STAT 5000

Statistical Methods I

Week 11

Fall 2024

Dr. Danica Ommen

Unit 3

SLR: Forbes Example

SLR: Forbes Example

Forbes Data

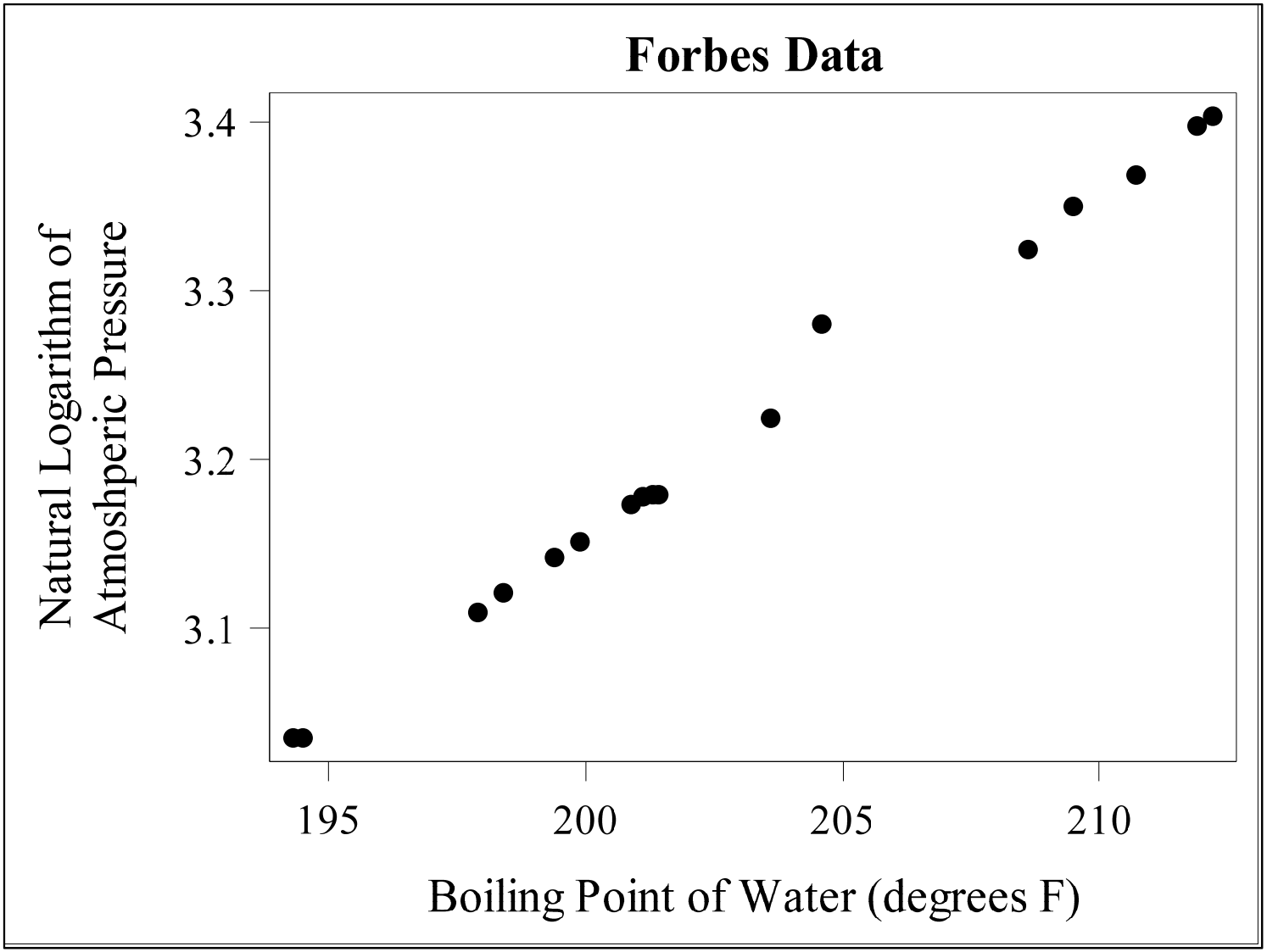
Weisberg, Sanford, *Applied Linear Regression*, Wiley, 1980.

- James D. Forbes collected data in the mountains of Scotland
- n=17 locations (at different altitudes)
- Objective: Predict barometric pressure (in inches of mercury) from boiling point of water (X) in ^{o}F.
- Use Y=log(barometric pressure)
- Motivation: Fragile barometers were difficult to transport

Forbes Data

Obs	Boil. Point of Water (°F)	Bara-metric Pressure (in Hg)	Nat.Log Bara-metric Pressure	Obs	Boil. Point of Water (°F)	Bara-metric Pressure (in Hg)	Nat.Log of Bara-metric Pressure
1	194.3	20.79	3.03447	10	201.4	24.02	3.17889
2	194.5	20.79	3.03447	11	203.6	25.14	3.22446
3	197.9	22.40	3.10906	12	204.6	26.57	3.27978
4	198.4	22.67	3.12104	13	208.6	27.76	3.32360
5	199.4	23.15	3.14199	14	209.5	28.49	3.34955
6	199.9	23.35	3.15060	15	210.7	29.04	3.36867
7	200.9	23.89	3.17346	16	211.9	29.88	3.39719
8	201.1	23.99	3.17764	17	212.2	30.06	3.40320
9	201.3	24.01	3.17847				

SLR: Forbes Example



Analysis of the Forbes Data

- Proposed regression model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

  where $\epsilon_i \sim N(0,\ \sigma^2)$, $i = 1, 2, ..., 17$
- Y_i=log(pressure)
- X_i=boiling point (^{o}F)
- β_1 is the increase in mean log(pressure) when boiling point of water increases by 1 ^{o}F
- β_0 is the mean log(pressure) when boiling point of water is 0 ^{o}F (Is this extrapolation realistic?)

Analysis of the Forbes Data

- Estimated regression model

$$\hat{Y} = b_0 + b_1 x = -0.97097 + 0.020623x$$

- Could have subtracted 212 ^{o}F from each boiling point. Then the estimated model is

$$\begin{aligned} \hat{Y} &= b_0 + 212 b_1 + b_1(x - 212) \\ &= 3.401106 + 0.020623(x - 212) \end{aligned}$$

- Then 3.401106 is an estimate of the mean log(pressure) at 212 ^{o}F.

Predicted Values

$$\hat{Y}_i = -0.97097 + 0.020623x$$

- Values on the estimated regression line.
- Predict values of Y_i for a given value of x_i
  - $x_i = 201.1\ ^o\text{F}$:

$$\hat{Y}_i = -0.97097 + 0.020623(201.1) = 3.176315$$

  - $x_i = 210.7\ ^o\text{F}$:

$$\hat{Y}_i = -0.97097 + 0.020623(210.7) = 3.374296$$

Residuals

$$e_i = Y_i - \hat{Y}_i$$

- Vertical distance between observed value of Y and predicted value of Y.
- Residuals:
  - $x_i = 201.1\ ^oF$ and $Y_i = 3.17764$:

$$e_i = 3.17764 - 3.176315 = 0.001325$$

  - $x_i = 210.7\ ^oF$ and $Y_i = 3.36867$:

$$e_i = 3.36867 - 3.374296 = -0.005626$$

ANOVA Table

Source	df	SS	MS	F	p-value
Model	1	0.22573	0.22573	2961.55	< 0.0001
Error	15	0.00114	0.00007622		
Total	16	0.22688			

ANOVA F-test

- $H_0 : \beta_1 = 0$
- $H_a : \beta_1 \neq 0$
- $F = 2961.55$ with p-value < 0.0001
- Reject $H_0 \implies$ There is a significant linear relationship between boiling point of water and log of barometric pressure.

Coefficient of Determination

$$R^2 = \frac{SS_{\text{model}}}{SS_{\text{Total}}} = \frac{0.22573}{0.22688} = 0.9950$$

99.50% of the variation in log(barometric pressure) can be explained by the linear regression model with boiling point of water.

Inference for Slope

- Test $H_o : \beta_1 = 0$ $(Y_i = \beta_0 + \epsilon_i)$
  versus $H_a : \beta_1 \neq 0$ $(Y_i = \beta_0 + \beta_1 x_i + \epsilon_i)$
- Evaluate
$$t = \frac{b_1 - 0}{S_{b_1}} = \frac{.020623 - 0}{0.000379} = 54.42$$
- The least squares estimate of the slope is 54 standard errors away from zero (p-value $<<$ .0001).
  - It is extremely unlikely that an estimate that far from zero could occur simply because of random errors when β_1 is actually zero.
  - Consequently, reject the null hypothesis and conclude that the slope is positive.

Inference for Slope

- A 95% confidence interval for the slope indicates that the slope is "very well" estimated from these data

$$
\begin{aligned}
& b_1 \pm t_{15,.975} S_{b_1} \\
\Rightarrow \quad & 0.020623 \pm (2.131)(0.00037895) \\
\Rightarrow \quad & (0.0198, 0.0214)
\end{aligned}
$$

Inference for Intercept

- Test $H_o : \beta_0 = 0$ $(Y_i = \beta_1 x_i + \epsilon_i)$

  versus $H_a : \beta_0 \neq 0$ $(Y_i = \beta_0 + \beta_1 x_i + \epsilon_i)$
- Evaluate $t = \dfrac{b_0 - 0}{S_{b_0}} = \dfrac{-0.971 - 0}{0.0769} = -12.6$
- The least squares estimate of the intercept is 12.6 standard errors away from zero (p-value $<<$ .0001).
  Reject the null hypothesis and conclude that the intercept is negative. (No practical motivation)
- A 95% confidence interval for the intercept is

  $$b_0 \pm t_{15,.975} S_{b_0} \Rightarrow -0.971 \pm (2.131)(0.0769) \Rightarrow (-1.135, -0.807)$$

Inference for Conditional Mean

- Construct a 95% confidence interval for the mean of possible log-pressure measurements when the boiling point of water is x=209 ^{o}F
- Estimated mean is
$$\hat{\mu}_{Y|x} = b_0 + b_1 x = -0.9710 + (.0206)(209) = 3.339$$
- Evaluate the standard error of this estimate
$$S_{\hat{\mu}_{Y|x}} = \sqrt{.0000762\left(\frac{1}{17} + \frac{(209 - 202.953)^2}{530.78}\right)} = 0.00312$$
- A 95% confidence interval is
$\hat{\mu}_{Y|x} \pm t_{15,.975} S_{\hat{\mu}_{Y|x}} \Rightarrow 3.339 \pm (2.131)(0.00312) \Rightarrow (3.333, 3.346)$

Inference for Conditional Mean

- Apply the exponential function to the end points to get an *approximate* confidence interval for the mean pressure

$$(28.02, 28.39) \text{ inches of Hg}$$

- This could be computed with either the REG procedure or the GLM procedure in SAS by adding an additional line to the data file with X=209 and a missing value for Y

Simultaneous Confidence Region

Scheffe procedure for constructing a 95% confidence region for a segment of the true regression line

Evaluate $(b_0 + b_1 x) \pm \sqrt{2F_{(2,n-2),1-\alpha}} S_{b_0+b_1 x}$

$$\Rightarrow (b_0 + b_1 x) \pm \sqrt{2F_{(2,15),0.95}} S_{b_0+b_1 x}$$

$$\Rightarrow (b_0 + b_1 x) \pm (2.713)\sqrt{.0000762\left(\frac{1}{17} + \frac{(x-202.953)^2}{530.78}\right)}$$

Prediction Interval

- Construct a 95% prediction interval for a log-pressure value when the boiling point of water is x=209 ^{o}F
- Prediction is the estimated mean

$$\hat{Y} = b_0 + b_1 x + error = -0.9710 + (.0206)(209) + 0 = 3.339$$

- Evaluate the standard error of the prediction (include the variation of the associated random error, estimated as $MS_{error} = .0000762$)

$$S_{pred} = \sqrt{.0000762\left(1 + \frac{1}{17} + \frac{(209 - 202.953)^2}{530.78}\right)} = 0.00927$$

Prediction Interval

- A 95% prediction interval is

$$\hat{y} \pm t_{15,.975} S_{pred} \quad \Rightarrow \quad 3.339 \pm (2.131)(0.00927)$$

$$\Rightarrow \quad (3.319, 3.359)$$

- Apply the exponential function to the end points to get an *approximate* prediction interval for barometric pressure:
  (27.63, 28.76) inches of Hg
- This could be computed with either the REG or GLM procedure in SAS by adding an additional line to the data file with X=209 and a missing value for Y

Unit 3

Simple Linear Regression (SLR)

Model Diagnostics

SLR Model and Assumptions

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad \text{where} \quad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

- Values of Y_i are independent (independent random errors)
- Values of x_i are fixed
- $\mu_{Y|x_i}$ is a linear function of x_i
- Homogeneous error variance: $Var(\epsilon_i) = \sigma^2$
- Normally distributed errors: $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$

Independence

- Check independence of observations through details of data collection
- Beware of
  - Observations over time
  - Clustering of observations
  - Spatial elements to observations
- Crucial assumption - must use other methods if violated

Fixed Values of x

- Assume x is measured without error
- Check through variable definition and through details of data collection
- If violated, model the error in x using a random effect

Linearity

- Scatterplot: Plot of Y_i versus $x_i \rightarrow$ linear pattern
- Residual Plot: Plot of residuals e_i versus $x_i \rightarrow$ no pattern
- Violations of linearity
  - Transform Y_i values so that relationship with x_i is linear
  - Common transformations: log and power (Y^2, Y^3, $\sqrt{Y}$, etc.)
  - Conduct analysis with transformed Y values
  - Undo transformation in drawing conclusions

Residual Normality and Homogeneous Variance

- Residuals are approximations for random errors:

$$e_i = Y_i - \hat{Y}_i = Y_i - (b_o + b_1 x_i) \quad \text{for } i = 1, 2, \ldots, n$$

- Important properties of residuals
  - $\sum_i e_i = 0$
  - $\sum_i x_i e_i = \sum_i \hat{Y}_i e_i = 0$
  - Residuals are negatively correlated

$$e_i \sim N\left(0, \sigma^2\left(1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}\right)\right)$$

Regression Analysis - Residuals

- Residuals do not have homogeneous variances

$$e_i \sim N\left(0, \sigma^2\left(1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}\right)\right)$$

- Sometimes use

$$r_i = e_i / \sqrt{MSE\left(1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}\right)}$$

(known as studentized residuals)

Residual Plots

- Plot residuals versus predicted values
  - Detect non-constant variance
    - Look for changes in variability around the horizontal line at 0
    - Megaphone shaped pattern: variability of e_i increases or decreases as x_i increases
  - Detect non-linearity
  - Detect outliers

Residual Plots

- Plot residuals versus X
  - In simple linear regression, this is the same as previous
  - In multiple regression, it will be useful
- Plot residuals versus other possible predictors (e.g., time)
  - Detect important lurking variable
- Plot residuals vs lagged residuals
  - Detect correlated errors
- Normal probability plot of residuals
  - Detect non-normality

Remedies for Model Violations

- Transformation of Y
- Adding/modifying predictors
- More sophisticated models and/or estimation procedures
  - Weighted least squares for nonhomogeneous variance
  - Time series models for correlated errors
  - Robust regression methods for nonnormality
- These will be described more fully under multiple regression

Case Diagnostics

- Leverage
- Outliers
- Influential Points

Case Diagnostics - Leverage

- Extreme values of x are called high leverage cases because they exert a large "pull" on SLR
- Measure of "potential" influence of observation on SLR
- Leverage of the ith observation is:

$$h_i = \left(\frac{1}{n-1}\right)\left(\frac{x_i - \bar{x}}{s_x}\right)^2 + \frac{1}{n}$$

- Properties of h_i
  - $1/n \leq h_i \leq 1$
  - $\sum_{i=1}^{n} h_i = 2 \qquad \Longrightarrow \qquad \bar{h} = 2/n$

Case Diagnostics - Leverage

- Often use $4/n$ or $6/n$ as a guide for determining large h_i
- In addition to an absolute cutoff, look for large h_i by examining the distribution of h_i values across observations

Case Diagnostics - Outliers

- Extreme Y_i value for a given x_i
- Three assessment methods
  - Residuals
  - Internally studentized residuals
  - Externally studentized residuals

Case Diagnostics - Residuals

- Residuals

$$e_i = Y_i - \hat{Y}_i$$

- $\text{Var}(e_i) = \sigma^2(1 - h_i)$
- Observations with higher leverage will have residuals with smaller variability.

Case Diagnostics - Residuals

- Internally studentized residuals

$$r_i = \frac{e_i}{\sqrt{MSE(1 - h_i)}}$$

- r_i will have mean zero and approximately equal variance
- Outliers will inflate MSE
- r_i is called STUDENT in SAS

Case Diagnostics - Residuals

- Externally studentized residuals

$$t_i = \frac{e_i}{\sqrt{MSE_{(-i)}(1 - h_i)}}$$

  where $\text{MSE}_{(-i)}$ is MSE without the ith observation
- t_i will have mean zero and approximately equal variance
- t_i is called RSTUDENT in SAS

Case Diagnostics - Outliers

Studentized residual values with absolute value

- Less than 2 are fine
- Between 2 and 3 indicate potential outliers
- Greater than 3 indicate outliers

Case Diagnostics - Outliers

- Outliers inflate value of $\hat{\sigma}^2$
- Will lower values of test statistics t and F
- Will inflate widths of confidence intervals for parameters and prediction intervals

Case Diagnostics - Influence

- Concerned about unusual cases that have a big influence on both:
  - $\hat{Y}_i$ for some x_i
  - estimated slope $\hat{\beta}_1$
- Could delete the case, refit model and examine the change

Case Diagnostics - Influence

- COOK'S D - effect of deleting the i-th case on the least squares regression model

$$D_i = \left(\frac{r_i^2}{2}\right)\left(\frac{h_i}{1-h_i}\right)$$

- D_i is large when r_i is large and h_i is large
- $D_i > 2 * \sqrt{2/n}$ indicates substantial influence

Unit 3

Simple Linear Regression (SLR)

Lack of Fit

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

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

- $SS_{Pure\ Error}$ 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

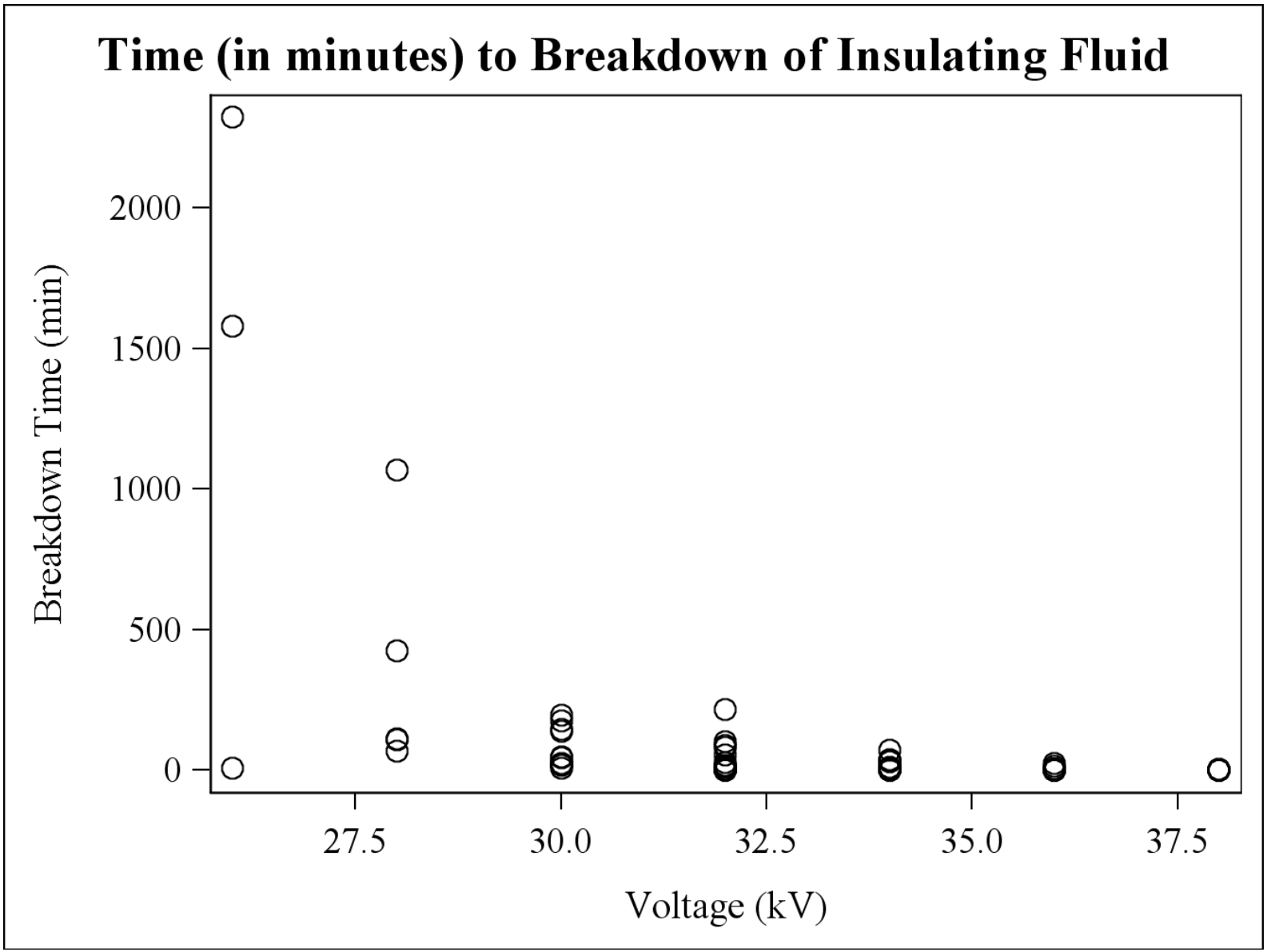
source of variation	degrees of freedom	sums of squares
Regression	1	$SS_{regression}$
Lack-of-Fit	$r-2$	$SS_{lack-of-fit}$
Pure Error	$n-r$	$SS_{pure\ error}$
Total	$n-1$	SS_{total}

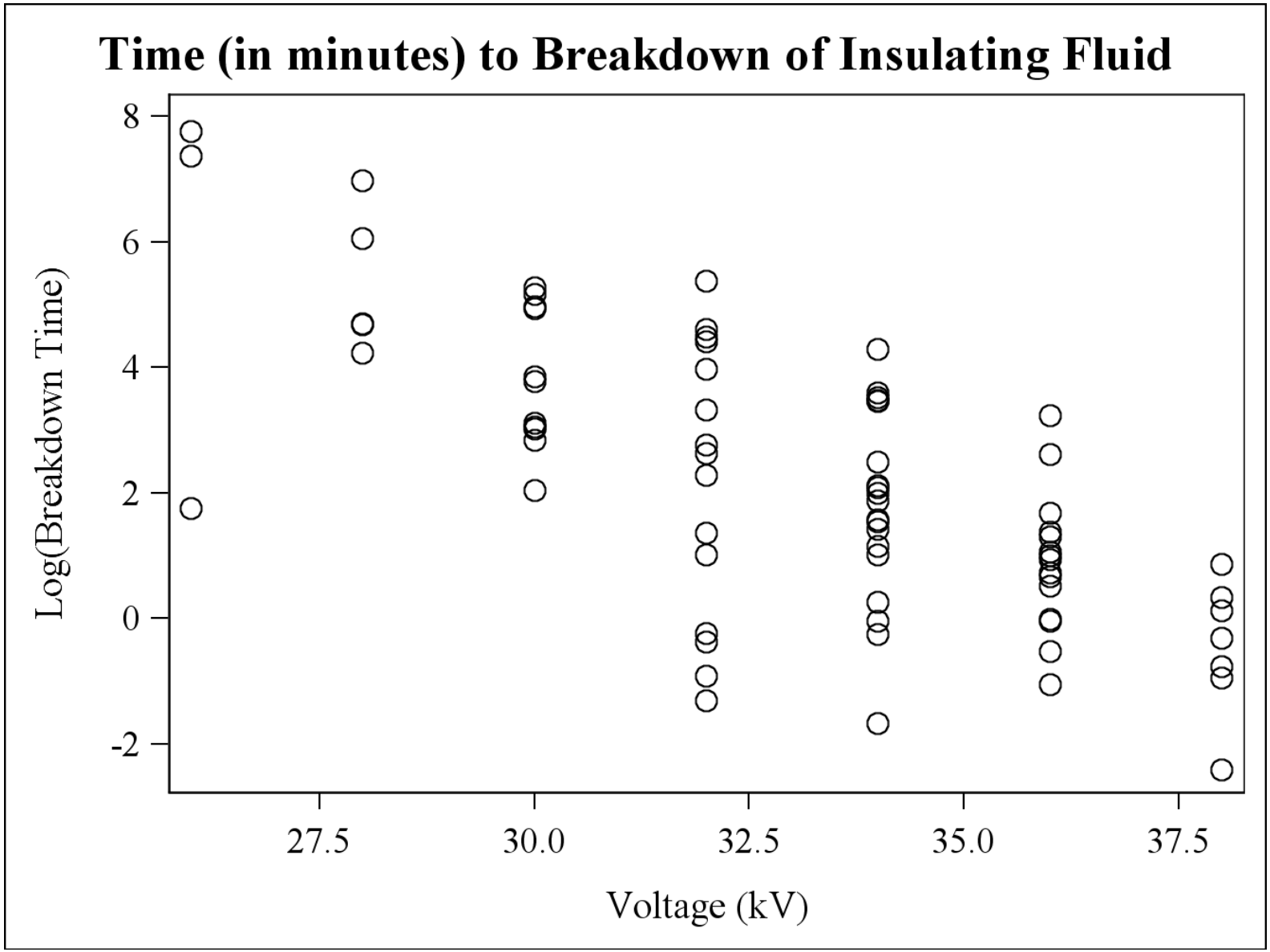
- $E(MS_{Pure\ Error}) = \sigma^2$
- $E(MS_{Lack-of-Fit}) = \sigma^2 + \dfrac{\sum_{i=1}^{r} n_i[\mu_i - (\beta_0 + \beta_1 x_i)]^2}{r-2}$
- $E(MS_{Regression}) = \sigma^2 + \beta_1^2 \sum_{i=1}^{r} n_i[x_i - \bar{x}]^2$

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
Log(Breakdown Time)
8
6
4
2
0
-2
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

$$Y_{ij} = \mu_i + \epsilon_{ij} \quad \text{where} \quad \epsilon_{ij} \sim N(0, \sigma^2)$$

source of variation	df	sums of squares	mean square
Voltage Levels	6	$\sum_{i=1}^{7} n_i(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2 = 190.43$	31.738
Pure Error	68	$\sum_{i=1}^{7}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i\cdot})^2 = 173.73$	2.555
Total	74	$\sum_{i=1}^{7}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{\cdot\cdot})^2 = 364.16$	

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

Example: Simple Linear Regression

Now consider the more restrictive regression model

$$Y_{ij} = \beta_0 + \beta_1 x_i + \eta_{ij} \quad \text{where} \quad \eta_{ij} \sim N(0, \sigma_\eta^2)$$

- Least squares estimates

$$\begin{aligned} b_1 &= \frac{\sum_{i=1}^{7} n_i(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})(x_i - \bar{x}_\cdot)}{\sum_{i=1}^{7} n_i(x_i - \bar{x}_\cdot)^2} = -0.5075 \\ b_0 = \bar{y}_{\cdot\cdot} - b_1\bar{x}_\cdot &= 2.17828 - (-0.5075)(33.06667) = 18.9605 \end{aligned}$$

- The least squares estimate of the line is

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

Example: Lack-of-Fit Test

Incorporate $SS_{regression} = \sum_{i=1}^{7} n_i(\hat{y}_i - \bar{y}_{..})^2 = 184.0856$ 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	$\sum_{i=1}^{7} n_i(\hat{y}_i - \bar{y}_{i.})^2 = 6.3427$	1.26854
Pure Error	68	$\sum_{i=1}^{7}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})^2 = 173.7316$	2.5549
Total	74	$\sum_{i=1}^{7}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{..})^2 = 364.1599$	

Example: Lack-of-Fit Test

- F-test for lack of fit

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

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

- $E(MS_{Pure\ Error}) = \sigma^2$
- $E(MS_{Lack-of-Fit}) = \sigma^2 + \frac{\sum_{i=1}^{l} n_i [g(x_i)]^2}{l-2}$
- 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.

Unit 3

Simple Linear Regression (SLR)

Correlation

Population Correlation Coefficient

- Measure of linear relationship between two quantitative variables (X and Y) in population
- Denoted as ρ
- Defined as

$$\rho = \frac{Cov(X, Y)}{\sigma_X \sigma_Y} = \frac{E[(X - E(X))(Y - E(Y))]}{\sigma_X \sigma_Y}$$

Properties of ρ

- $-1 \leq \rho \leq 1$
  - Perfect linear relationship $\rho = -1$ or $\rho = 1$
  - No linear relationship: $\rho = 0$
- Sign of ρ indicates direction of relationship
  - Negative linear relationship between X and Y: $\rho < 0$
  - Positive linear relationship between X and Y: $\rho > 0$
- Strength of relationship indicated by $|\rho|$
- ρ is invariant to the choice of scale for X and/or Y.
  - X = height, Y = weight
  - Same ρ whether X is measured in in. or cm.
  - Same ρ whether Y is measured in lbs. or kg.

Sample Correlation Coefficient

- Estimate ρ by taking a sample from population and calculating

$$r = \frac{1}{n-1}\left(\frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{S_X S_Y}\right)$$

- r has the same properties as ρ

Hypothesis Test for ρ

To determine whether two variables X and Y have a linear relationship, we can also use r to conduct a hypothesis test for ρ:
$H_0 : \rho = 0$ vs. $H_a : \rho \neq 0$

- Note that b_1 is a function of r

$$b_1 = r\left(\frac{S_Y}{S_X}\right)$$

- b_1 and r have same sign
- Inference for β_1 and ρ produce same test statistic, distribution, p-value, decision, and conclusion
- We can do a t-test for this null hypothesis

Differences between Correlation and Slope

- Correlation
  - Focus is relationship between X and Y
  - Use when there is not a clear response variable
- Slope
  - Focus is explaining change in values of Y with x
  - Use when there is a clear response variable

r and R^2

- r is a function of R^2

$$r = \pm\sqrt{R^2} \qquad r^2 = R^2$$

- r is a numerical summary of the direction and strength of the linear relationship between X and Y
- R^2 is a numerical summary of the percentage of variability in Y that can be explained by the linear regression with x

Questions?

Contact me:

Email: dmommen@iastate.edu

Student Office Hours: Thursdays @ 10-11 am