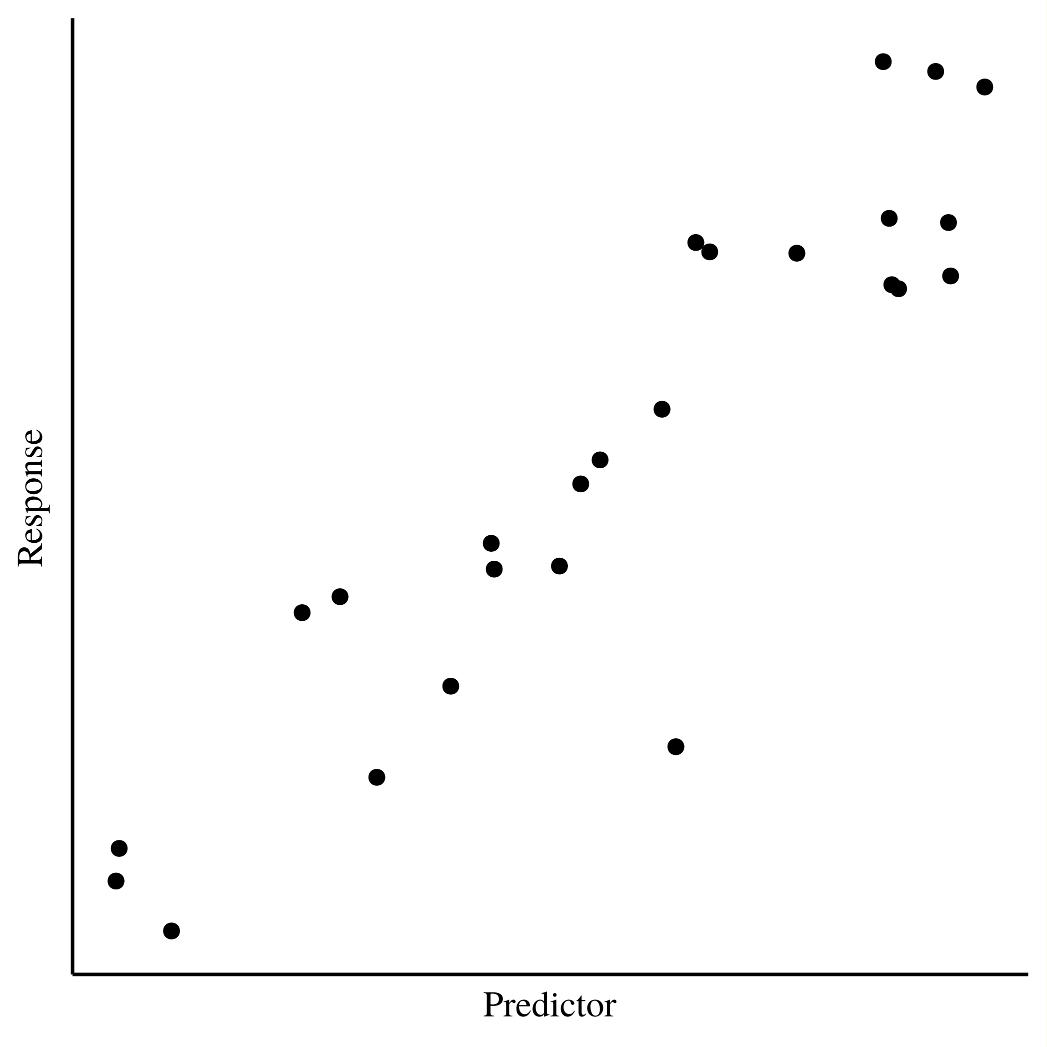
SLR: Transforming variables

Stat 230

April 08 2022

Overview



Today:

Transforming variables in SLR

- trial-and-error
- common models
- interpretation

What if model assumptions are violated?

Linearity, Variance, Normality: Transform one or both variables

Check if a linear relationship exists on a transformed scale

- $\log(y)$ vs x
- $\text{sqrt}(y)$ vs x
- $1/y$ vs x
- $\log(y)$ vs $\log(x)$
- y vs $\log(x)$

What transformation to use?

Application specific!

Trial-and-error

- Use scatter/residual plots to explore different transformations of y and/or x
- Common transformation forms: **logarithm, square root and reciprocal**

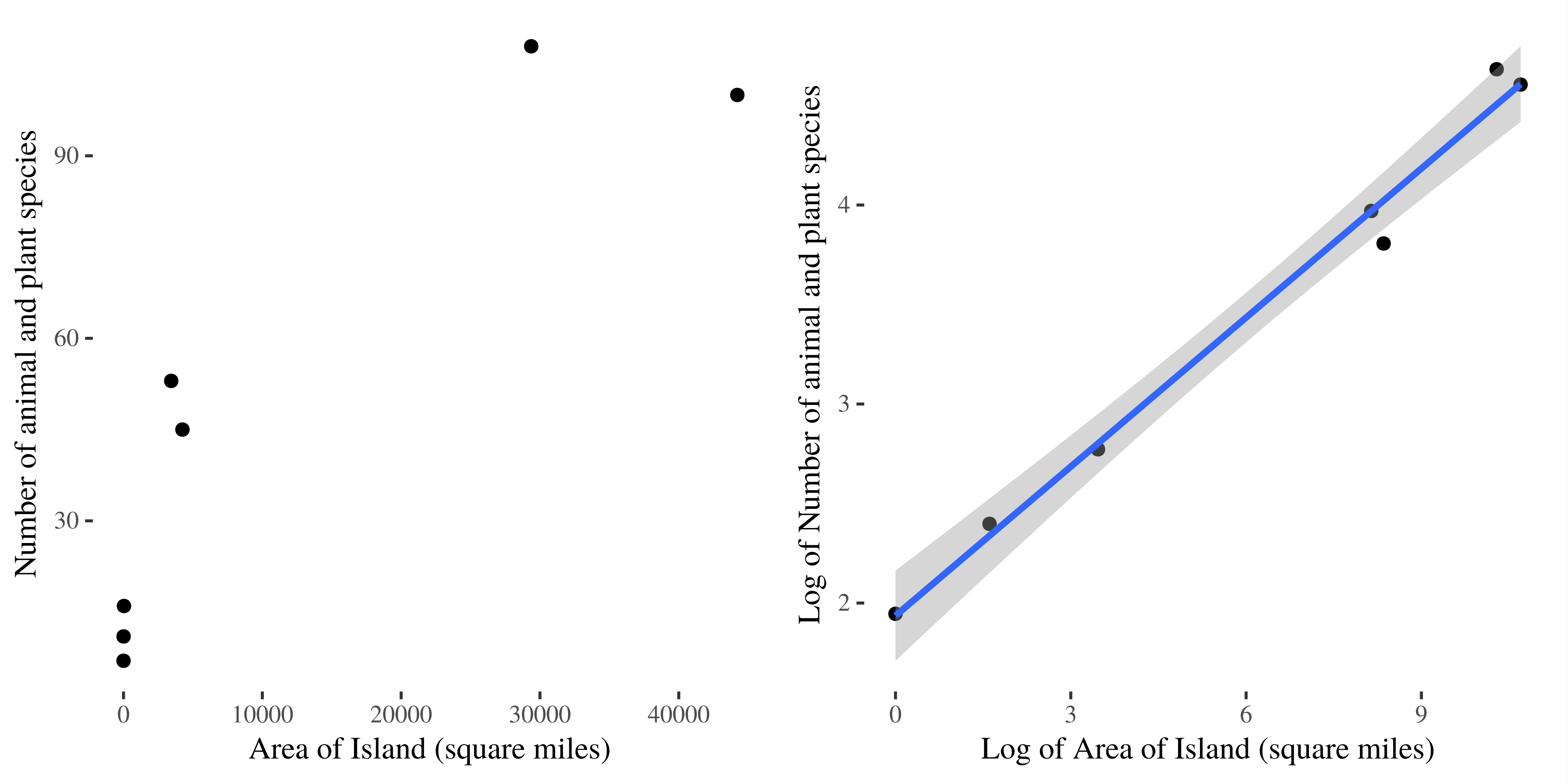
Case Study 1: Island Area and Number of Species

The data are the numbers of reptile and amphibian species and the island areas for seven islands in the West Indies.

```
summary(case0801)
```

```
      Area              Species     
 Min.   :    1.0   Min.   :  7.00  
 1st Qu.:   18.5   1st Qu.: 13.50  
 Median : 3435.0   Median : 45.00  
 Mean   :11615.1   Mean   : 48.57  
 3rd Qu.:16807.5   3rd Qu.: 76.50  
 Max.   :44218.0   Max.   :108.00  
```

Both Predictor and Response Transformation



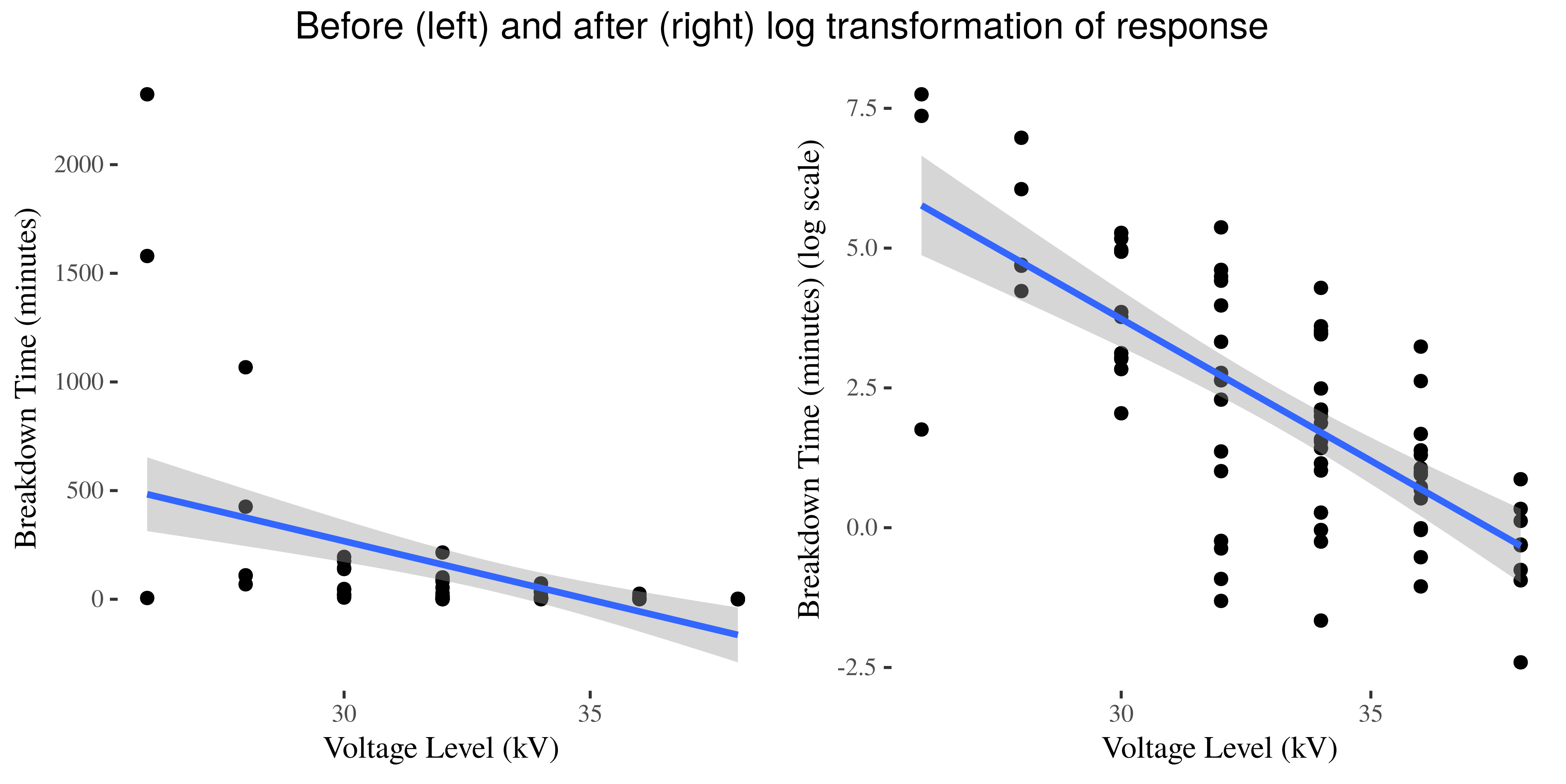
The data does not lie close to a straight line, but it does if both variables are log-transformed.

Case Study 2: Breakdown Times for Insulating Fluid Under Different Voltages

Batches of electrical insulating fluid were subjected to several constant voltages until the insulating property of the fluids broke down and the times taken were recorded

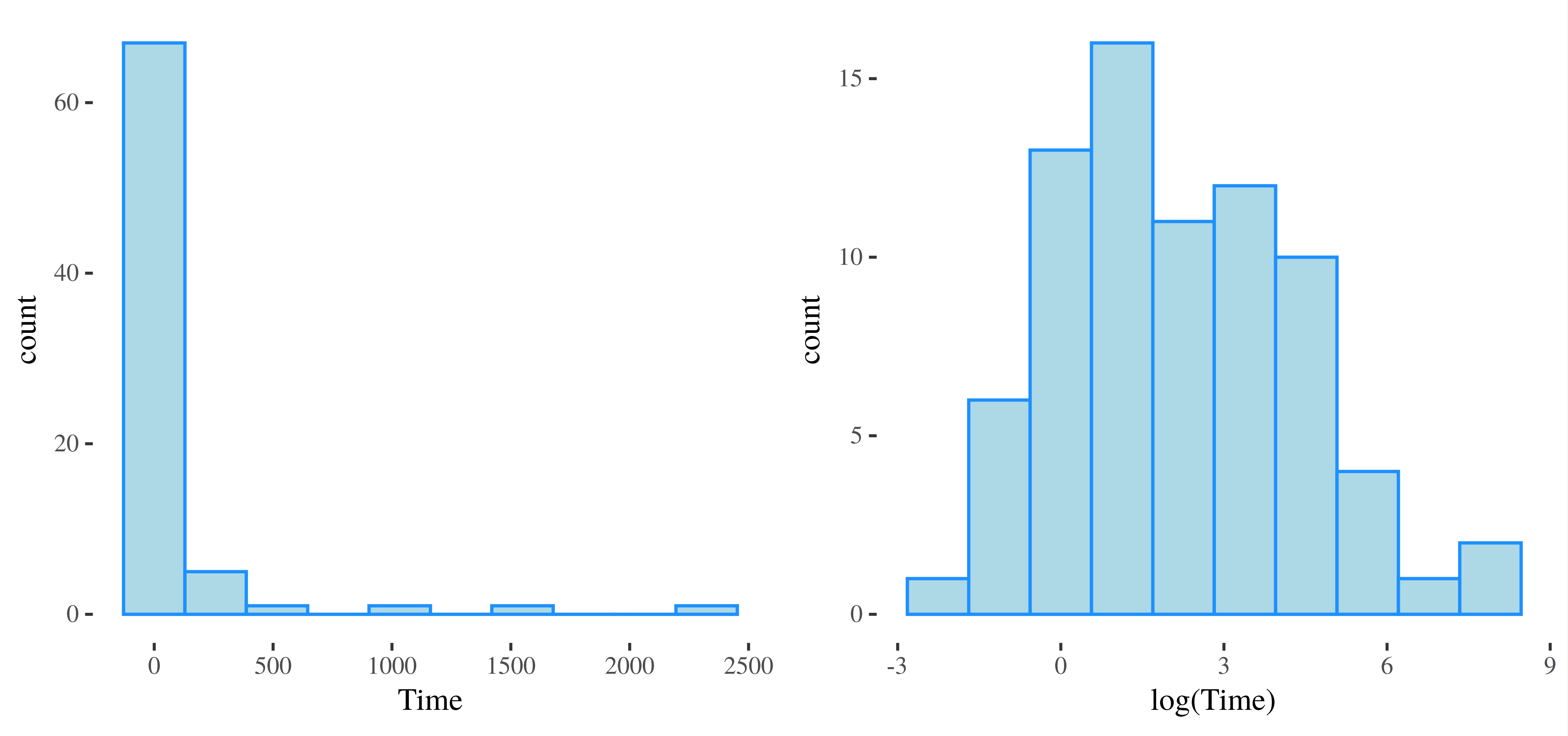
```
summary(case0802)
```

```
      Time                Voltage           Group   
 Min.    :   0.090   Min.    :26.00   Group1: 3  
 1st Qu.:   1.617   1st Qu.:31.50   Group2: 5  
 Median :   6.925   Median :34.00   Group3:11  
 Mean   :  98.558   Mean   :33.13   Group4:15  
 3rd Qu.:  38.383   3rd Qu.:36.00   Group5:19  
 Max.   :2323.700   Max.    :38.00   Group6:15  
                                    Group7: 8  
```



Taking the log of the y scale results in a fairly linear relationship

Distribution of response before (left) and after (right) log transformation



Right-skewed response seems to be "normalized" after the log trnsformation

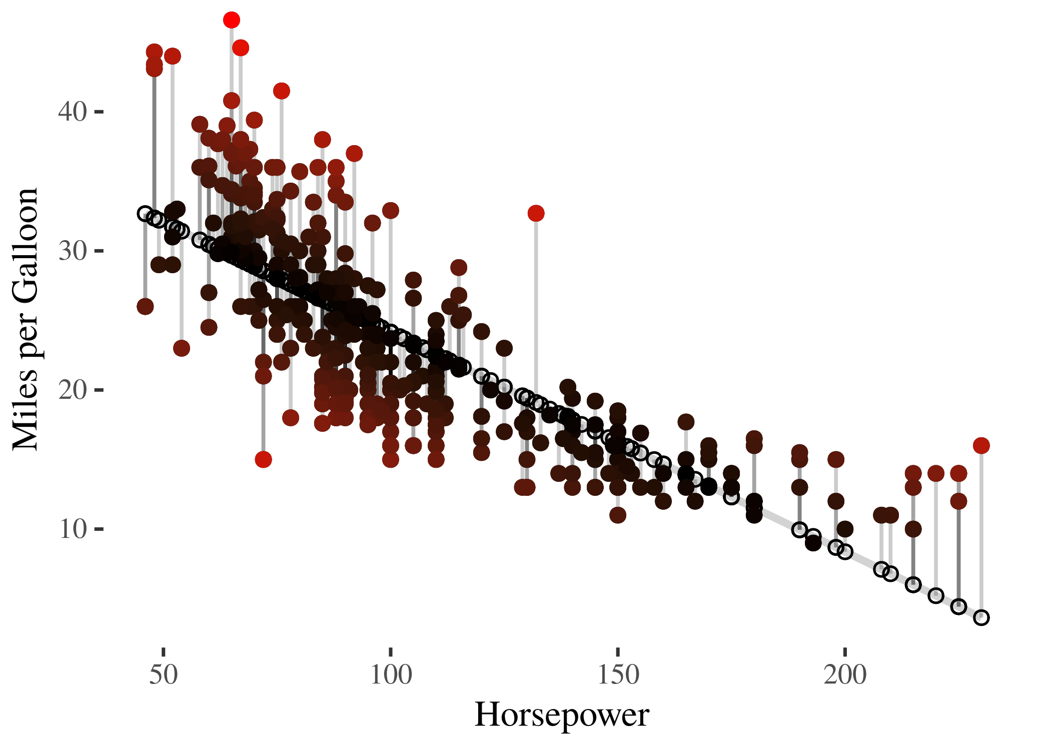
Case Study 3: `autompg` dataset

```r
library(dplyr)
glimpse(autompg)
```

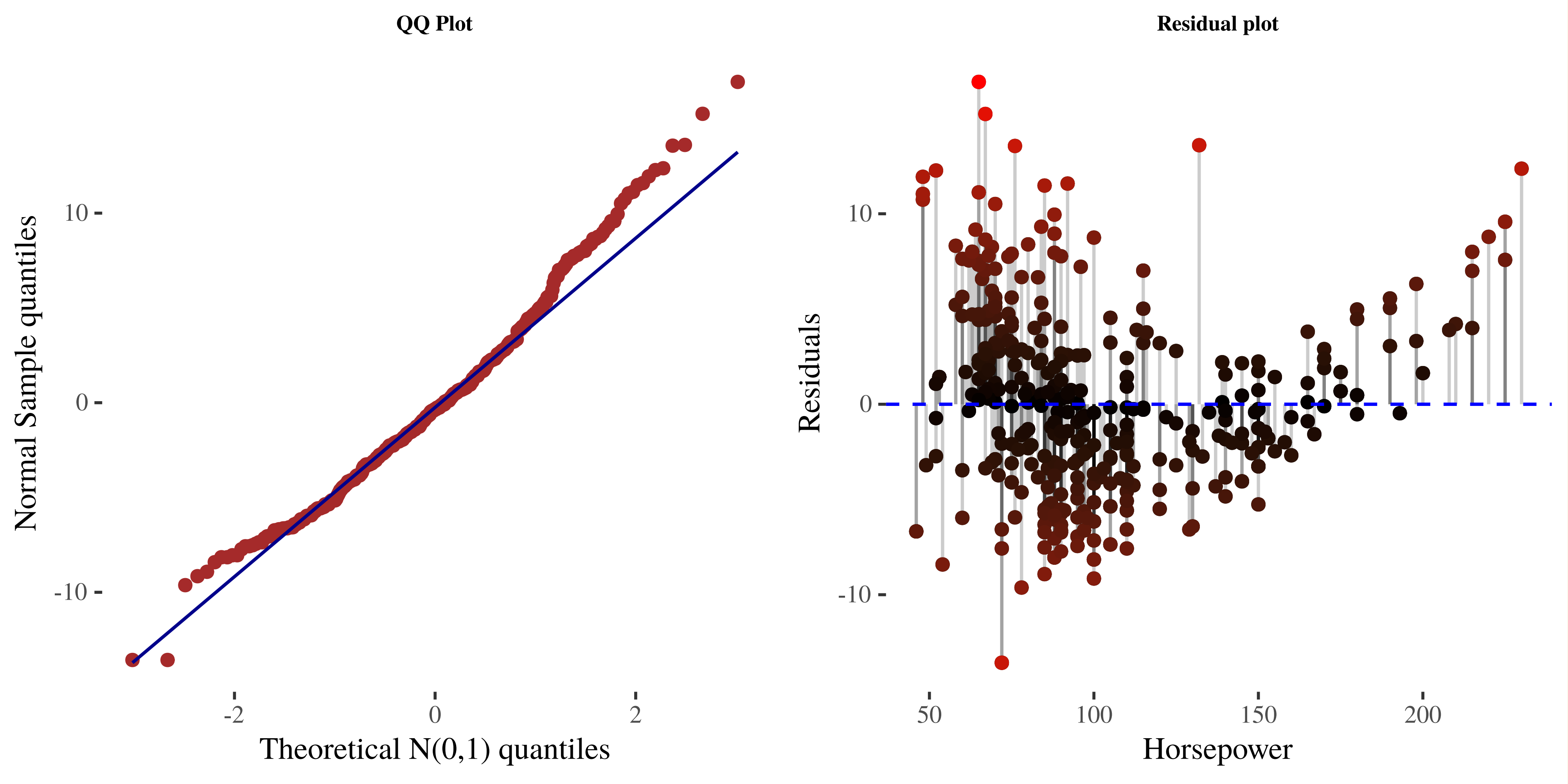
```
Rows: 398
Columns: 9
$ mpg          <dbl> 18, 15, 18, 16, 17, 15, 14, 14, 14, 15, 15, 14, 15, 14, 2…
$ cylinders    <dbl> 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 4, 6, 6, 6, 4, …
$ displacement <dbl> 307, 350, 318, 304, 302, 429, 454, 440, 455, 390, 383, 34…
$ horsepower   <dbl> 130, 165, 150, 150, 140, 198, 220, 215, 225, 190, 170, 16…
$ weight       <dbl> 3504, 3693, 3436, 3433, 3449, 4341, 4354, 4312, 4425, 385…
$ acceleration <dbl> 12.0, 11.5, 11.0, 12.0, 10.5, 10.0, 9.0, 8.5, 10.0, 8.5, …
$ `model year` <dbl> 70, 70, 70, 70, 70, 70, 70, 70, 70, 70, 70, 70, 70, 70, 7…
$ origin       <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 3, 1, 1, 1, 3, …
$ `car name`   <chr> "chevrolet chevelle malibu", "buick skylark 320", "plymou…
```

UCI Machine Learning (2017). Auto-mpg dataset [Dataset]. https://www.kaggle.com/uciml/autompg-dataset

```
lm_1 <- lm(mpg ~ horsepower, data = automp

# Get regression table
reg_table_auto_mpg_hp <-
  get_regression_table(lm_1)

# Get regression points
reg_points_auto_mpg_hp <-
  get_regression_points(lm_1)
```



The residuals are unevenly scattered for different values of the predictor.

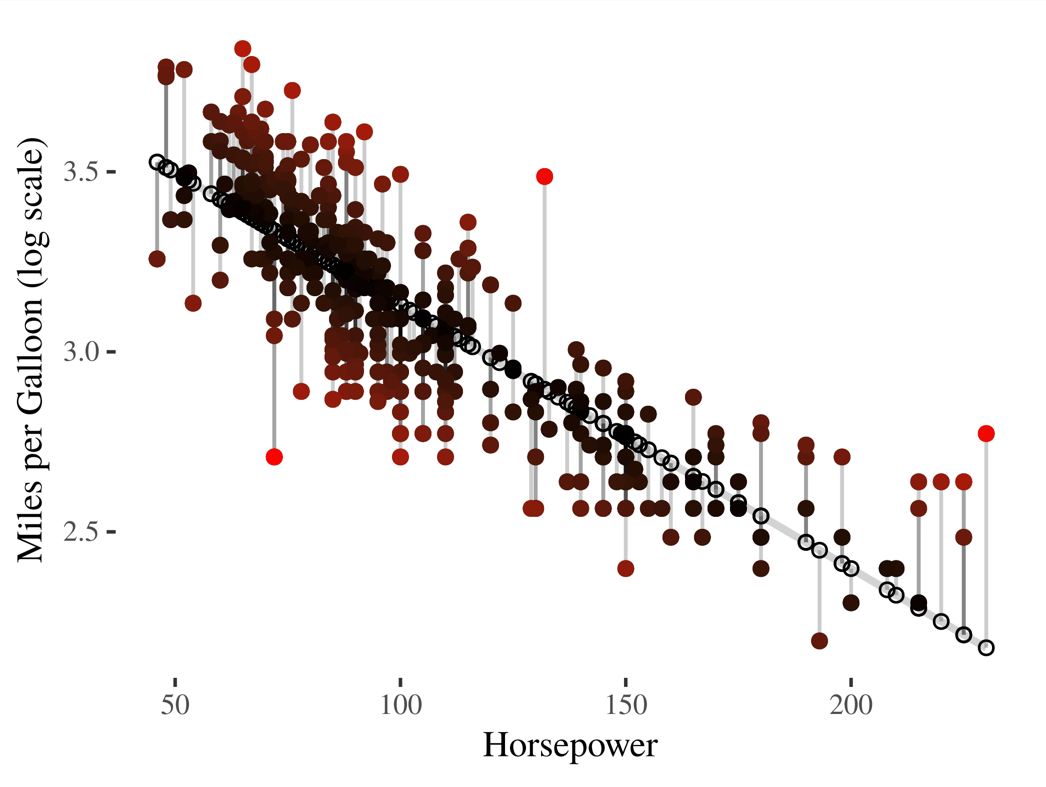


The constant variance assumptions seems to have been violated. Q-Q plot shows some deviations.

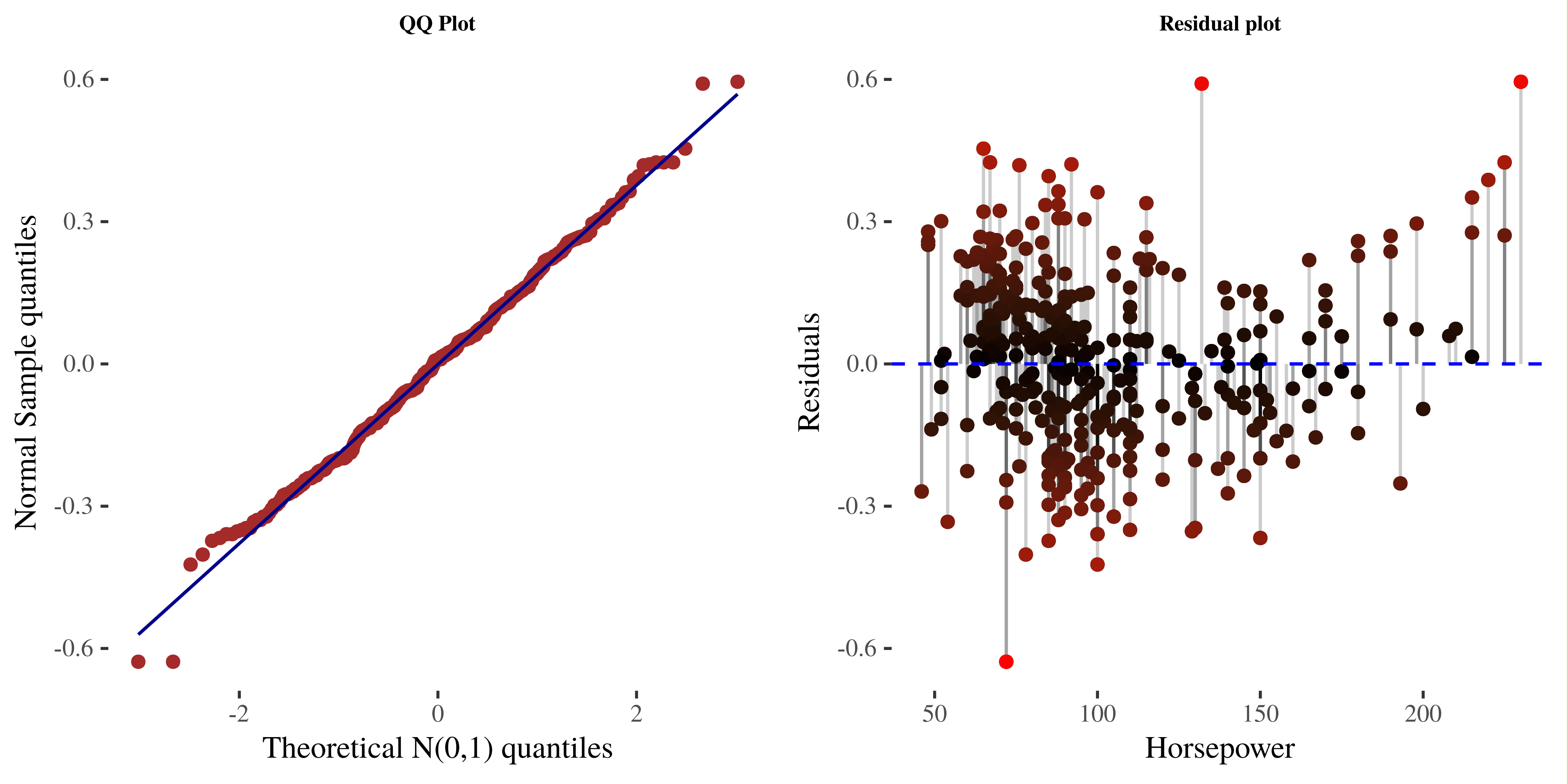
Transforming the response

```r
# make a new variable with mutate
autompg <-  autompg %>%
  mutate(logmpg = log(mpg))
```

```r
lm_2 <- lm(logmpg ~ horsepower, data = auto

# Get regression table
reg_table_auto_logmpg_hp <-
  get_regression_table(lm_2)

# Get regression points
reg_points_auto_logmpg_hp <-
  get_regression_points(lm_2)
```



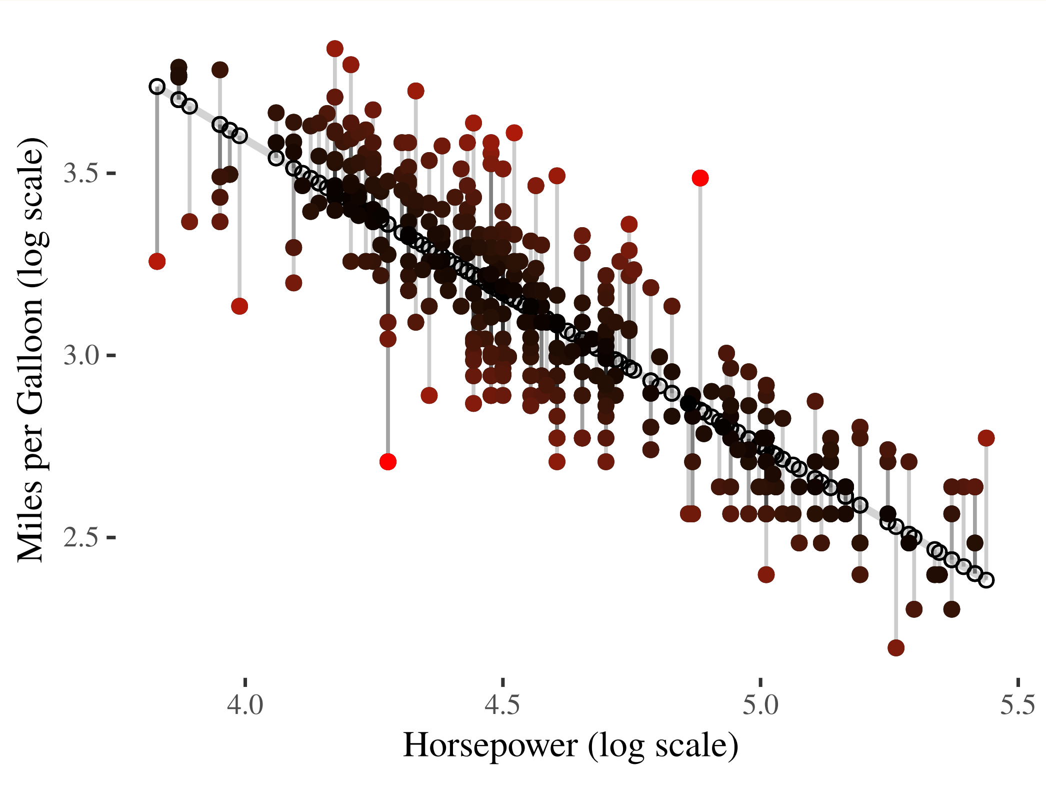
The plot looks a lot more linear than before after the log transformation of the response, but pattern in the residuals persists



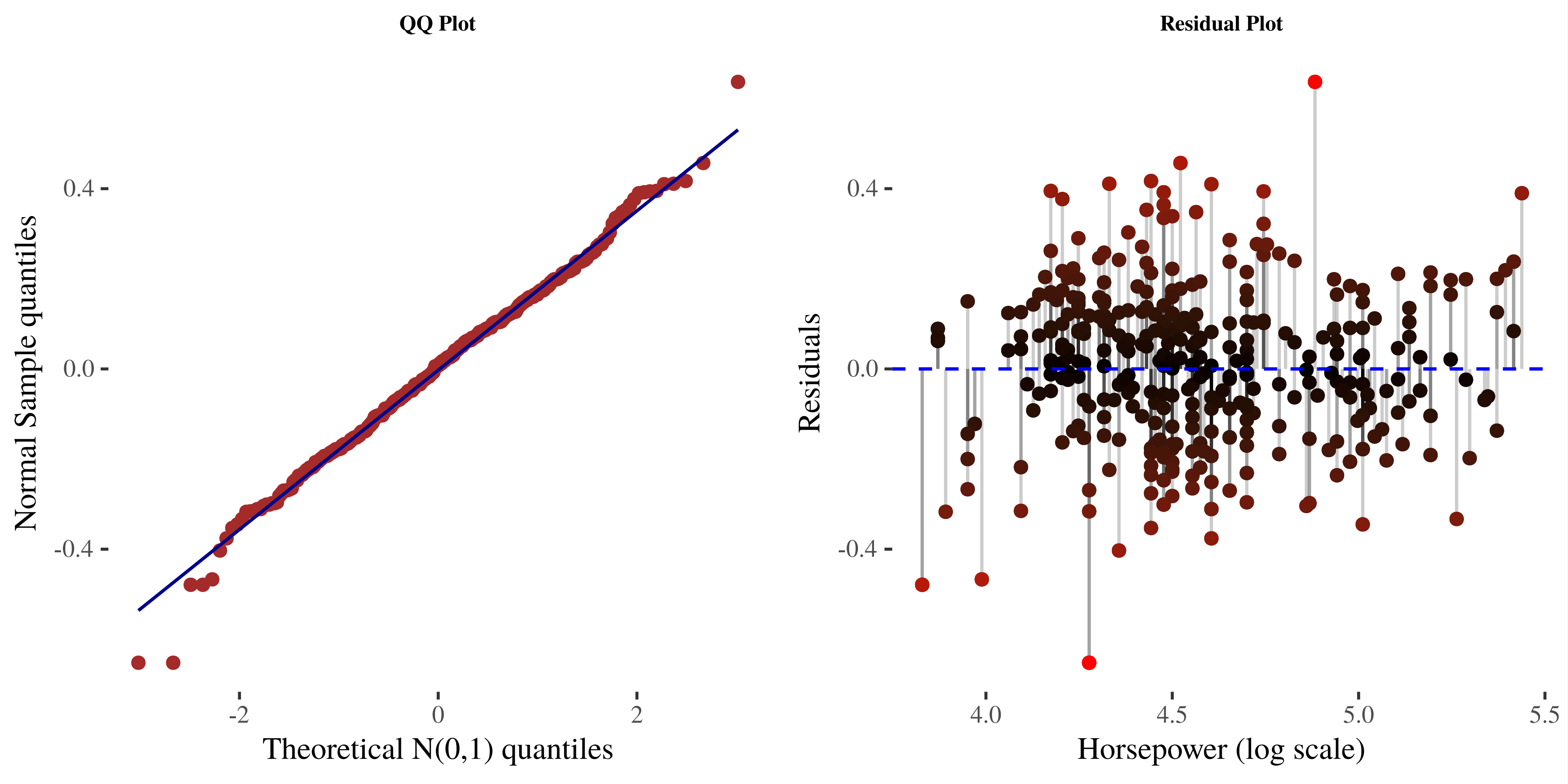
Q-Q plot has improved. We still see non-random pattern in the residual plot. So, a transformation of predictor would help.

Transforming both response and predictor

```
autompg <-  autompg %>%
  mutate(loghp = log(horsepower))

lm_3 <- lm(logmpg ~ loghp, data = autompg)

# Get regression table
reg_table_auto_logmpg_loghp <- get_regress

# Get regression points
reg_points_auto_logmpg_loghp <- get_regres
```

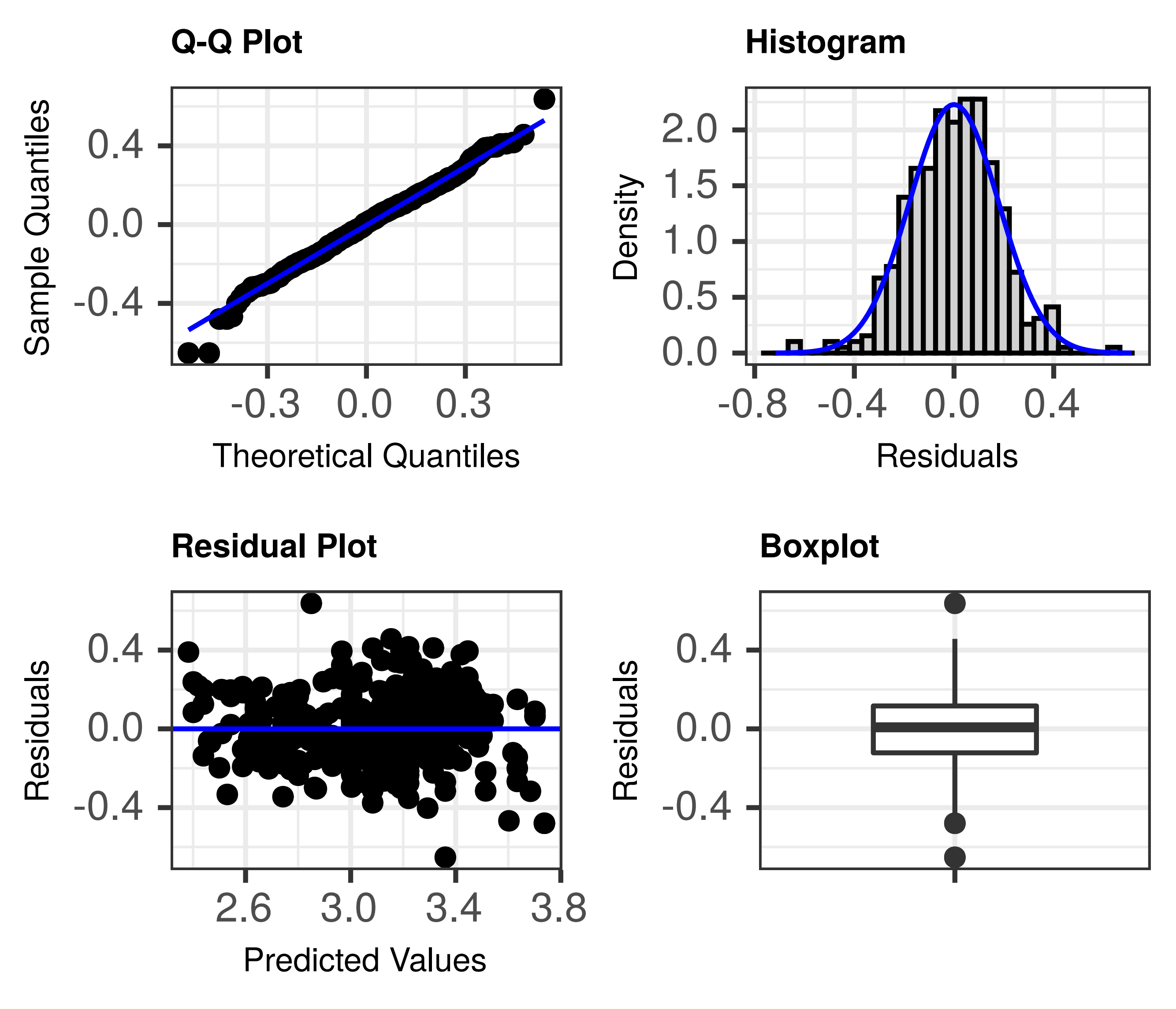


The linearity assumption seems to be satisfied



Finally, both the residual plot and the Q-Q plot look good.

Diagnostics using `ggResidpanel`



Example: autompg

$$\hat{\mu}(\,\log(\text{miles per gallon}) \mid \log(\text{horsepower})) = 6.961 - 0.842\,\log(\text{horsepower})$$

```r
library(moderndive)

# Get regression table
regression_table <- get_regression_table(lm_3)
knitr::kable(regression_table, digits = 4, format = "html")
```

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	6.961	0.121	57.296	0	6.722	7.199
loghp	-0.842	0.026	-31.881	0	-0.894	-0.790

How do we interpret the parameters in the original response scale?

Brief interlude and practice session to brush up your power, exponential, and logarithm knowledge!!

Logarithm Transformations

The logarithm base b of $x > 0$ is denoted $\log_b(x)$ and equal to

$$\log_b(x) = a$$

where a tells us what power we must raise b to to obtain the value x :

$$b^a = x$$

- $\log_{10} 10 = 1$ since $10^1 = 10$
- $\log_{10} 100 = 2$ since $10^2 = 100$
- $\log_2 0.5 = -1$ since $2^{-1} = 0.5$
- $\log_b 1 = 0$ since $b^0 = 1$

Logarithm Transformations

Multiplicative changes in x result in additive changes in $\log_b(x)$. If $m > 0$, then

$$\log_b(mx) = \log_b(m) + \log_b(x)$$

- We can "remove" a log transformation by applying the inverse function:

  $$b^{\log_b(x)} = x$$

  - e.g. $e^{\ln y} = y$

Logarithm model types

Logarithmic model: The regression of Y on $\log(x)$ has the mean function

$$\mu(Y \mid \log(x)) = \beta_0 + \beta_1 \log(x)$$

Multiplying x by a factor m is associated with a mean function change of

$$\mu(Y \mid \log(mx)) - \mu(Y \mid \log(x)) = \beta_1 \log(m)$$

- e.g. Using base-2: a doubling of $x (m = 2)$ is associated with a change in mean y of

$$\beta_1 \log_2(2) = \beta_1$$

Logarithm of the response

Suppose $\log(y) \mid x \sim N\left(\mu_{\log(y)|x}, \sigma\right)$

- By symmetry of the normal model,

$$\mu(\log(Y) \mid x) = \text{median}(\log(Y) \mid x)$$

- Because median only depends on the order of responses,

$$\text{median}(\log(Y) \mid x) = \log(\text{median}(Y \mid x))$$

- Putting these two facts together (using natural log)

$$\text{median}(Y \mid x) = e^{\log(\text{median}(Y|x))} = e^{\mu(\log(Y)|x)}$$

- Note $\text{median}(Y \mid x) \neq \mu(Y \mid x)$

Logarithm model types

- **Exponential model:** The regression of $\log(Y)$ on x has the mean function

$$\mu(\log(Y) \mid x) = \text{median}(\log(Y) \mid x) = \beta_0 + \beta_1 x$$

- On the original response scale,

$$\text{median}(Y \mid x) = e^{\mu(\log(Y)|x)} = e^{\beta_0} e^{\beta_1 x}$$

- A one unit increase in x is associated with a e^{β_1}-factor change in the median response

$$\text{median}(Y \mid x+1) = e^{\beta_0} e^{\beta_1(x+1)} = e^{\beta_0} e^{\beta_1 x} e^{\beta_1} = \text{median}(Y \mid x) e^{\beta_1}$$

- This result assumes a natural log transformation of y. If log base b is used, the one unit change in x effect is b^{β_1}.

Logarithm model types

- **Power model:** The regression of $\log(Y)$ on $\log(x)$ has the mean function

$$\mu(\log(Y) \mid \log(x)) = \text{median}(\log(Y) \mid \log(x)) = \beta_0 + \beta_1 \log(x)$$

- On the original response scale,

$$\text{median}(Y \mid x) = e^{\mu(\log(Y)|\log(x))} = e^{\beta_0} e^{\beta_1 \log(x)} = e^{\beta_0} x^{\beta_1}$$

- An m-fold (multiplicative) change in x is associated with a m^{β_1}-factor change in the median function since

$$\text{median}(Y \mid mx) = e^{\beta_0} e^{\beta_1 \log(mx)} = \text{median}(Y \mid x) m^{\beta_1}$$

Factor changes vs. percent changes

- A m-fold change in x is equivalent to a $(m-1)100\%$ percentage change

$$\begin{aligned}\text{percentage change} &= 100\% \times \frac{mx - x}{x} \\ &= 100\% \times \left(\frac{mx}{x} - \frac{x}{x}\right) \\ &= 100\% \times (m-1)\end{aligned}$$

Interpretation after log transformations (Summary)

Use $\text{median}(Y|x)$:

- When X is logged: a **doubling of** X is associated with adding a constant amount to Y.
- When only Y is logged: When **X increases by 1**, Y's median is multiplied.
- When both X and Y are logged: a **doubling of X** is associated with multiplying Y's median by a constant factor.

autompg log-log transformation power model interpretation

$$\hat{\mu}(\ \log(\text{miles per gallon}) \mid \log(\text{horsepower})) = 6.961 - 0.842 \log(\text{horsepower})$$

```
knitr::kable(regression_table, digits = 4, format = "html")
```

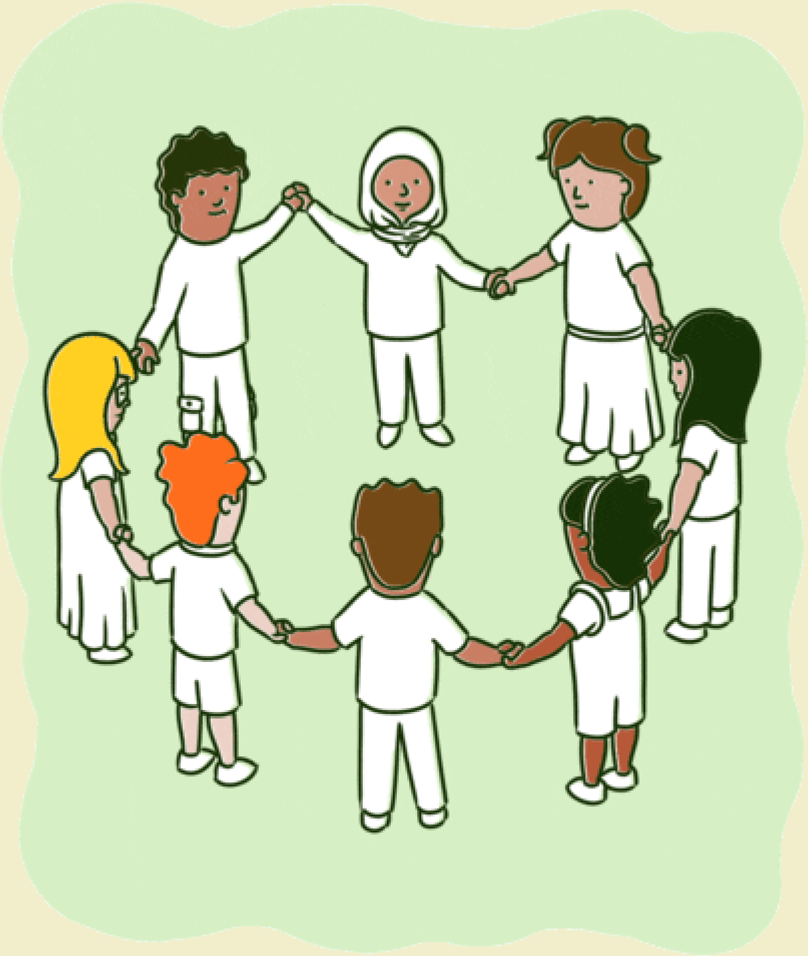
term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	6.961	0.121	57.296	0	6.722	7.199
loghp	-0.842	0.026	-31.881	0	-0.894	-0.790

We are 95% confident that a factor of 2 increase in horsepower of a car is associated with a reduction of 0.5382 to 0.5784 fold of median miles per gallon.

```
2^confint(lm_3)
                  2.5 %      97.5 %
(Intercept) 105.5512084 146.9823446
loghp         0.5382086   0.5783716
```

Your Turn 1

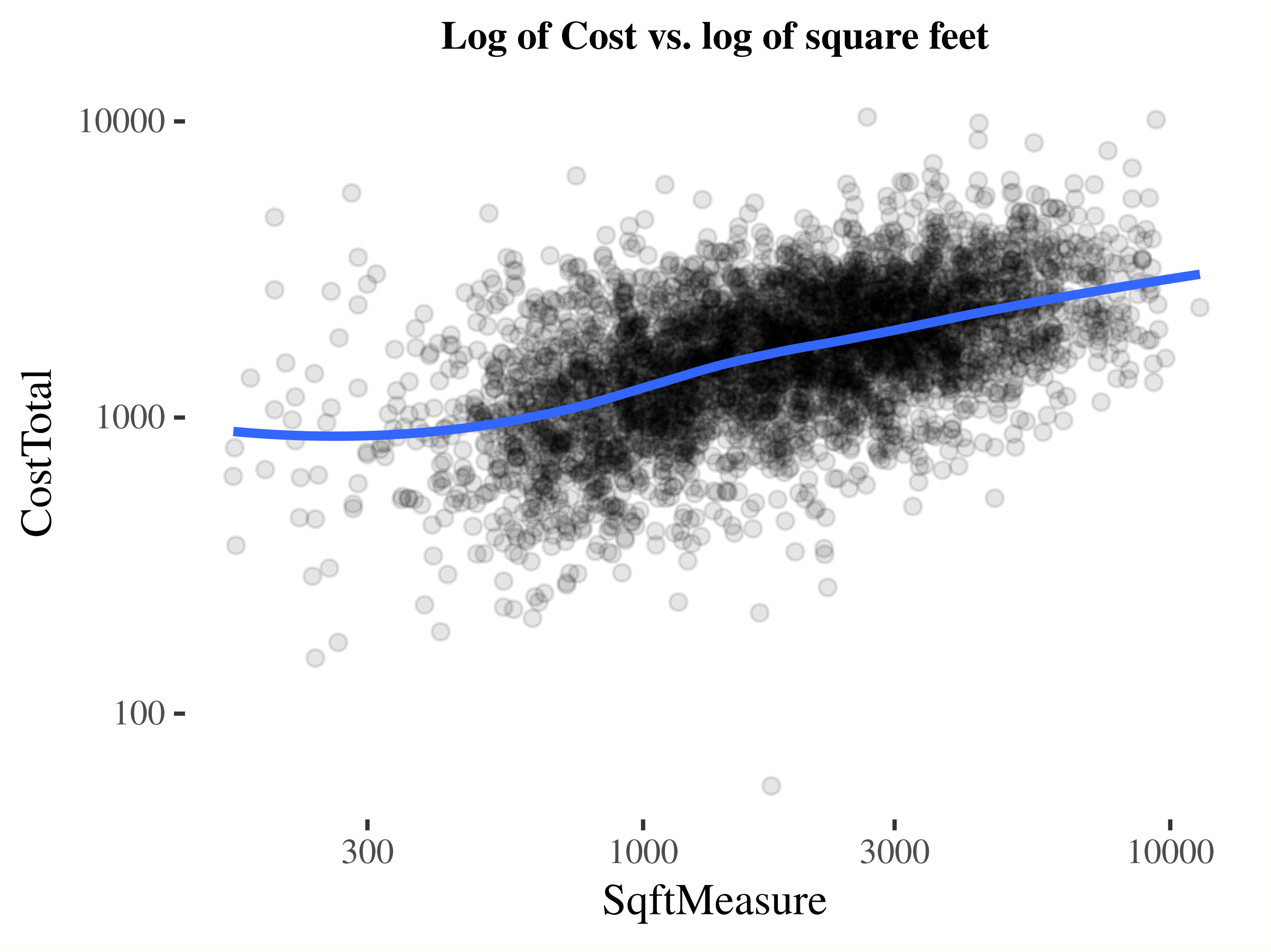
05:00



- Get the in class activity file from moodle
- We will do a case study of residential energy survey
- Please skim through the activity .Rmd file in your group

RECS

```
ggplot(energy, aes(x = SqftMeasure, y = CostTotal))
 geom_point(alpha = .1) +
 geom_smooth(method = "loess", se = FALSE) +
 scale_x_log10() +
 scale_y_log10() +
 labs(title = "Log of Cost vs. log of square feet")
theme(plot.title = element_text(hjust=0.5, size=9,
```



RECS

```r
cost_lm <- lm(log(CostTotal) ~ log(SqftMeasure), data = energy)
regression_table <- get_regression_table(cost_lm)
knitr::kable(regression_table, digits = 4)
```

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	4.358	0.074	59.083	0	4.214	4.503
log(SqftMeasure)	0.403	0.010	41.135	0	0.384	0.422

- What is the fitted model?
- How does cost change if house size is doubled? Get a CI for this effect.
- How does cost change if house size decreases by 10% ? Get a CI for the effect.