Stats_101A_hw7_anna_piskun

Anna Piskun

2/20/2020

Part A:

Load the waistweightheight dataframe into R. The objective of this exercise is to understand why we need hypothesis tests to help us decide whether to add new variables and why we need adjusted R-squared.

```
wwh <- read.table("waistweightheight.txt", header = T, sep = "\t", fill = FALSE)
```

a) First, fit a model that predicts Weight using waist size and height.
   i) Find and report SYY, SSReg, and RSS
   ii) Report R-squared and adjusted R-squared
   iii) Just for fun, interpret the slope for height.
   iv) Does interpreting the slope for height using the phrase "as height increases…" make sense? How about if we replace height with waist size?

```
model1 <- lm(Weight~Waist + Height, data = wwh)
summary(model1)
```

```
##
## Call:
## lm(formula = Weight ~ Waist + Height, data = wwh)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.760  -6.405  -0.420   5.656  45.474
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -165.5332     8.2517  -20.06   <2e-16 ***
## Waist          4.9605     0.1229   40.37   <2e-16 ***
## Height         2.4884     0.1438   17.30   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.986 on 504 degrees of freedom
## Multiple R-squared:  0.8853, Adjusted R-squared:  0.8848
## F-statistic:  1945 on 2 and 504 DF,  p-value: < 2.2e-16
```

```
anova(model1)
```

```
## Analysis of Variance Table
##
## Response: Weight
##            Df Sum Sq Mean Sq F value    Pr(>F)
## Waist       1 358074  358074 3590.77 < 2.2e-16 ***
## Height      1  29843   29843  299.26 < 2.2e-16 ***
## Residuals 504  50259     100
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

SYY = 438176, SSReg = 387917, and RSS = 5025. R-squared = 0.8853, and adjusted R-squared = 0.8848. Interpretation of slope for height: On average, a one inch increase in height is associated with a 2.4884 pound increase in weight. Interpreting the slope for height using the phrase as height increases does not make sense because that would mean that with every inch you grew you would gain 2.4884 pounds which is not necessarily the case. Likewise there can be other confounding variables that affect weight, such as waist size and whenever we interpret the slope with multiple regression we must keep all variables in mind. Both height and waist size can have affects on weight so even replacing height with waist size in the above statement won't be accurate because it won't take into account all of the present variables.

b) We're now going to add a variable to our model that is useless.

c) Find and report SYY, SSReg, and RSS, and SSreg due to the variable worthless

ii) Comment on how these have changed from (a).
iii) How has R-squared and adjusted R-squared changed from (a)?

```r
set.seed(23)
new.df <- transform(wwh, worthless = rnorm(dim(wwh)[1],0,5))
#Add this variable to the model as the LAST variable
model2 <- lm(Weight~Waist + Height + worthless, data = new.df)
anova(model2)
```

```
## Analysis of Variance Table
##
## Response: Weight
##            Df Sum Sq Mean Sq   F value Pr(>F)
## Waist       1 358074  358074 3584.4800 <2e-16 ***
## Height      1  29843   29843  298.7400 <2e-16 ***
## worthless   1     12      12    0.1176 0.7318
## Residuals 503  50247     100
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
summary(model2)
```

```
##
## Call:
## lm(formula = Weight ~ Waist + Height + worthless, data = new.df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.981  -6.384  -0.350   5.800  45.435
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -165.54777    8.25903 -20.044   <2e-16 ***
## Waist          4.95999    0.12300  40.325   <2e-16 ***
## Height         2.48874    0.14397  17.286   <2e-16 ***
## worthless      0.02992    0.08724   0.343    0.732
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.995 on 503 degrees of freedom
## Multiple R-squared:  0.8853, Adjusted R-squared:  0.8846
## F-statistic:  1294 on 3 and 503 DF,  p-value: < 2.2e-16
```

SYY = 438176, SSReg = 388039, RSS = 50137, SSreg due to worthless = 122. SYY remained the same, while SSReg increased and RSS decreased. R-squared = 0.8856 and adjusted R-squared = 0.8849 so both increased slightly with R-squared increasing by 0.0003 and adjusted R-squared increasing by 0.0001.

c) Repeat (b) but this time put worthless in the model first. Comment on how the terms have changed from (b).

```
model3 <- lm(Weight~worthless + Waist + Height, data = new.df)
anova(model3)
```

```
## Analysis of Variance Table
##
## Response: Weight
##            Df Sum Sq Mean Sq   F value Pr(>F)
## worthless   1     58      58    0.5828 0.4456
## Waist       1 358020  358020 3583.9463 <2e-16 ***
## Height      1  29850   29850  298.8086 <2e-16 ***
## Residuals 503  50247     100
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(model3)
```

```
##
## Call:
## lm(formula = Weight ~ worthless + Waist + Height, data = new.df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.981  -6.384  -0.350   5.800  45.435
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -165.54777    8.25903 -20.044   <2e-16 ***
## worthless      0.02992    0.08724   0.343    0.732
## Waist          4.95999    0.12300  40.325   <2e-16 ***
## Height         2.48874    0.14397  17.286   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.995 on 503 degrees of freedom
## Multiple R-squared:  0.8853, Adjusted R-squared:  0.8846
## F-statistic:  1294 on 3 and 503 DF,  p-value: < 2.2e-16
```

SYY stayed the same (at 438176) as well as overall SSReg (388039) and RSS (50137), however individual SSRegs for the Waist and Height variables both decreased and the SSReg for the worthless variable increased.

d) Which do you think is a more reliable guide as to whether a new variable should be added, Rsquared or adjusted Rsquared? Why?

Whenever a new variable is added to a model R-squared will go up regardless. Since R-squared will always increase this can lead to potential overfitting (just adding random noise). Adjusted R-squared is a better measure of whether adding a new variable is an improvement since it compares estimated variability in the noise compared to the overall variability. Thus if adjusted R-squared goes down or stays the same, then the new variable wasn't signifcant, but if it goes up then it's probably useful.

e) Why can't we just look at SSreg to decide whether to add a new variable (following the rule if SSreg gets bigger, add the new variable)? Why do you think partial tests are useful for telling us whether we should add a new variable?

We can't just look at SSreg to decide whether to add a new variable or not because order matters and the amount of variability that is left to explain depends on how much was explained by variables already entered into the model. Partial tests allow us to see whether or not adding a new variable is useful because it allows us to test whether or not one variable is signifcant assuming that other variables are. This way we can control for each new variable that is added and determine on a case-by-case basis whether or not including it will improve our model.

Part B:

Again, let's work with the cars data set. (cars04). Fit a model to predict Suggested RetailPrice using all of the remaining numerical variables. In other words, exclue Vehicle.Name and Hybrid from your model. (This is the same model you fit for last week's homework.)

```
cars04 <- read.csv('cars04.csv')
cars04_clean <- dplyr::select(cars04, SuggestedRetailPrice:Width)
model1_cars <- lm(SuggestedRetailPrice~., data = cars04_clean)
summary(model1_cars)
```

```
##
## Call:
## lm(formula = SuggestedRetailPrice ~ ., data = cars04_clean)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1403.85  -276.86   -55.03   257.55  2584.11
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 349.97628 1461.40052   0.239 0.810953
## DealerCost    1.05418    0.00564 186.923  < 2e-16 ***
## EngineSize  -32.24720  123.05642  -0.262 0.793523
## Cylinders   228.32952   71.99492   3.171 0.001730 **
## Horsepower    2.36212    1.42851   1.654 0.099624 .
## CityMPG     -16.74239   21.46286  -0.780 0.436181
## HighwayMPG   46.75754   24.17910   1.934 0.054403 .
## Weight        0.69920    0.20751   3.370 0.000887 ***
## WheelBase    27.05345   16.36168   1.653 0.099644 .
## Length       -7.32019    7.12296  -1.028 0.305209
## Width       -84.70850   30.21238  -2.804 0.005496 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 532.3 on 223 degrees of freedom
## Multiple R-squared:  0.9989, Adjusted R-squared:  0.9989
## F-statistic: 2.073e+04 on 10 and 223 DF,  p-value: < 2.2e-16
```

```
anova(model1_cars)
```

```
## Analysis of Variance Table
##
## Response: SuggestedRetailPrice
##             Df     Sum Sq    Mean Sq    F value    Pr(>F)
## DealerCost   1 5.8714e+10 5.8714e+10 2.0724e+05 < 2.2e-16 ***
## EngineSize   1 7.7453e+06 7.7453e+06 2.7338e+01 3.925e-07 ***
## Cylinders    1 2.7222e+06 2.7222e+06 9.6084e+00  0.002186 **
## Horsepower   1 7.0394e+05 7.0394e+05 2.4847e+00  0.116377
## CityMPG      1 2.1856e+05 2.1856e+05 7.7150e-01  0.380714
## HighwayMPG   1 2.1052e+05 2.1052e+05 7.4310e-01  0.389601
## Weight       1 1.2563e+06 1.2563e+06 4.4344e+00  0.036341 *
## WheelBase    1 3.9621e+04 3.9621e+04 1.3990e-01  0.708785
## Length       1 1.6483e+06 1.6483e+06 5.8179e+00  0.016673 *
## Width        1 2.2271e+06 2.2271e+06 7.8611e+00  0.005496 **
## Residuals  223 6.3178e+07 2.8331e+05
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

a) write the equation of the fitted model

SuggestedRetailPrice = 349.97628 + 1.05418(DealerCost) - 32.24720(EngineSize) + 228.32952(Cylinders) + 2.36212(Horsepower) - 16.74239(CityMPG) + 46.75754(HighwayMPG) + 0.69920(Weight) + 27.05345(WheelBase) - 7.32019(Length) - 84.70850(Width)

b) Using the summary command, report the estimated slope, the t-statistic and the p-value for the Cylinders variable. What can we conclude from this t-statistic and p-value? (Assume that all necessary model conditions are valid.)

```r
summary(model1_cars)
```

```
##
## Call:
## lm(formula = SuggestedRetailPrice ~ ., data = cars04_clean)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1403.85  -276.86   -55.03   257.55  2584.11
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  349.97628 1461.40052   0.239 0.810953
## DealerCost     1.05418    0.00564 186.923  < 2e-16 ***
## EngineSize   -32.24720  123.05642  -0.262 0.793523
## Cylinders    228.32952   71.99492   3.171 0.001730 **
## Horsepower     2.36212    1.42851   1.654 0.099624 .
## CityMPG      -16.74239   21.46286  -0.780 0.436181
## HighwayMPG    46.75754   24.17910   1.934 0.054403 .
## Weight         0.69920    0.20751   3.370 0.000887 ***
## WheelBase     27.05345   16.36168   1.653 0.099644 .
## Length        -7.32019    7.12296  -1.028 0.305209
## Width        -84.70850   30.21238  -2.804 0.005496 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 532.3 on 223 degrees of freedom
## Multiple R-squared:  0.9989, Adjusted R-squared:  0.9989
## F-statistic: 2.073e+04 on 10 and 223 DF,  p-value: < 2.2e-16
```

For the Cylinders variable, the estimated slope = 228.32952, t-statistic = 3.171, and p-value = 0.001730. Since the p-value is less than 0.05, the variable Cylinders is statistically significant given that DealerCost and EngineSize are already included in the model. In other words, adding the variable Cylinders improves our model and allows us to explain more of the variation.

c) Show how to get the t-statistic value for Cylinders using the anova() command.

```
model2_cars <- lm(SuggestedRetailPrice~DealerCost + EngineSize + Horsepower + CityMPG + HighwayMPG + Wei
anova(model2_cars)
```

```
## Analysis of Variance Table
##
## Response: SuggestedRetailPrice
##              Df     Sum Sq    Mean Sq    F value    Pr(>F)    
## DealerCost    1 5.8714e+10 5.8714e+10 2.0724e+05 < 2.2e-16 ***
## EngineSize    1 7.7453e+06 7.7453e+06 2.7338e+01 3.925e-07 ***
## Horsepower    1 1.0860e+06 1.0860e+06 3.8331e+00  0.051496 .  
## CityMPG       1 1.9693e+05 1.9693e+05 6.9510e-01  0.405327    
## HighwayMPG    1 5.4432e+04 5.4432e+04 1.9210e-01  0.661576    
## Weight        1 1.3086e+06 1.3086e+06 4.6190e+00  0.032697 *  
## WheelBase     1 6.4650e+04 6.4650e+04 2.2820e-01  0.633335    
## Length        1 1.9825e+06 1.9825e+06 6.9977e+00  0.008742 ** 
## Width         1 1.4838e+06 1.4838e+06 5.2374e+00  0.023043 *  
## Cylinders     1 2.8496e+06 2.8496e+06 1.0058e+01  0.001730 ** 
## Residuals   223 6.3178e+07 2.8331e+05                           
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#square root the F-value for Cylinders from the anova table to find the t-statistic value for Cylinders
t.stat <- sqrt(1.0058e+01)
t.stat
```

```
## [1] 3.171435
```

d) Report and interpret the F-statistic from the summary() command.

Since the F-statistic is large and the p-value is small (less than 0.05), it's statistically significant and therefore the full model explains more variation in suggested retail price than the null model.

e) Carry out a test to determine whether the full model is better than a model that excludes both CityMPG and HighwayMPG. In otherwords, test the hypothesis that fuel consumption has no affect on the suggested retail price.

```
null.model <- lm(SuggestedRetailPrice~DealerCost + EngineSize + Cylinders + Horsepower + Weight + Wheel
anova(null.model)
```

```
## Analysis of Variance Table
##
## Response: SuggestedRetailPrice
##              Df     Sum Sq    Mean Sq    F value    Pr(>F)    
## DealerCost    1 5.8714e+10 5.8714e+10 2.0204e+05 < 2.2e-16 ***
## EngineSize    1 7.7453e+06 7.7453e+06 2.6651e+01 5.353e-07 ***
## Cylinders     1 2.7222e+06 2.7222e+06 9.3670e+00  0.002478 ** 
## Horsepower    1 7.0394e+05 7.0394e+05 2.4223e+00  0.121028    
## Weight        1 5.3446e+05 5.3446e+05 1.8391e+00  0.176418    
## WheelBase     1 7.3600e+02 7.3600e+02 2.5000e-03  0.959900    
## Length        1 1.4322e+06 1.4322e+06 4.9281e+00  0.027421 *  
## Width         1 1.4236e+06 1.4236e+06 4.8985e+00  0.027885 *  
## Residuals   225 6.5388e+07 2.9061e+05                           
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
full.model <- lm(SuggestedRetailPrice~., data = cars04_clean)
anova(full.model)
```

```
## Analysis of Variance Table
##
## Response: SuggestedRetailPrice
##              Df     Sum Sq    Mean Sq    F value    Pr(>F)
## DealerCost    1 5.8714e+10 5.8714e+10 2.0724e+05 < 2.2e-16 ***
## EngineSize    1 7.7453e+06 7.7453e+06 2.7338e+01 3.925e-07 ***
## Cylinders     1 2.7222e+06 2.7222e+06 9.6084e+00  0.002186 **
## Horsepower    1 7.0394e+05 7.0394e+05 2.4847e+00  0.116377
## CityMPG       1 2.1856e+05 2.1856e+05 7.7150e-01  0.380714
## HighwayMPG    1 2.1052e+05 2.1052e+05 7.4310e-01  0.389601
## Weight        1 1.2563e+06 1.2563e+06 4.4344e+00  0.036341 *
## WheelBase     1 3.9621e+04 3.9621e+04 1.3990e-01  0.708785
## Length        1 1.6483e+06 1.6483e+06 5.8179e+00  0.016673 *
## Width         1 2.2271e+06 2.2271e+06 7.8611e+00  0.005496 **
## Residuals   223 6.3178e+07 2.8331e+05
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
full.model2 <- lm(SuggestedRetailPrice~DealerCost + EngineSize + Cylinders + Horsepower + HighwayMPG + C
anova(full.model2)
```

```
## Analysis of Variance Table
##
## Response: SuggestedRetailPrice
##              Df     Sum Sq    Mean Sq    F value    Pr(>F)
## DealerCost    1 5.8714e+10 5.8714e+10 2.0724e+05 < 2.2e-16 ***
## EngineSize    1 7.7453e+06 7.7453e+06 2.7338e+01 3.925e-07 ***
## Cylinders     1 2.7222e+06 2.7222e+06 9.6084e+00  0.002186 **
## Horsepower    1 7.0394e+05 7.0394e+05 2.4847e+00  0.116377
## HighwayMPG    1 3.8004e+05 3.8004e+05 1.3414e+00  0.248019
## CityMPG       1 4.9040e+04 4.9040e+04 1.7310e-01  0.677774
## Weight        1 1.2563e+06 1.2563e+06 4.4344e+00  0.036341 *
## WheelBase     1 3.9621e+04 3.9621e+04 1.3990e-01  0.708785
## Length        1 1.6483e+06 1.6483e+06 5.8179e+00  0.016673 *
## Width         1 2.2271e+06 2.2271e+06 7.8611e+00  0.005496 **
## Residuals   223 6.3178e+07 2.8331e+05
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Preforming a partial F-test, we see that when accounting for DealerCost, EngineSize, Cylinders,and Horsepower in our model, CityMPG and HighwayMPG are not statistically significant (p-values greater than 0.05) and therefore fuel consumption may not have a significant effect on the suggested retail price.

Part C:

The file realty.txt combines house prices for four neighborhoods in Los Angeles. (You've seen subsets of this data set already). Upload this file. Transform as needed until you get 1555 observations and 10 variables.

```
realty <- read.table("realty.txt", header = T, sep = "\t", fill = FALSE)
table(realty$type)
```

```
##
##            Condo/Twh       Land     Mobile        SFR
##        39       654        24          8       951
```

```
new.frame <- subset(realty, type == "Condo/Twh"| type == "SFR")
new2.frame <- subset(new.frame, sqft>0 & bath>0)
realty.new <- transform(new2.frame, lprice=log(price))
```

i) Fit a model that predicts the log of price with city, bed, bath, and sqft. Assuming conditions of the model hold, interpret the intercept. (Be sure to specify data=realty.new in your lm command.)

```
model4 <- lm(lprice~city + bed + bath + sqft, data = realty.new)
summary(model4)
```

```
## 
## Call:
## lm(formula = lprice ~ city + bed + bath + sqft, data = realty.new)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -3.5421 -0.3024 -0.0145  0.2777  1.8701 
## 
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)    
## (Intercept)       1.327e+01  5.519e-02 240.444  < 2e-16 ***
## cityLong Beach   -1.226e+00  4.252e-02 -28.832  < 2e-16 ***
## citySanta Monica -3.118e-01  5.094e-02  -6.121 1.18e-09 ***
## cityWestwood     -6.161e-01  6.232e-02  -9.887  < 2e-16 ***
## bed               1.744e-01  1.632e-02  10.686  < 2e-16 ***
## bath              2.825e-02  1.788e-02   1.580    0.114    
## sqft              1.731e-04  1.433e-05  12.076  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.4726 on 1548 degrees of freedom
## Multiple R-squared:  0.7967, Adjusted R-squared:  0.7959 
## F-statistic:  1011 on 6 and 1548 DF,  p-value: < 2.2e-16
```

```
intercept <- exp(1.327e+01)
intercept
```

```
## [1] 579545.8
```

```
anova(model4)
```

```
## Analysis of Variance Table
## 
## Response: lprice
##             Df Sum Sq Mean Sq F value    Pr(>F)    
## city         3 973.47  324.49 1453.06 < 2.2e-16 ***
## bed          1 304.71  304.71 1364.49 < 2.2e-16 ***
## bath         1  43.56   43.56  195.04 < 2.2e-16 ***
## sqft         1  32.56   32.56  145.82 < 2.2e-16 ***
## Residuals 1548 345.69    0.22                      
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The intercept tells us that the base price of a home in Los Angeles is $579,545.80 (in other words this is the cost for just the land).

ii) Interpret cityWestwood. (Hint: what cities are in the dataset?) Which city is most expensive, on average? Which least?

cityWestwood is the variable that looks at prices for homes in Westwood, Los Angeles. Since its p-value is less than 0.05, the variable is statistically significant and should be included in our model. On average, Westwood is the most expensive and Long Beach is the least expensive.

iii) Are more bedrooms more valuable? Interpret the meaning of the bed variable.

Since bedrooms have a very large F-value but small p-value the variable is statistically significant and therefore should be included in our model. When already accounting for the city varaible, from looking at the F-values we see that bedrooms are more valuable than bath and sqft.

iv) The p-value for bath is high. What does this mean?

Since the p-value = 0.114, the bath variable is not statistically significant when taking city and bedrooms into account, and therefore it is not necessary to add to our model to improve its ability to explain variation in the data.

v) Fit the model again without the variable “bed”. Why is “bath” now significant? (hint: try the update() command.)

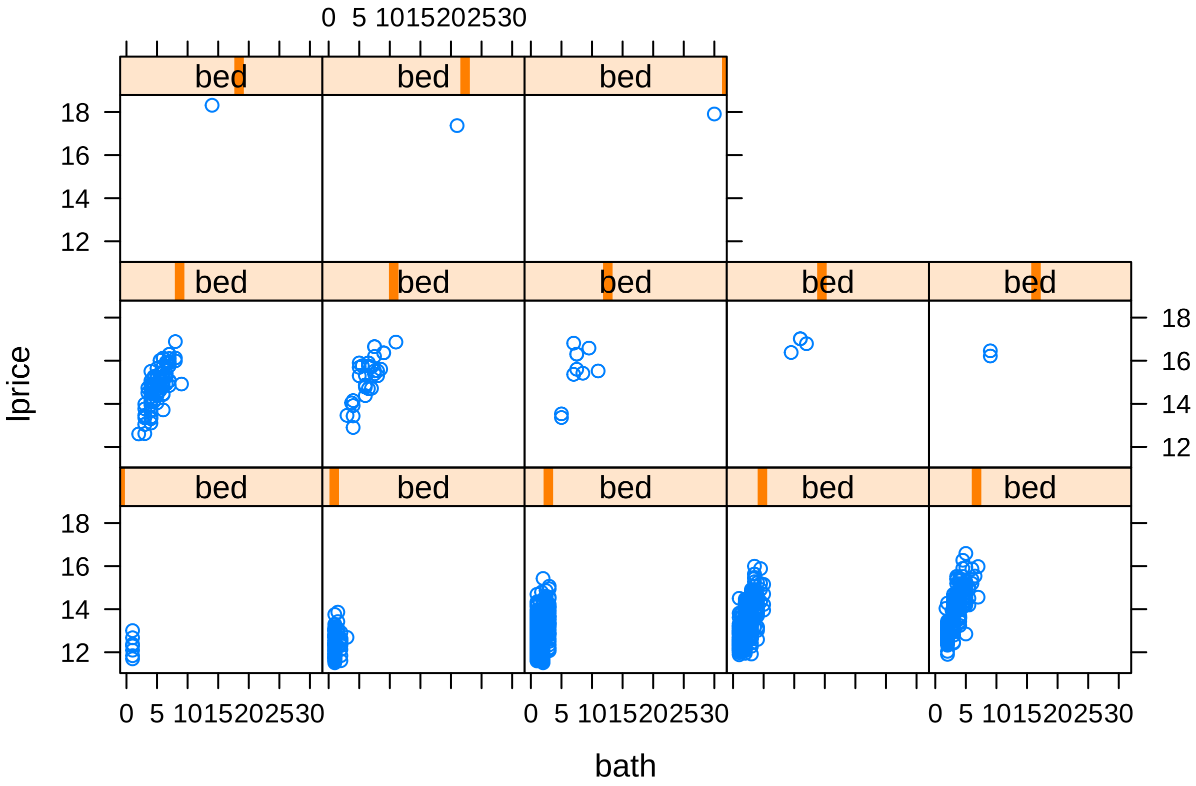
```
model5 <- lm(lprice~city + bath + sqft, data = realty.new)
summary(model5)
```

```
##
## Call:
## lm(formula = lprice ~ city + bath + sqft, data = realty.new)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8757 -0.3086 -0.0177  0.3070  1.8754
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      13.5055934  0.0524483 257.503  < 2e-16 ***
## cityLong Beach   -1.2087082  0.0440188 -27.459  < 2e-16 ***
## citySanta Monica -0.3574888  0.0525854  -6.798 1.51e-11 ***
## cityWestwood     -0.6685917  0.0643523 -10.390  < 2e-16 ***
## bath              0.1067374  0.0168902   6.319 3.42e-10 ***
## sqft              0.0002012  0.0000146  13.781  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4895 on 1549 degrees of freedom
## Multiple R-squared:  0.7816, Adjusted R-squared:  0.7809
## F-statistic:  1109 on 5 and 1549 DF,  p-value: < 2.2e-16
```

Bath is now significant because we are no longer accounting for bedrooms in our model. Since bedrooms often times serve as a way to quantify and add value to a home, by excluding this information the number of bathrooms become more essential to our model. Normally the more bathrooms there are the nicer and therefore more expensive the home is. Bedrooms serve as a better predictor which is why when they were included in our model the bathroom variable was not significant, however, without it the bathrooms variable becomes more valuable in explaining the variation in price and therefore statistically significant.

vi) Make a lattice plot of log(price) against bathrooms, controlling for the number of bedrooms and write a sentence or two interpreting the plot. What does this plot tell us about the need for including both bed and bath in the same model? (Hint: See below)

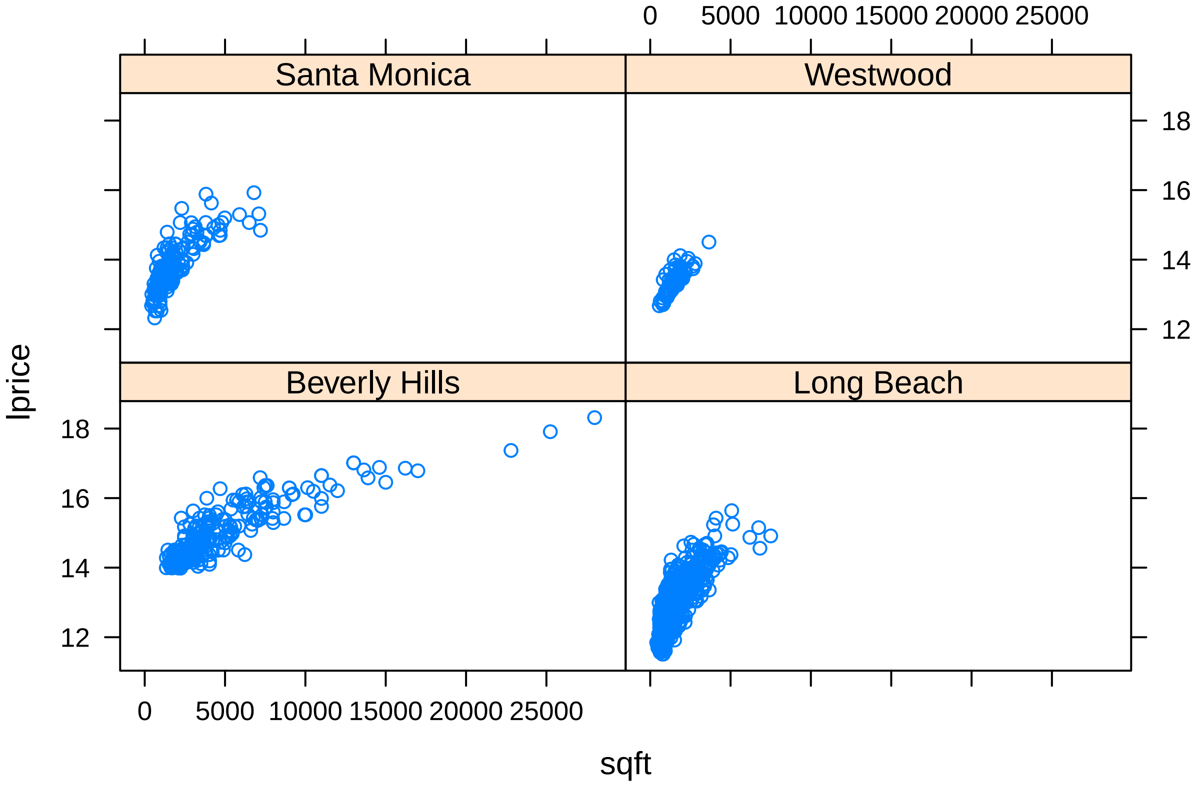
```
library(lattice)
xyplot(lprice~bath| bed, realty.new)
```



Looking at the lattice plot, we are trying to see if there is any variability in bath while controlling for the number of beds. Since all of the plots follow a vertical line pattern we can conclude that there isn't much variability and so the two variables (bed and bath) must explain a similar amount of the variation. Therefore, we do not need to include both in our model.

vii) The model we've fit so far assumes that the relation between log(price) and size (measured by sqft) is the same in each city. Does this seem like a valid assumption? To check make and interpret the lattice plot:

```
library(lattice)
xyplot(lprice~sqft| city, realty.new)
```



No, this is not a valid assumption since by looking at the plots we see that the variability is different for the different cities. Beverly Hills has a much more spread out scattered plot showing more variation than when compared to Westwood. While Santa Monica and Long Beach have similar looking plots, we cannot assume that all four cities have the same relation between log(price) and size.