "black", cex = 1.1, pch = 21,frame = FALSE)
abline(h = 0, lwd = 2)
```



The residuals from our model look ok. The upper quantiles seem to be wandering off from the qqline so our

linear model may not work so well as our predictor increases.

#3

Let x_i = RSpd, the wind speed at the reference site for $i = 1,\ldots,n$. Let y_i = CSpd, the wind speed at the reference site for $i = 1,\ldots,n$.

The equation for the sample regression model is

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad for\ 0.2221 \leq x \leq 21.6015$$

β_0 is the model parameter known as the intercept. This is the value for y_i when x_i is equal to zero. We do not know this parameter so we estimate it using the sample data. The model parameter β_1 is the slope

The equation for the errors is given by $\hat{e}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$.

And here are the derived formulas for $\hat{\beta}_1$ and $\hat{\beta}_0$:

$$\hat{\beta}_1 = Cor(Y, X)\frac{Sd(Y)}{Sd(X)} \qquad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

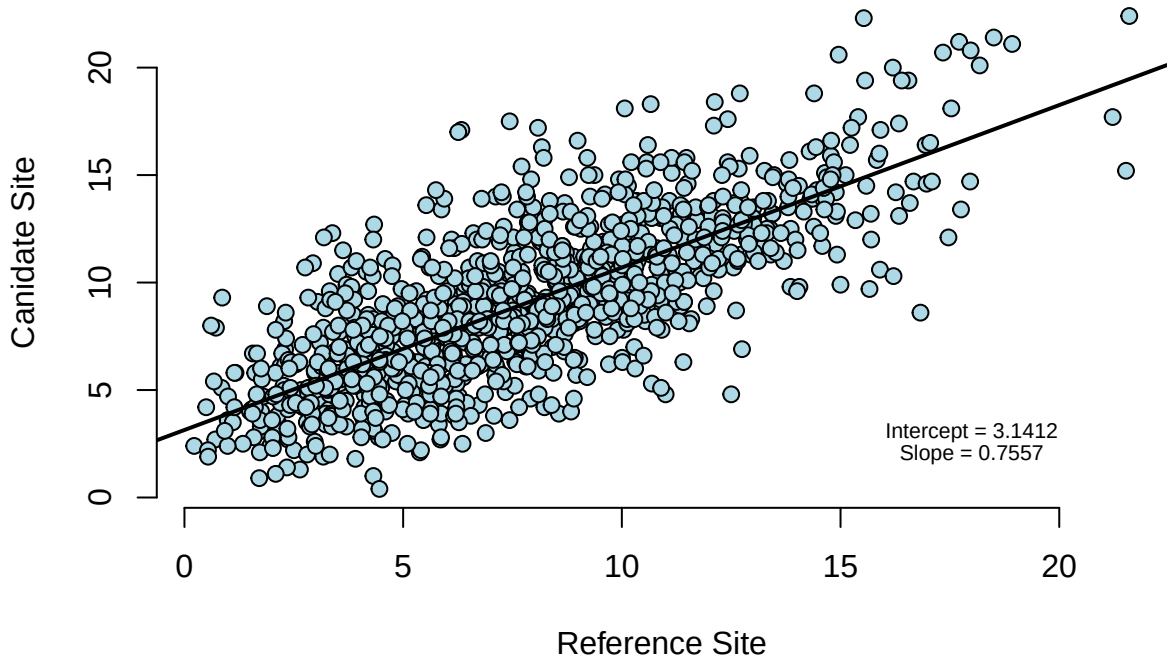
So, thinking aout this in the context of wind speed, we want an equation that for a given reference site wind speed (predictor) we can predict with the least amount of error at the site of interest.

```
# calculate ssx, ssy, sxy, b1, b0, sse, mse, and rmse in R using our sample
# data:
ssx <- sum((x - mean(x))^2)
ssy <- sum((y - mean(y))^2)
sxy <- sum((x-mean(x))*(y-mean(y)))
b1 <- sxy/ssx
b0 <- mean(y) - (b1 * mean(x))
sse <- ssy - (sxy^2/ssx)
mse <- sse/(length(x) - 2)
rmse <- sqrt(mse) #same value as our models Residual square error.
```

#4

- Fit a linear model and plot the data with fitted least squares regression.

```
#fit a linear model
wm <- lm(y ~ x)
plot(x, y, bg = "lightblue", col = "black", cex = 1.1, pch = 21,
     frame = FALSE, xlab = "Reference Site", ylab = "Canidate Site")
abline(wm, lwd = 2)
#add some text to the plot
text(x=18, y=3, paste0("Intercept = ", round(coef(wm)[1], 4)), cex = 0.7)
text(x=18, y=2, paste0("Slope = ", round(coef(wm)[2], 4)), cex = 0.7)
```



Equation of regression line:

$$CanidateSite = 0.7557 * ReferenceSite + 3.1412$$

Our equation is that of a line of the form $y = mx + b$. Our slope is equal to 0.7557. The intercept is 3.1412. This means that if we had zero m/s wind speed at the reference site we would expect approx. 3.14 m/s wind speed at the canidate site. We must keep in mind $0.2221 \leq x \leq 21.6015$.

```
# summary of the model
summary(wm)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## 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 ***
## x            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
```

#5

- If we are to use our model to make predictions in wind speed at the canidate site, we would think about for every unit increase in wind speed at reference site there will be an approximate 75% increase at the canidate site. Below we plug 12 for ReferenceSite into our model.

```
b1 * 12 + b0
```

```
## [1] 12.21003
```

#6

- Use the model to predict wind speed at canidate site when reference site is 30 m/s.

```
b1 * 30 + b0
```

```
## [1] 25.81323
```

This is extrapolating and should not be done. The maximum value of our predictor variable is 21.6015. Thirty is well beyond this value. Also, 30 is greater than our prediction of 25.8132322 when the mean of the canidate site was almost 1.5 m/s greater than the refernce site. If the model were any good in this range than we would expect the prediction to be greater than 30 m/s or at least a lot closer to it.