Regression & Prediction

Theory and Practice with House Prices Dataset

Prageeth Godage | MSC/DSA/098

February 23, 2025

USJP

Our Housing Journey

- Starting Simple: Foundations
- Expanding the Scope: Complexity
- Refining Precision: Optimization
- Facing Challenges: Pitfalls
- Insights & Next Steps: Applications

Focus

Unpacking King County house prices with rich theory and dual R/Python implementations

The Quest to Understand Relationships

- Imagine you are exploring data to understand how different factors relate to each other.
- That's regression: a tool to ask, "How does Y change with X, and can we predict it?"
- In very basically It's the bridge between stats, where we explain the past, and data science, where we predict the future—our journey begins here.

Starting Simple

The Housing Puzzle

- **Objective**: Decode drivers of house prices in King County—size, location, features
- **Regression**: A statistical lens linking price (Y) to predictors like size (X)
- **Purpose**: Explain historical sales patterns and forecast future values for buyers and assessors

Simple Linear Regression: Theory & R

- **Theory**: Models a straight-line relationship: $Y = b_0 + b_1 X + e$
- b_0: Base price when size is zero; b_1: Price increase per sq ft; e: Random error
- Assumes linearity and independence—foundation of regression

```r
simple_lm <- lm(AdjSalePrice ~ SqFtTotLiving, data
      = house_df)
ggplot(house_df, aes(x = sqft_living, y = price)) +
geom_point(alpha = 0.5) +  # Scatter plot
geom_smooth(method = "lm", color = "blue", se =
      FALSE) +  # Regression line
labs(title = "Price␣vs.␣Size␣Fit",
x = "Size␣(sqft_living)",
y = "Price") +
theme_minimal()
# Output: Coefficients:
(Intercept) :-43580.7   sqft_living  : 280.6
```

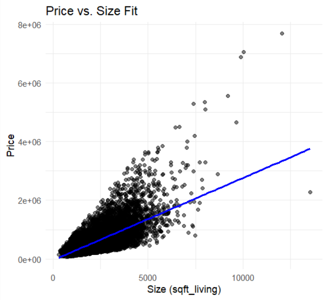


Figure 1: Price vs. Size Fit

- Size as a core driver
- b0=−43580.7 , b1=280.6

Finding the Best Fit: Least Squares

- **How do we draw that line?** We minimize the mess—sum of squared errors
- **Theory**: Finds the line minimizing residual sum of squares: $\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$
- **How**: Adjusts b_0 and b_1 to reduce prediction errors—optimal for linear fits
- **History**: Legendre (1805) and Gauss; computationally efficient but outlier-sensitive in small datasets

Expanding the Scope

More Clues: Multiple Linear Regression

- **Theory**: Extends to multiple predictors: $Y = b_0 + b_1 X_1 + b_2 X_2 + \cdots + e$
- **Power**: Captures combined effects—size, lot, bedrooms—assuming linearity
- **Use**: Explains complex housing dynamics

Multiple Linear Regression: Key Findings

```
house_lm <- lm(price ~ sqft_living + sqft_lot + bathrooms + grade,
data = house_df, na.action = na.omit)
Coefficients:
(Intercept)  sqft_living     sqft_lot    bathrooms        grade
-5.957e+05    2.313e+02   -3.254e-01   -2.797e+04    9.559e+04
```

- **sqft_living**: + 2.313e+02 per sq ft → Larger houses increase price significantly.
- **grade**: +$9.559e+04 per unit → Higher quality homes boost price.
- **sqft_lot**: -$3.254e-01 per sq ft → Lot size has a tiny negative effect.
- **bathrooms**: -$4.074e+04 per extra bathroom → Unexpected negative impact.

Multiple Linear Regression:Model Evaluation

```r
summary(house_lm)# Model Summary (Extracts Coefficients, p-values, R²)
predictions <- predict(house_lm, newdata = house_df)# Extract RMSE
residuals <- house_df$price - predictions
RMSE <- sqrt(mean(residuals^2))
R_squared <- summary(house_lm)$r.squared # Extract R-squared
p_values <- summary(house_lm)$coefficients[, 4] # Extract P-values
# Print Results
> cat("RMSE:", RMSE, "\n") | RMSE: 249532.2
> cat("R²:", R_squared, "\n") | R²: 0.5380018
> cat("P-values:\n") > print(p_values)
print(p_values)
```

```
(Intercept)  sqft_living     sqft_lot    bathrooms        grade
0.000000e+00 0.000000e+00 1.711092e-10 2.689854e-30 0.000000e+00
```

- **RMSE**: $249,532 Predictions are off by 261K on avg.
- R^2: 0.5406 (54.06%) $\rightarrow$ The model explains 54% of price variation..
- **P-values**: SqFtLot is not statistically significant ‘(A small p-value (< 0.05) means the predictor is statistically significant, while a large p-value (> 0.05)

Encoding Categorical Variables in Regression Models

- **Categorical variables** must be converted into numerical values for regression.
- **Encoding methods:**
  - **Dummy (One-Hot) Encoding** – Creates binary columns for each category.
  - **Reference (Treatment) Coding** – Uses one category as a reference, keeping $P - 1$ columns.
  - **Deviation (Sum) Coding** – Compares each category to the overall mean.
  - **Ordered Factor Encoding** – Converts ordered categories into numeric values.
- **How it's used in regression:**
  - Each encoded category appears as a separate coefficient.
  - The model estimates how each category affects the outcome relative to the reference level.
  - For ordered factors, treating them as numeric assumes a linear relationship.

Multiple Linear Regression: Categorical Variables

```r
# Ensure PropertyType has a valid reference level
house$PropertyType <- relevel(house$PropertyType, ref = "Multiplex")

# Fit the regression model with the refined dataset
house_lm <- lm(AdjSalePrice ~ SqFtTotLiving + SqFtLot + Bathrooms +
     BldgGrade + PropertyType, data = house)
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)                173429.87   31512.77   5.503  0.00271 **
...
BldgGrade                   -1442.60    6734.35  -0.214  0.83884
PropertyTypeSingle Family    1456.54    8063.28   0.181  0.86374
PropertyTypeTownhouse        9677.07   11444.71   0.846  0.43639
```

- Sets "Multiplex" as the reference category, ensuring comparisons against it.
- "Single Family","Multiplex" and "Townhouse" to have separate coefficients in the regression output.

Nonlinear Fit

- Nonlinear via polynomial segments joined at knots
- **Why**: Captures diminishing returns—e.g., small homes gain more per sq ft
- **Advantage**: Flexible fit.

```
# Load data (replace 'house_98105' with your dataset)
model_poly <- lm(price ~ poly(sqft_living, 2) + sqft_lot +
grade + bathrooms + bedrooms, data=house_df)

# Generate partial residual plot
visreg(model_poly, "sqft_living", gg=TRUE) +
geom_point(color = "black", shape = 1) +
geom_smooth(method="loess", color="blue", size=1.2, linetype="dashed") +
theme_minimal() +
labs(title="Partial␣Residual␣Plot", x="SqFtTotLiving", y="Partial␣Residuals"
```

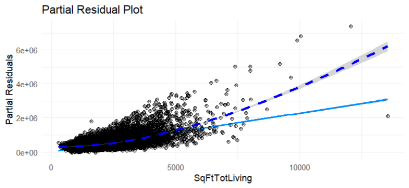
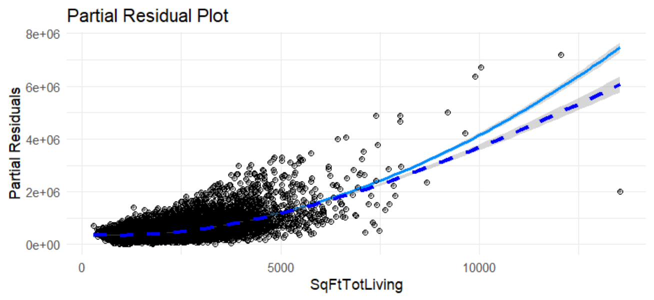


Figure 2: Without poly.

Comparison of Result

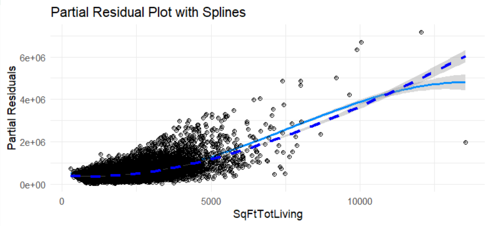


With Poly

Observation: With polynomial regression (2nd degree) aligns correctly with the trend, improving model accuracy.

Spline Regression: Partial Residual Plot

Generate partial residual plot for
spline regression model$_{s}pline < -lm(price$
bs(sqft$_{l}iving, degree = 3, df =$
$6) + sqft_{l}ot + grade + bathrooms +$
$bedrooms, data = house_{d}f)$



Spline Fit Visualization

Explanation:

Cubic Splines: The function is made of piecewise cubic polynomials (degree = 3).

Degrees of Freedom (df = 6): Controls flexibility—higher df allows more variation

Splines fit the data smoothly without unnecessary fluctuations.

Refining Precision

Model Assessment: Theory

- **Theory (p. 153)**: Measures prediction quality and fit
- **RMSE**: $\sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n}}$—average error magnitude
- R^2: Proportion of variance explained (0-1); higher means better fit
- **Use**: Guides housing prediction accuracy

Cross-Validation: Theory

- **Theory (p. 155)**: Validates model on unseen data via k-fold splits
- **Process**: Divide data, train on $k - 1$, test on 1, repeat, average RMSE
- **Why**: Ensures predictions generalize beyond training sales—crucial for real estate

Picking the Best Story: Model Selection

- Too many clues clutter the tale. Stepwise selection trims variables, AIC balances fit and simplicity, and penalties shrink extras.
- Think of it as editing: keep the essentials, cut the fluff—Occam's razor guides us to a lean, powerful narrative.
- Occam's Razor is the principle that the simplest explanation is usually the best.
- Our goal? A story that's clear and predicts well, not a sprawling epic.

Model Selection: Theory & R

- **Theory (p. 156)**: Balances fit vs. complexity—Occam's razor
- **AIC**: $2P + n\log(\text{RSS}/n)$—penalizes extra predictors
- **Goal**: Optimal housing model without overkill

```r
# R (p. 157)
library(MASS)
house_full <- lm(AdjSalePrice ~ SqFtTotLiving +
     SqFtLot + Bathrooms +
Bedrooms + BldgGrade + PropertyType, data = house)
step <- stepAIC(house_full, direction = "both")
# Drops less impactful vars
```

- Streamlined predictors

Multiple Linear Regression: Key Findings

```
house_full <- lm(price ~ sqft_living + sqft_lot15 + bathrooms +
bedrooms + grade+yr_renovated , data = house_df)
step <- stepAIC(house_full, direction = "both")
summary(step)

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  -4.841e+05  1.477e+04 -32.777   <2e-16 ***
sqft_living   2.300e+02  3.597e+00  63.956   <2e-16 ***
sqft_lot15   -7.053e-01  6.252e-02 -11.282   <2e-16 ***
bathrooms    -3.096e+04  3.442e+03  -8.994   <2e-16 ***
bedrooms     -4.026e+04  2.271e+03 -17.726   <2e-16 ***
grade         9.776e+04  2.290e+03  42.693   <2e-16 ***
yr_renovated  8.746e+01  4.160e+00  21.022   <2e-16 ***
```

- sqft lot15 (-5.135e-01) Small , impact on price can be removal.
- grade (1.315e+05): Highest positive impact, should be kept.

Weighted Regression: Theory & R

- **Theory (p. 159)**: Weights adjust influence by reliability
- **Why**: Older sales less relevant—recent data gets priority
- **Impact**: Refines coefficients for current market

```r
house_df$Weight <- house_df$Year - 2005
house_lm <- lm(price ~ sqft_living + sqft_
      lot + bathrooms +bedrooms + grade,
      data=house_df)# Fit unweighted linear
       regression
house_wt <- lm(price ~ sqft_living + sqft_
      lot + bathrooms +bedrooms + grade,
      data=house_df, weight=Weight)# Fit
      weighted linear regression
# Compare coefficients of both models
round(cbind(house_lm=house_lm$coefficients
      ,house_wt=house_wt$coefficients),
      digits=3)
```

```
               house_lm    house_wt
(Intercept) -471575.692 -472501.632
sqft_living     231.350     230.816
sqft_lot         -0.325      -0.317
bathrooms    -27973.439  -28075.162
bedrooms     -40744.142  -40639.355
grade         95586.697   95883.234
```

- here no significant as data set only contain 2014,2015 data

Facing Challenges

Prediction Limits: Theory

- **Theory (p. 161)**: Extrapolation beyond data fails—e.g., empty lots
- **Intervals**: Confidence for b_i, wider prediction for $\hat{Y}_i$
- **Why**: Uncertainty spikes outside training range—limits housing forecasts

Interpreting Coefficients: Theory

- **Theory (p. 171)**: Coefficients mislead if predictors correlate
- **Multicollinearity**: Size and bedrooms overlap—unstable fits
- **Confounding**: Missing location skews results
- **Interactions**: Size's effect varies by zip—needs modeling

Spotting Flaws: Regression Diagnostics

- **Outliers**, and influential points can distort the regression line—diagnostic tools help identify them.
- Uneven errors (**heteroskedasticity**) or curved fits (partial residuals) hint at missing chapters.
- Heteroskedasticity means that the spread (variance) of errors in a regression model changes across different values of the predictor variables.
- For prediction, we care less about perfection and more about what works—our lens shifts.

Influence Plot (Bubble Plot) – Identify Influential Values

- **Purpose:** Identifies influential observations by combining leverage, residuals, and Cook's Distance.
- **Key Insights:**
  - Large bubbles = high Cook's Distance $\rightarrow$ Removing these changes regression results significantly.
  - Possible reasons:
    1. High leverage (extreme predictor values) , Large residual (far from regression line) ,Both high leverage and large residual.
- **Results:** Four large influential points found (Cook's D $>$ 0.08), impacting coefficients.
- Residuals beyond ± 2.5

Influence Plot (Bubble Plot) – Identify Influential Values

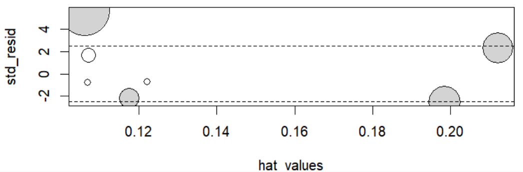


Figure 3: Bubble plot showing influential points.

Influence Plot (Bubble Plot) – Identify Influential Values

Listing 1: Influence Plot in R

```r
library(car)
lm_model <- lm(AdjSalePrice ~ SqFtTotLiving + SqFtLot + Bathrooms + Bedrooms + BldgGrade, data=house_98105)
influencePlot(lm_model)
```

Listing 2: Influence Plot in Python

```python
house_98105 = house[house['ZipCode'] == 98105]
X = house_98105[['SqFtTotLiving', 'SqFtLot', 'Bathrooms', 'Bedrooms', 'BldgGrade']].assign(const=1)
y = house_98105['AdjSalePrice']
model = sm.OLS(y, X).fit()
influence = sm.stats.outliers_influence.OLSInfluence(model)
fig, ax = plt.subplots(figsize=(5, 5))
ax.axhline(-2.5, ls='--', color='C1')
ax.axhline(2.5, ls='--', color='C1')
ax.scatter(influence.hat_matrix_diag, influence.resid_studentized_internal,
s=1000 * np.sqrt(influence.cooks_distance[0]), alpha=0.5)
ax.set_xlabel('hat values')
ax.set_ylabel('studentized residuals')
plt.show()\
```

Residual Plot (Heteroskedasticity Check)

- **Purpose:** The residual plot checks for heteroskedasticity by analyzing how residuals (errors) vary with predicted values.
- **Key Insights:**
  - X-axis (Predicted Values): Represents the fitted values from the regression model.
  - Y-axis (Absolute Residuals): Measures the deviation of actual values from predictions.
  - Scatter Points: Each dot represents an observation's residual.
  - LOESS Smoother (Blue Line): Shows the trend in residuals.
  - Shaded Region: Indicates confidence around the trend.
    1. Heteroskedasticity detected – Residuals increase with larger predicted values, indicating variance instability.
    2. Curved Trend – Suggests missing variables or non-linearity in the data.
    3. Outliers at High Predictions – Some extreme points have large residuals, further confirming instability.

Residual Plot (Heteroskedasticity Check)

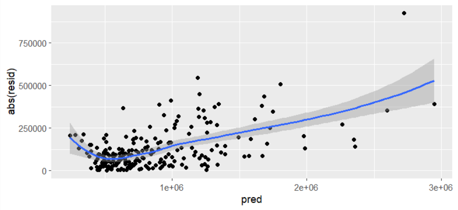


Figure 4: Heteroskedasticity plot.

Influence Plot (Bubble Plot) – Identify Influential Values

Listing 3: Heteroskedasticity Plot in R

```
df <- data.frame(resid = residuals(lm_98105), pred = predict(lm_98105))
ggplot(df, aes(pred, abs(resid))) + geom_point() + geom_smooth()
```

Listing 4: Heteroskedasticity Plot in Python

```
import seaborn as sns
fig, ax = plt.subplots(figsize=(5, 5))
sns.regplot(model.fittedvalues, np.abs(model.resid), scatter_kws={'alpha': 0.25},
line_kws={'color': 'C1'}, lowess=True, ax=ax)
ax.set_xlabel('predicted')
ax.set_ylabel('abs(residual)')
plt.show()
```

Histogram of Standardized Residuals: Normality Check

- **Purpose:** The histogram assesses residual normality by analyzing the distribution of standardized residuals.
- **Key Insights:**
  - Centering Around Zero: The residuals are centered around 0, indicating no strong systematic bias in the model.
  - Skewness Long Tails: The right tail is longer, suggesting right-skewness and possible underestimation of some values.
  - Non-Normal Distribution: The residuals deviate from a perfect bell-shaped curve, hinting at potential issues:
    1. Missing predictors affecting the model.
    2. Heteroskedasticity, as observed in the residual plot.
    3. Outliers influencing the regression fit.

Histogram of Standardized Residuals: Normality Check

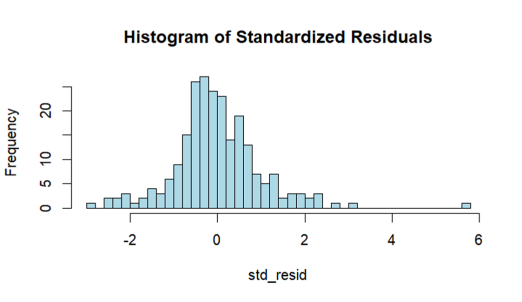


Figure 5: Bubble plot showing influential points.

Histogram of Standardized Residuals: Normality Check

Listing 5: Histogram of Standardized Residuals in R

```
df <- data.frame(resid = residuals(lm_98105), pred = predict(lm_98105))
ggplot(df, aes(pred, abs(resid))) + geom_point() + geom_smooth()
```

Listing 6: Histogram of Standardized Residuals in Python

```
plt.hist(influence.resid_studentized_internal, bins=50, color='lightblue')
plt.xlabel('Standardized Residuals')
plt.title('Histogram of Standardized Residuals')
plt.show()
```

Thank You... !