STAT011 Statistical Methods I

Lecture 22 Simple Linear Regression I

Lu Chen
Swarthmore College
4/18/2019

Review

Statistical Inference		No Explanatory	Explanatory		
			Binary	Categorical	Quantitative
Response	Binary	Inference of a proportion *(Lecture 18)*	Inference of two proportions *(Lecture 19)*		——
	Categorical	Goodness-of-fit test *(Lecture 20)*	Chi-squared test *(Lecture 20)*		——
	Quantitative	One-sample t test *(Lecture 15)*	Two-sample t test *(Lecture 16~17)*	——	Linear regression *(Lecture 22~25)*

Outline

- Least-squares regression review
  - Scatterplot and correlation
  - Least-squares regression
  - Assessing the regression line: residual plot and r^2
- Simple linear regression
  - Idea
  - Model
- Inference for the regression line
  - Confidence intervals of intercept and slope
  - Significance test for the slope

Relationship btw two quantitative variables

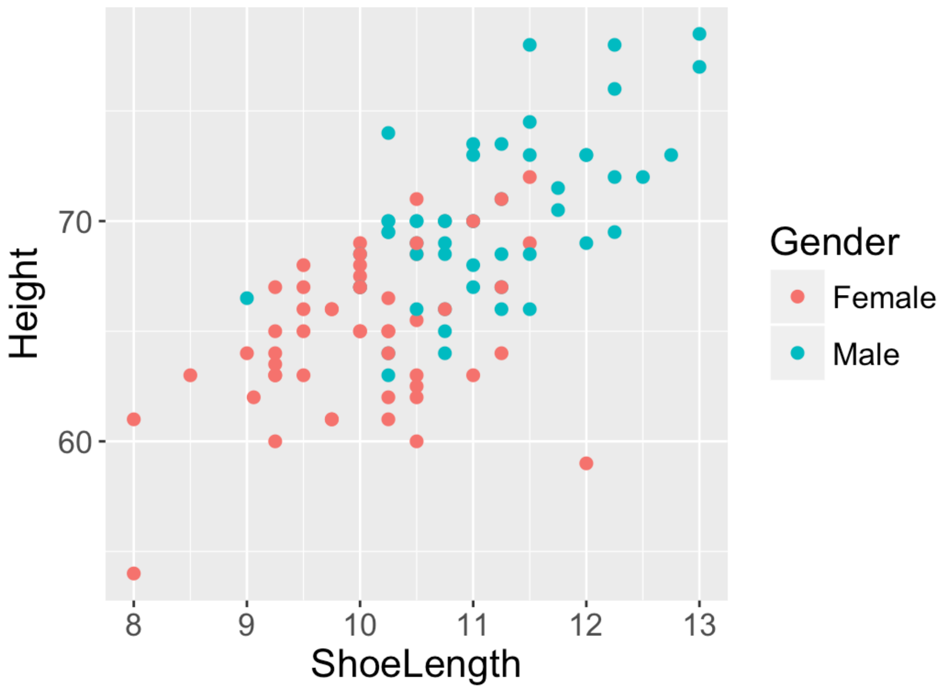
```r
head(Survey[, c("Height", "ShoeLength")])
```

```
##   Height ShoeLength
## 1   61.0       8.00
## 2   66.0      10.75
## 3   70.0      11.00
## 4   63.0       9.50
## 5   67.5      10.00
## 6   62.0       9.06
```

Is *Height* related to *ShoeLength*? Let's take *Height* as the response variable and *ShoeLength* as the explanatory variable.

Scatterplot

```
library(ggplot2)
theme_update(text=element_text(size=16)) # Set larger text size
gp <- ggplot(data=Survey, aes(x=ShoeLength, y=Height))+ # Specify data and variables
      geom_point(aes(colour=Gender), size=2) # Scatterplot
gp
```



- Scatterplot: relationship btw two quantitative variables
  - y-axis: response variable
  - x-axis: explanatory variable
- Description
  - Form: linear or curved or none?
  - Direction: positive or negative?
  - Strength: strong or weak?
  - Any outlier?

Correlation

```r
cor(Survey$ShoeLength, Survey$Height, use = "complete.obs")
```

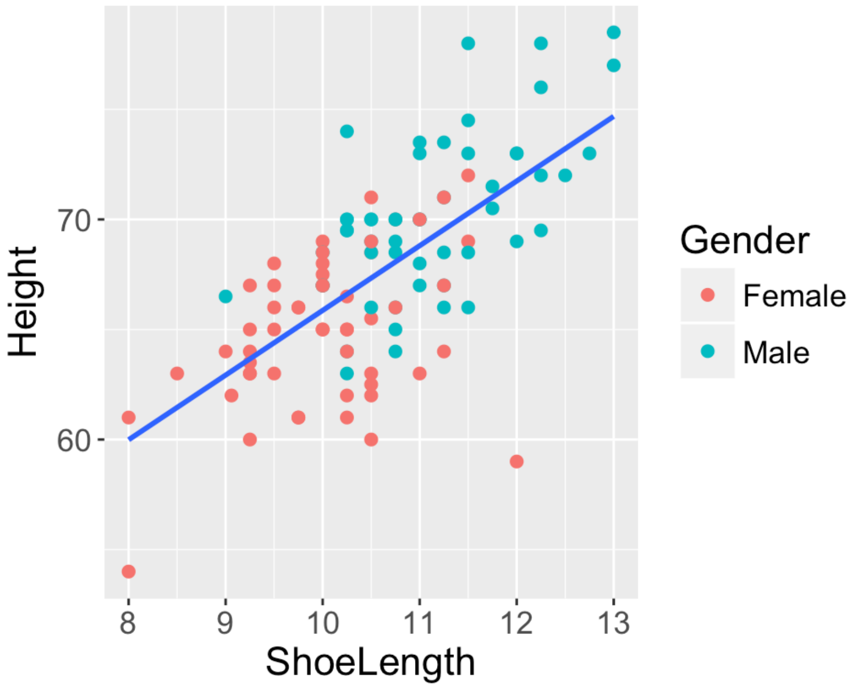
```
## [1] 0.6730721
```

> The **correlation** measures the *direction* and *strength* of the **linear relationship** between two quantitative variables. Correlation is usually written as ***r***.

- $-1 \leq r \leq 1$
- $r > 0$: positive relationship
- $r < 0$: negative relationship
- $r = 0$: no relationship
- $r = \pm 1$: perfect relationship

Least-squares regression

```r
# Add regression line
gp+geom_smooth(method="lm", se=F)
```



$$\hat{y} = b_0 + b_1 x$$
$$y = b_0 + b_1 x + e = \hat{y} + e$$

- Y: response variable (*Height*)
  - y: observed values of variable Y
  - $\hat{y}$: predicted values of variable Y
- X: explanatory variable (*ShoeLength*)
  - x: observed values of variable X
- e: difference between the observed and the predicted values of Y
- b_0: **intercept**. The value of $\hat{y}$ when $x = 0$
- b_1: **slope**. The amount by which $\hat{y}$ changes when x increases by one unit.

Least-squares regression

> The **least-squares regression** line of y on x is the line that **minimizes** the sum of the squares of the vertical distances from the data points to the line.

In least-squares regression, we minimize

$$\sum e^2 = \sum \left(y - \hat{y}\right)^2 = \sum (y - b_0 - b_1 x)^2$$

where

$$e = y - \hat{y}$$

is difined as **residual**, the difference between the observed and the predicted y.

Least-squares regression

Minimizing $\sum (y - b_0 - b_1 x)^2$, we find

$$\text{Slope } b_1 = r\frac{s_y}{s_x}, \text{ Intercept } b_0 = \bar{y} - b_1\bar{x},$$

- $\bar{x}$ and $\bar{y}$: mean of X and Y
- s_x and s_y: standard deviatioin of X and Y
- r: correlation between X and Y

Least-squares regression

```r
mymodel <- lm(Height ~ ShoeLength, data=Survey)
mymodel
```

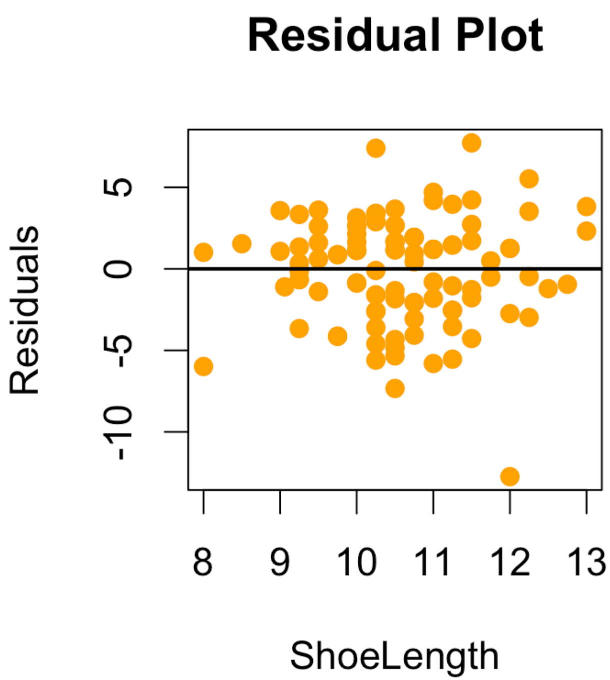
```
##
## Call:
## lm(formula = Height ~ ShoeLength, data = Survey)
##
## Coefficients:
## (Intercept)   ShoeLength
##      36.476        2.939
```

"Write down the regression line": $\hat{y} = 36.5 + 2.9x$.

- **Interpret the intercept**: when *ShoeLength* is 0, *Height* is 36.5 inches.
  - The interpretation does not have practical meaning in this case.
- **Interpret the slope**: when *ShoeLength* increases 1 inch, *Height* increases 2.9 inches.

Assessment: residual plot

```
Residuals <- mymodel$residuals
plot(Survey$ShoeLength, Residuals, pch=19, col="orange",
     xlab="ShoeLength", main="Residual Plot")
abline(h=0, lwd=2) # Add a horizontal y=0 line.
```



A **residual plot** is a scatterplot of the regression residuals against the explanatory variable. Residual plots help us assess the fit of a regression line.

- If the regression line catches the overall linear pattern of the data, there should be *no pattern* in the residual plot.
- If the residual plot shows *any pattern*, the regression line is NOT the best way to describing the data.

Assessment: coefficient of determination

> **Coefficient of determination r^2** is the **fraction of the variation** in the values of y that is explained by the least squares regression of y on x.

- **The value of r^2**: correlation squared.

```
cor(Survey$ShoeLength, Survey$Height, use = "complete.obs")^2
```

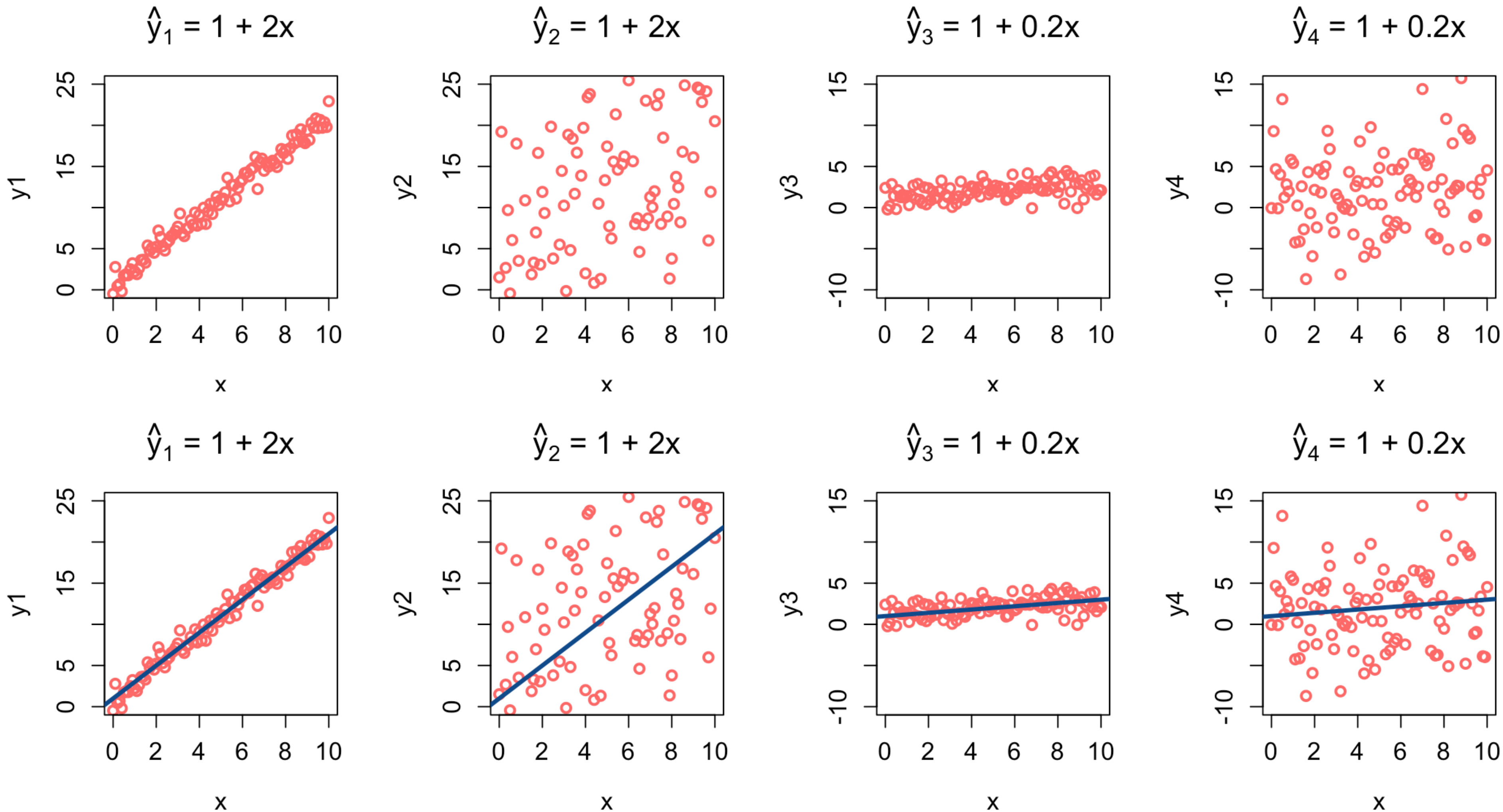
```
## [1] 0.4530261
```

- **The interpretation of r^2**: the fraction of the variation in the values of y that is explained by $\hat{y} = b_0 + b_1 x$ ("the least squares regression of y on x").

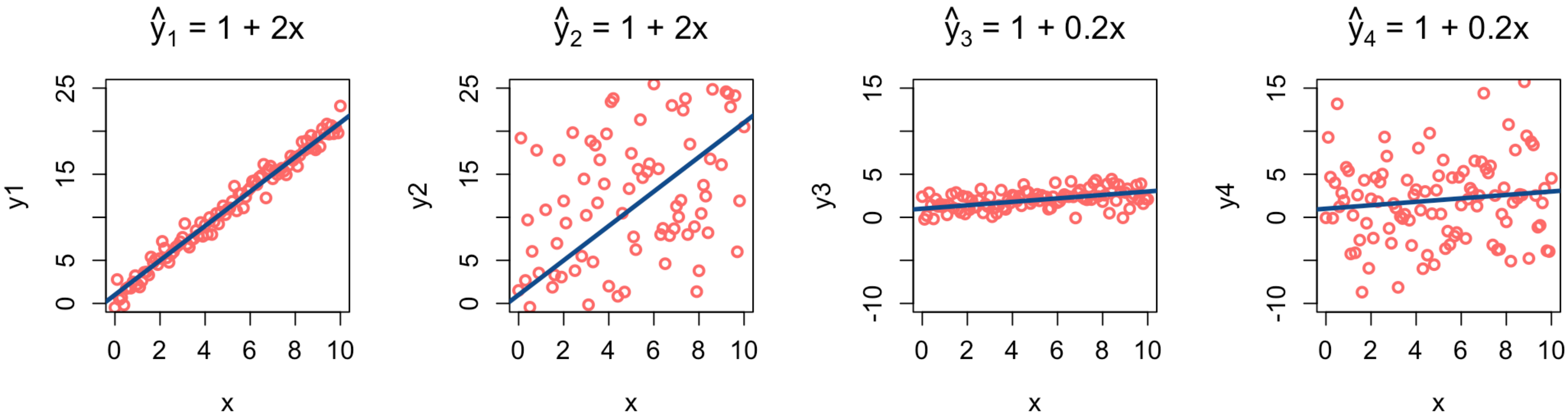
$$r^2 = \frac{\text{Variance}(\hat{y})}{\text{Variance}(y)}$$

  - 45% of the variation in *Height* is explained by the least squares regression line that involves *ShoeLength*.

Several regression lines



Several regression lines



Two important factors in making inference about regression lines:

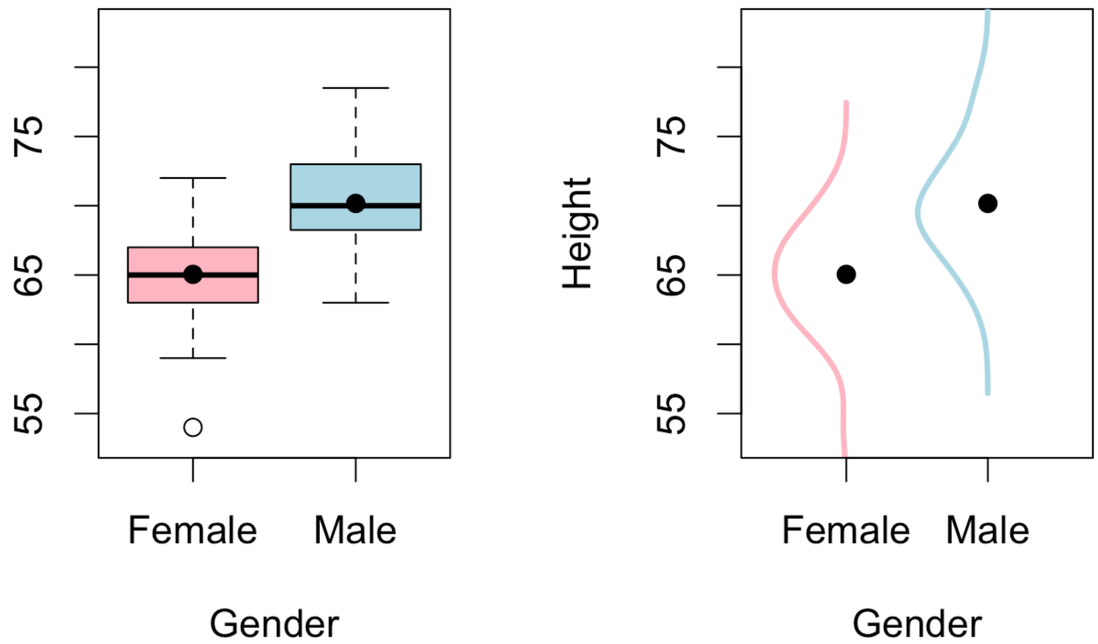
1. The variability of the residuals (how scattered the points are)
2. The slope of the regression line (how steep the trend is)

Which regression line do you think is the most significant one?

- $\hat{y}_1 = 1 + 2x$

Simple linear regression

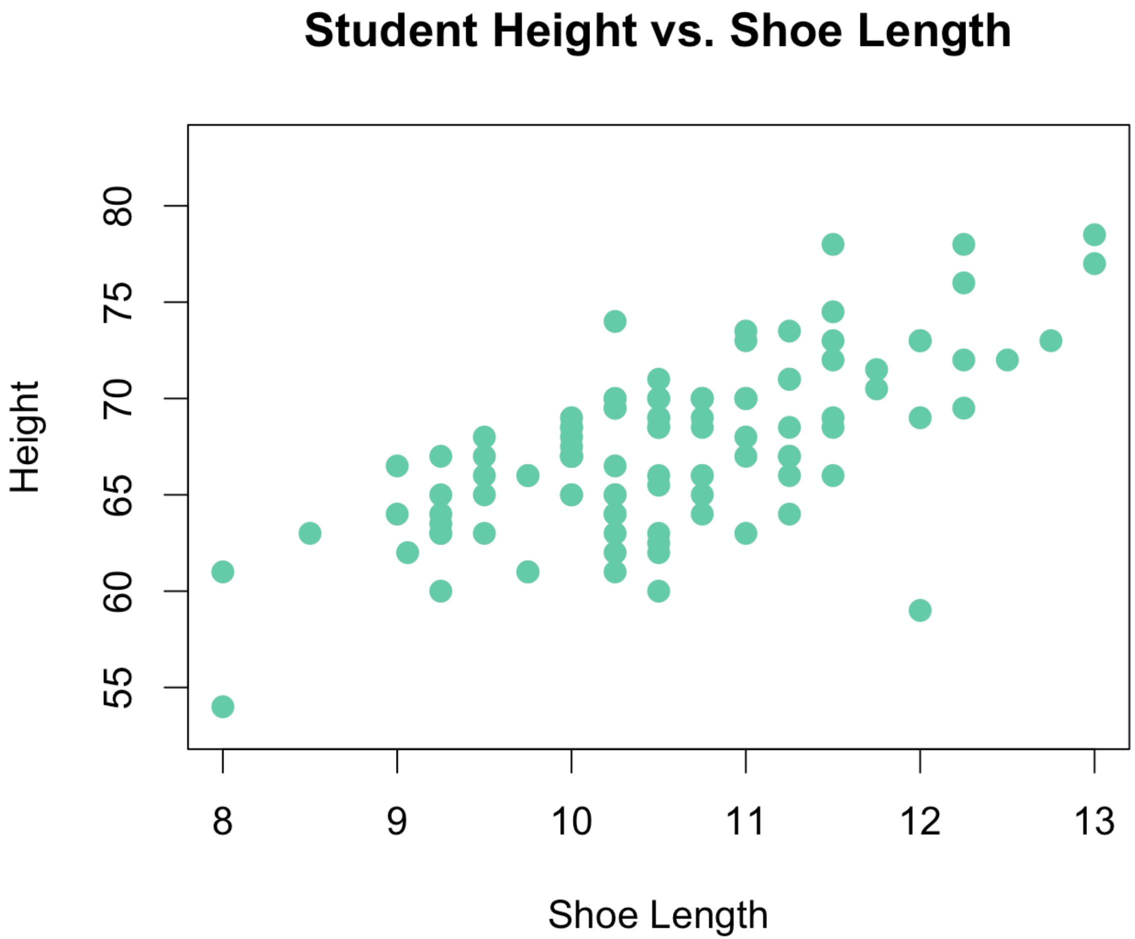
Quantitative versus binary: two-sample problem.



To study the relationship btw *Height* (quantitative) and *Gender* (binary), we compare the mean height of females with mean height of males.

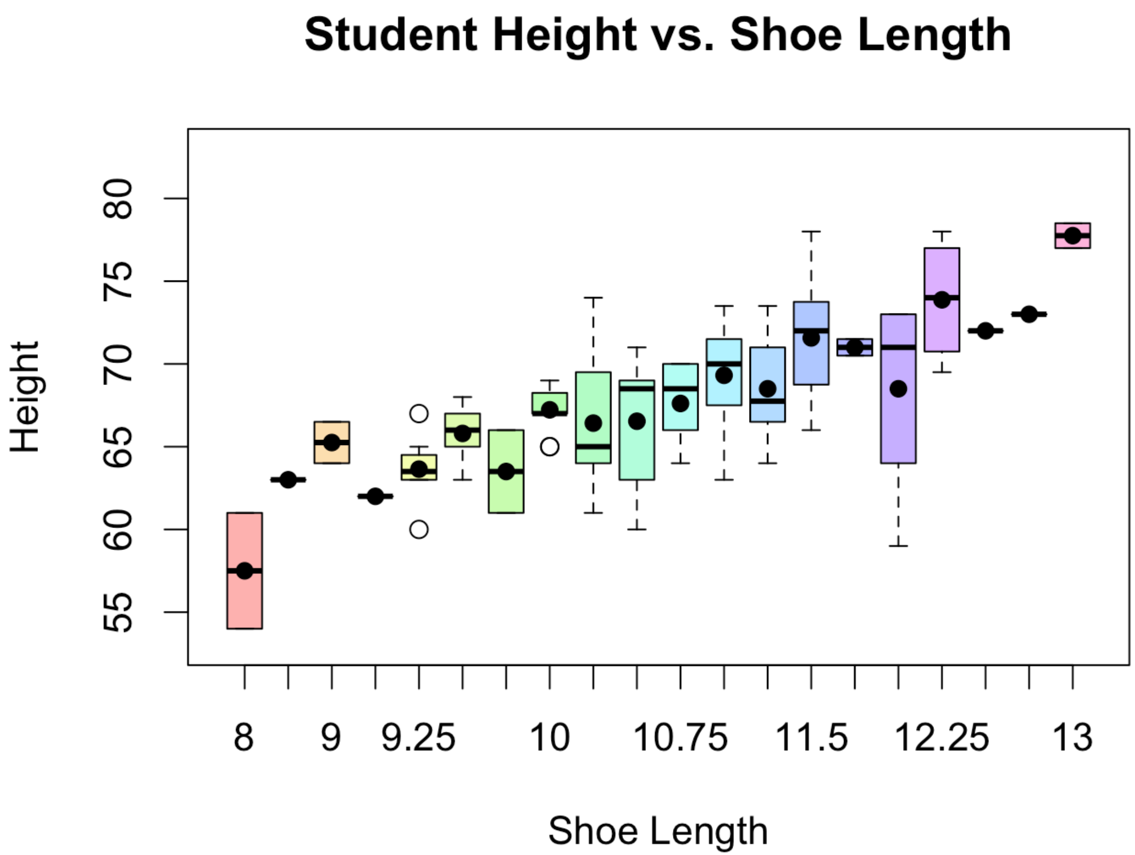
- Assume each sample to be Normally distributed.
- Use Normal distribution to make inference about the difference in means.
- By CLT, even the data is not Normal, when sample size is large, the inference methods still work well.

Simple linear regression



If we treat *ShoeLength* as a "categorical" variable, for each category (i.e., each possible value of *ShoeLength*), we may assume the *Height* values to be Normally distributed.

Simple linear regression



For **simple linear regression**, the **assumptions** are:

- For each *ShoeLength* value, *Height* follows a Normal distribution with mean μ and SD σ.
- Mean μ of *Height* is different for different values of *ShoeLength*.
- SD σ measures the variability of *Height* about the mean and is constant for different values of *ShoeLength*.

Simple linear regression - Model

- Denote μ_y as the mean of y for a given x and

$$\mu_y = \beta_0 + \beta_1 x$$

- Denote ϵ as the difference between the observed y and μ_y,

$$y = \mu_y + \epsilon$$

and

$$\epsilon \sim N(0, \sigma)$$

- Then $y = \mu_y + \epsilon \sim N(\mu_y, \sigma)$
  - y follows a Normal distribution with mean $\mu_y = \beta_0 + \beta_1 x$ and SD σ.
- Here μ_y, β_0, β_1 and σ are population parameters.

Simple linear regression - Model

$$\begin{array}{ccccc} y & = & \mu_y & + & \epsilon \\ \\ y & = & \beta_0 + \beta_1 x & + & \epsilon \\ \\ \text{Data} & = & \text{Fit} & + & \text{Residual} \end{array}$$

The data can be explain by two parts:

1. **Fit**: $\mu_y = \beta_0 + \beta_1 x$ is the **population regression line.** μ_y is the mean response at x.
2. **Residual**: ϵ is the variation of observed y about μ_y and $\epsilon \sim N(0, \sigma)$.
   - ϵ represents the "noise" around μ_y.

Simple linear regression - Model

Given n observations of the explanatory variable x and the response variable y,

$$(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$$

the **statistical model for simple linear regression** states that the observed response y when the explanatory variable takes the value x is

$$y = \beta_0 + \beta_1 x + \epsilon$$

Here $\beta_0 + \beta_1 x$ is the mean response at x. The deviations ϵ are assumed to be independent and distributed as $N(0, \sigma)$.

The **parameters of the model** are β_0, β_1 and σ.

Simple linear regression - Model inferences

	Intercept	Slope	SD	Mean response
Parameter	β_0	β_1	σ	μ_y
Statistic	b_0	b_1	s	$\hat{\mu}_y$

For simple linear regression, we are specifically interested in the inference for the slope β_1 because when $\beta_1 = 0$, it suggests no relationship between x and y (changing in x does not affect y).

- The estimator of β_1 is b_1, where b_1 is found using the least-squares method.
- Just like the inference for population mean μ is based on the distribution of $\bar{x}$, we also need to find the distribution of b_1 to make inference about β_1.

Simple linear regression - Model inferences

For $y = \beta_0 + \beta_1 x + \epsilon$, where $\epsilon \sim N(0, \sigma)$

$$b_1 = r\frac{s_y}{s_x} \qquad b_0 = \bar{y} - b_1\bar{x} \qquad s = \sqrt{\frac{\sum(y_i - b_0 - b_1 x_i)^2}{n-2}}$$

$$\sigma_{b_1} = \frac{\sigma}{\sqrt{\sum(x_i - \bar{x})^2}} \qquad \text{SE}_{b_1} = \frac{s}{\sqrt{\sum(x_i - \bar{x})^2}}$$

We have

$$\frac{b_1 - \beta_1}{\text{SE}_{b_1}} \overset{approx.}{\sim} t(n-2)$$

Simple linear regression - Model inferences

Comparing the inferences for population mean and slope

	Populaiton mean	Slope
Parameter of interest	μ	β_1
Estimate (statistic)	$\bar{x}$	b_1
Mean of estimate	μ	β_1
SD of estimate	$\frac{\sigma}{\sqrt{n}}$	$\frac{\sigma}{\sqrt{\sum(x_i - \bar{x})^2}}$
SE of estimate	$\frac{s}{\sqrt{n}}$	$\frac{s}{\sqrt{\sum(x_i - \bar{x})^2}}$
Distribution of test statistic	$t = \frac{\bar{x}-\mu}{\mathrm{SE}_{\bar{x}}} \overset{approx.}{\sim} t(n-1)$	$t = \frac{b_1-\beta_1}{\mathrm{SE}_{b_1}} \overset{approx.}{\sim} t(n-2)$

Inference for the regression line

A **level** C **confidence intervals for the intercept** β_0 **and slope** β_1 are

$$b_0 \pm t^* \mathrm{SE}_{b_0} \text{ and } b_1 \pm t^* \mathrm{SE}_{b_1}$$

In this expression t^* is the value for the $t(n-2)$ density curve with area C between $-t^*$ and t^*.

To test $H_0 : \beta_1 = 0$, compute the **test statistic**

$$t = \frac{b_1 - 0}{\mathrm{SE}_{b_1}} \overset{approx.}{\sim} t(n-2)$$

The **degrees of freedom** are $n-2$. In terms of a random variable T having the $t(n-2)$ distribution, the P-value for a test of H_0 against

$$H_a : \beta_1 > 0 \text{ is } P(T \geq t)$$

$$H_a : \beta_1 < 0 \text{ is } P(T \leq t)$$

$$H_a : \beta_1 \neq 0 \text{ is } 2P(T \geq |t|)$$

Inference for the regression line

```r
mymodel <- lm(Height ~ ShoeLength, data=Survey)
mymodel
```

```
##
## Call:
## lm(formula = Height ~ ShoeLength, data = Survey)
##
## Coefficients:
## (Intercept)   ShoeLength
##      36.476        2.939
```

"State the statistical model for simple linear regression":

Denote *Height* as y and *ShoeLength* as x, the simple linear regression model of *Height* on *ShoeLength* is $y = \beta_0 + \beta_1 x + \epsilon$, where $\epsilon \sim N(0, \sigma)$.

"Write down the estimated regression line":

$\hat{y} = 36.5 + 2.9x$

Inference for the regression line

```r
summary(mymodel)
```

```
## Call:
## lm(formula = Height ~ ShoeLength, data = Survey)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.7443  -1.8357   0.8686   1.9295   7.7253
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  36.4759     3.3775  10.800  < 2e-16 ***
## ShoeLength    2.9390     0.3182   9.236 3.69e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.251 on 103 degrees of freedom
## Multiple R-squared:  0.453,  Adjusted R-squared:  0.4477
## F-statistic: 85.31 on 1 and 103 DF,  p-value: 3.688e-15
```

$b_0 = 36.5, \mathrm{SE}_{b_0} = 3.4$

$b_1 = 2.9, \mathrm{SE}_{b_1} = 0.3$

$s = 3.3$

$\mathrm{df} = 103 = 105 - 2$

$r^2 = 0.453$

Confidence intervals for intercept and slope

```r
summary(mymodel)
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  36.4759     3.3775  10.800  < 2e-16 ***
## ShoeLength    2.9390     0.3182   9.236 3.69e-15 ***
```

```r
qt(0.975, df=103) # t*
```

```
## [1] 1.983264
```

$b_0 = 36.5, \mathrm{SE}_{b_0} = 3.4, b_1 = 2.9, \mathrm{SE}_{b_1} = 0.3$

- **95% confidence intervals** for β_0:

  $b_0 \pm t^* \mathrm{SE}_{b_0} = 36.5 \pm 1.98 \times 3.4 = 36.5 \pm 6.7$
  We are 95% confident that the true population intercept is btw 29.8 and 43.2.

- **95% confidence intervals** for β_1:

  $b_1 \pm t^* \mathrm{SE}_{b_1} = 2.9 \pm 1.98 \times 0.3 = 2.9 \pm 0.6$
  We are 95% confident that the true population slope is btw 2.3 and 3.5.

Confidence intervals for intercept and slope

```r
confint(mymodel)
```

```
##                  2.5 %    97.5 %
## (Intercept) 29.777456 43.174326
## ShoeLength   2.307946  3.570116
```

```r
confint(mymodel, level=0.99)
```

```
##                  0.5 %    99.5 %
## (Intercept) 27.612010 45.339772
## ShoeLength   2.103931  3.774131
```

The R function `confint.lm()` calculates the 95% confidence intervals for linear regression models by default.

Significance test for the slope

```r
summary(mymodel)
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   36.4759     3.3775  10.800  < 2e-16 ***
## ShoeLength     2.9390     0.3182   9.236 3.69e-15 ***
```

$b_1 = 2.9$, $\text{SE}_{b_1} = 0.3$. Test $H_0 : \beta_1 = 0$ vesus $H_a : \beta_1 \neq 0$

- $t = 9.236$ or $t = \frac{b_1 - 0}{\text{SE}_{b_1}} = \frac{2.9}{0.3}$
- $P = 2P(T \geq |t|) =$ `2*(1-pt(9.236, df=103))` $3.7 \times 10^{-15} < 0.05$
- **Conclusion**: We reject H_0 at level 0.05. There is a highly signifiantly linear relationship between *Height* and *ShoeLength*.
- **Note**: In simple linear regression, usually we do NOT test about the intercept. R by default tests $\beta_0 = 0$ and returns a P-value. But it does not have any practical meaning most of the time.

Conducting simple linear regression analysis

Steps:

1. State the statistical model for simple linear regression
2. Do exploratory data analysis: scatterplot and correlation
3. Obtain the least-squares regression line and add the line to the scatterplot
4. Check assumptions (Lecture 23)
   - If assumptions are violated, try transformation (Lecture 24)
5. Assess the fitting of the model: r^2
6. Make inferences:
   - Confidence intervals of both intercept and slope
   - Significance test for the slope
7. Predictions (Lecture 23):
   - Mean response and its confidence interval
   - Individual response and its Prediction interval

Summary

- Least-squares regression review
  - Scatterplot and correlation
  - Least-squares regression
  - Assessing the regression line: residual plot and r^2
- Simple linear regression
  - Idea
  - Model $y = \mu_y + \epsilon = \beta_0 + \beta_1 x + \epsilon$ where $\epsilon \sim N(0, \sigma)$
- Inference for the regression line

  Confidence intervals $b_0 \pm t^* \mathrm{SE}_{b_0}$ and $b_1 \pm t^* \mathrm{SE}_{b_1}$

  Significance test $t = \frac{b_1 - 0}{\mathrm{SE}_{b_1}} \overset{approx.}{\sim} t(n-2)$