Windmill (Assignment 1)

Cody Frisby

January 13, 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?

```r
# 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)
```

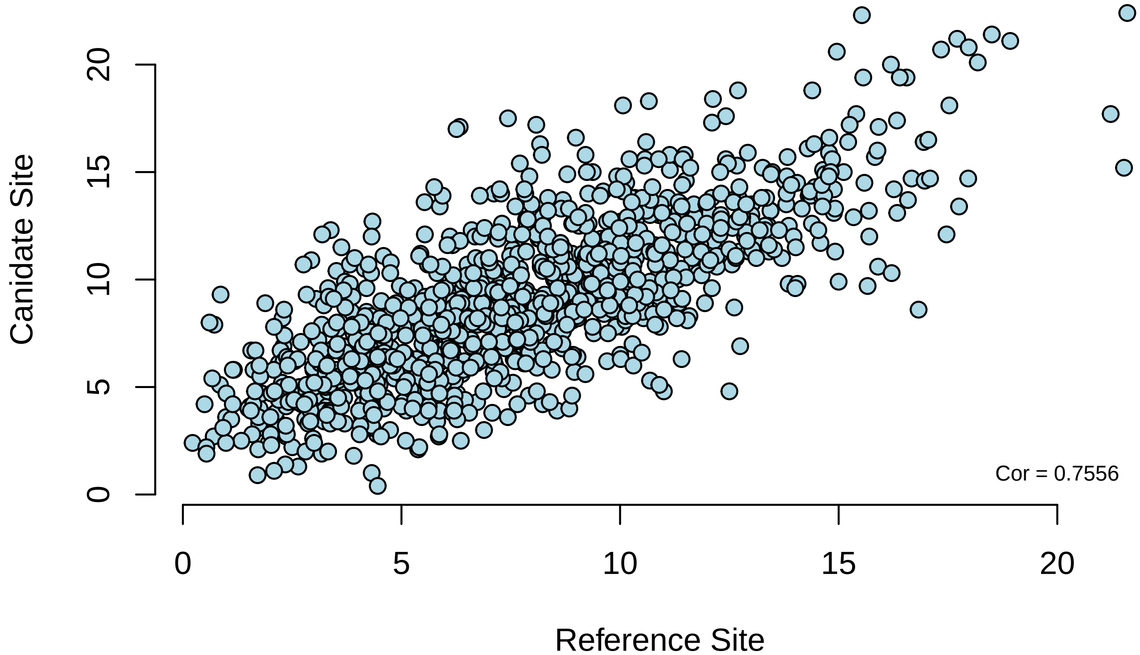
```
##       CSpd             RSpd
##  Min.   : 0.400   Min.   : 0.2221
##  1st Qu.: 6.100   1st Qu.: 4.7769
##  Median : 8.800   Median : 7.5477
##  Mean   : 9.019   Mean   : 7.7773
##  3rd Qu.:11.500   3rd Qu.:10.2096
##  Max.   :22.400   Max.   :21.6015
```

```r
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.

```r
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)
```

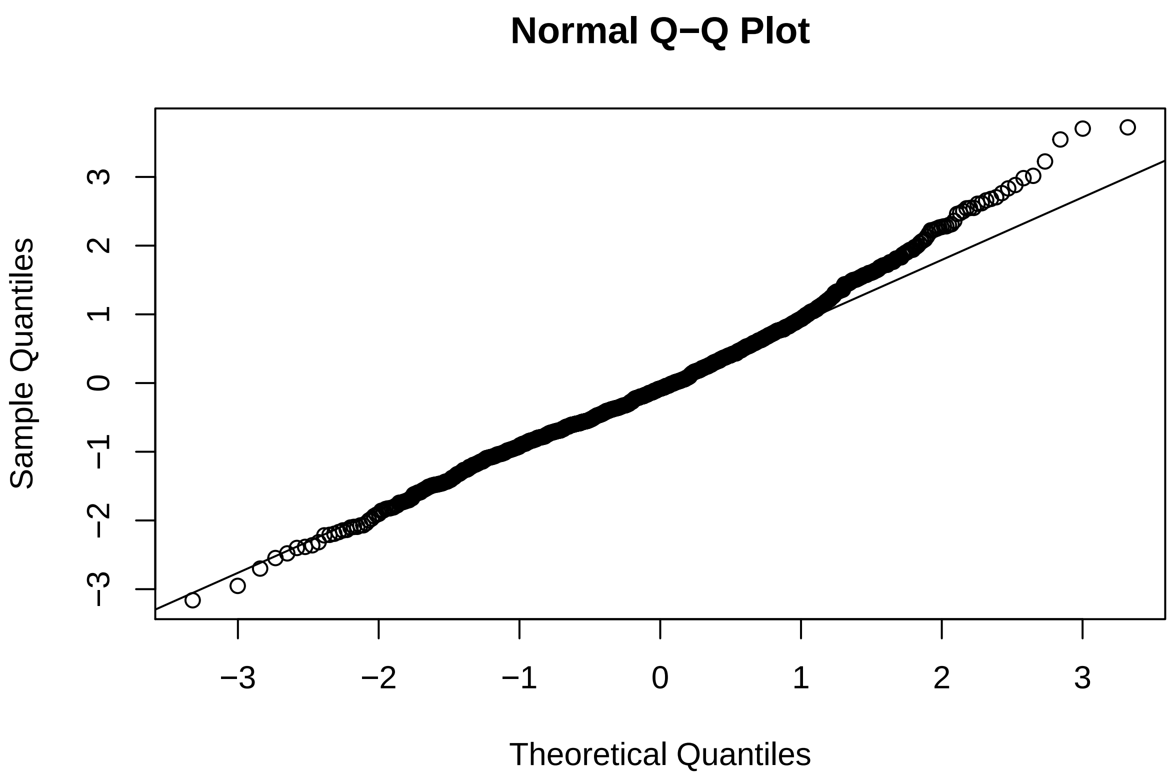


And now to look at some diagnostics of the standardized residuals from our model. This residual plot doesn't appear to violate any linearity assumptions. There are a few outliers beyond ± 3 standard deviations.

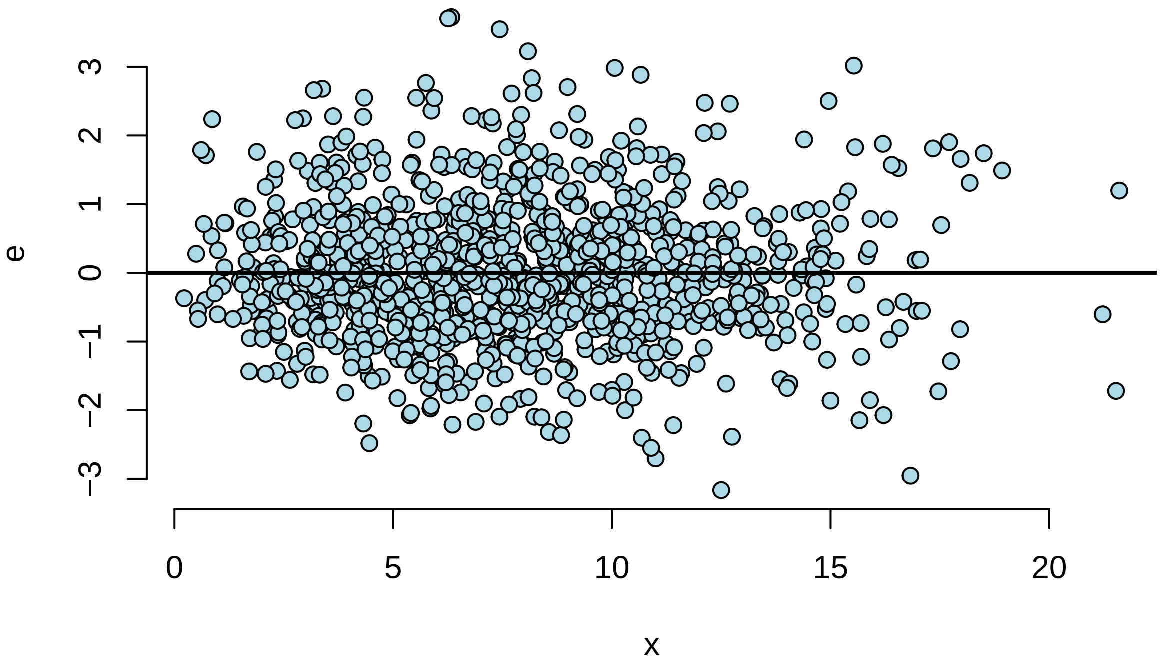
```
e <- rstandard(lm(y ~ x))
summary(e)
```

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

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




```r
# 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

The equation for the population regression line is $\mu_Y(x) = \beta_0 + \beta_1 x_i + \epsilon_i$

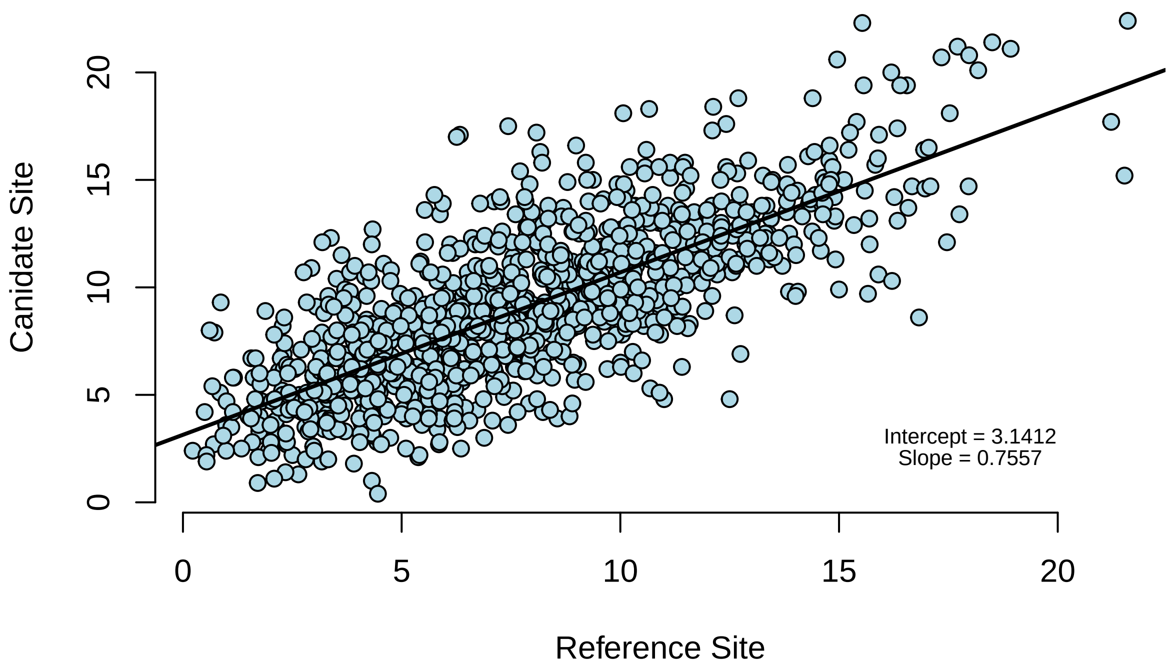
And here are the derived formulas for $\hat{\beta}_1$ and $\hat{\beta}_0$:

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

#4

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

```
#fit a linear model
wm <- lm(W$CSpd ~ W$RSpd)
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:

$$CanidateSite = 0.7557 * ReferenceSite + 3.1412$$

#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 a 75% increase at the canidate site. Below we plug 12 into our model.

```r
b0 <- coef(wm)[1]
b1 <- coef(wm)[2]
yhat <- b1 * 12 + b0
names(yhat) <- c("Predicted Wind Speed (m/s)")
yhat
```

```
## Predicted Wind Speed (m/s)
##                   12.21003
```

#6

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

```r
yhat <- b1 * 30 + b0
names(yhat) <- c("Predicted Wind Speed (m/s)")
yhat
```

```
## Predicted Wind Speed (m/s)
##                   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.