Exercise04

Manuel Bauder

```
#install.packages("lmtest")
#install.packages("ggplot2")

library(ggplot2)
library(lmtest)
```

```
Loading required package: zoo
```

```

Attaching package: 'zoo'
```

```
The following objects are masked from 'package:base':

    as.Date, as.Date.numeric
```

Linear Regression

Reading the data into R:

```
house.data <- read.csv("kc_house_data.csv")
```

(a)

Model Estimation

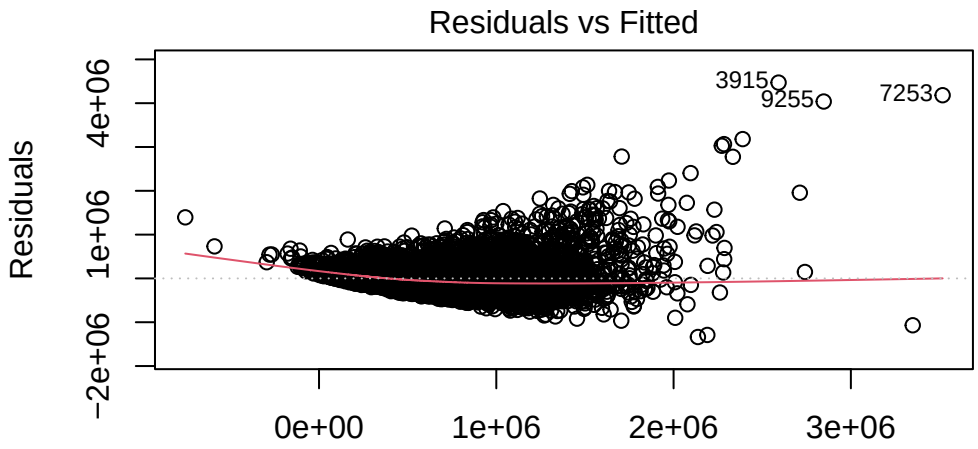
```
model_a <- lm(house.data$price ~ house.data$bedrooms + house.data$bathrooms + house.data$sqft
              + house.data$view + house.data$condition + house.data$grade + house.data$yr_bu

summary(model_a)
```

```
Call:
lm(formula = house.data$price ~ house.data$bedrooms + house.data$bathrooms +
    house.data$sqft_living + house.data$floors + house.data$view +
    house.data$condition + house.data$grade + house.data$yr_built)

Residuals:
     Min       1Q   Median       3Q      Max
-1337280  -111873   -10359    90133  4470268

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             6.356e+06  1.326e+05  47.950  < 2e-16 ***
house.data$bedrooms    -4.065e+04  2.066e+03 -19.671  < 2e-16 ***
house.data$bathrooms    4.769e+04  3.487e+03  13.677  < 2e-16 ***
house.data$sqft_living  1.693e+02  3.307e+00  51.199  < 2e-16 ***
house.data$floors       2.832e+04  3.496e+03   8.101 5.72e-16 ***
house.data$view         7.138e+04  2.107e+03  33.885  < 2e-16 ***
house.data$condition    1.815e+04  2.519e+03   7.205 6.01e-13 ***
house.data$grade        1.228e+05  2.187e+03  56.160  < 2e-16 ***
house.data$yr_built    -3.650e+03  6.824e+01 -53.484  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 221600 on 21604 degrees of freedom
Multiple R-squared:  0.6359,    Adjusted R-squared:  0.6358
F-statistic:  4717 on 8 and 21604 DF,  p-value: < 2.2e-16
```

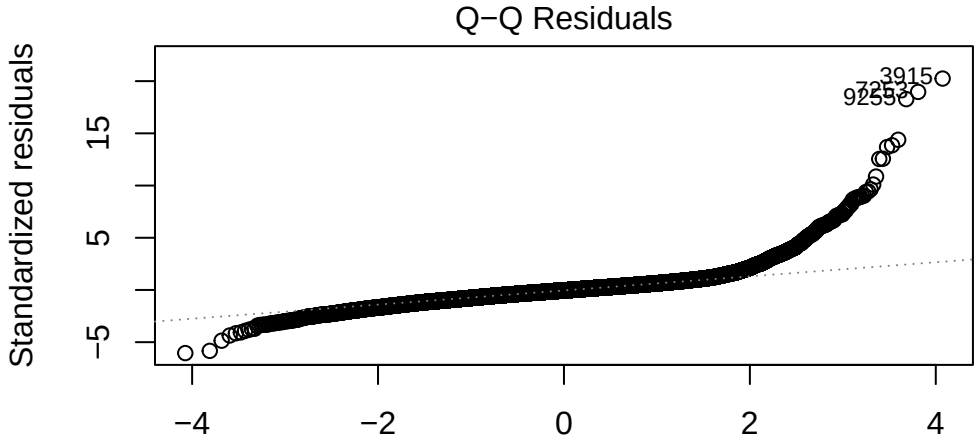
All variables in the model are significant. $\mathbf{R^2 = 0.6359}$, which indicates that $\mathbf{63.59\%}$ of the variance of the variable price is explained by the chosen covariates.

Residual Analysis

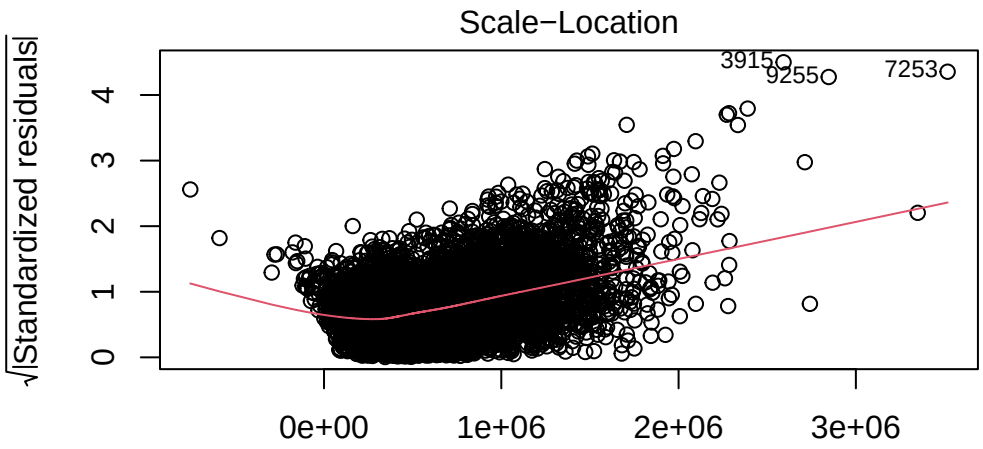
```
plot(model_a)
```

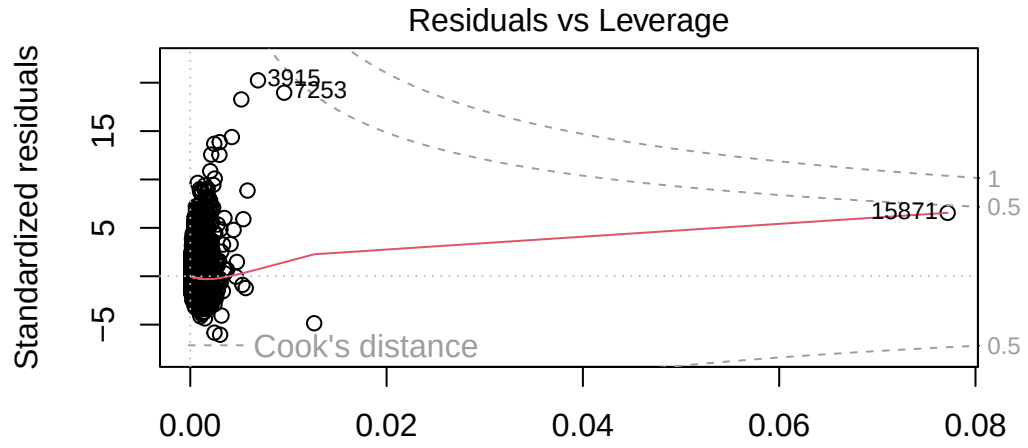


(house.data$price ~ house.data$bedrooms + house.data$bathrooms + hou



(house.data$price ~ house.data$bedrooms + house.data$bathrooms + hou





Linearity

The plot "Residuals vs. Fitted values" indicates if the relationship between the predictors and

the outcome variable is linear or not. The residuals should be randomly scattered around the horizontal line (y=0) with now visual pattern like a curve. That would indicate a non-linear relationship.

In this model, a linear relationship can be assumed.

Normality

The Q-Q Plot indicates if the residuals are normally distributed. For a normal distribution, the points should nearly follow the diagonal line.

In this model, this is not the case. In particular the right tail is deviating from the diagonal. Therefore, it can be assumed that the residuals are not normally distribute (positive skewness).

Linear Model Assumptions: The normality assumption is violated as the residuals are not normally distributed.

Homoscedasticity

The plot "Standardized residuals vs. Fitted values" indicates homoscedasticity. The variance of the residuals should be roughly the same for all levels of fitted values.

In this model, the variance increases with higher values for the fitted values. Therefore, heteroscedasticity can be assumed.

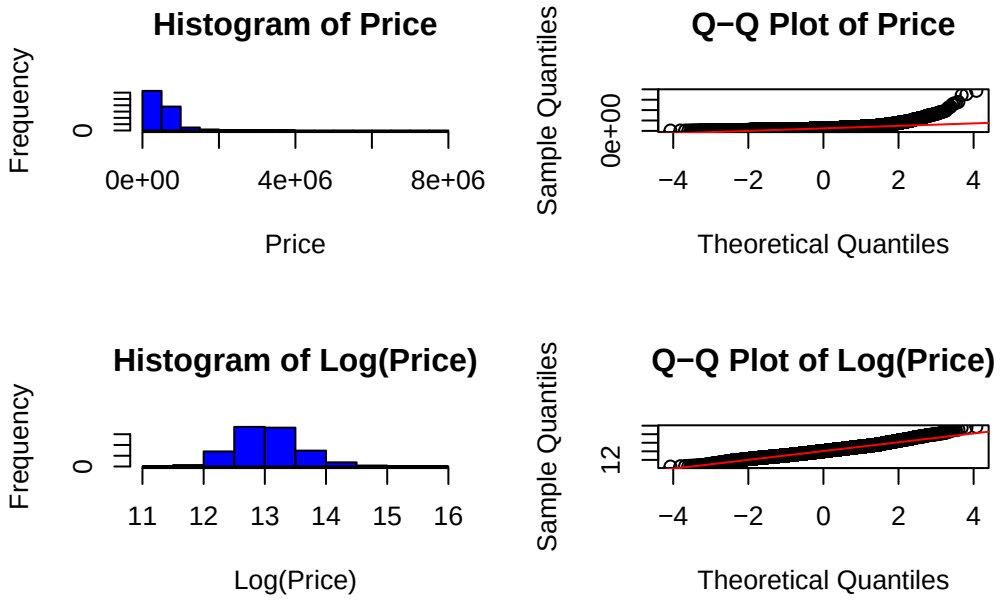
Linear Model Assumptions: The homoscedasticity assumption is violated, indicating heteroscedasticity.

b)

QQ-Plot and Histogramm

```
par(mfrow = c(2, 2))

# Plot Histogram of price
hist(house.data$price, main = "Histogram of Price", xlab = "Price", col = "blue", border = "

# Plot Q-Q Plot of price
qqnorm(house.data$price, main = "Q-Q Plot of Price")
qqline(house.data$price, col = "red")

# Log-transform the price
```

```r
log_price <- log(house.data$price)

# Plot Histogram of log(price)
hist(log_price, main = "Histogram of Log(Price)", xlab = "Log(Price)", col = "blue", border =

# Plot Q-Q Plot of log(price)
qqnorm(log_price, main = "Q-Q Plot of Log(Price)")
qqline(log_price, col = "red")
```




```r
# Reset the plotting area to default (optional)
par(mfrow = c(1, 1))
```

Observation: In the Histogram of price it becomes visible that most of the data points are concentrated at the lower end of the price scale, indicating that low prices are more common. The Q-Q plot of price shows that the points at the right end of the tail (representing the higher price values) are far away from the diagonal line, indicating that the actual distribution has heavier tails than the normal distribution.

Interpretation: The variable price is positively skewed. The log transformation of price is reducing skewness.

Log transformation of the model

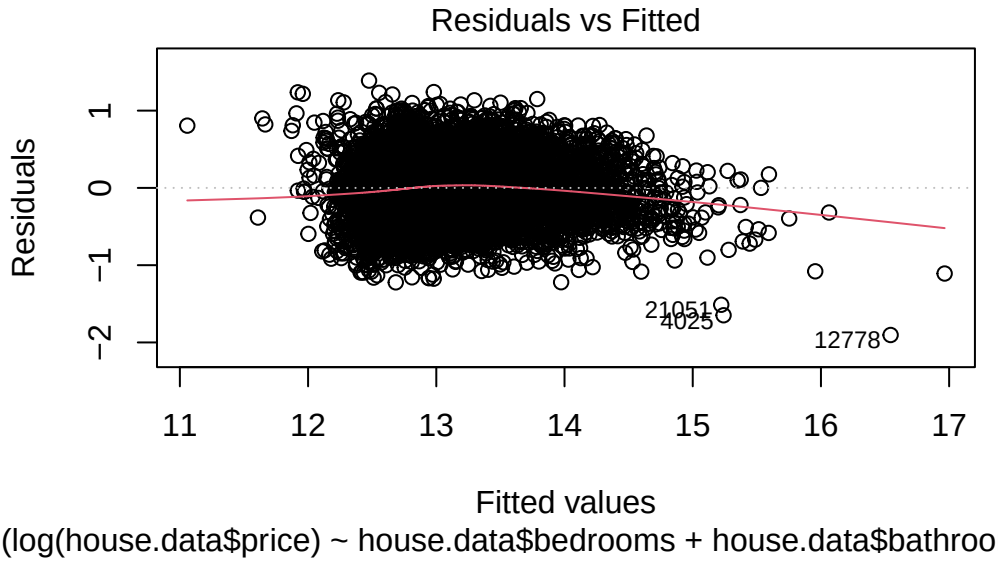
```
model_b <- lm(log(house.data$price) ~ house.data$bedrooms + house.data$bathrooms + house.data
              + house.data$view + house.data$condition + house.data$grade + house.data$yr_bu

summary(model_b)
```

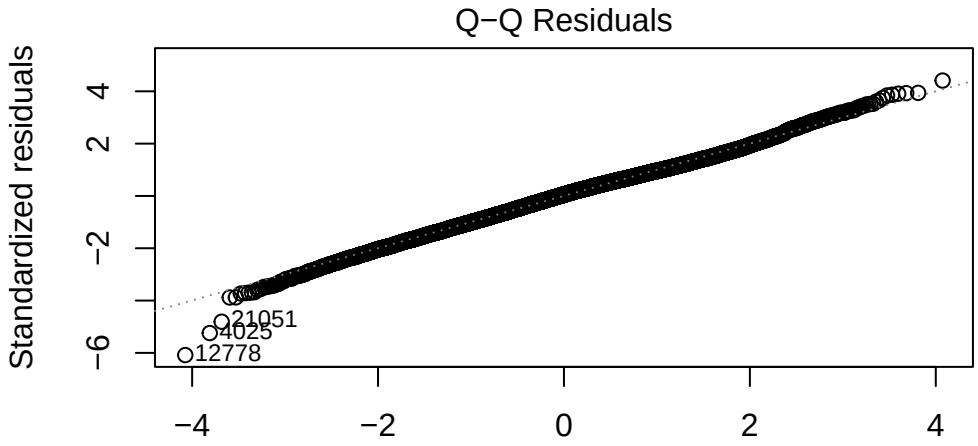
```
Call:
lm(formula = log(house.data$price) ~ house.data$bedrooms + house.data$bathrooms +
    house.data$sqft_living + house.data$floors + house.data$view +
    house.data$condition + house.data$grade + house.data$yr_built)

Residuals:
     Min       1Q   Median       3Q      Max
-1.90374 -0.21157  0.01624  0.21288  1.38880

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)             2.151e+01  1.884e-01 114.155  < 2e-16 ***
house.data$bedrooms    -2.366e-02  2.937e-03  -8.056 8.28e-16 ***
house.data$bathrooms    8.500e-02  4.956e-03  17.152  < 2e-16 ***
house.data$sqft_living  1.664e-04  4.701e-06  35.402  < 2e-16 ***
house.data$floors       8.569e-02  4.968e-03  17.246  < 2e-16 ***
house.data$view         6.740e-02  2.994e-03  22.510  < 2e-16 ***
house.data$condition    4.226e-02  3.580e-03  11.803  < 2e-16 ***
house.data$grade        2.218e-01  3.108e-03  71.359  < 2e-16 ***
house.data$yr_built    -5.526e-03  9.699e-05 -56.980  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3149 on 21604 degrees of freedom
Multiple R-squared:  0.6426,    Adjusted R-squared:  0.6425
F-statistic:  4856 on 8 and 21604 DF,  p-value: < 2.2e-16
```

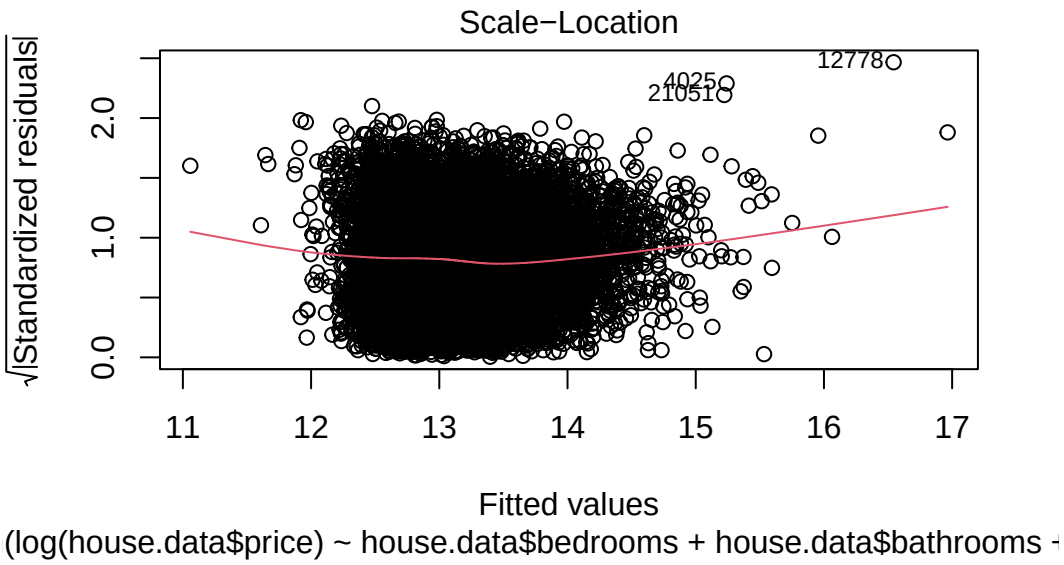
All variables in the model are significant. $\mathbf{R^2 = 0.6426}$, which indicates that **64.25%** of the variance of the variable price is explained by the chosen covariates. This R^2 is higher than in the model where price was not log-transformed. Therefore, in the second model more variance is explained by the predictor variables.

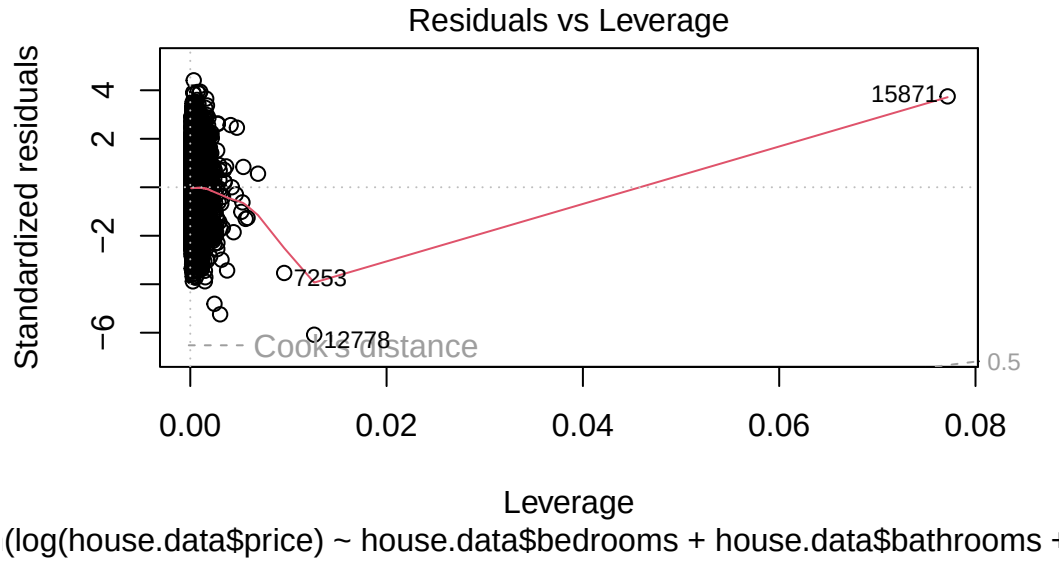
Residual Analysis

```
plot(model_b)
```









The log transformation of price mitigated the effects of positive skewness. The residuals are now more closer to a normal distribution (QQ-plot) and heteroscedasticity is reduced (Standardized residuals vs. Scale-Location plot). Normality and Homoscedasticity assumptions are no longer violated.

This, and a higher R^2, is a strong indication that the log transformation of price was beneficial for modeling the relationship. The second model is more adequate.

c)

Interpretation of parameters

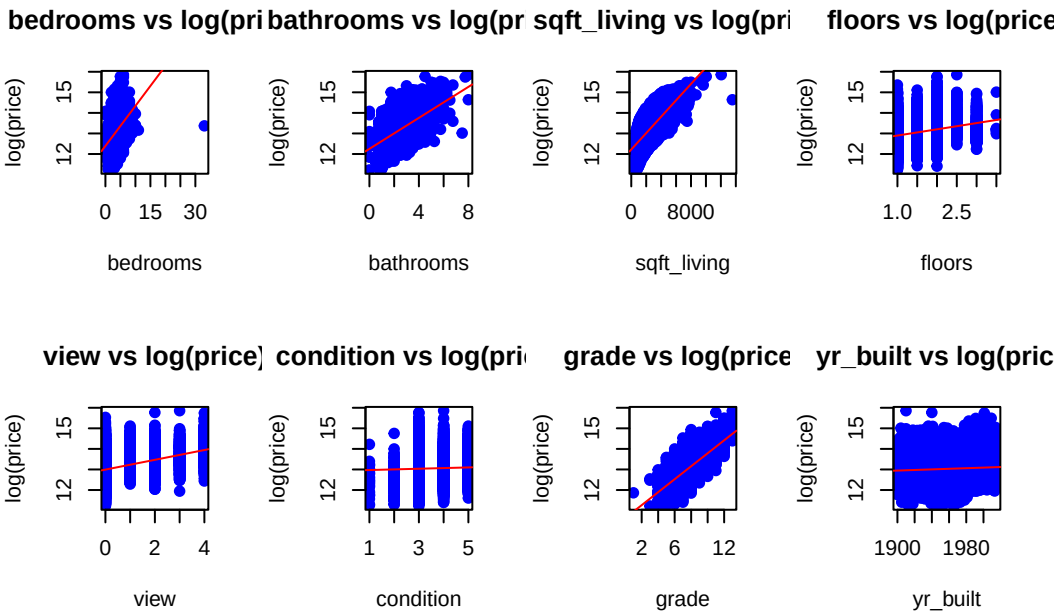
In this model, predictors such as the number of bedrooms and the year built have a negative impact on the log-transformed house prices, while factors such as the number of bathrooms, square footage of living space, number of floors, view rating, condition, and grade have a positive impact.

Price decreasing factors: - For every additional bedroom, the price decreases by approximately 2.366%. - For every additional year, the price decreases by approximately 0.5526%.

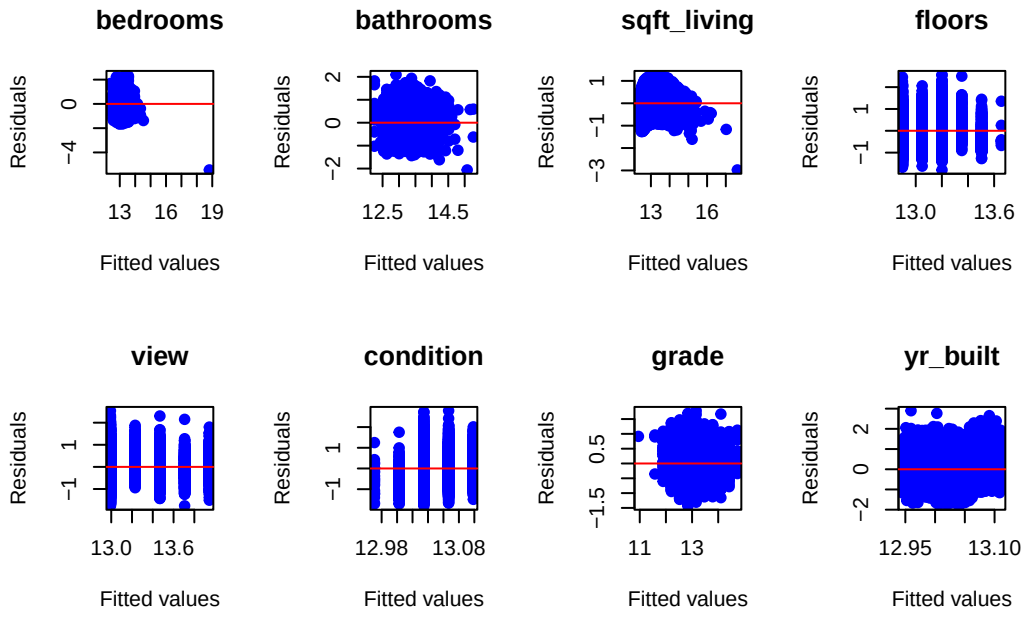
Price increasing factors - For every additional bathroom, the price increases by approximately 8.5%. - For every additional square foot of living space, the price increases by approximately 0.01664%. - For every additional floor, the price increases by approximately 8.569%. - For every unit increase in the view rating, the price increases by approximately 6.74%. - For every unit increase in the condition rating, the price increases by approximately 4.226%. - For every unit increase in the grade rating, the price increases by approximately 22.18%

Linear dependence log(price) ~ covariates

```
par(mfrow = c(2, 4))
covariates <- c("bedrooms", "bathrooms", "sqft_living", "floors", "view", "condition", "grade

for (covariate in covariates) {
  plot(house.data[[covariate]], log(house.data$price),
       main = paste(covariate, "vs log(price)"),
       xlab = covariate, ylab = "log(price)",
       pch = 19, col = "blue")
  abline(lm(log(price) ~ house.data[[covariate]], data = house.data), col = "red")
}
```




```
par(mfrow = c(2, 4))
for (covariate in covariates) {
  model_single <- lm(log(price) ~ house.data[[covariate]], data = house.data)
  plot(model_single$fitted.values, model_single$residuals,
       main = paste(covariate),
       xlab = "Fitted values", ylab = "Residuals",
       pch = 19, col = "blue")
  abline(h = 0, col = "red")
}
```




```
log_price <- log(house.data$price)
covariates <- c("bedrooms", "bathrooms", "sqft_living", "floors", "view", "condition", "grade

correlation_results <- sapply(covariates, function(covariate) {
  cor(house.data[[covariate]], log_price)
})

correlation_results
```

```
   bedrooms   bathrooms sqft_living      floors        view   condition
 0.34356114  0.55080193  0.69534060  0.31055811  0.34652193  0.03955750
      grade    yr_built
 0.70363412  0.08065456
```

Checking the scattered plots with the covariates against the log(price), the residual plots, and the correlation coefficients (cor) are indicating that: - bathrooms, sqft_living, and grade have a clear linear dependence to log(price). - bedrooms, floors, and view have a slight linear dependence to log(price). - condition and yr_built do not have a linear dependence to log(price).

```r
house.data$yr_built_sq <- house.data$yr_built^2
house.data$sqft_living_sq <- house.data$sqft_living^2

model_c <- lm(log(house.data$price) ~ house.data$bedrooms + house.data$bathrooms + house.data
              house.data$floors + house.data$view + house.data$condition + house.data$grade +
summary(model_c)
```

```
Call:
lm(formula = log(house.data$price) ~ house.data$bedrooms + house.data$bathrooms +
    house.data$sqft_living + I(house.data$sqft_living^2) + house.data$floors +
    house.data$view + house.data$condition + house.data$grade +
    house.data$yr_built + I(house.data$yr_built^2))

Residuals:
    Min      1Q  Median      3Q     Max
-1.2644 -0.2113  0.0138  0.2107  1.4160

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  1.674e+02  1.082e+01  15.470   <2e-16 ***
house.data$bedrooms         -2.978e-02  3.009e-03  -9.895   <2e-16 ***
house.data$bathrooms         7.317e-02  4.959e-03  14.754   <2e-16 ***
house.data$sqft_living       2.792e-04  8.690e-06  32.132   <2e-16 ***
I(house.data$sqft_living^2) -1.782e-08  1.172e-09 -15.212   <2e-16 ***
house.data$floors            4.733e-02  5.613e-03   8.432   <2e-16 ***
house.data$view              7.222e-02  2.979e-03  24.247   <2e-16 ***
house.data$condition         4.640e-02  3.591e-03  12.920   <2e-16 ***
house.data$grade             2.176e-01  3.092e-03  70.355   <2e-16 ***
house.data$yr_built         -1.545e-01  1.106e-02 -13.978   <2e-16 ***
I(house.data$yr_built^2)     3.802e-05  2.823e-06  13.468   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.312 on 21602 degrees of freedom
Multiple R-squared:  0.6492,    Adjusted R-squared:  0.649
F-statistic:  3998 on 10 and 21602 DF,  p-value: < 2.2e-16
```

Both squared terms are significant. R^2 increased from 0.6426 to 0.6492.

d)

Model comparison b) and c)

Divide house.data randomly into training and test set:

```
# Set seed for reproducibility
set.seed(1122)

# Total number of observations
n <- nrow(house.data)

# Number of observations for the training set
n_train <- 10806

# Randomly sample indices for the training set
train_indices <- sample(1:n, n_train)

# Create the training set
train_set <- house.data[train_indices, ]

# Create the test set with the remaining indices
test_set <- house.data[-train_indices, ]
```

Fit both models on the training set and make predictions on the test set

```
# Model Fit
model_b_train <- lm(log(price) ~ bedrooms + bathrooms + sqft_living + floors
              + view + condition + grade + yr_built, data = house.data)

model_c_train <- lm(log(price) ~ bedrooms + bathrooms + sqft_living + I(sqft_living^2) +
              floors + view + condition + grade + yr_built + I(yr_built^2), data = house.data

# Predictions
predictions_b <- predict(model_b_train, newdata = test_set)
predictions_c <- predict(model_c_train, newdata = test_set)

# Calculate mean squared difference
msd_b <- mean((predictions_b- log(test_set$price))^2)
msd_c <- mean((predictions_c - log(test_set$price))^2)

cat("Mean Squared Difference for Model 1:", msd_b, "\n")
```

```
Mean Squared Difference for Model 1: 0.09758539
```

```
cat("Mean Squared Difference for Model 2:", msd_c, "\n")
```

```
Mean Squared Difference for Model 2: 0.09630774
```

In this comparison, model_c has a lower prediction error.

Improve prediction

Continue here