

R03 - Regression: using logarithms

STAT 5870 (Engineering)
Iowa State University

August 28, 2024

Parameter interpretation in regression

If

$$E[Y|X] = \beta_0 + \beta_1 X,$$

then

- β_0 is the expected response when X is zero and
- $d\beta_1$ is the expected change in the response for a d unit change in the explanatory variable.

For the following discussion,

- Y is always going to be the **original** response and
- X is always going to be the **original** explanatory variable.

Corn yield example

Suppose

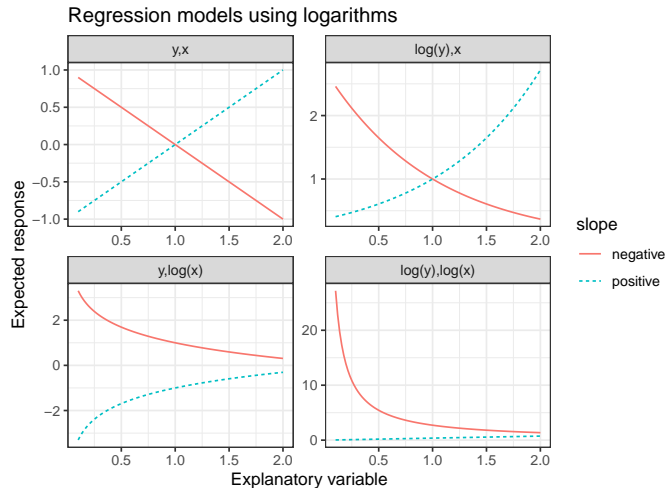
- Y is corn yield (bushels/acre)
- X is fertilizer level in lbs/acre

Then, if

$$E[Y|X] = \beta_0 + \beta_1 X$$

- β_0 is the **expected** corn yield (bushels/acre) when fertilizer level is zero and
- $d\beta_1$ is the **expected** change in corn yield (bushels/acre) when fertilizer is increased by d lbs/acre.

Regression with logarithms



Response is logged

If

$$E[\log(Y)|X] = \beta_0 + \beta_1 X,$$

then we have

$$\text{Median}[Y|X] = e^{\beta_0 + \beta_1 X} = e^{\beta_0} e^{\beta_1 X}$$

then

- e^{β_0} is the **median** of Y when X is zero
- $e^{d\beta_1}$ is the **multiplicative change** in the **median** of Y for a d unit change in the explanatory variable.

Response is logged

Let Y be corn yield (bushels/acre) and X is fertilizer level in lbs/acre.
If we assume

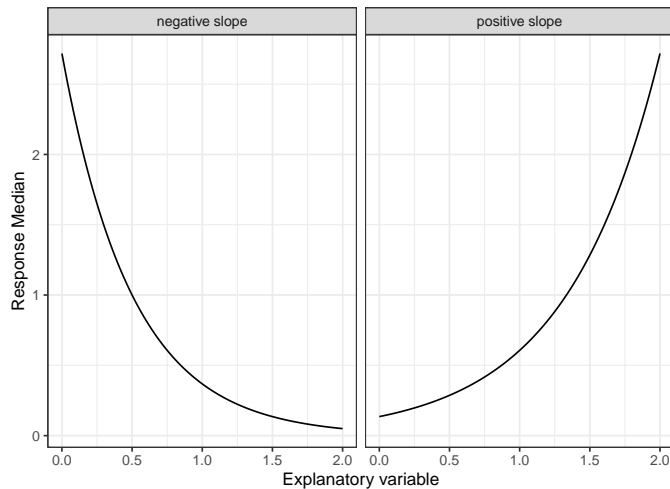
$$E[\log(Y)|X] = \beta_0 + \beta_1 X$$

then

$$\text{Median}[Y|X] = e^{\beta_0} e^{\beta_1 X}$$

- e^{β_0} is the **median** corn yield (bushels/acre) when fertilizer level is 0 and
- $e^{d\beta_1}$ is the **multiplicative change** in median corn yield (bushels/acre) when fertilizer is increased by d lbs/acre.

Response is logged



Explanatory variable is logged

If

$$E[Y|X] = \beta_0 + \beta_1 \log(X),$$

then,

- β_0 is the **expected** response when X is 1 and
- $\beta_1 \log(d)$ is the **expected** change in the response when X increases **multiplicatively** by d , e.g.
 - $\beta_1 \log(2)$ is the **expected** change in the response for each **doubling** of X or
 - $\beta_1 \log(10)$ is the **expected** change in the response for each **ten-fold increase** in X .

Explanatory variable is logged

Suppose

- Y is corn yield (bushels/acre)
- X is fertilizer level in lbs/acre

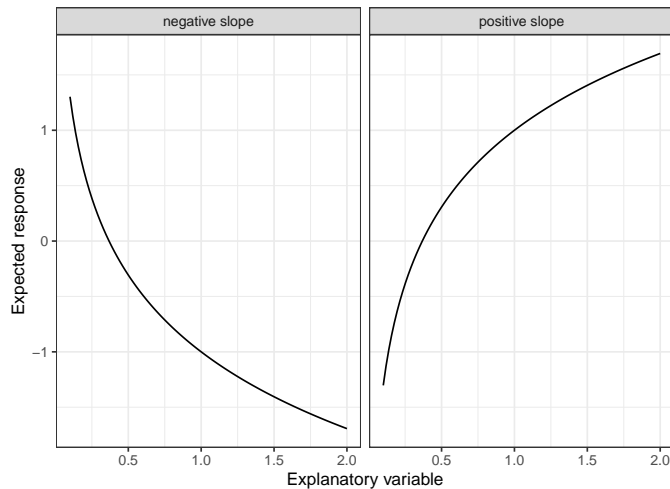
If

$$E[Y|X] = \beta_0 + \beta_1 \log(X)$$

then

- β_0 is the **expected** corn yield (bushels/acre) when fertilizer amount is 1 lb/acre and
- $\beta_1 \log(2)$ is the **expected** change in corn yield when fertilizer amount is **doubled**.

Explanatory variable is logged



Both response and explanatory variable are logged

If

$$E[\log(Y)|X] = \beta_0 + \beta_1 \log(X),$$

then

$$\text{Median}[Y|X] = e^{\beta_0} X^{\beta_1},$$

and thus

- e^{β_0} is the **median** of Y when X is 1 and
- d^{β_1} is the **multiplicative** change in the **median** of the response when X increases **multiplicatively** by d , e.g.
 - 2^{β_1} is the **multiplicative** change in the **median** of the response for each **doubling** of X or
 - 10^{β_1} is the **multiplicative** change in the **median** of the response for each **ten-fold increase** in X .

Both response and explanatory variables are logged

Suppose

- Y is corn yield (bushels/acre)
- X is fertilizer level in lbs/acre

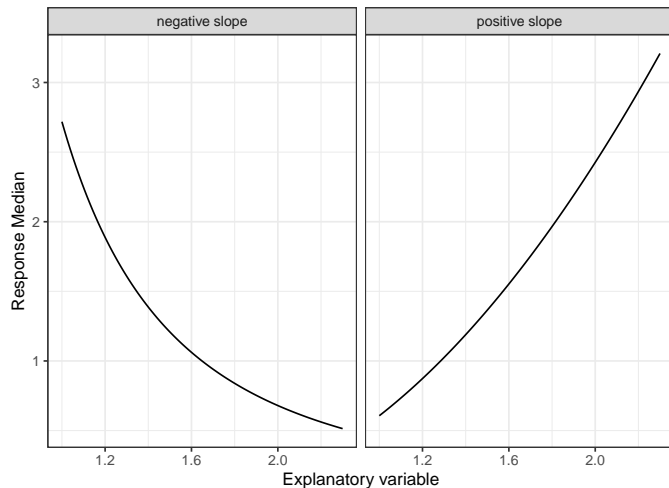
If

$$E[\log(Y)|X] = \beta_0 + \beta_1 \log(X) \quad \text{or} \quad \text{Median}[Y|X] = e^{\beta_0} e^{\beta_1 \log(X)} = e^{\beta_0} X^{\beta_1},$$

then

- e^{β_0} is the **median** corn yield (bushels/acre) at 1 lb/acre of fertilizer and
- 2^{β_1} is the **multiplicative change** in median corn yield (bushels/acre) when fertilizer is **doubled**.

Both response and explanatory variables are logged



Why use logarithms

The most common transformation of either the response or explanatory variable(s) is to take logarithms because

- linearity will often then be approximately true,
- the variance will likely be approximately constant,
- influence of some observations may decrease, and
- there is a (relatively) convenient interpretation.

Summary of interpretations when using logarithms

- When using the log of the response,
 - β_0 determines the **median** response
 - β_1 determines the **multiplicative** change in the median response
- When using the log of the explanatory variable (X),
 - β_0 determines the response when $X = 1$
 - β_1 determines the change in the response when there is a **multiplicative** increase in X

Constructing credible intervals

Recall the model

$$Y_i \stackrel{ind}{\sim} N(\beta_0 + \beta_1 X_i, \sigma^2).$$

Let (L, U) be a $100(1 - a)\%$ credible interval for β .

For ease of interpretation, it is often convenient to calculate functions of β , e.g.

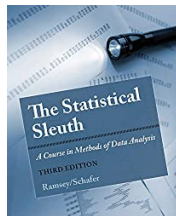
$$f(\beta) = d\beta \quad \text{and} \quad f(\beta) = e^\beta.$$

A $100(1 - a)\%$ credible interval for $f(\beta)$ (when f is monotonic) is

$$(f(L), f(U)).$$

Breakdown times

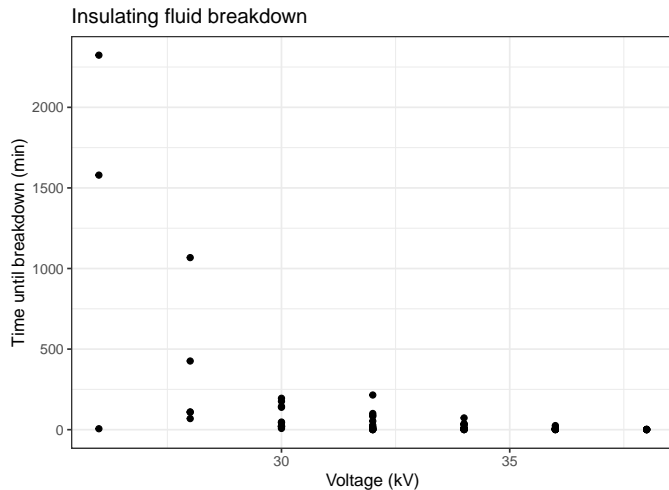
In an industrial laboratory, under uniform conditions, batches of electrical insulating fluid were subjected to constant voltages (kV) until the insulating property of the fluids broke down. Seven different voltage levels were studied and the measured responses were the times (minutes) until breakdown.



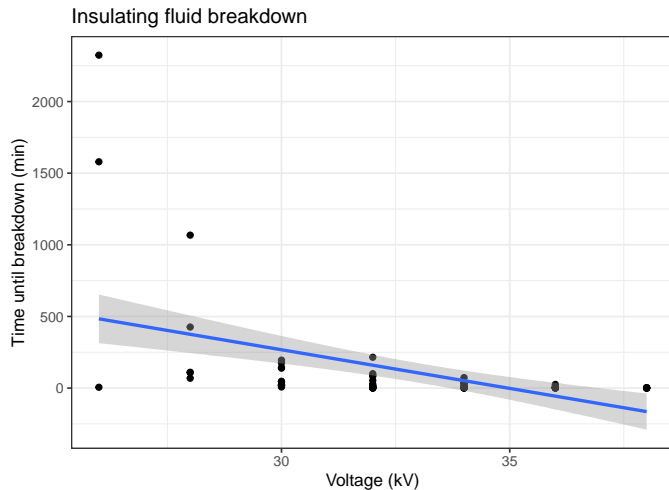
```
summary(Sleuth3::case0802)
```

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

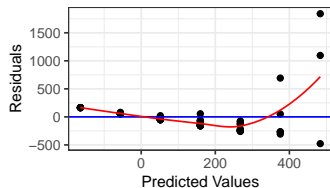
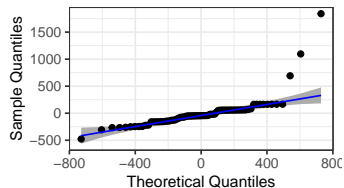
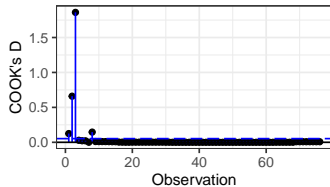
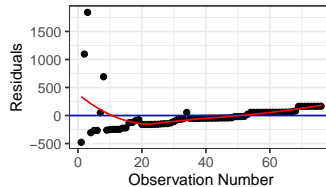
Insulating fluid breakdown



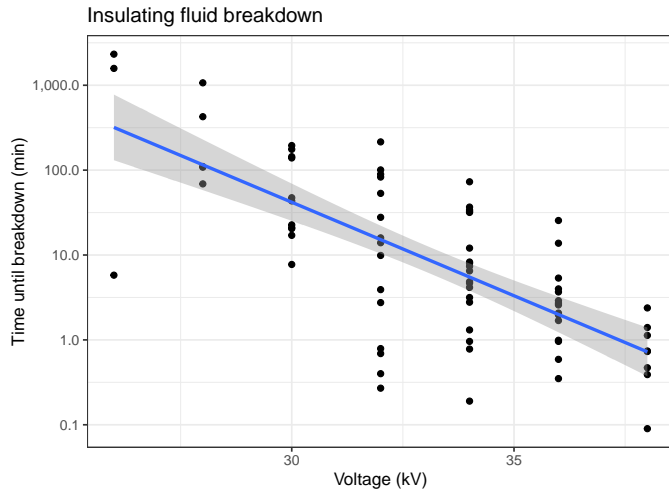
Insulating fluid breakdown



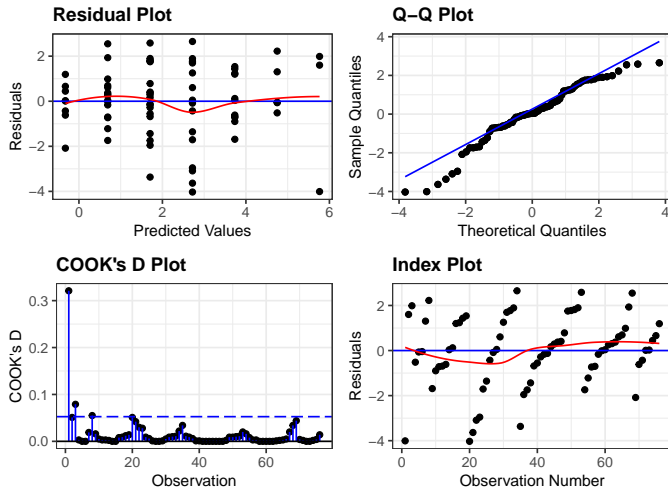
Run the regression and look at diagnostics

Residual Plot**Q-Q Plot****COOK's D Plot****Index Plot**

Logarithm of time (response)



Logarithm of time (response): residuals



Summary

```
m <- lm(log(Time) ~ I(Voltage-30), Sleuth3::case0802)
exp(m$coefficients)
```

```
(Intercept) I(Voltage - 30)
41.86752      0.60208
```

```
exp(confint(m))
```

```
              2.5 %      97.5 %
(Intercept) 25.2582342 69.3987157
I(Voltage - 30) 0.5370152 0.6750281
```

- At 30 kV, the median breakdown time is estimated to be 42 minutes with a 95% credible interval of (25, 69).
- Each 1 kV increase in voltage was associated with a 40% (32%, 46%) reduction in median breakdown time.