# YK_Assignment8

*Yinan Kang*

*4/15/2019*

```
# Packages
install.packages("dplyr")
require(dyplr)
```

## Problem 8.22

(1) What is the meaning of Beta3?

Response: Beta3 indicates how much higher (or lower) the response function for tool model M3 is than the one for tool model M1.

(2) What is the meaning of Beta4 - Beta3?

Response: Beta4 - Beta3 measures how much higher (or lower) the response function for tool models M4 is than the response function for tool models M3 for any given level of tool speed.

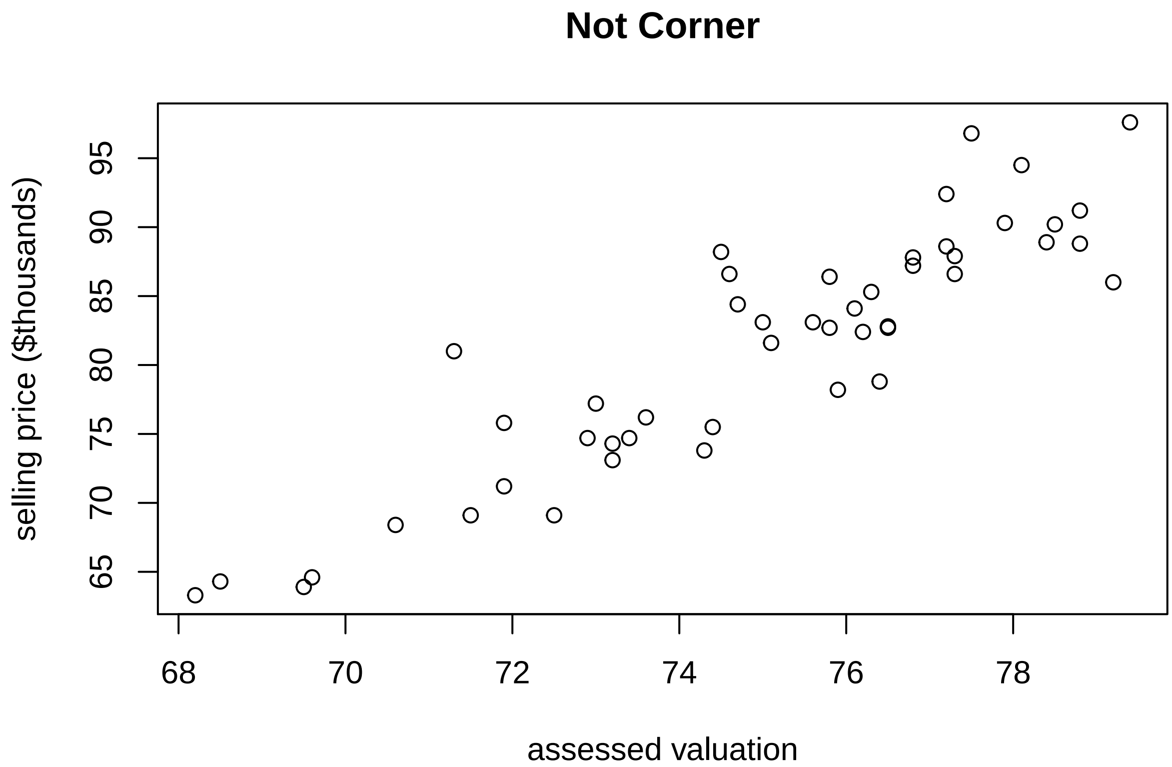
(3) What is the meaning of Beta1?

Response: Beta1 represents the effect of tool speed (X1) on tool wear (Y).

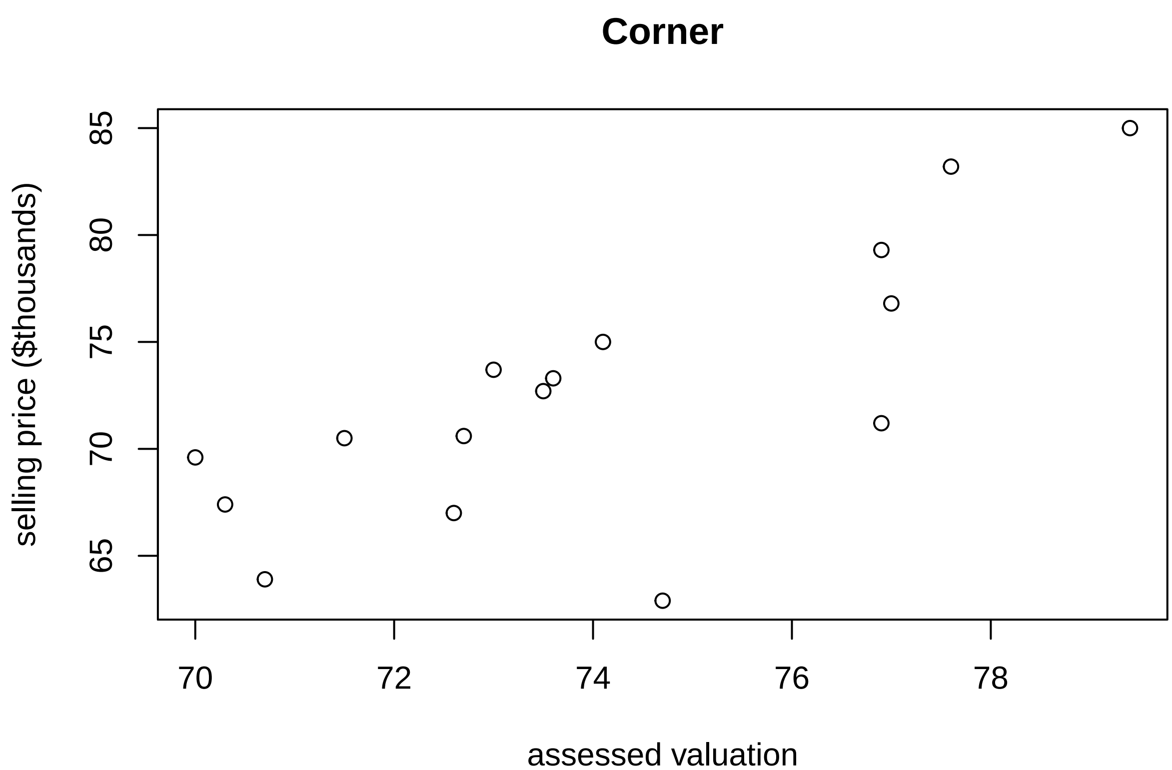
## Problem 8.24 - Assessed Valuations

### (a) Scatterplots

```
rm(list=ls())
colnames <- c("y","x1","x2")
df <- read.table(url("http://users.stat.ufl.edu/~rrandles/sta4210/Rclassnotes/data/textdatasets/KutnerDa
n <- nrow(df)
attach(df)

model <- lm(y~x1+x2+x1:x2)

notcorner.subset <- dplyr::filter(df,df$x2==0)
corner.subset <- dplyr::filter(df,df$x2==1)

plot(notcorner.subset$x1,notcorner.subset$y,  main="Not Corner", xlab = "assessed valuation", ylab = "s
```




```
plot(corner.subset$x1,corner.subset$y,  main="Corner", xlab = "assessed valuation", ylab = "selling pric
```




Do the potential regressions appear similar?

Analysis: From a purely visual standpoint, the two potential regressions SEEM TO BE DIFFERENT. While they would likely both have an intercept value approximately just under $65K selling price, the slope for

X2=0 (non-corner) data points seem to be noticeably higher than the slope for X2=1 (corner).

### (b) Test for identity of the regression functions (corner and non-corner)

```
corner.model <- lm(y~ x1 + x2 + x1:x2, data=corner.subset)
notcorner.model <- lm(y~ x1 + x2 + x1:x2, data=notcorner.subset)
anova(corner.model)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## x1         1 334.91  334.91  18.849 0.0006769 ***
## Residuals 14 248.75   17.77
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(notcorner.model)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## x1         1 3030.46 3030.46   211.1 < 2.2e-16 ***
## Residuals 46  660.36   14.36
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Equation for F Statistic: F = [ (RSS of Reduced) - (RSS of Full) / (dfR-dfF) ] / [(RSS of Full) / dfF]
As Corner dwellings have much less observations, it is the "Restricted" model technically

Alternatives:
H0: X2 is significant to model (corner status is significant) Ha: X2 is not significant to model

Decisions:
If F-stat <= F-crit (calculated below), conclude H0
If F-stat > F-crit, conclude Ha

Calculating F statistic:

```
rssr <- 248.75
rssf <- 660.36
dfr <- 15
dff <- 47
f.stat <- ((rssr-rssf)/(dfr-dff))/(rssf/dff)
f.crit <- qf(0.95,63,47)

print(f.stat)
```

```
## [1] 0.9154888
```

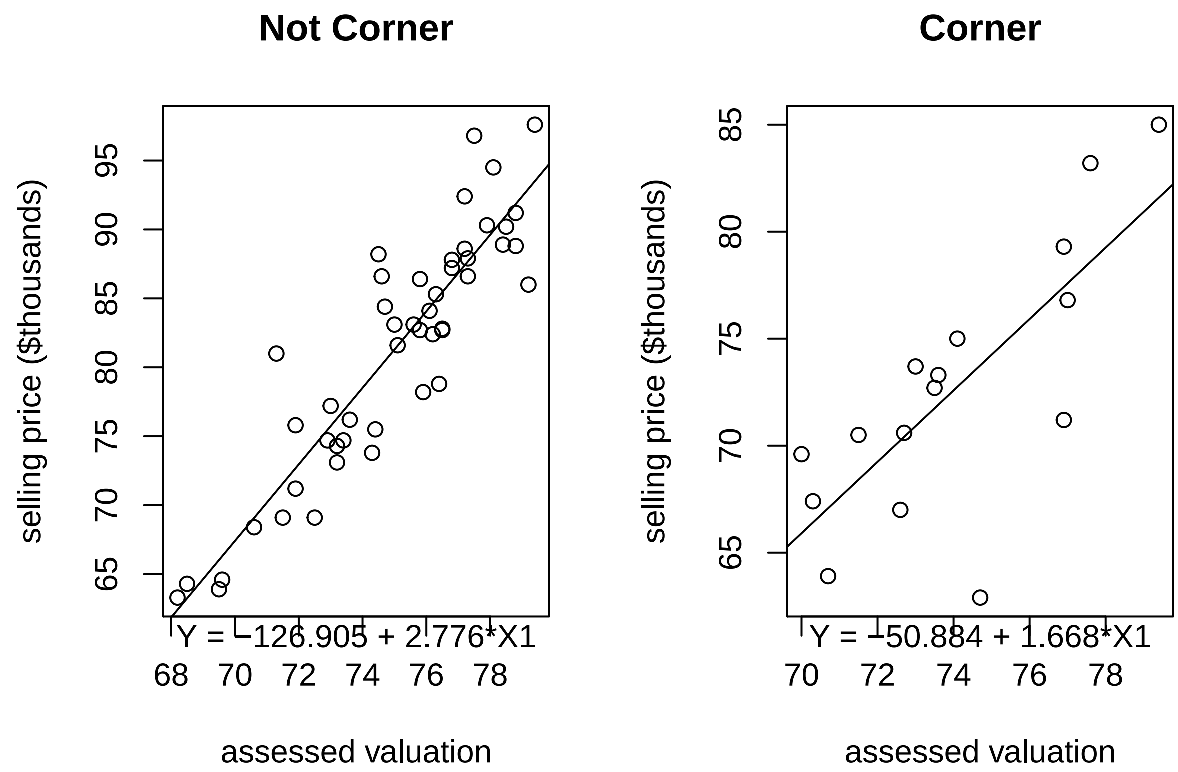
```
print(f.crit)
```

```
## [1] 1.58568
```

Conclusion: As f.stat < f.crit, we conclude H0, that X2 is significant, and MODEL SHOULD KEEP X2.

### (c) Plot regressions

```
par(mfrow=c(1,2))
plot(notcorner.subset$x1,notcorner.subset$y,  main="Not Corner", xlab = "assessed valuation", ylab = "se
mtext("Y = -126.905 + 2.776*X1",1)
abline(notcorner.model)
plot(corner.subset$x1,corner.subset$y,  main="Corner", xlab = "assessed valuation", ylab = "selling pri
abline(corner.model)
mtext("Y = -50.884 + 1.668*X1",1)
```




The Regression Equations:
Corner: Y = -50.884 + 1.668*X1*
*Not Corner: Y = -126.905 + 2.776*X1

We see that the slope for non-corner dwelling is higher, that selling price rices with assessed valuation more so than in corner dwellings.

## Problem 8.29 - Second Order

### Inputting sets of X

```
set1 <- c(1,1.5,1.1,1.3,1.9,.8,1.2,1.4)
set2 <- c(12,1,123,17,415,71,283,38)
```

### Calculation coefficients of correlation

```
set1.sq <- set1^2
set1.cu <- set1^3
cor(set1, set1.sq)
```

```
## [1] 0.9902871
```

```
cor(set1, set1.cu)
```

```
## [1] 0.9659484
```

```
set2.sq <- set2^2
set2.cu <- set2^3
cor(set2, set2.sq)
```

```
## [1] 0.9699782
```

```
cor(set2, set2.cu)
```

```
## [1] 0.9290059
```

### Summary of results

Coefficients of Correlation in Set 1:
X and X^2: 0.9902871
X and X^3: 0.9659484
x and x^2 (little x's): 0

Coefficients of Correlation in Set 2:
X and X^2: 0.9699782
X and X^3: 0.9290059
x and x^2 (little x's): 0

Analysis: We see that there are low multicollinearity levels, and no curvature and interaction effects are needed.

## Problem 8.36 - CDI

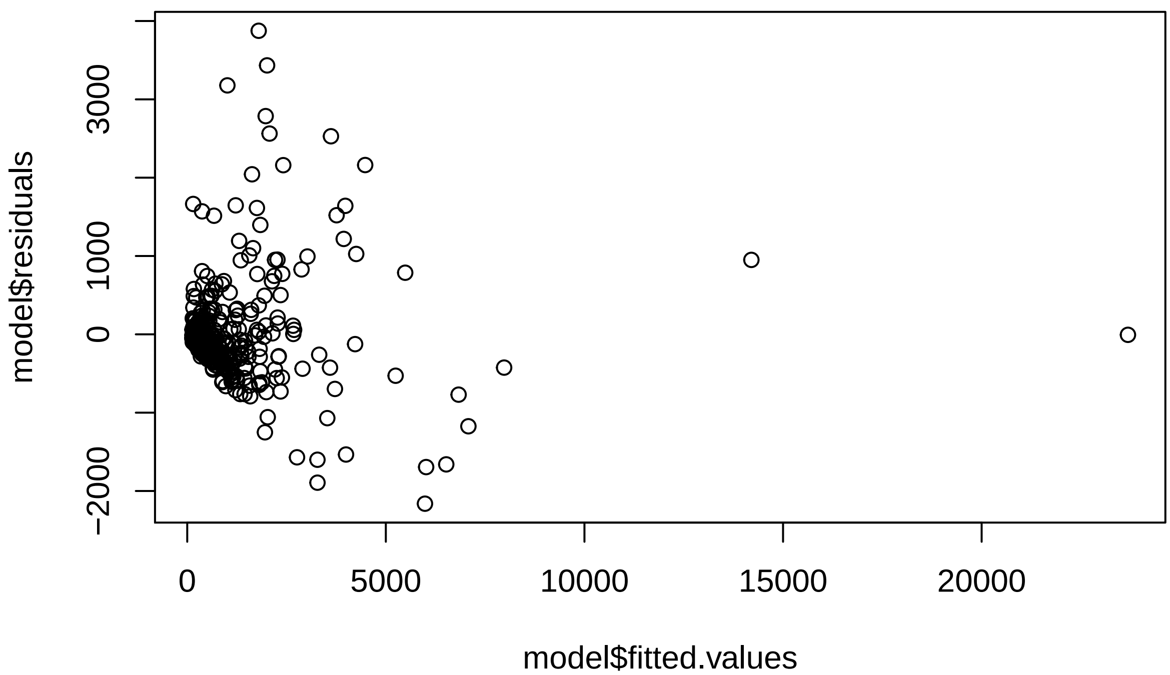
```
# Import Data
rm(list=ls())
cdi.df <- read.csv("CDI.csv")
n <- nrow(cdi.df)

# Create X^2 (X = total population)
cdi.df$popsq <- (cdi.df$Total.Population)^2
attach(cdi.df)
```

### (a) Fit 2nd degree model

```
# Create 2nd degree model
model <- lm(Number.Active.Physicians~Total.Population+popsq)
summary(model)
```

```
##
## Call:
## lm(formula = Number.Active.Physicians ~ Total.Population + popsq)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2161.9  -201.3   -59.6    48.1  3875.4
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -1.674e+02  4.214e+01  -3.971 8.36e-05 ***
## Total.Population  2.983e-03  9.313e-05  32.031  < 2e-16 ***
## popsq            -3.295e-11  1.400e-11  -2.353   0.0191 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 606.9 on 437 degrees of freedom
## Multiple R-squared:  0.8855, Adjusted R-squared:  0.885
## F-statistic:  1690 on 2 and 437 DF,  p-value: < 2.2e-16
```

```r
plot(model$fitted.values, model$residuals)
```




Analysis: Judging by the plot above, for the majority of fitted values, the residuals lie between -1000 and 1000, which is not outstanding, but due to the scope of the dataset, it DOES appear to show representation with the data.

### (b)

R-squared for the 2nd order model = 0.886

```r
# Fitting first order model
model1 <- lm(Number.Active.Physicians~Total.Population)
summary(model1)
```

```
##
## Call:
## lm(formula = Number.Active.Physicians ~ Total.Population)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1969.4  -209.2   -88.0    27.9  3928.7
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -1.106e+02  3.475e+01  -3.184  0.00156 **
## Total.Population  2.795e-03  4.837e-05  57.793  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 610.1 on 438 degrees of freedom
## Multiple R-squared:  0.8841, Adjusted R-squared:  0.8838
## F-statistic:  3340 on 1 and 438 DF,  p-value: < 2.2e-16
```

Analysis: R-squared for 1st order is 0.884. 2nd order DID NOT significantly increase R-squared.

## (c)

Alternatives:
H0: X^2 is statistically significant (helpful to model)
Ha: X^2 is not significant and can be dropped

Decision Rules:
If p-value <= alpha, conclude H0
If p-value > alpha, conclude Ha

alpha = 0.5

```r
anova(model1, model)
```

```
## Analysis of Variance Table
##
## Model 1: Number.Active.Physicians ~ Total.Population
## Model 2: Number.Active.Physicians ~ Total.Population + popsq
##   Res.Df       RSS Df Sum of Sq      F  Pr(>F)
## 1    438 163025135
## 2    437 160985454  1   2039681 5.5368 0.01906 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As p = 0.1906 < 0.5, we fail to reject H0 and CANNOT DROP 2nd order term.