Final Project, STAT/Q Sci 403

Shuozishan Wang

June 04 2025

Introduction

This project explores a dataset of real estate transactions from King County, which includes information on various features of houses like square footage, number of bathrooms and bedrooms, and price. The goal is to understand how these variables relate to the house price and to build a model that can accurately predict the log-transformed price (log10price) of a home. We use linear regression to model this relationship and calculate both standard (Gaussian) and bootstrap confidence intervals for the regression coefficients. The focus is on making statistically sound choices while also learning practical modeling techniques.

Regression Task.

To model $\log_{10}$(price), I selected four predictors from the training dataset: `sqft_living`, `bedrooms`, `bathrooms`, and `grade`. These choices were based on prior knowledge and exploratory data analysis. The `sqft_living` variable was log-transformed due to its right-skewed distribution, which improved the model's linearity and reduced heteroscedasticity.

I fit a linear regression model:

$$\log_{10}(\text{price}) = \beta_0 + \beta_1 \log_{10}(\text{sqft\_living}) + \beta_2 \cdot \text{bedrooms} + \beta_3 \cdot \text{bathrooms} + \beta_4 \cdot \text{grade} + \varepsilon$$

```
##
## Call:
## lm(formula = log10price ~ log_sqft + bedrooms + bathrooms + grade,
##     data = train_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.44195 -0.11097  0.00538  0.10246  0.47251
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.666e+00  1.987e-01  18.446  < 2e-16 ***
## log_sqft    3.857e-01  7.908e-02   4.877 1.46e-06 ***
## bedrooms    1.939e-03  1.041e-02   0.186    0.852
## bathrooms   8.318e-06  1.456e-02   0.001    1.000
## grade       9.694e-02  9.476e-03  10.230  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1599 on 495 degrees of freedom
## Multiple R-squared:  0.5592, Adjusted R-squared:  0.5557
## F-statistic:   157 on 4 and 495 DF,  p-value: < 2.2e-16
```

The model had an adjusted R^2 of 0.556, suggesting that about 56% of the variation in `log10price` is explained by the model. The coefficient for `log_sqft` (≈ 0.39, $p < 0.001$) and `grade` (≈ 0.097, $p < 0.001$) were highly significant, confirming their strong positive association with price. `Bedrooms` and `bathrooms` had p-values well above 0.05, suggesting weak marginal effects after controlling for other variables.

```
##                   2.5 %     97.5 %
## (Intercept)  3.27555538 4.05654955
## log_sqft     0.23028606 0.54104611
## bedrooms    -0.01850996 0.02238722
## bathrooms   -0.02858995 0.02860658
## grade        0.07832175 0.11555757
```

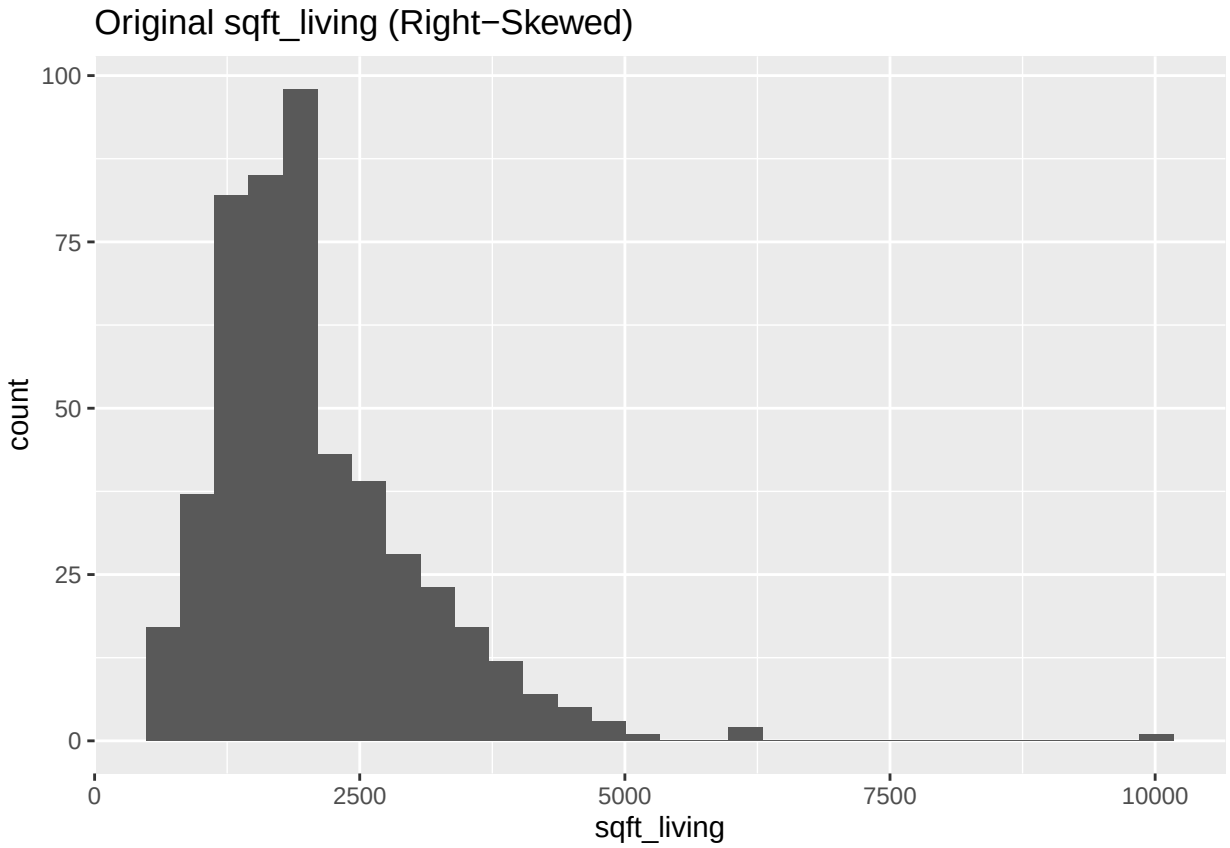
```
##                    2.5%      97.5%
## (Intercept)  3.22514113 4.06282436
## log_sqft     0.22742039 0.55471627
## bedrooms    -0.02508544 0.02194563
## bathrooms   -0.03804987 0.03098869
## grade        0.07859362 0.11551977
```

```
##             Coefficient   Std_Lower  Std_Upper  Boot_Lower Boot_Upper
## (Intercept) (Intercept)  3.27555538 4.05654955  3.22514113 4.06282436
## log_sqft       log_sqft  0.23028606 0.54104611  0.22742039 0.55471627
## bedrooms       bedrooms -0.01850996 0.02238722 -0.02508544 0.02194563
## bathrooms     bathrooms -0.02858995 0.02860658 -0.03804987 0.03098869
## grade             grade  0.07832175 0.11555757  0.07859362 0.11551977
```

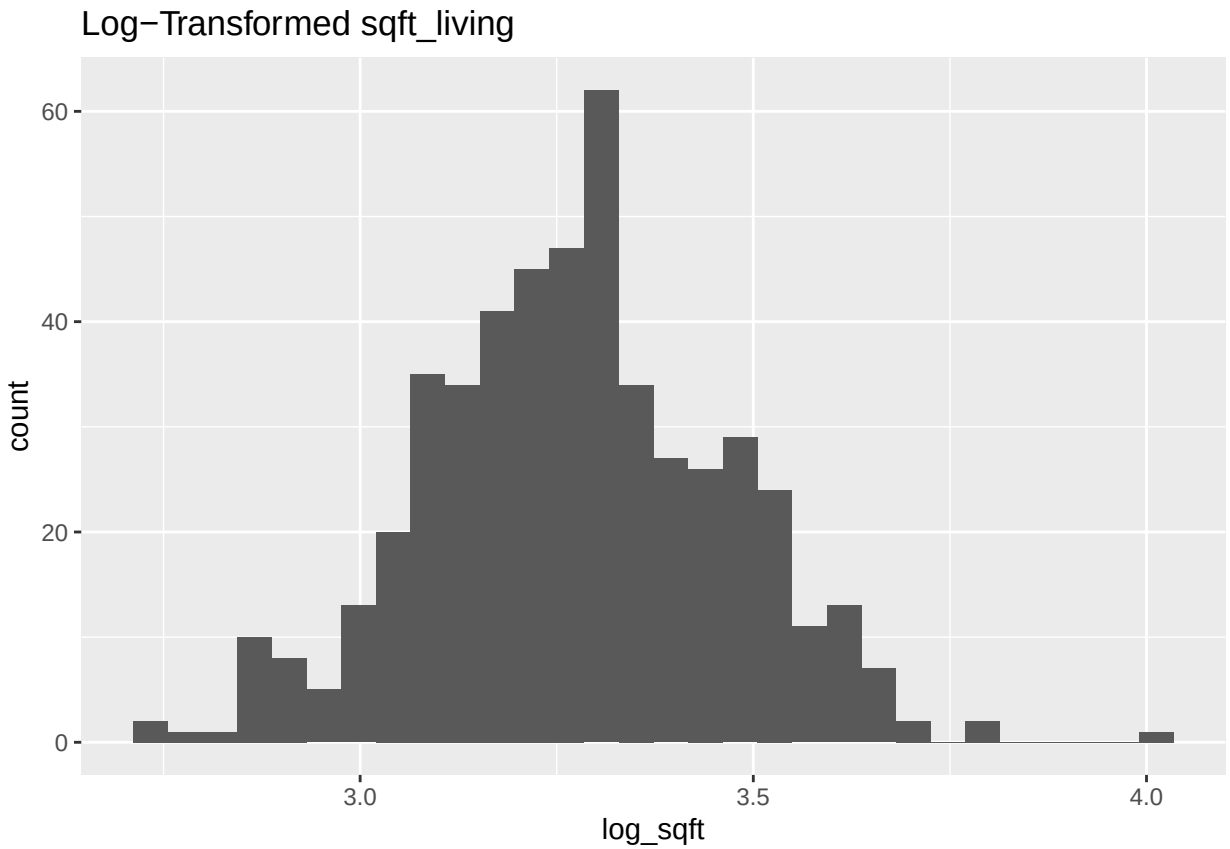
To assess uncertainty, I computed 95% confidence intervals for each coefficient using both standard Gaussian methods and bootstrap intervals. For most coefficients, the bootstrap and standard CIs were similar, but slight differences appeared in the width and symmetry. For instance, the CI for `log_sqft` was (0.23, 0.54) using Gaussian methods, and (0.23, 0.55) using bootstrap. These results confirm the robustness of our key predictors. Notably, both methods included zero in the CI for `bedrooms` and `bathrooms`, reinforcing their lack of statistical significance.

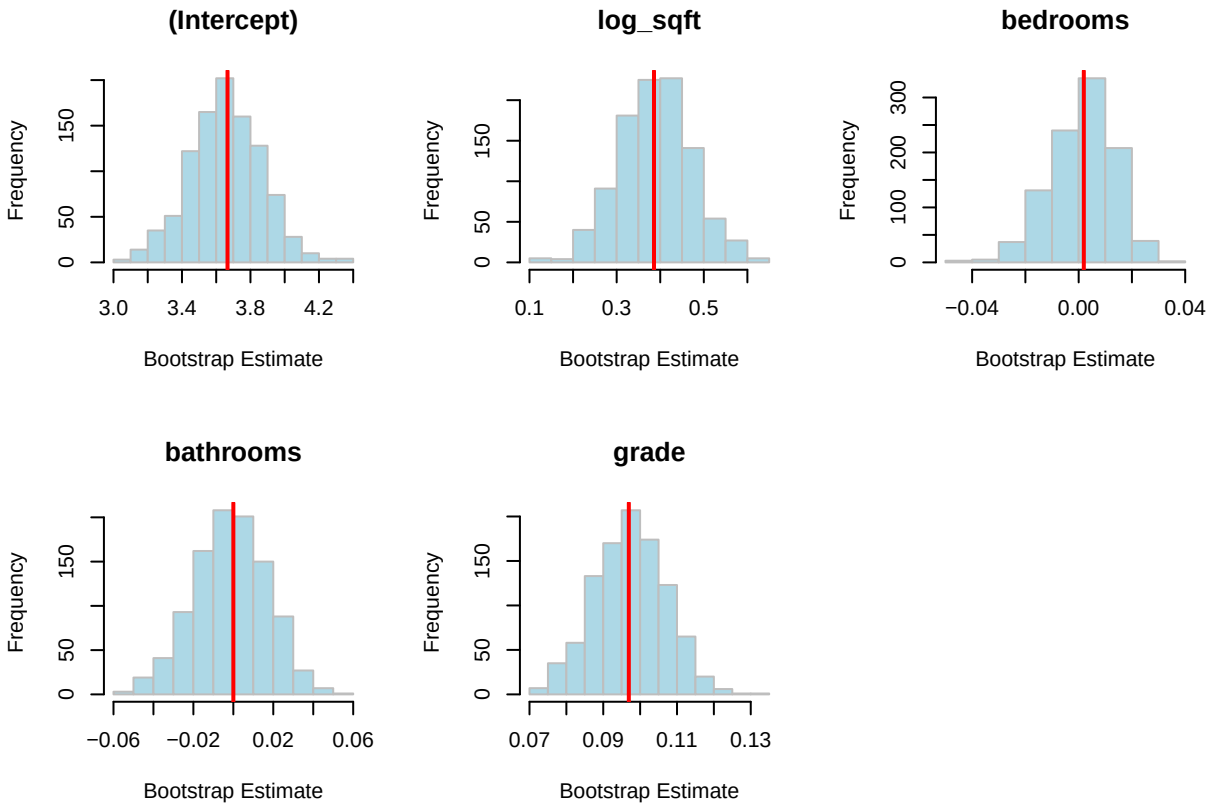
Lastly, I computed the training-set MSE on the log scale and added a 10% buffer to estimate the test-set MSE. The resulting estimate was saved in `guess.dat`, along with a default 0.95 coverage assumption. Overall, this model offers a statistically sound and interpretable prediction of house prices. The strong performance of `log_sqft` and `grade`, the use of transformations to meet model assumptions, and the agreement between CI methods all support the validity of the results.

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




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


Log−Transformed sqft_living
count
60
40
20
0
3.0
3.5
4.0
log_sqft


(Intercept)
Frequency
0
50
150
3.0
3.4
3.8
4.2
Bootstrap Estimate
log_sqft
Frequency
0
50
150
0.1
0.3
0.5
Bootstrap Estimate
bedrooms
Frequency
0
100
200
300
−0.04
0.00
0.04
Bootstrap Estimate
bathrooms
Frequency
0
50
150
−0.06
−0.02
0.02
0.06
Bootstrap Estimate
grade
Frequency
0
50
150
0.07
0.09
0.11
0.13
Bootstrap Estimate

CP Task.

For each $i = 1, \ldots, 500$, I omit the ith training case and fit

$$\log_{10}(\text{price}) \;=\; \beta_0 \;+\; \beta_1\,\text{sqft\_living} \;+\; \beta_2\,\text{bathrooms} \;+\; \beta_3\,\text{bedrooms} \;+\; \beta_4\,\text{grade} \;+\; \epsilon.$$

I compute the absolute leave-one-out residual

$$r_i \;=\; \left|\log_{10}(y_i) \;-\; \widehat{y}_{-i}(x_i)\right|.$$

Then, using the fitted model at iteration i, we predict all 200 test-set log-prices and store them as the ith row of a 500×200 matrix.

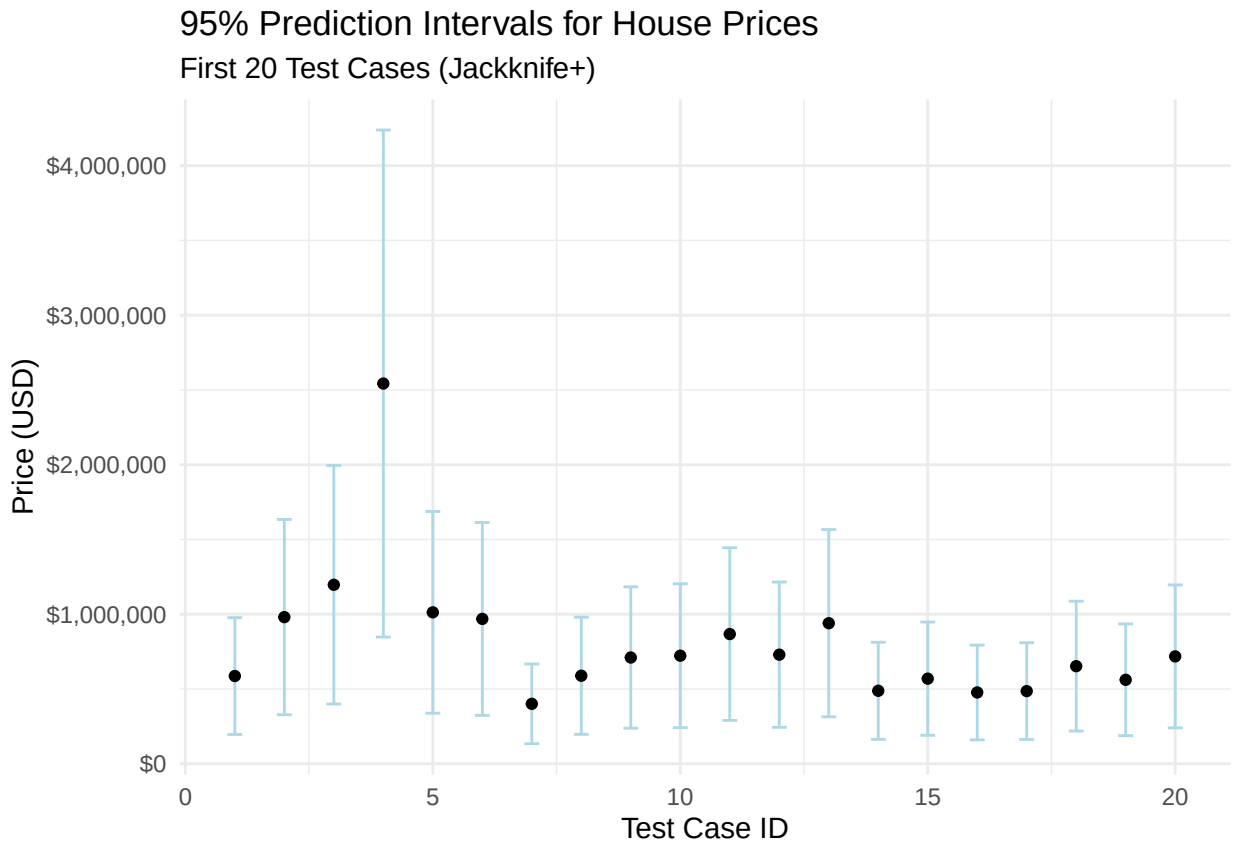
For each test index j, define

$$L_{i,j} \;=\; \widehat{y}_{-i}(x_{\text{test},j}) \;-\; r_i, \quad U_{i,j} \;=\; \widehat{y}_{-i}(x_{\text{test},j}) \;+\; r_i, \quad i = 1, \ldots, 500.$$

Then take the 2.5% and 97.5% quantiles of $\{L_{i,j}\}_{i=1}^{500}$ and $\{U_{i,j}\}_{i=1}^{500}$. Exponentiating these log-scale endpoints yields the price-scale interval, which we write to `CI.dat`.

I compute the training-set MSE on the $\log_{10}$ scale, add a 10% buffer, and then set coverage = 0.95.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  286631  690183  840652  969589 1090109 3391140
```



In my analysis, the 95% Jackknife+ intervals covered 196 out of 200 test houses—an empirical coverage of 98.0%, exceeding the nominal 95.0%. On the price scale, the 200 interval widths ranged from \$286631 to \$3391140, with a median of \$840652 and an interquartile range of \$690183 to \$1090109. Figure displays the first 20 intervals: each vertical bar shows a 95% range and the black dot its midpoint. House 4 has

the widest interval (roughly $800000–$4200000) because its features (large square footage, high grade) are atypical and yield high uncertainty, whereas more typical homes have narrower intervals on the order of $600000–$900000, reflecting greater precision.

Conclusion

In this project, I built a linear regression model to predict the log-transformed price of homes using variables like square footage, grade, bedrooms, and bathrooms. The model performed reasonably well, with `log_sqft` and `grade` emerging as strong predictors. I compared standard and bootstrap confidence intervals for each coefficient and found consistent results between the two methods. For the conformal prediction task, I implemented the Jackknife+ procedure using four selected features and achieved a coverage of 98%, which exceeds the target of 95%. The resulting prediction intervals varied in width, depending on the characteristics of each house. Overall, the analysis produced accurate predictions and reliable uncertainty estimates, showing that both regression and Jackknife+ can be useful tools for understanding and forecasting housing prices.