Regression analysis and diagnostics for Albuquerque housing prices

AUTHOR
Steve Simon and Leroy Wheeler

PUBLISHED
August 18, 2024

This program reads data on housing prices in Albuquerque, New Mexico in 1993. Find more information in the data dictionary.

This code is placed in the public domain.

Load the tidyverse library

For most of your programs, you should load the tidyverse library. The broom package provides a nice way to compute residuals and predicted values. The messages and warnings are suppressed.

```
library(broom)
library(tidyverse)
```

Read the data and view a brief summary

Use the read_csv function to read the data. The glimpse function will produce a brief summary.

```
alb <- read_csv(
  file="../data/albuquerque-housing.csv",
  col_names=TRUE,
  col_types="nnnnccc",
  na=".")
glimpse(alb)
```

```
Rows: 117
Columns: 7
$ price        <dbl> 205000, 208000, 215000, 215000, 199900, 190000, 180000, 1…
$ sqft         <dbl> 2650, 2600, 2664, 2921, 2580, 2580, 2774, 1920, 2150, 171…
$ age          <dbl> 13, NA, 6, 3, 4, 4, 2, 1, NA, 1, 4, 8, 15, 14, 18, NA, 16…
$ features     <dbl> 7, 4, 5, 6, 4, 4, 4, 5, 4, 3, 5, 6, 3, 5, 8, 3, 4, 3, 4, …
$ northeast    <chr> "yes", "yes", "yes", "yes", "yes", "yes", "yes", "yes", "…
$ custom_build <chr> "yes", "yes", "yes", "yes", "yes", "no", "no", "yes", "no…
$ corner_lot   <chr> "no", "no", "no", "no", "no", "no", "no", "no", "no", "no…
```

m1: regression analysis using square feet to predict price

You might expect that a house with more square feet would have a higher sales price. Your first steps are to compute simple descriptive statistics for both the independent variable (features) and the dependent variable (price). Then you should plot the data.

m1: Question 1: Calculate descriptive statistics (mean, standard deviation, minimum, and maximum for sqft. Interpret these numbers

```
     alb |>
       summarise(
         sqft_mn=mean(sqft, na.rm=TRUE),
         sqft_sd=sd(sqft, na.rm=TRUE),
         sqft_min=min(sqft, na.rm=TRUE),
         sqft_max=max(sqft, na.rm=TRUE),
         n_missing=sum(is.na(sqft)))
```

```
# A tibble: 1 × 5
  sqft_mn sqft_sd sqft_min sqft_max n_missing
    <dbl>   <dbl>    <dbl>    <dbl>     <int>
1   1654.    524.      837     3750         0
```

The average square footage of the homes in this data set are small compared to today’s standards. The the standard deviation of 524 square feet indicates little variation. There are no missing values in this data set of 117 houses.

m1: Calculate descriptive statistics for price

```
alb |>
  summarize(
    price_mn=mean(price, na.rm=TRUE),
    price_sd=sd(price, na.rm=TRUE),
    price_min=min(price, na.rm=TRUE),
    price_max=max(price, na.rm=TRUE),
    n_missing=sum(is.na(price)))
```

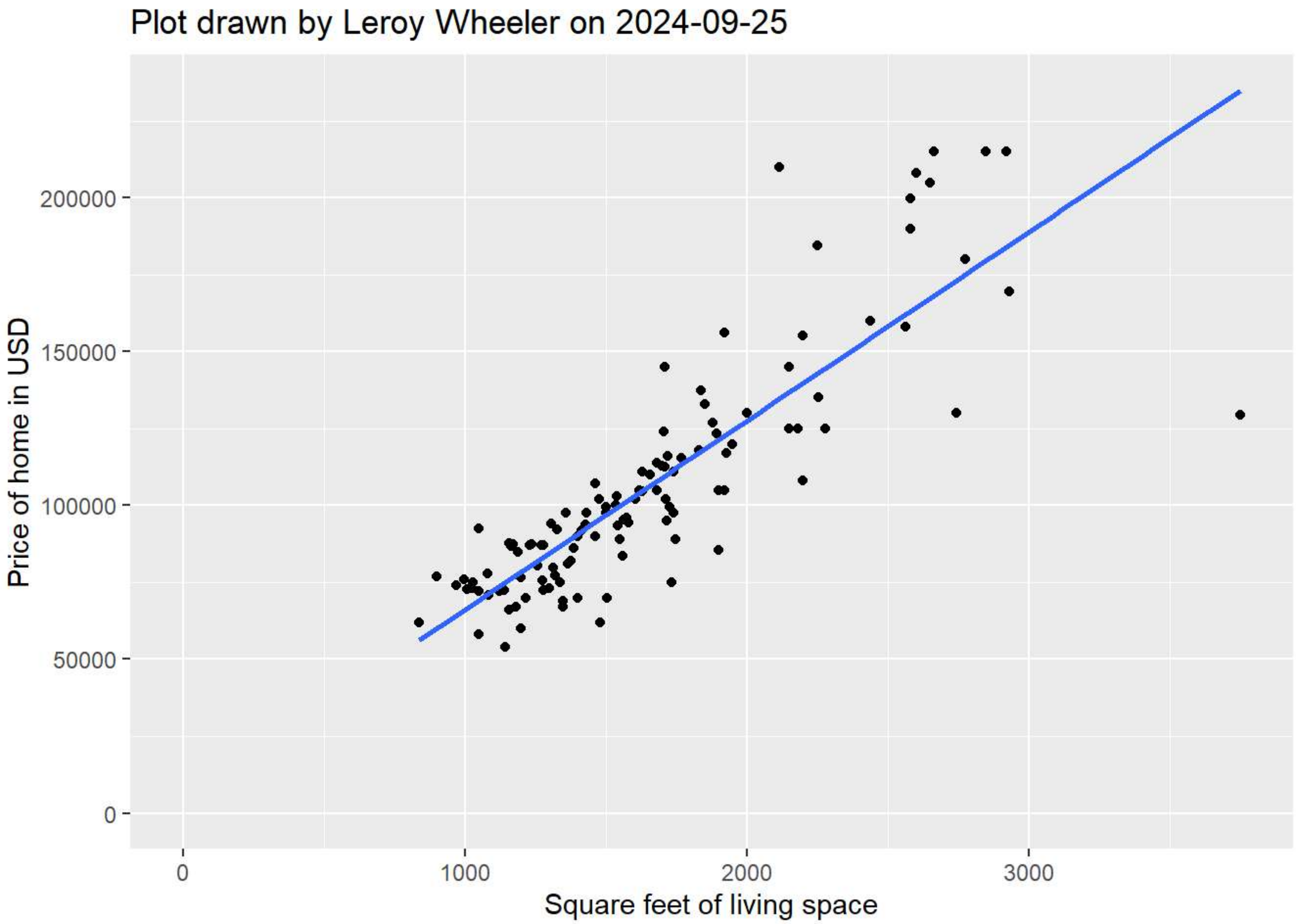
```
# A tibble: 1 × 5
  price_mn price_sd price_min price_max n_missing
     <dbl>    <dbl>     <dbl>     <dbl>     <int>
1  106274.   38044.     54000    215000         0
```

The average price is low 106,000 USD but the standard deviation 38,000 USD shows a fair amount of variation.

m1: Question 2: Draw a plot with price on the y-axis and sqft on the x-axis. Include a linear regression line, but do not extend it beyond the range of the data. Interpret this plot

```
alb |>
  ggplot(aes(sqft, price)) +
    geom_point() +
    geom_smooth(method="lm", se=FALSE) +
      ggtitle("Plot drawn by Leroy Wheeler on 2024-09-25") +
      xlab("Square feet of living space") +
      ylab("Price of home in USD")+
  expand_limits(x=0,y=0)
```

```
`geom_smooth()` using formula = 'y ~ x'
```



There is a strong positive relationship between the square feet of home and the listed price.

m1: Question 3: Calculate a linear regression model using sqft to predict price. Interpret the slope and intercept

```
m1 <- lm(price~sqft, data=alb)
m1
```

```
Call:
lm(formula = price ~ sqft, data = alb)

Coefficients:
(Intercept)         sqft
    4781.93        61.37
```

The estimated average sales price for a house with no living space (i.e. zero square feet) is about 4800 USD. This value of the y-intercept is meaningless because I assume buyers will be only willing to pay for a home with some indoor living space, unless they are looking for an empty lot to build a house on. As seen from the slope of this linear model also predicts that the buyers will have to pay approximately 61 USD for each square foot of home. Compared to today's standards, this is pretty cheap.

Skip some of the functions for hypothesis tests and p-values

Normally, you would follow this up with various functions like anova(), confint(), or tidy(). This program skips those steps to focus on the diagnostic plots of the residuals.

m1: Calculate residuals and predicted values

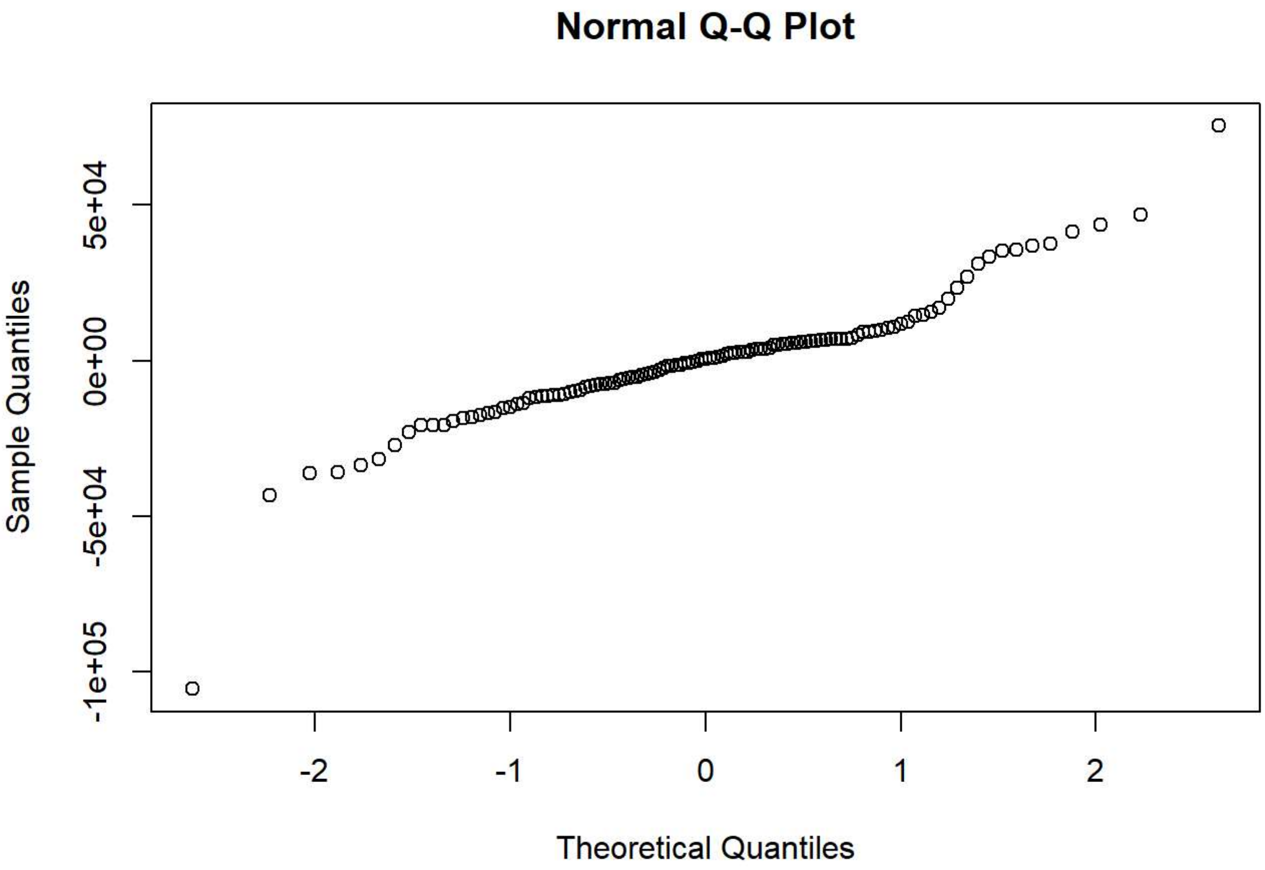
```
r1 <- augment(m1)
glimpse(r1)
```

```
Rows: 117
Columns: 8
$ price      <dbl> 205000, 208000, 215000, 215000, 199900, 190000, 180000, 156…
$ sqft       <dbl> 2650, 2600, 2664, 2921, 2580, 2580, 2774, 1920, 2150, 1710,…
$ .fitted    <dbl> 167403.63, 164335.30, 168262.77, 184034.01, 163107.97, 1631…
$ .resid     <dbl> 37596.3651, 43664.6991, 46737.2315, 30965.9946, 36792.0327,…
$ .hat       <dbl> 0.039734786, 0.036682514, 0.040617583, 0.059012195, 0.03550…
$ .sigma     <dbl> 20217.75, 20107.41, 20042.38, 20315.77, 20232.60, 20373.81,…
$ .cooksd    <dbl> 7.285696e-02, 9.015129e-02, 1.153048e-01, 7.644246e-02, 6.1…
$ .std.resid <dbl> 1.87655228, 2.17598631, 2.33387459, 1.56136142, 1.83237492,…
```

You could have also used the resid() and predict() functions. No interpretation is needed here, as these numbers are better reviewed using various graphical displays.

m1: Question 4a: Draw a normal probability plot for the residuals (.resid). Interpret these plots.

```
qqnorm(r1$.resid)
```

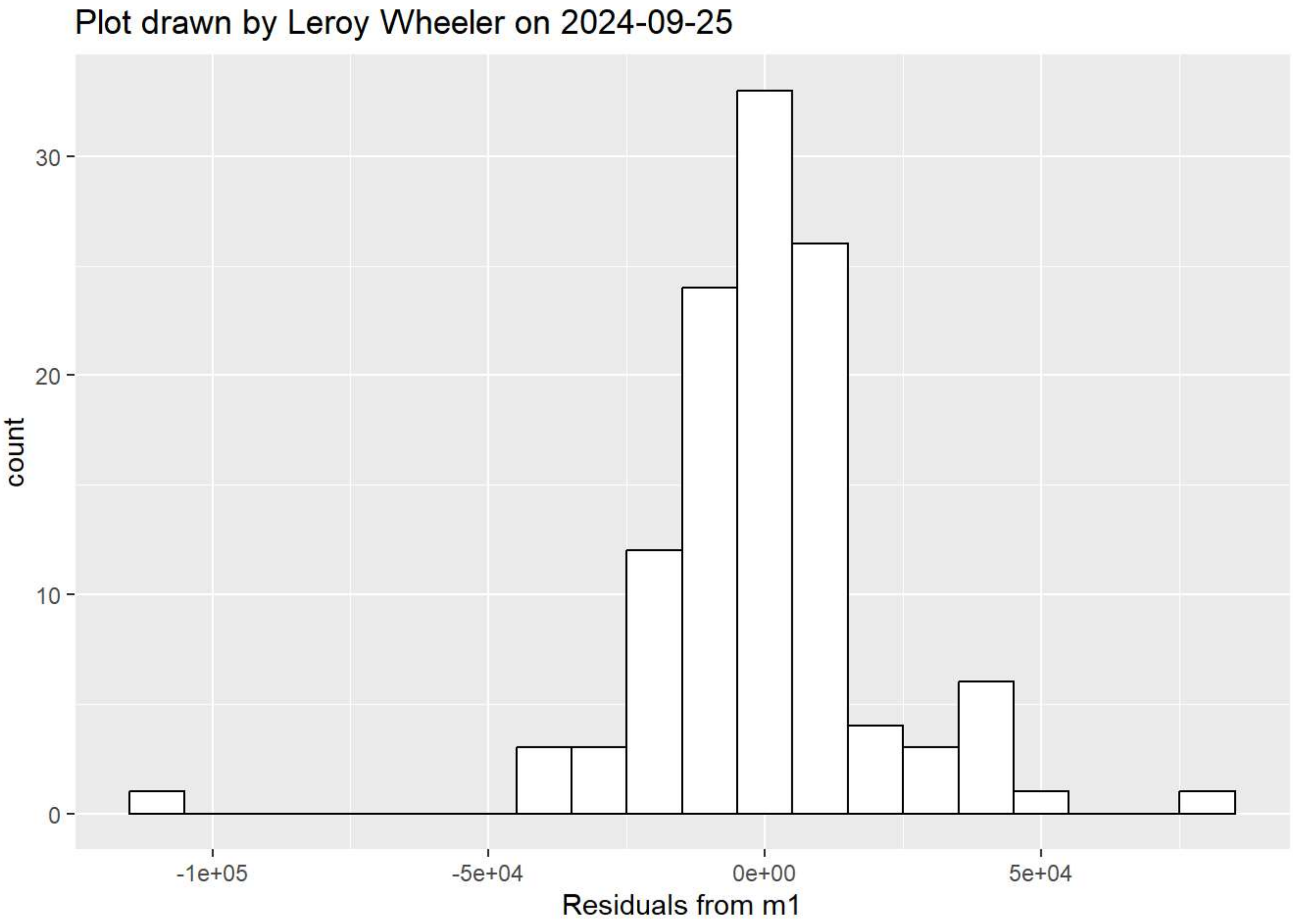


The normal probability plot reminds me of a straight line suggesting the residuals are normal, but I will be more convinced when I look at the histogram.

Note that you cannnot use ggtitle, xlab, or ylab with the qqnorm function.

m1: Question 4b: Draw a histogram for the residuals (.resid). Interpret these plots.

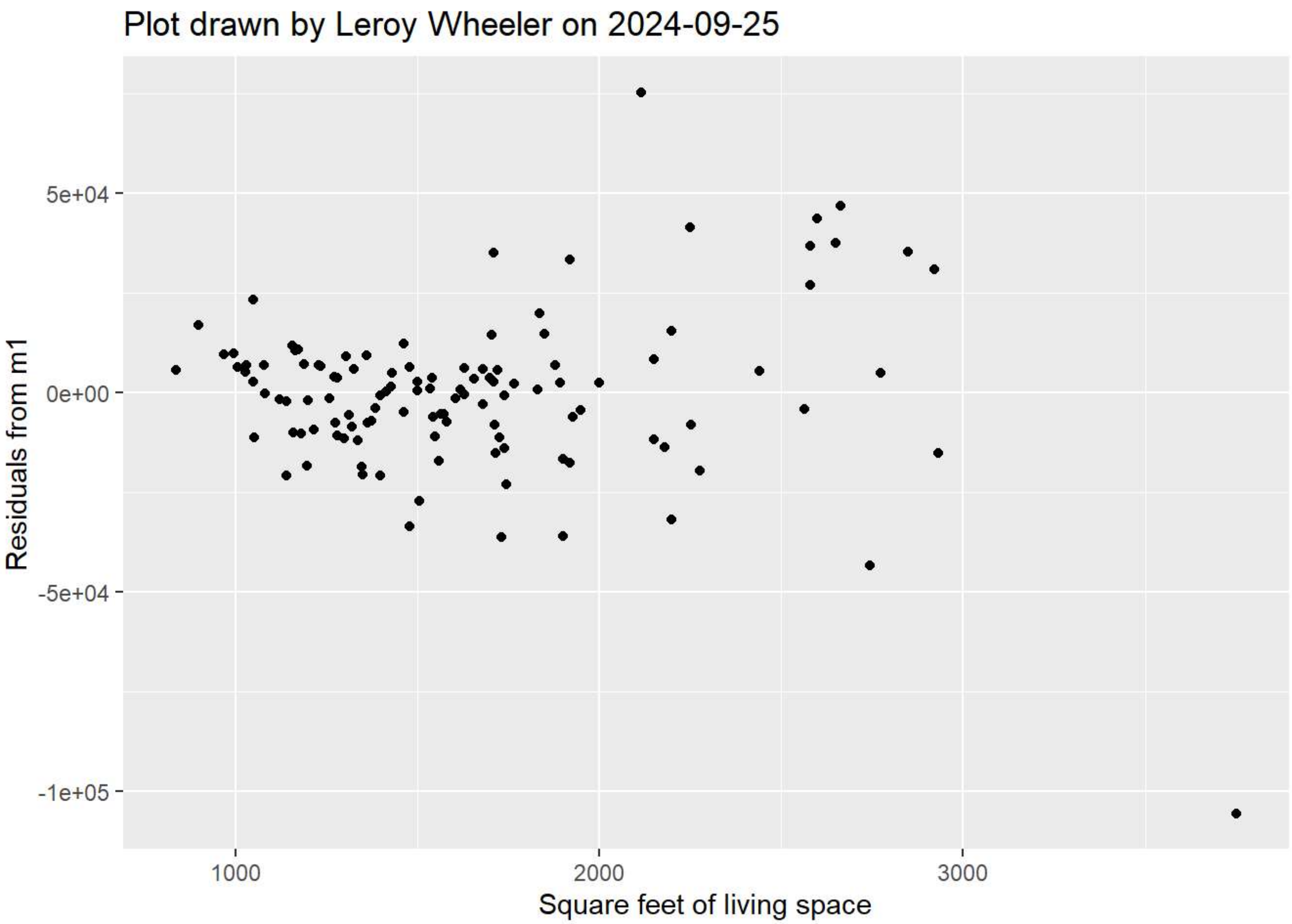
```r
r1 |>
  ggplot(aes(.resid)) +
    geom_histogram(
      binwidth=10000,
      color="black",
      fill="white") +
      ggtitle("Plot drawn by Leroy Wheeler on 2024-09-25") +
      xlab("Residuals from m1")
```



The histogram of the residuals of square feet versus price looks normal so I will be happy with my linear regression model of square feet predicting the price of the homes.

m1: Question 5: Draw a scatterplot of sqft on the x-axis and the residuals on the y-axis. Is there evidence of non-linearity or heterogeneity?

```
r1 |>
  ggplot(aes(sqft, .resid)) +
             geom_point() +
               ggtitle("Plot drawn by Leroy Wheeler on 2024-09-25") +
               xlab("Square feet of living space") +
               ylab("Residuals from m1")
```



There may be some heterogeneity where larger houses tend to exhibit more variation in their residuals. There is no evidence of non-linearity as a horizontal trend line can be drawn through the data. There is one data point from the largest house that could be a problem.

m1: Question 6: Display the data (if any) for leverage values greater than 3*2/n. Describe where these leverage values are found relative to the independent and/or dependent variables.

```
n <- nrow(r1)
r1 |> filter(.hat > 3*2/n)
```

```
# A tibble: 4 × 8
   price  sqft .fitted   .resid   .hat .sigma .cooksd .std.resid
   <dbl> <dbl>   <dbl>    <dbl>  <dbl>  <dbl>   <dbl>      <dbl>
1 215000  2921 184034.   30966. 0.0590 20316.  0.0764      1.56
2 169500  2931 184648.  -15148. 0.0598 20482.  0.0186     -0.764
3 215000  2848 179554.   35446. 0.0534 20249.  0.0895      1.78
4 129500  3750 234907. -105407. 0.147  17535.  2.68       -5.58
```

There are four data points with high leverage, meaning more than 6/n. These correspond to the four houses with the largest square footage. The house with the largest value of square foot actually has an observed price which is relatively far away from its predicted price value.

m1: Queston 7: Display the data (if any) for studentized deleted residuals (.std.resid) values greater than 3. Describe where these leverage values are found relative to the independent and/or dependent variables.

```
r1 |>
  filter(abs(.std.resid) > 3)
```

```
# A tibble: 2 × 8
   price  sqft .fitted   .resid   .hat .sigma .cooksd .std.resid
   <dbl> <dbl>   <dbl>    <dbl>  <dbl>  <dbl>   <dbl>      <dbl>
1 129500  3750 234907. -105407. 0.147  17535.   2.68       -5.58
2 210000  2116 134634.   75366. 0.0153 19263.   0.107       3.71
```

There are two houses with a studentized deleted residual greater than 3. The largest house has an observed price far below what the model would predict. This house was also one of the houses with a lot of leverage. The second house with 2116 square feet had an observed price far above what the model would predict.

m1: Question 8: Display the data (if any) for Cook's distance (.cooksd) values greater than 1. Describe where these leverage values are found relative to the independent and/or dependent variables.

```
r1 |>
  filter(.cooksd > 1)
```

```
# A tibble: 1 × 8
   price  sqft .fitted   .resid  .hat .sigma .cooksd .std.resid
   <dbl> <dbl>   <dbl>    <dbl> <dbl>  <dbl>   <dbl>      <dbl>
1 129500  3750 234907. -105407. 0.147 17535.    2.68      -5.58
```

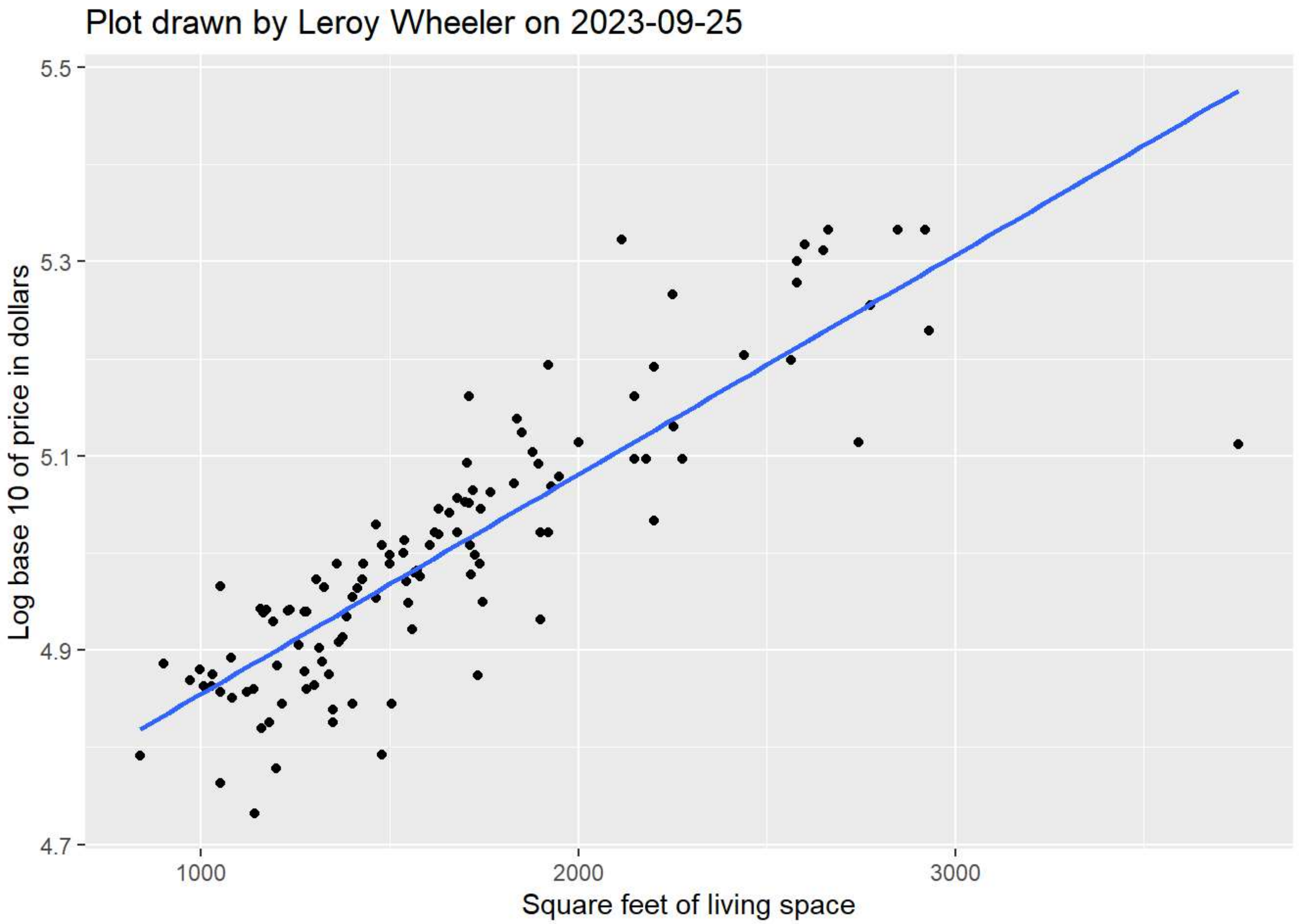
The largest house in the data set (3750 sqft) had a large value for Cook's distance. This single data point has unusually high influence on the predicted values.

m2: Because there are some concerns about non-normality and heterogeneity, you might consider using a log transformation for price. In this example, a base 10 logarithm is a reasonable choice.

m2: scatterplot

```
alb$log_price <- log10(alb$price)
alb |>
  ggplot(aes(sqft, log_price)) +
    geom_point() +
    geom_smooth(method="lm", se=FALSE) +
      ggtitle("Plot drawn by Leroy Wheeler on 2023-09-25") +
            xlab("Square feet of living space") +
            ylab("Log base 10 of price in dollars")
```

```
`geom_smooth()` using formula = 'y ~ x'
```



There is a strong positive linear relationship between log price and square feet of living space.

m2: Question 9a: Calculate the regression equation predicting log10 of price using sqft.

```r
m2 <- lm(log_price~sqft, data=alb)
m2
```

```

Call:
lm(formula = log_price ~ sqft, data = alb)

Coefficients:
(Intercept)         sqft  
  4.6294697    0.0002258  
```

The estimated average log price is 4.6 for a house with no living space. This Y-intercept value is meaningless. The estimated average price increases by 0.00023 log for each additional square foot of living space. These numbers are easier to interpret when transformed back to the original scale.

m2: Question 9b: Transform the coefficients back to the original scale of measurement and interpret these values.

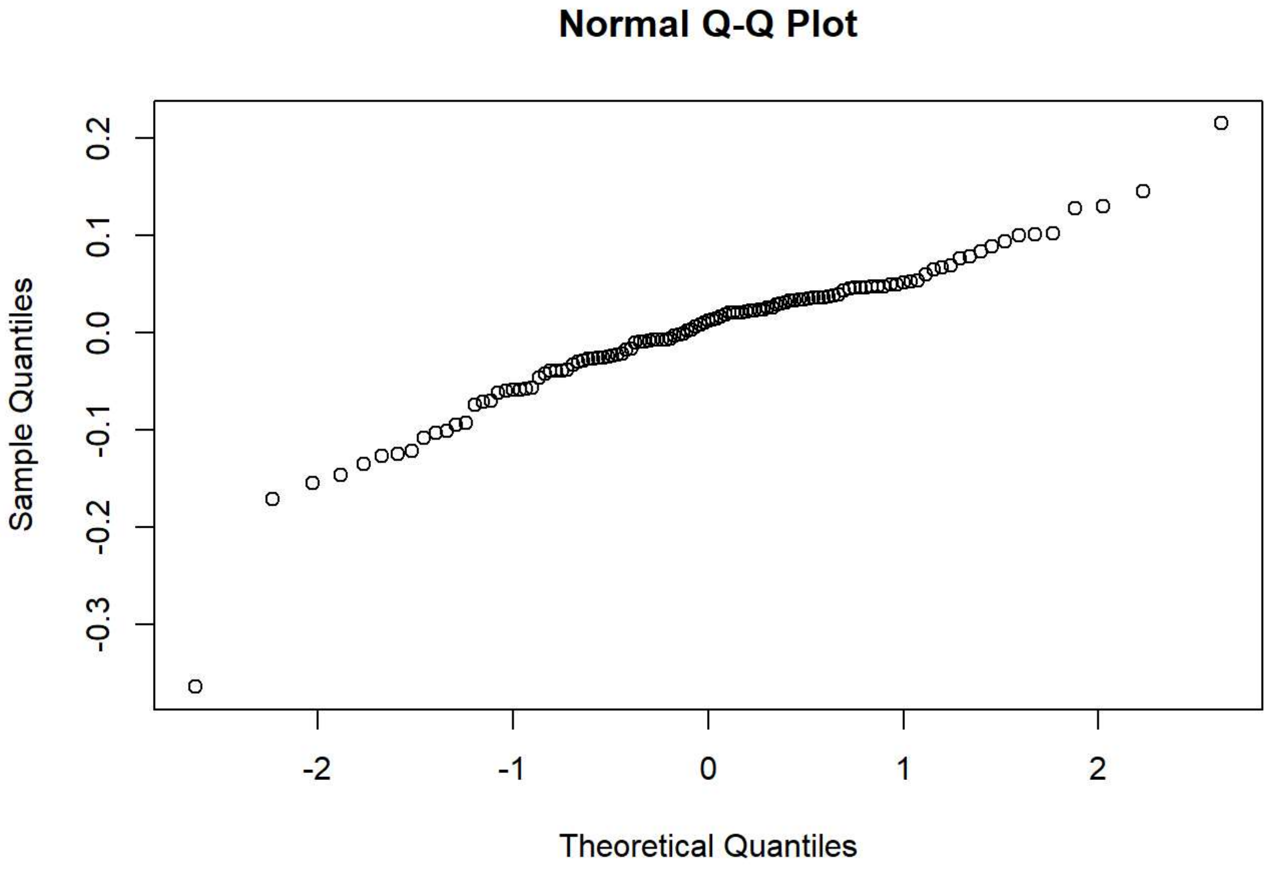
```r
10^(coef(m2))
```

```
(Intercept)        sqft 
42605.89143     1.00052 
```

The estimated average price is about 43,000 USD for a house with no living space, which is meaningless. The house will increase in price by 5% per additional 100 square feet, which is pretty cheap.

m2: Question 10: Calculate diagnostic plots (normal probability plot, histogram, and sqft versus residuals). Do these plots show that a model using log10 price better meets the assumptions for linear regression?

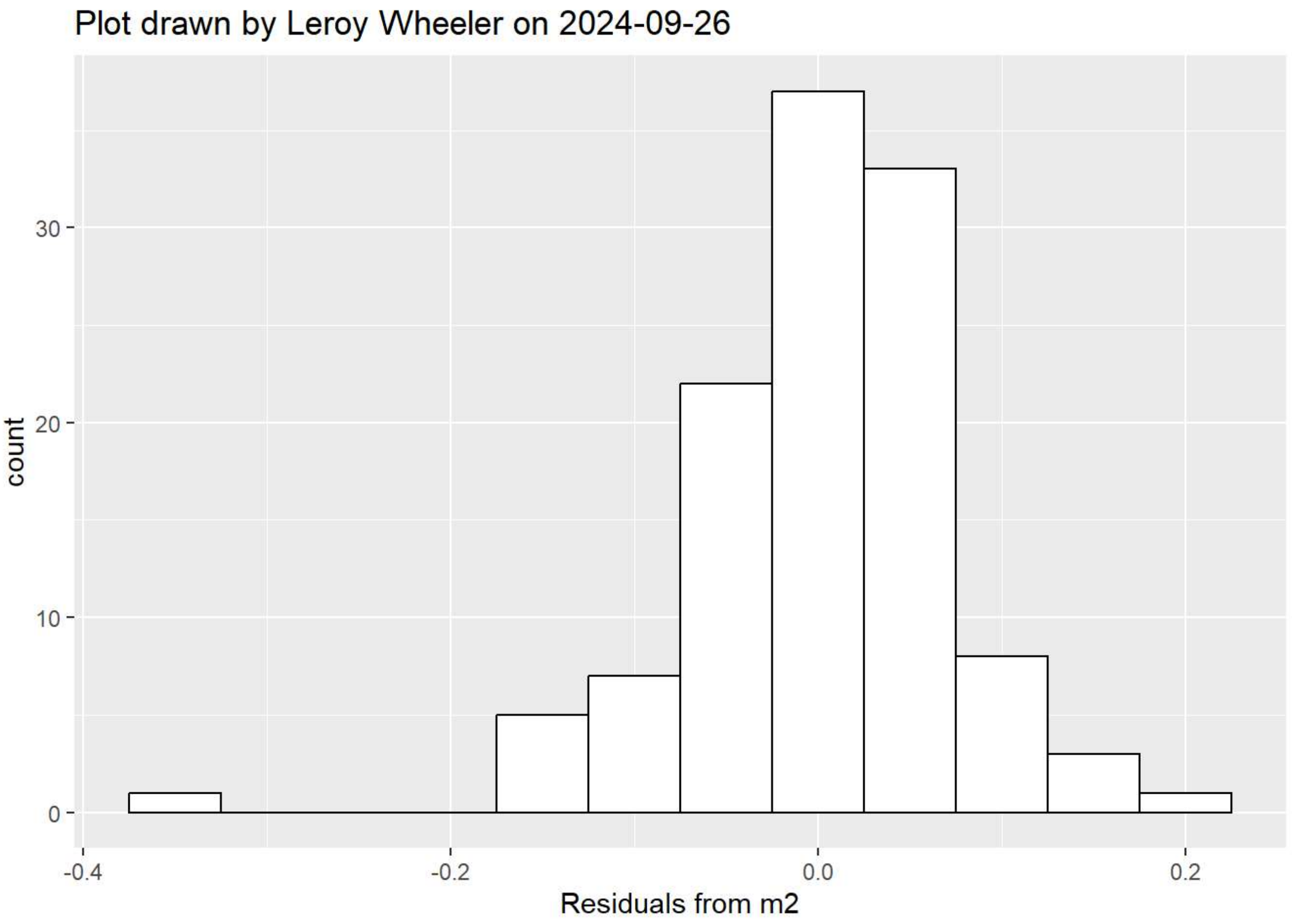
```r
r2 <- augment(m2)
qqnorm(r2$.resid)
```



The normal probability plot of the residuals looks like a straight line, indicating a reasonably close fit to a normal distribution.

m2: Histogram of residuals

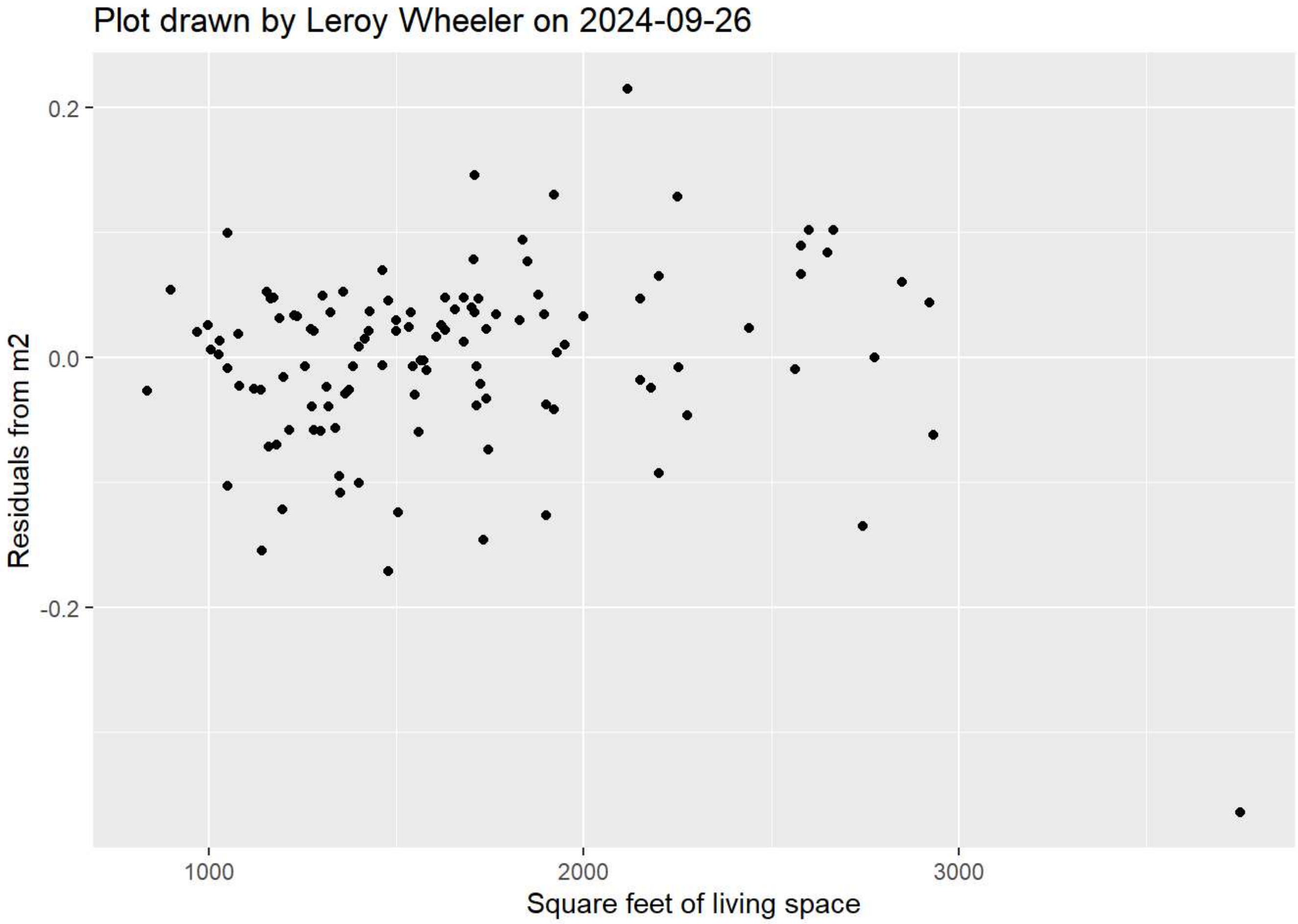
```
r2 |>
  ggplot(aes(.resid)) +
    geom_histogram(
      binwidth=0.05,
      color="black",
      fill="white") +
      ggtitle("Plot drawn by Leroy Wheeler on 2024-09-26") +
      xlab("Residuals from m2")
```



The histogram of residuals also indicates a close fit to a normal distribution. The regression model using log price does a better job meeting the normality assumption.

m2: Plot residuals versus square feet

```
r2 |>
  ggplot(aes(sqft, .resid)) +
    geom_point() +
      ggtitle("Plot drawn by Leroy Wheeler on 2024-09-26") +
      xlab("Square feet of living space") +
      ylab("Residuals from m2")
```



Variation in the residuals seem to be independent of square feet, with no obvious increase or decrease as the square feet increases. Overall performing the log transformation of the price before calculating the linear regression model may have improved the model a little bit.