Inference for Simple Linear Regression (Ch. 9.1)

Yifan Zhu

Iowa State University

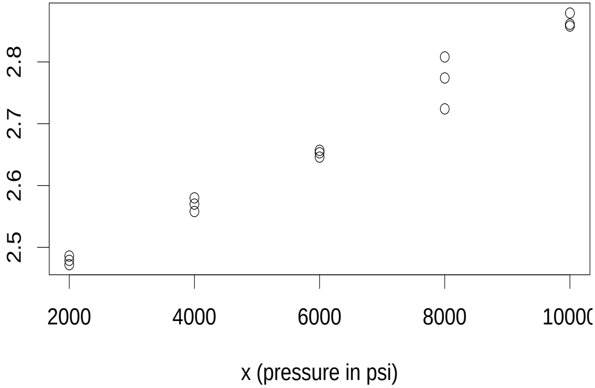
Outline

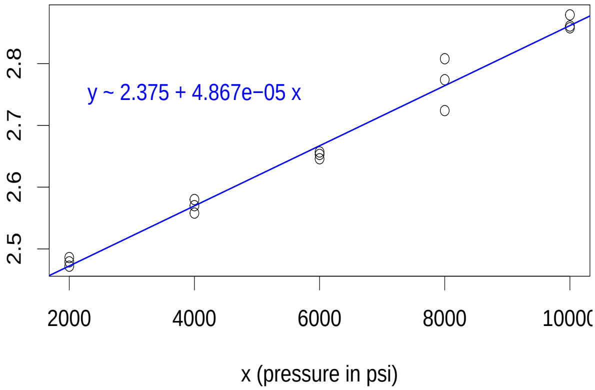
Pressing pressures and specimen densities for a ceramic compound

A mixture of Al_2O_3, polyvinyl alcohol, and water was prepared, dried overnight, crushed, and sieved to obtain 100 mesh size grains. These were pressed into cylinders at pressures from 2,000 psi to 10,000 psi, and cylinder densities were calculated.

x (pressure in psi)	y (density in g/cc)
2000.00	2.49
2000.00	2.48
2000.00	2.47
4000.00	2.56
4000.00	2.57
4000.00	2.58
6000.00	2.65
6000.00	2.66
6000.00	2.65
8000.00	2.72
8000.00	2.77
8000.00	2.81
10000.00	2.86
10000.00	2.88
10000.00	2.86

Scatterplot: ceramics data





- The line, $y \approx 2.375 + 4.867 \times 10^{-5}x$, is the **regression line** fit to the data.

Why fit a regression line?

1. To predict future values of y based on x.
   - I.e., a new ceramic under pressure $x = 5000$ psi should have a density of $2.375 + 4.867 \times 10^{-5} \cdot 5000 = 2.618$ g/cc.
2. To characterize the relationship between x and y in terms of strength, direction, and shape.
   - In the ceramics data, density has a strong, positive, linear association with x.
   - On average, the density increases by 4.867×10^{-5} g/cc for every increase in pressure of 1 psi.

Fitting a linear regression line

- For a response variable y and a predictor variable x, we declare:

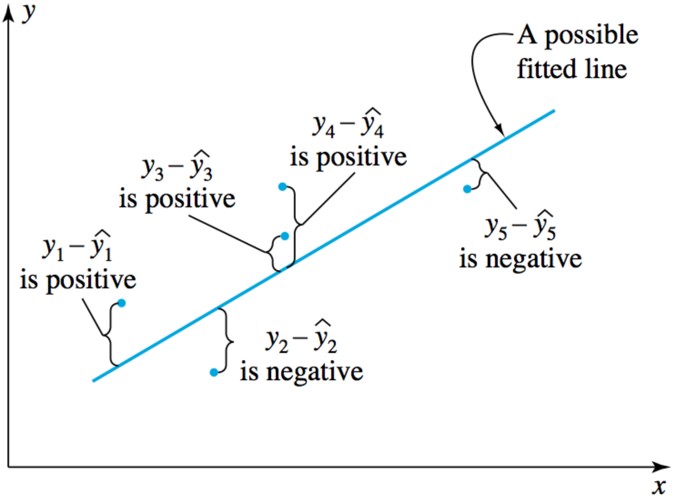
$$y \approx b_0 + b_1 x$$

- and then calculate the intercept b_0 and slope b_1 using **least squares**.
  - We apply the **principle of least squares**: that is, the best-fit line is given by minimizing the **loss function** in terms of b_0 and b_1:

$$S(b_0, b_1) = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2$$

  - Here, $\widehat{y}_i = b_0 + b_1 x_i$

Minimize $\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$ to get the line as close as possible to the points.



How to apply least squares to get the regression line

- From the principle of least squares, one can derive the **normal equations**:

$$nb_0 + b_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

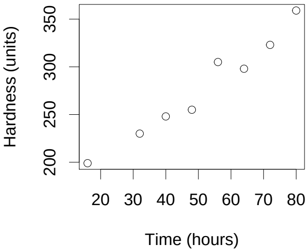
- and then solve for b_0 and b_1:

$$b_1 = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sum (x_i - \overline{x})^2} \qquad b_0 = \overline{y} - b_1 \overline{x}$$

Example: plastics hardness data

Eight batches of plastic are made. From each batch one test item is molded. At a given time (in hours), it hardness is measured in units (assume freshly-melted plastic has a hardness of 0 units). The following are the 8 measurements and times.

time	hardness
32.00	230.00
72.00	323.00
64.00	298.00
48.00	255.00
16.00	199.00
40.00	248.00
80.00	359.00
56.00	305.00



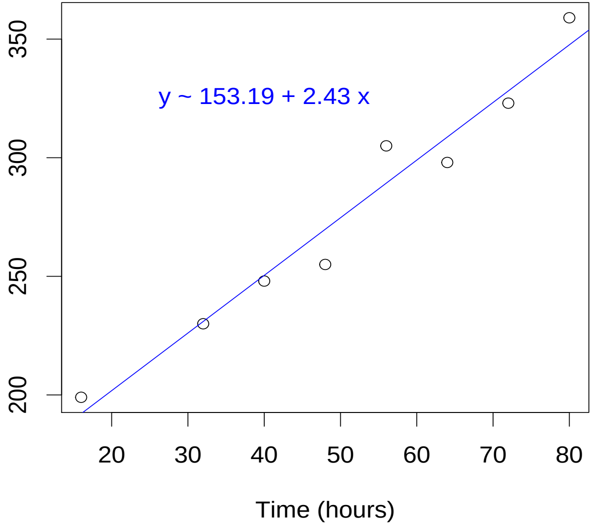
Fitting the line

- $\overline{x} = 51$
- $\overline{y} = 277.125$

x	y	$x_i - \overline{x}$	$y_i - \overline{y}$	$(x_i - \overline{x})(y_i - \overline{y})$	$(x_i - \overline{x})^2$
32.00	230.00	-19.00	-47.12	895.38	361.00
72.00	323.00	21.00	45.88	963.38	441.00
64.00	298.00	13.00	20.88	271.38	169.00
48.00	255.00	-3.00	-22.12	66.38	9.00
16.00	199.00	-35.00	-78.12	2734.38	1225.00
40.00	248.00	-11.00	-29.12	320.38	121.00
80.00	359.00	29.00	81.88	2374.38	841.00
56.00	305.00	5.00	27.88	139.38	25.00

- $\sum(x_i - \overline{x})(y_i - \overline{y}) = 895.38 + 963.38 + \cdots 139.38 = 7765$
- $\sum(x_i - \overline{x})^2 = 361 + 441 + \cdots 25 = 3192$
- $b_1 = \frac{7765}{3192} = 2.43$
- $b_0 = \overline{y} - b_1\overline{x} = 277.125 - 2.43 \cdot 51 = 153.19$

Plot the line to check the fit.



Interpret the model terms

- $b_1 = 2.43$ means that on average, the plastic hardens 2.43 more units for every additional hour it is allowed to harden.
- $b_0 = 153.19$ means that if the model is completely true, at the very beginning of the hardening process (time = 0 hours), the plastics had a hardness of 153.19 on average.
  - But we know that the plastics were completely molten at the very beginning, with a hardness of 0.
  - Don't **extrapolate**: i.e., predict y values beyond the range of the x data.

Linear correlation: a measure of usefulness

- **Linear correlation**: a measure of usefulness of a fitted line, defined by:

$$r = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum(x_i - \overline{x})^2 \sum(y_i - \overline{y})^2}}$$

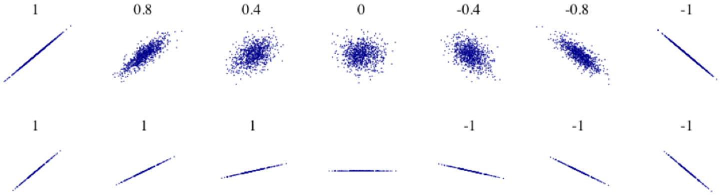
- As it turns out:

$$r = b_1 \frac{s_x}{s_y}$$

where s_x is the standard deviation of the x_i's and x_y is the standard deviation of the y_i's.

Facts about linear correlation

- $-1 \leq r \leq 1$
- $r < 0$ means a negative slope, $r > 0$ means a positive slope
- High $|r|$ means x and y have a strong linear relationship (high correlation), and low $|r|$ implies a weak linear relationship (low correlation).



Coefficient of determination

- **Coefficient of determination**: another measure of the usefulness of a fitted line, defined by:

$$R^2 = \frac{\sum(y_i - \overline{y})^2 - \sum(y_i - \widehat{y}_i)^2}{\sum(y_i - \overline{y})^2}$$

  where $\widehat{y}_i = b_0 + b_1 x_i$.
- Fortunately,

$$R^2 = r^2$$

- Interpretation: R^2 is the fraction of variation in the response variable (y) explained by the fitted line.
- Ceramics data: $R^2 = r^2 = 0.9911^2 = 0.9822792$, so 98.227921% of the variation in density is explained by pressure. Hence, the line is useful for predicting density from pressure.
- Plastics data: $R^2 = r^2 = 0.9796^2 = 0.9596162$, so 95.961616% of the variation in hardness is explained by time. Hence, so the line is useful for predicting hardness from time.

Outline

The informal simple linear regression model

- Up until now, we have looked at fitted lines of the form:

$$y_i = b_0 + b_1 x_i + e_i$$

where:

  - $y_1, \ y_2, \ldots, y_n$ are the fixed, observed values of the response variable.
  - $x_1, \ x_2, \ldots, x_n$ are the fixed, observed values of the predictor variable.
  - b_0 is the estimated slope of the line based on *sample* data.
  - b_1 is the estimated intercept of the line based on *sample* data.
  - e_i is the residual of the i'th unit of the sample.

The formal simple linear regression model

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

- $Y_1, \; Y_2, \ldots, Y_n$ are random variables that describe the response.
- $x_1, \; x_2, \ldots, x_n$ are still fixed, observed values of the predictor variable.
- β_0 is a parameter denoting the *true* intercept of the line if we fit it to the population.
- β_1 is a parameter denoting the *true* slope of the line if we fit it to the population.
- $\varepsilon_1, \; \varepsilon_2, \ldots, \varepsilon_n$ are random variables called **error terms**.

The formal simple linear regression model

- We assume:

$$\varepsilon_1,\ \varepsilon_2, \ldots, \varepsilon_n \overset{\text{iid}}{\sim} N(0, \sigma^2)$$

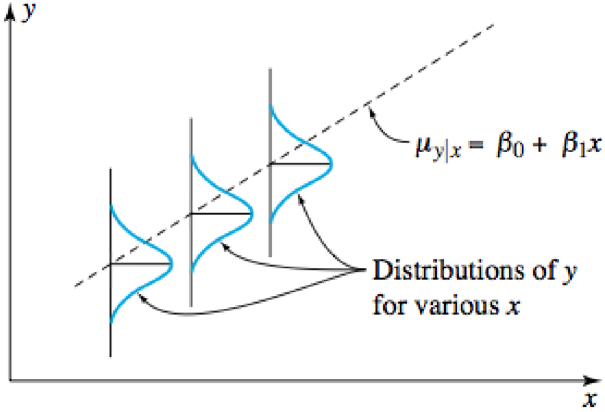
- Which means that for all i:

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

- We often say:

$$\mu_{y|x} = \beta_0 + \beta_1 x$$

The formal simple linear regression model



Outline

The line-fitting sample variance

- Recall:
  - $\widehat{y}_i = b_0 + b_1 x_i$
  - $e_i = y_i - \widehat{y}_i$
- The **line-fitting sample variance**, also called **mean squared error** (MSE) is:

$$s_{LF}^2 = \frac{1}{n-2} \sum_i (y_i - \widehat{y}_i)^2 = \frac{1}{n-2} \sum_i e_i^2$$

  and it satisfies:

$$E(s_{LF}^2) = \sigma^2$$

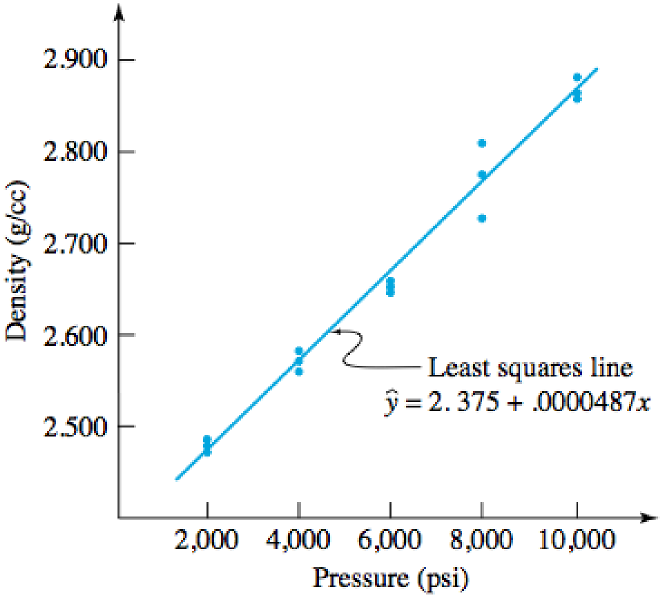
- The line-fitting sample standard deviation is just $s_{LF} = \sqrt{s_{LF}^2}$

Example: ceramics

- A mixture of Al_2O_3, polyvinyl alcohol, and water was prepared, dried overnight, crushed, and sieved to obtain 100 mesh size grains. These were pressed into cylinders at pressures from 2,000 psi to 10,000 psi, and cylinder densities were calculated.

x, Pressure (psi)	y, Density (g/cc)
2,000	2.486
2,000	2.479
2,000	2.472
4,000	2.558
4,000	2.570
4,000	2.580
6,000	2.646
6,000	2.657
6,000	2.653
8,000	2.724
8,000	2.774
8,000	2.808
10,000	2.861
10,000	2.879
10,000	2.858

Example: ceramics



Example: ceramics

- The fitted least squares line is $\widehat{y}_i = 2.375 + 0.0000487x_i$.
- The fitted values $\widehat{y}_i$ are:

Fitted Density Values

x, Pressure	$\hat{y}$, Fitted Density
2,000	2.4723
4,000	2.5697
6,000	2.6670
8,000	2.7643
10,000	2.8617

- And $\sum(y_i - \widehat{y}_i)^2$ is:

$$\begin{aligned}\sum(y_i - \hat{y}_i)^2 &= (2.486 - 2.4723)^2 + (2.479 - 2.4723)^2 + (2.472 - 2.4723)^2 \\ &\quad + (2.558 - 2.5697)^2 + \cdots + (2.879 - 2.8617)^2 \\ &\quad + (2.858 - 2.8617)^2 \\ &= .005153\end{aligned}$$

- Thus, $s^2_{LF} = \frac{1}{n-2}\sum(y_i - \widehat{y}_i)^2 = \frac{1}{15-2} \cdot 0.005153 = 0.00396(g/cc)^2$
- $s_{LF} = \sqrt{s^2_{LF}} = 0.0199g/cc$

Outline

Standardized residuals

- Recall that we assume $\varepsilon_1, \ldots, \varepsilon_n \overset{\text{iid}}{\sim} N(0, \sigma^2)$.
- We also have $E(e_i) = 0$, but because we're *estimating* the slope and intercept instead of using the true slope and intercept,

$$Var(e_j) = \sigma^2 \left(1 - \frac{1}{n} - \frac{(x_j - \overline{x})^2}{\sum_i (x_i - \overline{x})^2}\right)$$

- We don't want $Var(e_j)$ to vary with j, so we define the j'th **standardized residual**:

$$e_j^* = \frac{e_j}{s_{LF}\sqrt{1 - \frac{1}{n} - \frac{(x_j - \overline{x})^2}{\sum_i (x_i - \overline{x})^2}}}$$

which, under our model assumptions, is $\approx N(0, 1)$.

Example: ceramics

- Since $\overline{x} = 6000$, we can calculate $\sum (x_i - \overline{x})^2 = 1.2 \times 10^8$.

Calculations for Standardized Residuals in the Pressure/Density Study

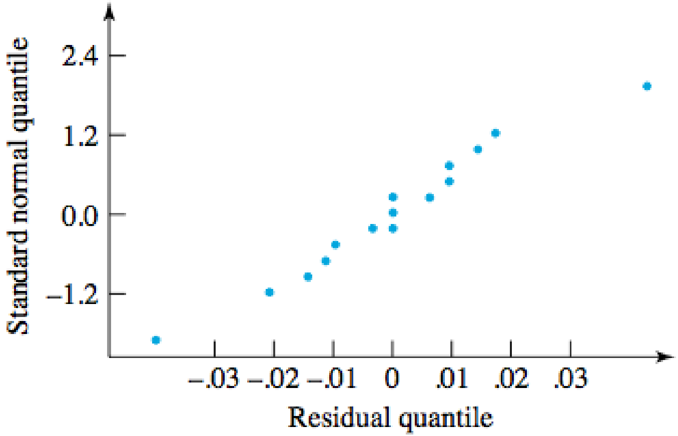
x	$\sqrt{1 - \frac{1}{15} - \frac{(x - 6{,}000)^2}{120{,}000{,}000}}$
2,000	.894
4,000	.949
6,000	.966
8,000	.949
10,000	.894

Example: ceramics

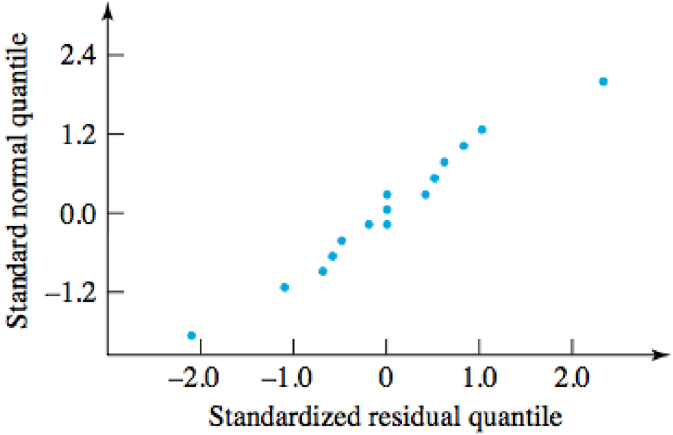
Residuals and Standardized Residuals for the Pressure/Density Study

x	e	Standardized Residual
2,000	.0137, .0067, −.0003	.77, .38, −.02
4,000	−.0117, .0003, .0103	−.62, .02, .55
6,000	−.0210, −.0100, −.0140	−1.09, −.52, −.73
8,000	−.0403, .0097, .0437	−2.13, .51, 2.31
10,000	−.0007, .0173, −.0037	−.04, .97, −.21

Example: ceramics



Example: ceramics



Outline

Inference for the slope parameter

- Since b_1 was estimated from the data, we can treat it as a random variable.
- Under the assumptions of the simple linear regression model,

$$b_1 \sim N\left(\beta_1, \ \frac{\sigma^2}{\sum_i (x_i - \overline{x})^2}\right)$$

- Thus:

$$Z = \frac{b_1 - \beta_1}{\frac{\sigma}{\sqrt{\sum_i (x_i - \overline{x})^2}}} \sim N(0, 1)$$

and

$$T = \frac{b_1 - \beta_1}{\frac{s_{LF}}{\sqrt{\sum_i (x_i - \overline{x})^2}}} \sim t_{n-2}$$

Inference for the slope parameter

- If we want to test $H_0 : \beta_1 = \#$, we can use the test statistic:

$$T = \frac{b_1 - \#}{\frac{s_{LF}}{\sqrt{\sum_i (x_i - \overline{x})^2}}} \sim t_{n-2}$$

  which has a t_{n-2} distribution if H_0 is true and the model assumptions are true.

- We can write a two-sided $1 - \alpha$ confidence interval as:

$$\left(b_1 - t_{n-2,\ 1-\alpha/2} \cdot \frac{s_{LF}}{\sqrt{\sum_i (x_i - \overline{x})^2}}, b_1 + t_{n-2,1-\alpha/2} \cdot \frac{s_{LF}}{\sqrt{\sum_i (x_i - \overline{x})^2}} \right)$$

- The one-sided confidence intervals are analogous.

Example: ceramics

- I will construct a two-sided 95% confidence interval for β_1 ($\alpha = 0.05$).
- From before, $b_1 = 0.0000487$ g/cc/psi, $\sum_i (x_i - \overline{x})^2 = 1.2 \times 10^8$, and $s_{LF} = 0.0199$.
- $t_{n-2,\ 1-\alpha/2} = t_{13,\ 0.975} = 2.16$.
- The confidence interval is then:

$$\left(0.0000487 - 2.16\frac{0.0199}{\sqrt{1.2 \times 10^8}},\ 0.0000487 + 2.16\frac{0.0199}{\sqrt{1.2 \times 10^8}}\right)$$

$$(0.0000448,\ 0.0000526)$$

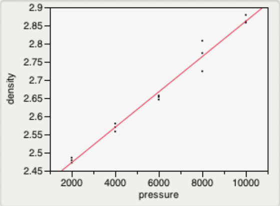
- We're 95% confident that for every unit increase in psi, the density of the next ceramic increases by anywhere between 0.0000448 g/cc and 0.0000526 g/cc.

Example: ceramics

- In JMP:
  - Open the data in a spreadsheet with:
    - 1 column for x
    - 1 column for y
  - For simple linear regression
    - Click Analyze $\rightarrow$ Fit Y by X
    - Y variable - in Y, Response
    - X variable - in X, Factor
    - Click red triangle - Fit line

Example: ceramics

Bivariate Fit of density By pressure



Linear Fit

Linear Fit

density = 2.375 + 4.8667e-5*pressure

Summary of Fit

RSquare	0.982193
RSquare Adj	0.980824
Root Mean Square Error	0.019909
Mean of Response	2.667
Observations (or Sum Wgts)	15

Lack Of Fit

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	0.28421333	0.284213	717.0604
Error	13	0.00515267	0.000396	Prob > F
C. Total	14	0.28936600		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	2.375	0.012055	197.01	<.0001*
pressure	4.8667e-5	1.817e-6	26.78	<.0001*

Example: ceramics

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	2.375	0.012055	197.01	<.0001*
pressure	4.8667e-5	1.817e-6	26.78	<.0001*

- I can construct the same confidence interval using the JMP output:
  - $b_1 = 4.87 \times 10^{-5}$, $t_{n-1,1-\alpha/2} = 2.16$, $\widehat{SD}(b_1) = 1.817 \times 10^{-6}$
  - $$(4.87 \times 10^{-5} - 2.16 \cdot 1.817 \times 10^{-6},\; 4.87 \times 10^{-5} + 2.16 \cdot 1.817 \times 10^{-6}) = (0.0000448,\; 0.0000526)$$

Your turn: ceramics

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	2.375	0.012055	197.01	<.0001*
pressure	4.8667e-5	1.817e-6	26.78	<.0001*

- At $\alpha = 0.05$, conduct a two-sided hypothesis test of $H_0 : \beta_1 = 0$ using the method of p-values.

Answers: ceramics

1. $H_0 : \beta_1 = 0$, $H_a : \beta_1 \neq 0$.
2. $\alpha = 0.05$
3. Use the test statistic:

$$T = \frac{b_1 - 0}{\frac{s_{LF}}{\sqrt{\sum(x_i - \overline{x})^2}}} = \frac{b_1}{\widehat{SD}(b_1)}$$

I assume:

- H_0 is true.
- The model, $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ with errors $\varepsilon_i \sim$ iid $N(0, \sigma^2)$, is correct.

Under these assumptions, $T \sim t_{n-2} = t_{15-2} = t_{13}$

Answers: ceramics

4. Observed test statistic:

$$t = \frac{4.87 \times 10^{-5}}{1.817 \times 10^{-6}} = 26.80 \quad (\text{“t Ratio” in JMP output})$$

$$\text{p-value} = P(|t_{13}| > |26.8|) = P(t_{13} > 26.8) + P(t_{13} < -26.8)$$
$$< 0.0001 \quad (\text{"Prob}> |t|\text{" in JMP output})$$

5. With a p-value $< 0.0001 < 0.05 = \alpha$, we reject H_0 and conclude H_a.
6. There is overwhelming evidence that the true slope of the line is different from 0.

Outline

ANOVA method for testing $H_0 : \beta_1 = 0,\ H_a : \beta_1 \neq 0$

- Another method for testing $H_0 : \beta_1 = 0,\ H_a : \beta_0 \neq 0$ is the "ANalysi Of VAriance" (ANOVA) method.
- Fact: the Total Sum of Squares can be decomposed into Error Sum of Squares and Regression Sum of Squares.

$$\underbrace{\sum_{i=1}^{n}(y_i - \bar{y})^2}_{SSTol} = \underbrace{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}_{SSE} + \underbrace{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})}_{SSR}$$

- Under the assumptions of SLR model, and assuming $H_0 : \beta_1 = 0$ is true, the test statistic

$$F = \frac{SSR/1}{SSE/(n-2)}$$

has a $F_{1,n-2}$ distribution. (Reiview F distribution in ch5part5_Mar_3.pdf.)

ANOVA method for testing $H_0 : \beta_1 = 0,\ H_a : \beta_1 \neq 0$

- We reject $H_0 : \beta_1 = 0$ in favor of $H_a : \beta_1 \neq 0$ when the test statistic F is large. So the p-value is

$$P(F_{1,n-2} > \text{observed } F)$$

- In fact, the square of the t statistic for testing $H_0 : \beta_1 = 0$ is

$$T^2 = \left(\frac{b_1 - 0}{\frac{s_{LF}}{\sqrt{\sum(x_i - \bar{x})^2}}} \right) = \frac{SSR/1}{SSE/(n-2)} = F$$

  which has an $F_{1,n-2}$ distribution if H_0 is true and tends to be large if H_0 is false. So counting large F as evidence against H_0 in favor of $H_a : \beta_1 \neq 0$ is a sensible significance testing method.

ANOVA table

Calculations in the ANOVA method can be summarized in the ANOVA table:

Source	SS	df	MS	F
Regression	SSR	1	MSR = SSR/1	F = MSR/MSE
Error	SSE	$n-2$	MSE = SSE/($n-2$)	
Total	SSTot	$n-1$		

Example: Ceramics

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	0.28421333	0.284213	717.0604
Error	13	0.00515267	0.000396	**Prob > F**
C. Total	14	0.28936600		<.0001*

- The p-value in the F-test is very small. So we reject H_0.
- There is significant evidence that the true slope is different from 0.