hw_05_REVISED

lisa liubovich

2024-02-22

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(ggplot2)
library(broom)
```

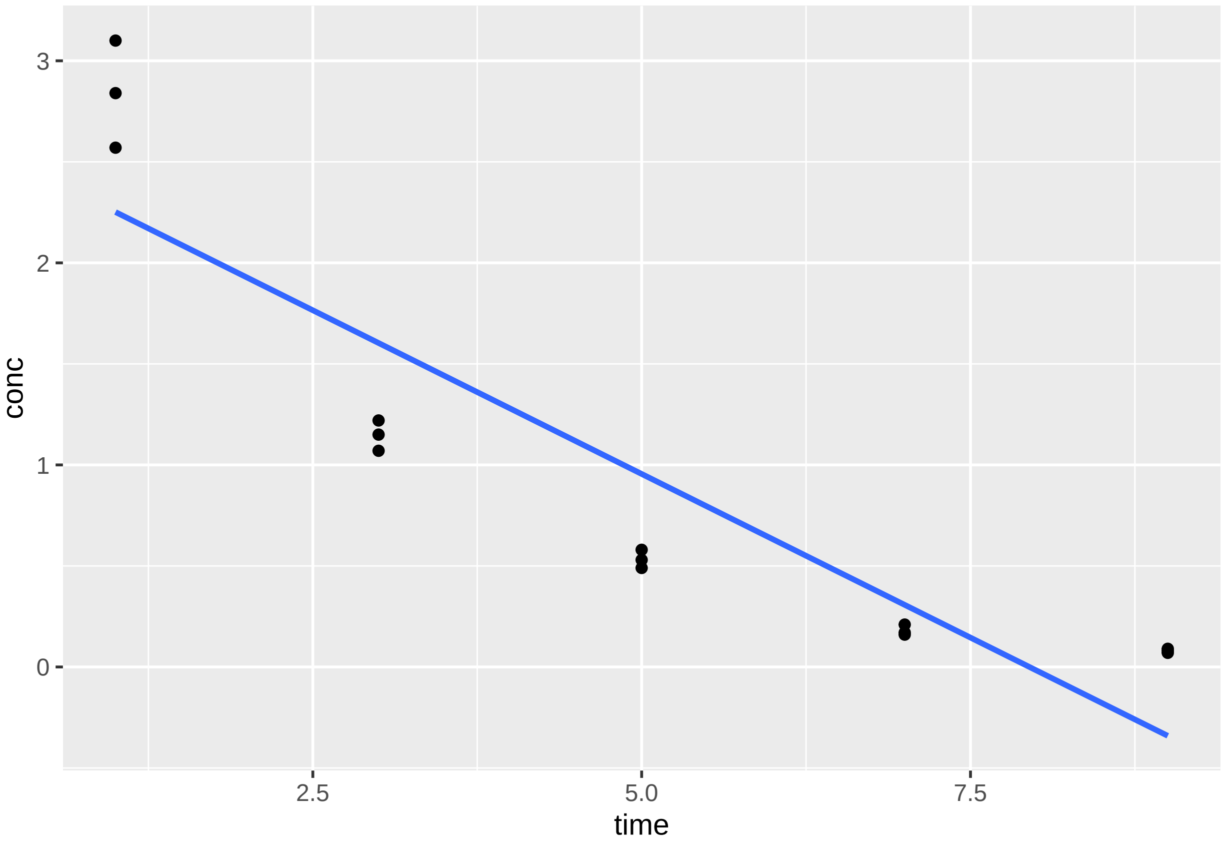
Solution Concentration

```
sol <- tribble(~conc, ~time,
0.07, 9,
0.09, 9,
0.08, 9,
0.16, 7,
0.17, 7,
0.21, 7,
0.49, 5,
0.58, 5,
0.53, 5,
1.22, 3,
1.15, 3,
1.07, 3,
2.84, 1,
2.57, 1,
3.10, 1
)
```

1. plotting data

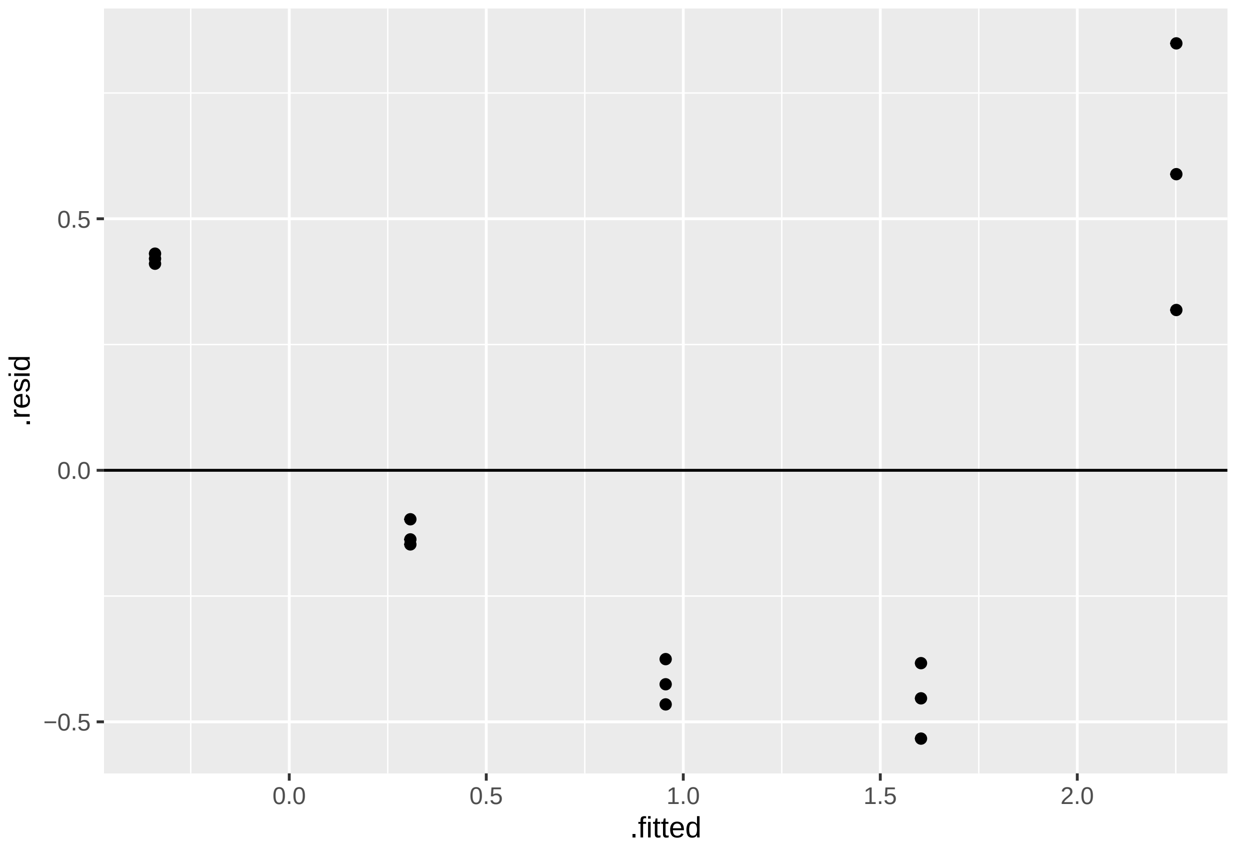
```
ggplot(sol, aes(x = time, y = conc)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)
```

```
## ‘geom_smooth()‘ using formula = ’y ~ x’
```



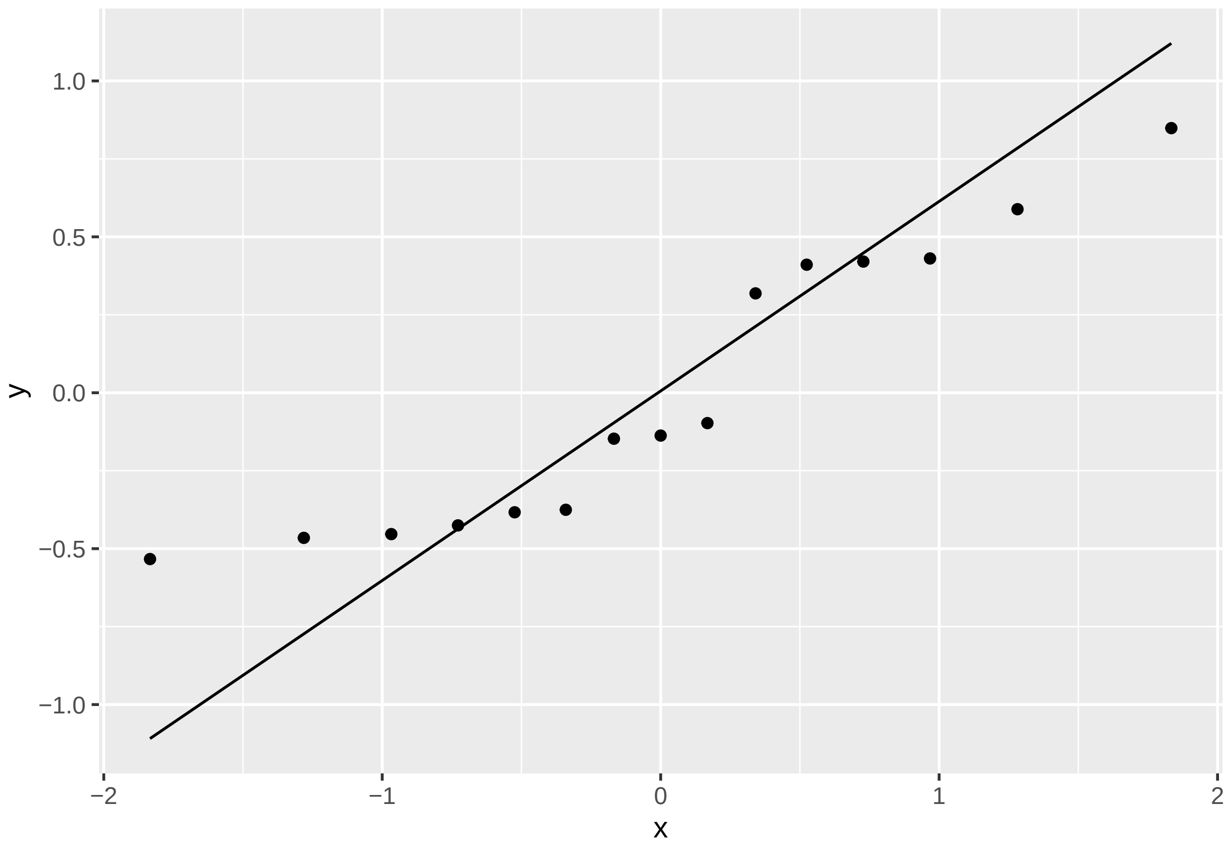
Relationship looks curved and vertical spread is inconsistent –> violation of linearity and constant variance assumptions

```
lmsol <- lm(conc ~ time, sol)
asol <- augment(lmsol)
ggplot(asol, aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0)
```



Very inconsistent vertical spread and clear pattern –> violation of constant variance and independence asssumptions

```
ggplot(asol, aes(sample = .resid)) +
  geom_qq() +
  geom_qq_line()
```



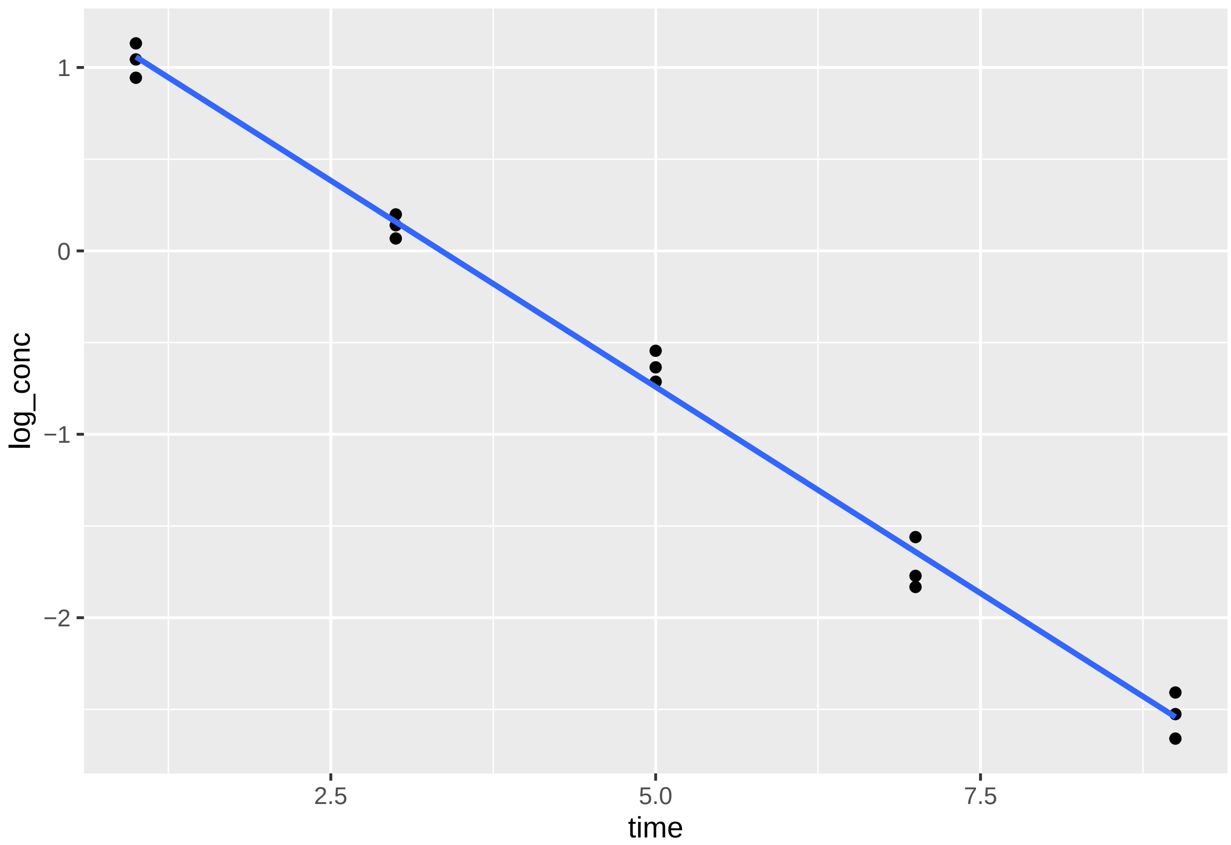
Points are quite far from the diagonal line –> violation of normality assumption

2. transformation

In order to fix our curved, heteroskedastic data, we can transform y with log(y)

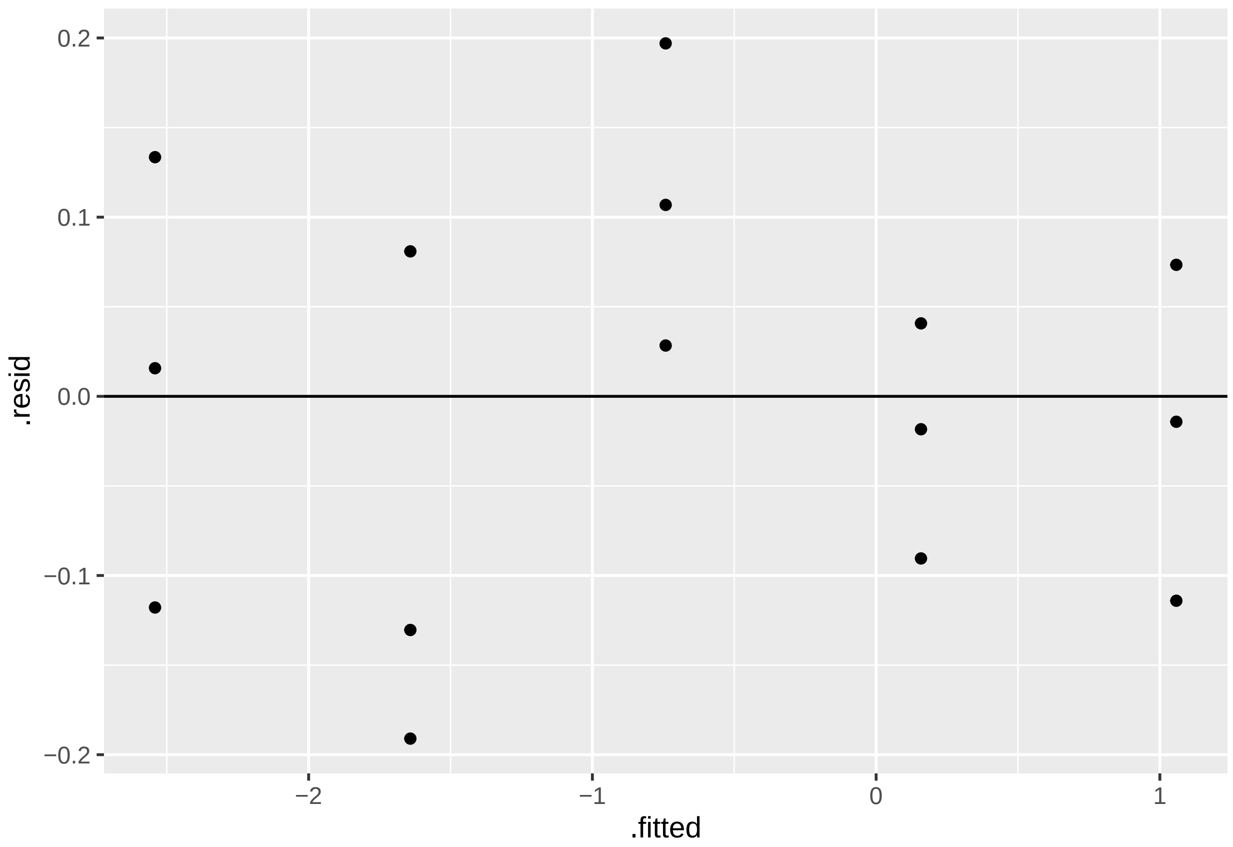
```
sol_log <- mutate(sol, log_conc = log(conc))
ggplot(sol_log, aes(x = time, y = log_conc)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)
```

```
## ‘geom_smooth()‘ using formula = ’y ~ x’
```



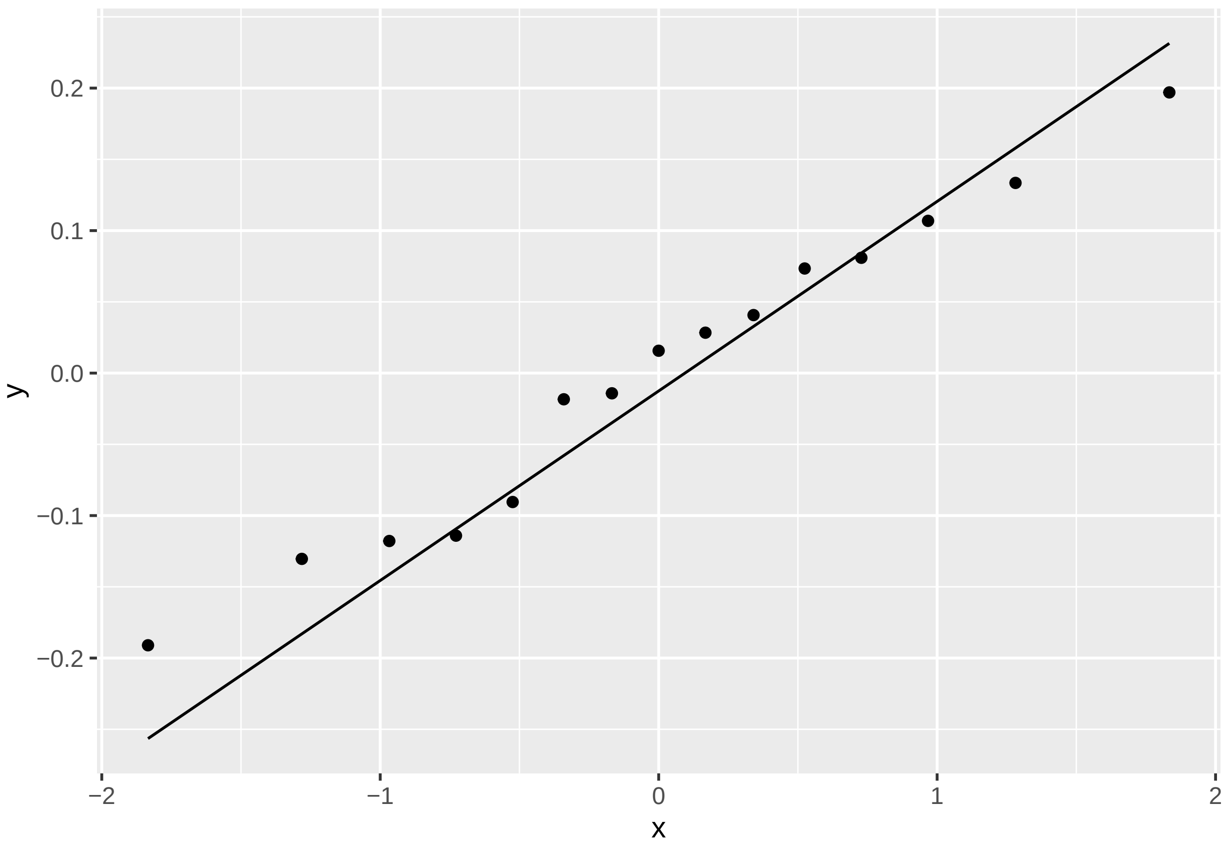
Much better linearity and constant variance –> linearity and constant variance validated

```
lmlogsol <- lm(log_conc ~ time, sol_log)
alogsol <- augment(lmlogsol)
ggplot(alogsol, aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0)
```



Much better vertical spread and more acceptable lack of pattern –> constant variance and independence validated

```
ggplot(alogsol, aes(sample = .resid)) +
  geom_qq() +
  geom_qq_line()
```



A bit skewed, but overall much more normal –> normality assumption validated

```
tidy(lmlogsol, conf.int = TRUE)
```

```
## # A tibble: 2 x 7
##   term        estimate std.error statistic  p.value conf.low conf.high
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
## 1 (Intercept)    1.51     0.0603      25.0 2.22e-12    1.38      1.64
## 2 time          -0.450    0.0105     -42.9 2.19e-15   -0.473    -0.427
```

need to exponentiate in order to return to original scale:

```
exp(-0.4499)
```

```
## [1] 0.6376919
```

```
exp(-0.4725)
```

```
## [1] 0.6234417
```

```
exp(-0.4272)
```

```
## [1] 0.6523331
```

Interpretation: Each hour, the concentration was 63.77% higher on average (95% CI of 62.23% higher to 65.23% higher).

Real Estate Data

```
est <- read_csv("../week_5/estate.csv")
```

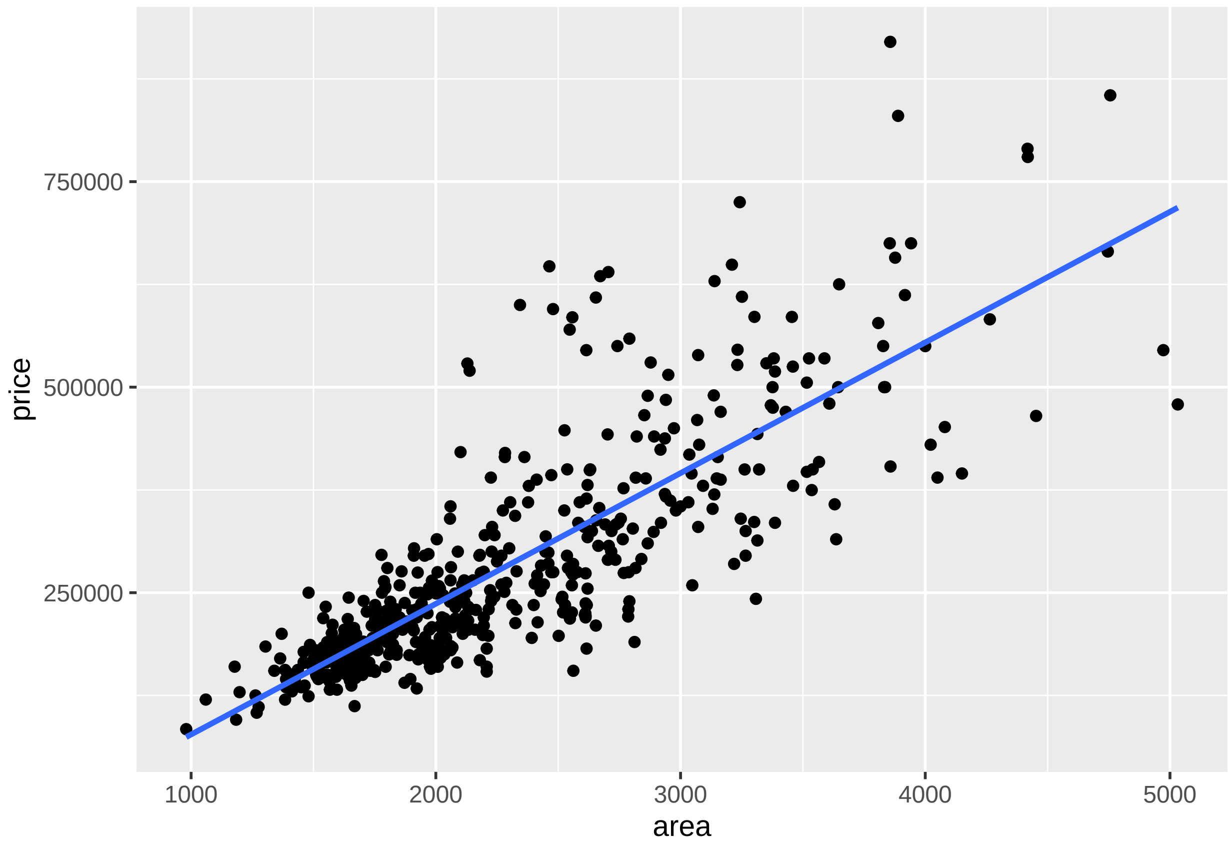
```
## Rows: 522 Columns: 13
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (4): ac, pool, quality, highway
## dbl (9): id, price, area, bed, bath, garage, year, style, lot
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

Analysis:

validating model assumptions:

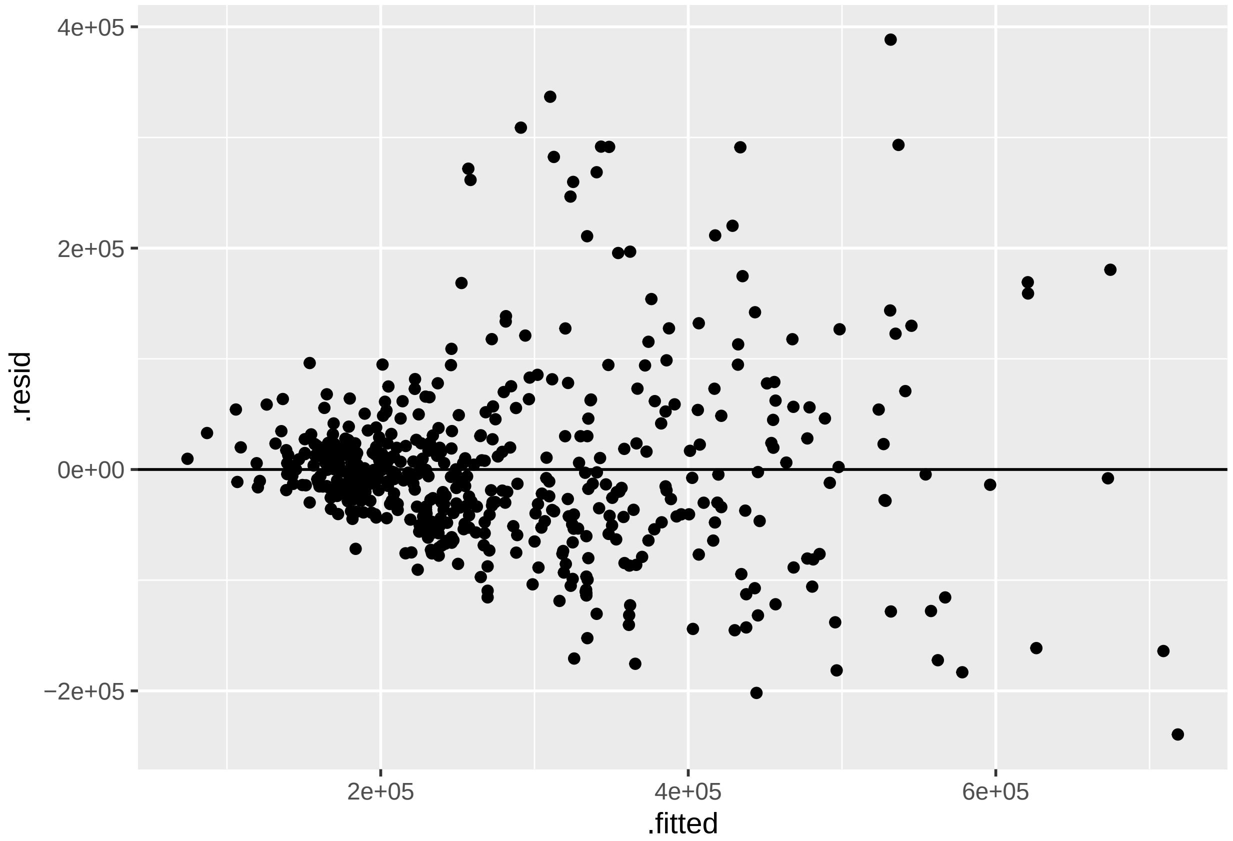
```
ggplot(est, aes(x = area, y = price)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)
```

```
## ‘geom_smooth()‘ using formula = ’y ~ x’
```



Seems pretty linear, but the vertical spread is not consistent. –> linear assumption validated, constant variance assumption violated

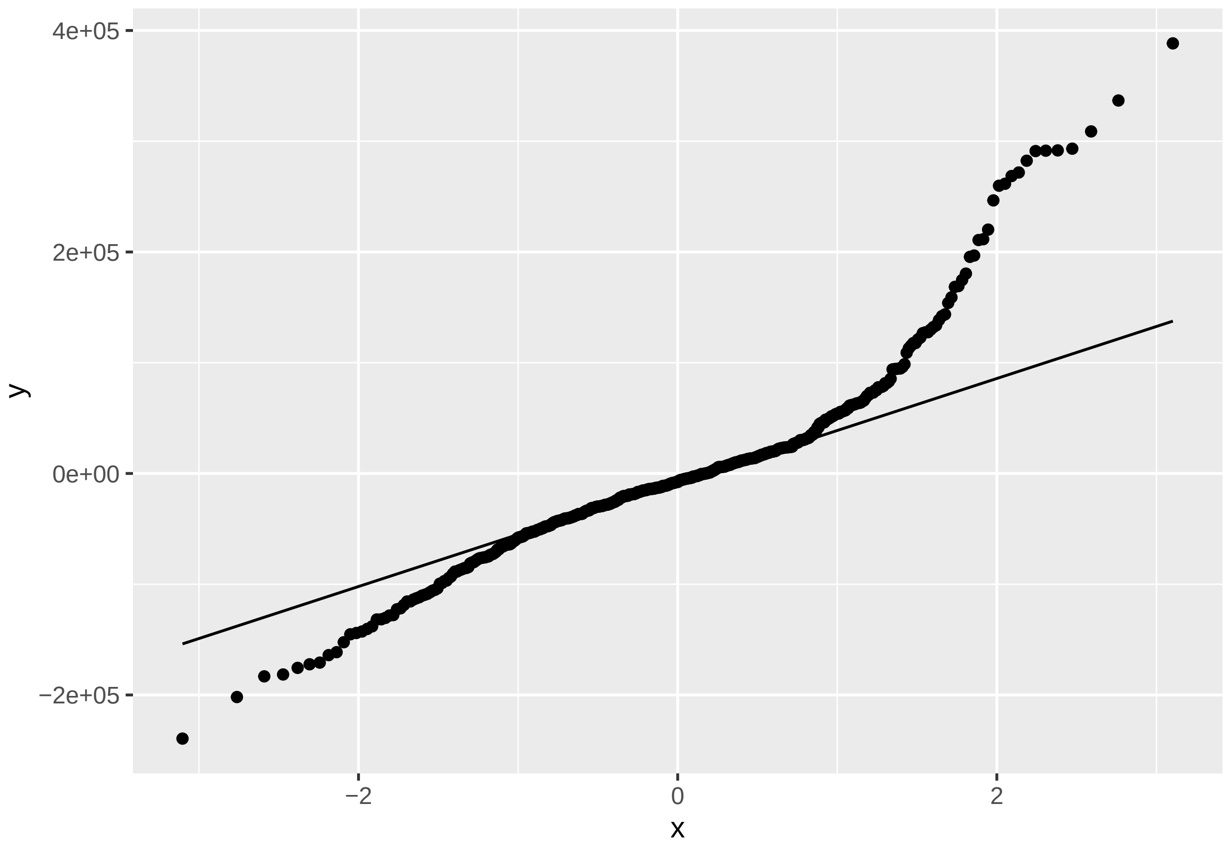
```r
lmest <- lm(price ~ area, est)
aest <- augment(lmest)
ggplot(aest, aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0)
```



Funnel shaped –> constant variance assumption violated

no clear pattern -> independence assumption validated

```r
ggplot(aest, aes(sample = .resid)) +
  geom_qq() +
  geom_qq_line()
```

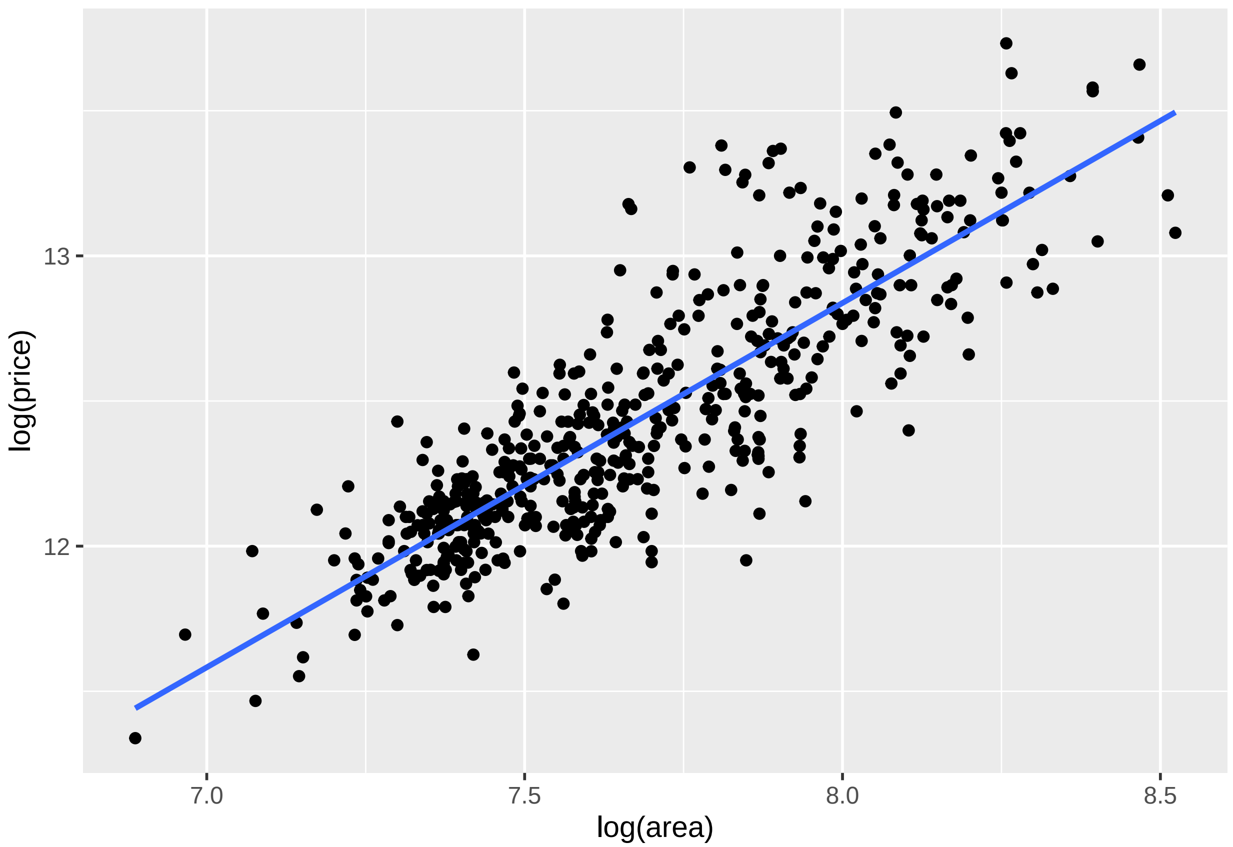


Bit of a skew, but not that bad –> normality assumption validated, especially given a large sample size of n = 522.

Transformation: log both x and y

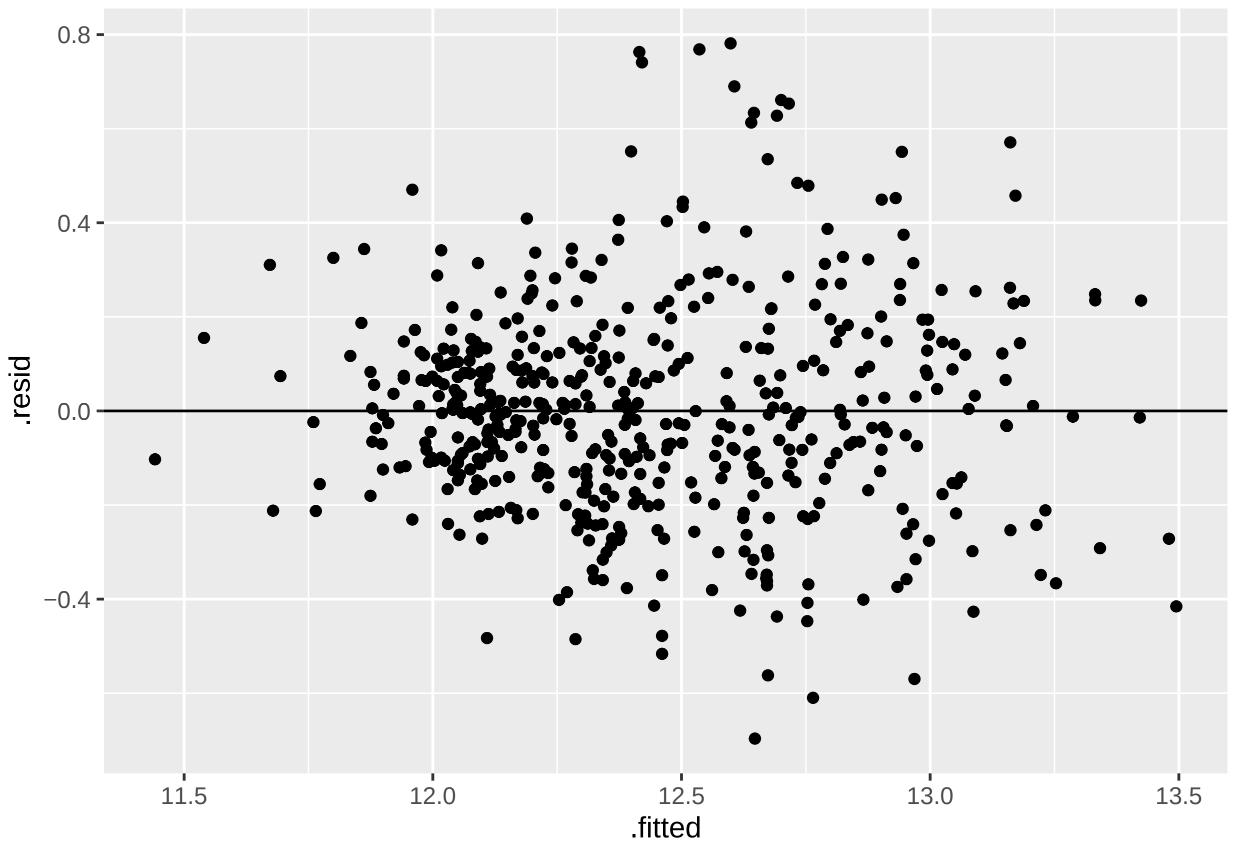
```
ggplot(est, aes(x = log(area), y = log(price))) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)
```

```
## ‘geom_smooth()‘ using formula = ’y ~ x’
```



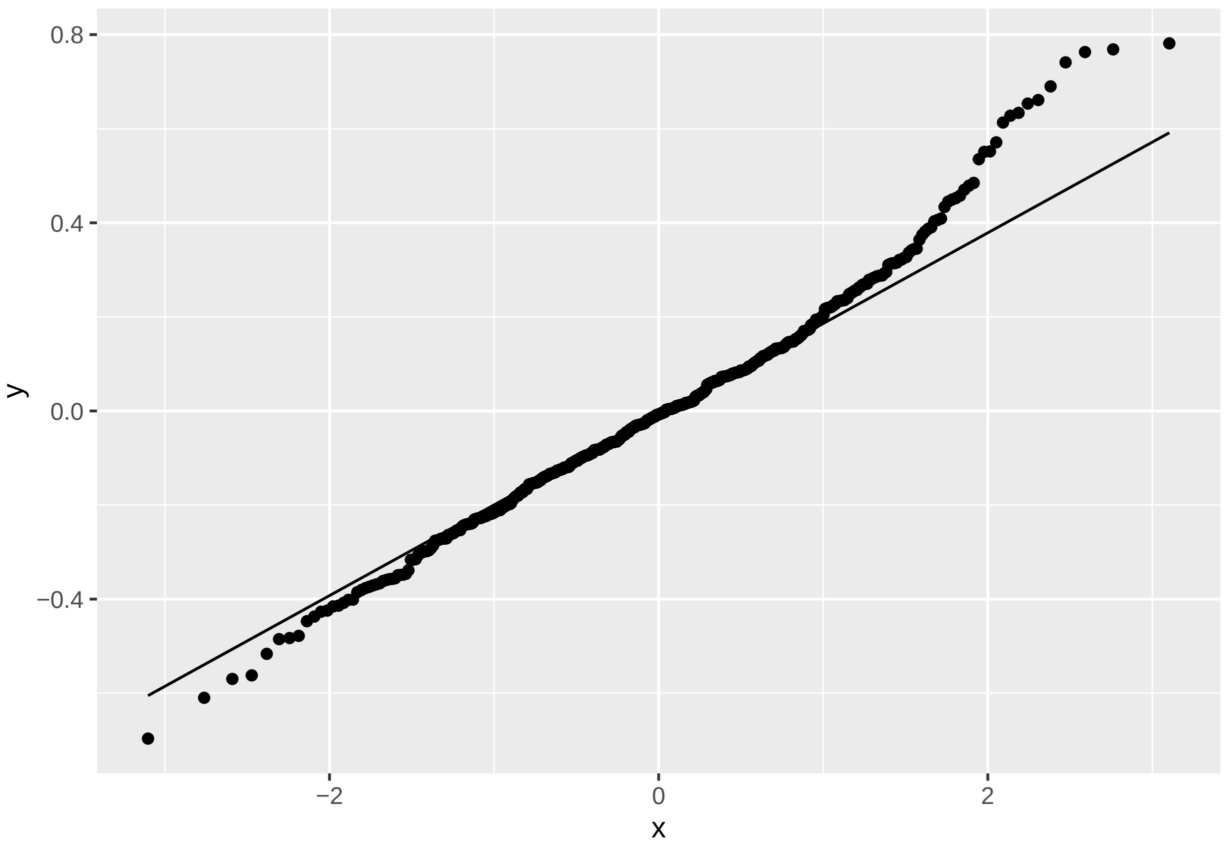
Much better linearity and vertical spread –> linearity and constant variance assumptions validated

```
logest <- mutate(est, log_price = log(price),
                 log_area = log(area))
lmlogest <- lm(log_price ~ log_area, logest)
alogest <- augment(lmlogest)
ggplot(alogest, aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0)
```



Much more consistent vertical spread and no clear pattern -> constant variance and independence assumption validated.

```
ggplot(alogest, aes(sample = .resid)) +
  geom_qq() +
  geom_qq_line()
```



Points are much closer to diagonal line -> no/minimal skew -> normality assumption validated.

interpretation and measures of uncertainty

fitting our linear model:

```
tidy(lmlogest, conf.int = TRUE)
```

```
## # A tibble: 2 x 7
##   term        estimate std.error statistic   p.value conf.low conf.high
##   <chr>          <dbl>     <dbl>     <dbl>     <dbl>    <dbl>     <dbl>
## 1 (Intercept)     2.80    0.259       10.8 1.31e- 24     2.29      3.31
## 2 log_area        1.26    0.0337      37.2 1.20e-148     1.19      1.32
```

This is a power law relationship, so the model is e^{B0} x^{B1} –> in this case, e$^{2.796}$ x$^{1.255}$

Let’s say c = 2 –> 2ˆ1.255 is the slope

```
2^1.255 -1
```

```
## [1] 1.386671
```

```
2^1.188 -1
```

```
## [1] 1.278367
```

```
2^1.321 -1
```

```
## [1] 1.498392
```

Houses with twice as much square footage are 38.66% more expensive on average (95% CI of 27.84% more expensive to 49.84% more expensive).

predictions:

```
range(est$area) # this is our range, since the power law equation is back in the original scale
```

```
## [1]  980 5032
```

576 is not in range, therefore we cannot make a prediction for this area size. The other two are in range, so we can make predictions for them.

Prediction interval:

```
newdf <- data.frame(area = 1020)
newdf <-mutate(newdf, log_area = log(area))
predict(object = lmlogest, newdata = newdf, interval = "prediction")
```

```
##        fit      lwr      upr
## 1 11.49164 11.04479 11.93849
```

Need to exponentiate to get back in original scale:

```
exp(11.492)
```

```
## [1] 97929.2
```

```
exp(11.045)
```

```
## [1] 62630.02
```

```
exp(11.938)
```

```
## [1] 152970.4
```

Houses that are 1020 square feet in area are, on average, about $97,929.20 with a prediction interval of $62,630.02 to $152,970.4

```
newdf_1 <- data.frame(area = 3067)
newdf_1 <-mutate(newdf_1, log_area = log(area))
predict(object = lmlogest, newdata = newdf_1, interval = "prediction")
```

```
##        fit     lwr      upr
## 1 12.87347 12.4288 13.31814
```

Need to exponentiate to return to original scale:

```r
exp(12.873)
```

```
## [1] 389648.4
```

```r
exp(12.429)
```

```
## [1] 249946
```

```r
exp(13.318)
```

```
## [1] 608042.5
```

Houses that are 3067 square feet in area are, on average, about $389,648.40 with a prediction interval of $249,946 to $608,042.50.

Report:

Modeling the raw data lead to a linear model that violated the assumptions of constant variance. In order to address this issue, a log transformation was applied to both price and area. With a model of log(price) = B0 + B1log(area), all four of the assumptions of the linear model (linearity, independence, constant variance, and normality), were validated. Our prediction equation is in power law form:

y-hat = $e^{2.796}$ $x^{1.255}$, where y-hat is predicted price and x is area in square feet. In this form, our model can be interpreted as follows:

Houses with twice as much square footage are 38.66% more expensive on average, with a 95% confidence interval of 27.84% more expensive to 49.84% more expensive. Essentially, we are testing the null hypothesis that B1 = 0 and the alternative hypothesis that B1 =/=0. The estimate of 38.66% has a p-value of less than 0.001, therefore we reject H0 and conclude that there is statistically significant evidence to suggest that there is an association between price and area. Moreover, our confidence interval does not contain 0, which is further evidence that the association between price and area is not a result of random chance alone.

When it comes to predicting the price of the given square footage (576, 1020, 3067), we cannot make a prediction for 576 square feet because it is out of range of our data (the range is 980 to 5032) and thus such a prediction would be inappropriate. However, 1020 square feet and 3067 square feet are within range, so we can make predictions for those areas using a prediction interval.

Houses that are 1020 square feet in area are, on average, about $97,929.20 with a prediction interval of $62,630.02 to $152,970.4. Houses that are 1020 square feet in area are, on average, about $97,929.20 with a prediction interval of $62,630.02 to $152,970.40.