Chapter 8: Introduction to linear regression

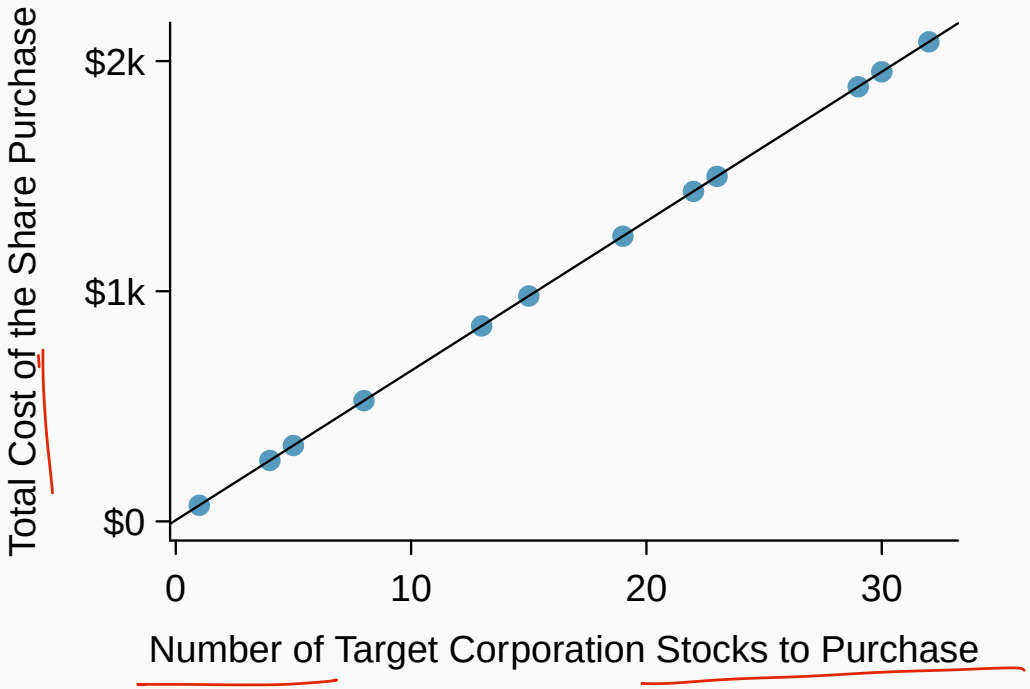
Yonghyun Kwon

Department of Mathematics, Korea Military Academy

Line fitting, residuals, and correlation

Perfect Linear Relationship

- Some data follow a perfect linear relationship.
- Example: Stock purchases computed as $y = 5 + 64.96x$.
- In real-world cases, perfect relationships are rare.



Linear Regression Model

Linear regression

[redacted]

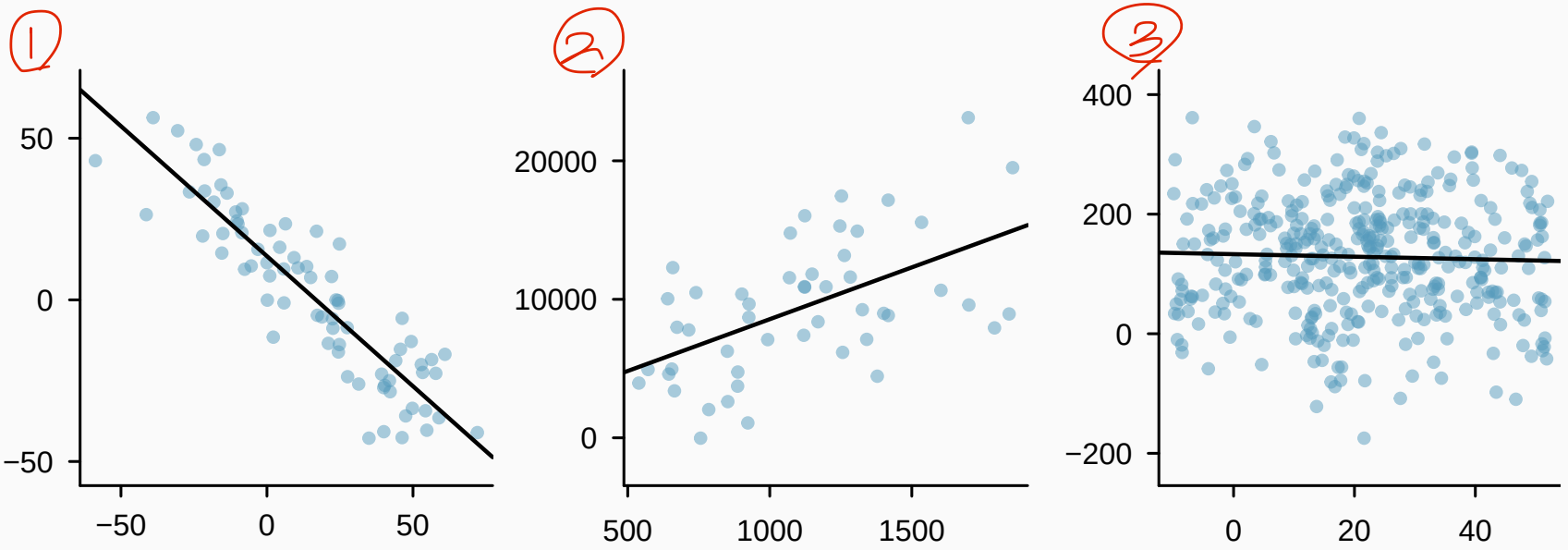
$$y = \beta_0 + \beta_1 x + \varepsilon,$$

[redacted]

- Parameters, β_0 and β_1, are estimated from data.
  - The pointe estimates are written as $\hat{\beta}_0$ and $\hat{\beta}_1$.
- Usually, we use x to predict y.
  - [redacted]

Imperfect Linear Relationships

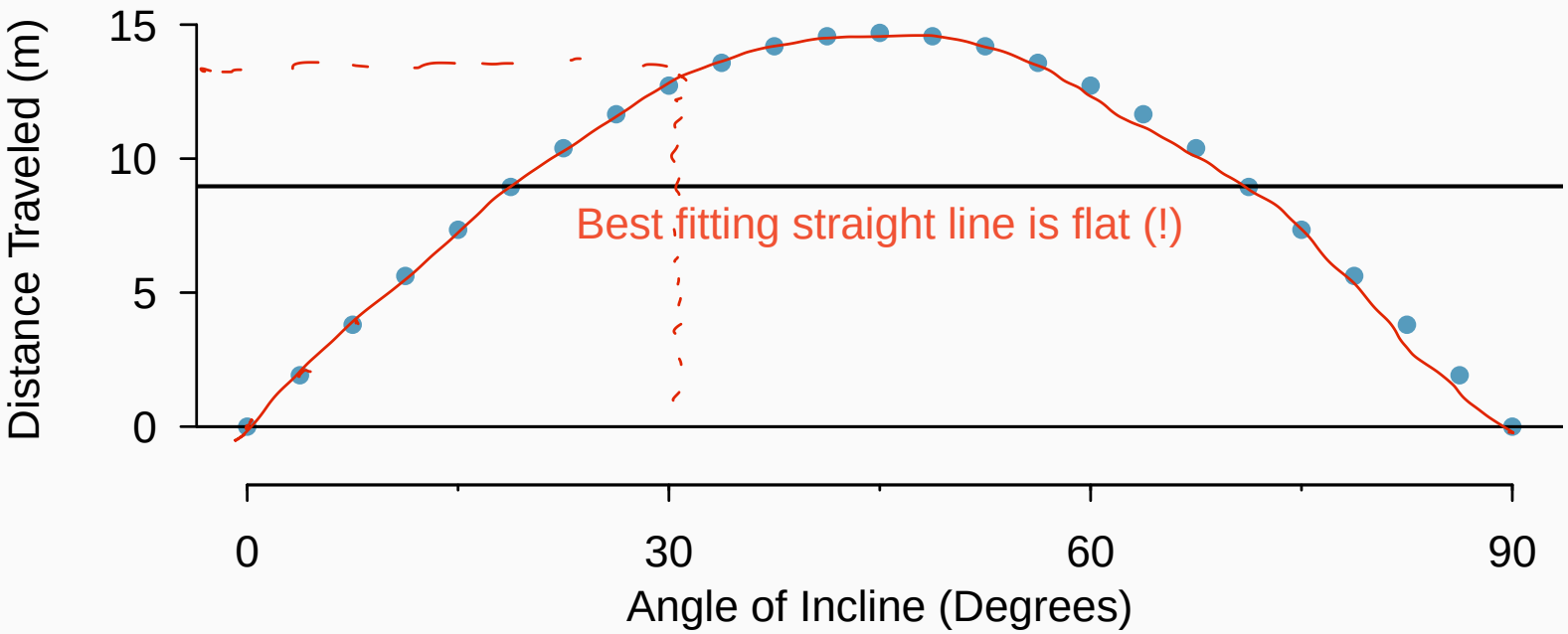
Data usually form a cloud of points around a trend.



Which of the above scatterplots has a (strong downward / moderate upward / weak downward) linear trend?

When a Linear Model is Not Useful

- Some relationships are nonlinear.
- Example: Projectile motion experiment.
- A linear fit is not appropriate.

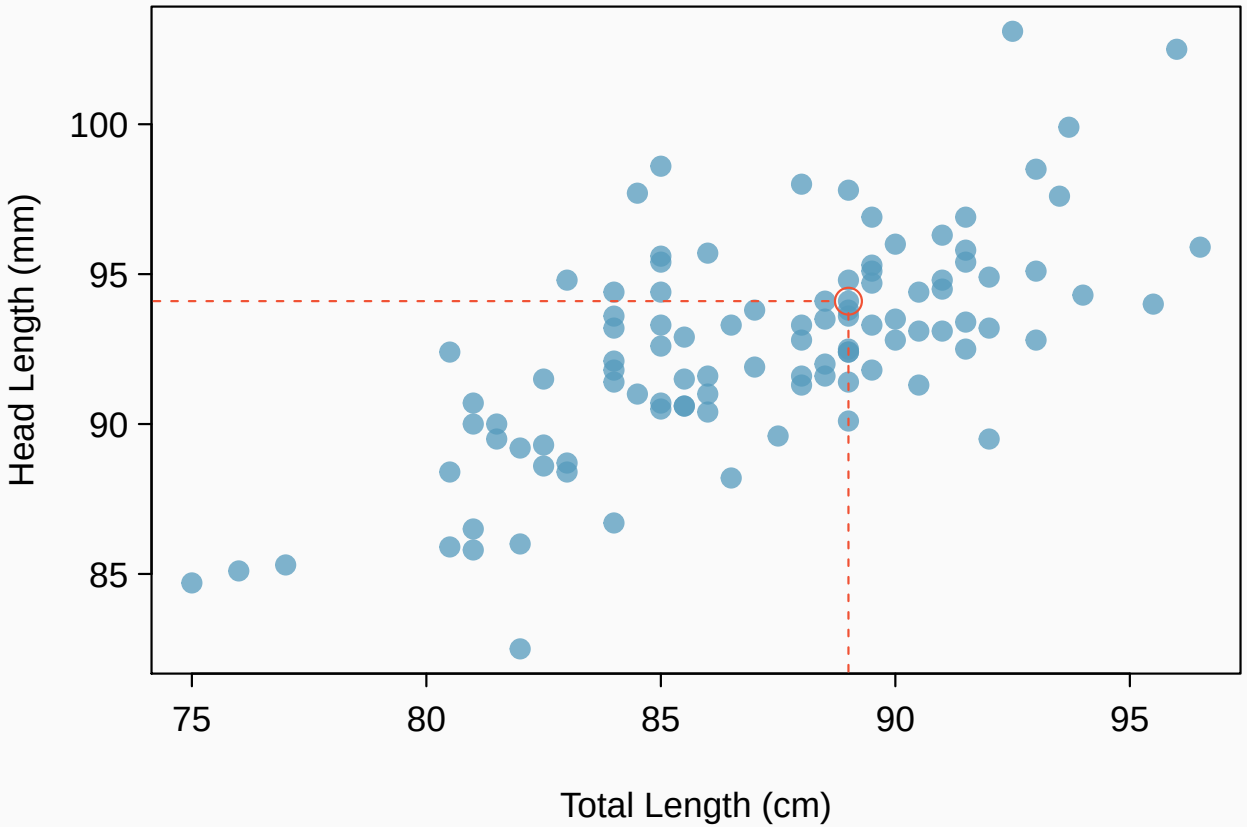


Example: brushtail possums

- Brushtail possums are marsupials found in Australia.
- Researchers measured 104 possums before releasing them.
- We examine the relationship between total length (head to tail) and head length.

Scatterplot of possum Data

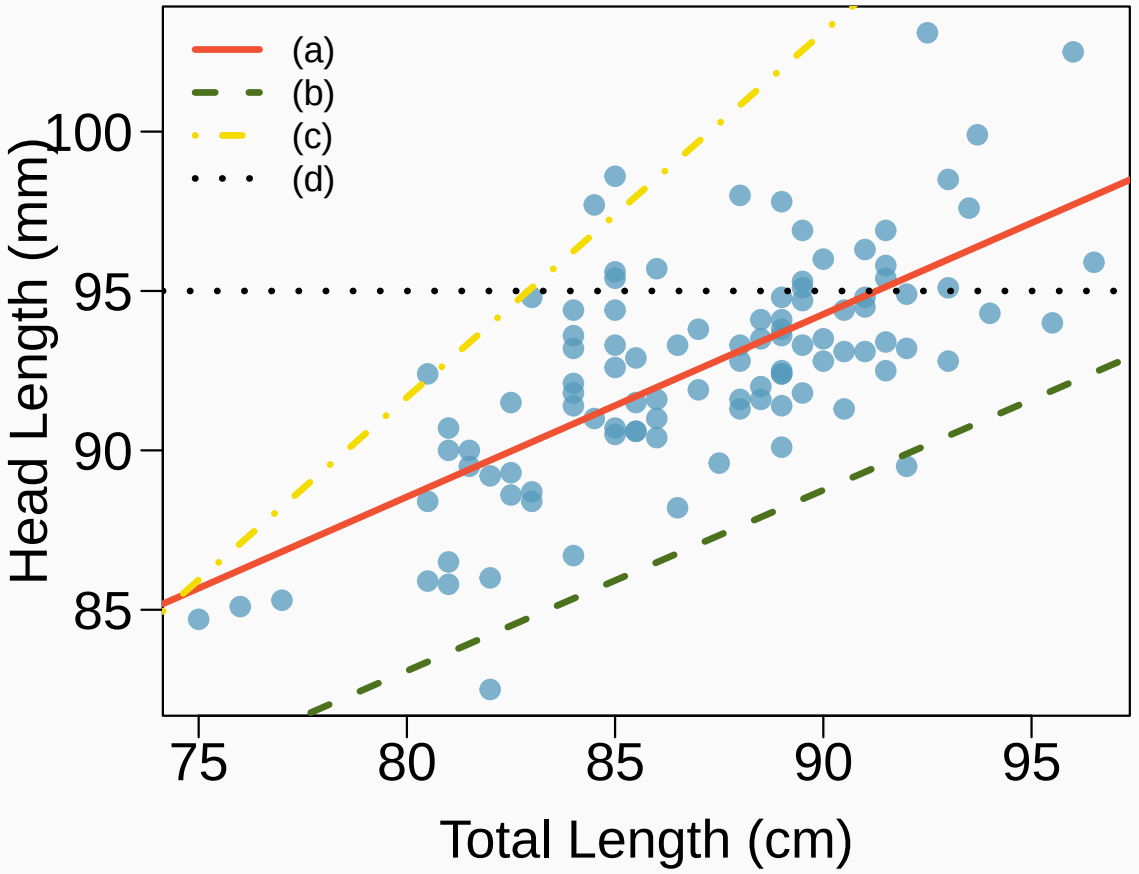
- Head length and total length are positively associated.
- While not perfectly linear, a straight-line model may be useful.



Eyeballing the line

Which of the following appears to be the line that best fits the linear relationship between head length and total length? Choose one.

(a)



Linear Regression Model

- We use total length (x) to predict head length (y).
- Estimated regression equation:

(predicted value, fitted value)

Estimated regression coefficients

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = 41 + 0.59x$$

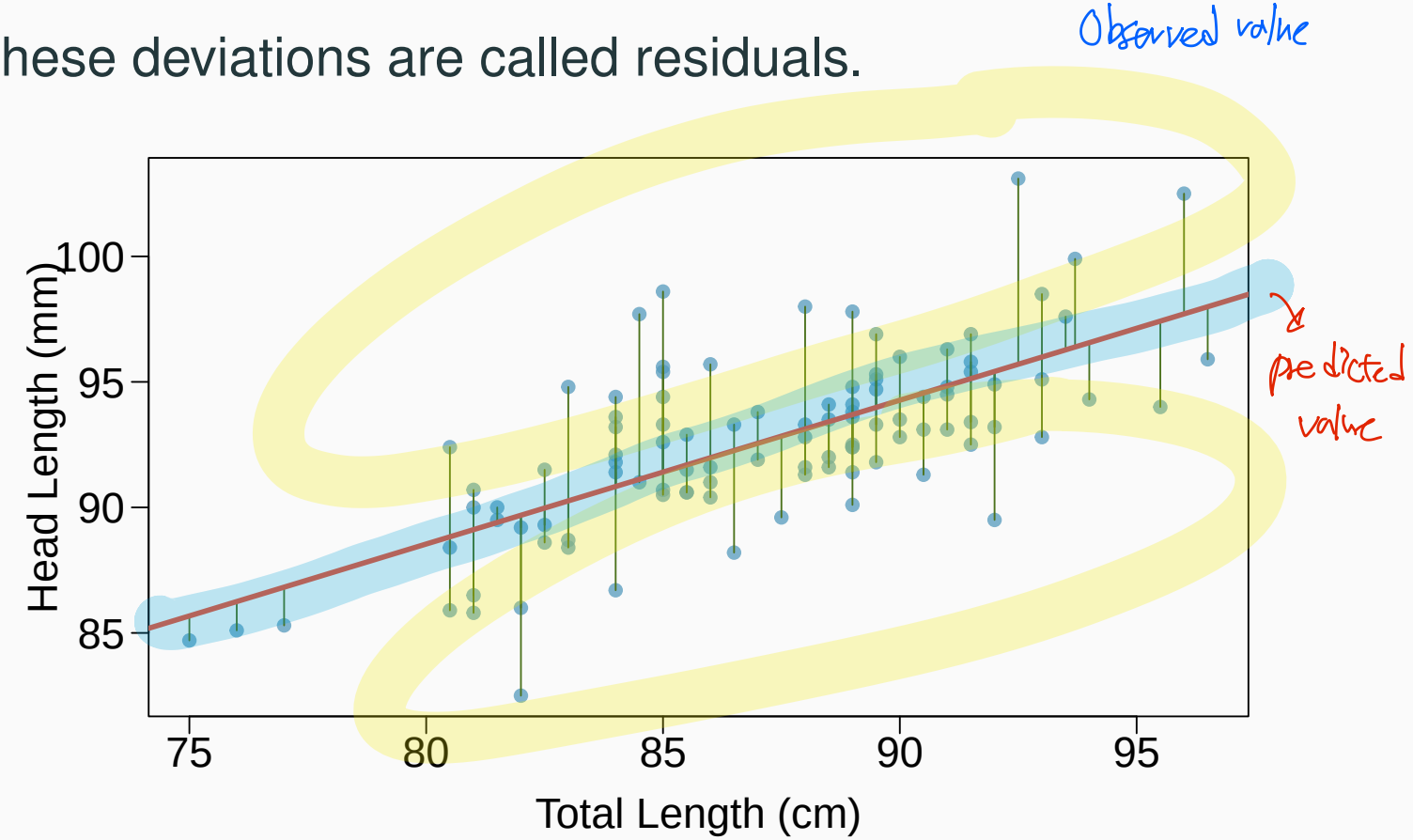
- Example: For a possum with a total length of 80 cm:

$$\hat{y} = 41 + 0.59 \times 80 = 88.2 \quad (\text{mm})$$

- The equation predicts that possums with a total length of 80cm will have an *average* head length of 88.2mm.

Regression Line and Residuals

- A fitted regression line represents the trend.
- Some observations fall above or below the line.
- These deviations are called residuals.



Residuals

Residuals are the leftovers from the model fit

- Data = Fit + Residual
- Each observation has a residual.

Residual: Difference between observed and expected

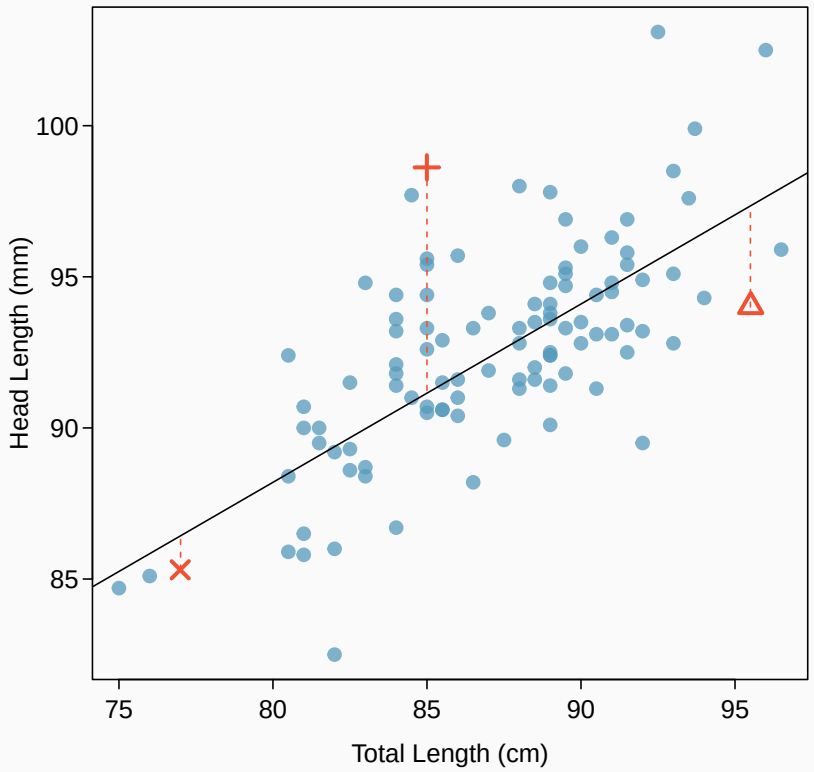
The residual of the i-th observation (x_i, y_i) is the difference between the observed (y_i) and predicted $(\hat{y}_i)$:

$$e_i = y_i - \hat{y}_i$$

observed − predicted

Residuals (cont.)

The linear fit is given as $\hat{y} = 41 + 0.59x$. Based on this line, formally compute the residual of the observation (85, 98.6), denoted by "+".



The predicted value is

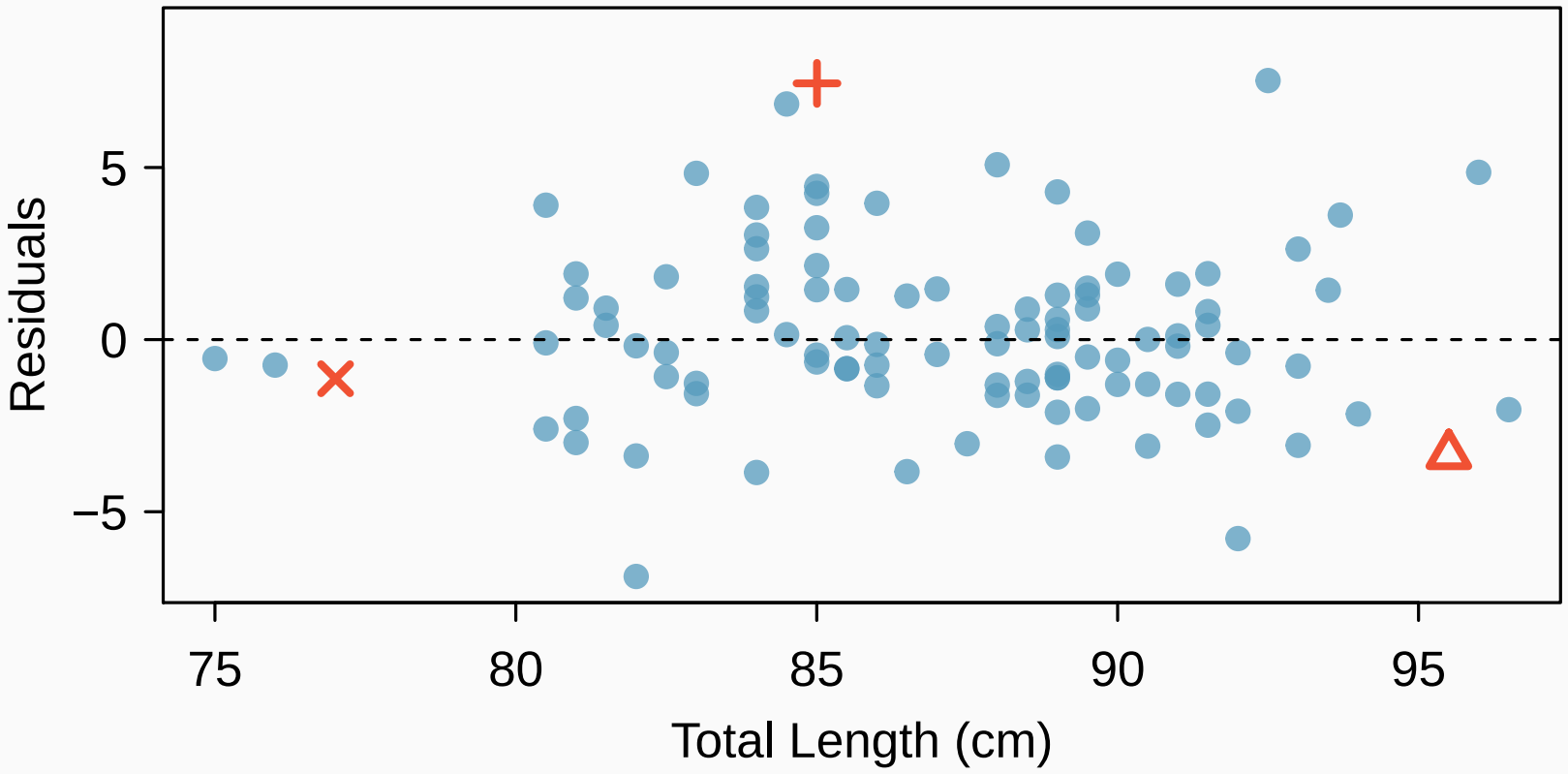
$$\hat{y}_i = 41 + 0.59 \times 85$$
$$= 91.15$$

The residual is given by

$$e_i = 98.6 - 91.15$$
$$= 7.45$$

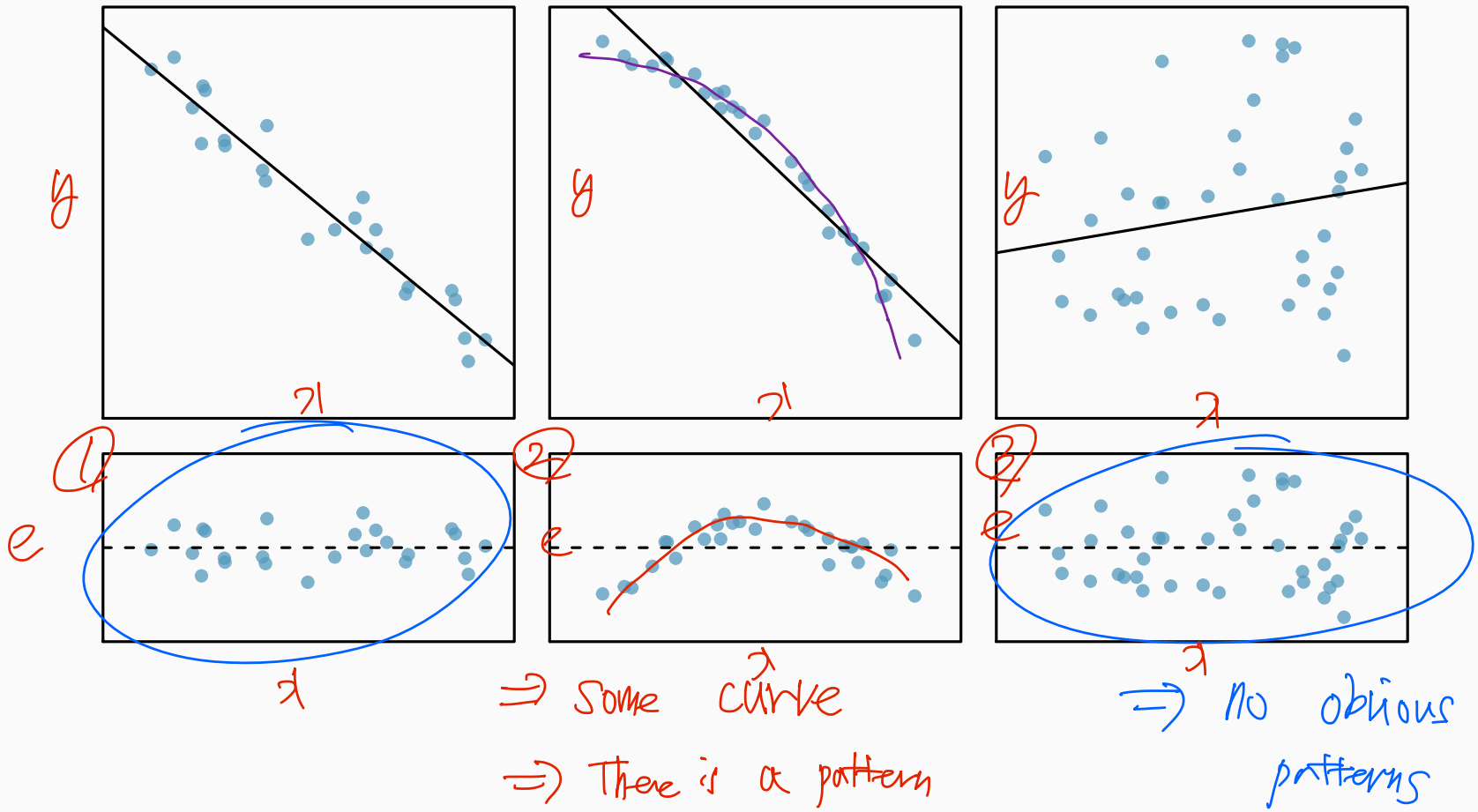
Residual Plot

- *Residual plot* is the scatter plot of (x_i, e_i).
- Residuals should be randomly scattered around the dashed line that represents 0.



Identifying Patterns in Residuals

We have three scatterplots with linear models in the first row and residual plots in the second row. Identify any patterns remaining in the residuals.



Correlation

Correlation

Correlation, or sample correlation coefficient, for observations $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$ is defined as

$$R = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}},$$

where $S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2$, $S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2$, and $S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$.

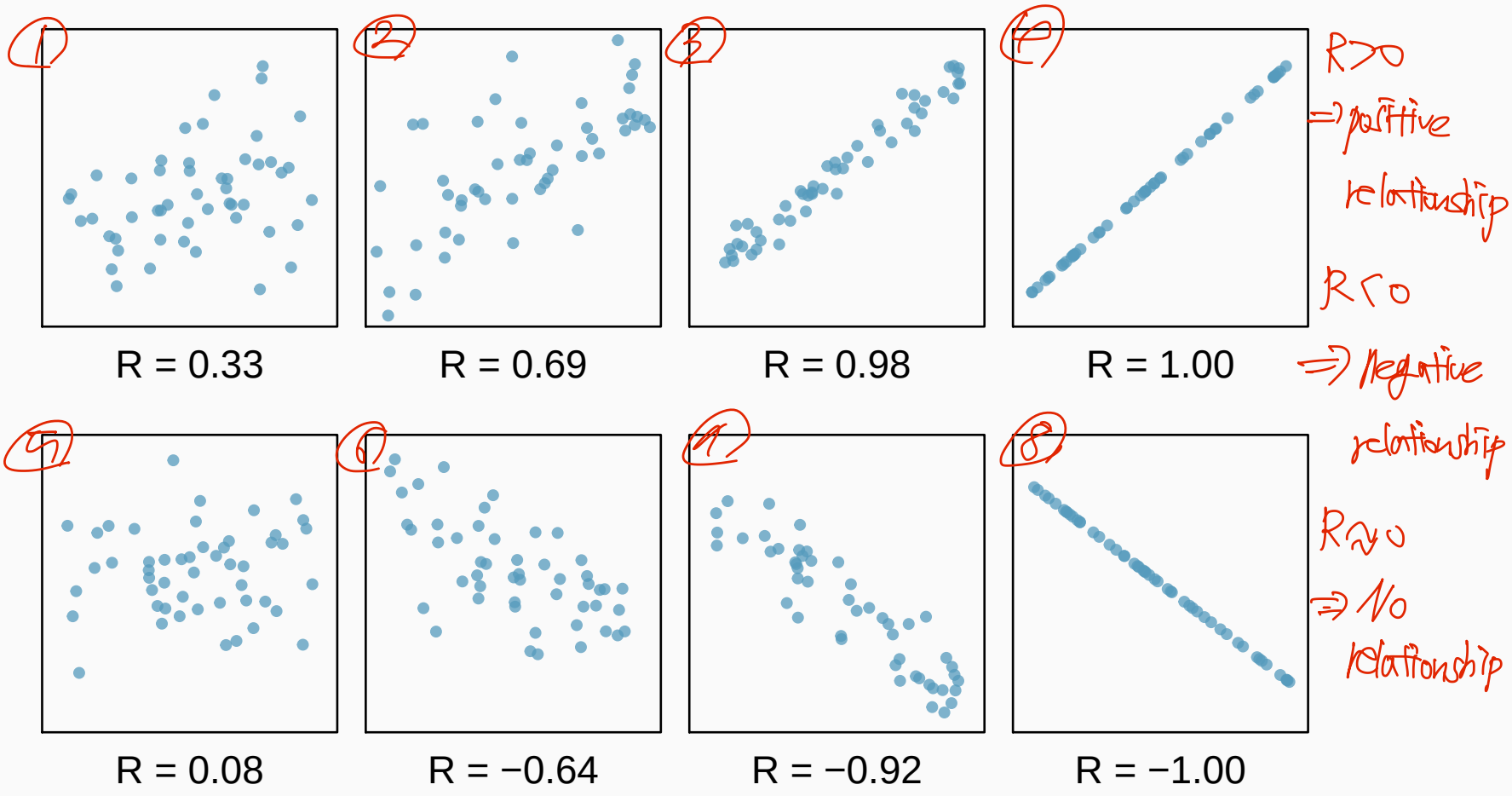
$\propto$ Sample variances

$\propto$ Sample covariance

- *Correlation* quantifies the strength of a *linear* relationship between two variables.
- It always takes values between -1 and 1.
- Perfect correlation: $R = \pm 1$, No correlation: $R = 0$.

Correlation in Practice

- Computed using means and standard deviations of variables.
- Example scatterplots with varying correlations:



Guessing the correlation

Which of the following is the best guess for the correlation between head length and total length?

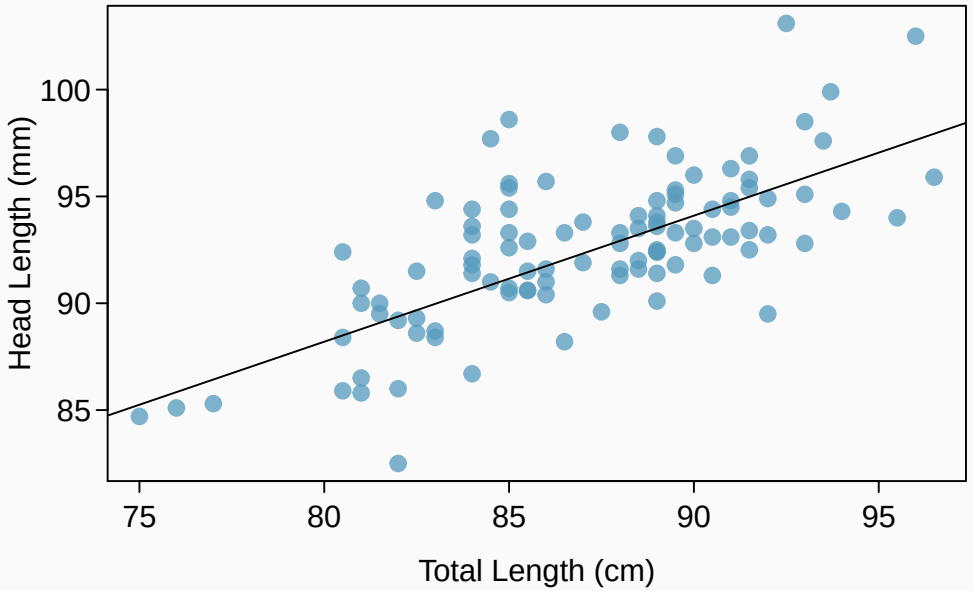
(a) -0.75

(b) 0.69

(c) -0.1

(d) 0.02

(e) -1.5



Does correlation depend on the unit of measure?

Recall: Correlation of two random variables X and Y

Correlation, or population correlation coefficient, of X and Y is

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} = \frac{E\left[(X-\mu_X)(Y-\mu_Y)\right]}{\sqrt{E\left[(X-\mu_X)^2\right]E\left[(Y-\mu_Y)^2\right]}},$$

where $\mu_X = E(X)$ and $\mu_Y = E(Y)$.

Sample correlation coefficient R can be used to estimate the population correlation coefficient $\rho = \rho(X,Y)$.

- For instance, in the possum example, $\hat{\rho} = r = 0.69$.

We can even test whether there exists linear relationship:

$$H_0 : \rho = 0, \qquad H_A : \rho \neq 0.$$

Consider testing $H_0 : \rho = 0$ versus $H_A : \rho \neq 0$. $\rho>0$, $\rho<0$

Correlation test using t-distribution

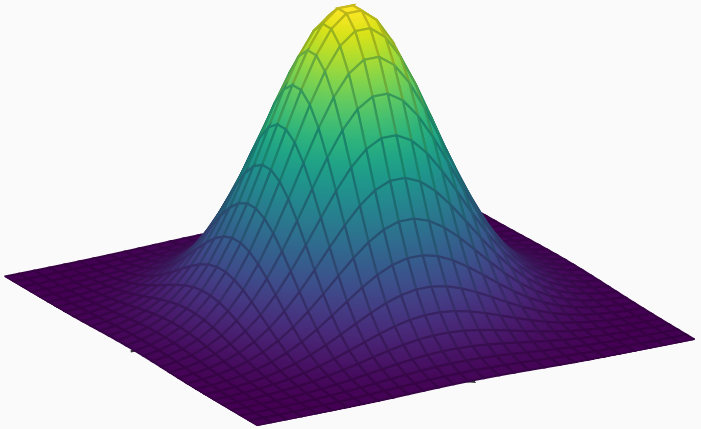
If $(X_1, Y_1), \cdots, (X_n, Y_n)$ are independent observations from a bivariate normal distribution, the test statistic and its null distribution are

$$T = \frac{R\sqrt{n-2}}{\sqrt{1-R^2}} \overset{H_0}{\sim} t(n-2),$$

where R is the sample correlation coefficient.

Note: Bivariate normal distribution is a generalization of normal distribution to two dimensions.

Pdf of normal distribution



Joint pdf of bivarite normal distribution

Back to the brushtail possums example

We try to determine if there is a correlation between total length and head length. The hypotheses are:

H_0 : $\rho = 0$ (No correlation between total and head length)

H_A : $\rho \neq 0$

Assumptions:

- Observations are independent.
- The data come from a bivariate normal distribution.

Computing the test statistic

- Summary statistics:

$$n = 104, \qquad r = 0.69$$

- Test statistic formula:

$$T = \frac{R\sqrt{n-2}}{\sqrt{1-R^2}} \sim t(n-2) \quad \text{under } H_0$$

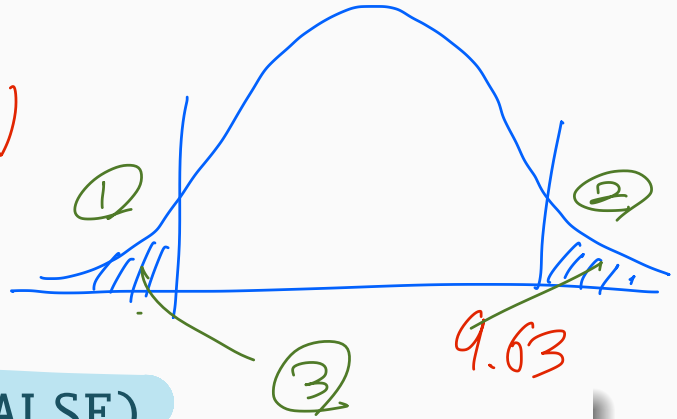
Observed test statistic:

$$\frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.69\sqrt{102}}{\sqrt{1-0.69^2}} = 9.63 \sim t(102)$$

Finding the P-value

P-value of the two-sided test is $2 \times P(T > 9.63) = 5.36 \times 10^{-16}$

$T \sim t(102)$




```
> pt(9.63, df = 102, lower.tail = FALSE)
[1] 2.684797e-16
```

- P-value is less than $\alpha = 0.05$, so we (reject / don't reject) H_0.
- There is (sufficient / insufficient) evidence to conclude that there is a correlation between total length and head length in brushtail possums.

R code

```
> cor.test(total_l,head_l, alternative = "two.sided")
```

R output

```
        Pearson's product-moment correlation

data:  total_l and head_l
t = 9.6569, df = 102, p-value = 4.681e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5750415 0.7798823
sample estimates:
      cor
0.6910937
```

= R

Practice ⇒ Homework

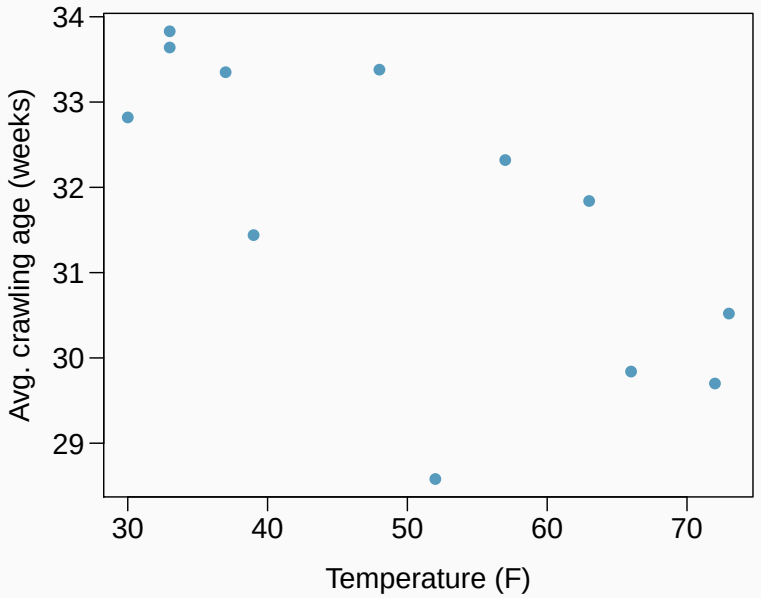
(Exercise 8.12*) A study recorded data on 12 groups of infants born throughout the year to investigate whether crawling age(x) is associated with environmental temperature(y). For each group, researchers measured the average crawling age (in weeks) and the average temperature (in °F) when the babies were six months old. Conduct a hypothesis test to determine if there is a **negative** correlation between temperature and crawling age. (The data come from a bivariate normal distribution.)

1. State the null and alternative hypotheses.(Use one-sided test)

2. Find the test statistic and its null distribution.

Practice

Summary statistics	
n	12
$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$	34.0961
$S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2$	2762.25
$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$	-214.735



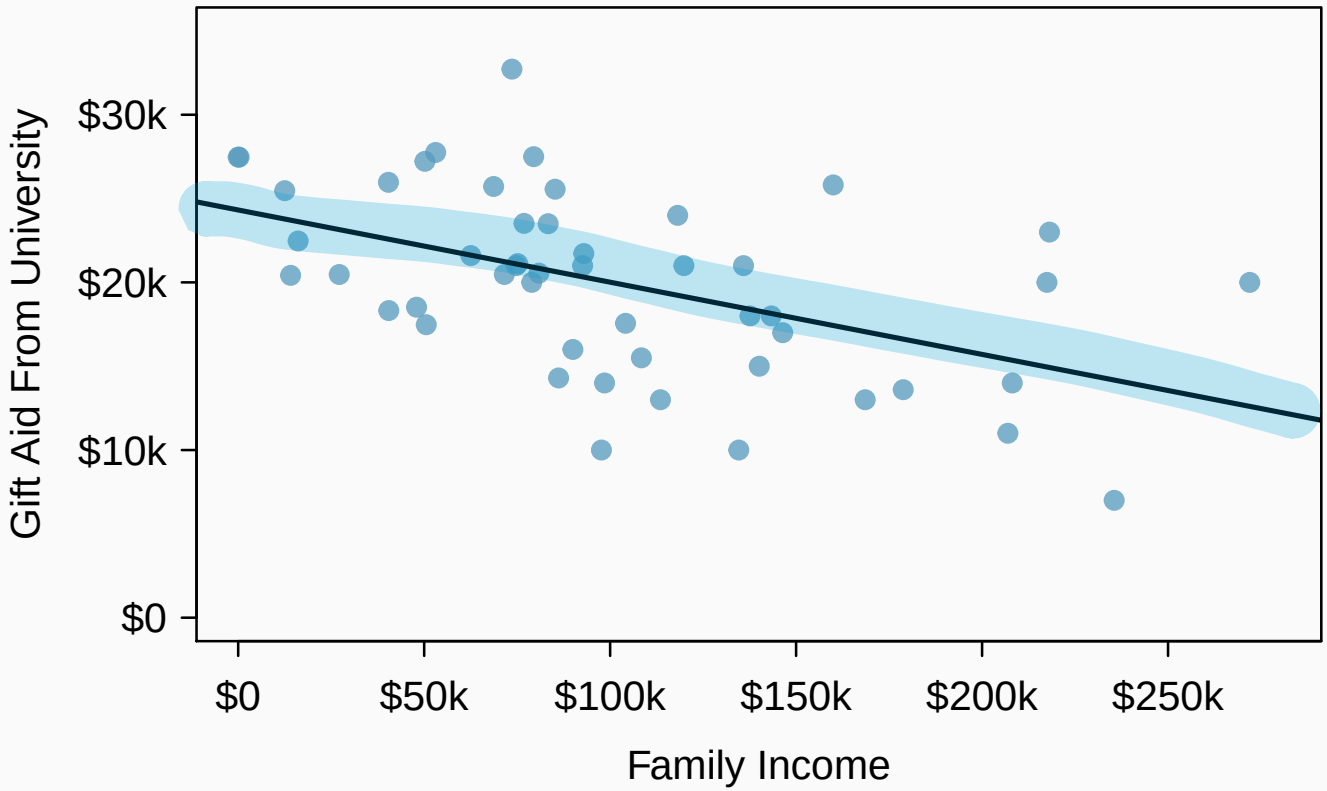
3. Compute the correlation coefficient and the observed test statistic.

4. Compute the p-value and complete the hypothesis test.

Fitting a line by least squares regression

Elmhurst College Data

- Study of 50 freshman students at Elmhurst College, Illinois.
- Examines the relationship between family income and gift aid.
- Gift aid is financial support that does not need to be paid back.



Least squares regression

- We try to find a regression line that minimizes residuals.
  One way is to minimize the sum of squared residuals:

$$e_1^2 + e_2^2 + \cdots + e_n^2,$$

  which is called *least square regression*.
- Squaring residuals gives a higher penalty to large errors.

The Least Squares Regression Line

Consider a linear regression model:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \cdots, n,$$

where β_0 is the *intercept* and β_1 is the *slope*.

Least squares estimation for regression coefficients

The least squares estimation for parameters β_0 and β_1 minimizes

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

with respect to β_0 and β_1.

$= \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2$

observed

$\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \leq \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$

where β_0, β_1 are any values

Least square estimates for the regression coefficients

The least square estimates that minimize $S(\beta_0, \beta_1)$ are

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{s_y}{s_x} R,$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

where

$$S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}), \quad S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2,$$

R is the sample correlation coefficient, $\bar{x}$, $\bar{y}$ are the sample means of x_i's, and y_i's, s_x and s_y are the sample standard deviations.

Note: To minimize $S(\beta_0, \beta_1)$, consider normal equations:

$$\frac{\partial S(\hat{\beta}_0, \hat{\beta}_1)}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \tag{1}$$

$$\frac{\partial S(\hat{\beta}_0, \hat{\beta}_1)}{\partial \beta_1} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i = 0 \tag{2}$$

(1) *gives* $\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$. *Plugging* $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ *into* (2),

$$\sum_{i=1}^{n} (y_i - \bar{y} - \hat{\beta}_1 x_i - \hat{\beta}_1 \bar{x}) x_i = \sum_{i=1}^{n} (y_i - \bar{y} - \hat{\beta}_1 x_i - \hat{\beta}_1 \bar{x})(x_i - \bar{x}) = 0$$

gives

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}} = \underbrace{\frac{\sqrt{S_{yy}}}{\sqrt{S_{xx}}}}_{s_y / s_x} \cdot \underbrace{\frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}}_{R} = \frac{s_y}{s_x} R.$$

Elmhurst Data Summary Statistics

Using the following summary statistics for Elmhurst data, find the least square estimates for regression coefficients.

	Family Income (x)	Gift Aid (y)
Mean	101,780	19,940
Standard Deviation	63,200	5,460
Correlation		$R = -0.499$

$$\hat{\beta}_1 = \frac{S_y}{S_x} R = \frac{5460}{63200}(-0.499) = -0.0431 = -4.31 \times 10^{-2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 19940 - (-0.0431) \times 101780$$

$$= 24326.72 \approx 2.4 \times 10^4$$

Regression Output from R

```
> g <- lm(gift_aid ~ family_income, data = elmhurst)
> summary(g)$coefficients
```

```
                  Estimate Std. Error t value   Pr(>|t|)
(Intercept)     2.4319e+04 1.2915e+03 18.8310 8.2810e-24
family_income -4.3072e-02 1.0809e-02 -3.9846 2.2887e-04
```

$2.4319 \times 10^{4} = \beta_0$

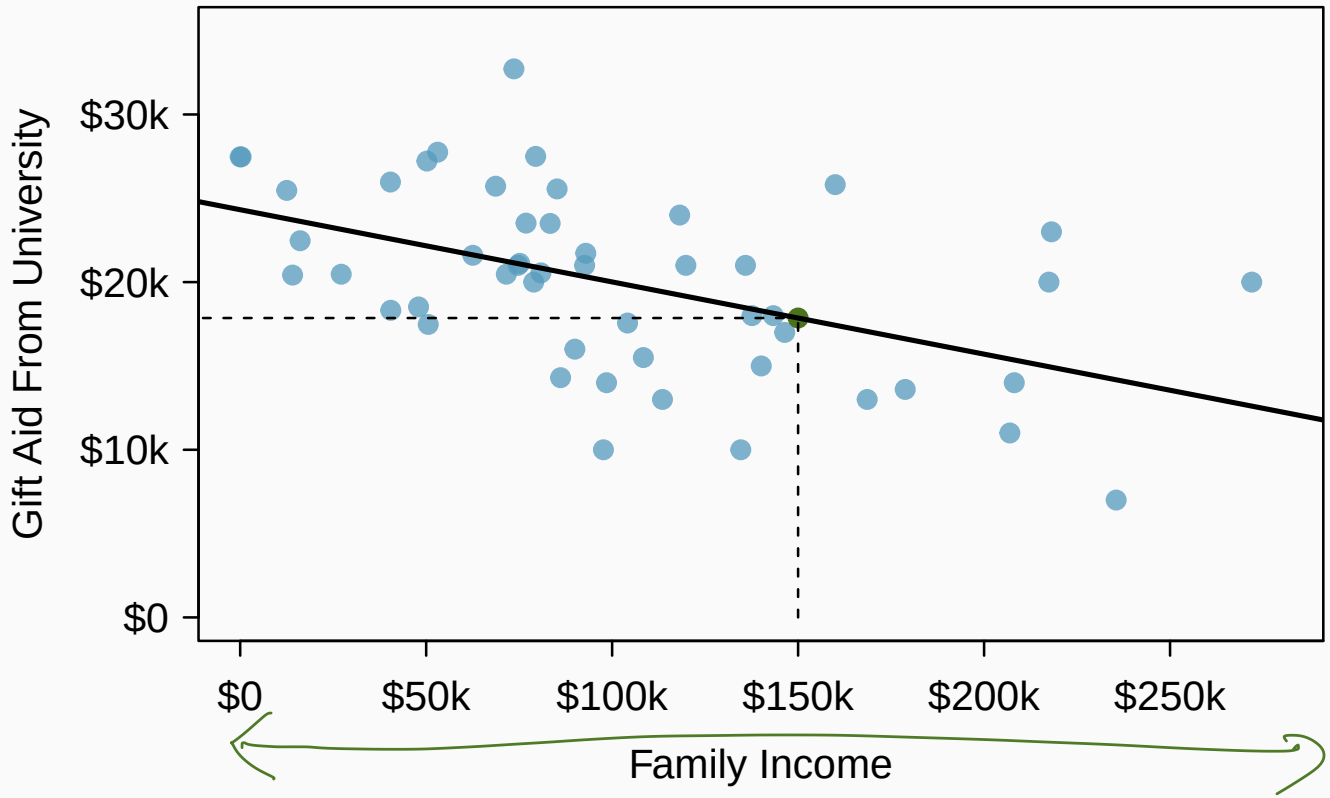
$-4.3072 \times 10^{-2} = \beta_1$

Interpreting the model parameter estimates

- The slope represents the *expected change in y per unit change in x.*
  - $\hat{\beta}_1 = -0.0431$: For each additional \$1,000 in family income, expected aid decreases by $1000 \times \hat{\beta}_1 = \$43.1$
- The intercept describes the *expected outcome of y if* $x = 0$.
  - $\hat{\beta}_0 = 24,319$: if a student's family had no income, the expected aid is $\$24,319$

Prediction

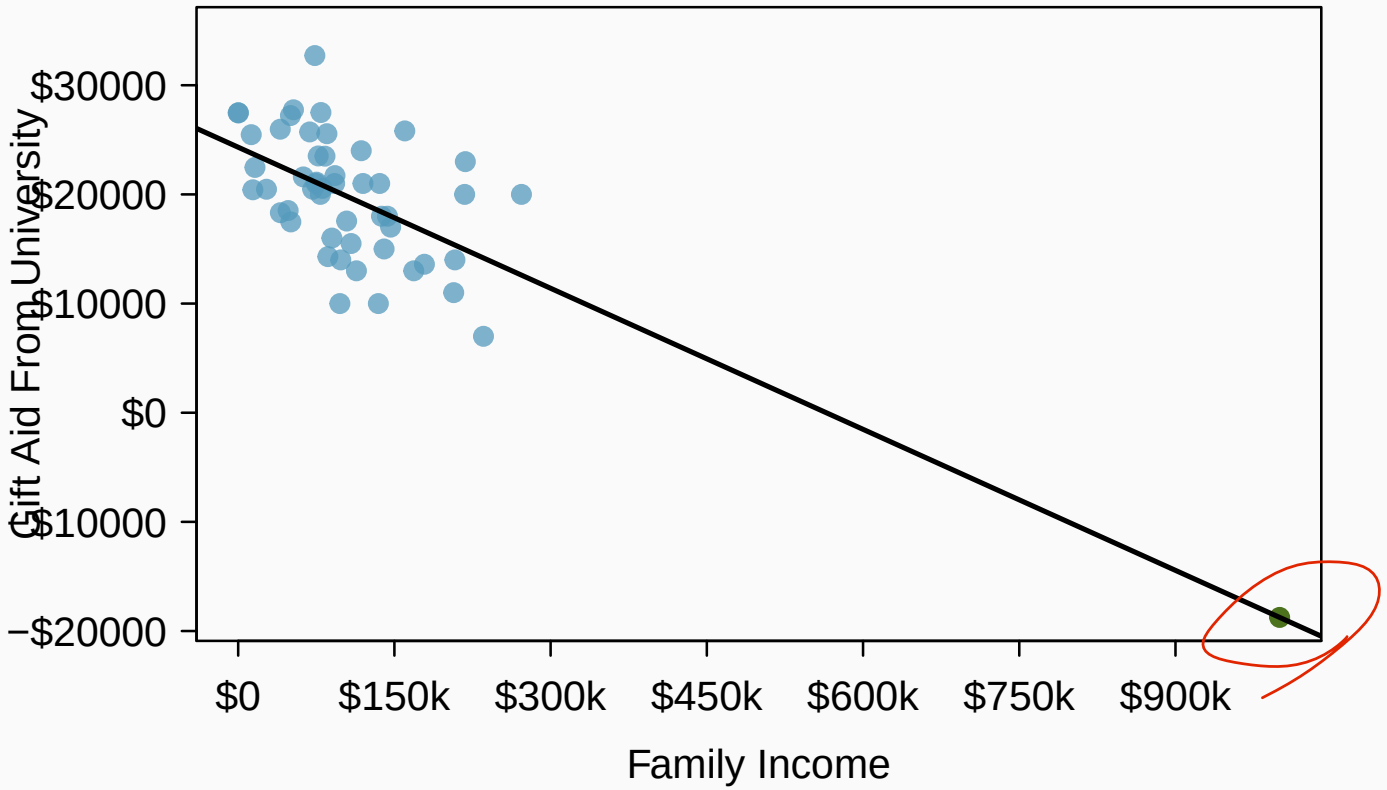
Using the linear model to predict the expected value of y for a given value of x is called *prediction*, and denoted as $\hat{y}$.



$x = 150{,}000 \quad \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 150000 = 17854$

Extrapolation

Extrapolation occurs when we apply a model beyond the observed data range.



Example: Extrapolation Mistake

- Using the model: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

$$\hat{y} = 24,319 - 0.0431 \times x$$

- Predict aid for a student with family income of \$1 million:

$$\begin{aligned} \hat{y} &= 24,319 - 0.0431 \times 1,000,000 \\ &= -18,781 \end{aligned}$$

- This prediction is unrealistic—aid cannot be negative!

$\Rightarrow$ Extrapolation

R^2 measures the proportion of variability in the response variable explained by the regression model.

Coefficient of determination, R^2

Coefficient of determination, R^2, is the square of the sample correlation coefficient, R, and can be computed by

$$R^2 = \frac{SSR}{SST} = \frac{SST - SSE}{SST}, \quad \text{where}$$

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 : \text{ Sum of Squares Regression}$$

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 : \text{ Sum of Squares Error}$$

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 = SSR + SSE : \text{ Total Sum of Squares}$$

Note: *From the definition of sample correlation coefficient R,*

$$R^2 = \frac{S_{xy}^2}{S_{xx}S_{yy}} = \frac{S_{xx}}{S_{yy}}\frac{S_{xy}^2}{S_{xx}^2} = \frac{S_{xx}}{S_{yy}}\hat{\beta}_1^2$$
$$= \frac{\sum_{i=1}^n(\hat{\beta}_1 x_i - \hat{\beta}_1\bar{x})^2}{S_{yy}} = \frac{\sum_{i=1}^n(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n(y_i - \bar{y})^2} = \frac{SSR}{SST}.$$

Furthermore, SST can be decomposed as the sum of SSR and SSE because

$$SST = \sum_{i=1}^n(y_i - \bar{y})^2 = \sum_{i=1}^n(y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2$$
$$= \underbrace{\sum_{i=1}^n(y_i - \hat{y}_i)^2}_{SSE} + \underbrace{\sum_{i=1}^n(\hat{y}_i - \bar{y})^2}_{SSR} + 2\underbrace{\sum_{i=1}^n(y_i - \hat{y}_i)(\hat{y}_i - \bar{y})}_{0},$$

and the last term vanishes since

$$\sum_{i=1}^n(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(\hat{\beta}_0 + \hat{\beta}_1 x_i - \bar{y}) = 0$$

from the normal equations.

Assumption 1: independence and constant variance

We assume a linear regression model:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \cdots, n,$$

where β_0 is the intercept, β_1 is the slope, and ε_i's are the *independent* error terms satisfying $E(\varepsilon_i) = 0$ and $Var(\varepsilon_i) = \sigma^2$.

Estimation of σ^2

To estimate the variance of error terms σ^2, we consider

$$\hat{\sigma}^2 = MSE = \frac{SSE}{n-2}: \text{ Mean Squared Error}$$

Note: *In fact, $\hat{\sigma}^2 = MSE$ is unbiased for σ^2 in the sense that $E(\hat{\sigma}^2) = \sigma^2$.*

Example: Elmhurst Data

$n = 50$

Suppose $SST = \sum_{i=1}^{50}(y_i - \bar{y})^2 = \$1.46\text{B}$ and $SSE = \sum_{i=1}^{50}(y_i - \hat{y}_i)^2 = \$1.07\text{B}$ for Elmhurst data. How much proportion of variability in aid received(y) does family income(x) explain?

$\Rightarrow$ skip

What is the estimated variance of the error, $\hat{\sigma}^2$?

$$\hat{\sigma}^2 = MSE = \frac{SSE}{n-2} = \frac{1.07 \times 10^9}{50-2} = 22.29 \times 10^6$$

R code

```
> g <- lm(gift_aid ~ family_income, data = elmhurst)
> summary(g)

Call:
lm(formula = gift_aid ~ family_income, data = elmhurst)

Residuals:
     Min       1Q   Median       3Q      Max
-10112.8  -3623.4   -216.1   3158.7  11570.7

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.432e+04  1.291e+03  18.831  < 2e-16 ***
family_income -4.307e-02  1.081e-02  -3.985 0.000229 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4783 on 48 degrees of freedom
Multiple R-squared:  0.2486,  Adjusted R-squared:  0.2329
F-statistic: 15.88 on 1 and 48 DF,  p-value: 0.0002289
```

$\hat{\beta}_0$

$\hat{\beta}_1$

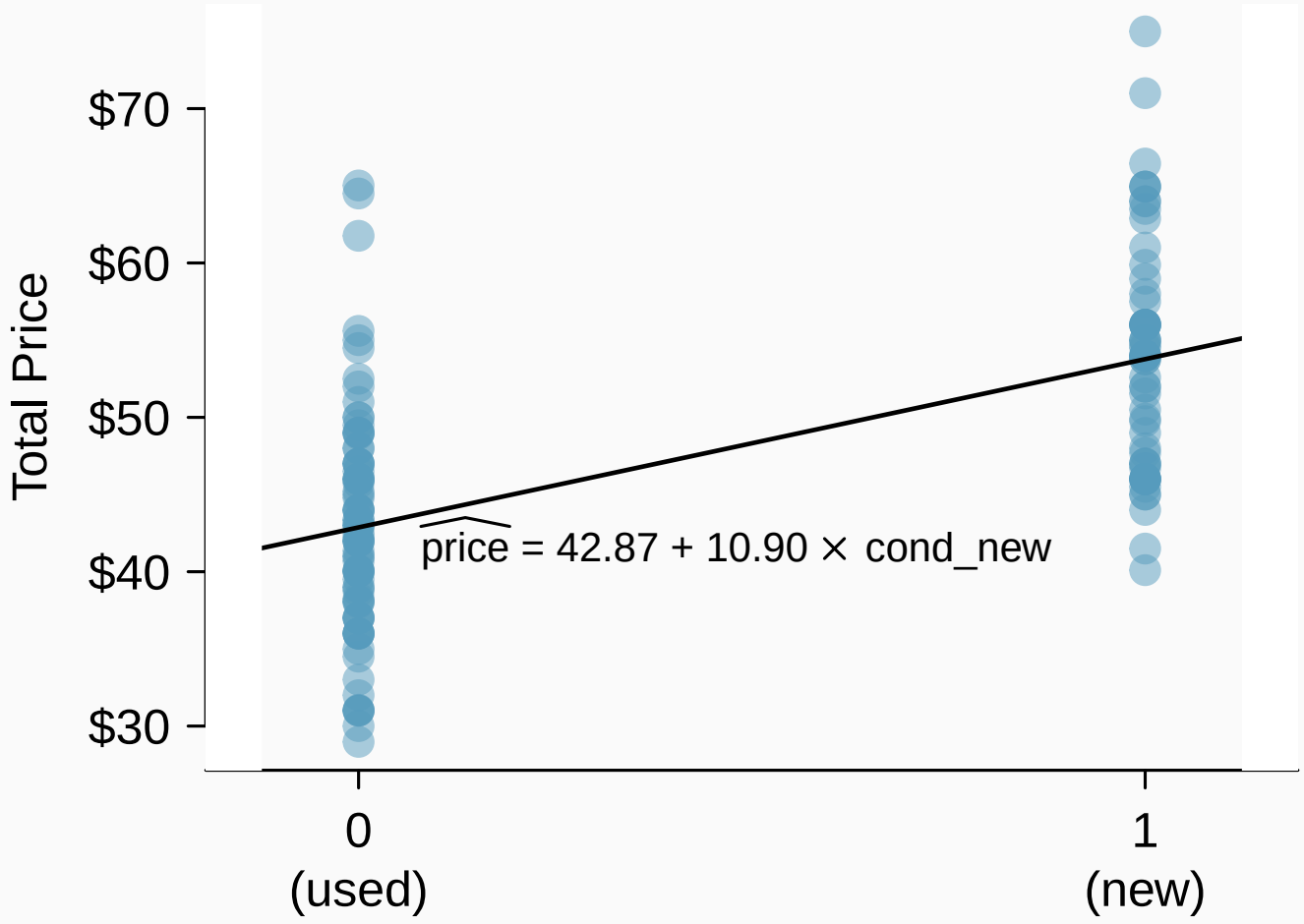
R^2

$\hat{\sigma}$

$MSE = \hat{\sigma}^2 = 4783^2$

Categorical Predictors in Regression

Example: Predicting auction prices of the game *Mario Kart* based on condition (new vs. used).



Interpreting Regression for Categorical Predictors

- The model:

$$\widehat{price} = \hat{\beta}_0 + \hat{\beta}_1 \times cond\_new$$

- Here, *cond_new* = 1 for new games, 0 for used games.
- From the regression output:

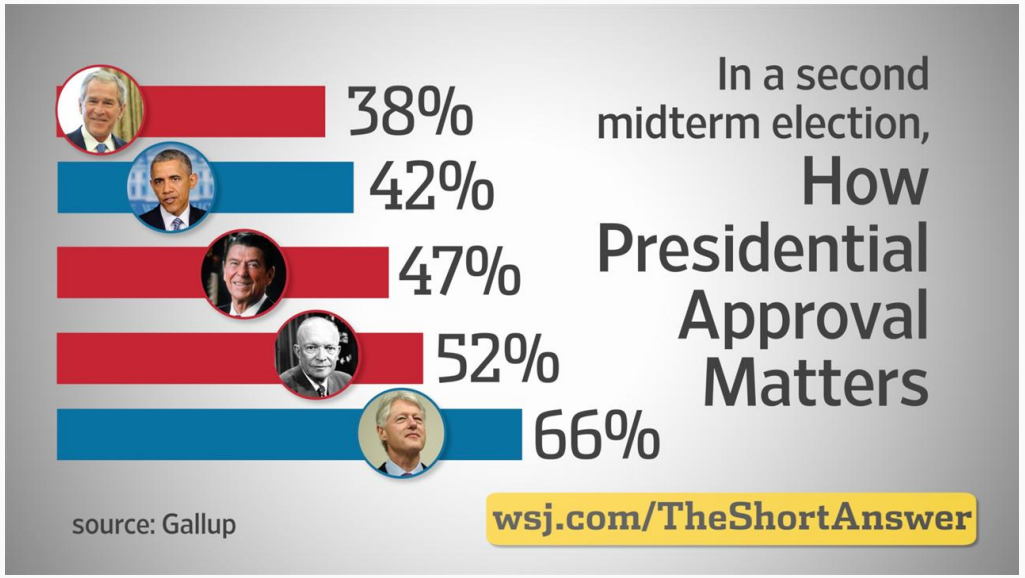
$$\widehat{price} = 42.87 + 10.90 \times cond\_new$$

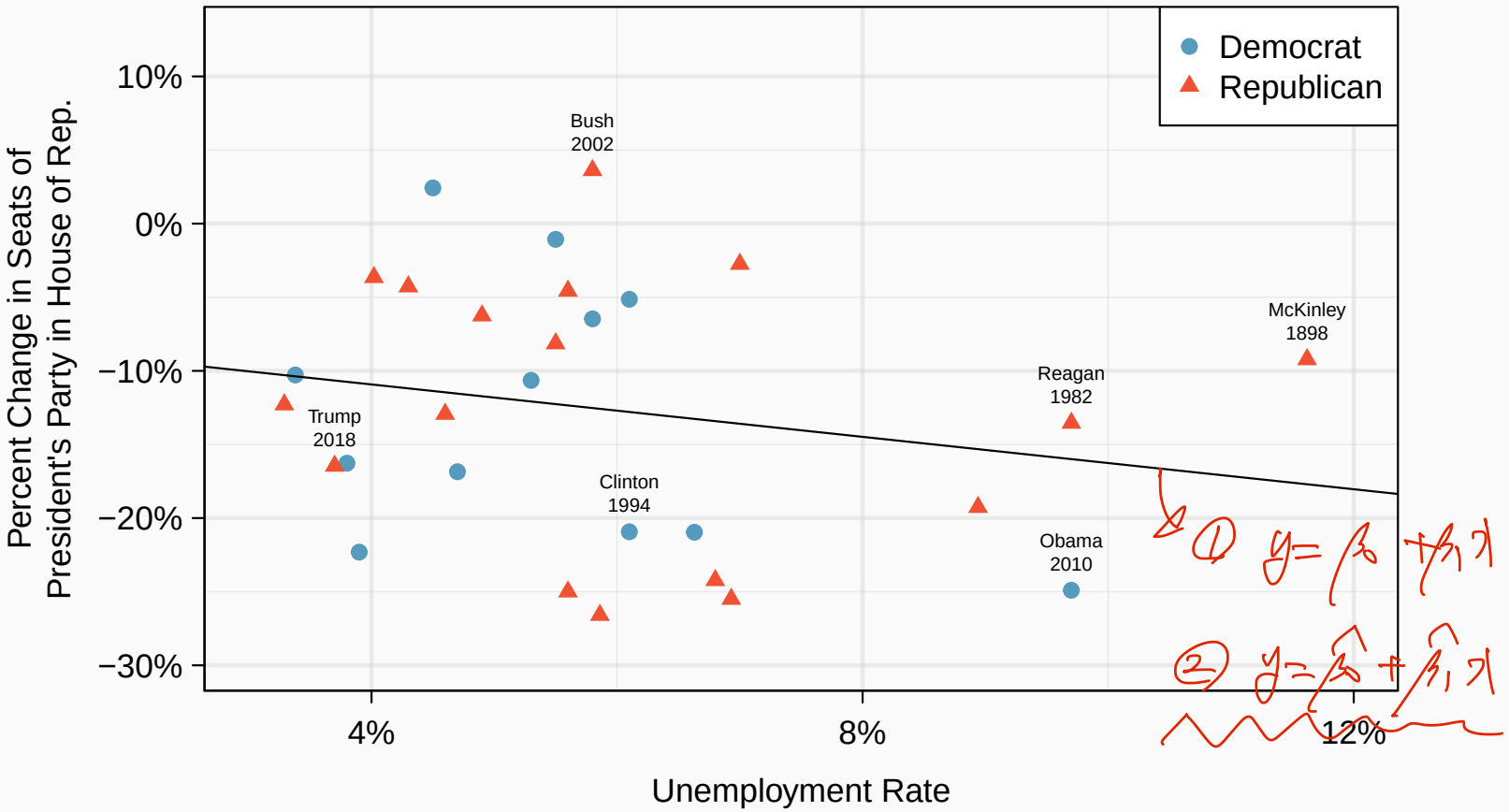
- Interpretation:
  - The intercept (42.87) is the average price of used games.
  - The slope (10.90) represents the price difference between new and used games.

Inference for linear regression

Midterm Elections and Economic Conditions

- U.S. House elections occur every two years.
- Midterm elections happen in the middle of a Presidential term.
- A political theory suggests that higher unemployment rates lead to worse outcomes for the President's party.





Data from 1934 and 1938 were excluded (unemployment rates: 21% and 18% in the Great Depression).

⇒ Outliers

Hypothesis Testing for the Slope

- We test whether unemployment(x) is a significant predictor of midterm election outcomes(y).
- Hypotheses:
  - $H_0 : \beta_1 = 0$ (No relationship)
  - $H_A : \beta_1 \neq 0$ (There is relationship)
- If we reject H_0, it suggests unemployment is a meaningful predictor.

Assumption 2: normality

Assumption 1

: Independence & constant variance

We assume *normality* on the error terms to conduct statistical inference, including hypothesis testing and confidence intervals. We consider a linear regression model:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \cdots, n,$$

where β_0 is the intercept, β_1 is the slope, and ε_i's are identically and independently distributed as $\varepsilon_i \sim N(0, \sigma^2)$.

$E(\varepsilon_i) = 0$, $Var(\varepsilon_i) = \sigma^2$

$SD(\varepsilon_i) = \sigma$

Testing for the slope β_1

Consider the following hypothesis test:

$$H_0 : \beta_1 = 0, \qquad H_A : \beta_1 \neq 0.$$

Test statistic for the slope β_1

The test statistic and its null distribution are

$$T = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)} = \frac{\hat{\beta}_1}{\hat{\sigma}/\sqrt{S_{xx}}} \overset{H_0}{\sim} t(n-2),$$

where $\hat{\beta}_1$ is the least-square estimate of β_1, $\hat{\sigma} = \sqrt{MSE}$ is the estimated standard deviation of error, and $S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$.

Note: Let $x_1, x_2, \cdots, x_n$ be fixed constants. Note that

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{1}{S_{xx}} \sum_{i=1}^{n} (x_i - \bar{x}) y_i$$

is normally distributed since $\hat{\beta}_1$ is a linear combination of $y_1, \cdots, y_n$. Also,

$$E(\hat{\beta}_1) = \frac{1}{S_{xx}} \sum_{i=1}^{n} (x_i - \bar{x}) E(y_i) = \frac{1}{S_{xx}} \left[\sum_{i=1}^{n} (x_i - \bar{x}) \beta_0 + \sum_{i=1}^{n} (x_i - \bar{x}) x_i \beta_1 \right] = \beta_1$$

$$Var(\hat{\beta}_1) = \frac{1}{S_{xx}^2} \sum_{i=1}^{n} (x_i - \bar{x})^2 Var(y_i) = \frac{1}{S_{xx}^2} S_{xx} \sigma^2 = \frac{\sigma^2}{S_{xx}}$$

It follows that $Z := \dfrac{\hat{\beta}_1 - \beta_1}{\sigma / \sqrt{S_{xx}}} \sim N(0,1)$. Furthermore, it is known that

$$V := \frac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-2)$$

and V is independent of Z. Therefore,

$$T = \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma} / \sqrt{S_{xx}}} = \underbrace{\frac{\hat{\beta}_1 - \beta_1}{\sigma / \sqrt{S_{xx}}}}_{Z} \left(\underbrace{\frac{(n-2)\hat{\sigma}^2}{\sigma^2}}_{V} \cdot \frac{1}{n-2} \right)^{-1/2} = \frac{Z}{\sqrt{V/(n-2)}} \sim t(n-2).$$

Back to the example

Using two-sided test, hypotheses are:

H_0 : $\beta_1 = 0$ (Unemployment is not predictive of elections.)

H_A : $\beta_1 \neq 0$ (Unemployment is predictive of election results.)

Checking conditions:

- use residual plots(discussed later).

Computing the Test Statistic

- Sample statistics:

$$n = 50, \quad S_{xx} = 113.9, \quad S_{xy} = -101.4, \quad \hat{\sigma}^2 = 79.4$$

- Compute the estimated slope, $\hat{\beta}_1$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{-101.4}{113.9} = -0.8903$$

- Compute standard error of $\hat{\beta}_1$:

$$SE = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} = \sqrt{\frac{79.4}{113.9}} = 0.8349$$

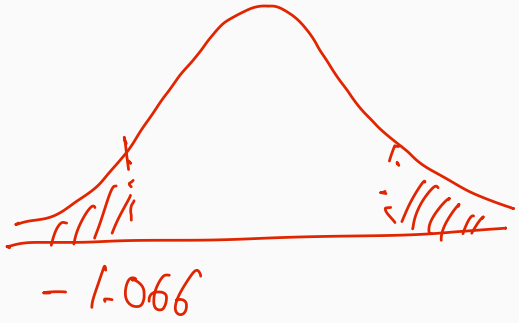
- Compute T-score:

$$T = \frac{\hat{\beta}_1 - 0}{SE} = \frac{-0.8903}{0.8349} = -1.066$$

Finding the P-value

P-value is $2 \times P(T < -1.066)$, $T \sim t(27)$

$= 2 \times 0.1458 = 0.2918$




```
> pt(-1.066, df = ____27____, lower.tail = F) T
[1] 0.1458782
```

- P-value is larger than $\alpha = 0.05$, so we (reject / don't reject) H_0.
- There is (significant / no significant) evidence to conclude that unemployment and midterm election losses are related.

```
> g <- lm(house_change ~ unemp, midterms_house)
> summary(g)

Call:
lm(formula = house_change ~ unemp, data = midterms_house)

Residuals:
     Min       1Q   Median       3Q      Max
-14.0124  -7.6989   0.0913   7.2974  16.1447

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -7.3644     5.1553  -1.429    0.165
unemp        -0.8897     0.8350  -1.066    0.296

Residual standard error: 8.913 on 27 degrees of freedom
Multiple R-squared:  0.04035, Adjusted R-squared:  0.004812
F-statistic: 1.135 on 1 and 27 DF,  p-value: 0.2961
```

β_0

β_1

$F = 8.913$

$R^2 = 0.04035$

Table below shows R output from fitting the least squares regression line in Elmhurst College data. Use this output to formally test the following hypotheses.

$$H_0 : \beta_1 = 0, \qquad H_A : \beta_1 \neq 0,$$

$2P(T < -3.9846)$

$H_A: \beta_1 < 0$, $P(T < -3.9846)$

```
> g <- lm(gift_aid ~ family_income, data = elmhurst)
> summary(g)$coefficients
```

```
                  Estimate Std. Error t value   Pr(>|t|)
(Intercept)    2.4319e+04 1.2915e+03 18.8310 8.2810e-24
family_income -4.3072e-02 1.0809e-02 -3.9846 2.2887e-04
```

p-Value $= 2.2887 \times 10^{-4} < 0.05$, Reject H_0.

Gift aid and family income are related.

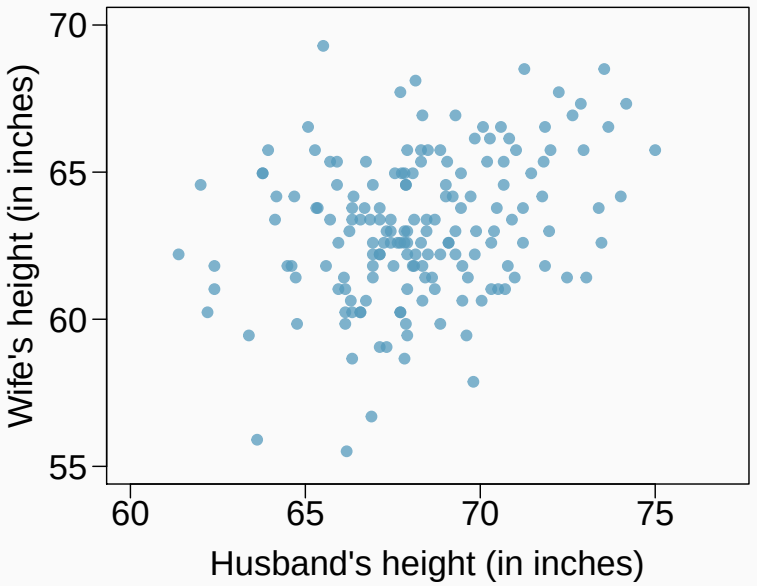
Practice ⇒ Homework

(Exercise 8.33*) A survey agency collected data on a random sample of 170 married couples in Britain, recording heights of husbands and wives. Let the husband's height be explanatory variable x and the wive's height be a response variable y, and consider a linear model, $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ for $i = 1, \cdots, n$, where $\varepsilon_i \sim N(0, \sigma^2)$. Conduct the following hypothesis test: $H_0 : \beta_1 = 0$, ~~H_0~~ $H_A : \beta_1 > 0$.

1. Find the test statistic and its null distribution.

Practice

Summary statistics	
n	170
$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$	1157.40
$S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2$	1010.43
$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$	331.37
$\hat{\sigma} = \sqrt{MSE}$	2.334



2. Compute the slope($\hat{\beta}_1$) and the observed test statistic.

3. Compute the p-value and complete the hypothesis test.

Confidence interval for the slope

A t-confidence interval for the slope

A $100(1-\alpha)\%$ confidence interval for the slope β_1 is

$$\text{point estimate} \pm t_{\alpha/2}(df) \times SE = \hat{\beta}_1 \pm t_{\alpha/2}(df) \times \frac{\hat{\sigma}}{\sqrt{S_{xx}}},$$

where $t_{\alpha/2}(df)$ is the $\alpha/2$-th upper quantile of the t-distribtuion with degree of freedom $df = n - 2$.

Example: A 95% confidence interval for β_1 in Elmhurst data is

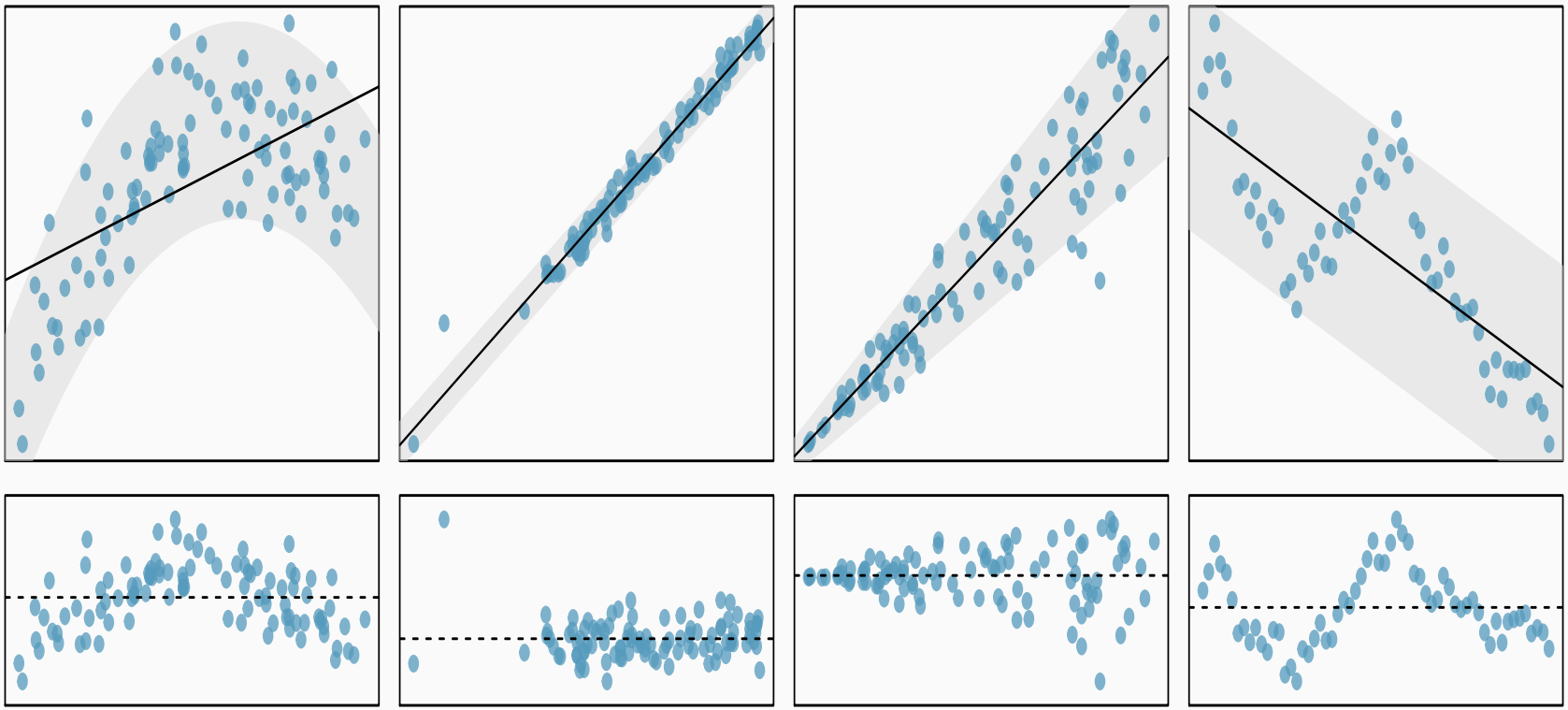
$$-0.0431 \pm 2.01 \times 0.0108 = (-0.0648, -0.0214).$$

```
> qt(p = 0.025, df = 48, lower.tail = F)
[1] 2.010635
```

Conditions for the least squares line

1. **Linearity**
   - The data should show a linear trend.
2. **Nearly normal residuals**
   - Generally, the residuals must be nearly normal.
3. **Constant variability**
   - The variability of residuals should remain roughly constant.
4. **Independent observations**
   - Successive observations should not be highly correlated, such as time series data.

What conditions are the following linear models obviously violating?



Note: *Consider a linear regression model:*

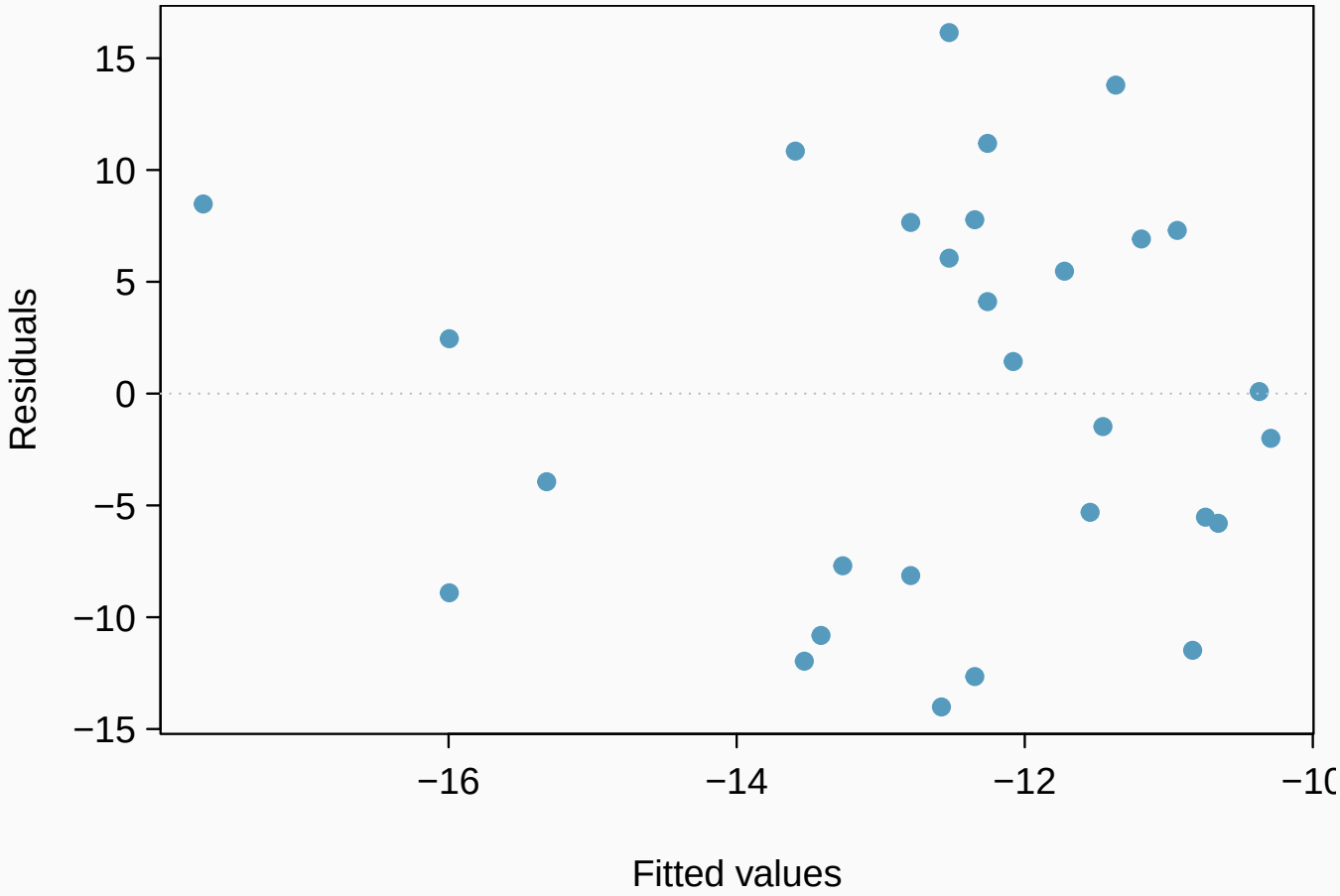
$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \cdots, n,$$

where ε_i *follows the standard normal distribution independently.*
Each condition can be formulated as follows:

1. ***Linearity***: $E(Y_i \mid x_i) = \beta_0 + \beta_1 x_i$.
2. ***Nearly normal residuals***: ε_i*'s are from normal distribution.*
3. ***Constant variability***: $Var(\varepsilon_i) = \sigma^2$.
4. ***Independent observations***: ε_i*'s are mutually independent.*

Practice

Do the midterm election data meet the conditions required for fitting a least squares line?



Exercises

Exercises in OpenIntro Statistics 4th ed.

- Fitting a line, residuals, and correlation: Exercise 8.15
- Least squares regression: Exercise 8.23 (a), (c), (d), (e), (f)
- Inference for linear regression: Exercise 8.33 (a), (b), (d)