Part 3

Nick George

3/21/2018

Nick George and Spencer Louie

Introduction:

Regression Model:

```
##              lprice grade lsqft_living view condition lsqft_lot waterfront
## lprice         1.00  0.70         0.67 0.35      0.04      0.14       0.17
## grade          0.70  1.00         0.74 0.25     -0.14      0.18       0.08
## lsqft_living   0.67  0.74         1.00 0.25     -0.05      0.32       0.08
## view           0.35  0.25         0.25 1.00      0.05      0.12       0.40
## condition      0.04 -0.14        -0.05 0.05      1.00      0.07       0.02
## lsqft_lot      0.14  0.18         0.32 0.12      0.07      1.00       0.07
## waterfront     0.17  0.08         0.08 0.40      0.02      0.07       1.00
## bedbath        0.50  0.57         0.79 0.15     -0.05      0.17       0.03
## floors         0.31  0.46         0.37 0.03     -0.26     -0.24       0.02
##              bedbath floors
## lprice          0.50   0.31
## grade           0.57   0.46
## lsqft_living    0.79   0.37
## view            0.15   0.03
## condition      -0.05  -0.26
## lsqft_lot       0.17  -0.24
## waterfront      0.03   0.02
## bedbath         1.00   0.37
## floors          0.37   1.00
```

[pairs(kc_data_3) The graphs really slow things down so I don't have them included for ease of processing right now.]

The trouble explanatory variables with high correlation are: grade and lsqft_living as well as lsqft_living and bedbath. lsqft_living and lqft_lot are surprisingly not particularly problematic.

Interaction Terms: [We are not currently using interaction terms, however we could. Ones that could be potentially viable in a logical sense are: lsqft_living and floors, capturing so affect on how big each floor is. lsqft_lot and waterfront to try and capture how much waterfront property you have. And finally bedbath and lsqft_living because it could matter how spread out your bed and bath are in the house. However they are already high correlated so I'm not sure how that would affect it. Perhaps an interaction term with lsqft_living and lsqft_lot would be able to capture the effect of having a larger house on a smaller lot or vice versa.]

```
##
## Call:
## lm(formula = log(price) ~ grade + log(sqft_living) + view + condition +
##     log(sqft_lot) + waterfront + bedbath + floors, data = kc_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.30948 -0.23720  0.01315  0.22246  1.38121
##
## Coefficients:
```

```
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       8.160236   0.062686 130.177  < 2e-16 ***
## grade             0.201385   0.003097  65.028  < 2e-16 ***
## log(sqft_living)  0.478005   0.011354  42.100  < 2e-16 ***
## view              0.087473   0.003395  25.767  < 2e-16 ***
## condition         0.094509   0.003669  25.759  < 2e-16 ***
## log(sqft_lot)    -0.054304   0.002949 -18.411  < 2e-16 ***
## waterfront        0.379934   0.028915  13.140  < 2e-16 ***
## bedbath          -0.018108   0.002541  -7.125 1.07e-12 ***
## floors           -0.014428   0.005373  -2.685  0.00725 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.336 on 21604 degrees of freedom
## Multiple R-squared:  0.5931, Adjusted R-squared:  0.5929
## F-statistic:  3936 on 8 and 21604 DF,  p-value: < 2.2e-16
```

Grade is asscoiated with an "x" increase in price. Sqft_living is associated with an "x" increase in price. Having a view is associated with an "x" increase in price. An increase in condition by 1 is associated with an "x" increase in price. Sqft_lot is associated with an "x" decrease in price [CHECK - seems very strange.] Having a waterfront home is associated with an "x" increase in price. Having an extra bedbath is associated with an "x" decrease in price [CHECK - seems very strange.] Having an extra floor is associated with an "x" decrease in price [CHECK - might be strange, might not be can see logic for both ways]. All of the variables are significant at .001 significancy level. [I don't think it makes any sense to do a t-test other than just by 0, because I don't think the scale of the effect really matters here.]

```
kc_2.lm_nested <- lm(log(price) ~ log(sqft_lot) + log(sqft_living) + grade + waterfront + view + conditi
anova(kc_2.lm_nested, kc_2.lm) #Took out bedbath and floors as they were the least significant in the d
```

```
## Analysis of Variance Table
##
## Model 1: log(price) ~ log(sqft_lot) + log(sqft_living) + grade + waterfront +
##     view + condition
## Model 2: log(price) ~ grade + log(sqft_living) + view + condition + log(sqft_lot) +
##     waterfront + bedbath + floors
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1  21606 2446.5
## 2  21604 2439.4  2    7.0829 31.364 2.504e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The bigger model, which includes bedbath and floors, is definitely significant at a near 0 significancy level. I would absolutely report the larger model not only because it is very siginificant, but even if it were only partially significant because this model is meant largely for prediction. The model being parismonioius is not very important because we want to predict future home sales based on the homes specifications, rather than just trying to find a handful key predictors.

```
summary(kc_2.lm_nested)$r.squared
```

```
## [1] 0.5919162
```

```
summary(kc_2.lm)$r.squared
```

```
## [1] 0.5930976
```

```
summary(kc_2.lm_nested)$adj.r.squared
```

```
## [1] 0.5918028
```

```
summary(kc_2.lm)$adj.r.squared
```

```
## [1] 0.5929469
```

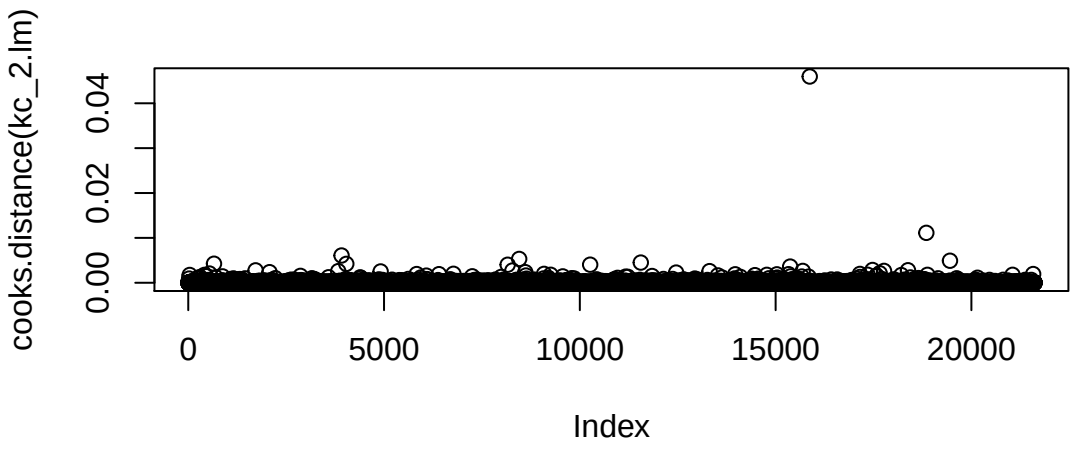
The nested model explains 59.2% of the variance in prices while the larger model explains 59.3% of the variance. The adjusted values are slightly smaller, but still in the same order. So the larger model definitely explains more even though it is less parismonous. Even though the R^2 is fairly large this is no guarantee that we have accurately described the population. An outlier could be throwing us off and pulling the line in a strange direction. Even if we impossibly had 100%, that only means that we have accurately described our sample, but the overall population might be quite different.

```
## [1] -5.042715  5.042715
```

Above 5.04 and below -5.04 are outliers. There are none that satisfy that requirement.

2p/n = 20/21603 = 9.26e-4. [Need to check this seems extreme. Will ask her.] One does seem like a notable outlier though, observation 15871.

	DFFITS	Intercept	Grade	lsq_liv	View	Condition	lsq_lot	Waterfront	BedBath	Floors	Cook's
15871	0.64	0.39	0.05	-0.42	0.02	0.03	0.05	0.02	0.64	-0.08	0.05



The only Cook's Distance that's significantly higher than the rest is at obesrvation 15871, but even that is still drasticly below 1. Overall it does not seem that any outlier is going to be problematic and so we do not need to look at data that excludes some points.

Partial Coefficent of Determination:

```
## [1] 0.07581951
```

```
## [1] 0.01544818
```

```
## [1] 0.0003336533
```

```
## [1] 0.007928433
```

```
## [1] 0.02981634
```

```
## [1] 0.02979712
```

```
## [1] 0.1636954
```

```
## [1] 0.002344566
```