Describing Relationships Between Variables (Ch. 4)

x, y

Yifan Zhu

Iowa State University

x_i, y_i

$i = 1, \ldots, n$

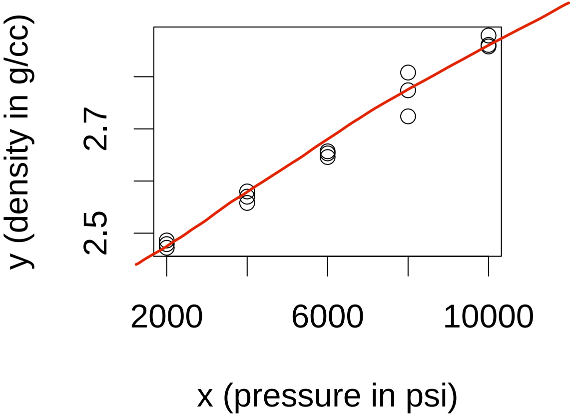
Outline

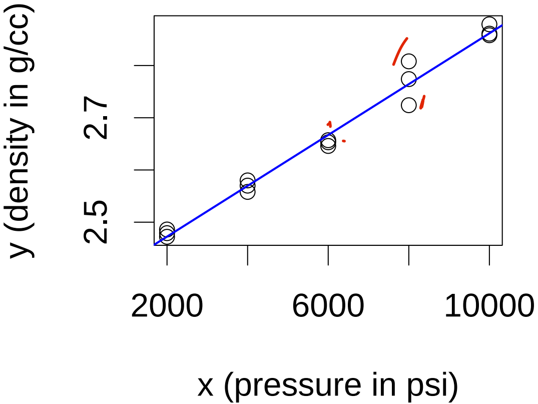
Pressing pressures and specimen densities for a ceramic compound

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 unobserved 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.

only for pressure $\in [2000, 10000]$

Outline

Introduction

Fitting a regression line

Is the model useful?

Is the model valid?

Fitting a linear regression line

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

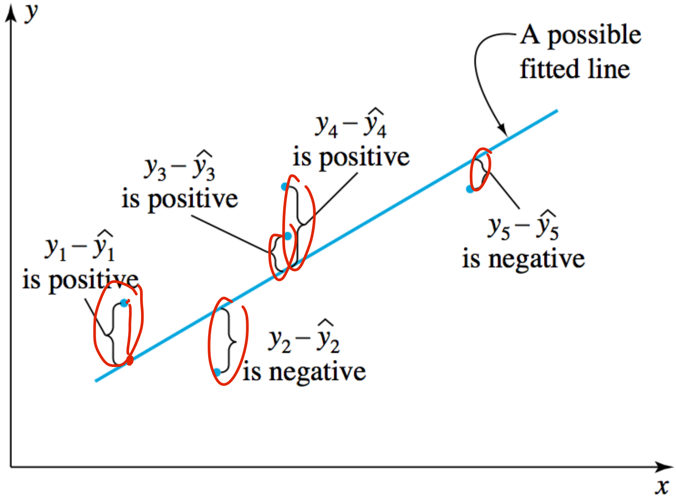
$$y \approx b_0 + b_1 x$$

- 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.



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

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

$$\begin{cases} \frac{\partial}{\partial b_0} S(b_0, b_1) = \sum_{i=1}^{n} 2(y_i - (b_0 + b_1 x_i)) \cdot (-1) = 0 \\ \frac{\partial}{\partial b_1} S(b_0, b_1) = \sum_{i=1}^{n} 2(y_i - (b_0 + b_1 x_i)) \cdot (-x_i) = 0 \end{cases}$$

$\Downarrow$ Normal Equations:

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

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

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:

$$= \hat{\sigma}_{xy}$$

$$b_1 = \frac{\frac{1}{n-1}\sum(x_i - \overline{x})(y_i - \overline{y})}{\underbrace{\frac{1}{n-1}\sum(x_i - \overline{x})^2}_{= s_x^2}} \qquad b_0 = \overline{y} - b_1\overline{x}$$

Describing Relationships Between Variables (Ch. 4)
Yifan Zhu
Introduction
Fitting a regression line
Is the model useful?
Is the model valid?

$$\begin{cases} n b_0 + b_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i & (1) \\ b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i & (2) \end{cases}$$

$$\begin{cases} b_0 + b_1 \bar{x} = \bar{y} \Rightarrow \boxed{b_0 = \bar{y} - b_1 \bar{x}} \\ b_0 \bar{x} + b_1 \cdot \frac{1}{n} \sum_{i=1}^{n} x_i^2 = \frac{1}{n} \sum_{i=1}^{n} x_i y_i \end{cases}$$

$$(\bar{y} - b_1 \bar{x}) \bar{x} + b_1 \cdot \frac{1}{n} \sum_{i=1}^{n} x_i^2 = \frac{1}{n} \sum_{i=1}^{n} x_i y_i$$

$$\bar{x}\bar{y} + \left(\frac{1}{n} \sum_{i=1}^{n} x_i^2 - \bar{x}^2 \right) b_1 = \frac{1}{n} \sum_{i=1}^{n} x_i y_i$$

$$b_1 = \frac{\frac{1}{n} \sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}}{\frac{1}{n} \sum_{i=1}^{n} x_i^2 - \bar{x}^2} = \boxed{\frac{\sum_{i=1}^{n} x_i y_i - n \bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n \bar{x}^2}}$$

$$b_1 = \boxed{\frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}} \overset{?}{=} \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}}$$

$$\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

$$= \sum_{i=1}^{n} (x_i y_i + \bar{x}\bar{y} - x_i\bar{y} - \bar{x} y_i)$$

$$= \sum_{i=1}^{n} x_i y_i + \underbrace{n\bar{x}\bar{y} - \bar{y}\underbrace{\boxed{\sum_{i=1}^{n} x_i}}_{= n\bar{x}} - \bar{x}\boxed{\sum_{i=1}^{n} y_i}}_{-n\bar{x}\bar{y}} = n\bar{y}$$

$$= \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}$$

$$\sum_{i=1}^{n}(x_i-\bar{x})^2$$

$$= \sum_{i=1}^{n}\left(x_i^2+\bar{x}^2-2x_i\bar{x}\right)$$

$$= \sum_{i=1}^{n} x_i^2 + \underbrace{n\bar{x}^2 - 2\bar{x}\boxed{\sum_{i=1}^{n} x_i}}_{-n\bar{x}^2} = n\bar{x}$$

$$= \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$

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

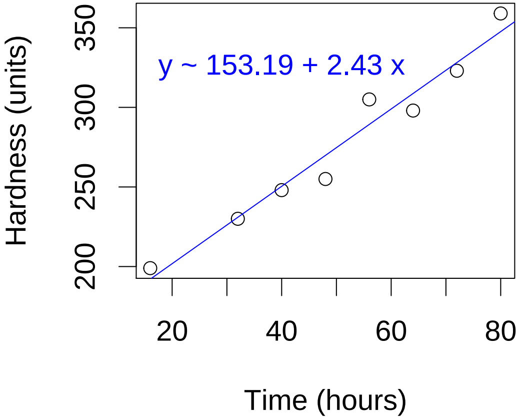
Fitting the line

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

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 at the very beginning of the hardening process (time $= 0$ hours), the plastics had a hardness of 153.19 on average, IF the model is still correct around time 0.
  - 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.

Checking a fitted line

1. Is the model useful?
   - How closely do the points cluster around the line?
   - How strong is the linear relationship between x and y?
   - How much variation in y can be explained by the fitted line?
   - How well can the fitted line predict future values of y?
   - Is the model *precise*?
2. Is the model valid?
   - Should we really be using a straight line to explain y using x, or would some other equation (like a parabola) be better?
   - Does y deviate from the fitted line in some systematic way?
   - Is the model *valid*?

Outline

Linear correlation: a measure of the usefulness of a fitted line

- **Linear correlation**:

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

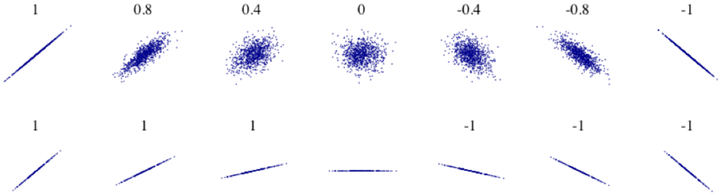
- As it turns out:

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

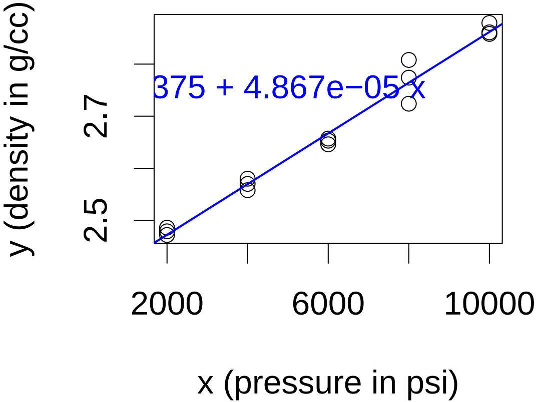
where s_x is the standard deviation of the x_i's and x_y is the standard deviation of the y_i's.

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).



Correlation in the ceramics data



Correlation in the plastics data

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

x	y	$x_i - \overline{x}$	$y_i - \overline{y}$	$(x_i - \overline{x})^2$	$(y_i - \overline{y})^2$	$\Delta x \Delta y$
32.00	230.00	-19.00	-47.12	361.00	2220.77	895.38
72.00	323.00	21.00	45.88	441.00	2104.52	963.38
64.00	298.00	13.00	20.88	169.00	435.77	271.38
48.00	255.00	-3.00	-22.12	9.00	489.52	66.38
16.00	199.00	-35.00	-78.12	1225.00	6103.52	2734.38
40.00	248.00	-11.00	-29.12	121.00	848.27	320.38
80.00	359.00	29.00	81.88	841.00	6703.52	2374.38
56.00	305.00	5.00	27.88	25.00	777.02	139.38

- $\sum(x_i - \overline{x})(y_i - \overline{y}) = 895.39 + 963.38 + \cdots + 139.38 = 7765$
- $\sum(x_i - \overline{x})^2 = 361 + 441 + \cdots + 25 = 3192$
- $\sum(y_i - \overline{y})^2 = 2220.77 + 2104.52 + \cdots + 777.02 = 19682.875$
- $r = \frac{(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{(x_i - \overline{x})^2 (y_i - \overline{y})^2}} = \frac{7765}{\sqrt{3192 \cdot 1.9683 \times 10^4}} = 0.979635179238839$

CAUTION: the data may be highly correlated even if the *linear* correlation, r, is low.

Coefficient of determination

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

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

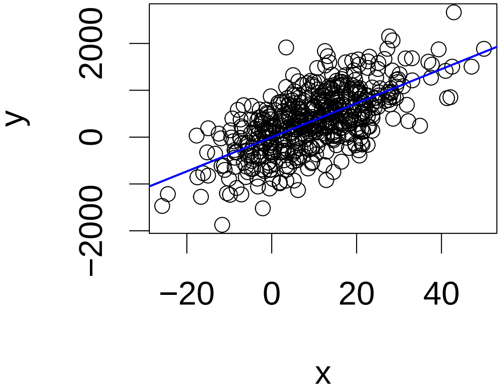
  where $y_i = b_0 + b_1 x_i$.
- Fortunately,

$$R^2 = r^2$$

- Interpretation: R^2 is the fraction of variation in the response variable (y) explained by the fitted line.
- Ceramics data: $R^2 = r^2 = 0.9911^2 = 0.98227921$, so 98.23% of the variation in density is explained by a linear equation in terms of pressure. Hence, the line is useful for predicting density from pressure.
- Plastics data: $R^2 = r^2 = 0.9796^2 = 0.95961616$, so 95.96% of the variation in hardness is explained by a linear equation in terms of time. Hence, so the line is useful for predicting hardness from time.

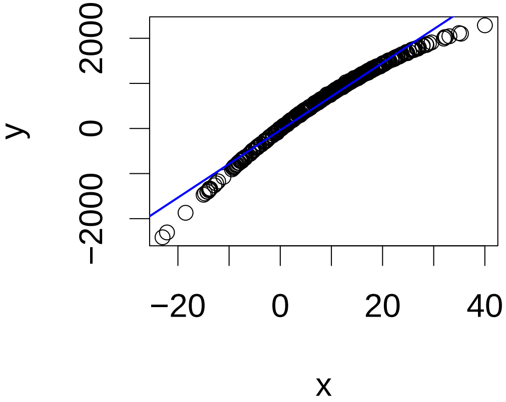
R^2 measures *usefulness* (or precision), not validity.

- x and y can have a true linear relationship despite a low R^2



Outline

CAUTION: Sometimes, the true relationship between x and y is not linear, despite a high R^2



Residuals: a way to check the validity of a fitted line

- **Residuals**: numbers e_i of the form:

$$\begin{aligned} e_i &= y_i - \widehat{y}_i \\ &= y_i - (b_0 + b_1 x_i) \end{aligned}$$

- Instead of:

$$y_i \approx b_0 + b_1 x_i$$

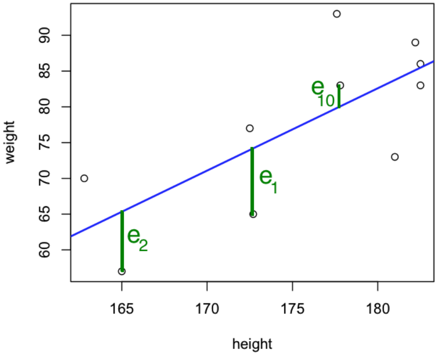
or:

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

you can now write:

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

What do residuals mean? (Scatterplot: heights and weights of 10 elderly men)

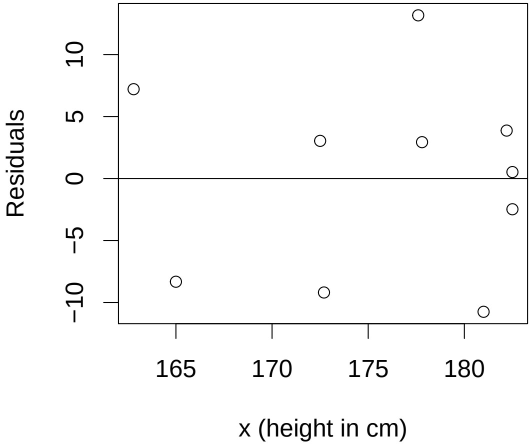


- Residuals are the vertical distances between the points and the fitted line.

Residuals: heights and weights of elderly men data

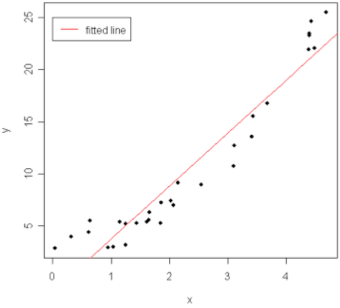
x_i (height in cm)	y_i (weight in kg)	$\widehat{y}_i$	$e_i = y_i - \widehat{y}_i$
172.70	65.00	74.19	-9.19
165.00	57.00	65.32	-8.32
172.50	77.00	73.96	3.04
182.20	89.00	85.13	3.87
177.60	93.00	79.83	13.17
181.00	73.00	83.75	-10.75
182.50	83.00	85.48	-2.48
182.50	86.00	85.48	0.52
162.80	70.00	62.79	7.21
177.80	83.00	80.06	2.94

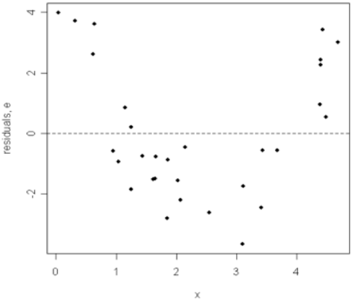
Plots of residuals



Residual plots and validity

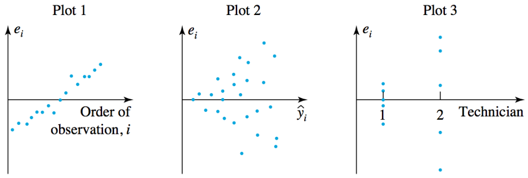
- Left: data that don't fit a line
- Right: the plot of residuals on x
  - The residuals show a nonlinear pattern in the residual plot.
  - Hence, the fitted line is not a valid model.



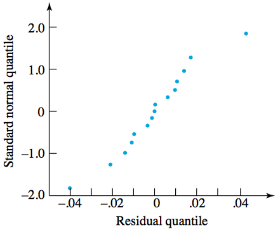


More residual plots and patterns

- All patterns are bad in plots of residual vs. fitted values, x, time, etc.



- When we get to inference, we want to make sure the residuals have a bell-shaped distribution:



- This normal QQ plot shows that the residuals are roughly bell-shaped, which is good.