

R03 - Using logarithms in regression

STAT 587 (Engineering) - Iowa State University

April 2, 2019

Parameter interpretation in regression

If

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

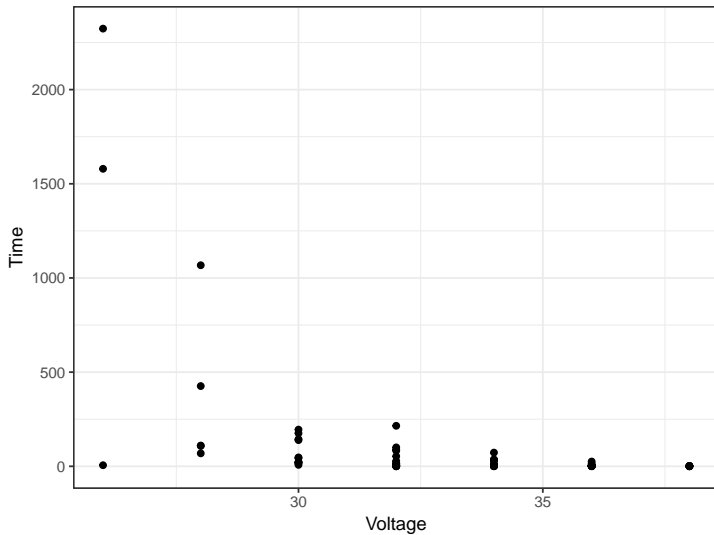
then

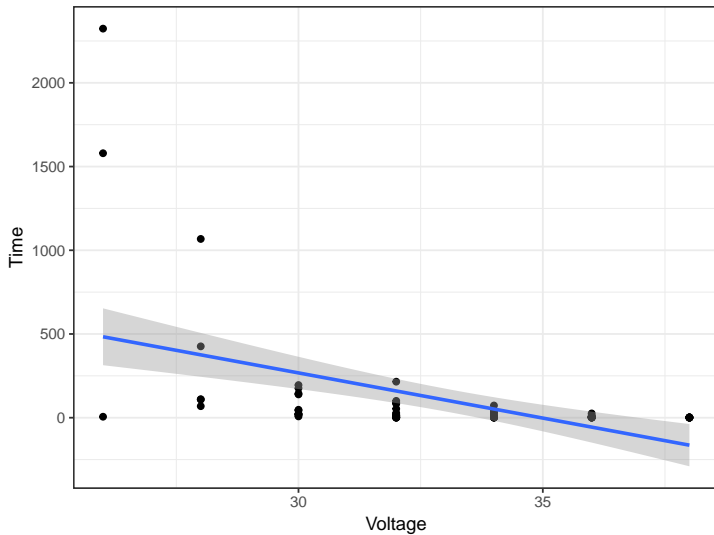
- β_0 is the expected response when X is zero and
- β_1 is the expected increase in the response for a one unit increase in the explanatory variable.
- $d\beta_1$ is the expected increase in the response for a d unit increase in the explanatory variable.

For the following discussion,

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

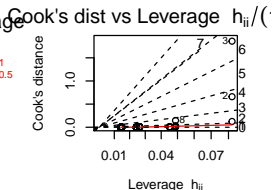
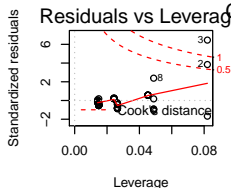
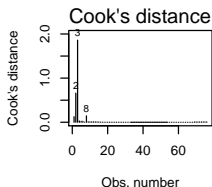
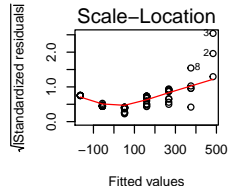
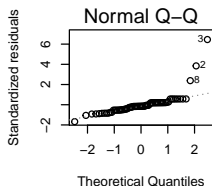
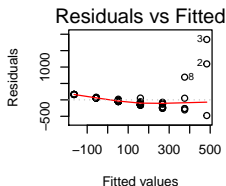
Sleuth3::case0802





Run the regression and look at diagnostics

```
m <- lm(Time ~ Voltage, insulating)
```



Interpretations using logs

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.

We will talk about interpretation of β_0 and β_1 when

- only the response is logged,
- only the explanatory variable is logged, and
- when both are logged.

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
- β_1 is the expected change in corn yield (bushels/acre) when fertilizer is increased by 1 lb/acre or
- $d\beta_1$ is the expected change in corn yield (bushels/acre) when fertilizer is increased by d lbs/acre.

Response is logged

If

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

then

- β_0 is the expected $\log(Y)$ when X is zero,
- β_1 is the expected change in $\log(Y)$ for a one unit increase in the explanatory variable, and
- $d\beta_1$ is the expected change in $\log(Y)$ for a d unit increase in the explanatory variable.

But since

$$E[\log(Y)|X] = \text{Median}[\log(Y)|X] = \log(\text{Median}[Y|X])$$

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^{β_1} is the multiplicative effect on the median of Y for a one unit increase in the explanatory variable, and
- $e^{d\beta_1}$ is the multiplicative effect on the median of Y for a d unit increase in the explanatory variable.

Response is logged

Suppose

- Y is corn yield (bushels/acre)
- 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,
- e^{β_1} is the multiplicative effect in median corn yield (bushels/acre) when fertilizer is increased by 1 lb/acre, and
- $e^{d\beta_1}$ is the multiplicative effect in median corn yield (bushels/acre) when fertilizer is increased by d lbs/acre.

Response is logged

```
Error in gather(., slope, 'Response Median', negative, positive): could not find function "gather"  
Error in ggplot(d, aes(x, 'Response Median')): object 'd' not found
```

Explanatory variable is logged

If

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

then

- β_0 is the expected response when $\log(X)$ is zero and
- β_1 is the expected change in the response for a one unit increase in $\log(X)$

Alternatively,

- 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

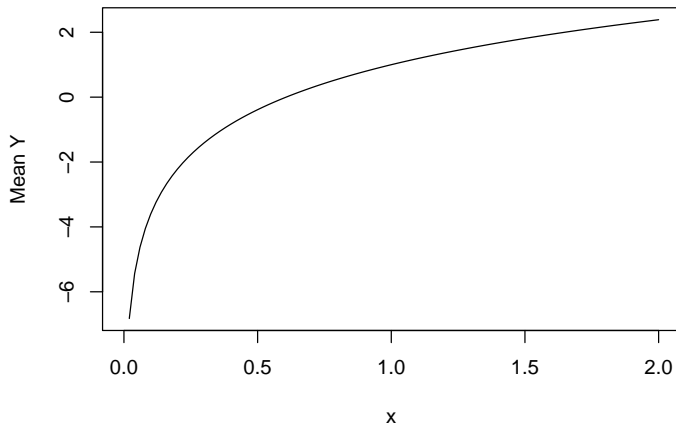
- Y is corn yield (bushels/acre)
- X is fertilizer level in lbs/acre

Then, if

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

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

Explanatory variable is logged



Both response and explanatory variable are logged

If we assume

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

then

- β_0 is the expected $\log(Y)$ when $\log(X)$ is zero and
- β_1 is the expected change in $\log(Y)$ for a one unit increase in $\log(X)$.

But we also have

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

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 effect on the median of the response for each doubling of X or
 - 10^{β_1} is the multiplicative effect on the median of the response for each ten-fold increase in X .

Both response and explanatory variables are logged

Suppose

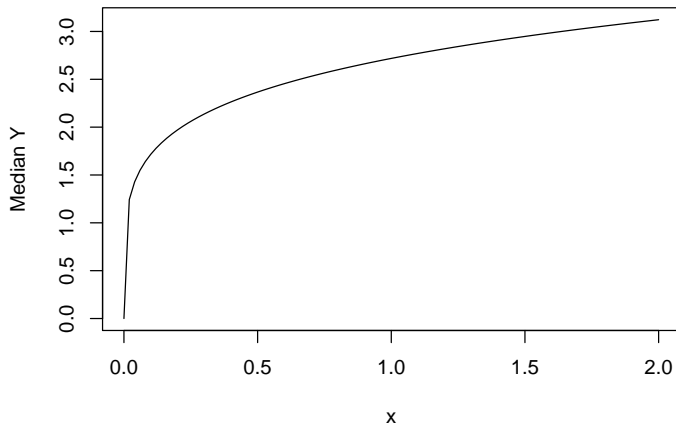
- Y is corn yield (bushels/acre)
- X is fertilizer level in lbs/acre

Then, if

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

- e^{β_0} is the median corn yield (bushels/acre) when fertilizer level is 1 lb/acre and
- 2^{β_1} is the multiplicative effect on median corn yield (bushels/acre) when fertilizer level doubles.

Both response and explanatory variables are logged



Summary of interpretations when using logarithms

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

To construct confidence intervals for $f(\beta)$ (when $f()$ is monotonic, e.g. $f(x) = dx$ or $f(x) = \exp(x)$), find a confidence interval for β and evaluate the function at the endpoints, i.e. if (L, U) is a confidence interval for β , then $(f(L), f(U))$ is a confidence interval for $f(\beta)$.

Breakdown times

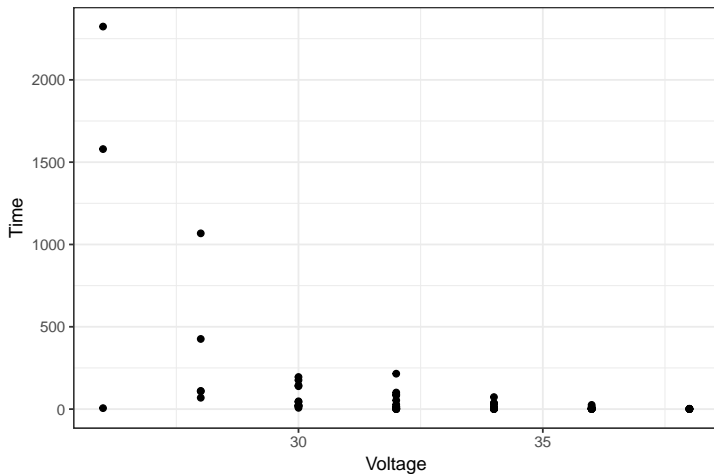
In an industrial laboratory, under uniform conditions, batches of electrical insulating fluid were subjected to constant voltages (kV) until the insulating property of the fluids broke down. Seven different voltage levels were studied and the measured responses were the times (minutes) until breakdown.

```
summary(insulating)
```

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

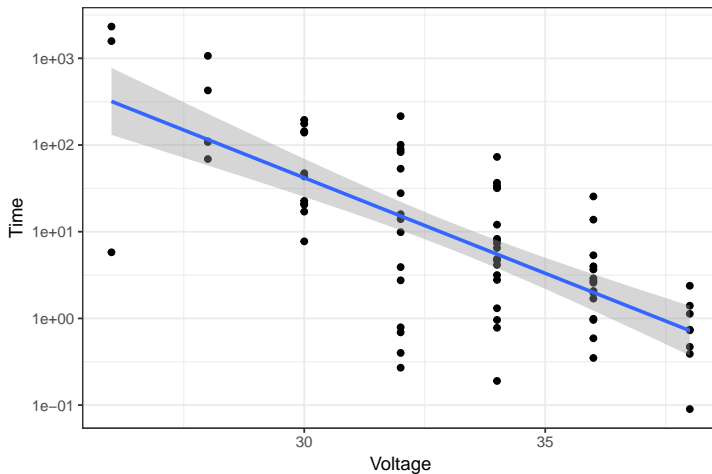
Natural scale

```
g <- ggplot(insulating, aes(Voltage, Time)) + geom_point() + theme_bw(); g
```



Take log of time

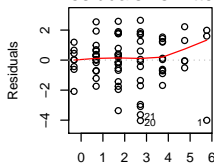
```
g + stat_smooth(method="lm") + scale_y_log10()
```



Take log of time

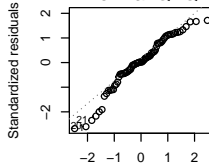
```
m <- lm(log(Time) ~ Voltage, insulating)
opar = par(mfrow=c(2,3)); plot(m, 1:6, ask=FALSE); par(opar)
```

Residuals vs Fitted



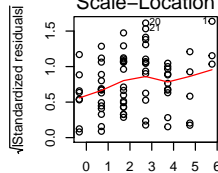
Fitted values

Normal Q-Q



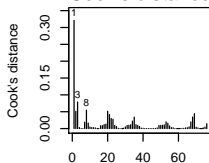
Theoretical Quantiles

Scale-Location

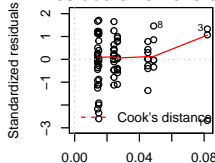


Fitted values

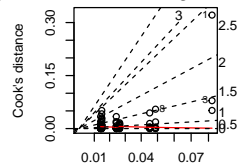
Cook's distance



Residuals vs Leverage



Cook's distance

Cook's dist vs Leverage $h_{ii}/(n-1)$ 

Summary

```
m$coefficients %>% exp
```

```
(Intercept)      Voltage
1.707069e+08 6.020800e-01
```

```
confint(m) %>% exp()
```

```
                2.5 %      97.5 %
(Intercept) 3.796778e+06 7.675154e+09
Voltage      5.370152e-01 6.750281e-01
```

```
lm(log(Time) ~ Voltage, insulating, subset= Time != 5.79) %>% confint() %>% exp() # remove first observation
```

```
                2.5 %      97.5 %
(Intercept) 1.658205e+07 3.219178e+10
Voltage      5.153150e-01 6.465834e-01
```

Summary:

Each 1 kV increase in voltage caused a multiplicative effect of 0.6 (0.5,0.7) on median breakdown time.