

# W271 Lab 2

Brandon Shurick, Alejandro J. Rojas, Olivier Zimmer

July 14 2016

## 1 Question 1 – Saratoga

**1a.** *Begin with a thorough exploratory data analysis. For each item presented, provide a discussion of any observations and insights you find.*

We begin by loading the data and then gathering the summary statistics (shown in figure 1). We observe a single NA value in the Acres variable. We will omit this single record in further analysis. From the summary we also see that the median house in this dataset has a price of \$153,000, living area of 1680 sq ft, 2 baths, 3 bedrooms, with a fireplace, 0.39 acres, and is 18 years old.

| Price          | Living.Area  | Baths         | Bedrooms    | Fireplace | Acres          | Age            |
|----------------|--------------|---------------|-------------|-----------|----------------|----------------|
| Min. : 16858   | Min. : 672   | Min. :1.000   | Min. :1.0   | No :428   | Min. :0.0000   | Min. : 0.00    |
| 1st Qu.:112579 | 1st Qu.:1344 | 1st Qu.:1.500 | 1st Qu.:3.0 | Yes:635   | 1st Qu.:0.2100 | 1st Qu.: 6.00  |
| Median :152786 | Median :1680 | Median :2.000 | Median :3.0 |           | Median :0.3900 | Median : 18.00 |
| Mean :170069   | Mean :1833   | Mean :1.937   | Mean :3.2   |           | Mean :0.5754   | Mean : 28.25   |
| 3rd Qu.:207128 | 3rd Qu.:2242 | 3rd Qu.:2.500 | 3rd Qu.:4.0 |           | 3rd Qu.:0.6100 | 3rd Qu.: 34.00 |
| Max. :882341   | Max. :5632   | Max. :5.500   | Max. :7.0   |           | Max. :9.0000   | Max. :247.00   |
|                |              |               |             |           | NA's :1        |                |

Figure 1: Saratoga Dataset Summary

Secondly, we create a scatterplot matrix of the dataset (shown in figure 2). From the scatterplot we see that Price as well as the Living.Area and Age have right-skewed distributions. Price and Living.Area seem to have strong correlation. Price also seems to have negative correlation with Age, and positive correlation with all other variables.

Next, we look at a correlation matrix of all variables in the dataset (shown in figure 3). From the correlation matrix, we see a strong correlation between living area and price (0.77) and baths (0.67). Baths and living area may have collinearity, as they have strong correlation together (0.74). Bedrooms also has a strong correlation with living area (0.67). Age is the only variable in the dataset that has a negative correlation with price (-0.26).

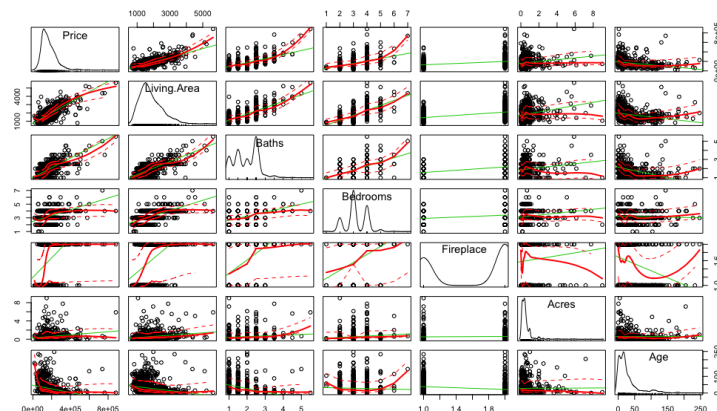


Figure 2: Saratoga Scatterplot Matrix

|               | Price      | Living.Area | Baths      | Bedrooms    | Acres       | Age          | has_fireplace |
|---------------|------------|-------------|------------|-------------|-------------|--------------|---------------|
| Price         | 1.0000000  | 0.7709239   | 0.6693772  | 0.46951869  | 0.179137672 | -0.261745659 | 0.40896547    |
| Living.Area   | 0.7709239  | 1.0000000   | 0.7436379  | 0.66815719  | 0.223243183 | -0.231709758 | 0.47324854    |
| Baths         | 0.6693772  | 0.7436379   | 1.0000000  | 0.51389903  | 0.130310668 | -0.401041501 | 0.44526513    |
| Bedrooms      | 0.4695187  | 0.6681572   | 0.5138990  | 1.0000000   | 0.145251648 | -0.039232377 | 0.30304337    |
| Acres         | 0.1791377  | 0.2232432   | 0.1303107  | 0.14525165  | 1.0000000   | 0.005947941  | 0.05771873    |
| Age           | -0.2617457 | -0.2317098  | -0.4010415 | -0.03923238 | 0.005947941 | 1.0000000    | -0.24296807   |
| has_fireplace | 0.4089655  | 0.4732485   | 0.4452651  | 0.30304337  | 0.057718729 | -0.242968072 | 1.0000000     |

Figure 3: Saratoga Correlation Matrix

**2a.** Fit a model that uses size to predict price, denote this as model #1. Is there evidence the line does not pass through the origin? Answer this question using a confidence interval.

Model #1 summary is shown in figure 4. There is no clear evidence that the intercept does not pass through zero. The model intercept is estimated as -3,086, with a 95% confidence interval between -19,298 and 13,125. Since this confidence interval crosses zero, it is not strong evidence that the actual population intercept does not cross zero.

**2b.** If the line passes through the origin, then the slope is a proxy for the price per square foot. Is there evidence the price per square foot is less than \$100 per square foot? Answer this question using a hypothesis test.

In this test, the hypothesis' can be stated as

$$H_0 : B_1 - 100 = 0$$

$$H_1 : B_1 - 100 < 0$$

```
lm(formula = saratoga$Price ~ saratoga$Living.Area)

Residuals:
    Min       1Q   Median       3Q      Max
-281329  -25370   -4390   17844  403379

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   -3116.957    4690.573   -0.665    0.507
saratoga$Living.Area    94.457      2.395   39.445 <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 53820 on 1061 degrees of freedom
Multiple R-squared:  0.5946,    Adjusted R-squared:  0.5942
F-statistic: 1556 on 1 and 1061 DF,  p-value: < 2.2e-16
```

Figure 4: Model #1 Summary

The t-statistic is estimated by subtracting 100 from the estimated coefficient and dividing by the standard error,

$$t = (B_1 - 100)/se(B_1)$$

. The resulting t-statistic is -2.32, which equates to a p-value of 0.01. This is evidence that the price per square foot is less than \$100.

**2c.** *Is there evidence the residuals do not have a Normal distribution? Answer this question with the appropriate visualization and hypothesis test.*

Based on the QQ plot (shown in figure 5), it is apparent that the residuals are not normally distributed. A Shapiro-Wilk test of model #1 results in a p-value  $< 2.2e-16$ , which is enough evidence to reject the null hypothesis that the residuals are normally distributed. However, since  $n > 30$  we can rely on the OLS asymptotic assumptions.

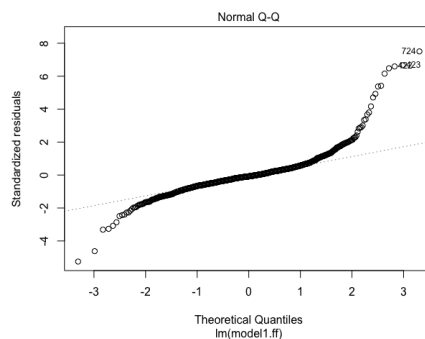


Figure 5: QQ Plot of Model #1

**2d.** *Is there evidence the fireplace variable is needed in the model? Answer this question with the appropriate visualization and numerical statistics. If you find that the fireplace variable is needed in the model, what condition is violated for model #1?*

To answer this question, we construct an alternative model that includes the fireplace variable, called model #1F. We conduct an F-test that compares the restricted model (model #1) to the unrestricted model (model #1F) using the formula:

$$F = \frac{(SSR_r - SSR_{ur})/q}{(SSR_{ur}/df_{ur})}$$

The resulting F-statistic is 6.59, which is significant, with p-value of 0.01.

**3a.** *Fit a model that uses the fireplace variable to predict price, denote this as model #2. What is the baseline or reference group?*

Summary of model #2 is included in figure 6. The baseline group in this model is the group of houses with no fireplace.

**3b.** *Is there evidence the change in the average price is not zero dollars when changing from homes without a fireplace to homes with a fireplace? Answer this question using a hypothesis test.*

```
lm(formula = saratoga$Price ~ saratoga$fire)

Residuals:
    Min       1Q   Median       3Q      Max
-150825  -44082  -14159   28573  683886

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  127955      3727    34.33  <2e-16 ***
saratoga$fire   70500      4822    14.62  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 77110 on 1061 degrees of freedom
Multiple R-squared:  0.1677,    Adjusted R-squared:  0.1669
F-statistic: 213.7 on 1 and 1061 DF,  p-value: < 2.2e-16
```

Figure 6: Model #2 Summary

Constructing a model #2 that uses having a fireplace as only predictor leads to a conclusion that having a fireplace adds \$70k to the price of the property. However, we know that this coefficient is inflated as the fact of having a fireplace is highly correlated to properties that are bigger in size. In any case, we can conclude that we have a reason to believe that moving to a place with fireplace will lead to higher value because places with fireplace are usually bigger in size. If Living Area remains constant then the evidence that having a fireplace will increase our property value diminishes as its Wald test exhibits a p-value that is significant only for 0.01 as we discussed in the prior question.

**3c.** Refer to the previous part. What statistical procedure is the hypothesis test equivalent to? Specify the corresponding competing hypotheses.

This is equivalent to a t-test of the coefficient for fireplace,  $\beta_f$ , where  $H_0 : \beta_f = 0$  and  $H_1 : \beta_f \neq 0$ .

**4a.** Fit a model that uses all of the numeric variables to predict the price, denote this as model #3. Is there evidence of collinear predictors? Answer this question with the appropriate visualization and numerical statistics.

We measure the Variance Inflation Factor for each coefficient (shown in figure 7), and find that Living.Area, Baths, and Bedrooms are moderately correlated; however, none of the VIFs are near 5, which is generally considered to be evidence of highly collinear predictors.

| Living.Area | Baths    | Bedrooms | Acres    | Age      |
|-------------|----------|----------|----------|----------|
| 3.063671    | 2.580597 | 1.871925 | 1.057860 | 1.250570 |

Figure 7: VIF Output for model #3

To observe which variables are collinear, we generate a correlation plot (shown in figure 8). We can observe that there is strong correlation between living area and baths and bedrooms.

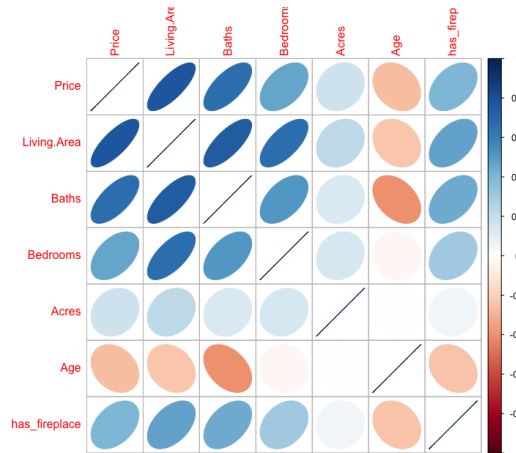


Figure 8: Saratoga Correlation Plot

**4b.** Is there evidence at least one of the acreage or age variables are needed in the model? Answer this question using a hypothesis test.

To answer this question we construct a restricted model, model #3R. We use the same formula from 2d to calculate the F-statistic, which is 1.27. This F value is not significant, with a p-value of 0.2821. This means that there is not evidence that adding both Acres and Age improves the model.

We then create restricted models, model #3R1 and model #3R2, which remove each of Acres and Age, respectively. For each model we find F-statistics of 0.718 and 1.9185, respectively. Neither of these models are significant improvements from the unrestricted model, with p-values of 0.397 and 0.1663, respectively.

We conclude that both Age and Acres can be excluded from the model without impacting the model accuracy.

**4c.** *Is there evidence the variation of the residuals is heteroskedastic? Answer this question with the appropriate visualization and hypothesis test.*

We observe from the residuals vs fitted plot for model #3 (shown in figure 9) that the residuals have decreasing thickness from left-to-right, which is an indicator that there may be heteroskedasticity.

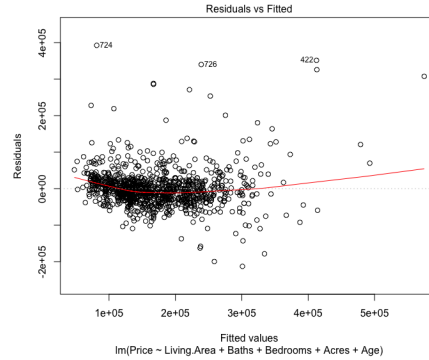


Figure 9: Residuals vs. Fitted Plot for Model #3

In the scale-location plot for model #3 (shown in figure 10), we observe that the standardized residuals are increasing with fitted values, which indicates that there is likely heteroskedasticity.

Lastly, we apply a Breusch-Pagan test to the fitted model, which results in a p-value  $< 2.2e-16$ . This is enough evidence to reject the null hypothesis that there is homoskedasticity.

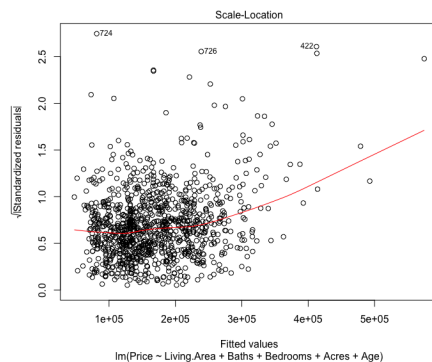


Figure 10: Scale-Location Plot for Model #3

**5a.** Fit a model that uses the size, number of baths, number of bedrooms and the fireplace variable to predict the price, denote this as model #4. Is there evidence the change in the average price is not zero dollars when changing from homes without a fireplace to homes with a fireplace? Answer this question using a hypothesis test.

There is not evidence that fireplace changes the price, as the coefficient estimate for fireplace is now 5,143, and the standard error is 3,756, which means that the 95% confidence interval contains zero.

The t-value of this estimate is 1.369, which is a p-value of 0.1712, so there is not enough evidence to reject the null hypothesis that the coefficient is zero.

**5b.** Refer to model #2 and part 3 (b). Explain why the results are different using model #4.

Compare to model #2, standard error actually decreases a bit but the size of the coefficient is now much lower since adding new variables removes positive bias of the omitted variables. However we're really seeing two different markets. Houses with fireplace are concentrated in areas that usually demand higher prices per sqft, while those without fireplaces are more spread out and consequently our linear model does worse at estimating the price based on the features selected.

**5c.** From model #4, identify any outliers. Explain what it means for an observation to be an outlier in this context.

The residuals indicate that the current model cannot properly explain these outlier points. This likely means that there are wide differences depending on the exact location of the property. Outliers represent very high priced properties.

**6a.** *Fit a model that uses the size, number of baths, number of bedrooms, fireplace variable, and an interaction between the size and fireplace variable to predict the price, denote this as model #5. For homes with a fireplace, what is the slope between size and price.*

The slope between size and price in this model is  $40.7 + 48.8 = 89.5$ .

**6b.** *Is there evidence the interaction term is needed in the model? Answer this question using a hypothesis test.*

To answer this question we created a restricted model without the interaction term and measured an F-statistic of 43.219, which is highly significant, with a p-value of 7.68e-11.

**6c.** *Explain what an interaction between the size and fireplace variables means in the context of the problem.*

The interaction term means that the effect of fireplace on price is dependent on the value of living area. As living area increases, the effect of fireplace increases by 40.73.

**6d.** *Are there any omitted variables that may create endogeneity bias? For each, indicate where the endogeneity bias may appear and if an instrument variable model would be appropriate to use in this situation.*

One can still suspect that there are some omitted variables creating endogenous bias to the model. For example, a small fancy loft in an urban area in a nice neighborhood with no fireplace would be more expensive than a large country side or suburban house with a fireplace far from the city. Other examples are school ratings, local crime, industry zoning, and taxes.

A possible instrument variable for Fireplace could be a regulation on having a fireplace in residential homes.

**7.** *How could the sales prices be spatially correlated? Explain.*

Sale prices could be spatially correlated, for example in a neighborhood on a golf course, or for houses with close proximity to a school, a train station, or houses along a waterfront. Also, houses in a city with a booming economy would probably be higher than nearby rural cities.



8. *How could the sales prices be temporally correlated? Explain.*

Fluctuations in the real-estate market could mean that house prices are temporally correlated, meaning that a house price in the same year is more correlated than house prices in different decades.

9. *Are these results useful for a real estate agent in San Luis Obispo CA? Explain.*

No, value of a fireplace on the price might be very different; same goes for other parameters such as the price per square feet.

## 2 Question 2 – Election

1. *Does Hillary Clinton rate relatively higher compared to Bernie Sanders among individuals who have a higher feeling thermometer rating for minority groups?*

To answer this question, we created a variable called `HC_over_BS`, which is the difference in rating between Hillary Clinton and Bernie Sanders for each observation in the dataset. We then created another variable, `ftminority`, which averages the ratings for both minority ratings, `ftblack` and `ftthisp`. Then, we created a fitted model with these two variables, with the dependent variable `HC_over_BS`.

From the resulting model, we do not see evidence that an increase in rating for minorities explains a relatively higher rating of Hillary Clinton compared to Bernie Sanders. The slope coefficient for our `ftminority` variable is 0.0072, with standard error of 0.046, which translates to a 95% confidence interval of -0.083 to 0.097. Since this includes zero, we cannot rule out that the coefficient is zero.

2. *How does the inclusion of respondents' perception of President Obama and the economy (well known predictors of presidential elections) impact your answer to the first question?*

Adding in parameters for rating of Obama and the economy (to a much lesser extent) reduces noise in the coefficient for minority and now signifies that – holding approval for Obama and rating of economy constant – that a higher rating for minorities decreases the likelihood of rating Hillary over Sanders.