STAT 500

Lack of Fit Test

Lack of Fit Test

- One method for model checking.
- Suppose we have multiple observations at one or more of the x_i values
- Notation: Y_{ij} is the jth observation at x_i
- Three models:

$$
\begin{aligned}
1)\ Y_{ij} &= \mu + \epsilon_i && \text{(common mean)} \\
2)\ Y_{ij} &= \beta_0 + \beta_1 X_i + \epsilon_i && \text{(regression)} \\
3)\ Y_{ij} &= \mu_i + \epsilon_i && \text{(separate means)}
\end{aligned}
$$

Lack of Fit Test

- SSE from regression model 2

$$\begin{aligned} SS_{error} &= \sum_i \sum_j (Y_{ij} - \widehat{Y}_i)^2 \\ &= \sum_i \sum_j (Y_{ij} - \bar{Y}_{i\cdot})^2 + \sum_i \sum_j (\bar{Y}_{i\cdot} - \widehat{Y}_i)^2 \\ &= SS_{pure\ error} + SS_{lack-of-fit} \end{aligned}$$

- $SS_{Pure\ Error}$ is the error sum of squares for model 3. It measures variability of observations about the mean response for each X. Does not assume the model fits.
- $SS_{Lack-of-Fit}$ measures lack of fit.
- Let $r =$ number of distinct x values

Lack of Fit Test

- New and improved ANOVA table

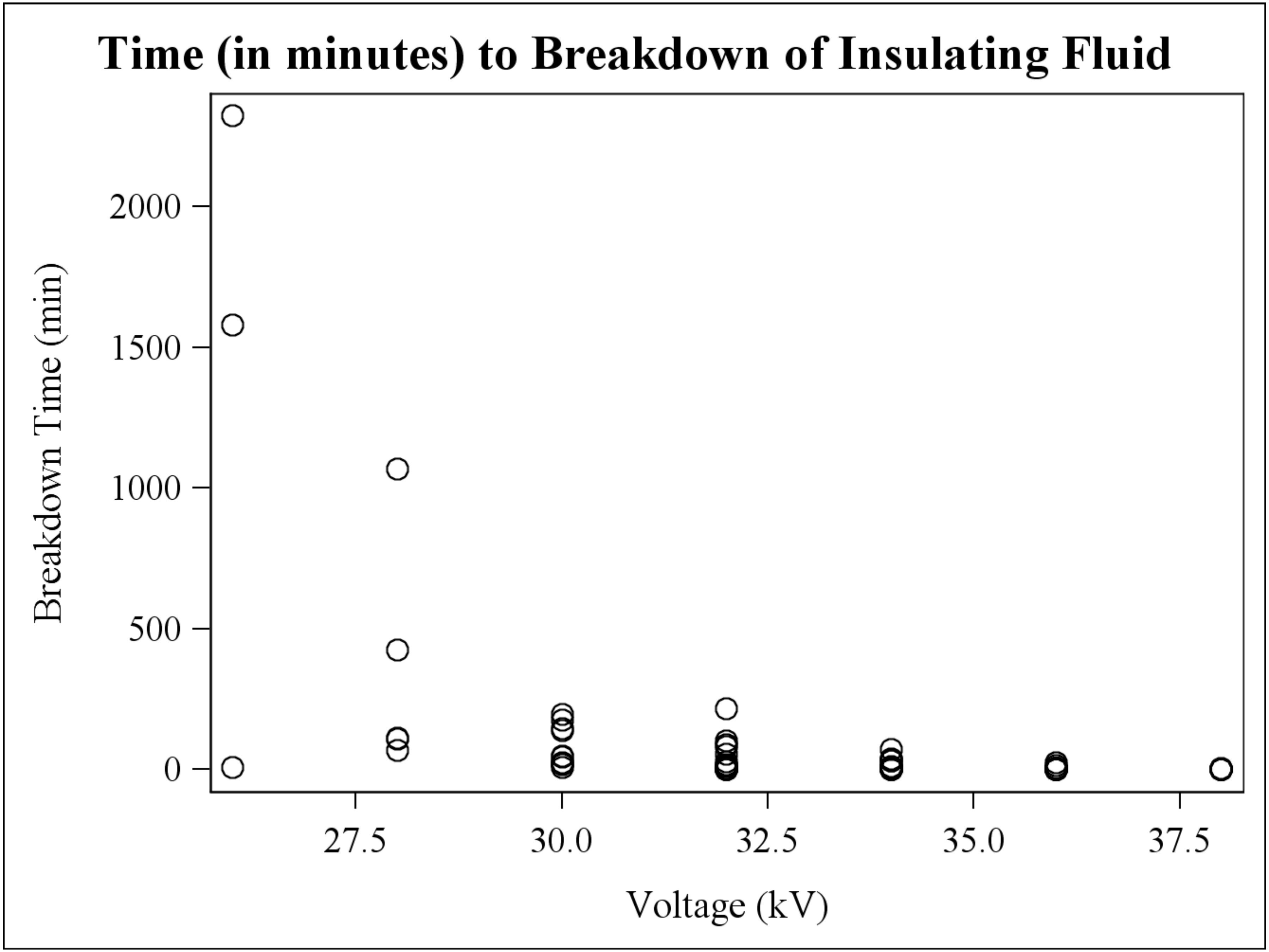
source of variation	degrees of freedom	sums of squares
Regression	1	$SS_{regression}$
Lack-of-Fit	$r-2$	$SS_{lack-of-fit}$
Pure Error	$n-r$	$SS_{pure\ error}$
Total	$n-1$	SS_{total}

- $E(MS_{Pure\ Error}) = \sigma^2$
- $E(MS_{Lack-of-Fit}) = \sigma^2 + \dfrac{\sum_{i=1}^{r} n_i[\mu_i - (\beta_0 + \beta_1 x_i)]^2}{r-2}$
- $E(MS_{Regression}) = \sigma^2 + \beta_1^2 \sum_{i=1}^{r} n_i[x_i - \bar{x}]^2$

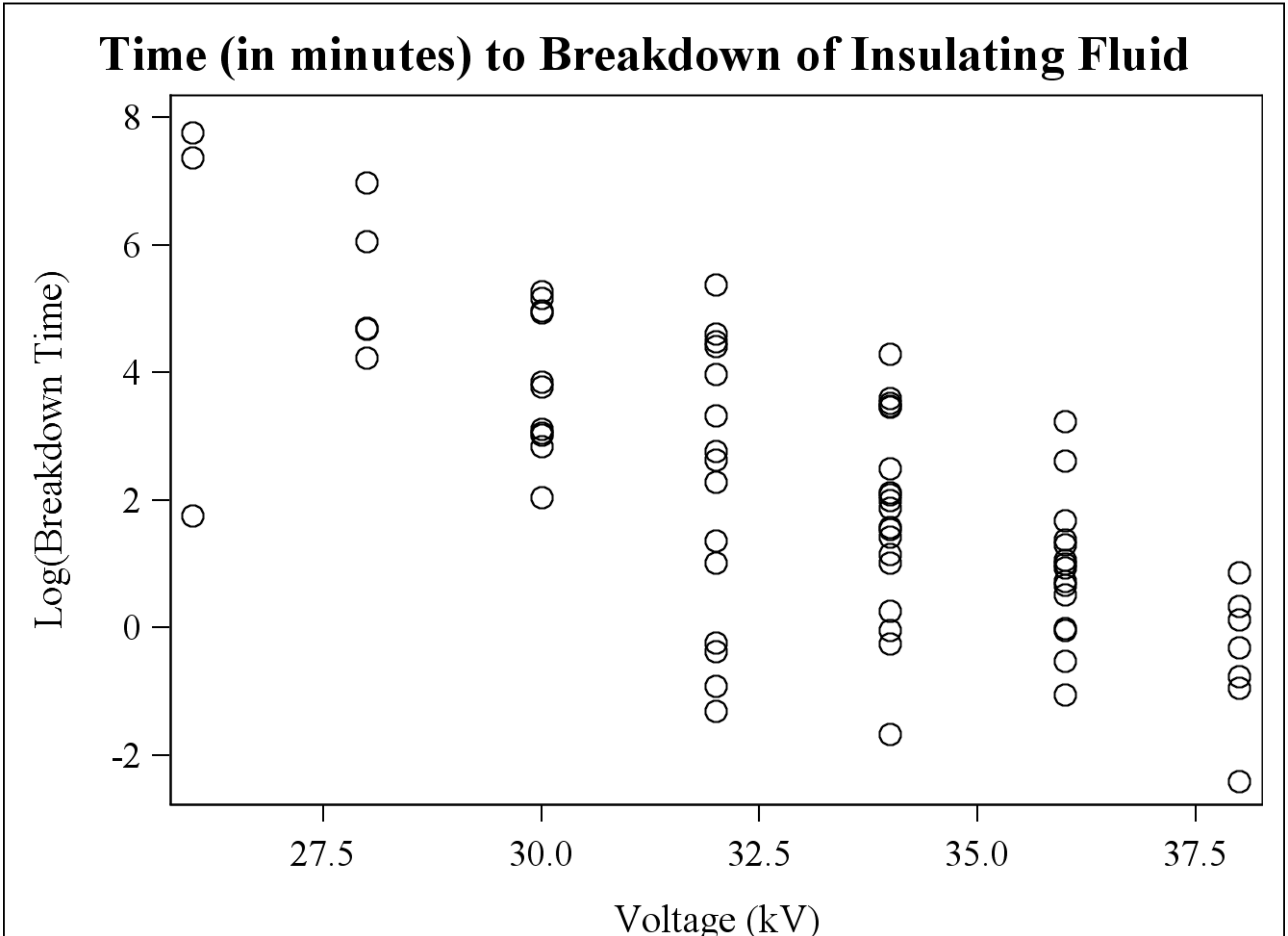
Breakdown Times of an Insulating Fluid

Chapter 8, *The Statistical Sleuth*

- Objective: Examine the relationship between voltage and breakdown time of insulating fluid
- Different batches of an insulating fluid were subjected to particular voltages until the insulating property of the fluid broke down
- Seven voltages were used, spaced 2 kV apart from 26 kV to 38 kV
- Measured time (in minutes) until the insulating property broke down
- More than one batch tested at each voltage level



Time (in minutes) to Breakdown of Insulating Fluid
Breakdown Time (min)
2000
1500
1000
500
0
27.5
30.0
32.5
35.0
37.5
Voltage (kV)


Time (in minutes) to Breakdown of Insulating Fluid
8
6
4
2
0
-2
Log(Breakdown Time)
27.5
30.0
32.5
35.0
37.5
Voltage (kV)

Summary Statistics Log(Breakdown times)

Level of voltage	N	Logy	
		Mean	Std Dev
26	3	5.62397487	3.35520660
28	5	5.32952567	1.14455914
30	11	3.82199830	1.11120182
32	15	2.22852317	2.19809924
34	19	1.78639275	1.52521123
36	15	0.90224550	1.10990142
38	7	-0.44192816	1.06980069

Example: One-way ANOVA

First consider a one-way ANOVA for the model with a different mean breakdown time at each voltage

$$Y_{ij} = \mu_i + \epsilon_{ij}$$

where $\epsilon_{ij} \sim N(0, \sigma^2)$

source of variation	df	sums of squares	mean square
Voltage Levels	6	$\sum_{i=1}^{9} n_i(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2 = 190.43$	31.738
Pure Error	68	$\sum_{i=1}^{9} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i\cdot})^2 = 173.73$	2.555
Total	74	$\sum_{i=1}^{9} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{\cdot\cdot})^2 = 364.16$	

Note that $E(MS_{Pure\ Error}) = \sigma^2$

Example: Simple Linear Regression

Now consider the more restrictive regression model

$$Y_{ij} = \beta_0 + \beta_1 x_i + \eta_{ij}$$

where $\eta_{ij} \sim N(0, \sigma^2_\eta)$

- Least squares estimates

$$b_1 = \frac{\sum_{i=1}^{9} n_i(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})(x_i - \bar{x}_\cdot)}{\sum_{i=1}^{9} n_i(x_i - \bar{x}_\cdot)^2} = -0.5075$$

$$b_0 = \bar{y}_{\cdot\cdot} - b_1\bar{x}_\cdot \;=\; 2.17828 - (-0.5075)(33.06667) = 18.9605$$

- The least squares estimate of the line is

$$\hat{Y}_i = 18.9605 - 0.5075 x_i$$

Example: Lack-of-Fit Test

Incorporate $SS_{regression} = \sum_{i=1}^{9} n_i(\widehat{y}_i - \bar{y}_{\cdot\cdot})^2 = 184.0856$ into the ANOVA table

source of variation	df	sums of squares	mean square
Regression on voltage	1	$SS_{regression} = 184.0856$	184.0856
Lack-of-Fit	5	$\sum_{i=1}^{9} n_i(\widehat{y}_i - \bar{y}_{i\cdot})^2 = 6.3427$	1.26854
Pure Error	68	$\sum_{i=1}^{9} \sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i\cdot})^2 = 173.7316$	2.5549
Total	74	$\sum_{i=1}^{9} \sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{\cdot\cdot})^2 = 364.1599$	

Example: Lack-of-Fit Test

- F-test for lack of fit

$$H_o : E(Y_{ij}|x_i) = \beta_0 + \beta_1 x_i$$

$$H_a : E(Y_{ij}|x_i) = \mu_i = \beta_0 + \beta_1 x_i + g(x_i)$$

- $E(MS_{Pure\ \ Error}) = \sigma^2$

- $E(MS_{Lack-of-Fit}) = \sigma^2 + \dfrac{\sum_{i=1}^{I} n_i [g(x_i)]^2}{I-2}$

- Reject H_o if $F = \frac{MS_{Lack-of-Fit}}{MS_{Pure\ \ Error}} > F_{(df_{LoF}, df_{PE}), 1-\alpha}$

- For the insulating fluid breakdown data,

  $F = \frac{1.26854}{2.5549} = 0.50$ on (5,68) df with p-value $= 0.778$

Example: Conclusion and Remarks

- Conclusion: Using $Y = Log(time)$ as the response, the data are consistent with a straight line model

$$Y_{ij} = \beta_0 + \beta_1 x_i + \epsilon_i$$

- This does not prove that

$$Y_{ij} = \beta_0 + \beta_1 x_i + \epsilon_i$$

with $\epsilon_{ij} \sim N(0, \sigma^2)$ is exactly correct, but it suggests that a straight line model is a reasonable approximation for $E(Y|X = x)$, the conditional mean of the natural logarithm of the breakdown time when the voltage is set at x.

- If the lack-of-fit test is significant, search for a better model.