Part 3

Nick George

3/21/2018

Nick George and Spencer Louie

Introduction

Our regression aims to examine the factors that drive house prices. Our dataset is from the Center for Spatial Data Science, which has collected data on house prices from the King County Washington area from 2014 to 2015 (May). On the left hand side, we have the log of house prices (we decided to log the house prices to correct heteroskedasticity issues) and we have a number of selections for RHS variables. RHS variables of interest include: the square footage of the living space, the lot size, number of floors, whether it is a waterfront property, the condition of the house (as based on King County grading system), and number of bedroom/bathrooms.

Regression Model:

The overall model we are looking at is one that include up to the log of living square footage, the log of lot square footage, number of floors, whether the house is waterfront or not, whether the house has a view or not, the condition of the house on a scale, the grade of the house on a scale, and the number of total beds and baths in the house.

```
##              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
```

An increase in the grade scale by 1 is asscoiated with an 20.1% increase in price. A 1% increase in sqft_living is associated with an .478% increase in price. Having a view is associated with an 8.7% increase in price. An increase in the condition scale by 1 is associated with an 9.5% increase in price. A 1% increase in sqft_lot is associated with an .054% decrease in price. Having a waterfront home is associated with an 37.9% increase in price. Having an extra bedbath is associated with a 1.8% decrease in price Having an extra floor is associated with a 1.4% decrease in price . All of the variables are significant at .001 significancy level. Some notable surprises are that having extra square footage in your lot is associated with a decrease in price, holding the other variables constant and having another bedroom or bathroom also is associated with a decrease in price holding the other factors constant. Having another floor associated with a decrease in price may seem strange, but makes sense when you consider that overall square footage is being held constant.

```
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
```

```
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
```

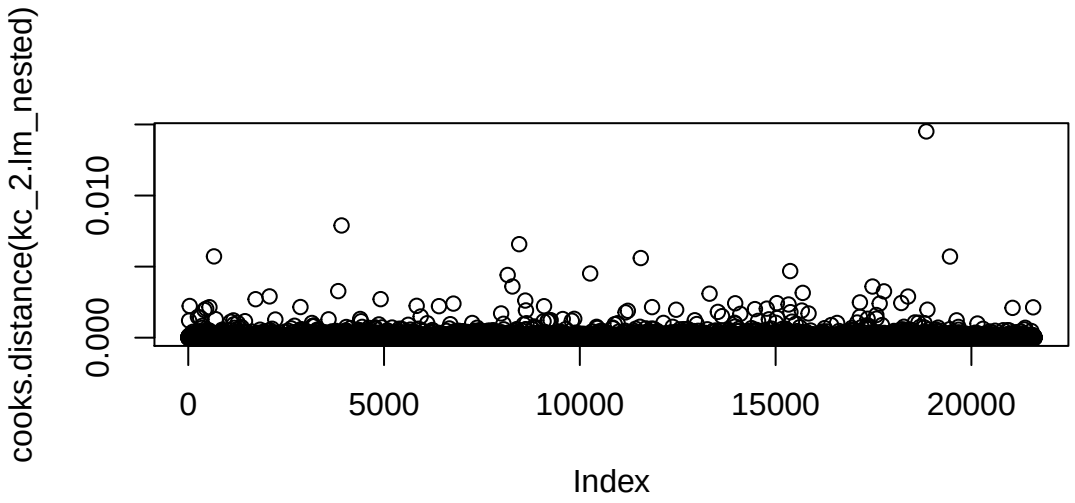
The nested model explains 59.2% of the variance in prices while the larger model explains 59.3% of the variance. The adjusted values are slightly smaller, but still in the same order. So the larger model definitely explains more, but it is less parsimonious. Since the R^2 change is so slight, we decided to use the nested model because it is more parsimonious and allows to more accrurately pinpoint driving factors of housing prices.

```
## [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. This likely too small of a cutoff point, however one observation does seem like a notable outlier though, observation 15871.

	DFFITS	Intercept	Grade	lsq_liv	View	Condition	lsq_lot	Waterfront	
15871	0.0206	-0.0025	-0.0033	0	-1e-04	4e-04	-0.0026	0.0186	
None of	these valu	s are partiul	arly out of	the ordina	ry so it d	oes not seem	that our da	ta is much inf	lu



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.

Model Selection:

This model was seleced using a step function going both forward and backward based on AIC. So first the function added a predictor, originally to the null model, that decreased the AIC the most significantly. It would do that again until it ran out of predictors or could no longer add a predictor that reduced AIC. After that it went backward seeing that if it could decrease AIC by removing any predictors. This way you can account for variables that may no longer be significant as others are added, such as through multicollinearity. This process implied that we should use the full possible model noted above. However, using the nested F-test we came to another conclusion, as we put more importance on parsimony than the process implies. So we decided to use the smaller model that did not include the sum of beds and baths as well as the number of floors in the home.

Model Interpretation:

```
summary(kc_2.lm_nested)
```

```
##
## Call:
## lm(formula = log(price) ~ log(sqft_lot) + log(sqft_living) +
##     grade + waterfront + view + condition, data = kc_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.32398 -0.23842  0.01451  0.22221  1.40680
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       8.429544   0.049190  171.37   <2e-16 ***
## log(sqft_lot)    -0.048306   0.002707  -17.85   <2e-16 ***
## log(sqft_living)  0.419464   0.008451   49.64   <2e-16 ***
## grade             0.200069   0.002983   67.07   <2e-16 ***
## waterfront        0.380848   0.028919   13.17   <2e-16 ***
## view              0.089536   0.003384   26.46   <2e-16 ***
## condition         0.096701   0.003591   26.93   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3365 on 21606 degrees of freedom
## Multiple R-squared:  0.5919, Adjusted R-squared:  0.5918
## F-statistic:  5223 on 6 and 21606 DF,  p-value: < 2.2e-16
```

All of our variables are significant at a near 0 significany level. We note that an increase in grade, living square footage, condition, and having a view or being waterfront are associated with increase in price. The only negative effect from a predictor is an increase in lot size leads to a decrease in price. Even in this smaller model we still find that to be the case, which again seems somewhat troublesome. None of these variables are overly correlated with each other with the exception of grade and the log of square foot living space at .74.

Condifence Intervals:

```
kc_3_pred.data <- data.frame(grade=8, sqft_living=1200, view=0, condition=3, sqft_lot=7500, waterfront=
kc_3_crit_val <- qt(.975, glance(kc_2.lm_nested)$df.resid)
kc_3_gl <- broom::glance(kc_2.lm_nested)
kc_3_sig <- dplyr::pull(kc_3_gl, sigma)
kc_3_pred <- broom::augment(kc_2.lm_nested, newdata=kc_3_pred.data, typepredict = "predict") %>% mutate
kc_3_pred
```

```
##   grade sqft_living view condition sqft_lot waterfront  .fitted    .se.fit
## 1     8        1200    0         3     7500          1 13.24406 0.02941195
##    .se.pred lower_PI upper_PI lower_CI upper_CI
## 1 0.3377833 12.58198 13.90614 13.18641 13.30171
```

PI: 12.58198 13.90614 CI: 13.18641 13.30171

```
exp(12.58198)
```

```
## [1] 291262.4
```

```
exp(13.90614)
```

```
## [1] 1094863
```

```
exp(13.18641)
```

```
## [1] 533071.1
```

```
exp(13.30171)
```

```
## [1] 598217.7
```

PI: 263050.2 1212288 CI: 533071.1 598217.7 95% of prices for a home with grade 8, 1200 living square footage, no view, a condition rating of 3, a 7500 square foot lot and a awater front view would be between 291,262 and 1,094,863. We are 95% confident that the average price of a home with those specifications would be between 533,071 and 598,217.

Partial Coefficent of Determination:

```
## [1] 0.1023593
```

```
## [1] 0.01452891
```

```
## [1] 0.007963239
```

```
## [1] 0.03139174
```

```
## [1] 0.03247452
```

```
## [1] 0.1723293
```

The higher partial coefficients of determination belong to the log of living square footage and grade, with .102 and .172 respectively. The partial coeffienct of determination measures the the marginal contribution of the variables when the others are included in the model. So the log of living square footage has a marginal effect of .10 and grade has a marginal contribution of .17. Neither are excesively large so neither is exclusively driving our model, however their effects are clearly important more so than the other predictors included.

Summary

We started by finding a model through the forwards-backwards selection with an AIC criterion. Through this we landed on a model with square footage of the lot and of the living space, its grade, its quality of view, whether it was a waterfront view, and its condition. Through a nested F test we eliminated the number of floors and the number of bedrooms/bathrooms. Through ANOVA, residual plots, and diagnostics, we see that the regression is well-fitted and not unduly influenced by extreme X or Y variables.

Aside