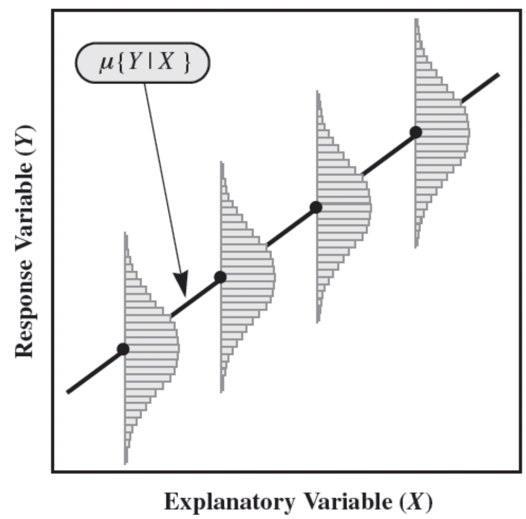
ST 411/511 Outline 8

Reading assignment: Chapter 8. This chapter continues the discussion of simple linear regression (SLR) and considers the assumptions for SLR.

Chapter 8 More on Simple Linear Regression: Assumptions and Models

Recall assumptions for SLR:

DISPLAY 7.5 The ideal normal, simple linear regression model

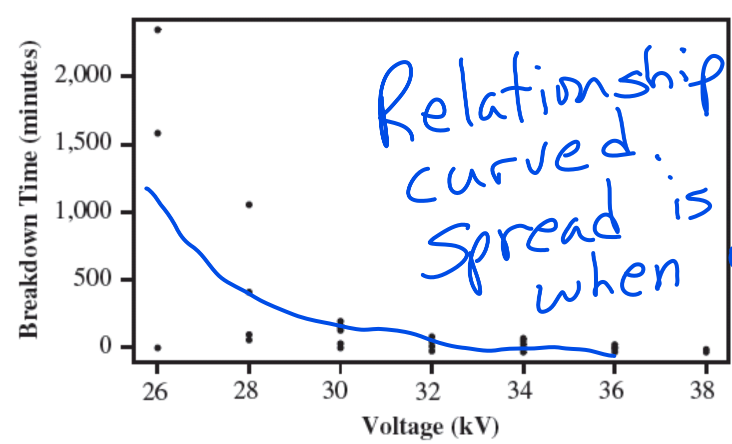


Model Assumptions

1. There is a normally distributed subpopulation of responses for each value of the explanatory variable.
2. The means of the subpopulations fall on a straight line function of the explanatory variable.
3. The subpopulation standard deviations are all equal (to σ).
4. The selection of an observation from any of the subpopulations is independent of the selection of any other observation.

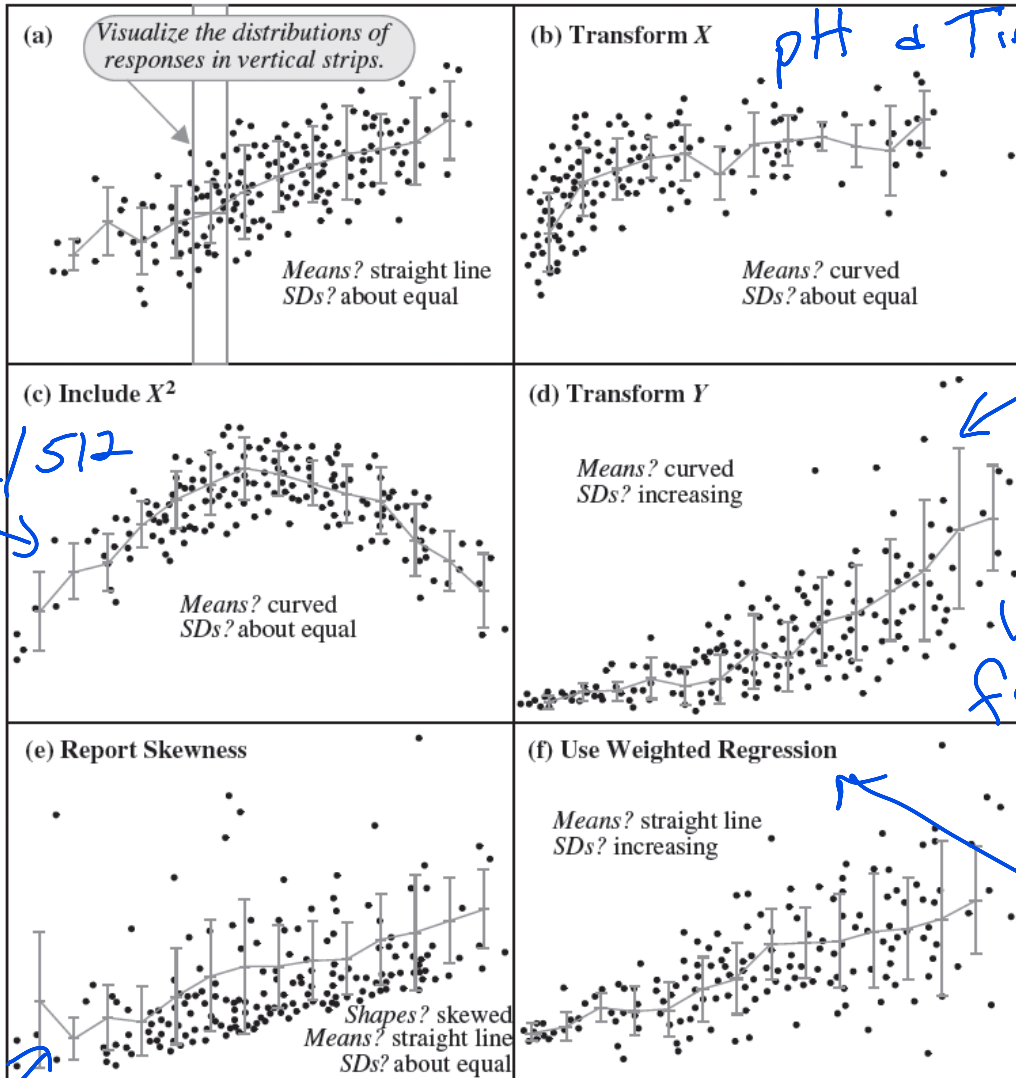
Case Study 8.1.2: Breakdown time of an insulating fluid.

DISPLAY 8.5 Scatterplot of breakdown times versus voltage



Relationship looks curved. Spread is much larger when response is larger

DISPLAY 8.6 Some hypothetical scatterplots of response versus explanatory variable with suggested courses of action; (a) is ideal



pH & Time study

ST412/512

Like insulating fluid.

Larger spread for larger means.

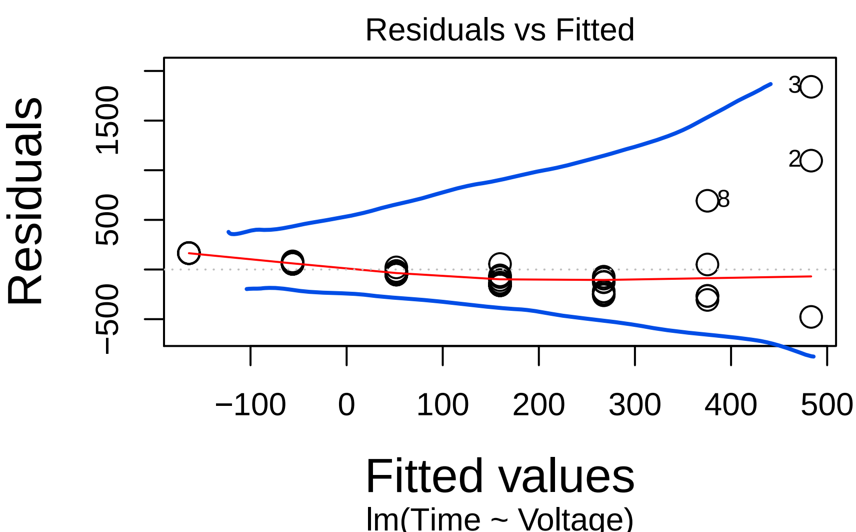
Mentioned in Ch. 11

skewness that is not too extreme is OK because SLR is robust.

response   expl. var

```
> head(case0802)
     Time Voltage  Group
1    5.79      26 Group1
2 1579.52      26 Group1
3 2323.70      26 Group1
4   68.85      28 Group2
5  108.29      28 Group2
6  110.29      28 Group2

> case0802_lm <- lm(Time~Voltage, data=case0802)
> plot(case0802_lm, which=1)
```

Same command as with aov object.



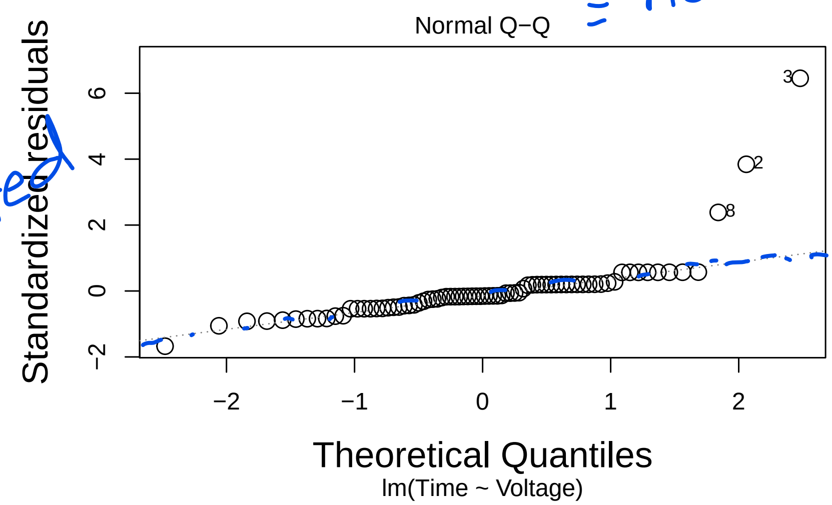
Larger spread for larger fitted values

"funnel shape" indicates we should try a log transformation of response.

Normal probability plot ← Great tool to assess normality

```
> plot(case0802_lm, which=2)
```

= normal prob. plot



Recall: Each residual is the difference between the observed data value and the estimated population mean for that observation. The estimated population means are called the "fitted values."

Residuals in one-way ANOVA: $Y_{ij} - \overline{Y_i}$ $\quad \overline{Y}_i$ estimates $\mu\{Y_{ij}\}$

Residuals in simple linear regression:

$$Y_i - \underbrace{(\hat{\beta}_0 + \hat{\beta}_1 X_i)}_{\text{estimates } \mu\{Y_i | X_i\}}$$

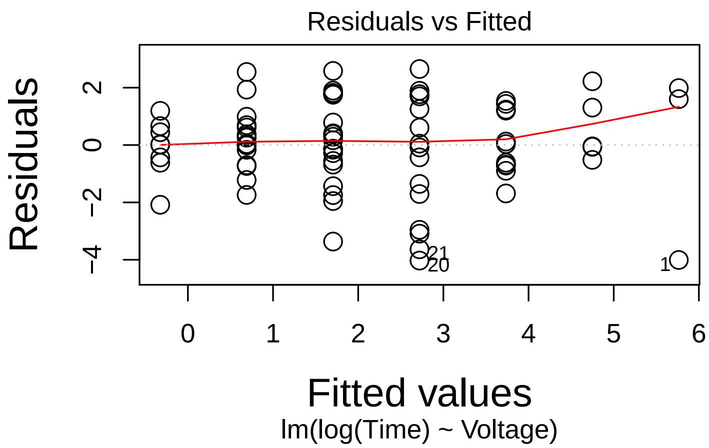
Standardized residuals: are divided by an estimate of their standard deviation.

std. res. have a std. dev. of about 1

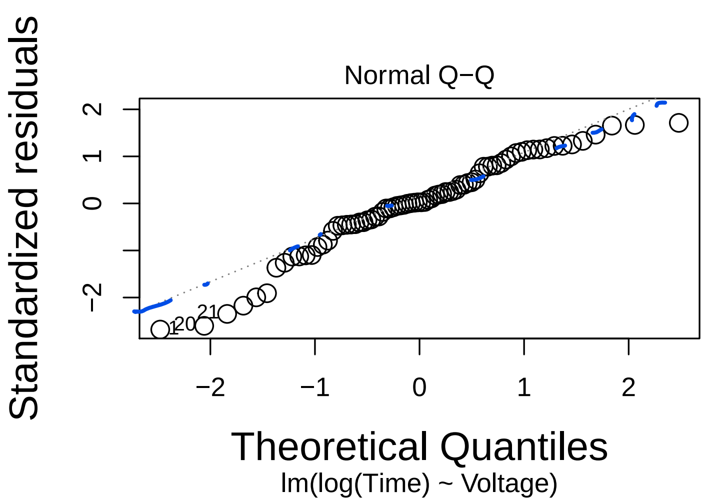
Theoretical quantiles: Expected ordered standardized residuals if data are actually normal.

If normal assumption is reasonable, then pts. in normal Q-Q plot fall approx. on a 45^0 line log response

```
> case0802_lm_log <- lm(log(Time)~Voltage, data=case0802)
> plot(case0802_lm_log, which=1)
> plot(case0802_lm_log, which=2)
```

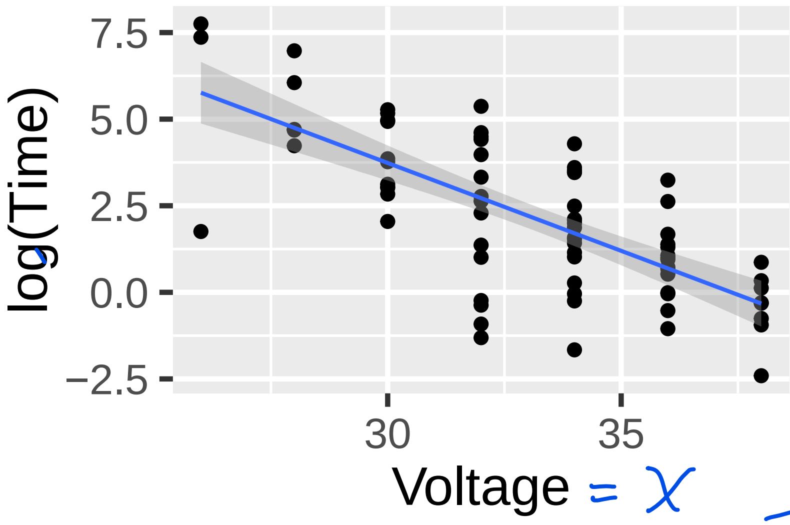


Looks OK



Looks better with log transformation

```
> ggplot(case0802, aes(x=Voltage, y=log(Time))) +
+   geom_point(size=3) +
+   geom_smooth(method=lm)
```



Slope parameter quantifies relationship between X and population mean of Y.

Estimate β_1 to estimate relationship between Y & X

Then deal with log transformation.

```
> summary(case0802_lm_log)

Call:
lm(formula = log(Time) ~ Voltage, data = case0802)

Residuals:
    Min      1Q  Median      3Q     Max
-4.0291 -0.6919  0.0366  1.2094  2.6513

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  18.9555     1.9100   9.924 3.05e-15
Voltage      -0.5074     0.0574  -8.840 3.34e-13
```

$\beta_1 \rightarrow$ Voltage

$\hat{\beta}_1$ $SE(\hat{\beta}_1)$

```
(Intercept) ***
Voltage     ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.56 on 74 degrees of freedom
Multiple R-squared:  0.5136,Adjusted R-squared:  0.507
F-statistic: 78.14 on 1 and 74 DF,  p-value: 3.34e-13
```

res. df to get t-quantile for CI.

$\hat{\beta}_1$ t-quantile $SE(\hat{\beta}_1)$

```
> # CI for beta_1 "by hand"
> -0.5074 - qt(0.975,74)*0.0574 # lower bound
[1] -0.621772
```

Conf. level is 95%

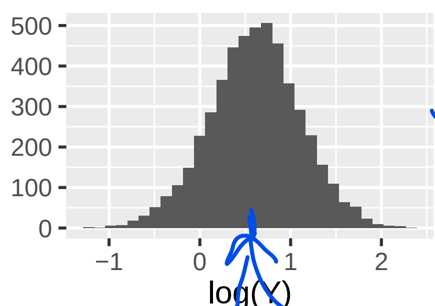
```
> -0.5074 + qt(0.975,74)*0.0574 # upper bound
[1] -0.393028
```

not OK. Need to back-transform

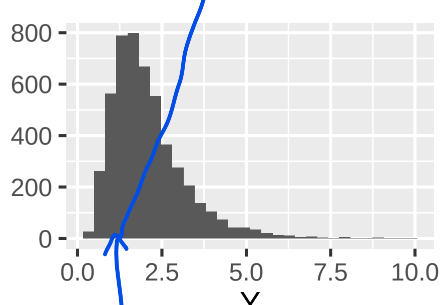
Not-quite-right statistical conclusion for confidence interval:

We estimate the decrease in pop'n mean log breakdown time for each 1 KV increase in voltage is 0.393 to 0.622 log minutes (95% CI, SLR).

Back-transforming from a log transformation in Chapter 3:



Symmetric distribution $\Rightarrow$
pop. mean($\log(Y)$) = pop. median($\log(Y)$)



median

Exponentiation preserves order $\Rightarrow$
$e^{\text{pop. median}(\log(Y))} = \text{pop. median}(Y)$

Back-transformation takes pop'n mean on log scale to pop'n median on original scale.

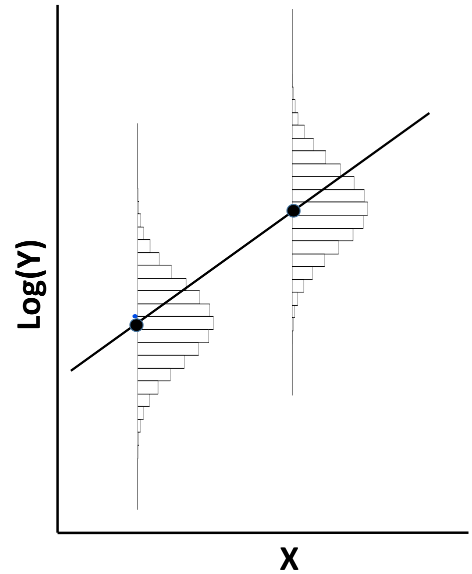
Interpreting slope β_1 when response Y was log-transformed

If the distribution of the log population is symmetric, then

pop. mean(log(Y)) = pop. median(log(Y))

$\mu\{\log(Y)|X\} =$

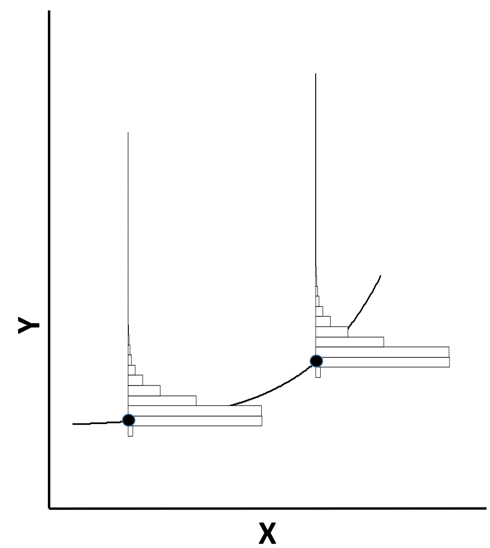
$\text{median}\{\log(Y)|X\}$



$e^{\text{pop. mean}(\log(Y))}$ = pop. median(Y)

$e^{\mu\{\log(Y)|X\}} =$

$\text{median}\{Y|X\}$



In the regression setting: $\mu\{\log(Y)|X\} = \beta_0 + \beta_1 X$

$$\underbrace{e^{\mu\{\log(Y)|X\}}}_{\text{median}\{Y|X\}} = e^{\beta_0 + \beta_1 X}$$

What happens to the median of Y when X is increased by one unit?

$\text{median}\{Y|X\} = e^{\beta_0 + \beta_1 X}$

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

$$= e^{\beta_0 + \beta_1 X + \beta_1}$$

$$= \underbrace{e^{\beta_0 + \beta_1 X}}_{\text{median}\{Y|X\}} \cdot \underbrace{e^{\beta_1}}$$

multiplicative change in median{Y|X} when X increased by one unit.

Statistical conclusion: We estimate popn median breakdown time decreases by a factor of $e^{0.393}$ to $e^{0.622}$ for each 1 kV increase in voltage (95% CI, SLR).

(Mult. change doesn't have units.)

Calculate these out in a report (but not a test).

ANOVA for Regression

Review: One-way analysis of variance (ANOVA) F-test

```
> case0501_aov <- aov(Lifetime~Diet, data=case0501)
> anova(case0501_aov)
Analysis of Variance Table

Response: Lifetime
           Df Sum Sq Mean Sq F value    Pr(>F)
Diet        5  12734  2546.8  57.104 < 2.2e-16 ***
Residuals 343  15297    44.6
Total     348  28,031
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

extra df + SS = diff. between reduced + full models

sep. means

equal means

full model: sep. means

reduced mod: equal means

Above F-test compares two models

Full model:

$$\mu\{Y_{ij}\} = \mu_i$$

Reduced model:

$$\mu\{Y_{ij}\} = \mu$$

Null hypothesis tested by $F = 57.104$ is the restriction on the parameters of the full model that yields the reduced model:

$$H_0: \mu_1 = \ldots = \mu_I = \mu$$

Extra SS (improvement in model fit)

extra df (extra complexity)

$$F\text{-statistic} = \frac{\{\text{residual SS(reduced)} - \text{residual SS(full)}\}/\{\text{df(reduced)} - \text{df(full)}\}}{\hat{\sigma}^2_{\text{full}}}$$

← MSE from full model

Is improvement in model fit worth extra complexity?

Residuals for any model: Observed value − Fitted value from model

Residuals for separate means model:

observed

$$Y_{ij} - \overline{Y}_i$$ ← est. pop'n mean

Residuals for equal means model:

$$Y_{ij} - \overline{Y}$$ ← est. pop'n mean if all Y_{ij}'s have same pop'n mean

df = residual degrees of freedom

$n -$ # mean parameters

Residual degrees of freedom for separate means model:

$n - I$ $\qquad$ $I =$ # groups $= 6$ for diet study

Residual degrees of freedom for equal means model:

$n - 1$

lm object

```
> case0802_lm_log <- lm(log(Time)~Voltage, data=case0802)
> anova(case0802_lm_log)
Analysis of Variance Table

Response: log(Time)
          Df Sum Sq Mean Sq F value    Pr(>F)
Voltage    1 190.15 190.151  78.141 3.34e-13 ***
Residuals 74 180.07   2.433
Total     75 370.22

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

apply anova() to lm object

extra SS & df

res df & SS for SLR model

MSE for full model

res. df & SS for equal means model

Above F-test compares two models

Full model: SLR

$$\mu\{Y_i \mid X_i\} = \beta_0 + \beta_1 X_i$$

Reduced model: equal means

$$\mu\{Y_i \mid X_i\} = \beta_0$$

here we will call the common mean β_0 instead of μ.

Null hypothesis tested by $F = 78.141$ is the restriction on the parameters of the full model that yields the reduced model:

$$H_0: \beta_1 = 0$$

This test null hypothesis of no relationship between explanatory var. and pop'n mean response.

extra SS $\qquad\qquad\qquad\qquad$ extra df

$$F\text{-statistic} = \frac{\{\text{residual SS(reduced)} - \text{residual SS(full)}\}/\{\text{df(reduced)} - \text{df(full)}\}}{\hat{\sigma}^2_{\text{full}}}$$

MSE from full model

Residuals for any model: Observed value − Fitted value from model

Residuals for simple linear regression (SLR) model:

obs.

$$Y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 X_i)$$

estimated pop'n mean of Y_i given X_i

Residuals for equal means model:

obs.

$$Y_i - \overline{Y}$$

estimated pop'n mean of Y_i under equal means model

$\overline{Y} \neq \hat{\beta}_0$, since $\hat{\beta}_0$ calculated with $\hat{\beta}_1$.

df = residual degrees of freedom

Residual degrees of freedom for SLR model:

$n-2$

β_0 & β_1 are the parameters

Residual degrees of freedom for equal means model:

$n-1$

extra df is diff. in # parameters

DISPLAY 8.8 Analysis of variances tables for the insulating fluid data from a simple linear regression analysis and from a separate-means (one-way ANOVA) analysis

We just did this

(a): Analysis of variance table from a simple linear regression analysis

Source	Sum of squares	d.f.	Mean square	*F*-statistic	*p*-value
Regression	190.1514	1	190.1514	78.14	<0.0001
Residual	180.0745	74	2.4334		
Total	370.2258	75			

Residual sum of squares, regression model

$\hat{\sigma}^2$ in regression model

compares regression and equal-means models

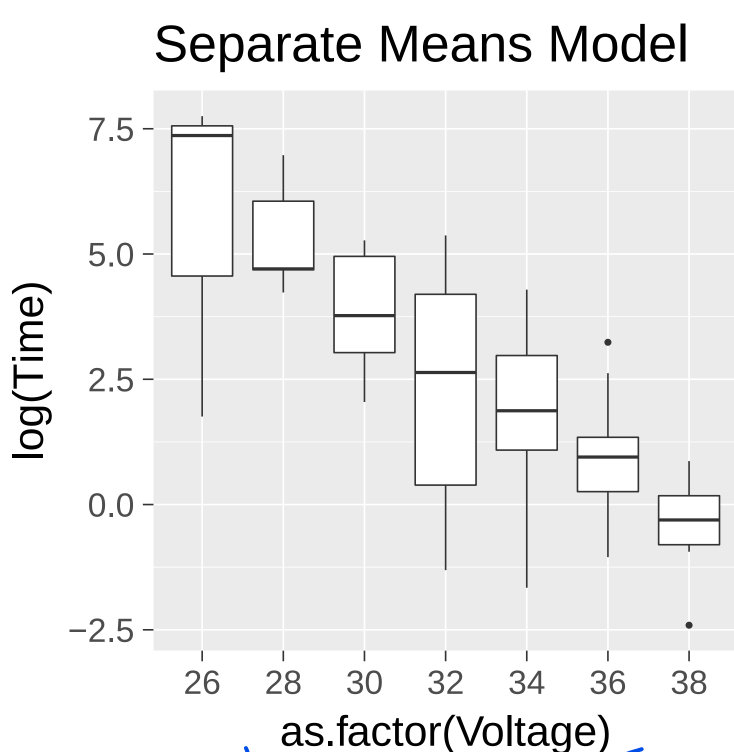
(b): Analysis of variance table from a one-way analysis of variance

Source	Sum of squares	d.f.	Mean square	*F*-statistic	*p*-value
Between groups	196.4774	6	32.7462	13.00	<0.0001
Within groups	173.7484	69	2.5181		
Total	370.2258	75			

Residual sum of squares, separate-means model

$\hat{\sigma}^2$ in separate-means model

compares separate-means and equal-means models



ANOVA (b) is one-way ANOVA with same data as (a).

use Voltage as a grouping variable

ANOVA (b)

grouping var.

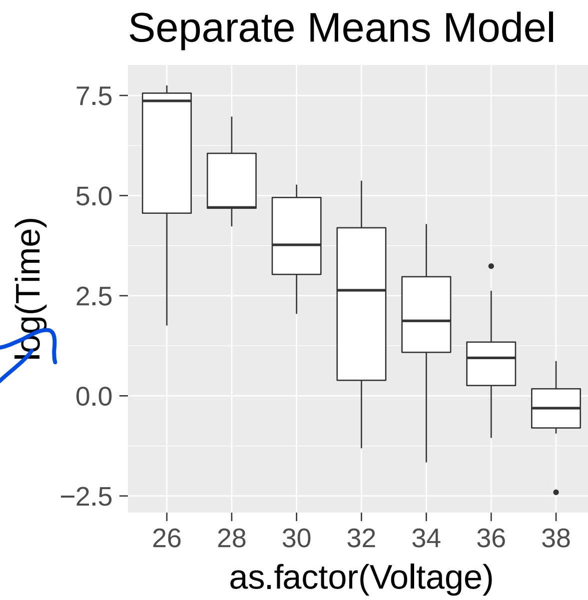
```
> case0802_aov_log <- aov(log(Time)~as.factor(Voltage), data=case0802)
> anova(case0802_aov_log)
Analysis of Variance Table

Response: log(Time)
                   Df Sum Sq Mean Sq F value   Pr(>F)
as.factor(Voltage)  6  196.5   32.75      13 8.87e-10 ***
Residuals          69  173.8    2.52
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

df + SS for sep. means

Total 75 370.3

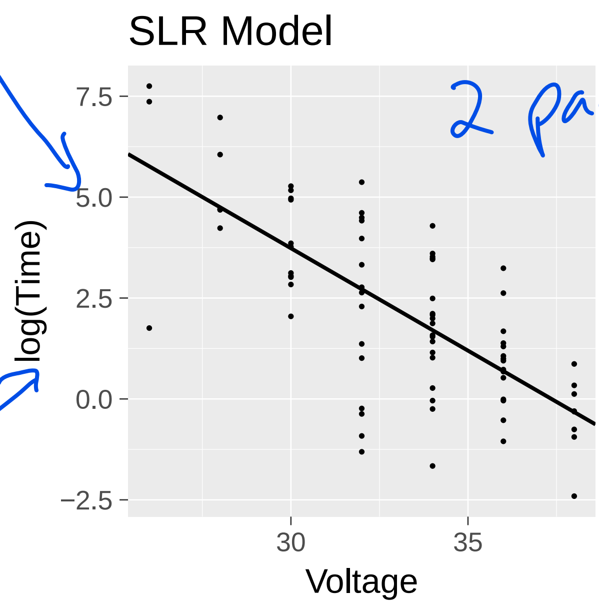
df & SS for equal means

7 parameters



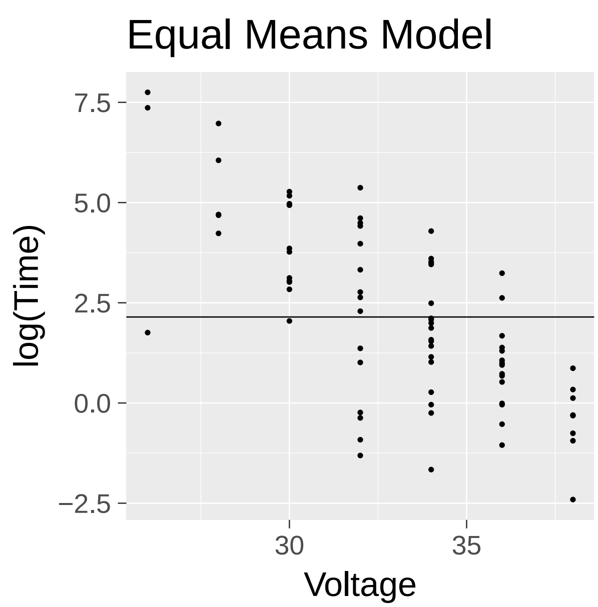
Lack of (linear) fit test

2 parameters



Compared in ANOVA (b)

Compared in ANOVA (a)



1 parameter

Lack of (Linear) Fit F-Test Not on final

Full model: Sep. means (more parameters)

Reduced model: SLR

Null hypothesis: H_0: group pop'n means fall on a line defined by a linear function of explanatory var.

$$F\text{-statistic} = \frac{\{\text{residual SS(SLR)} - \text{residual SS(sep. means)}\}/\{\text{df(SLR)} - \text{df(sep. means)}\}}{\hat{\sigma}^2_{\text{sep. means}}}$$

residual SS(SLR)= 180.0745 (res. SS ANOVA (a))

residual df(SLR)= 79

residual SS(sep. means)= 173.7484 (res. SS ANOVA (b))

residual df(sep. means)= 69

$\hat{\sigma}^2_{\text{sep. means}}$ = 2.5181

$$F = \frac{(180.0745 - 173.7484)/(79-69)}{2.5181}$$

$$\approx 0.502$$

DISPLAY 8.10 Composite analysis of variance table with *F*-test for lack of fit

Source of variation	Sum of squares	d.f.	Mean square	*F*-statistic	*p*-value
Between groups	*196.4774*	*6*	*32.7462*	*13.00*	<0.0001
Regression	190.1514	1	190.1514	75.51	<0.0001
Lack of fit	**6.3260**	**5**	**1.2652**	**0.50**	**0.78**
Within groups	*173.7484*	*69*	*2.5181*		
Total	370.2258	75			

By subtraction

LEGEND

Normal type items come from regression analysis (a).
Italicized items come from separate-means analysis (b).
Boldface items are new and calculated here.

0.502

extra SS & df for lack of linear fit test

Stat. conclusion for lack of fit test:

No evidence that pop mean log* breakdown time is not linearly related to voltage ($p \approx 0.78$, lack of fit test).

* Not going to back-transform because log was to satisfy linearity assumption.