Applied Regression and Time Series Analysis (2016 Fall): HW2 - Week 4

Jeffrey Yau

September 17, 2016

- **Due: 10/9/2016 (11:59pm PST)**

Overview:

Use the dataset, housePrice.Rdata, a very simple dataset, for this exercise. It contains only few variables. Consider the following scenario. You work in the strategic data science team supporting the executive management team of your company. This team has the mandate to provide data- and analytic-drive recommendations to guide corporate strategies and decisions. The company's management team is considering buying residential properties in the areas near the corporate headquarter to accomodate its remote employees traveling to the headquarter for long-term (i.e. more than 4 weeks) projects. Specifically, the management want to understand the local housing market and how selected characteristics of a house affects its price. Your job in this assignment is to build different linear regression models to answer various questions (to be specified below) asked by the management.

Description of the Data:

The file **birthweight_w271.Rdata** contains data from the *1988 National Health Interview Survey*, which is modified by the instructor. This survey is conducted by the U.S. Census Bureau and has collected data on individual health metrics since 1957. Like all surveys, a full analysis would require advanced techniques such as those provided by the R survey package. For this homework, however, you are to treat the data as a true random sample. You will use this dataset to practice interpreting OLS coefficients.

1. Question 1 - The Usual (Part I):

a. Load the data housePrie.Rdata

```
load('housePrice.rdata')
```

b. Examine the structure of the data

```
sapply(data,class)
```

```
##     price    assess     bdrms   lotsize     sqrft  colonial    lprice
## "numeric" "numeric" "integer" "numeric" "integer" "integer" "numeric"
##   lassess  llotsize    lsqrft
## "numeric" "numeric" "numeric"
```

c. Provide descriptive statistics of the data

```
describe(data)
```

```
## data
##
##  10  Variables      88  Observations
## ---------------------------------------------------------------------------
## price
##       n missing  unique    Info    Mean     .05     .10     .25     .50
##      88       0      71       1   293.5   192.4   209.0   230.0   265.5
##     .75     .90     .95
##   326.2   408.7   475.3
##
## lowest : 111.0 150.0 180.0 190.0 191.0
## highest: 477.5 495.0 575.0 713.5 725.0
## ---------------------------------------------------------------------------
## assess
##       n missing  unique    Info    Mean     .05     .10     .25     .50
##      88       0      88       1   315.7   214.3   228.8   253.9   290.2
##     .75     .90     .95
##   352.1   441.6   503.5
##
## lowest : 198.7 202.4 208.0 212.1 212.5
## highest: 515.1 518.1 543.6 655.4 708.6
## ---------------------------------------------------------------------------
## bdrms
##       n missing  unique    Info    Mean
##      88       0       6    0.84   3.568
##
##           2  3  4 5 6 7
## Frequency 4 42 33 7 1 1
## %         5 48 38 8 1 1
## ---------------------------------------------------------------------------
## lotsize
##       n missing  unique    Info    Mean     .05     .10     .25     .50
##      88       0      84       1    9020    4145    5087    5733    6430
##     .75     .90     .95
##    8583   15092   17787
##
## lowest :  1000  2892  3500  3597  4054
## highest: 18838 20700 28231 31000 92681
## ---------------------------------------------------------------------------
## sqrft
##       n missing  unique    Info    Mean     .05     .10     .25     .50
##      88       0      85       1    2014    1380    1446    1660    1845
##     .75     .90     .95
##    2227    2751    3360
##
## lowest : 1171 1185 1294 1374 1376, highest: 3375 3529 3662 3733 3880
## ---------------------------------------------------------------------------
## colonial
##       n missing  unique    Info     Sum    Mean
```

```
##      88       0       2    0.64      61  0.6932
## ------------------------------------------------------------------------
## lprice
##       n missing  unique    Info    Mean     .05     .10     .25     .50
##      88       0      71       1   5.633   5.260   5.342   5.438   5.582
##     .75     .90     .95
##   5.788   6.013   6.164
##
## lowest : 4.710 5.011 5.193 5.247 5.252
## highest: 6.169 6.205 6.354 6.570 6.586
## ------------------------------------------------------------------------
## lassess
##       n missing  unique    Info    Mean     .05     .10     .25     .50
##      88       0      88       1   5.718   5.367   5.433   5.537   5.671
##     .75     .90     .95
##   5.864   6.090   6.221
##
## lowest : 5.292 5.310 5.338 5.357 5.359
## highest: 6.244 6.250 6.298 6.485 6.563
## ------------------------------------------------------------------------
## llotsize
##       n missing  unique    Info    Mean     .05     .10     .25     .50
##      88       0      84       1   8.905   8.329   8.534   8.654   8.769
##     .75     .90     .95
##   9.058   9.622   9.783
##
## lowest :  6.908  7.970  8.161  8.188  8.307
## highest:  9.844  9.938 10.248 10.342 11.437
## ------------------------------------------------------------------------
## lsqrft
##       n missing  unique    Info    Mean     .05     .10     .25     .50
##      88       0      85       1   7.573   7.229   7.276   7.415   7.520
##     .75     .90     .95
##   7.708   7.920   8.120
##
## lowest : 7.066 7.077 7.165 7.225 7.227
## highest: 8.124 8.169 8.206 8.225 8.264
## ------------------------------------------------------------------------
```

d. Identify if there are unreasonable values, top-coding, and bottom-coding. If any of these is found, propose your strategy to handle them.

None of the variable values seem unreasonable, top coded, or bottom coded. There's a suspicious outlier though in lotsize (with 92681 lotsize):

```
data[data$lotsize == 92681,]
```

```
##    price assess bdrms lotsize sqrft colonial   lprice  lassess llotsize
## 77   318  295.2     4   92681  1696        1 5.762052 5.687653 11.43692
##      lsqrft
## 77 7.436028
```

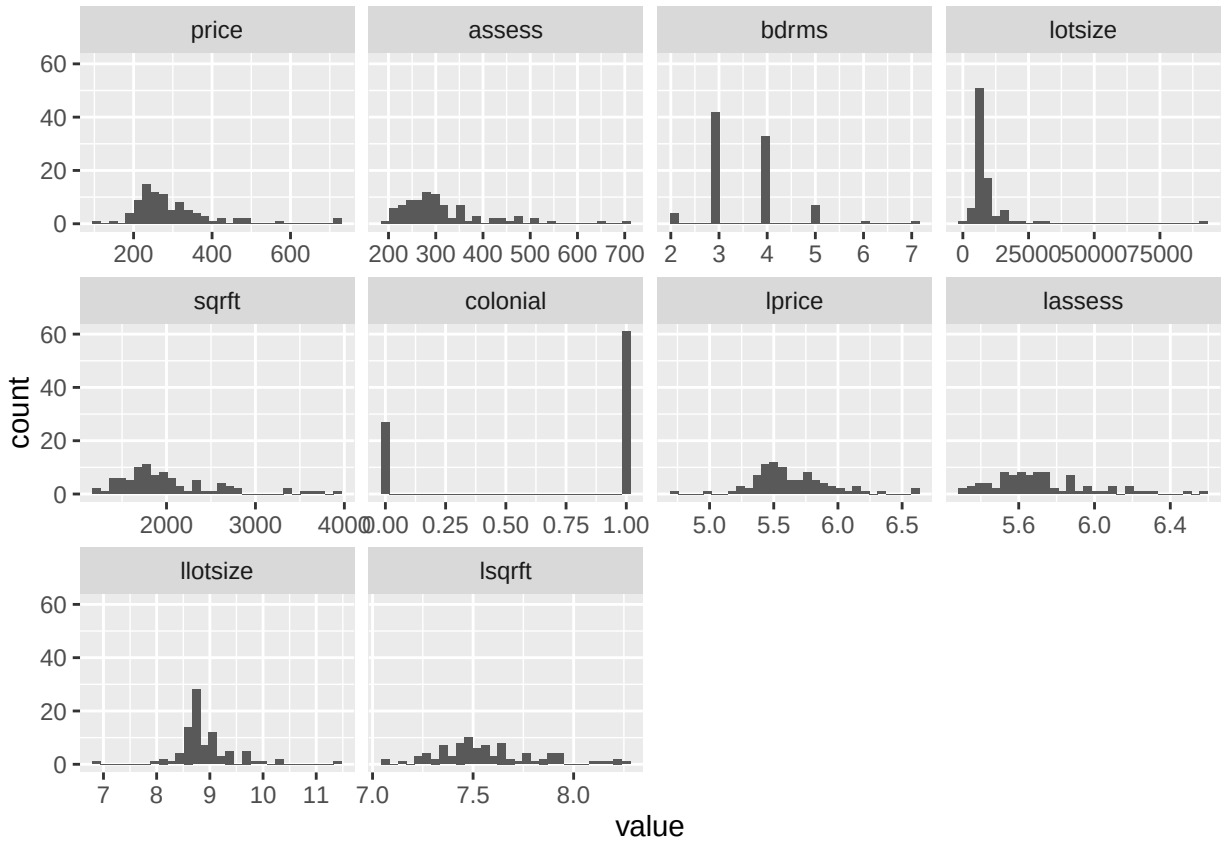
While this value does not it in with the rest of the data, it's still plausible. I'm going to leave it as is.

2. Question 2 - The Usual (Part II):

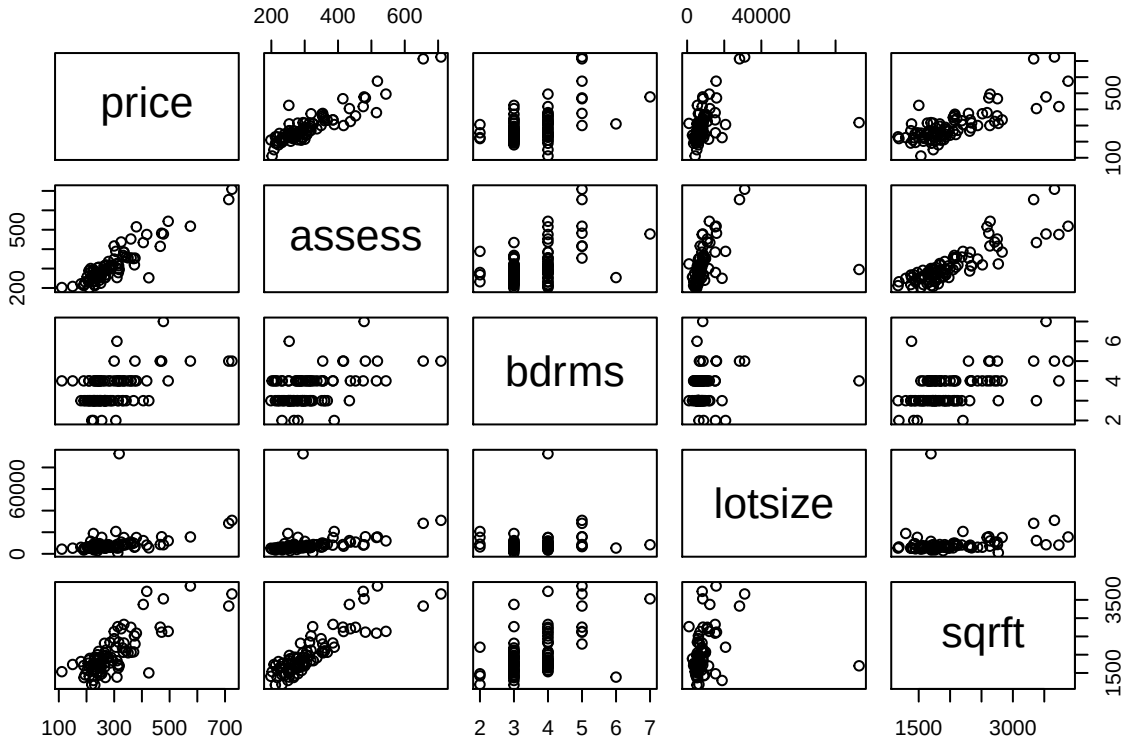
- **Conduct EDA, including both univariate and multivariate analyses, on this dataset.**

```
EDA = function(data){
  print("Summary")
print (summary(data))

library(reshape2)
library(ggplot2)
d <- melt(data)
print("")
print("Histograms of each variable")
 print(ggplot(d,aes(x = value)) +
    facet_wrap(~variable,scales = "free_x") +
    geom_histogram())

print("")
print("Plots of data (only metric columns)")
plot(data[0:5])

}
EDA(data)
```

```
## [1] "Summary"
##      price            assess           bdrms          lotsize
##  Min.   :111.0   Min.   :198.7   Min.   :2.000   Min.   : 1000
##  1st Qu.:230.0   1st Qu.:253.9   1st Qu.:3.000   1st Qu.: 5733
##  Median :265.5   Median :290.2   Median :3.000   Median : 6430
##  Mean   :293.5   Mean   :315.7   Mean   :3.568   Mean   : 9020
##  3rd Qu.:326.2   3rd Qu.:352.1   3rd Qu.:4.000   3rd Qu.: 8583
##  Max.   :725.0   Max.   :708.6   Max.   :7.000   Max.   :92681
##      sqrft          colonial          lprice          lassess
##  Min.   :1171   Min.   :0.0000   Min.   :4.710   Min.   :5.292
##  1st Qu.:1660   1st Qu.:0.0000   1st Qu.:5.438   1st Qu.:5.537
##  Median :1845   Median :1.0000   Median :5.582   Median :5.671
##  Mean   :2014   Mean   :0.6932   Mean   :5.633   Mean   :5.718
##  3rd Qu.:2227   3rd Qu.:1.0000   3rd Qu.:5.788   3rd Qu.:5.864
##  Max.   :3880   Max.   :1.0000   Max.   :6.586   Max.   :6.563
##     llotsize          lsqrft
##  Min.   : 6.908   Min.   :7.066
##  1st Qu.: 8.654   1st Qu.:7.415
##  Median : 8.769   Median :7.520
##  Mean   : 8.905   Mean   :7.573
##  3rd Qu.: 9.058   3rd Qu.:7.708
##  Max.   :11.437   Max.   :8.264

## No id variables; using all as measure variables

## [1] ""
## [1] "Histograms of each variable"

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
## [1] ""
## [1] "Plots of data (only metric columns)"
```



- **For each of the variables, discuss potential feature engineering (i.e. transformation of original variables, creation of a categorical variable from a numeric variable, creation of interaction among variables, etc) you may perform in later regression model building steps.**
  - Since *price* and *assess* are highly correlated, we could respecify the model to only capture *price* and the difference between *price* and *assess*.
- *bdrms* is technically a numeric variable, but since there aren't many unique values it takes on, we may want to consider treating it as ordinal and transforming it as such.
- *lotsize* and *sqrtft* are also pretty well correlated with *price* and *assess*, so it may be helpful to regress *price* on each and then replace each variable in the original regression with its residual.
- *colonial* is fine the way it is.

3. Build a regression model using price as the response variable and sqrft, bdrms, lotsize as explanatory variables. That is, a linear regression model of the following specification:

$$price = \beta_0 + \beta_1 sqrft + \beta_2 bdrms + \beta_3 lotsize + \epsilon$$

- **Interpret the coefficient estimates**

```
model = lm(price ~ sqrft + bdrms + lotsize,data)
summary(model)
```

```
##
## Call:
## lm(formula = price ~ sqrft + bdrms + lotsize, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -120.026  -38.530   -6.555   32.323  209.376
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.177e+01  2.948e+01  -0.739  0.46221
## sqrft        1.228e-01  1.324e-02   9.275 1.66e-14 ***
## bdrms        1.385e+01  9.010e+00   1.537  0.12795
## lotsize      2.068e-03  6.421e-04   3.220  0.00182 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 59.83 on 84 degrees of freedom
## Multiple R-squared:  0.6724, Adjusted R-squared:  0.6607
## F-statistic: 57.46 on 3 and 84 DF,  p-value: < 2.2e-16
```

- The coefficient on *sqrft* means that you'd expect two otherwise equal houses to be priced $120 more per unit difference in square feet
- The coefficient on *bdrms* means that you'd expect two otherwise equal houses to be priced $13000 more per difference in number of bedrooms
- The coefficient on *lotsize* means that you'd expect two otherwise equal houses to be priced $2 more per unit difference in lot size square feet
- **Respecify the model to use log(price) as the dependent variable. Interpret the coefficient estimate associated with the variable bdrms.**

```
model = lm(lprice ~ sqrft + bdrms + lotsize,data)
summary(model)
```

```
##
## Call:
## lm(formula = lprice ~ sqrft + bdrms + lotsize, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.73389 -0.10792 -0.01595  0.11181  0.63914
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.759e+00  9.354e-02  50.883  < 2e-16 ***
## sqrft       3.641e-04  4.201e-05   8.668 2.77e-13 ***
## bdrms       2.524e-02  2.859e-02   0.883  0.37992
## lotsize     5.602e-06  2.038e-06   2.749  0.00732 **
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1899 on 84 degrees of freedom
## Multiple R-squared:  0.6223, Adjusted R-squared:  0.6088
## F-statistic: 46.13 on 3 and 84 DF,  p-value: < 2.2e-16
```

- The coefficient on *sqrft* means that you'd expect two otherwise equal houses to be priced .036% more per unit difference in square feet
- The coefficient on *bdrms* means that you'd expect two otherwise equal houses to be priced 2.5% more per difference in number of bedrooms
- The coefficient on *lotsize* means that you'd expect two otherwise equal houses to be priced .00056% more per unit difference in lot size square feet

For all of the questions below, use log(price) as the dependent variable.

4. The management suspects that colonial-style properties (variable colonial = 1) have higher prices. Respecify the regression above and re-estimate the regression model to address this particular question raised by the management. Intrepret the coefficient(s) of interest.

```
model = lm(lprice~colonial,data)
summary(model)
```

```
##
## Call:
## lm(formula = lprice ~ colonial, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.95990 -0.16410 -0.07821  0.16930  1.03489
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.55128    0.05780  96.050   <2e-16 ***
## colonial     0.11815    0.06942   1.702   0.0924 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3003 on 86 degrees of freedom
## Multiple R-squared:  0.03259,    Adjusted R-squared:  0.02134
## F-statistic: 2.897 on 1 and 86 DF,  p-value: 0.09237
```

Management is correct. Colonial houses are on average about 12% higher priced than non-colonial houses

5. The management suspects that the effect of the number of bedrooms on price is nonlinear. Respecify the regression above and re-estimate the regression model to address this particular question raised by the management. Note that there are a few ways to capture nonlinear effect. You are asked to experiment to at least 2 approaches to capture the nonlinear effect. Note also that this question is slightly open-ended. So, please explain your approach and the results clearly.

Since we are trying to predict log of price, we aren't in the the position to observe a linear relationship between price and number of bedrooms. However, we can check if there is a linear relationship between log of price and number of bedrooms.

First approach: Transform *bdrms* to ordinal indicator variables and compare new model to old model

```
model1 = lm(lprice ~ sqrft + bdrms + lotsize,data)
summary(model1)
```

```
##
## Call:
## lm(formula = lprice ~ sqrft + bdrms + lotsize, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.73389 -0.10792 -0.01595  0.11181  0.63914
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.759e+00  9.354e-02  50.883  < 2e-16 ***
## sqrft       3.641e-04  4.201e-05   8.668 2.77e-13 ***
## bdrms       2.524e-02  2.859e-02   0.883  0.37992
## lotsize     5.602e-06  2.038e-06   2.749  0.00732 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1899 on 84 degrees of freedom
## Multiple R-squared:  0.6223, Adjusted R-squared:  0.6088
## F-statistic: 46.13 on 3 and 84 DF,  p-value: < 2.2e-16
```

The p value for the *bdrms* coefficient is not significant, so that implies there is no linear relationship.

Next lets transform bdrms to an ordinal variable and compare the two models:

```
bdrms2 = data$bdrms > 2
bdrms3 = data$bdrms > 3
bdrms4 = data$bdrms > 4
bdrms5 = data$bdrms > 5
bdrms6 = data$bdrms > 6
model2 = lm(lprice ~ sqrft + bdrms + bdrms2 + bdrms3 + bdrms4 + bdrms5 + bdrms6  + lotsize,data)
summary(model2)
```

```
##
## Call:
## lm(formula = lprice ~ sqrft + bdrms + bdrms2 + bdrms3 + bdrms4 +
##     bdrms5 + bdrms6 + lotsize, data = data)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.68019 -0.09330 -0.00215  0.08694  0.61934
##
## Coefficients: (1 not defined because of singularities)
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.578e+00  5.386e-01  10.356  < 2e-16 ***
## sqrft        3.536e-04  4.512e-05   7.837 1.69e-11 ***
## bdrms       -3.418e-01  2.755e-01  -1.241   0.2183
## bdrms2TRUE   3.203e-01  2.884e-01   1.111   0.2701
## bdrms3TRUE   2.935e-01  2.748e-01   1.068   0.2886
## bdrms4TRUE   5.691e-01  2.746e-01   2.072   0.0414 *
## bdrms5TRUE   5.087e-01  4.477e-01   1.136   0.2592
## bdrms6TRUE          NA         NA      NA       NA
## lotsize      5.276e-06  2.002e-06   2.635   0.0101 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1825 on 80 degrees of freedom
## Multiple R-squared:  0.6678, Adjusted R-squared:  0.6387
## F-statistic: 22.97 on 7 and 80 DF,  p-value: < 2.2e-16
```

```
anova(model1,model2)
```

```
## Analysis of Variance Table
##
## Model 1: lprice ~ sqrft + bdrms + lotsize
## Model 2: lprice ~ sqrft + bdrms + bdrms2 + bdrms3 + bdrms4 + bdrms5 +
##     bdrms6 + lotsize
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     84 3.0284
## 2     80 2.6637  4   0.36474 2.7386 0.03431 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since the p value is significant, the model with the ordinal variables is a better fit than the model without the ordinal variables, which implies again that there is a nonlinear effect between number of bedrooms and log of price.

6. The management suspects that the effect of the number of bedrooms on price depends the size of house in square feet (i.e. sqrft). Respecify the regression above and re-estimate the regression model to address this particular question raised by the management. Does management's "intuition" about the price effect of the number of bedrooms and house size correct? Please explain your answer.

To test this, let's create an interaction term for *bdrms* and *sqrft*:

```
bdrmssqrft = data$bdrms * data$sqrft
model2  = lm(lprice ~ sqrft + bdrms + lotsize + bdrmssqrft,data)
summary(model2)
```

```
##
## Call:
## lm(formula = lprice ~ sqrft + bdrms + lotsize + bdrmssqrft, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.72573 -0.10288 -0.01039  0.10416  0.62715
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.015e+00  2.853e-01  17.579  < 2e-16 ***
## sqrft        2.425e-04  1.348e-04   1.799  0.07570 .
## bdrms       -3.977e-02  7.422e-02  -0.536  0.59353
## lotsize      5.499e-06  2.042e-06   2.693  0.00856 **
## bdrmssqrft   2.983e-05  3.142e-05   0.949  0.34526
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.19 on 83 degrees of freedom
## Multiple R-squared:  0.6263, Adjusted R-squared:  0.6083
## F-statistic: 34.78 on 4 and 83 DF,  p-value: < 2.2e-16
```

```
anova(model1,model2)
```

```
## Analysis of Variance Table
##
## Model 1: lprice ~ sqrft + bdrms + lotsize
## Model 2: lprice ~ sqrft + bdrms + lotsize + bdrmssqrft
##   Res.Df    RSS Df Sum of Sq     F Pr(>F)
## 1     84 3.0284
## 2     83 2.9959  1  0.032523 0.901 0.3453
```

So the coefficient on the interaction term is not significant, and the model with the interaction term is not significantly better than the original model. Therefore, there is no evidence to support the intuition that *sqrft* affects the coefficient of *bdrms*