Chapter 9: Linear regression

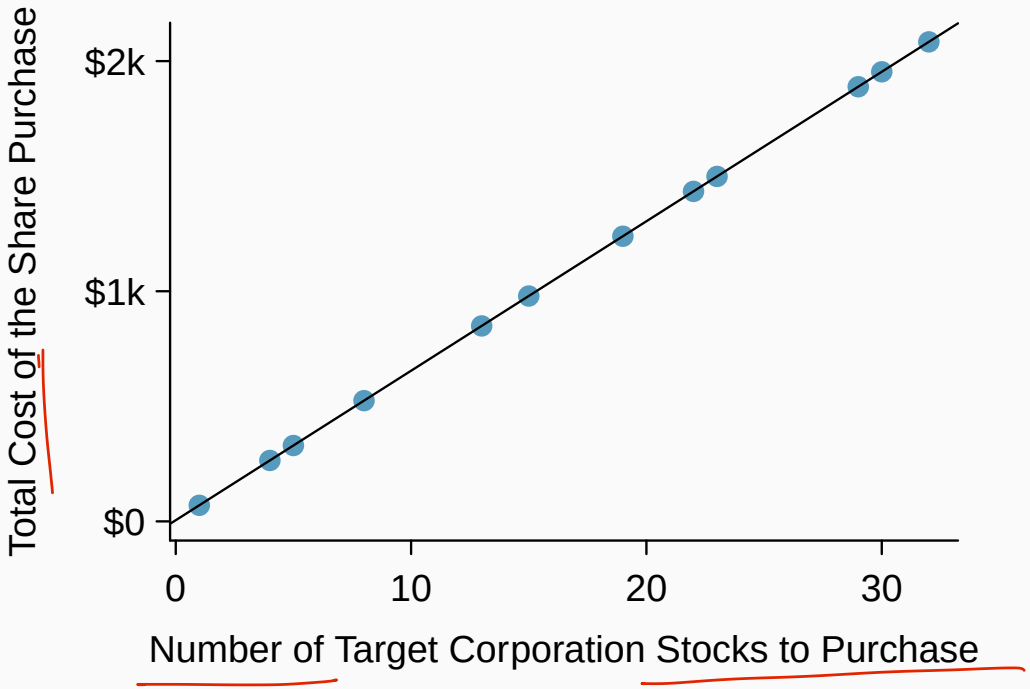
Yonghyun Kwon

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]

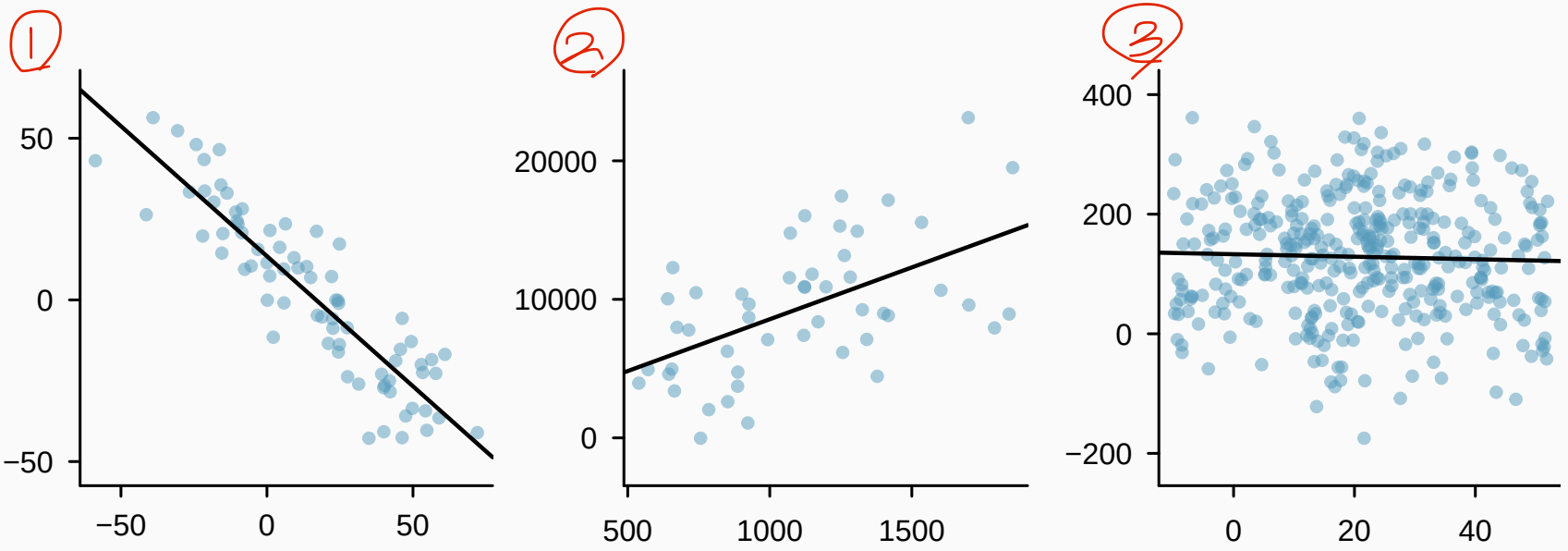
$$y = \beta_0 + \beta_1 x + \varepsilon,$$

[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

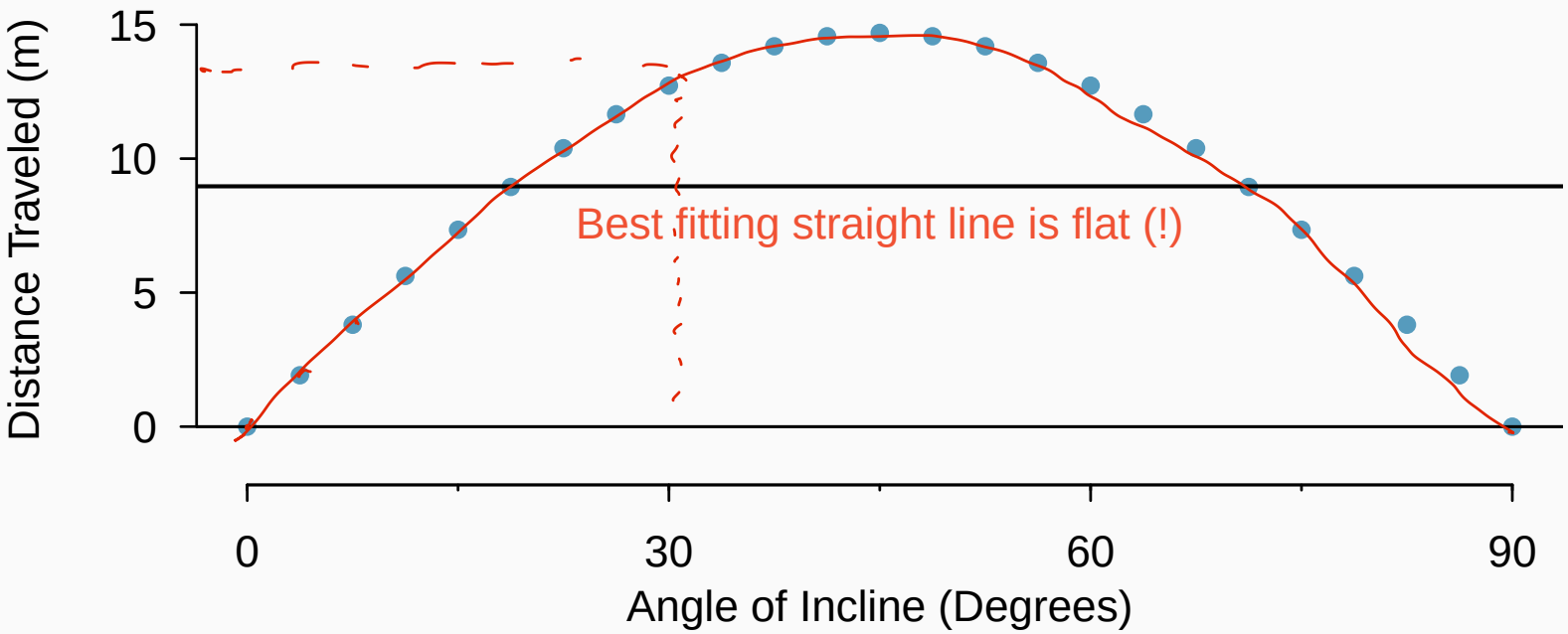
Data usually form a cloud of points around a trend.



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.

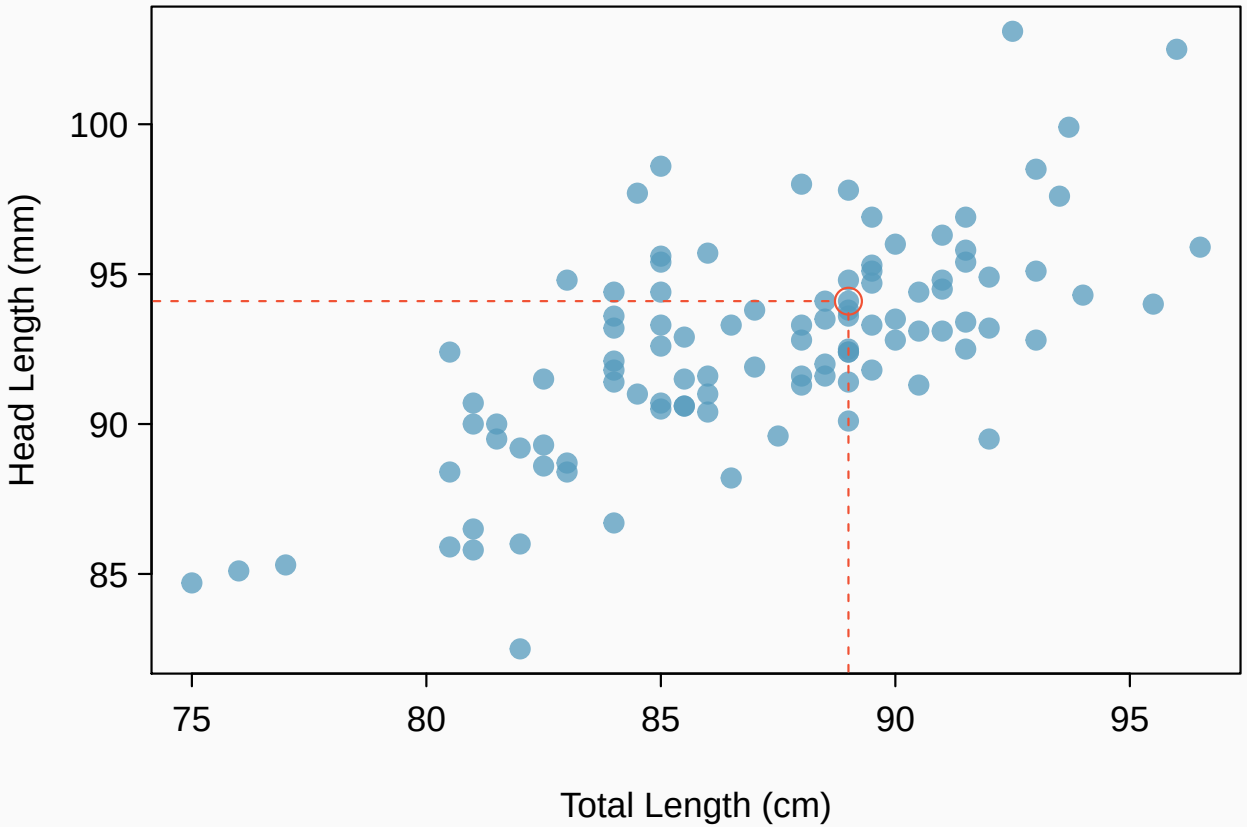


Example: brushtail possums

- 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

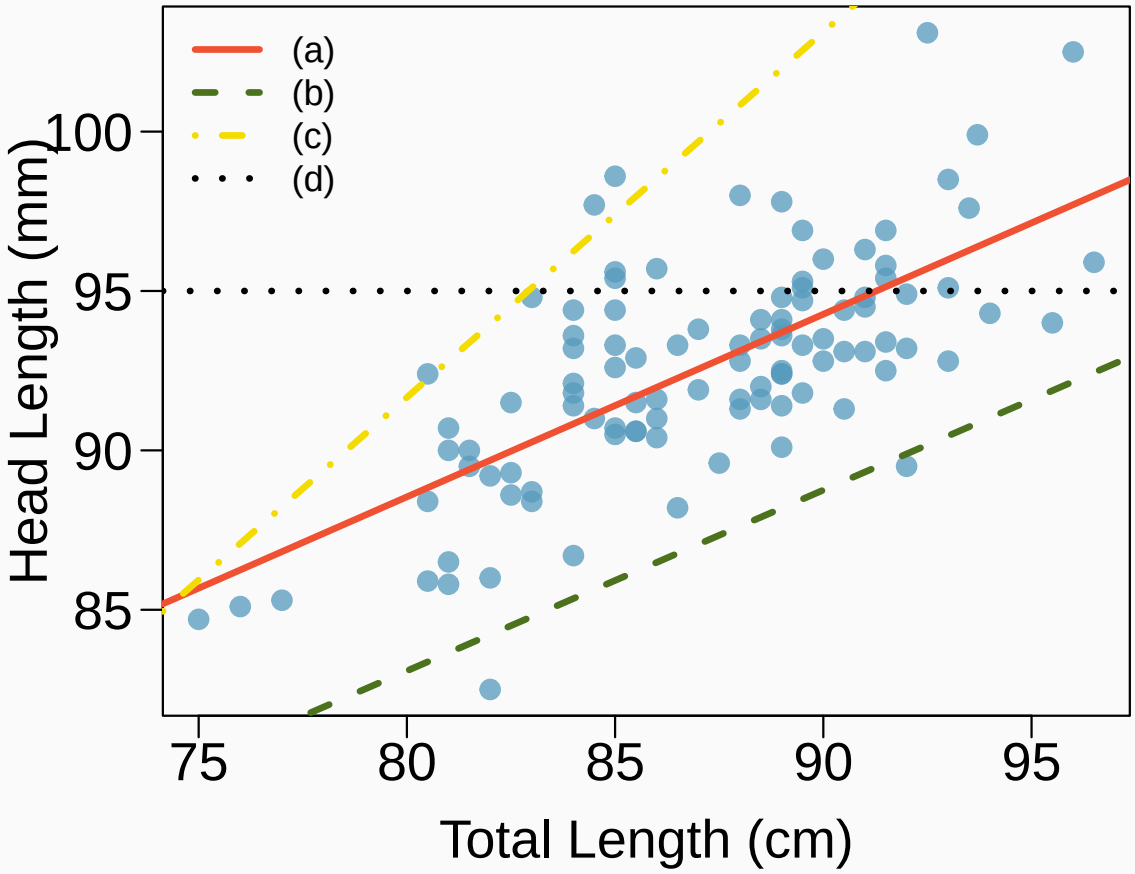
- Head length and total length are positively associated.
- While not perfectly linear, a straight-line model may be useful.



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

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

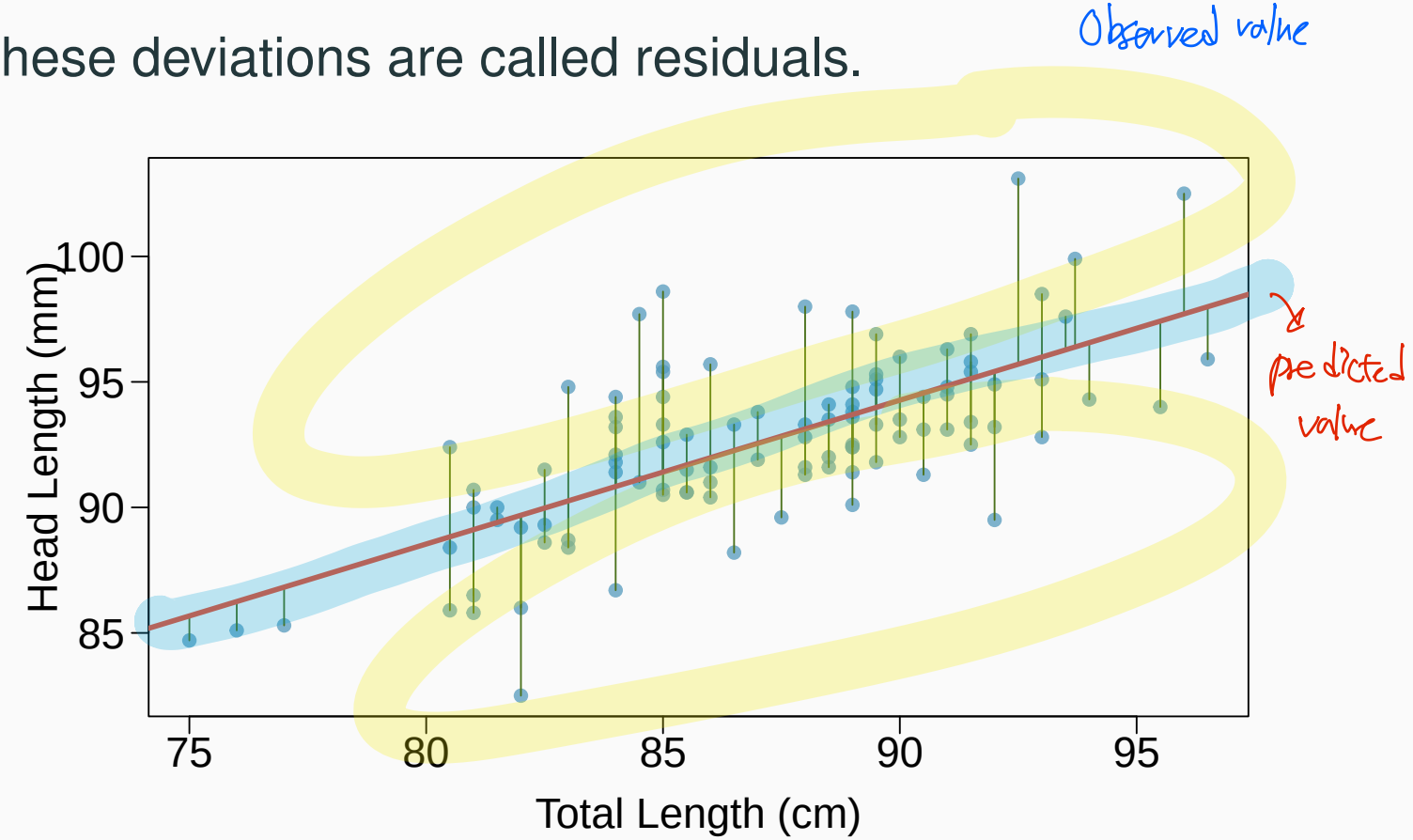
- Example: For a possum with a total length of 80 cm:

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

- 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

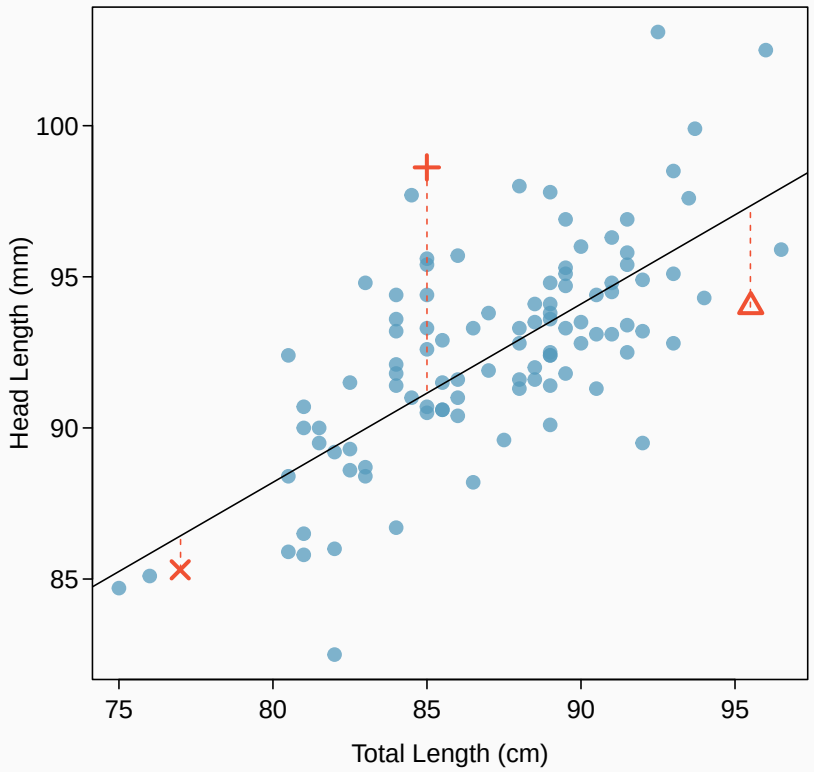
The residual of the i-th observation (x_i, y_i) is the difference between the observed (y_i) and predicted $(\hat{y}_i)$:

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

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

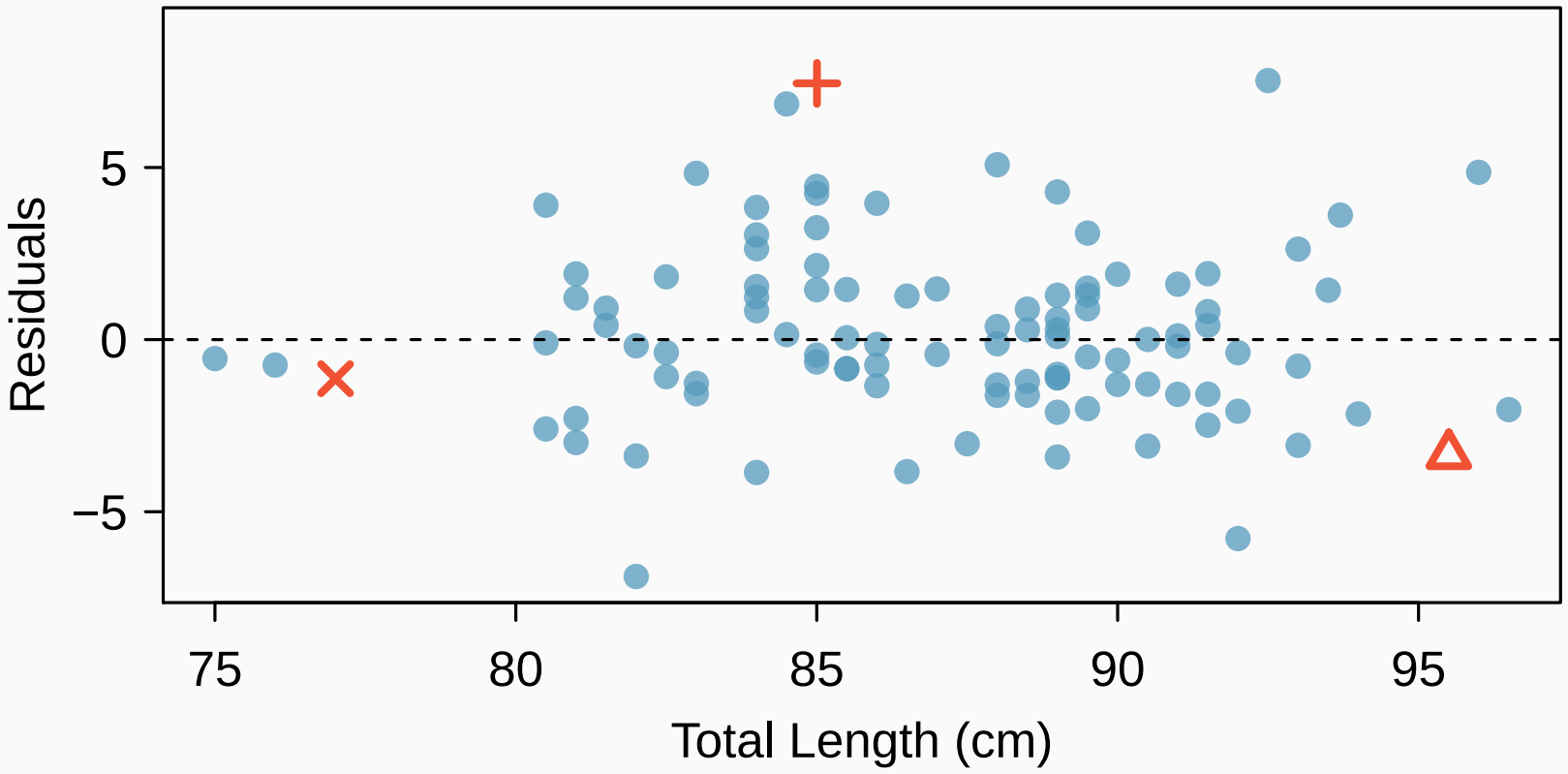
$$\hat{y}_i = 41 + 0.59 \times 85$$
$$= 91.15$$

The residual is given by

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

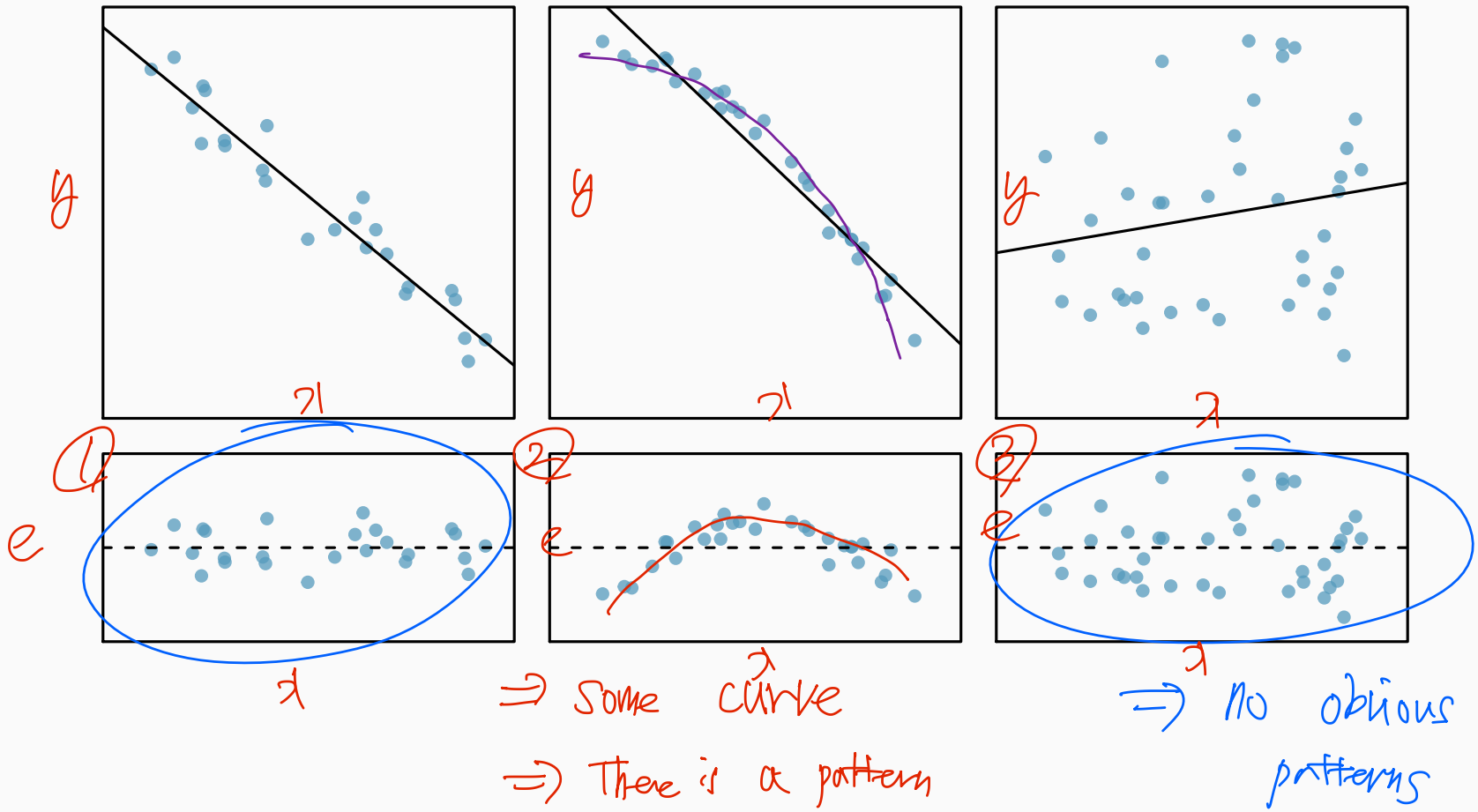
Residual Plot

- *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.



Sample correlation coefficient

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

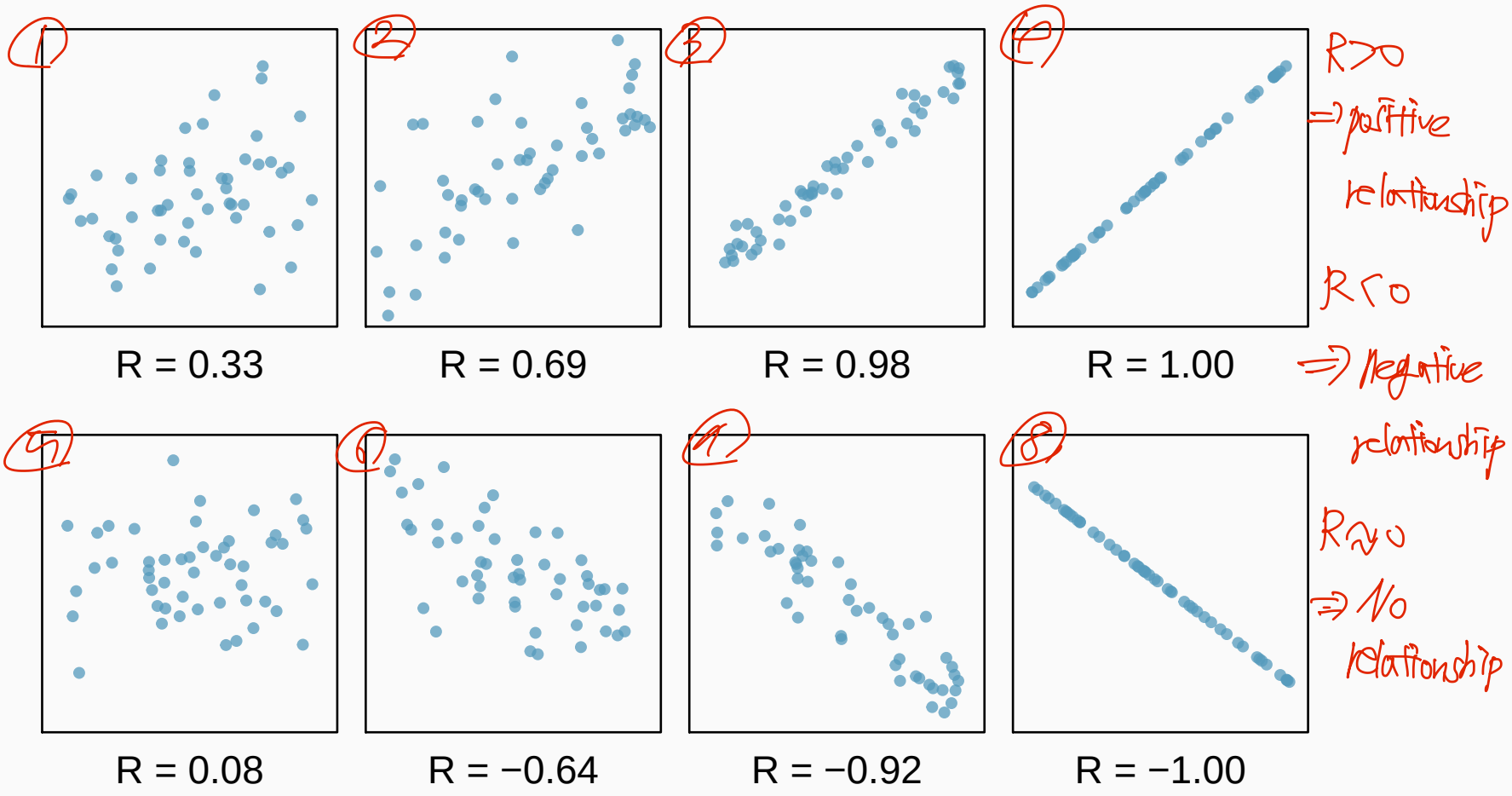
$\propto$ Sample variances

$\propto$ Sample covariance

- *Sample correlation coefficient* 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?

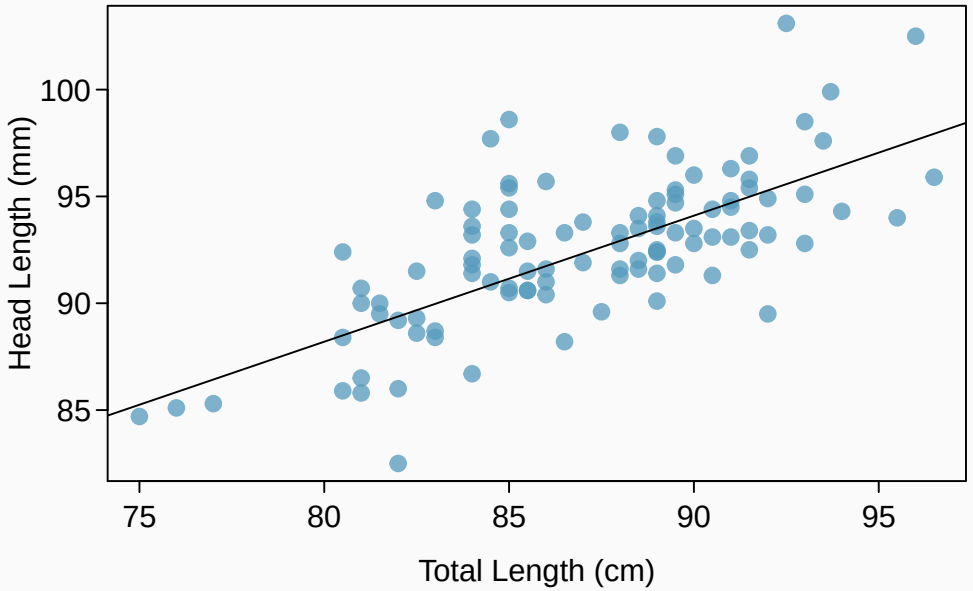
(a) -0.75

(b) 0.69

(c) -0.1

(d) 0.02

(e) -1.5



Does correlation depend on the unit of measure?

Recall: Correlation coefficient of random variables X and Y

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 for the population correlation coefficient ρ

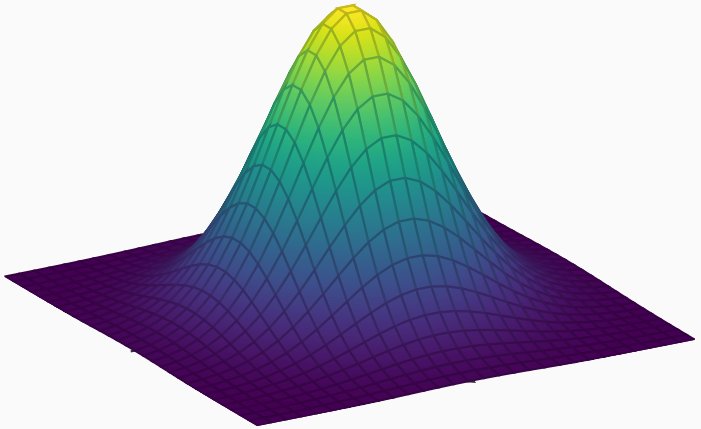
If $(X_1, Y_1), \cdots, (X_n, Y_n)$ are independent observations from a bivariate normal distribution, the test statistic and its null distribution are

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

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

Note: Since correlation coefficient is invariant of scaling and translation, without loss of generality, assume that $(X_i, Y_i), i = 1, \cdots, n$ are random samples from a bivariate normal distribution with mean 0 and standard deviation 1. Notice that under $H_0 : \rho = 0$, and X_i and Y_i are independent because they follow the bivariate normal distribution. Conditioning on $X_1 = x_1, \cdots, X_n = x_n$,

$$Z := \frac{S_{xY}}{S_{xx}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})Y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \sim N(0,1).$$

Furthermore, conditioning on $X_1 = x_1, \cdots, X_n = x_n$, it is known that

$$V := S_{YY} - \frac{S_{xY}^2}{S_{xx}} \sim \chi^2(n-2),$$

and V is independent of Z. Thus, under $H_0 : \rho = 0$, conditioning on $x_1, \cdots, x_n$,

$$T := \frac{R}{\sqrt{1-R^2}/\sqrt{n-2}} = \underbrace{\frac{S_{xy}}{S_{xx}}}_{Z} \left(\underbrace{\left(S_{YY} - \frac{S_{xY}^2}{S_{xx}} \right)}_{V} \cdot \frac{1}{n-2} \right)^{-1/2} = \frac{Z}{\sqrt{V/(n-2)}} \sim t(n-2).$$

Since conditional distribution does not depend on x's, $T = \dfrac{R\sqrt{n-2}}{\sqrt{1-R^2}} \sim t(n-2)$.

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:

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

- Compute the observed T-score:

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

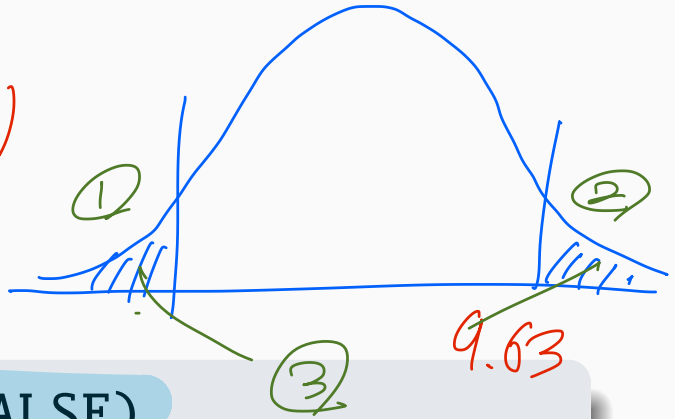
Observed test statistic:

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

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

R output

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

 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

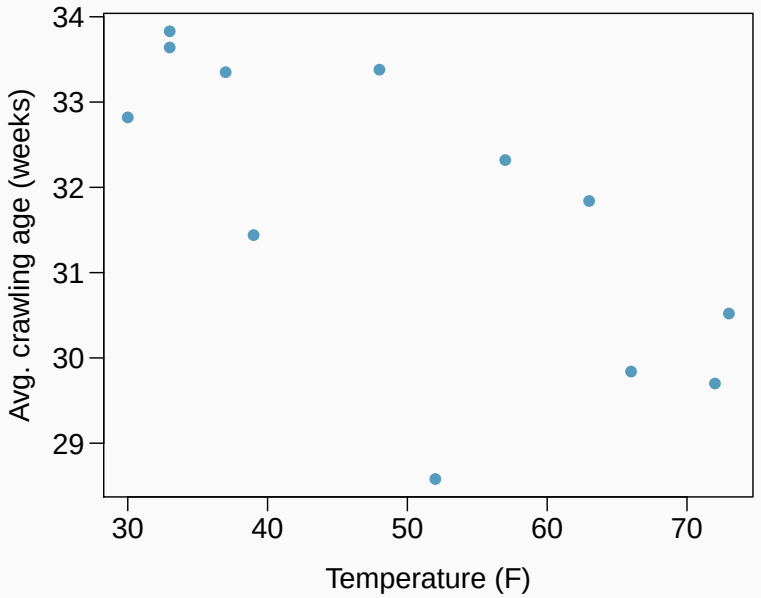
(Exercise 8.12*) A study recorded data on 12 groups of infants born throughout the year to investigate whether crawling age(x) is associated with environmental temperature(y). For each group, researchers measured the average crawling age (in weeks) and the average temperature (in °F) when the babies 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



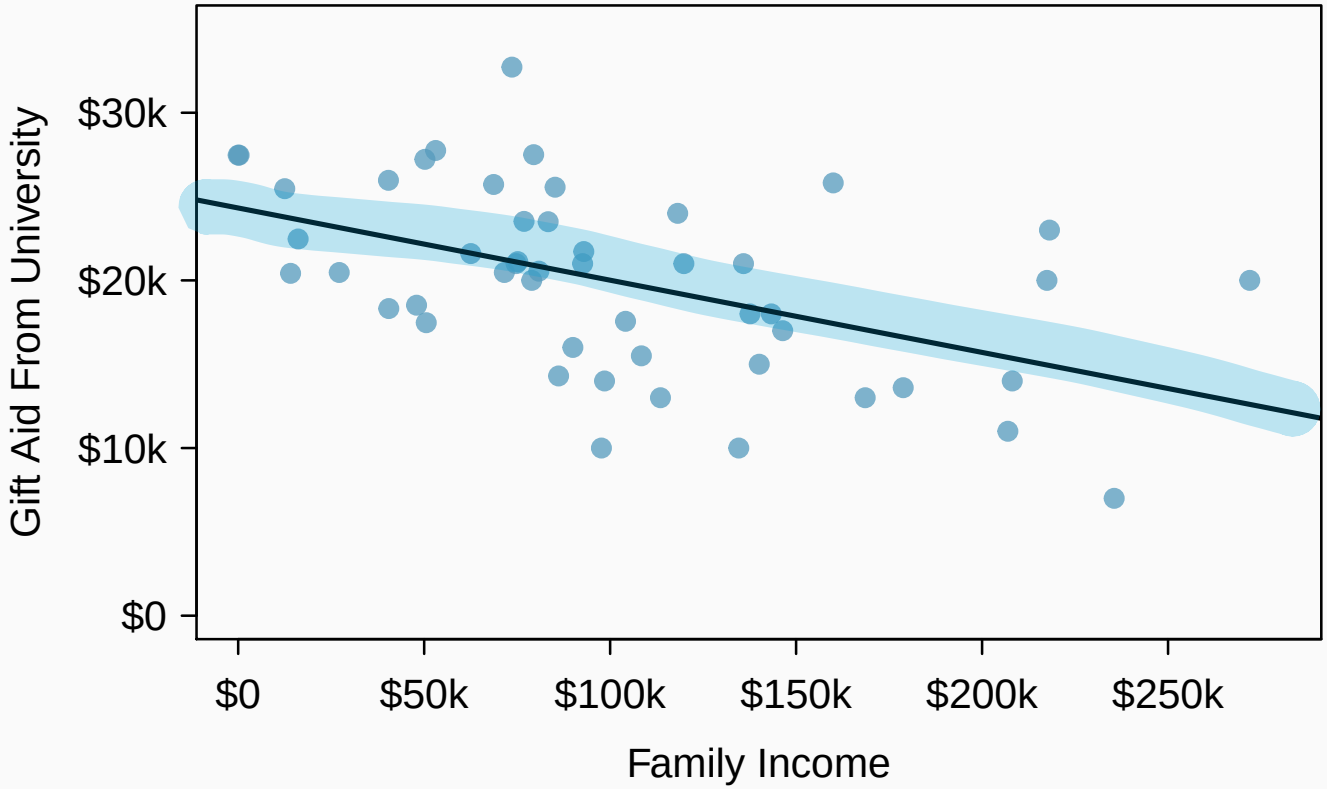
3. Compute the correlation coefficient and the observed test statistic.

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

Fitting a line by least squares

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 errors.
  One way is to minimize the sum of squared errors:

$$\varepsilon_1^2 + \varepsilon_2^2 + \cdots + \varepsilon_n^2,$$

  which is called *least square regression*.
- Squaring errors 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
```

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

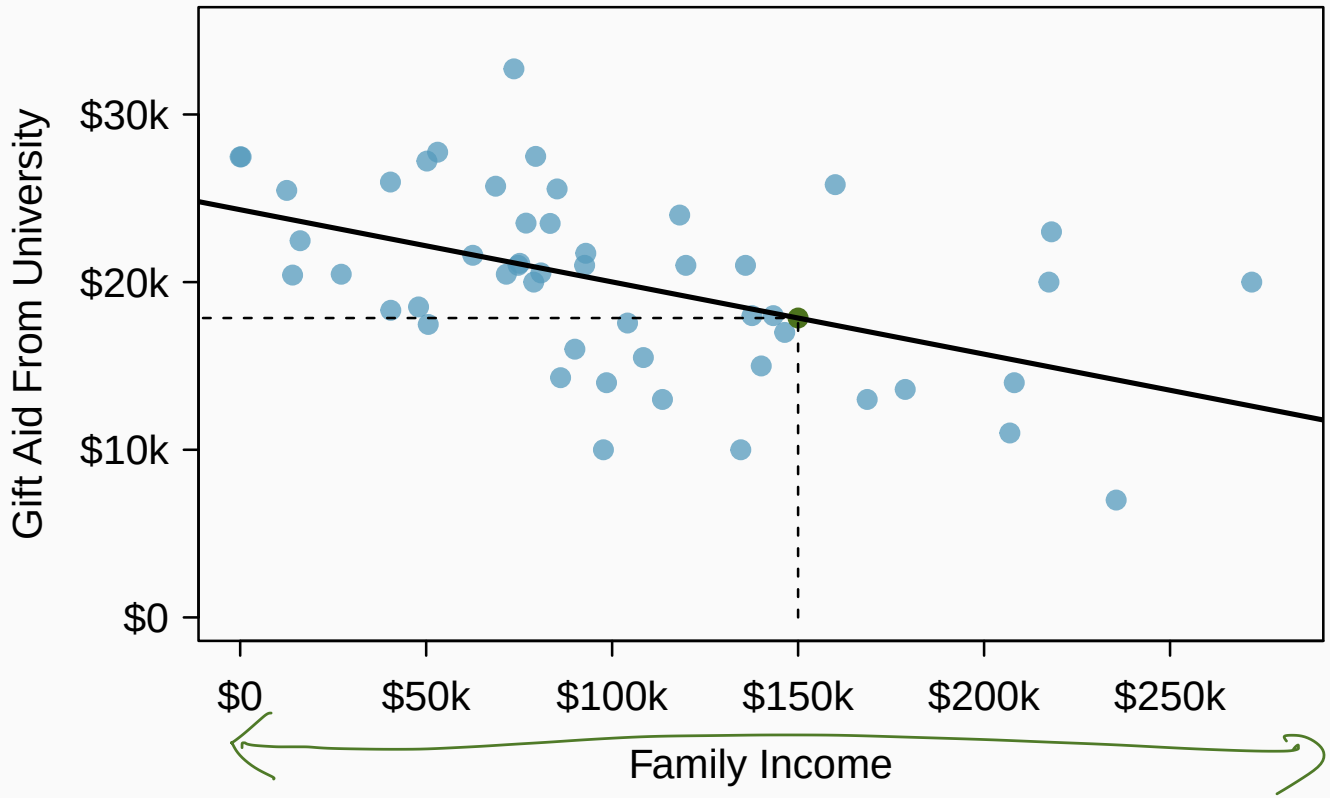
$-4.3072 \times 10^{-2} = \beta_1$

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

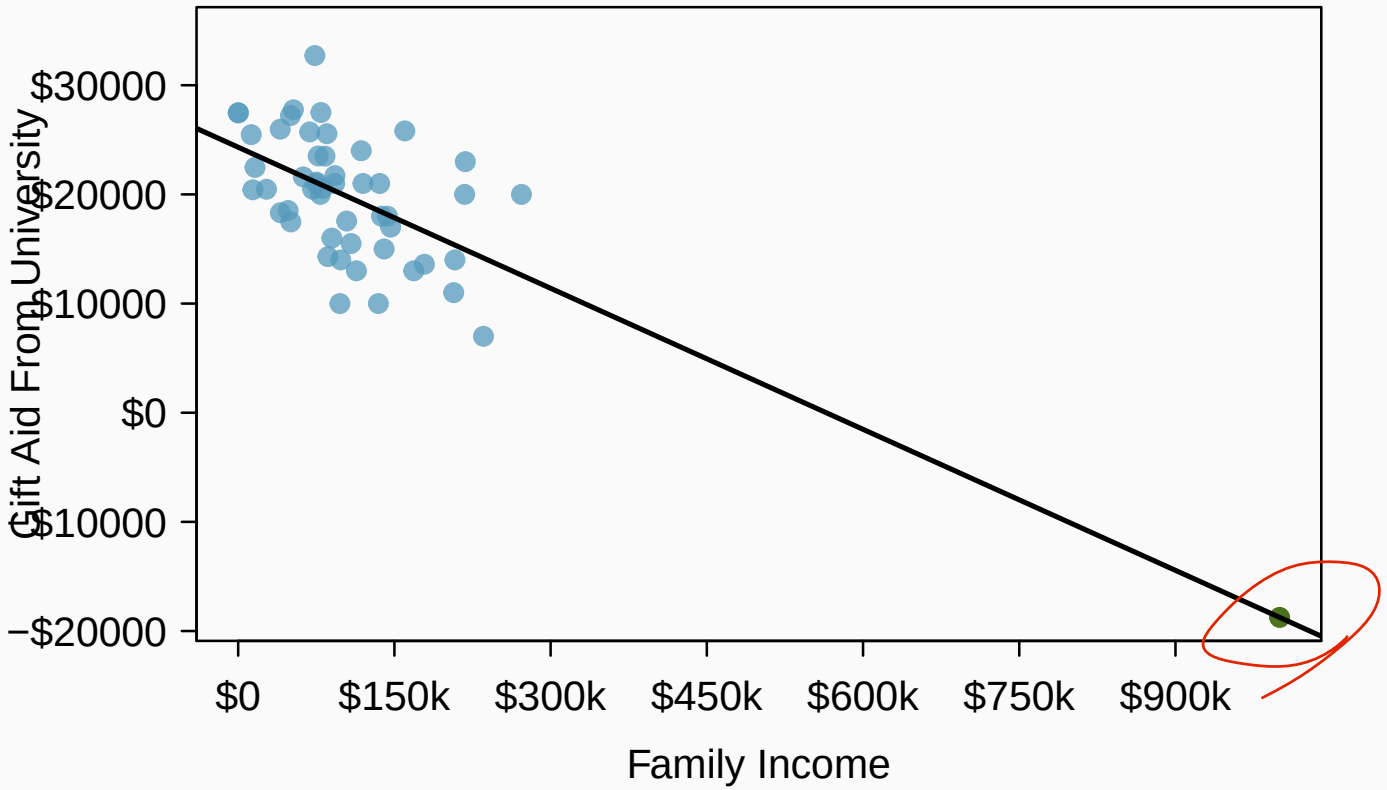
Using the linear model to predict the expected value of y for a given value of x is called *prediction*, and denoted as $\hat{y}$.



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

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

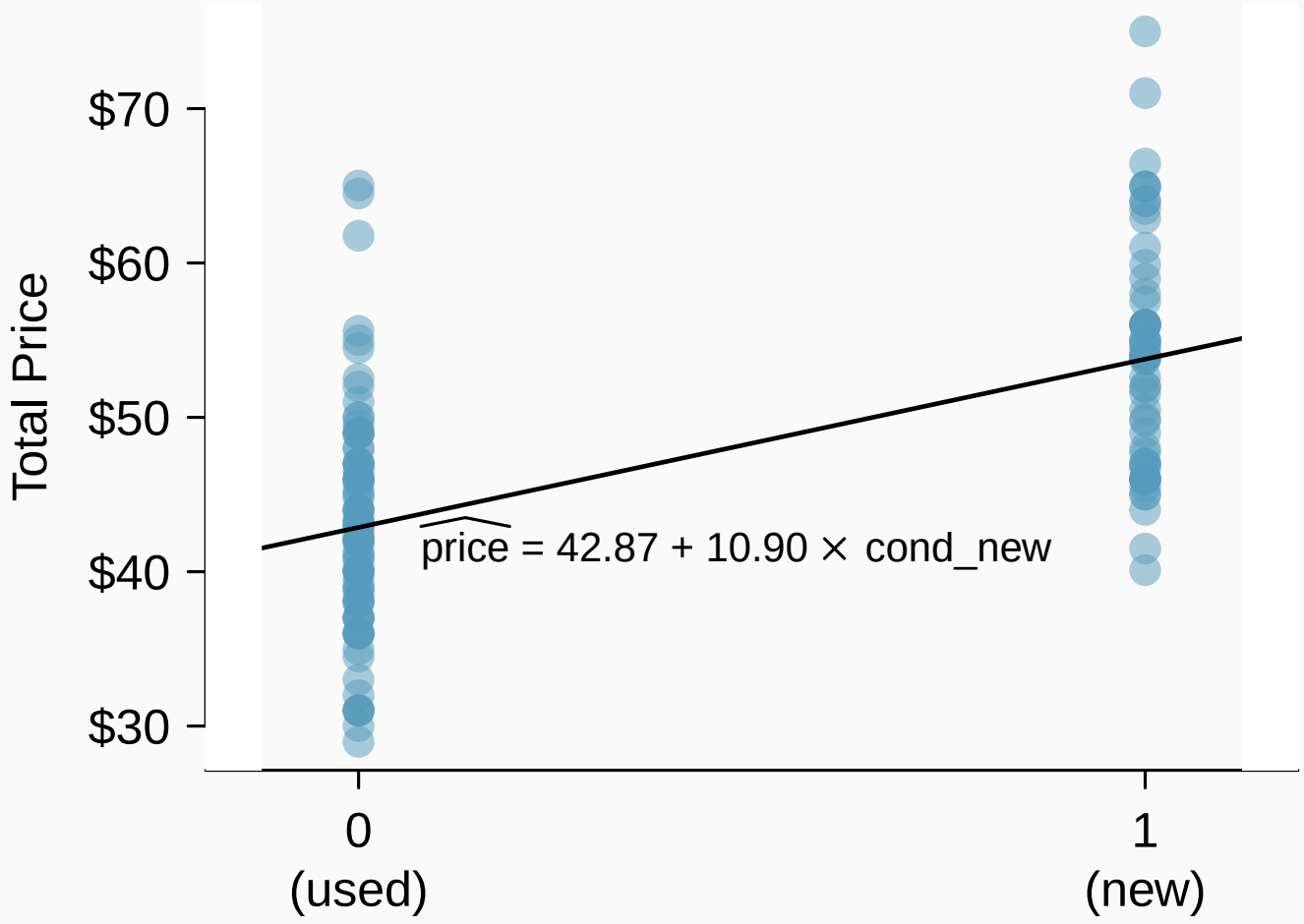
```
> g <- lm(gift_aid ~ family_income, data = elmhurst)
> summary(g)
Call:
lm(formula = gift_aid ~ family_income, data = elmhurst)

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:

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

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

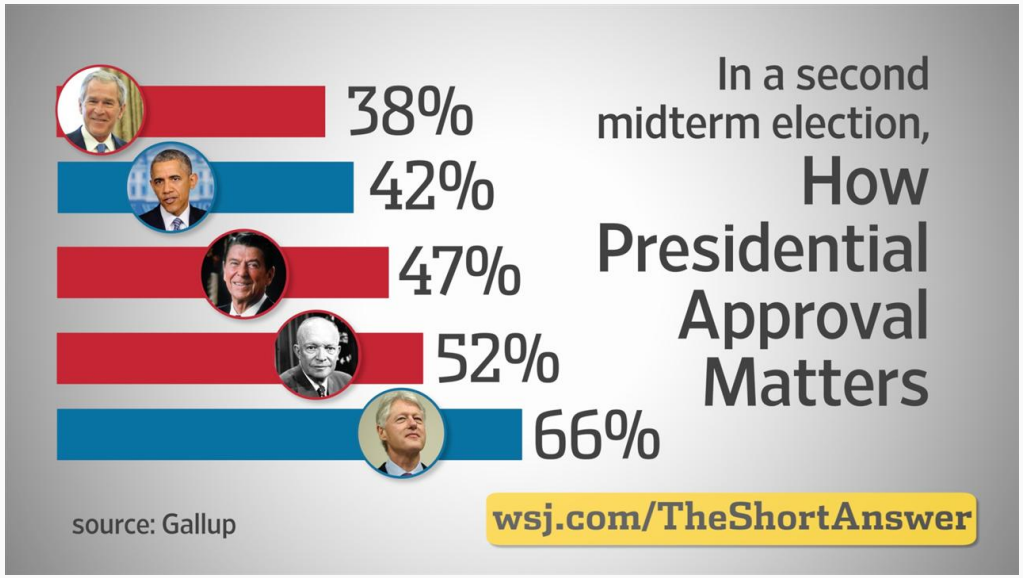
$$\widehat{price} = 42.87 + 10.90 \times cond\_new$$

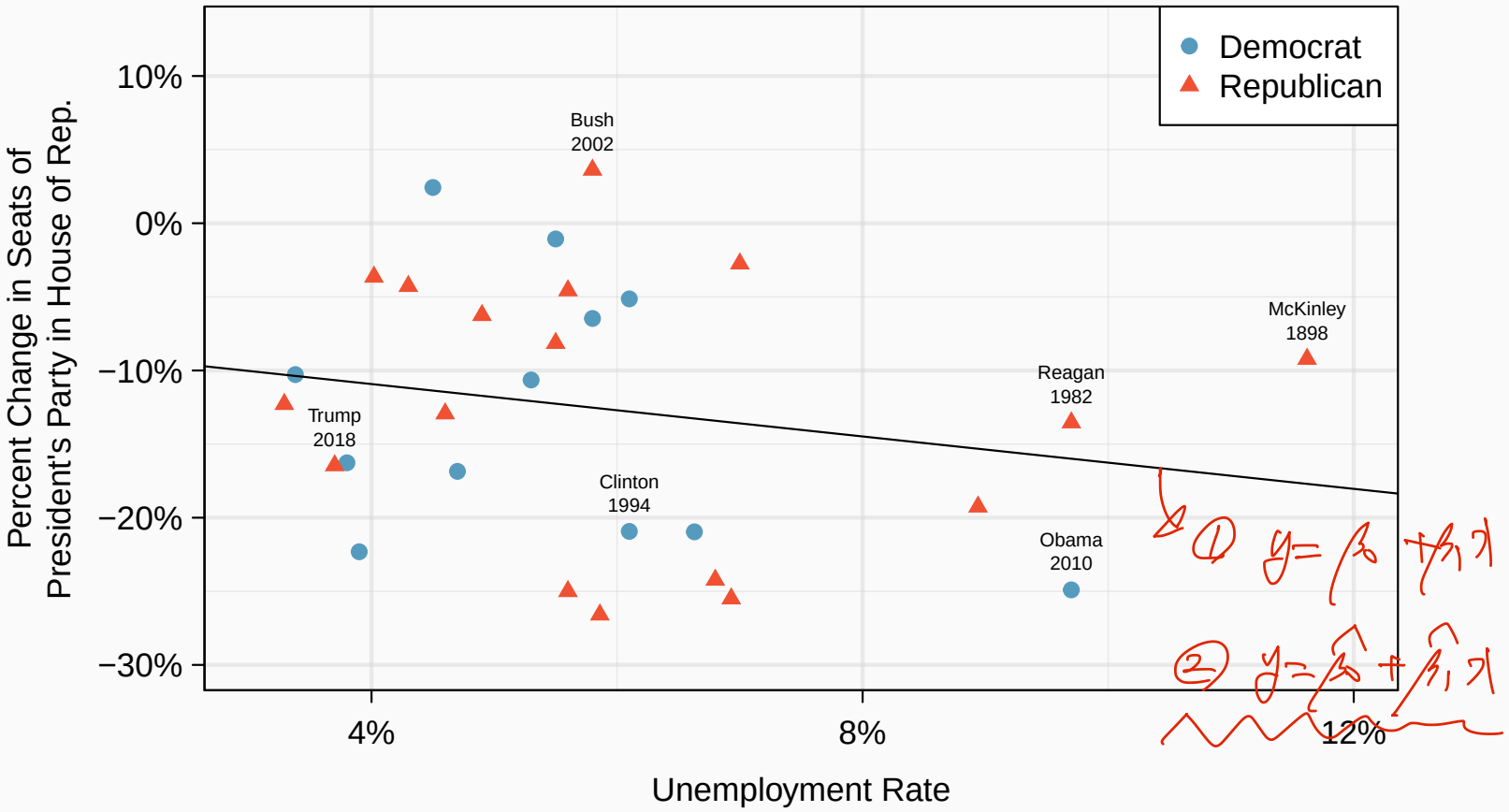
- Interpretation:
  - The intercept (42.87) is the average price of used games.
  - The slope (10.90) represents the price difference between new and used games.

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:
  - $H_0 : \beta_1 = 0$
  - $H_A : \beta_1 \neq 0$
- 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:

$$n = 29, \quad S_{xx} = 113.9, \quad S_{xy} = -101.4, \quad \hat{\sigma}^2 = 79.4$$

- 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 $\hat{\beta}_1$:

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

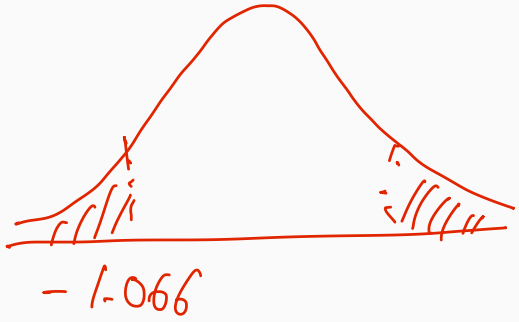
- Compute the observed T-score:

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

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 = TRUE)
[1] 0.1479318
```

- P-value is larger than $\alpha = 0.05$, so we (reject / don't reject) H_0.
- Unemployment (is / is not) a significant predictor variable for midterm election outcomes.

```
> 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

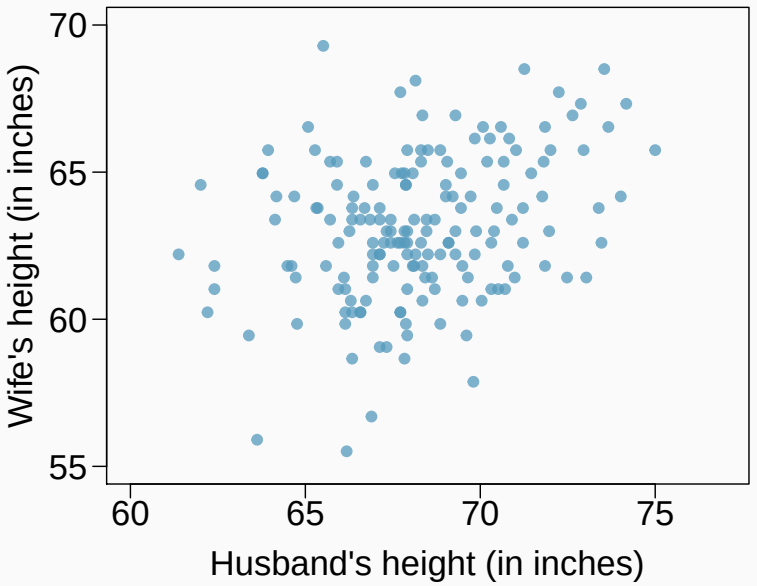
⇒ Homework

(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_A~~ $H_A : \beta_1 > 0$.

1. Find the test statistic and its null distribution(Use a **one-sided** test).

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

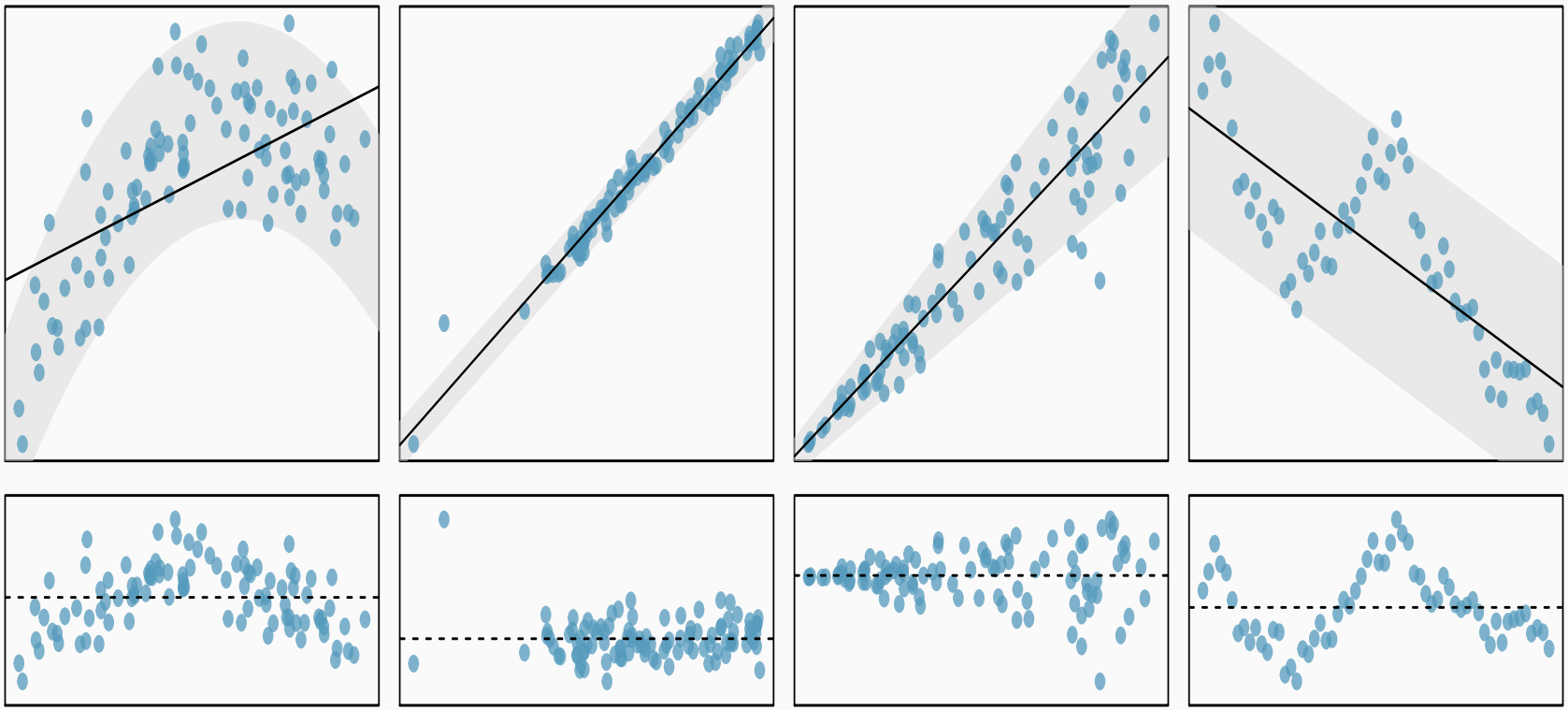
$$-0.0431 \pm 2.01 \times 0.0108 = (-0.0648, -0.0214).$$

```
> qt(p = 0.025, df = 48, lower.tail = FALSE)
[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:*

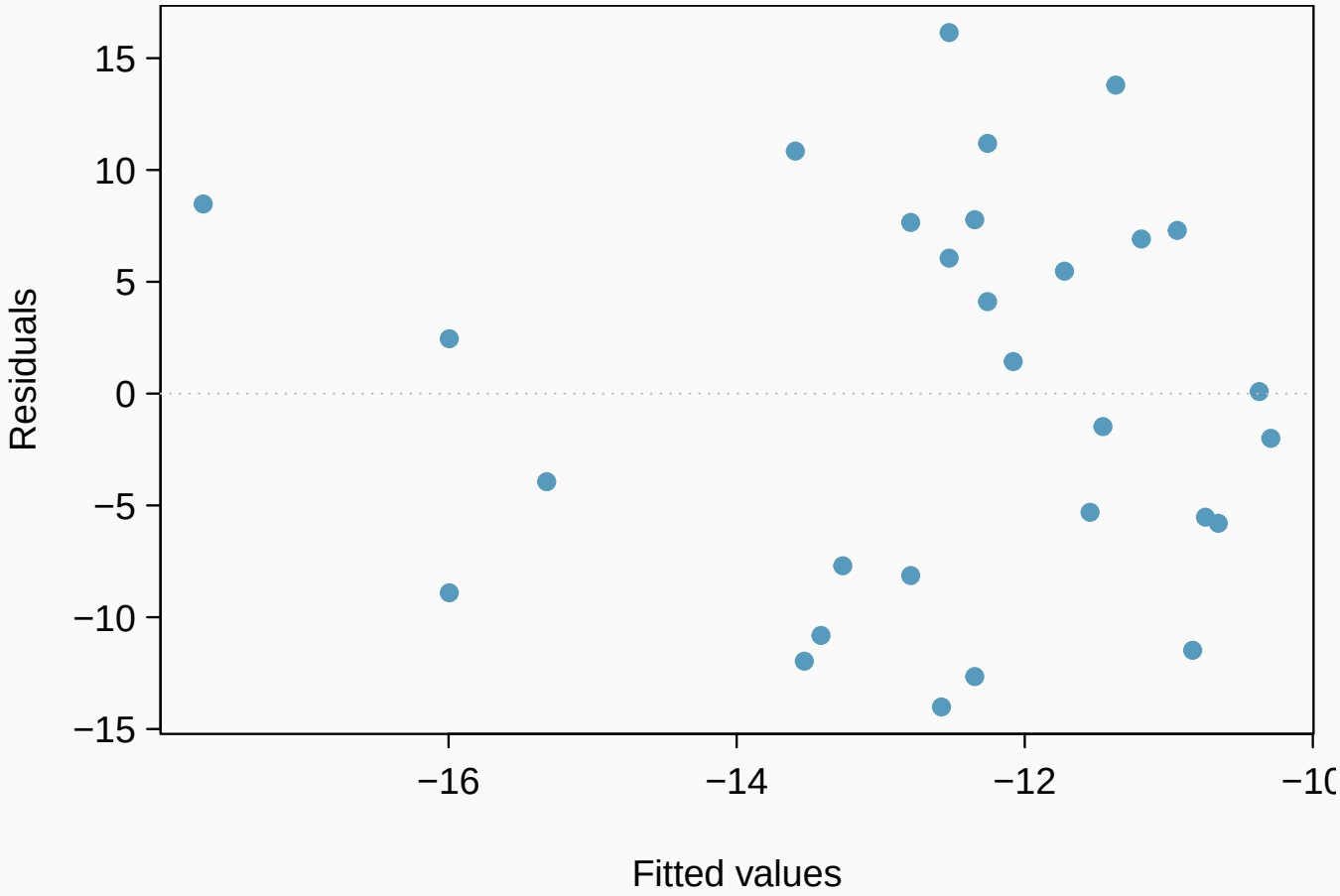
$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \cdots, n,$$

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?



Multiple linear regression

Multiple regression

- Simple linear regression: Bivariate - two variables: y and x
- Multiple linear regression: Multiple variables: y and $x_1, x_2, \cdots$

Poverty vs. region (east, west)

$$\widehat{poverty} = 11.17 + 0.38 \times west$$

- Explanatory variable: region, *reference level:* east
- *Intercept:* The estimated average poverty percentage in eastern states is 11.17%
  - This is the value we get if we plug in *0* for the explanatory variable
- *Slope:* The estimated average poverty percentage in western states is 0.38% higher than eastern states.
  - Then, the estimated average poverty percentage in western states is 11.17 + 0.38 = 11.55%.
  - This is the value we get if we plug in *1* for the explanatory variable

Poverty vs. region (northeast, midwest, west, south)

Which region (northeast, midwest, west, or south) is the reference level?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.50	0.87	10.94	0.00
region4midwest	0.03	1.15	0.02	0.98
region4west	1.79	1.13	1.59	0.12
region4south	4.16	1.07	3.87	0.00

(a) northeast

(b) midwest

(c) west

(d) south

(e) cannot tell

Poverty vs. region (northeast, midwest, west, south)

Which region (northeast, midwest, west, or south) has the lowest poverty percentage?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.50	0.87	10.94	0.00
region4midwest	0.03	1.15	0.02	0.98
region4west	1.79	1.13	1.59	0.12
region4south	4.16	1.07	3.87	0.00

(a) northeast

(b) midwest

(c) west

(d) south

(e) cannot tell

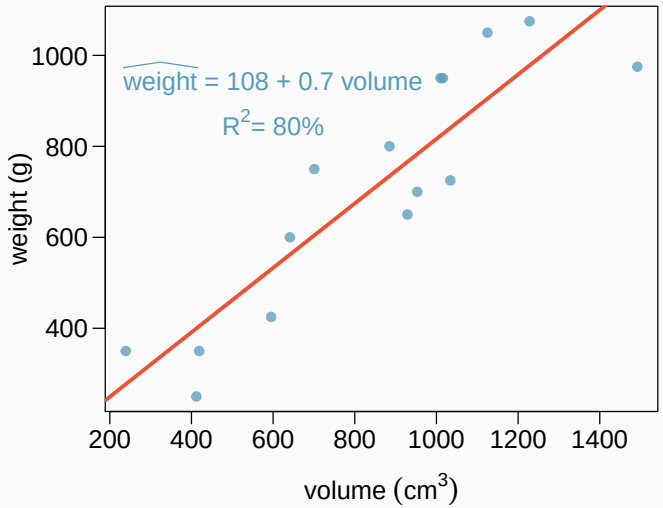
Weights of books

	weight (g)	volume (cm^3)	cover
1	800	885	hc
2	950	1016	hc
3	1050	1125	hc
4	350	239	hc
5	750	701	hc
6	600	641	hc
7	1075	1228	hc
8	250	412	pb
9	700	953	pb
10	650	929	pb
11	975	1492	pb
12	350	419	pb
13	950	1010	pb
14	425	595	pb
15	725	1034	pb

From: Maindonald, J.H. and Braun, W.J. (2nd ed., 2007) “Data Analysis and Graphics Using R”

Weights of books (cont.)

The scatterplot shows the relationship between weights and volumes of books as well as the regression output. Which of the below is correct?



(a) Weights of 80% of the books can be predicted accurately using this model.

(b) Books that are 10 cm^3 over average are expected to weigh 7 g over average.

(c) The correlation between weight and volume is $R = 0.80^2 = 0.64$.

(d) The model underestimates the weight of the book with the highest volume.

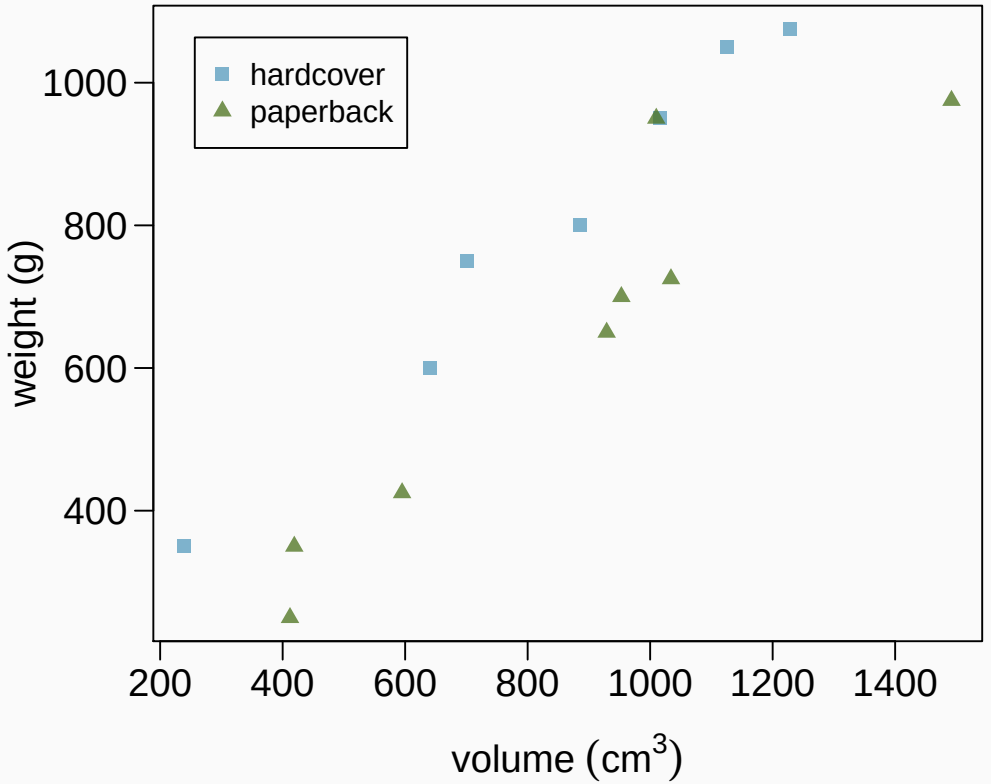
Modeling weights of books using volume

somewhat abbreviated output...

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 107.67931   88.37758   1.218    0.245
volume        0.70864    0.09746   7.271 6.26e-06


Residual standard error: 123.9 on 13 degrees of freedom
Multiple R-squared: 0.8026, Adjusted R-squared: 0.7875
F-statistic: 52.87 on 1 and 13 DF,  p-value: 6.262e-06
```

Weights of hardcover and paperback books

Can you identify a trend in the relationship between volume and weight of hardcover and paperback books?



Modeling weights of books using volume and cover type

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  197.96284   59.19274   3.344 0.005841 **
volume         0.71795    0.06153  11.669  6.6e-08 ***
cover:pb    -184.04727   40.49420  -4.545 0.000672 ***


Residual standard error: 78.2 on 12 degrees of freedom
Multiple R-squared: 0.9275, Adjusted R-squared: 0.9154
F-statistic: 76.73 on 2 and 12 DF,  p-value: 1.455e-07
```

Determining the reference level

Based on the regression output below, which level of `cover` is the reference level? Note that `pb`: paperback.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	197.9628	59.1927	3.34	0.0058
volume	0.7180	0.0615	11.67	0.0000
cover:pb	-184.0473	40.4942	-4.55	0.0007

(a) paperback

(b) hardcover

Determining the reference level

Which of the below correctly describes the roles of variables in this regression model?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	197.9628	59.1927	3.34	0.0058
volume	0.7180	0.0615	11.67	0.0000
cover:pb	-184.0473	40.4942	-4.55	0.0007

(a) response: weight, explanatory: volume, paperback cover

(b) response: weight, explanatory: volume, hardcover cover

(c) response: volume, explanatory: weight, cover type

(d) response: weight, explanatory: volume, cover type

Linear model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	197.96	59.19	3.34	0.01
volume	0.72	0.06	11.67	0.00
cover:pb	-184.05	40.49	-4.55	0.00

$$\widehat{weight} = 197.96 + 0.72\ volume - 184.05\ cover:pb$$

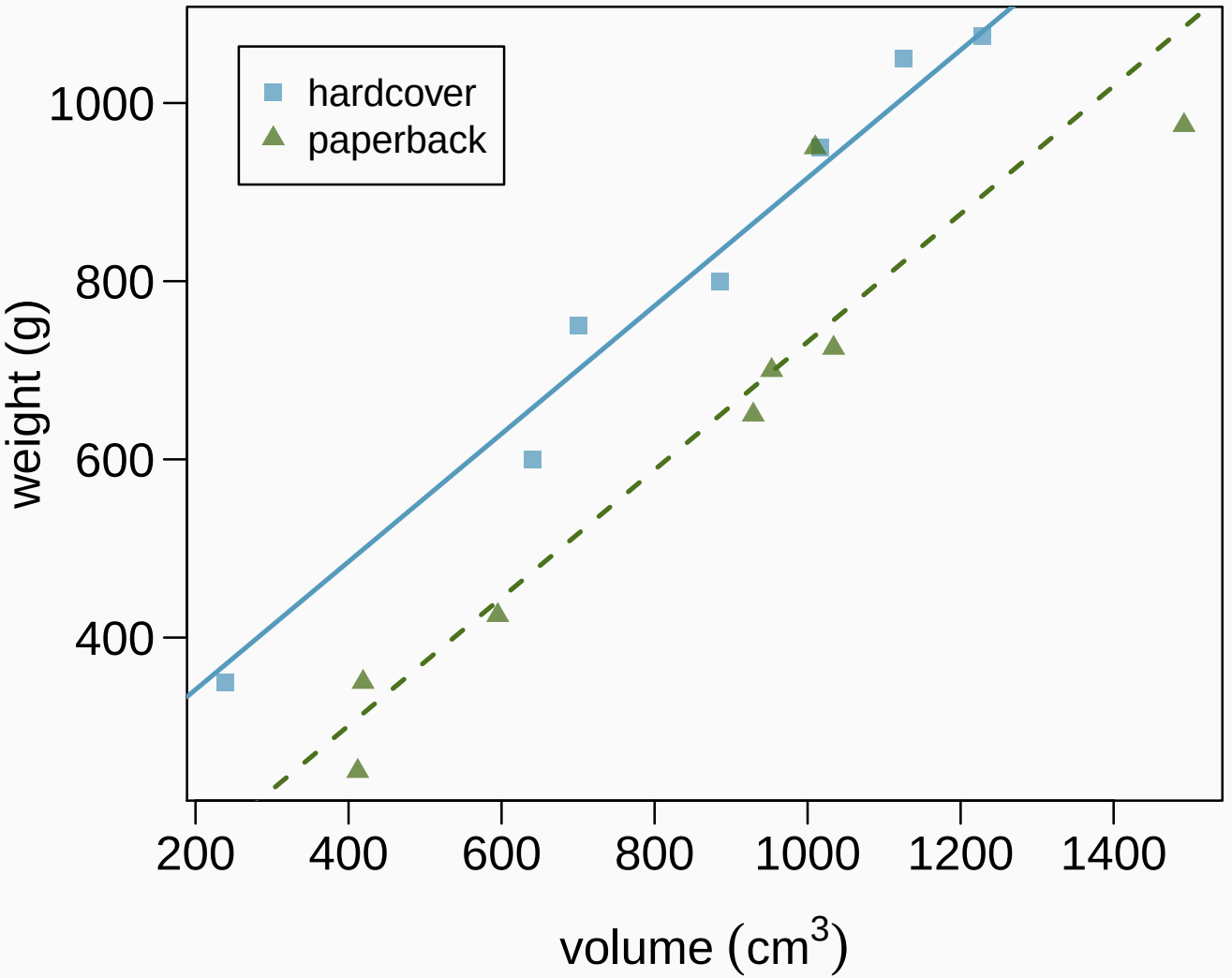
1. For *hardcover* books: plug in *0* for `cover`

$$\begin{aligned} \widehat{weight} &= 197.96 + 0.72\ volume - 184.05 \times 0 \\ &= 197.96 + 0.72\ volume \end{aligned}$$

2. For *paperback* books: plug in *1* for `cover`

$$\begin{aligned} \widehat{weight} &= 197.96 + 0.72\ volume - 184.05 \times 1 \\ &= 13.91 + 0.72\ volume \end{aligned}$$

Visualising the linear model



Interpretation of the regression coefficients

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	197.96	59.19	3.34	0.01
volume	0.72	0.06	11.67	0.00
cover:pb	-184.05	40.49	-4.55	0.00

- *Slope of volume:* All else held constant, books that are 1 more cubic centimeter in volume tend to weigh about 0.72 grams more.
- *Slope of cover:* All else held constant, the model predicts that paperback books weigh 184 grams lower than hardcover books.
- *Intercept:* Hardcover books with no volume are expected on average to weigh 198 grams.
  - Obviously, the intercept does not make sense in context. It only serves to adjust the height of the line.

Prediction

Which of the following is the correct calculation for the predicted weight of a paperback book that is 600 cm^3?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	197.96	59.19	3.34	0.01
volume	0.72	0.06	11.67	0.00
cover:pb	-184.05	40.49	-4.55	0.00

(a) 197.96 + 0.72 * 600 - 184.05 * 1

(b) 184.05 + 0.72 * 600 - 197.96 * 1

(c) 197.96 + 0.72 * 600 - 184.05 * 0

(d) 197.96 + 0.72 * 1 - 184.05 * 600

Another example: Modeling kid's test scores

Predicting cognitive test scores of three- and four-year-old children using characteristics of their mothers. Data are from a survey of adult American women and their children - a subsample from the National Longitudinal Survey of Youth.

	kid_score	mom_hs	mom_iq	mom_work	mom_age
1	65	yes	121.12	yes	27
⋮					
5	115	yes	92.75	yes	27
6	98	no	107.90	no	18
⋮					
434	70	yes	91.25	yes	25

Gelman, Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. (2007) Cambridge University Press.

Interpreting the slope

What is the correct interpretation of the slope for mom's IQ?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	19.59	9.22	2.13	0.03
mom_hs:yes	5.09	2.31	2.20	0.03
mom_iq	0.56	0.06	9.26	0.00
mom_work:yes	2.54	2.35	1.08	0.28
mom_age	0.22	0.33	0.66	0.51

Interpreting the slope

What is the correct interpretation of the intercept?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	19.59	9.22	2.13	0.03
mom_hs:yes	5.09	2.31	2.20	0.03
mom_iq	0.56	0.06	9.26	0.00
mom_work:yes	2.54	2.35	1.08	0.28
mom_age	0.22	0.33	0.66	0.51

Interpreting the slope

What is the correct interpretation of the slope for `mom_work`?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	19.59	9.22	2.13	0.03
mom_hs:yes	5.09	2.31	2.20	0.03
mom_iq	0.56	0.06	9.26	0.00
mom_work:yes	2.54	2.35	1.08	0.28
mom_age	0.22	0.33	0.66	0.51

All else being equal, kids whose moms worked during the first three years of the kid's life

(a) are estimated to score 2.54 points lower

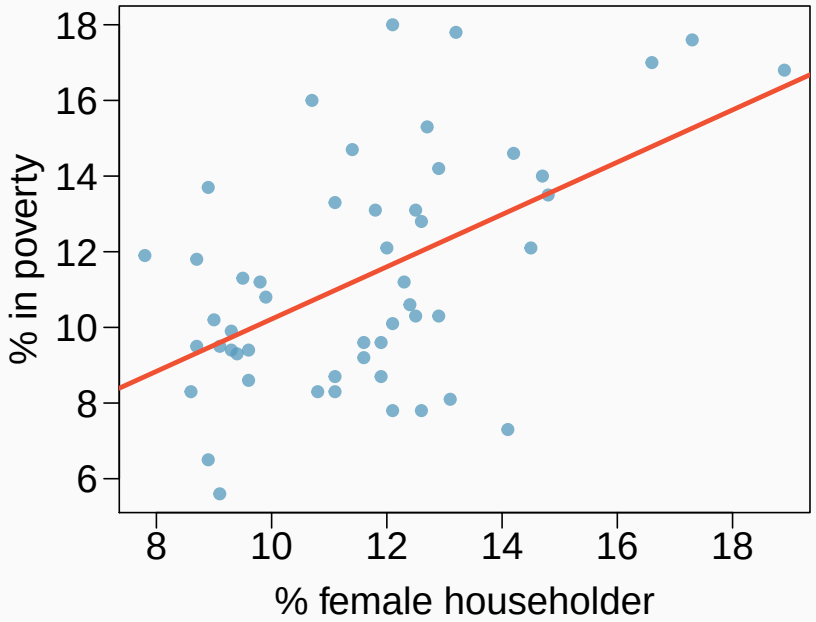
(b) are estimated to score 2.54 points higher

than those whose moms did not work.

Revisit: Modeling poverty

Predicting poverty using % female householder

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.31	1.90	1.74	0.09
female_house	0.69	0.16	4.32	0.00



$$R = 0.53$$

$$R^2 = 0.53^2 = 0.28$$

Another look at R^2

R^2 can be calculated in three ways:

1. square the correlation coefficient of x and y (how we have been calculating it)
2. square the correlation coefficient of y and $\hat{y}$
3. based on definition:

$$R^2 = \frac{\text{explained variability in } y}{\text{total variability in } y}$$

Using *ANOVA* we can calculate the explained variability and total variability in y.

Sum of squares

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
female_house	1	132.57	132.57	18.68	0.00
Residuals	49	347.68	7.10		
Total	50	480.25			

$$\begin{aligned}
\text{Sum of squares of } y\text{: } SS_{Total} &= \sum (y - \bar{y})^2 = 480.25 \rightarrow \textit{total variability} \\
\text{Sum of squares of residuals: } SS_{Error} &= \sum e_i^2 = 347.68 \rightarrow \textit{unexplained variabilit} \\
\text{Sum of squares of } x\text{: } SS_{Model} &= SS_{Total} - SS_{Error} \rightarrow \textit{explained variability} \\
&= 480.25 - 347.68 = 132.57
\end{aligned}$$

$$R^2 = \frac{\text{explained variability}}{\text{total variability}} = \frac{132.57}{480.25} = 0.28 \checkmark$$

Why bother?

Why bother with another approach for calculating R^2 when we had a perfectly good way to calculate it as the correlation coefficient squared?

Predicting poverty using % female hh + % white

Linear model:	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-2.58	5.78	-0.45	0.66
female_house	0.89	0.24	3.67	0.00
white	0.04	0.04	1.08	0.29

ANOVA:	Df	Sum Sq	Mean Sq	F value	Pr(>F)
female_house	1	132.57	132.57	18.74	0.00
white	1	8.21	8.21	1.16	0.29
Residuals	48	339.47	7.07		
Total	50	480.25			

$$R^2 = \frac{\text{explained variability}}{\text{total variability}} = \frac{132.57 + 8.21}{480.25} = 0.29$$

Does adding the variable `white` to the model add valuable information that wasn't provided by `female_house`?

Collinearity between explanatory variables

poverty vs. % female head of household

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.31	1.90	1.74	0.09
female_house	0.69 *0.69*	0.16	4.32	0.00

poverty vs. % female head of household and % female hh

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-2.58	5.78	-0.45	0.66
female_house	0.89 *0.89*	0.24	3.67	0.00
white	0.04	0.04	1.08	0.29

Collinearity between explanatory variables (cont.)

- Two predictor variables are said to be collinear when they are correlated, and this *collinearity* complicates model estimation.

  Remember: Predictors are also called explanatory or independent variables. Ideally, they would be independent of each other.

- We don't like adding predictors that are associated with each other to the model, because often times the addition of such variable brings nothing to the table. Instead, we prefer the simplest best model, i.e. *parsimonious* model.
- While it's impossible to avoid collinearity from arising in observational data, experiments are usually designed to prevent correlation among predictors.

R^2 vs. adjusted R^2

	R^2	Adjusted R^2
Model 1 (Single-predictor)	0.28	0.26
Model 2 (Multiple)	0.29	0.26

- When any variable is added to the model R^2 increases.
- But if the added variable doesn't really provide any new information, or is completely unrelated, adjusted R^2 does not increase.

Adjusted R^2

Adjusted R^2

$$R^2_{adj} = 1 - \left(\frac{SS_{Error}}{SS_{Total}} \times \frac{n-1}{n-p-1} \right)$$

where n is the number of cases and p is the number of predictors (explanatory variables) in the model.

- Because p is never negative, R^2_{adj} will always be smaller than R^2.
- R^2_{adj} applies a penalty for the number of predictors included in the model.
- Therefore, we choose models with higher R^2_{adj} over others.

Calculate adjusted R^2

ANOVA:	Df	Sum Sq	Mean Sq	F value	Pr(>F)
female_house	1	132.57	132.57	18.74	0.0001
white	1	8.21	8.21	1.16	0.2868
Residuals	48	339.47	7.07		
Total	50	480.25			

$$
\begin{aligned}
R^2_{adj} &= 1 - \left(\frac{SS_{Error}}{SS_{Total}} \times \frac{n-1}{n-p-1} \right) \\
&= 1 - \left(\frac{339.47}{480.25} \times \frac{51-1}{51-2-1} \right) \\
&= 1 - \left(\frac{339.47}{480.25} \times \frac{50}{48} \right) \\
&= 1 - 0.74 \\
&= 0.26
\end{aligned}
$$

Model selection

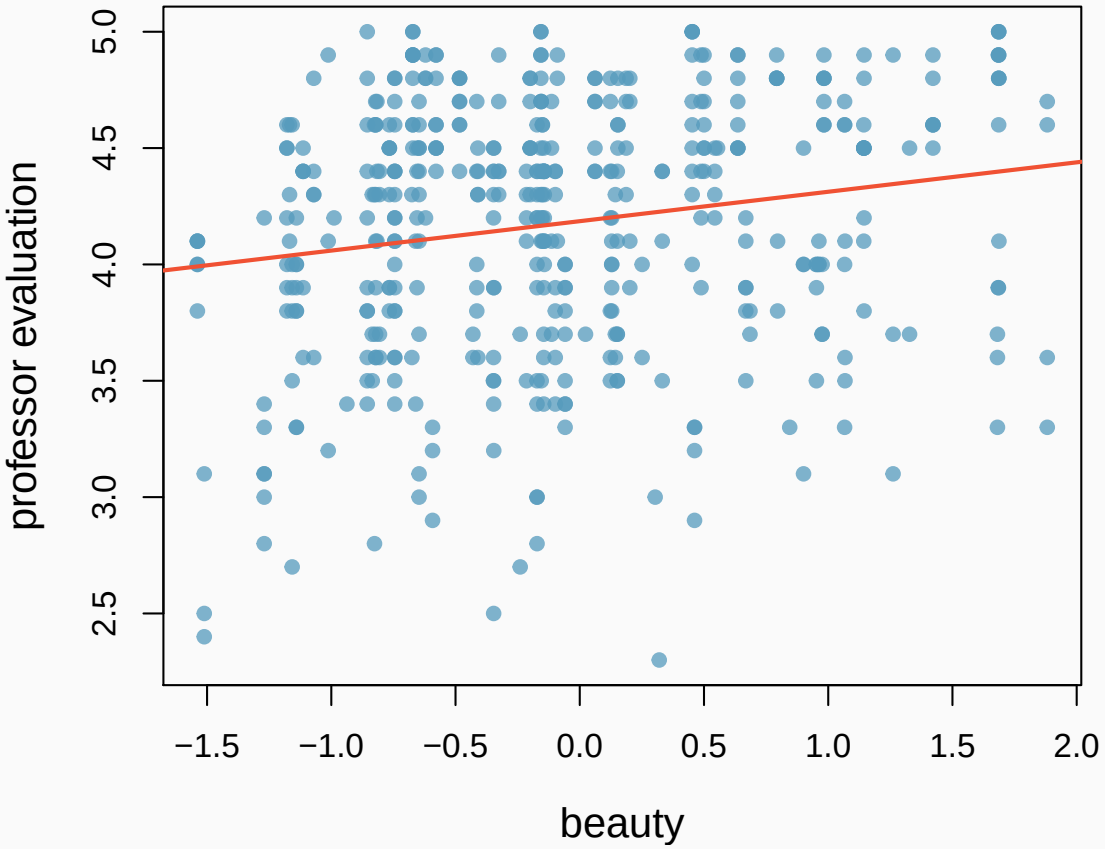
Beauty in the classroom

- Data: Student evaluations of instructors' beauty and teaching quality for 463 courses at the University of Texas.
- Evaluations conducted at the end of semester, and the beauty judgements were made later, by six students who had not attended the classes and were not aware of the course evaluations (2 upper level females, 2 upper level males, one lower level female, one lower level male).

Hamermesh & Parker. (2004)"Beauty in the classroom: instructors? pulchritude and putative pedagogical productivity? Economics Education Review.

Professor rating vs. beauty

Professor evaluation score (higher score means better) vs. beauty score (a score of 0 means average, negative score means below average, and a positive score above average):



Which of the below is correct based on the model output?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.19	0.03	167.24	0.00
beauty	0.13	0.03	4.00	0.00

$R^2 = 0.0336$

(a) Model predicts 3.36% of professor ratings correctly.

(b) Beauty is not a significant predictor of professor evaluation.

(c) Professors who score 1 point above average in their beauty score are tend to also score 0.13 points higher in their evaluation.

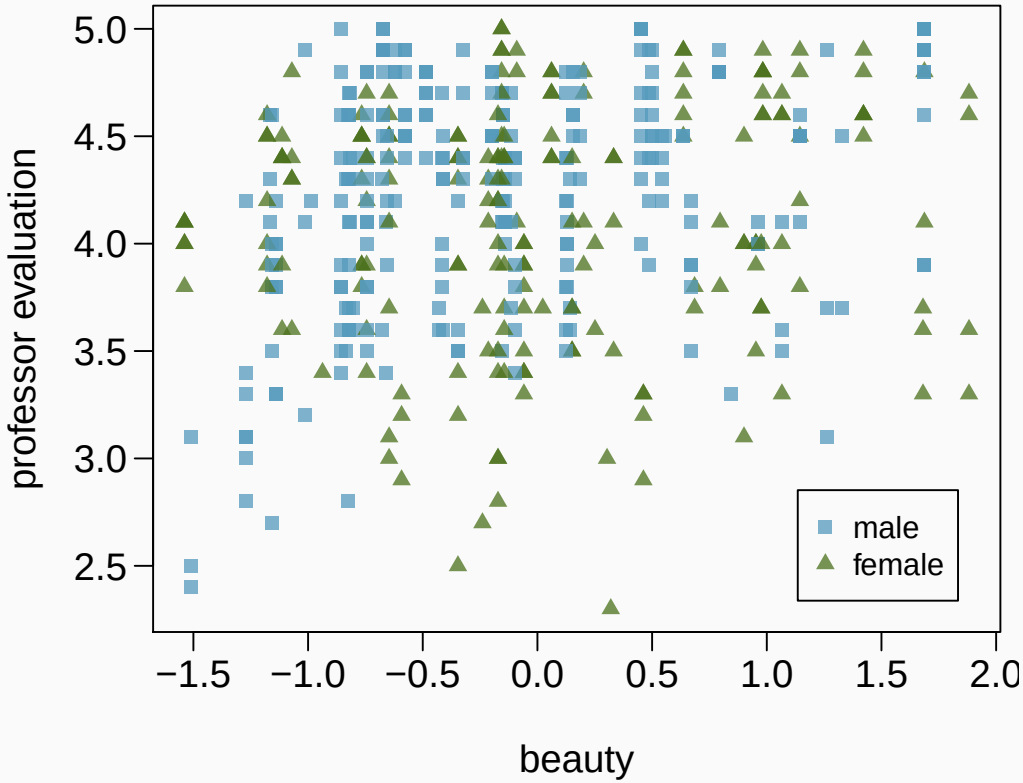
(d) 3.36% of variability in beauty scores can be explained by professor evaluation.

(e) The correlation coefficient could be $\sqrt{0.0336} = 0.18$ or -0.18, we can't tell which is correct.

Exploratory analysis

Any interesting features?

For a given beauty score, are male professors evaluated higher, lower, or about the same as female professors?



Professor rating vs. beauty + gender

For a given beauty score, are male professors evaluated higher, lower, or about the same as female professors?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.09	0.04	107.85	0.00
beauty	0.14	0.03	4.44	0.00
gender.male	0.17	0.05	3.38	0.00

$R^2_{adj} = 0.057$

(a) higher

(b) lower

(c) about the same

Full model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.6282	0.1720	26.90	0.00
beauty	0.1080	0.0329	3.28	0.00
gender.male	0.2040	0.0528	3.87	0.00
age	-0.0089	0.0032	-2.75	0.01
formal.yes [1]	0.1511	0.0749	2.02	0.04
lower.yes [2]	0.0582	0.0553	1.05	0.29
native.non english	-0.2158	0.1147	-1.88	0.06
minority.yes	-0.0707	0.0763	-0.93	0.35
students [3]	-0.0004	0.0004	-1.03	0.30
tenure.tenure track [4]	-0.1933	0.0847	-2.28	0.02
tenure.tenured	-0.1574	0.0656	-2.40	0.02

[1] `formal`: picture wearing tie&jacket/blouse, levels: `yes`, `no`

[2] `lower`: lower division course, levels: `yes`, `no`

[3] `students`: number of students

[4] `tenure`: tenure status, levels: `non-tenure track`, `tenure track`, `tenured`

Hypotheses

Just as the interpretation of the slope parameters take into account all other variables in the model, the hypotheses for testing for significance of a predictor also takes into account all other variables.

$H_0 : B_i = 0$ when other explanatory variables are included in the model.

$H_A : B_i \neq 0$ when other explanatory variables are included in the model.

Assessing significance: numerical variables

The p-value for age is 0.01. What does this indicate?

	Estimate	Std. Error	t value	Pr(>|t|)
...				
age	-0.0089	0.0032	-2.75	0.01
...				

(a) Since p-value is positive, higher the professor's age, the higher we would expect them to be rated.

(b) If we keep all other variables in the model, there is strong evidence that professor's age is associated with their rating.

(c) Probability that the true slope parameter for age is 0 is 0.01.

(d) There is about 1% chance that the true slope parameter for age is -0.0089.

Assessing significance: categorical variables

Tenure is a categorical variable with 3 levels: non tenure track, tenure track, tenured. Based on the model output given, which of the below is false?

	Estimate	Std. Error	t value	Pr(>|t|)
...				
tenure.tenure track	-0.1933	0.0847	-2.28	0.02
tenure.tenured	-0.1574	0.0656	-2.40	0.02

(a) Reference level is non tenure track.

(b) All else being equal, tenure track professors are rated, on average, 0.19 points lower than non-tenure track professors.

(c) All else being equal, tenured professors are rated, on average, 0.16 points lower than non-tenure track professors.

(d) All else being equal, there is a significant difference between the average ratings of tenure track and tenured professors.

Assessing significance

Which predictors do not seem to meaningfully contribute to the model, i.e. may not be significant predictors of professor's rating score?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.6282	0.1720	26.90	0.00
beauty	0.1080	0.0329	3.28	0.00
gender.male	0.2040	0.0528	3.87	0.00
age	-0.0089	0.0032	-2.75	0.01
formal.yes	0.1511	0.0749	2.02	0.04
lower.yes	0.0582	0.0553	1.05	0.29
native.non english	-0.2158	0.1147	-1.88	0.06
minority.yes	-0.0707	0.0763	-0.93	0.35
students	-0.0004	0.0004	-1.03	0.30
tenure.tenure track	-0.1933	0.0847	-2.28	0.02
tenure.tenured	-0.1574	0.0656	-2.40	0.02

Model selection strategies

Based on what we've learned so far, what are some ways you can think of that can be used to determine which variables to keep in the model and which to leave out?

Backward-elimination

1. Start with the full model
2. Drop one variable at a time and record R^2_{adj} of each smaller model
3. Pick the model with the highest increase in R^2_{adj}
4. Repeat until none of the models yield an increase in R^2_{adj}

Backward-elimination

Full	beauty + gender + age + formal + lower + native + minority + students + tenure	*0.0839*
Step 1	gender + age + formal + lower + native + minority + students + tenure	0.0642
	beauty + age + formal + lower + native + minority + students + tenure	0.0557
	beauty + gender + formal + lower + native + minority + students + tenure	0.0706
	beauty + gender + age + lower + native + minority + students + tenure	0.0777
	beauty + gender + age + formal + native + minority + students + tenure	0.0837
	beauty + gender + age + formal + lower + minority + students + tenure	0.0788
	beauty + gender + age + formal + lower + native + students + tenure	*0.0842*
	beauty + gender + age + formal + lower + native + minority + tenure	0.0838
	beauty + gender + age + formal + lower + native + minority + students	0.0733
Step 2	gender + age + formal + lower + native + students + tenure	0.0647
	beauty + age + formal + lower + native + students + tenure	0.0543
	beauty + gender + formal + lower + native + students + tenure	0.0708
	beauty + gender + age + lower + native + students + tenure	0.0776
	beauty + gender + age + formal + native + students + tenure	*0.0846*
	beauty + gender + age + formal + lower + native + tenure	0.0844
	beauty + gender + age + formal + lower + native + students	0.0725
Step 3	gender + age + formal + native + students + tenure	0.0653
	beauty + age + formal + native + students + tenure	0.0534
	beauty + gender + formal + native + students + tenure	0.0707
	beauty + gender + age + native + students + tenure	0.0786
	beauty + gender + age + formal + students + tenure	0.0756
	beauty + gender + age + formal + native + tenure	*0.0855*
	beauty + gender + age + formal + native + students	0.0713
Step 4	gender + age + formal + native + tenure	0.0667
	beauty + age + formal + native + tenure	0.0553
	beauty + gender + formal + native + tenure	0.0723
	beauty + gender + age + native + tenure	0.0806
	beauty + gender + age + formal + tenure	0.0773
	beauty + gender + age + formal + native	0.0713

step function in R

The step function in R does a similar backward elimination process, however it uses a different metric called AIC (Akaike Information Criterion) instead of adjusted R^2 to do the model selection.

```
Call:
lm(formula = profevaluation ~ beauty + gender + age + formal +
    native + tenure, data = d)

Coefficients:
       (Intercept)              beauty          gendermale
          4.628435            0.105546            0.208079
                age          formalyes   nativenon english
         -0.008844            0.132422           -0.243003
tenuretenure track       tenuretenured
         -0.206784           -0.175967
```

Best model: beauty + gender + age + formal + native + tenure

Forward selection

1. Start with regressions of response vs. each explanatory variable
2. Pick the model with the highest R^2_{adj}
3. Add the remaining variables one at a time to the existing model, and once again pick the model with the highest R^2_{adj}
4. Repeat until the addition of any of the remanning variables does not result in a higher R^2_{adj}

- Backward elimination with the p-value approach:
  1. Start with the full model
  2. Drop the variable with the highest p-value and refit a smaller model
  3. Repeat until all variables left in the model are significant
- Forward selection with the p-value approach:
  1. Start with regressions of response vs. each explanatory variable
  2. Pick the variable with the lowest significant p-value
  3. Add the remaining variables one at a time to the existing model, and pick the variable with the lowest significant p-value
  4. Repeat until any of the remaining variables does not have a significant p-value

Backward-elimination: *p – value* approach

Step	Variables included & p-value									
Full	beauty	gender	age	formal	lower	native	minority	students	tenure	tenure
		male		yes	yes	nonenglish	yes		tenure track	tenured
	0.00	0.00	0.01	0.04	0.29	0.06	*0.35*	0.30	0.02	0.02
Step 1	beauty	gender	age	formal	lower	native		students	tenure	tenure
		male		yes	yes	nonenglish			tenure track	tenured
	0.00	0.00	0.01	0.04	*0.38*	0.03		0.34	0.02	0.01
Step 2	beauty	gender	age	formal		native		students	tenure	tenure
		male		yes		nonenglish			tenure track	tenured
	0.00	0.00	0.01	0.05		0.02		*0.44*	0.01	0.01
Step 3	beauty	gender	age	formal		native			tenure	tenure
		male		yes		nonenglish			tenure track	tenured
	0.00	0.00	0.01	*0.06*		0.02			0.01	0.01
Step 4	beauty	gender	age			native			tenure	tenure
		male				nonenglish			tenure track	tenured
	0.00	0.00	0.01			*0.06*			0.01	0.01
Step 5	beauty	gender	age						tenure	tenure
		male							tenure track	tenured
	0.00	0.00	0.01						0.01	0.01

Best model: beauty + gender + age + tenure

Adjusted R^2 vs. p-value approaches

- The two approaches are similar, but they sometimes lead to different models, with the adjusted R^2 approach tending to include more predictors in the final model.
- When the sole goal is to improve prediction accuracy, use R^2 . This is commonly the case in machine learning applications.
- When we care about understanding which variables are statistically significant predictors of the response, or if there is interest in producing a simpler model at the potential cost of a little prediction accuracy, then the p-value approach is preferred.
- Regardless of the approach we use, our job is not done after variable selection – we must still verify the model conditions are reasonable.

Checking model conditions using graphs

Modeling conditions

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

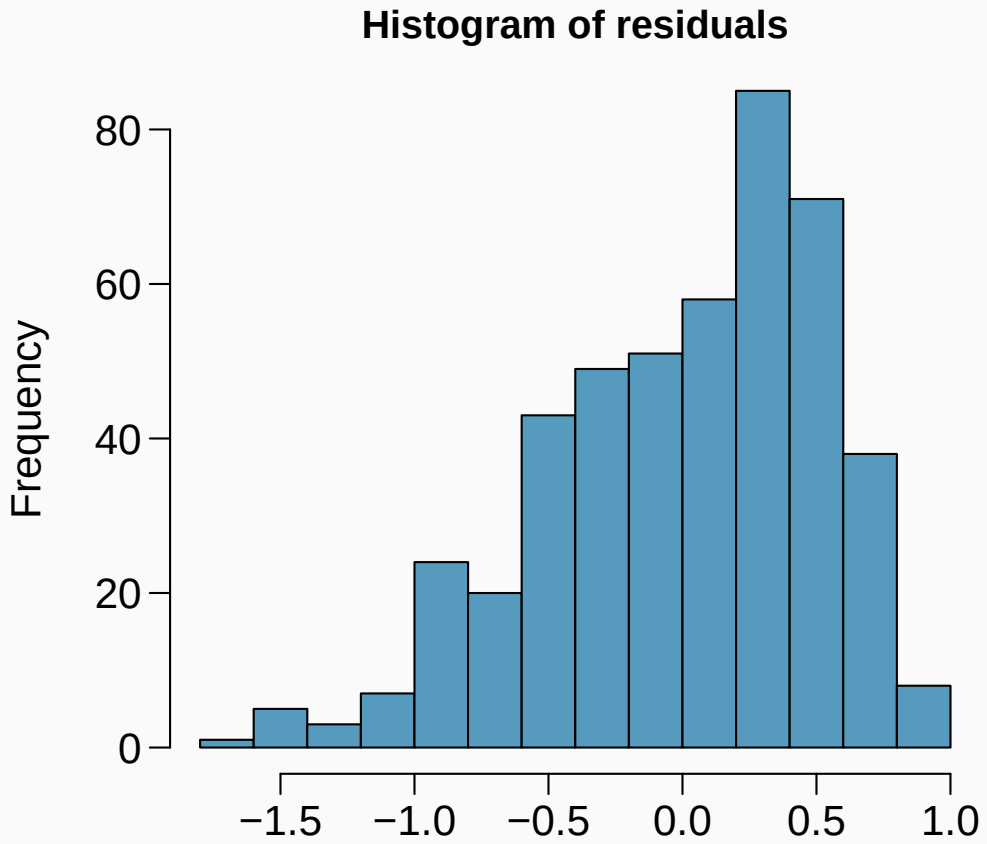
The model depends on the following conditions

1. residuals are nearly normal (less important for larger data sets)
2. residuals have constant variability
3. residuals are independent
4. each variable is linearly related to the outcome

We often use graphical methods to check the validity of these conditions, which we will go through in detail in the following slides.

(1) nearly normal residuals

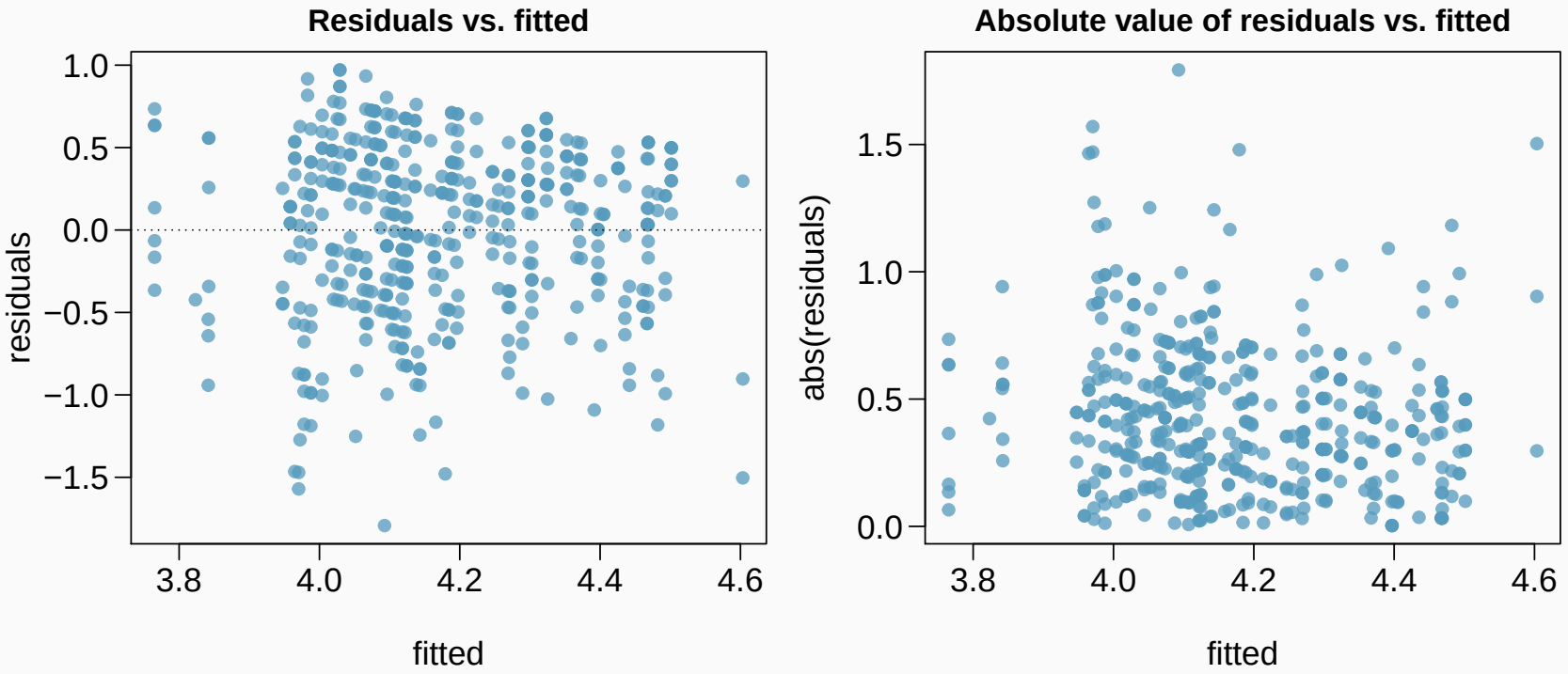
normal probability plot and/or histogram of residuals:



Does this condition appear to be satisfied?

(2) constant variability in residuals

scatterplot of residuals and/or absolute value of residuals vs. fitted (predicted):



Does this condition appear to be satisfied?

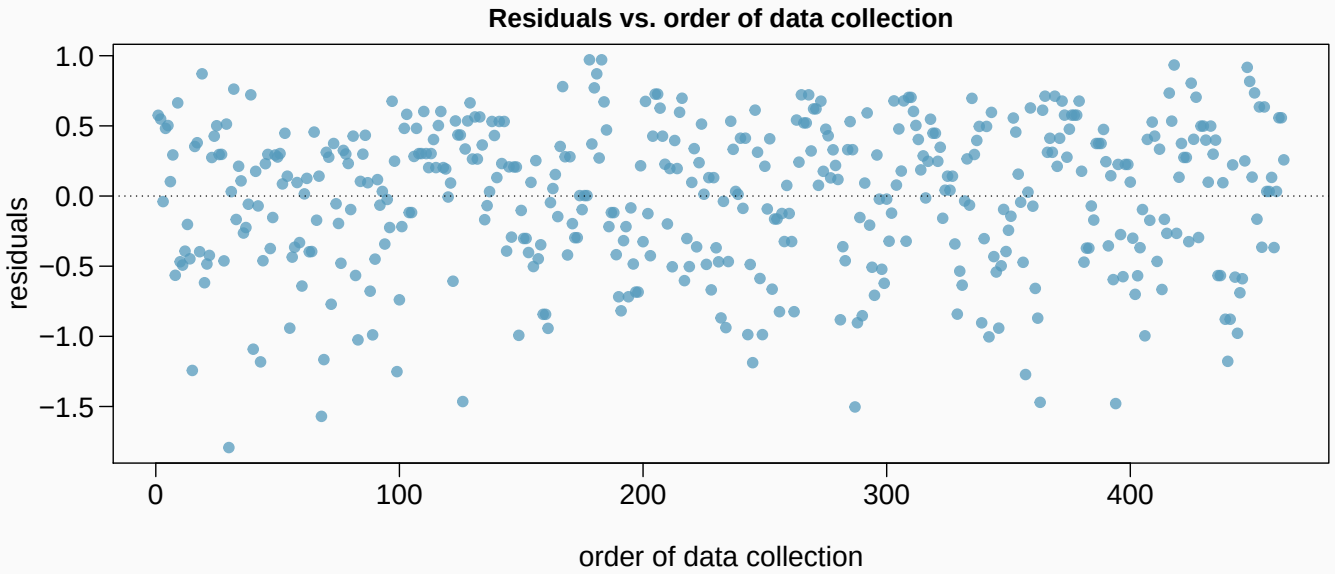
Checking constant variance - recap

- When we did simple linear regression (one explanatory variable) we checked the constant variance condition using a plot of *residuals vs. x*.
- With multiple linear regression (2+ explanatory variables) we checked the constant variance condition using a plot of *residuals vs. fitted*.

Why are we using different plots?

(3) independent residuals

scatterplot of residuals vs. order of data collection:



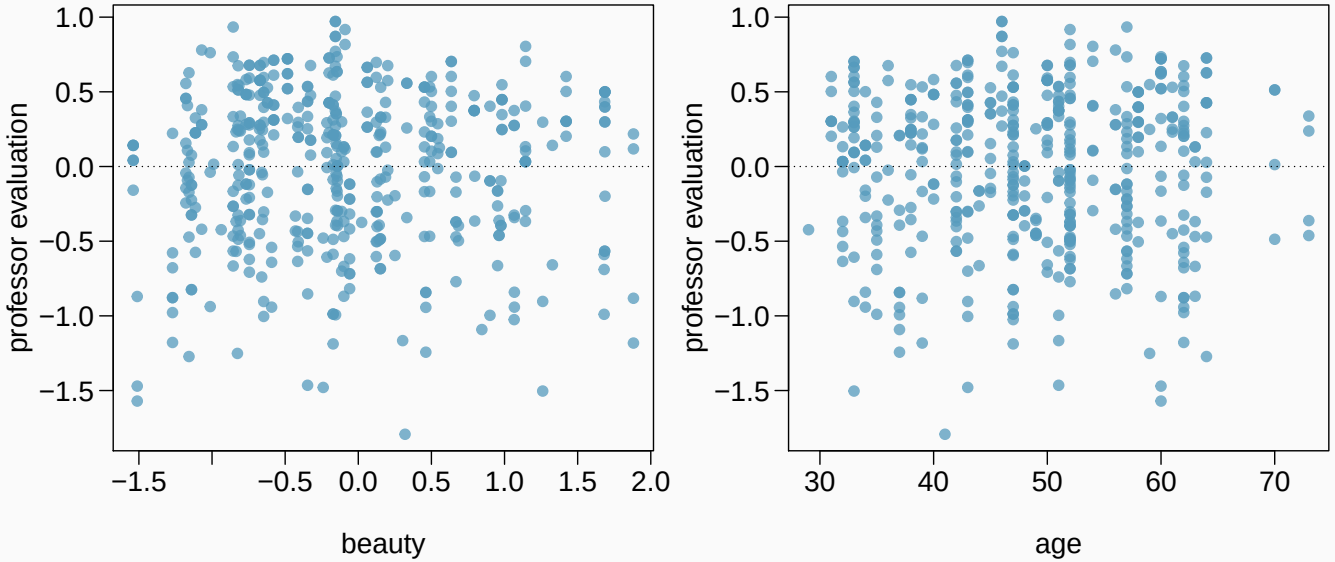
Does this condition appear to be satisfied?

More on the condition of independent residuals

- Checking for independent residuals allows us to indirectly check for independent observations.
- If observations and residuals are independent, we would not expect to see an increasing or decreasing trend in the scatterplot of residuals vs. order of data collection.
- This condition is often violated when we have time series data. Such data require more advanced time series regression techniques for proper analysis.

(4) linear relationships

scatterplot of residuals vs. each (numerical) explanatory variable:



Does this condition appear to be satisfied?

Note: We use residuals instead of the predictors on the y-axis so that we can still check for linearity without worrying about other possible violations like collinearity between the predictors.

Several options for improving a model

- Transforming variables
- Seeking out additional variables to fill model gaps
- Using more advanced methods that would account for challenges around inconsistent variability or nonlinear relationships between predictors and the outcome

Transformations

If the concern with the model is non-linear relationships between the explanatory variable(s) and the response variable, transforming the response variable can be helpful.

- Log transformation ($\log y$)
- Square root transformation ($\sqrt{y}$)
- Inverse transformation ($1/y$)
- Truncation (cap the max value possible)

It is also possible to apply transformations to the explanatory variable(s), however such transformations tend to make the model coefficients even harder to interpret.

Models can be wrong, but useful

All models are wrong, but some are useful. - George Box

- No model is perfect, but even imperfect models can be useful, as long as we are clear and report the model's shortcomings.
- If conditions are grossly violated, we should not report the model results, but instead consider a new model, even if it means learning more statistical methods or hiring someone who can help.

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)