SHC 798 Assignment 1, 2025

Richard Lubega

2025-07-14

SHC 798 Assignment 1, 2025

Part 3: Simple regression

Question 2

```
# Load the housing.rda data file
load(file.choose())
head(housing) # View first few rows of the dataset
```

```
##   size  price
## 1  125 132358
## 2  139 153827
## 3  237 154427
## 4  152 143527
## 5  154 131707
## 6  248 159368
```

```
summary(housing) # Get an overview of the dataset
```

```
##       size           price
##  Min.   : 74.0   Min.   :109141
##  1st Qu.:128.0   1st Qu.:133684
##  Median :151.5   Median :146803
##  Mean   :158.3   Mean   :148389
##  3rd Qu.:182.2   3rd Qu.:159955
##  Max.   :286.0   Max.   :208648
```

```
str(housing)
```

```
## 'data.frame':    100 obs. of  2 variables:
##  $ size : num  125 139 237 152 154 248 170 102 121 130 ...
##  $ price: num  132358 153827 154427 143527 131707 ...
```

```
# use regression analysis to explore the relationship between house size and price.
cat("\n A SLR between house price and house size \n")
```

```
##
##  A SLR between house price and house size
```

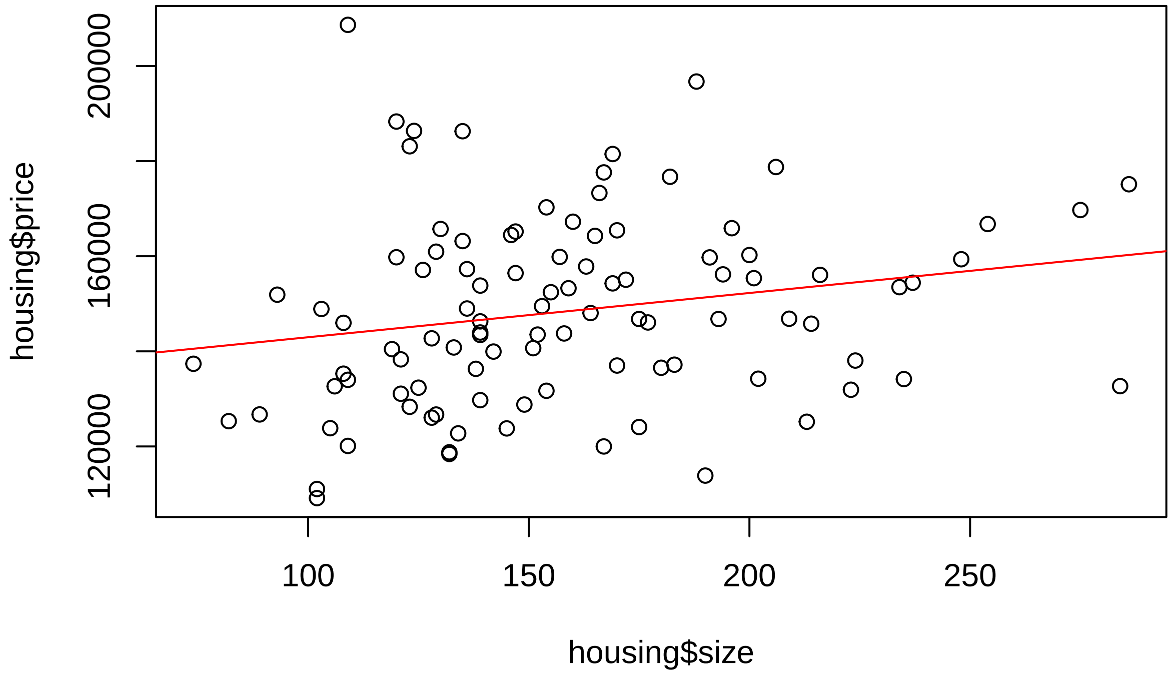
```r
lm_p.s <- lm(price ~ size, data = housing)
summary(lm_p.s)
```

```
##
## Call:
## lm(formula = price ~ size, data = housing)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -37465 -14199  -2284  11120  64842
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 133667.87    7271.00  18.384   <2e-16 ***
## size            93.01      44.26   2.102   0.0381 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19500 on 98 degrees of freedom
## Multiple R-squared:  0.04313,    Adjusted R-squared:  0.03336
## F-statistic: 4.417 on 1 and 98 DF,  p-value: 0.03814
```

```r
cat("\n ===  SLR Model Plot  === \n")
```

```
##
##  ===  SLR Model Plot  ===
```

```r
plot(housing$size, housing$price) +
  abline(lm_p.s, col = "red")
```



```
## integer(0)
```

(a) Comment on the Model Summary

Regression Coefficients:

- **Intercept** (133667.87): This means that for a theoretical house size of 0 units the predicted house price is 133667.87 units. This is not practically useful but ensures the regression line fits the data best within the observed size range. It is not meaningful to extrapolate to house size = 0.
  - It's *p-value* (< 2e-16) is small and statistically significant at the 5% level, but its practical value is limited.
- **Slope** (93.01): For every 1 unit increase in house size, the house price increases by about 93.01 units Bigger houses cost higher to buy.
  - The *p-value* (0.0381) is smaller than 0.05 (even at a 1% significance level). We reject the null hypothesis at the 5% significance level. This means that house size has a statistically significant effect on house price, and we can be fairly confident (with 95% confidence) that the relationship isn't due to chance.

Statistical Significance:

- From the F-statistic (4.417), the p-value for size is small (0.03814 < 0.05), meaning the model is statistically significant at the 5% level.

Model Goodness of Fit:

- The **multiple R-squared** (0.04313) and **adjusted R-squared** (0.03336) indicate that only 4.313% of the variability in house prices is explained by house size. This suggests the model is very weak in explanatory power and that other factors likely have a much bigger influence.

(b) Residual Diagnostics

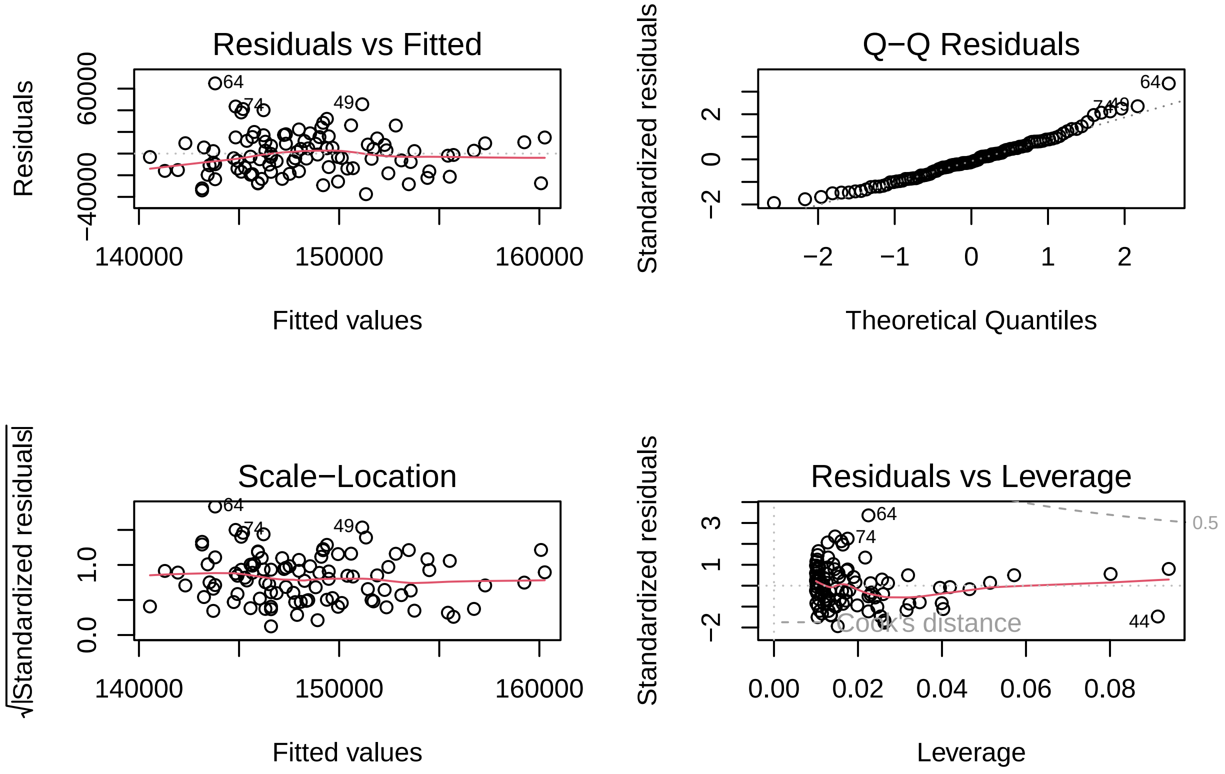
```r
# Perform residual diagnostics and comment on model assumptions
cat("Performing Model Diagnostics \n")
```

```
## Performing Model Diagnostics
```

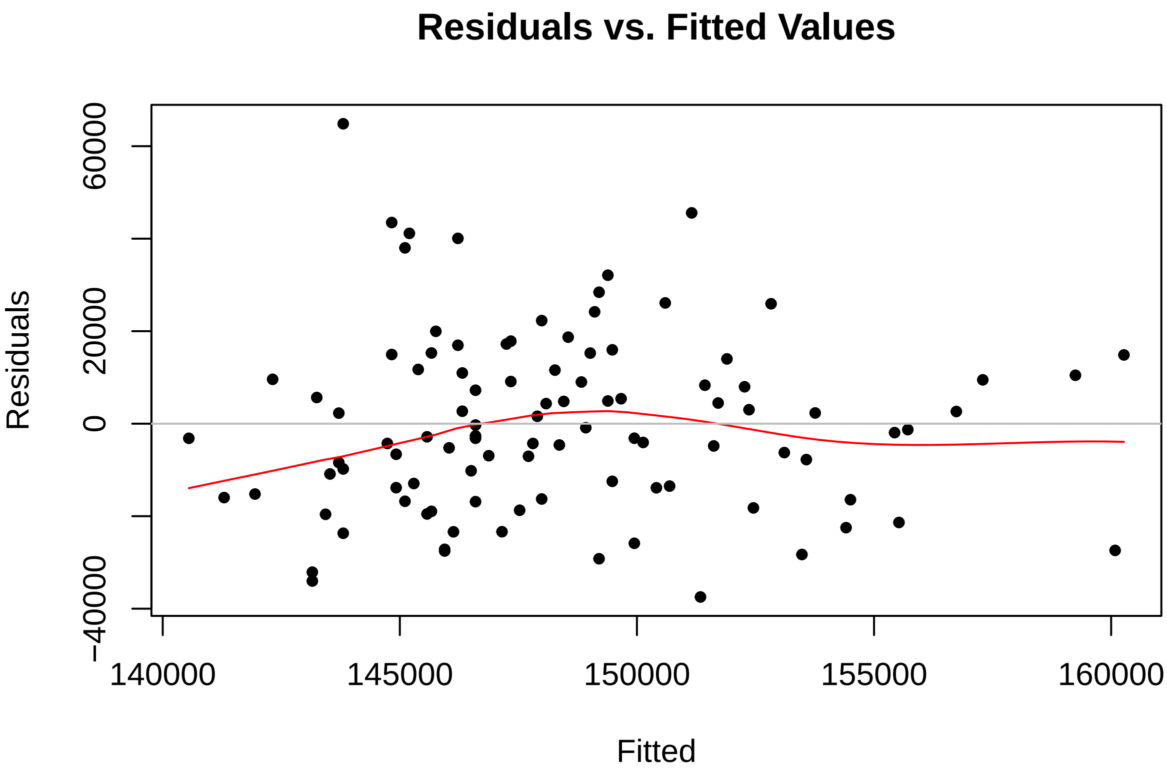
```r
cat("\n ===  Model Diagnostics Plots  === \n")
```

```
##
##  ===  Model Diagnostics Plots  ===
```

```r
# Diagnostics plots
par(mfrow = c(2,2))
plot(lm_p.s)
```

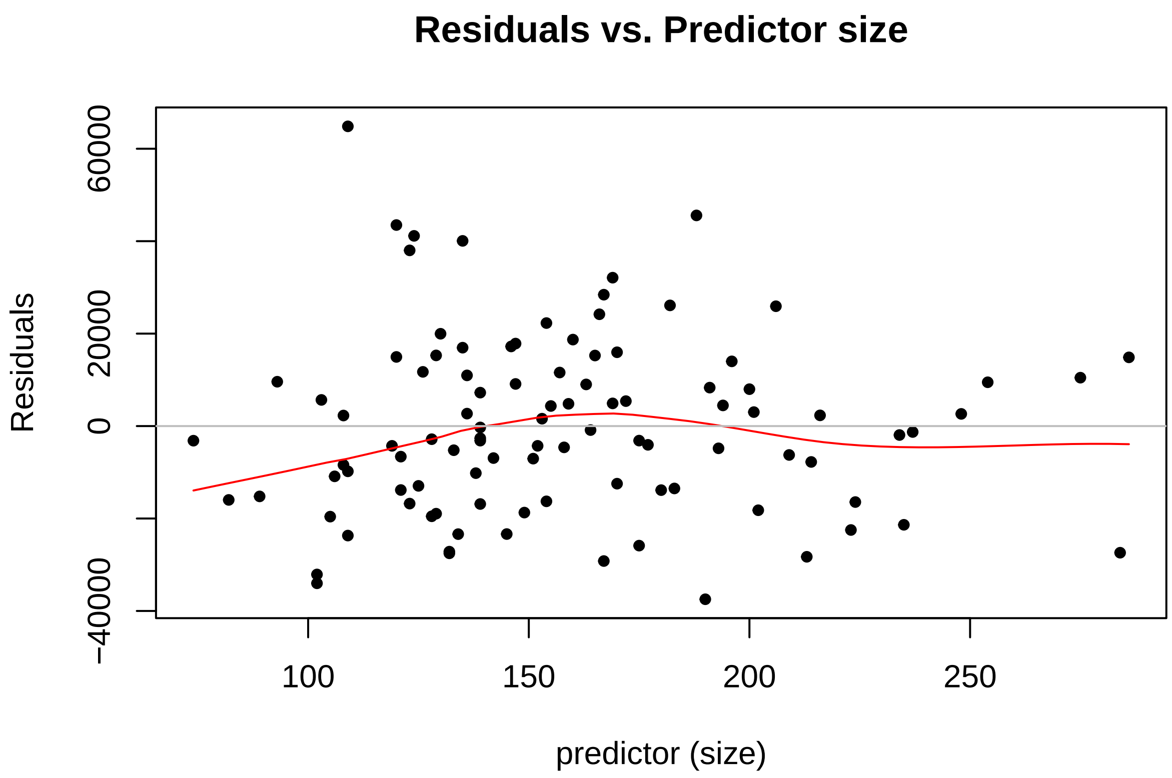


```r
par(mfrow = c(1,1))
# Tukey-Anscombe Plot
plot(lm_p.s$fitted.values, lm_p.s$residuals, xlab="Fitted", ylab="Residuals", pch=20) +
  title("Residuals vs. Fitted Values") +
  lines(loess.smooth(lm_p.s$fitted.values, lm_p.s$residuals),col="red") +
  abline(h=0, col="grey")
```



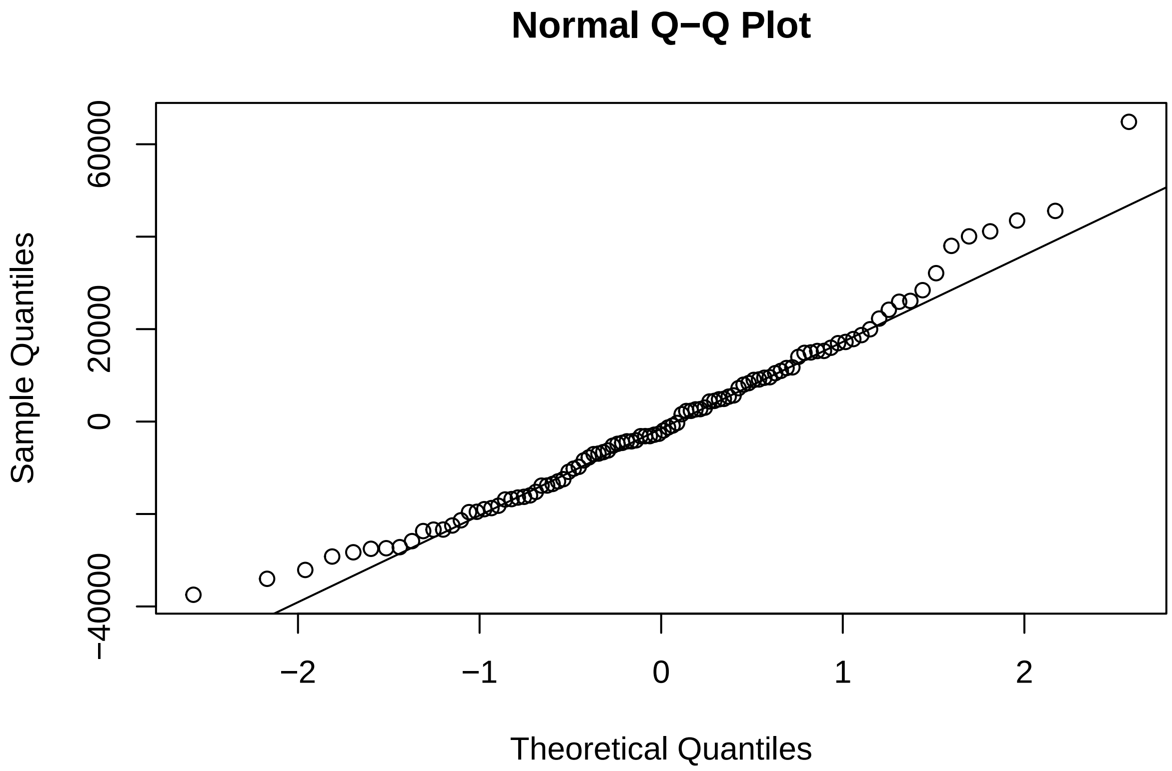
```
## integer(0)
```

```r
# Residuals vs. Predictor Plot
plot(housing$size, lm_p.s$residuals, xlab="predictor (size)", ylab="Residuals", pch=20) +
  title("Residuals vs. Predictor size") +
  lines(loess.smooth(housing$size, lm_p.s$residuals),col="red") +
  abline(h=0, col="grey")
```



```
## integer(0)
```

```
# Quantile-Quantile Plot
qqnorm(lm_p.s$residuals) #Quantile-Quantile Plot
qqline(lm_p.s$residuals) # adds the diagonal line
```



```
# Evaluating Model Assumptions
cat("Model Assumption Evaluation \n")
```

```
## Model Assumption Evaluation
```

Comments on Model Assumptions

1. **Linearity** — *From the Tukey-Anscombe Plot (Residuals vs. Fitted)*:

- From the plot, the residuals generally hover around the zero line which suggests that the $E[E_i] = 0$ is approximately met. However, LOESS smoother line has a kink in the middle and largely deviates from the horizontal. The residuals for low and high house size (and respective fitted house price) values are systematically negative and they are positive for medium values. The linearity assumption is violated and a straight line is not the correct fit to the data. The model may be improved by variable transformation.

2. **Homoskedasticity** — *From the Tukey-Anscombe plot and the Scale-Location Plot*:

- The Tukey-Anscombe plot indicates a more or less constant scatter for the entire range of house size (& fitted) values. There is no obvious violation of homoskedasticity. The red line in the Scale-Location Plot is fairly horizontal which implies constant variance with fitted values (no heteroscedasticity).

3. **Independence**

- The residuals can be considered independent and uncorrelated.

4. **Normality** — *From the Q-Q Plot*:

- The bulk of the residuals (in the central region) lie on the 45° line and thereby follow the Gaussian distribution. There are slight deviations (or outliers) at the lower and upper tail which indicate right skewness. The assumption of Gaussian errors is slightly violated by the model due to this moderate non-normality. Despite this, the approximation to normality in the center may be sufficient to validate this model.

(c) Check if a log transformation improves the model fit. Are any of the models useful?

```
# Any right-skewness in the data?
# View distribution (histogram)
cat("\n Viewing Paramter Distributions: Check for skewness\n")
```

```
##
##  Viewing Paramter Distributions: Check for skewness
```

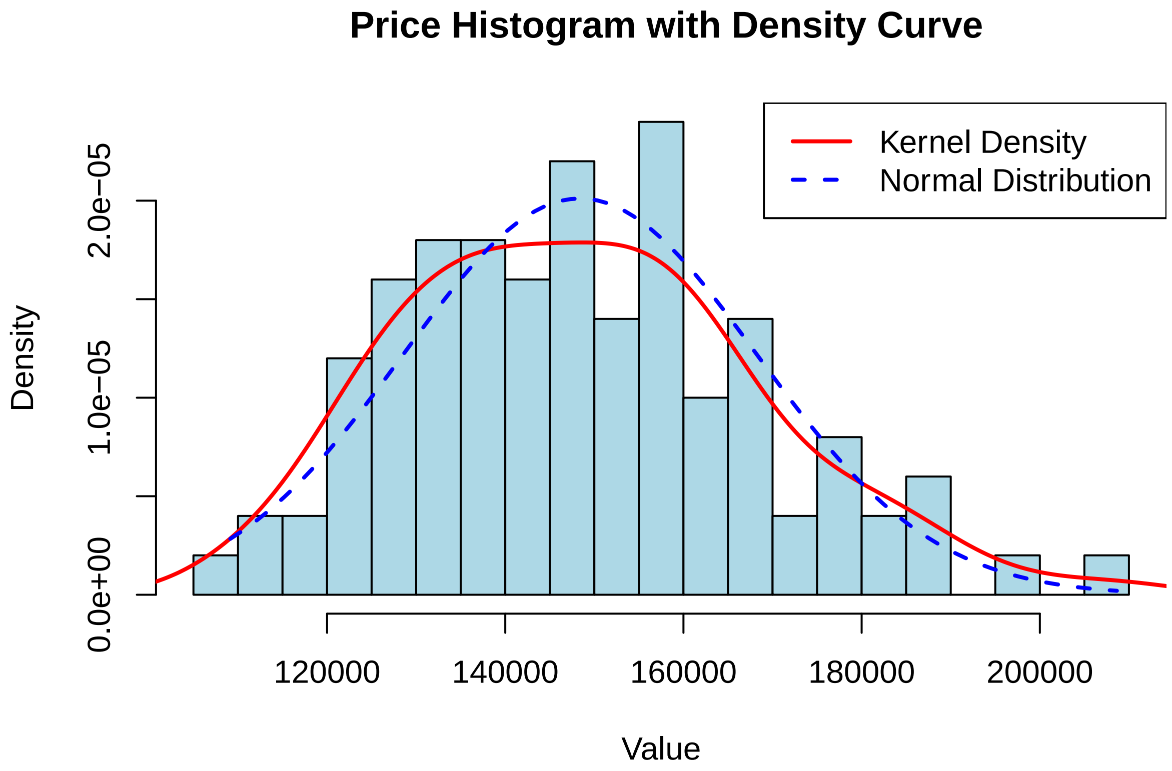
```
# Viewing House prices
hist(housing$price, freq = FALSE, breaks = 30, col = "lightblue",
     main = " Price Histogram with Density Curve", xlab = "Value", ylab = "Density",
     border = "black")

lines(density(housing$price, na.rm = TRUE), col = "red",lwd = 2) # Add density curve



# Adding a normal distribution curve for comparison
h_price <- seq(min(housing$price, na.rm = TRUE), max(housing$price, na.rm = TRUE), length.out = 100)
normal_price <- dnorm(h_price, mean = mean(housing$price, na.rm = TRUE), sd = sd(housing$price, na.rm =
lines(h_price, normal_price, col = "blue", lwd = 2, lty = 2)

# Add legend
legend("topright", legend = c("Kernel Density", "Normal Distribution"), col = c("red", "blue"),
       lwd = 2, lty = c(1, 2))
```

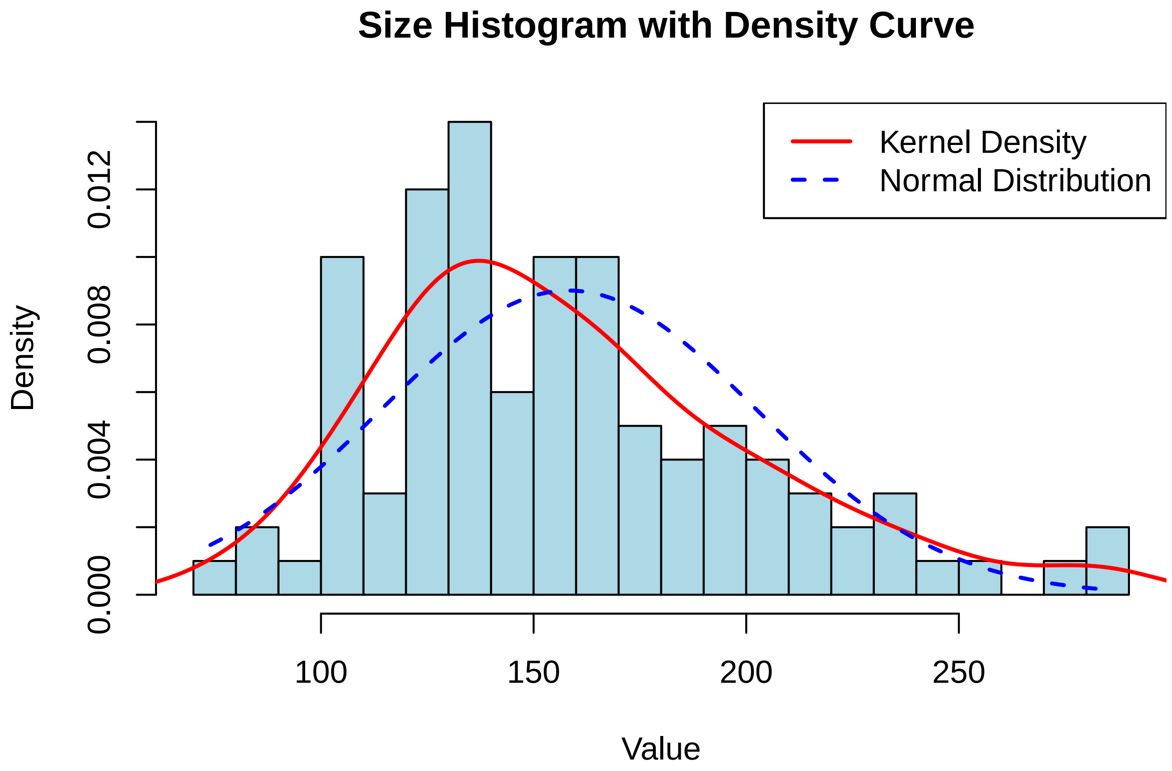


```r
# Viewing House Sizes
hist(housing$size, freq = FALSE, breaks = 30, col = "lightblue",
     main = " Size Histogram with Density Curve", xlab = "Value", ylab = "Density",
     border = "black")

lines(density(housing$size, na.rm = TRUE), col = "red",lwd = 2) # Add density curve

# Adding a normal distribution curve for comparison
h_size <- seq(min(housing$size, na.rm = TRUE), max(housing$size, na.rm = TRUE), length.out = 100)
normal_size <- dnorm(h_size, mean = mean(housing$size, na.rm = TRUE), sd = sd(housing$size, na.rm = TRU
lines(h_size, normal_size, col = "blue", lwd = 2, lty = 2)

# Add legend
legend("topright", legend = c("Kernel Density", "Normal Distribution"), col = c("red", "blue"),
       lwd = 2, lty = c(1, 2))
```



From the **plots**, the house **price** data is only *slightly right-skewed* (in agreement with the Q-Q plot) while the house **size** data is **clearly right-skewed**. A log-log transform is appropriate for the pair (which are right-skewed variables taking on only positive values).

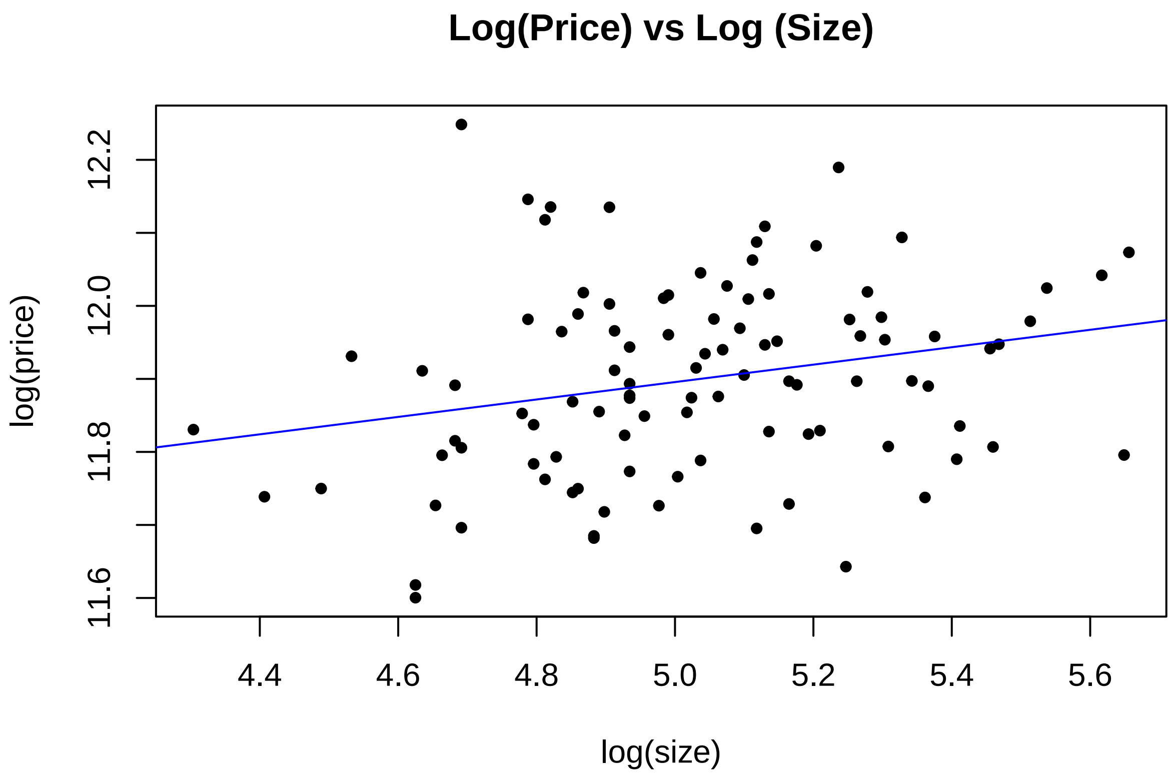
```
# 1. Log-log Transformation
lg.lg <- lm(log(price) ~ log(size), data = housing)
summary(lg.lg)
```

```
##
## Call:
## lm(formula = log(price) ~ log(size), data = housing)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.28226 -0.08861 -0.00627  0.08258  0.38958
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.29914    0.23830   47.42   <2e-16 ***
## log(size)    0.11930    0.04733    2.52   0.0133 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.129 on 98 degrees of freedom
## Multiple R-squared:  0.06088,    Adjusted R-squared:  0.05129
## F-statistic: 6.353 on 1 and 98 DF,  p-value: 0.01334
```

```
cat("\n ===  Log-log SLR Model Plot  === \n")
```

```
##
## ===  Log-log SLR Model Plot  ===
```

```
plot(log(price) ~ log(size), data = housing,  main = "Log(Price) vs Log (Size)", pch=20) +
  abline(lg.lg, col = "blue")
```



```
## integer(0)
```

```
# 2. Logged-Response Model Transformation
lm_lg <- lm(log(price) ~ size, data = housing)
summary(lm_lg)
```

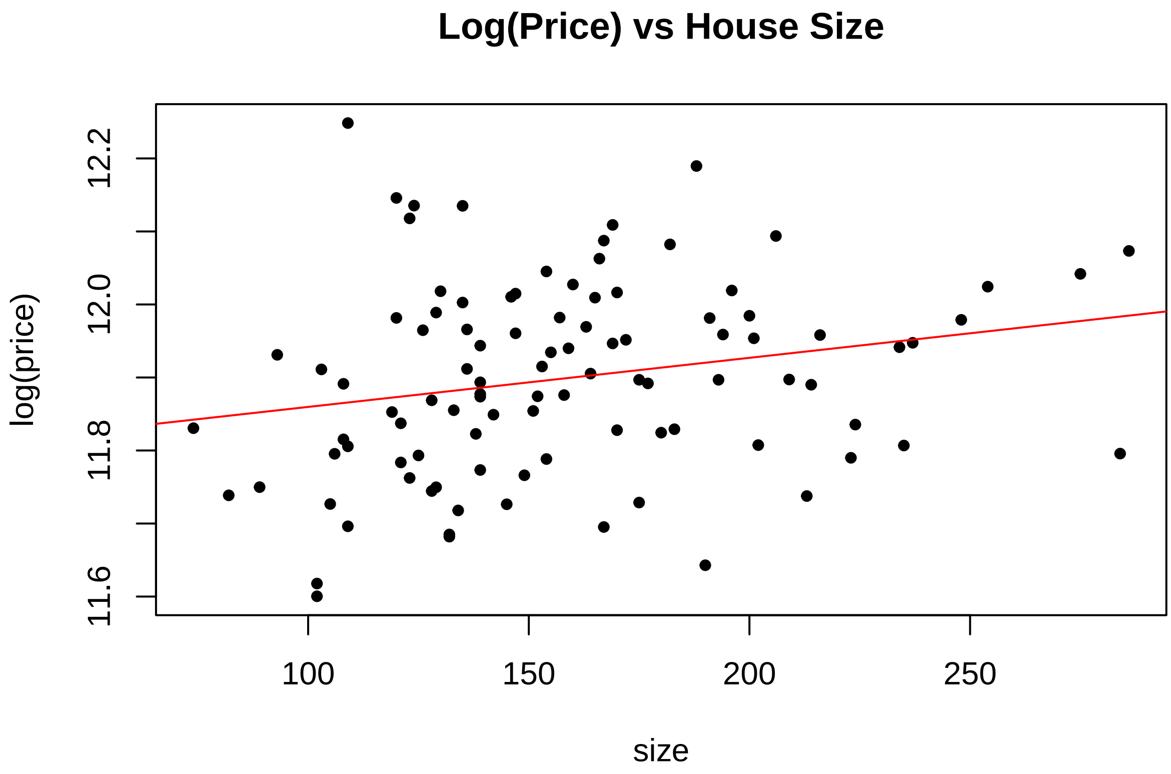
```
##
## Call:
## lm(formula = log(price) ~ size, data = housing)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.27734 -0.09326 -0.00847  0.08242  0.38265
##
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.179e+01  4.837e-02 243.821   <2e-16 ***
## size        6.722e-04  2.944e-04   2.283   0.0246 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1297 on 98 degrees of freedom
## Multiple R-squared:  0.05051,    Adjusted R-squared:  0.04082
## F-statistic: 5.213 on 1 and 98 DF,  p-value: 0.02457
```

```
cat("\n ===  Log SLR Model Plot  === \n")
```

```
##
## ===  Log SLR Model Plot  ===
```

```
plot(log(price) ~ size, data = housing, main = "Log(Price) vs House Size", pch=20) +
  # lines(loess.smooth(lm_lg$fitted.values, housing$size),col="red") +
  abline(lm_lg, col = "red")
```



```
## integer(0)
```

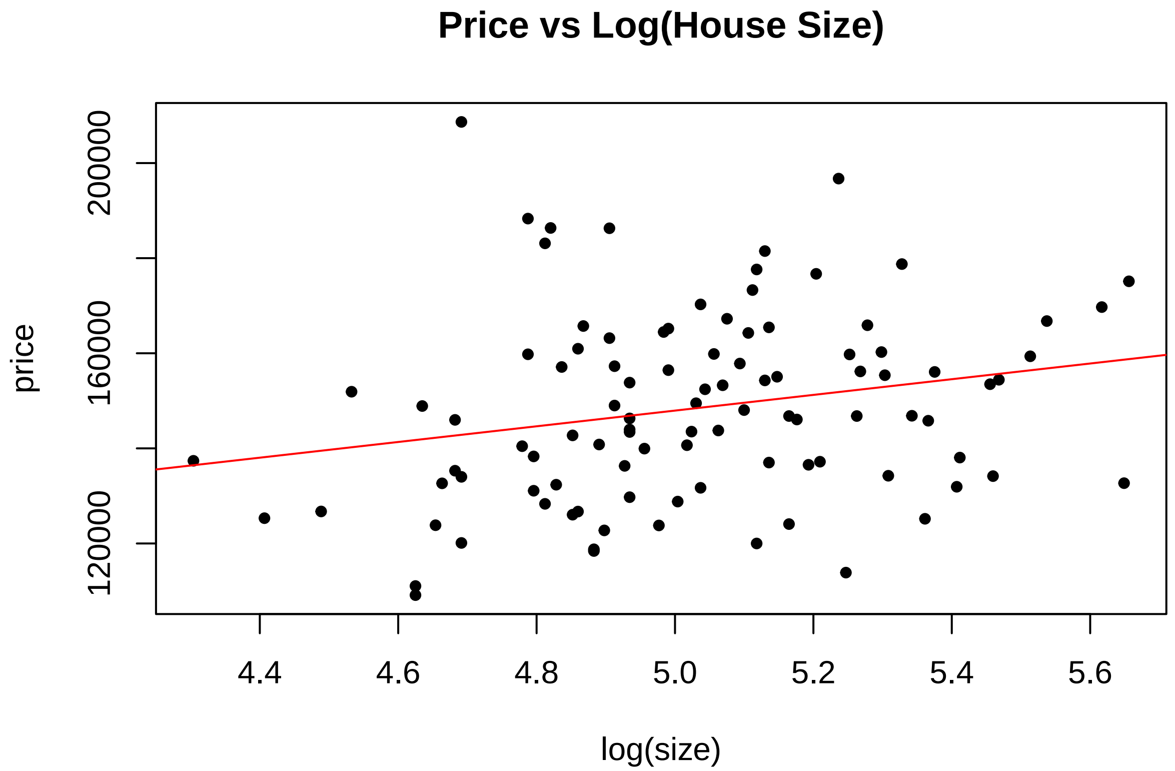
```
# 3.Transforming the Predictor
lm_lgh <- lm(price ~ log(size), data = housing)
summary(lm_lgh)
```

```
##
## Call:
## lm(formula = price ~ log(size), data = housing)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -38144 -13665  -1748  11615  65798
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    65431      35856   1.825   0.0711 .
## log(size)      16503       7122   2.317   0.0226 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19410 on 98 degrees of freedom
## Multiple R-squared:  0.05194,    Adjusted R-squared:  0.04226
## F-statistic: 5.369 on 1 and 98 DF,  p-value: 0.02258
```

```
cat("\n ===  Log SLR Model Plot  === \n")
```

```
##
##  ===  Log SLR Model Plot  ===
```

```
plot(price ~ log(size), data = housing, main = "Price vs Log(House Size)", pch=20) +
  # lines(loess.smooth(lm_lg$fitted.values, housing$size),col="red") +
  abline(lm_lgh, col = "red")
```



```
## integer(0)
```

```
cat("\n Evaluating Model Fit Improvements \n")
```

```
##
##  Evaluating Model Fit Improvements
```

Model Fit of the Log-log Model:

- The **multiple R-squared** (0.06088) and **adjusted R-squared** (0.05129) of the log-log model indicate that only 6.088% of the variability in house prices is explained by house size. This model is also very weak in explanatory power.

Model Fit of the Logged Response Model:

- The **multiple R-squared** (0.05051) and **adjusted R-squared** (0.04082) of the logged response model indicate that only 5.051% of the variability in house prices is explained by house size. This is also very weak model.

Model Fit of the Logged Predictor Model:

- The **multiple R-squared** (0.05194) and **adjusted R-squared** (0.04226) of this model suggest that only 5.194% of the variability in house prices can be explained by house size.

Are any of the models useful?

The residual diagnostics of all four models indicate no assumption violation. But all of them have very low model fits (explanatory power), i.e; 0.04313 (on the original scale), 0.06088 (for the log-log model), 0.05051 (for the logged response model) and 0.05194 (logged predictor model) , and thus **neither of them is useful**.