Correlations & Regressions

Antonio Vazquez Brust

November 17, 2015

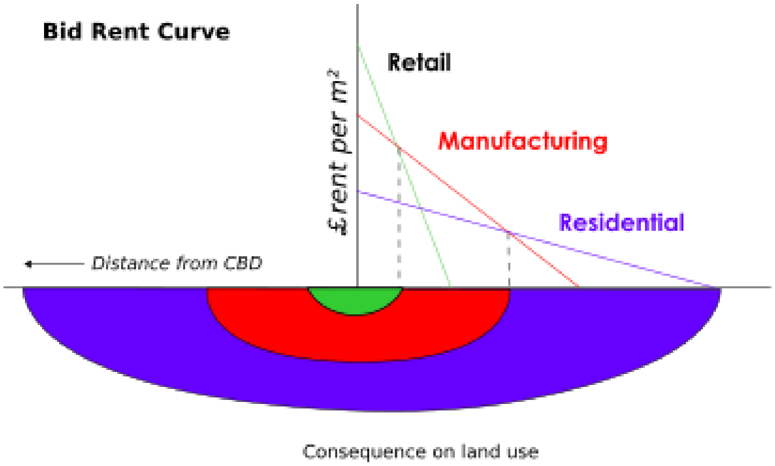
```
options(scipen = 99)
```

```
TAdata <- read.csv('c:/Users/havb/Dropbox/MSUI/Big Data for Cities - PPUA 5262 - 01/R/dat
a/Tax Assessor/TAdata.csv')

write.csv(TAdata, file = 'c:/Users/havb/Dropbox/MSUI/Big Data for Cities - PPUA 5262 - 0
1/R/data/Tax Assessor/TAdata.csv', row.names = F)

library(dplyr)
```

```
## 
## Attaching package: 'dplyr'
## 
## The following objects are masked from 'package:stats':
## 
##     filter, lag
## 
## The following objects are masked from 'package:base':
## 
##     intersect, setdiff, setequal, union
```

```
TAdata <- mutate(TAdata,
                 BLDG_RANK = ceiling(rank(AV_BLDG_PER_SF,na.last="keep")/
                                      length(which(!is.na(AV_BLDG_PER_SF)))*100) )
```

This week, I would like to compare the spatial distribution of building value in Boston against the now classical concentric zone model developed in the University of Chicago in the ’20s. The concentic zone model explains the demand (and hence, value) of land as a function of its closeness to the city’s center, or central business district.



from https://en.wikipedia.org/wiki/Concentric_zone_model (https://en.wikipedia.org/wiki/Concentric_zone_model)

This model, while nowadays considered too simplistic, is a good starting point for our analysis of building value distribution.

Having categorized the building of Boston in a 1 to 100 “value ranking”, we can check if there’s a correlation between the value ranking and the distance to eeh city’s center. We also have a variable that measures the distance between every parcel in our dataset and Boston Common, so that will be our central point.

First, let’s see if there’s a correlation between distance to the Common and building value ranking:

```
require(Hmisc)

rcorr(as.matrix(TAdata[c("BLDG_RANK","DIST_TO_DX_MI")]))
```

```
##               BLDG_RANK DIST_TO_DX_MI
## BLDG_RANK          1.00         -0.55
## DIST_TO_DX_MI     -0.55          1.00
##
## n
##               BLDG_RANK DIST_TO_DX_MI
## BLDG_RANK        137827        137827
## DIST_TO_DX_MI    137827        168146
##
## P
##               BLDG_RANK DIST_TO_DX_MI
## BLDG_RANK                0
## DIST_TO_DX_MI  0
```

There’s a substantive negative correlation; the higher the distance from the Common, the lower the value ranking. Let’s run a regression analysis to see if the correlation is statistically significant:

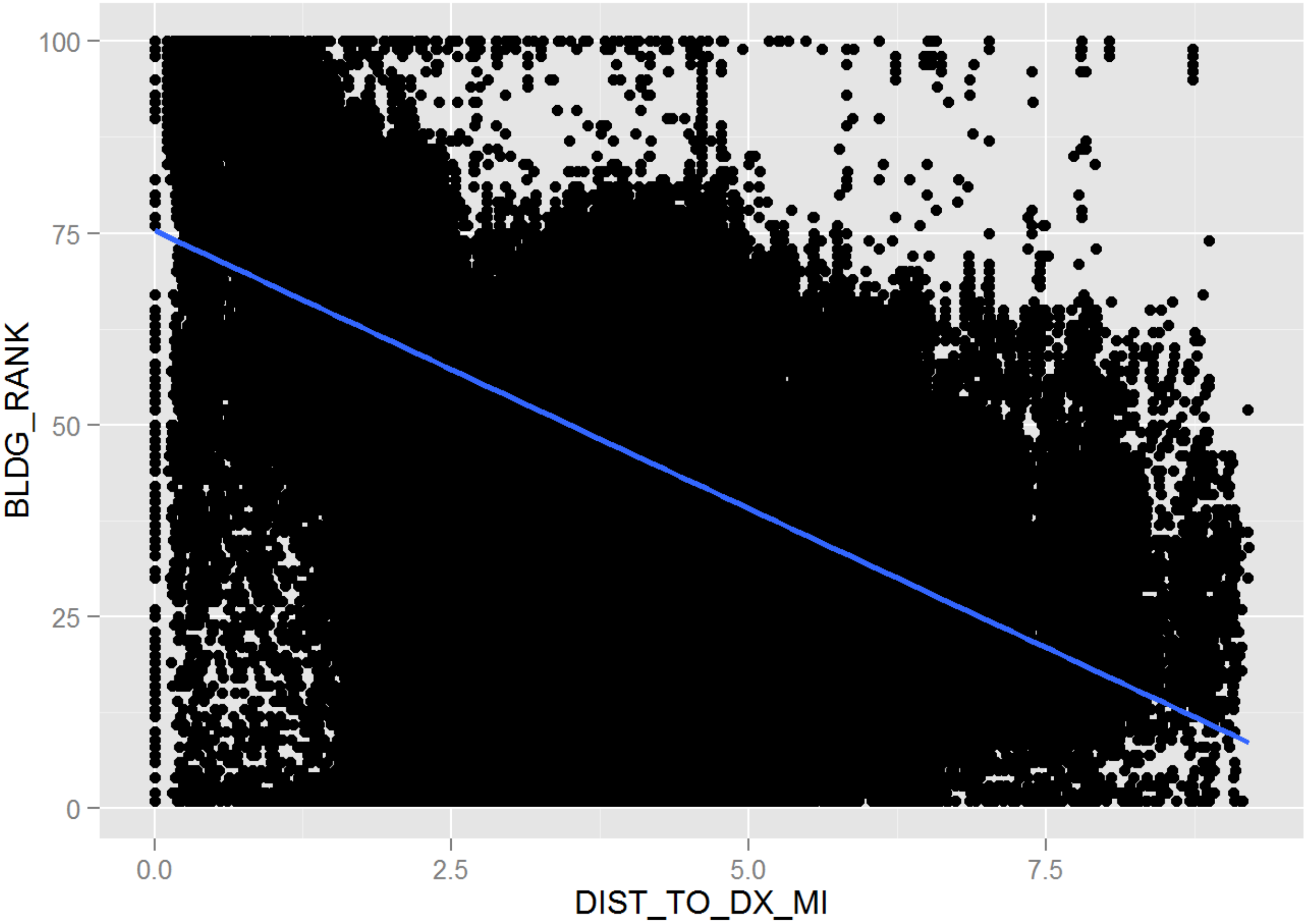
```
regression.bldg.rank <- lm(data = TAdata, BLDG_RANK ~ DIST_TO_DX_MI)
summary(regression.bldg.rank)
```

```
##
## Call:
## lm(formula = BLDG_RANK ~ DIST_TO_DX_MI, data = TAdata)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -74.34 -16.55   4.86  19.53  86.96
##
## Coefficients:
##               Estimate Std. Error t value            Pr(>|t|)
## (Intercept)   75.34371    0.12127   621.3 <0.0000000000000002 ***
## DIST_TO_DX_MI -7.24919    0.02986  -242.8 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24.16 on 137825 degrees of freedom
##   (30319 observations deleted due to missingness)
## Multiple R-squared:  0.2995, Adjusted R-squared:  0.2995
## F-statistic: 5.894e+04 on 1 and 137825 DF,  p-value: < 0.00000000000000022
```

It definitely is! The resulting p-value of 0.00000000000000022 indicates that such a correlation of value vs distance from the common, has les than a chance in a trillion to be explained by pure chance. The adjusted R^2 value of 0.2995 means than we can explain almost 30% of the variation in building ranking by its distance to the center alone.

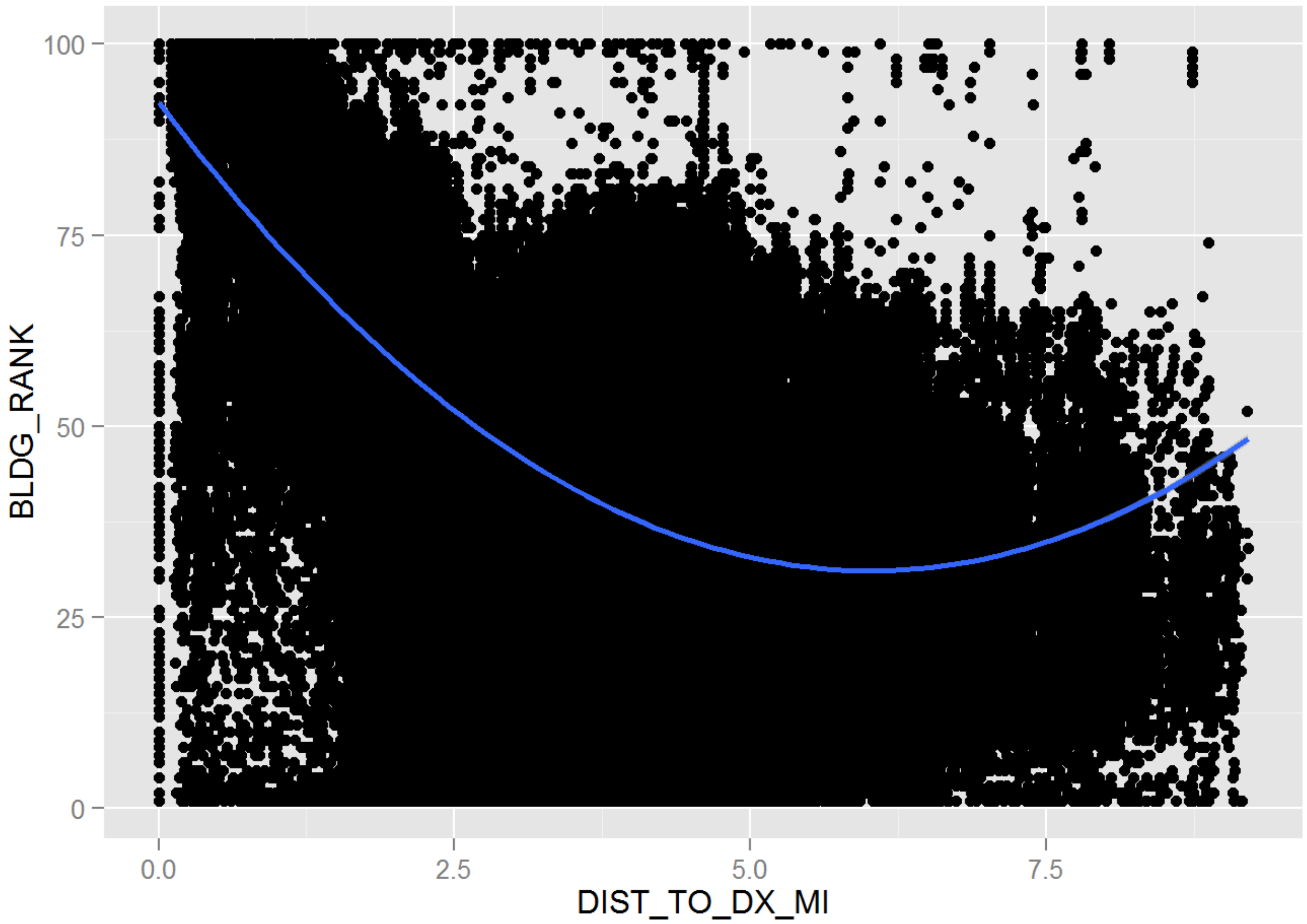
This is how the correlation between fistance and value looks when plotted:

```
ggplot(data = TAdata, aes(x = DIST_TO_DX_MI, y = BLDG_RANK)) +  geom_point() + stat_smoot
h(method = "lm", formula = y ~ x, size = 1)
```



The plot is interesting, because we can see that there something going on with distance that is not linear: buildings with really low value rankings are frequent at mid-distance from the center, but start to decrease as distance increases. This looks more like a quadratic (ie “curved” correlation). Let’s repeat our plot, this time fitting a quadratic regression line:

```
ggplot(data = TAdata, aes(x = DIST_TO_DX_MI, y = BLDG_RANK)) +  geom_point() + stat_smoot
h(method = "lm", formula = y ~ x + I(x^2), size = 1)
```



The fit now looks better, but is it significant? Let’s re run our linear model, this time using a quadratic formula:

```
regression.bldg.rank <- lm(data = TAdata, BLDG_RANK ~ DIST_TO_DX_MI+I(DIST_TO_DX_MI^2))
summary(regression.bldg.rank)
```

```
##
## Call:
## lm(formula = BLDG_RANK ~ DIST_TO_DX_MI + I(DIST_TO_DX_MI^2),
##     data = TAdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -91.374 -16.377   3.513  16.811  68.862
##
## Coefficients:
##                     Estimate Std. Error t value            Pr(>|t|)
## (Intercept)         92.37365    0.17315   533.5 <0.0000000000000002 ***
## DIST_TO_DX_MI      -20.37063    0.10405  -195.8 <0.0000000000000002 ***
## I(DIST_TO_DX_MI^2)   1.69381    0.01293   131.0 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.78 on 137824 degrees of freedom
##   (30319 observations deleted due to missingness)
## Multiple R-squared:  0.3771, Adjusted R-squared:  0.3771
## F-statistic: 4.172e+04 on 2 and 137824 DF,  p-value: < 0.00000000000000022
```

Well, that was an improvement. The model statistical significance is still huge, and our new adjusted R^2 value indicates that we can explain more than 37% of a building value ranking by it's distance alone! For such a simplistic model, the concentric zone model sure packs a punch.

The temptation of continuing to pick the low hanging fruit when it comes to explaining Boston property value is unresistable. Can we improve or predictive model even more if we add the percentage of minority population in the building's area as a variable?

```
demographics <- read.csv('C:/Users/havb/Dropbox/MSUI/Big Data for Cities - PPUA 5262 - 0
1/R/data/Census/Tract Census Data.csv', stringsAsFactors = FALSE)
```

Note to Professor O'Brian: The next step is to write a simple function that adds a new column in the tax assessor dataset with the "minority population percentage" variable, extracted by fetching the census tract demographic information for each parcel, and inversing the white population proportion.

I'll then see if I can continue improving the regression by adding the minority peprcentage variable to the linear model.