Real estate analysis

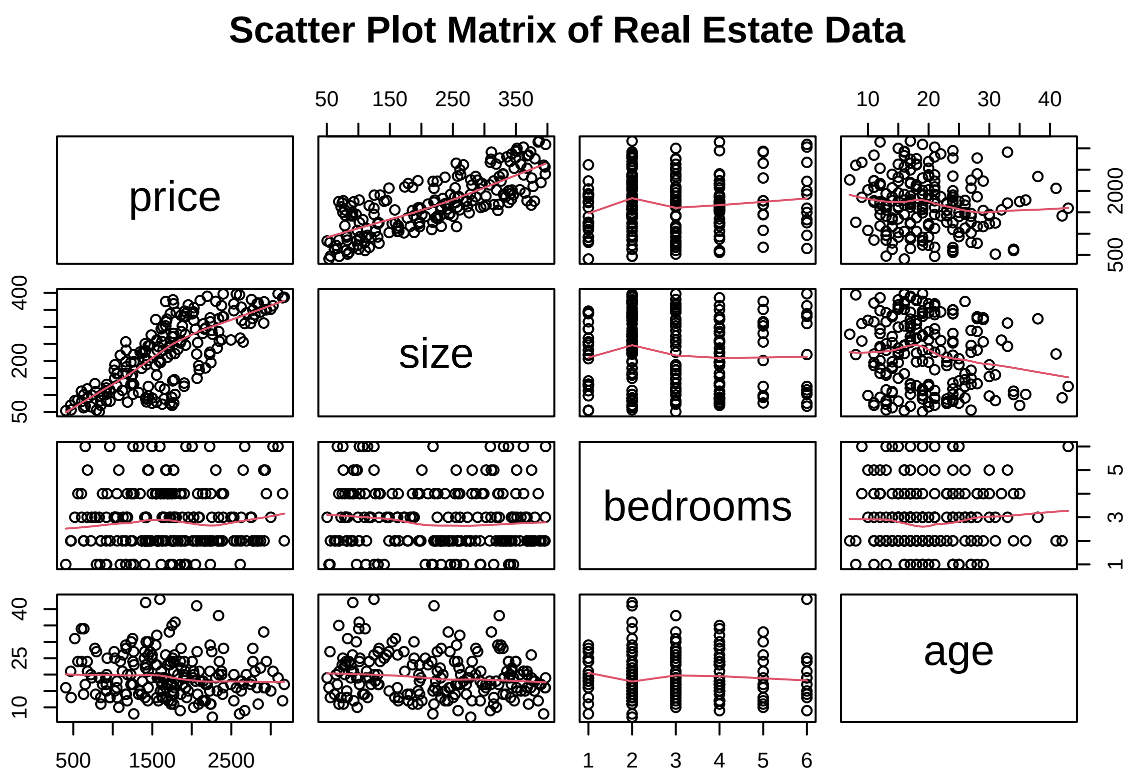
Hatim Hussaini

2024-11-22

Exploring Real Estate Data

Understanding Relationships Between Variables

```r
pairs(real_estate, main = "Scatter Plot Matrix of Real Estate Data", panel = panel.smooth)
```



Insights:

- Price vs. size has the strongest positive correlation, as bigger properties tend to be more expensive.
- Slight positive correlations exist between price and both bedrooms and age, suggesting newer homes or homes with more bedrooms are valued higher.

```r
cor(real_estate)
```

```
##                price        size    bedrooms         age
## price     1.00000000  0.77994644  0.05560245 -0.12347514
## size      0.77994644  1.00000000 -0.07285563 -0.16695401
## bedrooms  0.05560245 -0.07285563  1.00000000  0.02850195
## age      -0.12347514 -0.16695401  0.02850195  1.00000000
```

Insights:

- Strong correlation between price and size (0.7799), indicating size is a key factor influencing price.
- Correlations with other predictors, like bedrooms and age, are minimal.

Building a Full Regression Model to Predict Property Price

```r
model <- lm(price ~ size + bedrooms + age, data = real_estate)
summary(model)
```

Estimating the Impact of Property Size on Price

```
##
## Call:
## lm(formula = price ~ size + bedrooms + age, data = real_estate)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -748.08 -318.57  -54.74  366.46  784.33
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 449.2955   133.7219   3.360  0.00094 ***
## size          4.9371     0.2819  17.514  < 2e-16 ***
## bedrooms     53.6872    21.1222   2.542  0.01182 *
## age           0.4821     4.3038   0.112  0.91092
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 398.7 on 193 degrees of freedom
## Multiple R-squared:  0.621,  Adjusted R-squared:  0.6152
## F-statistic: 105.4 on 3 and 193 DF,  p-value: < 2.2e-16
```

```r
summary_model <- summary(model)
```

$$\hat{\beta}_{size} \pm t_{n-p,1-\alpha/2} \cdot s.e.(\hat{\beta}_{size})$$

$$= \hat{\beta}_{size} \pm t_{193,0.975} \cdot s.e.(\hat{\beta}_{size})$$

$$= 4.9371 \pm 1.97 \times 0.2819$$
$$= (4.38, 5.49)$$

We are 95% confident that for every unit increase in size, the price of the property will increase by between 4.38 and 5.49 thousand dollars on average.

Testing the Overall Significance of the Model

Conducting an F-Test   Theoretical Model:

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon_i, \quad i = 1, 2, \ldots, n$$

- Y_i is the response variable (price).
- β_0 is the intercept.
- $\beta_1, \beta_2, \beta_3$ are coefficients for the predictors X_1 (size), X_2 (bedrooms), and X_3 (age).
- $\epsilon_i \sim N(0, \sigma^2)$ represents random variation.

Null Hypothesis: $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ (no relationship between predictors and price).

```r
anova(model)
```

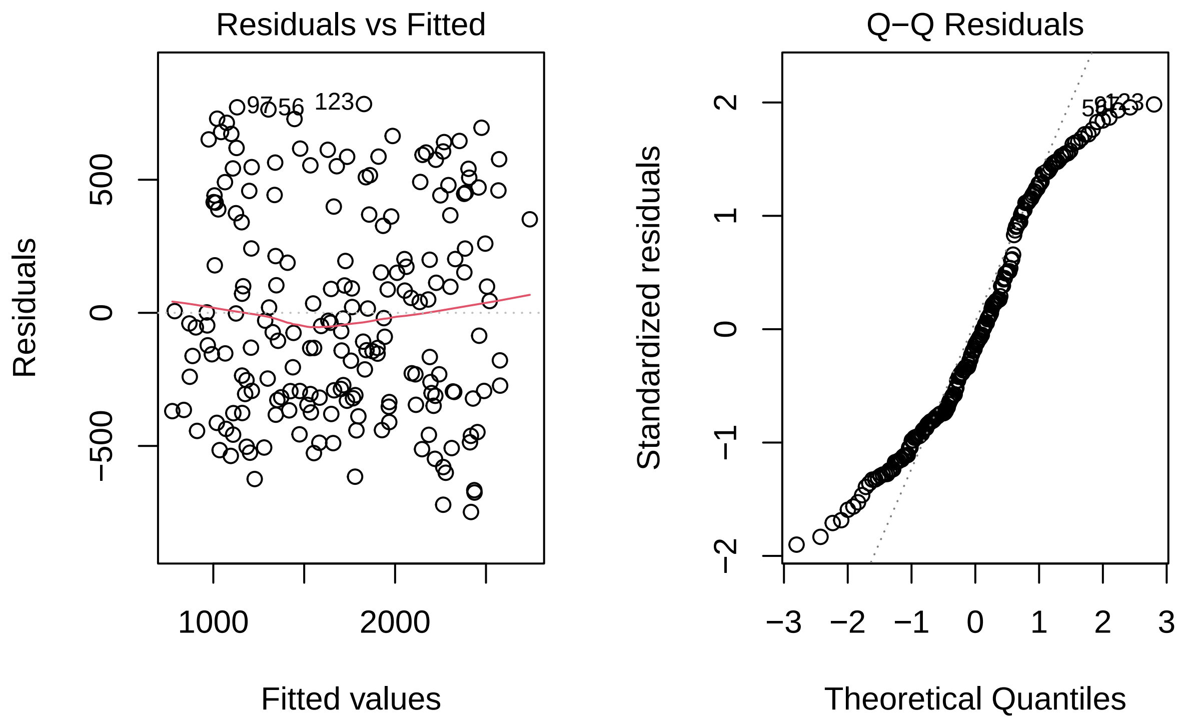
```
## Analysis of Variance Table
##
## Response: price
##            Df   Sum Sq  Mean Sq  F value  Pr(>F)
## size        1 49256631 49256631 309.8153 < 2e-16 ***
## bedrooms    1  1028915  1028915   6.4717 0.01174 *
## age         1     1995     1995   0.0125 0.91092
## Residuals 193 30684511   158987
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Regression SS = 50287541.
- Regression MS = 16762514.
- F-statistic: 105.43323.
- p-value: 1.9e-40 < 0.05.

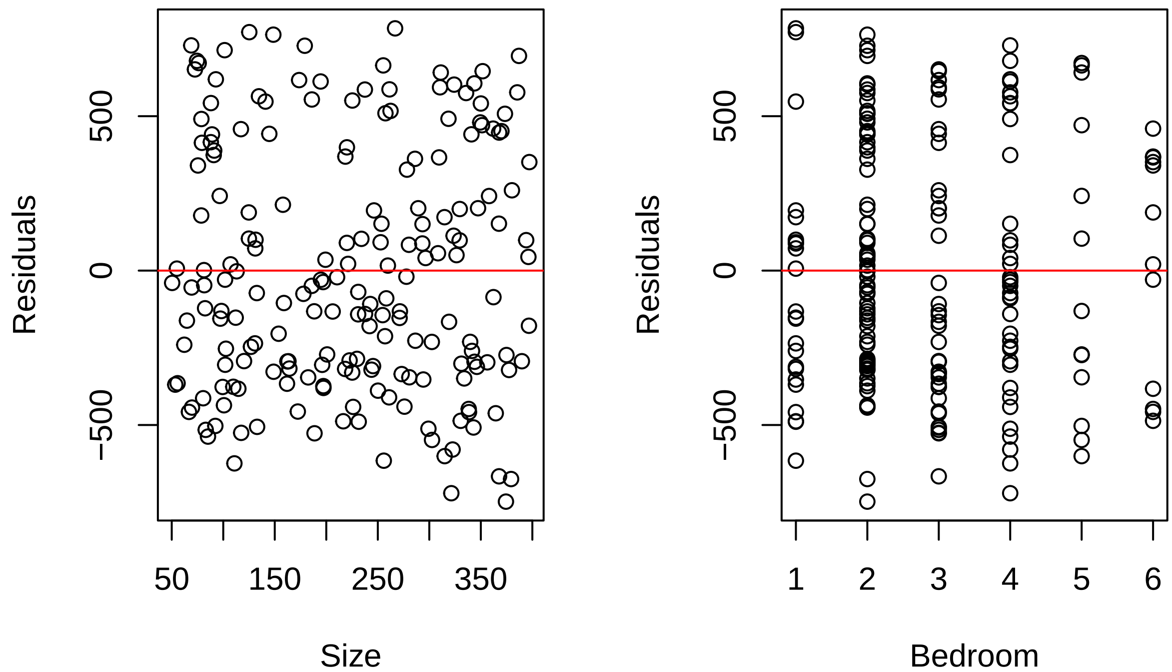
Conclusion: - There is significant evidence to reject H_0. - At least one predictor (size, bedrooms, age) significantly impacts price.

Validating the Regression Model

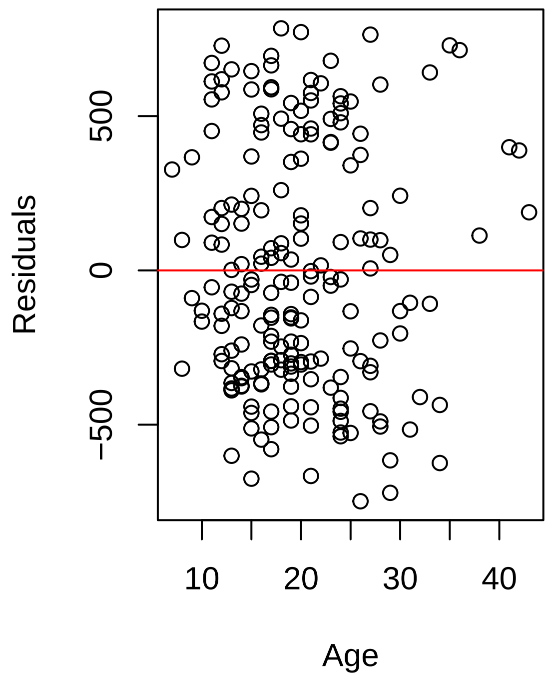
```r
par(mfrow=c(1,2))
plot(model, which=1:2)
```




```
plot(resid(model) ~ real_estate$size, xlab = 'Size', ylab = 'Residuals')
abline(h=0, col="red")
plot(resid(model) ~ real_estate$bedrooms, xlab = 'Bedroom', ylab = 'Residuals')
abline(h=0, col="red")
```




```
plot(resid(model) ~real_estate$age, xlab = 'Age', ylab = 'Residuals')
abline(h=0, col="red")
```



Insights:

- Residual vs. fitted plots show mostly random scatter with some imbalance in homoscedasticity.
- Q-Q plot shows deviations from normality.
- Residuals vs. predictors indicate linearity.

Overall, the model meets assumptions for linear regression with minor deviations.

Evaluating Model Fit (R^2)

$$R^2 = 1 - \frac{SS_{Res}}{SS_{Total}} = 1 - \frac{30684511}{50287541} = 0.39$$

An R^2 of 0.39 suggests the model explains 39% of the variability in property prices, indicating potential missing predictors like location.

Refining the Model with Feature Selection

```r
summary_model <- summary(model)
summary_model
```

```
##
## Call:
```

```
## lm(formula = price ~ size + bedrooms + age, data = real_estate)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -748.08 -318.57  -54.74  366.46  784.33
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 449.2955   133.7219   3.360  0.00094 ***
## size          4.9371     0.2819  17.514  < 2e-16 ***
## bedrooms     53.6872    21.1222   2.542  0.01182 *
## age           0.4821     4.3038   0.112  0.91092
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 398.7 on 193 degrees of freedom
## Multiple R-squared:  0.621,  Adjusted R-squared:  0.6152
## F-statistic: 105.4 on 3 and 193 DF,  p-value: < 2.2e-16
```

Removing the least significant predictor (age):

```r
model2 <- update(model, . ~ . - age)
summary_m2 <- summary(model2)
summary_m2
```

```
##
## Call:
## lm(formula = price ~ size + bedrooms, data = real_estate)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -744.25 -321.86  -59.73  362.39  783.79
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 459.8715    94.4581   4.869 2.33e-06 ***
## size          4.9318     0.2773  17.785  < 2e-16 ***
## bedrooms     53.7265    21.0655   2.550   0.0115 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 397.7 on 194 degrees of freedom
## Multiple R-squared:  0.621,  Adjusted R-squared:  0.6171
## F-statistic:   159 on 2 and 194 DF,  p-value: < 2.2e-16
```

Final Model Equation:

$$\hat{Y} = 459.8715 + 53.7265\text{Bedrooms} + 4.9318\text{Size}$$

Comparing R^2 and Adjusted R^2

```r
print(summary_model$r.squared)
```

```
## [1] 0.6210481
```

```r
print(summary_m2$r.squared)
```

```
## [1] 0.6210235
```

```r
print(summary_model$adj.r.squared)
```

```
## [1] 0.6151577
```

```r
print(summary_m2$adj.r.squared)
```

```
## [1] 0.6171165
```

- R^2 decreases slightly, showing reduced overall accuracy.
- Adjusted R^2 increases, indicating the model is now more efficient.