R03 - Using logarithms in regression

STAT 401 (Engineering) - Iowa State University

March 17, 2018

Parameter interpretation in regression

If

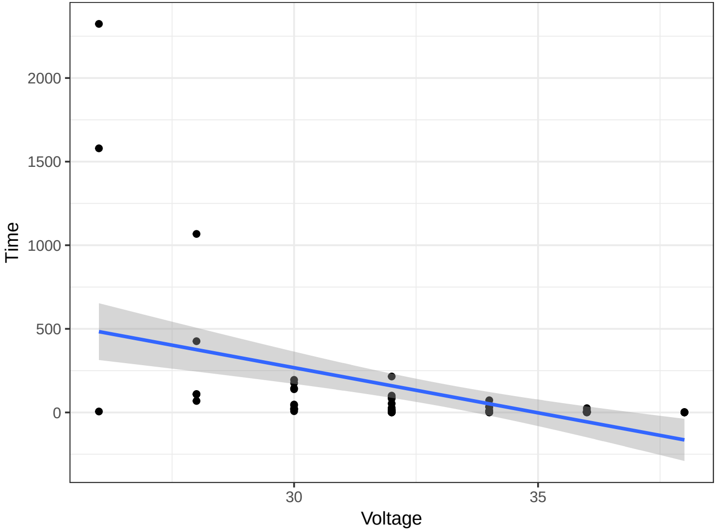
$$E[Y|X] = \beta_0 + \beta_1 X,$$

then

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

For the following discussion,

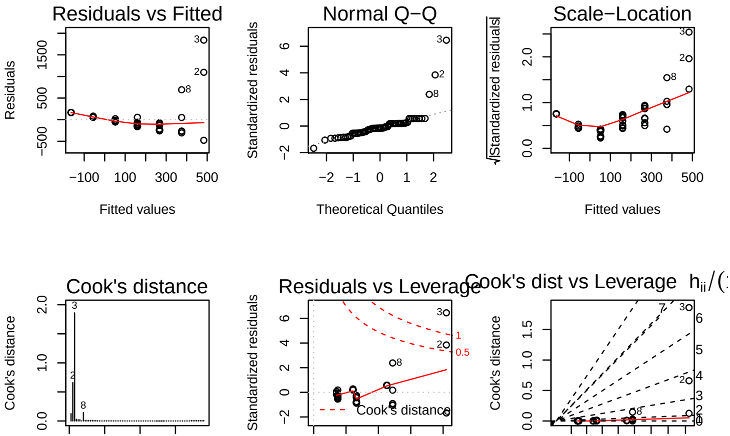
- Y is always going to be the original response and
- X is always going to be the original explanatory variable.


2000
1500
1000
500
0
Time
30
35
Voltage

Run the regression and look at diagnostics

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

```r
opar = par(mfrow=c(2,3)); plot(m, 1:6, ask=FALSE); par(opar)
```



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, 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 per 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 per acre when fertilizer level is zero and
- β_1 is the expected change in corn yield per acre when fertilizer is increase by 1 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 and
- β_1 is the expected change in $\log(Y)$ for a one unit increase in the explanatory variable.

But since

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

we have

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

then

- e^{β_0} is the median of Y when X is zero and
- e^{β_1} is the multiplicative effect on the median of Y for a one unit increase in the explanatory variable.

since

Response is logged

Suppose

- Y is corn yield per 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 per acre when fertilizer level is 0 and
- e^{β_1} is the multiplicative effect in median corn yield per acre when fertilizer is increased by 1 lbs/acre.

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 increase 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 per 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 per 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.

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

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

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 increase 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 per 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 per acre when fertilizer level is 1 lb/acre and
- 2^{β_1} is the multiplicative effect on median corn yield per acre when fertilizer level doubles.

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 e^β, find a confidence interval for β and exponentiate the endpoints, i.e. if (L, U) is a confidence interval for β, then (e^L, e^U) is a confidence interval for e^β.

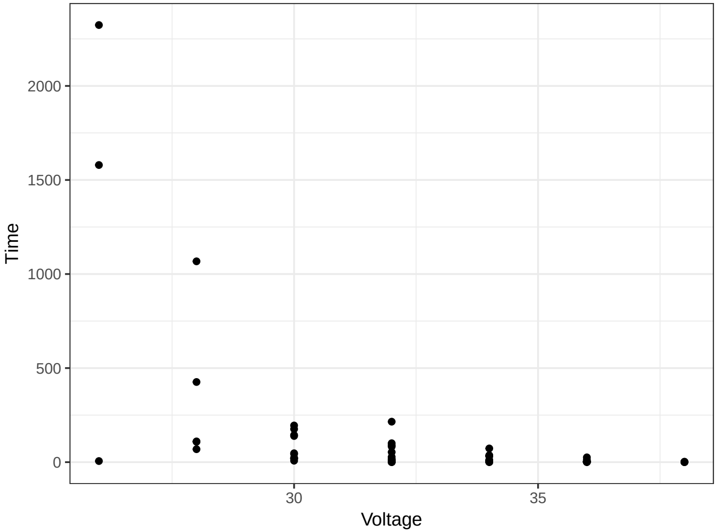
Breakdown times

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 reponses were the times (minutes) until breakdown.

```
summary(insulating)
```

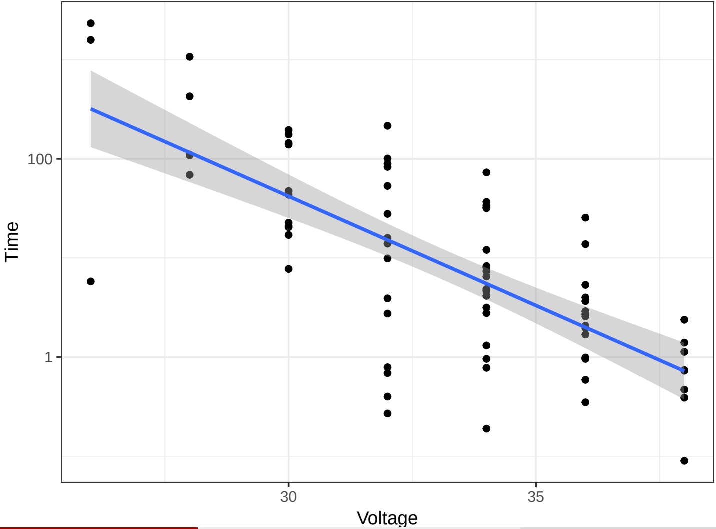
```
     Time             Voltage          Group
 Min.   :   0.090   Min.   :26.00   Group1: 3
 1st Qu.:   1.617   1st Qu.:31.50   Group2: 5
 Median :   6.925   Median :34.00   Group3:11
 Mean   :  98.558   Mean   :33.13   Group4:15
 3rd Qu.:  38.383   3rd Qu.:36.00   Group5:19
 Max.   :2323.700   Max.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

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



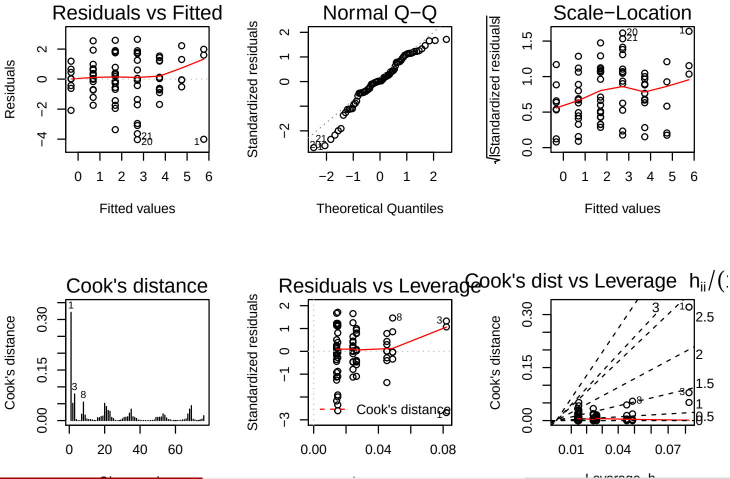
Take log of time

```r
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)
```



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.