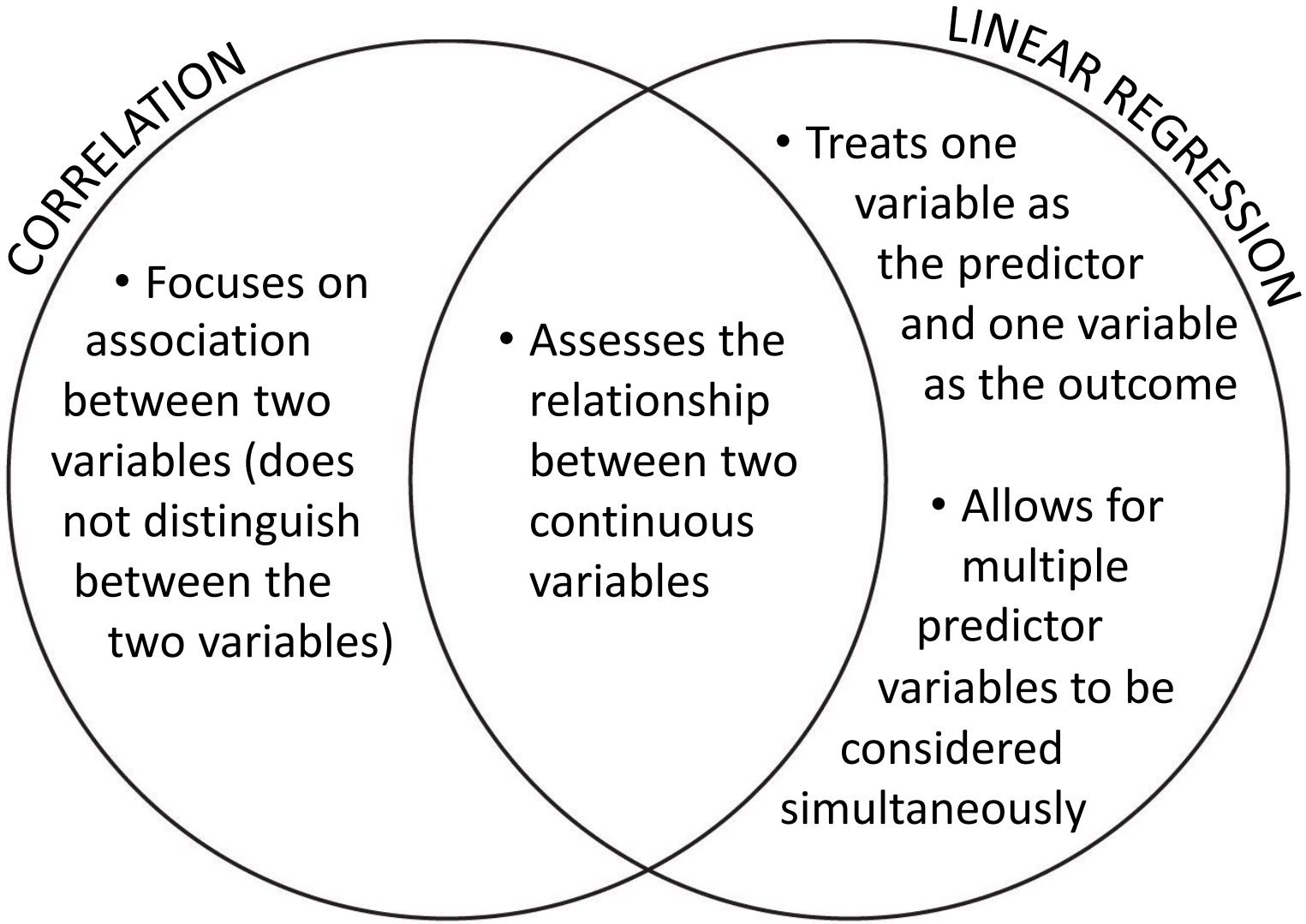
Continuous data – simple linear regression

Types of Variables/Number of Groups

Variable(s)	Analysis
Continuous	One-group t-test
Continuous vs. Categorical (2 categories)	Two-group t-test
Continuous vs. Categorical (>2 categories)	ANOVA
Continuous vs. Continuous	Correlation/linear regression

Correlation vs. Linear Regression



Predictor and Outcome Variables

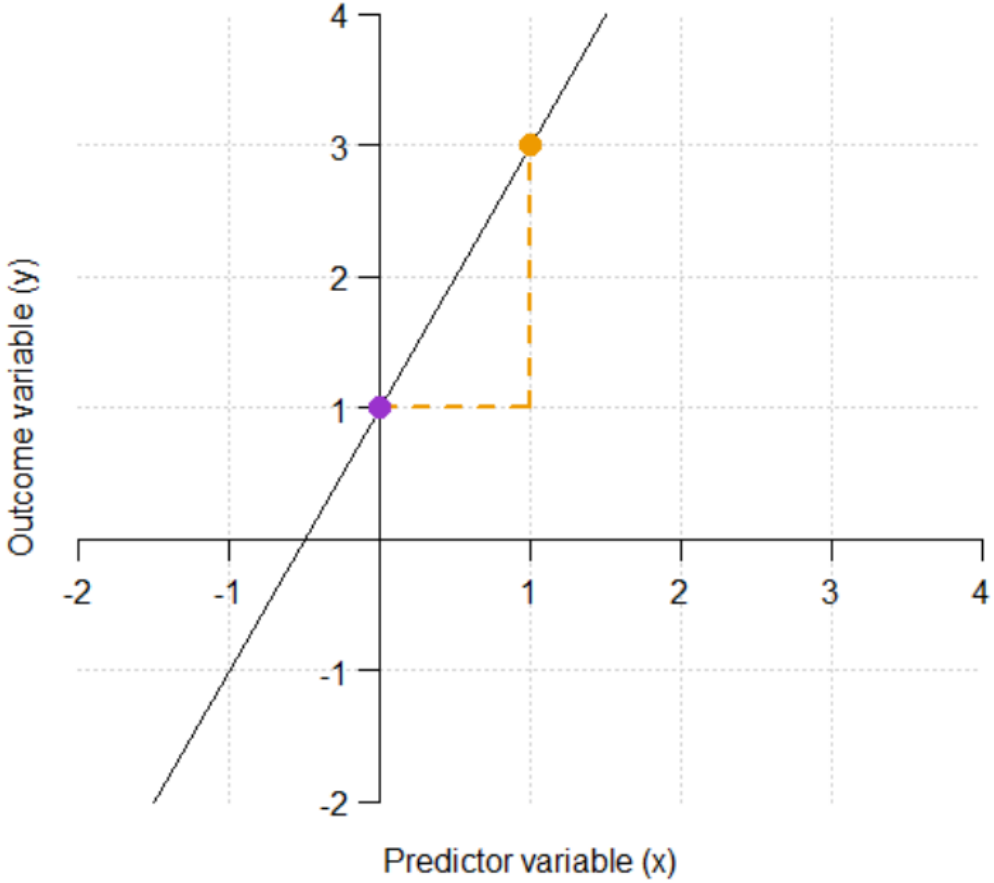
- To examine the association between two continuous variables with linear regression, you must identify which variable is the predictor and which is the outcome
- Predictor = independent variable
  - Plotted on x axis
- Outcome = dependent variable
  - Plotted on y axis

Simple Linear Regression

- Predictor variable = x
  Outcome variable = y
- Goal: Find the straight line that best fits the points
- Line is characterized by the **slope** and the **intercept**
  - Intercept = the value of y when x=0
  - Slope = number of units y increases for every 1 unit increase in x

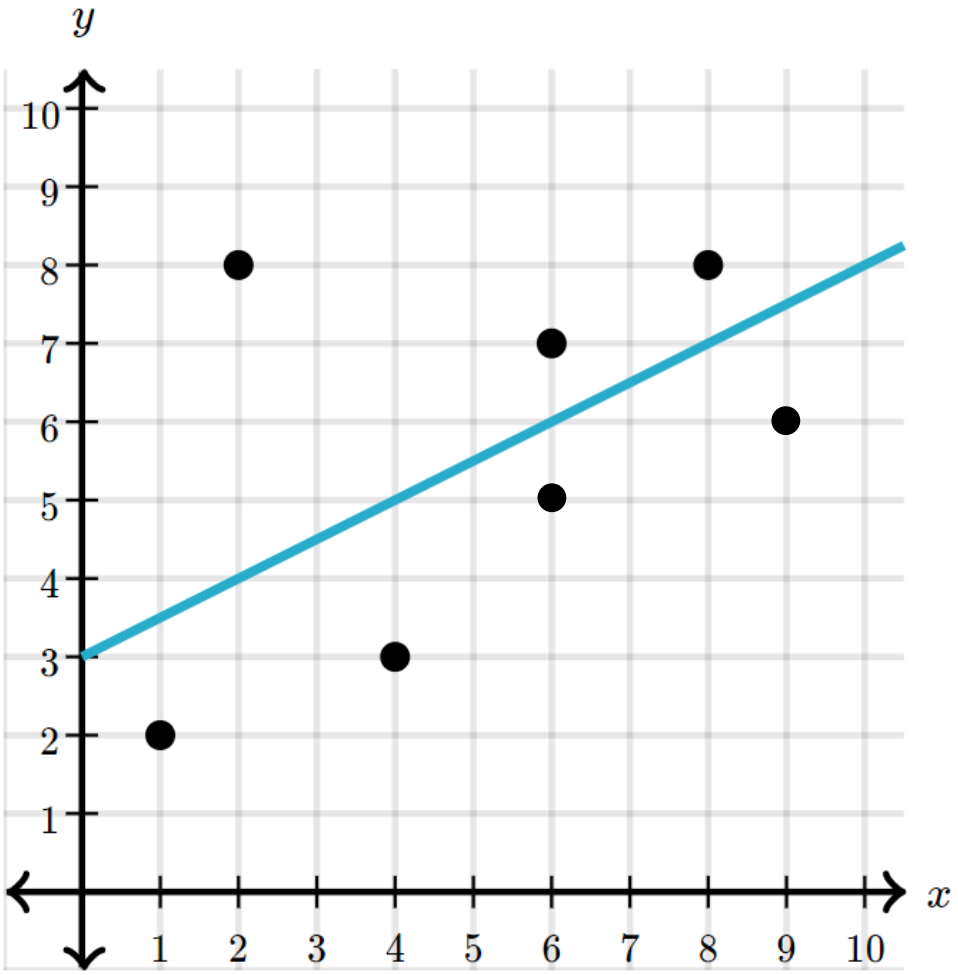
$$y = \text{intercept} + (\text{slope} \times x)$$

Example: $y = 1 + 2x$

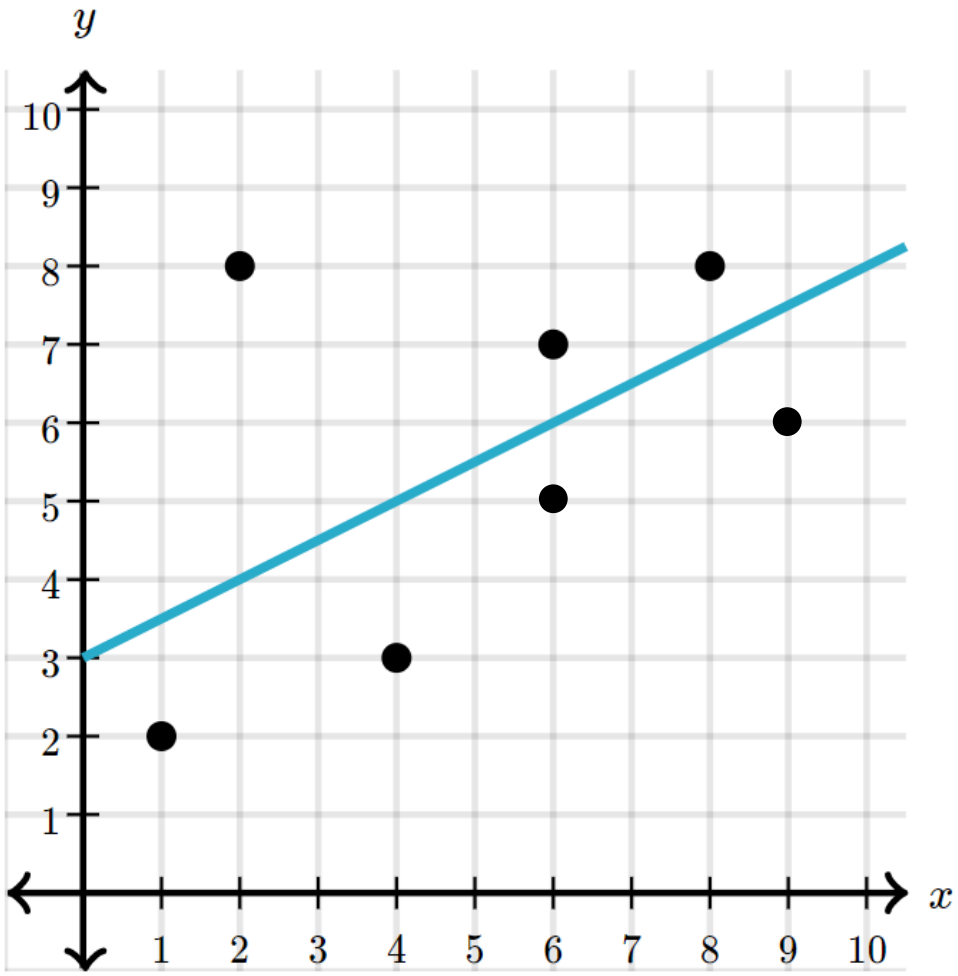


Regression Line and Residuals

- Statistical software finds the “best fitting” line by minimizing the vertical distance between each point and the line
  - Called the **least squares** method
- The vertical distances between the points and the line are called **residuals**



Making Predictions



- Regression line:

$$y = 3 + 0.5x$$

- Plug in value of x to the equation to get the predicted value of y
  - Example: For someone with x=3, what is our best prediction for their value of y?

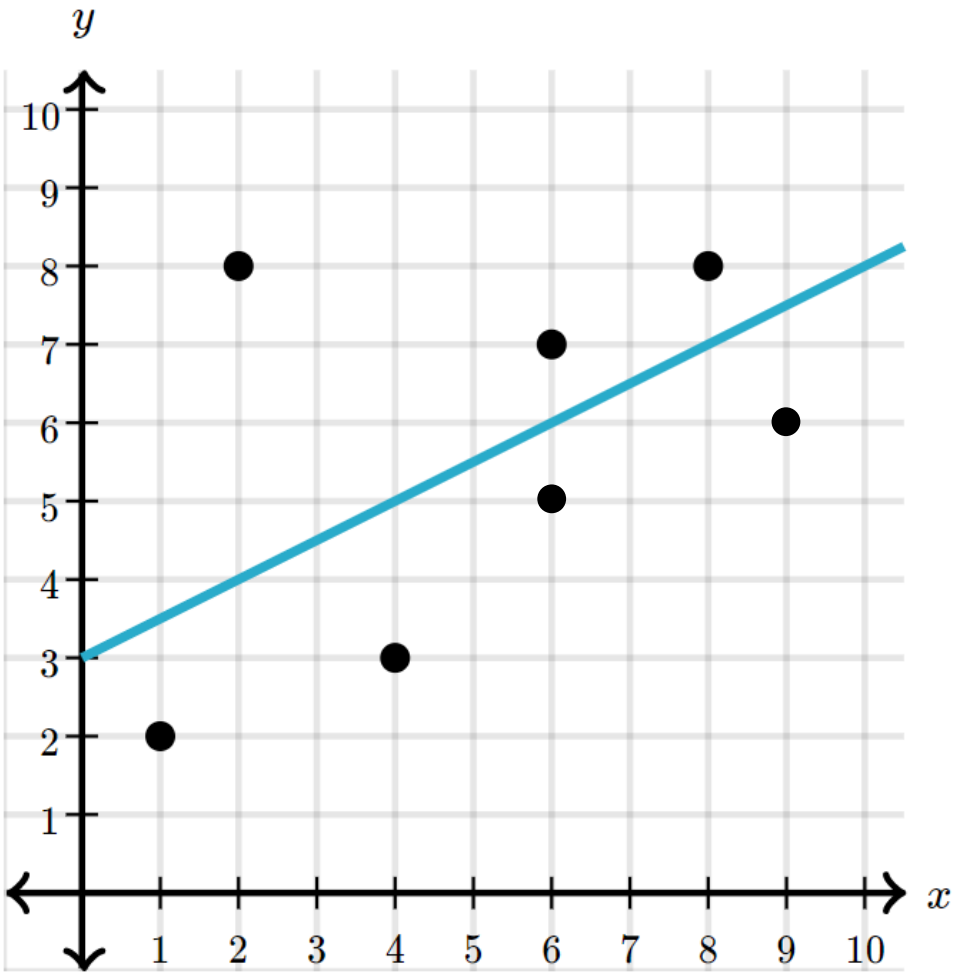
Making Predictions

- The predictions for subjects in our dataset are called **fitted values**

Data:

x	y	fitted value
1	2	
2	8	
4	3	
6	7	
6	5	
8	8	
9	6	

Regression line: $y = 3 + 0.5x$



Vocabulary and Notation

- Each subject has an observed value, a fitted value, and a residual

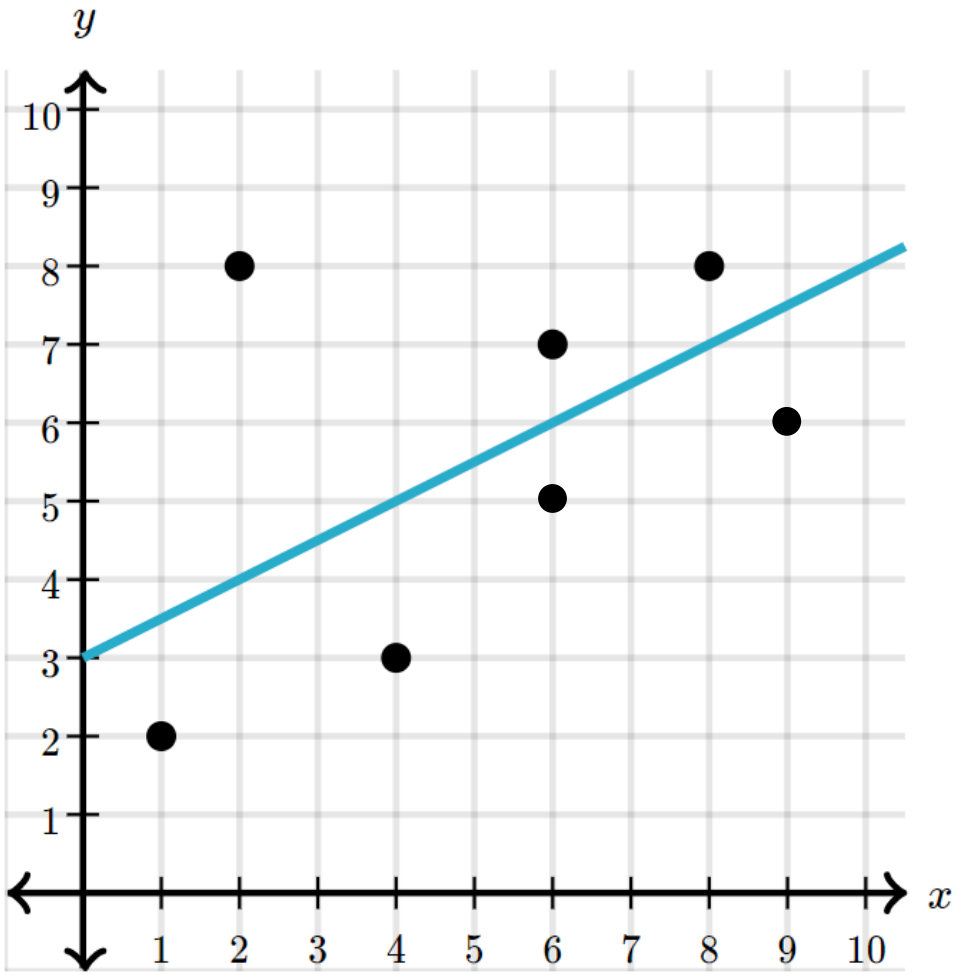
- Observed value = y
  - The actual observed value of the outcome variable in the dataset
- Fitted value = $\hat{y}$
  - The predicted value of the outcome variable (on the line)
- Residual = e
  - The vertical distance between the point and the regression line
  - Difference between y and $\hat{y}$

Vocabulary and Notation

- Proper notation for regression line:

$$\hat{y} = 3 + 0.5x \quad \text{or} \quad y = 3 + 0.5x + e$$

x	y	$\hat{y}$	e
1	2	3.5	-1.5
2	8	4	4
4	3	5	-2
6	7	6	1
6	5	6	-1
8	8	7	1
9	6	7.5	-1.5

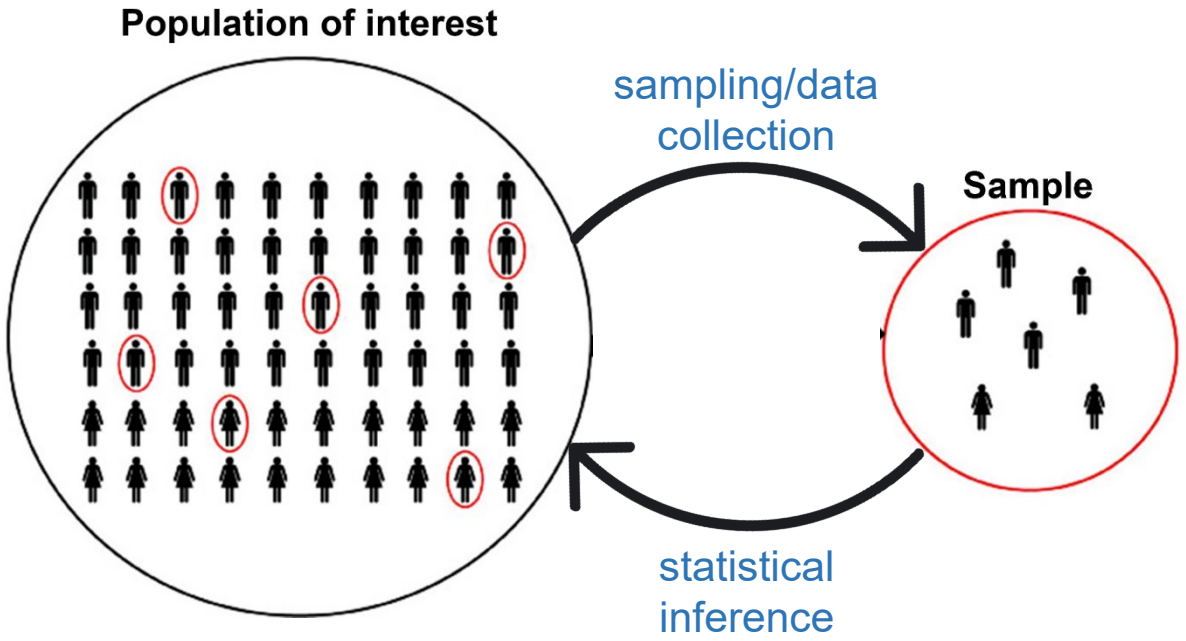


Vocabulary and Notation

Slope and intercept are called **coefficients**

- β_0 = true (population) intercept
- β_1 = true (population) slope
- $\hat{\beta}_0$ = estimated (sample) intercept
- $\hat{\beta}_1$ = estimated (sample) slope

β_0 and β_1 are fixed



$\hat{\beta}_0$ and $\hat{\beta}_1$ vary (would be different with a different sample of the population)

Intercept (β_0)

- Tells you the average value of y when x=0
  - Oftentimes it doesn't make sense
  - Usually not our focus – only needed to make predictions
- Interpretation of intercept:
  - "When (predictor variable) is 0 units, (outcome variable) is (intercept) units, on average."
  - "When (predictor variable) is 0 units, we expect (outcome variable) to be (intercept) units."

Important: Must say either "on average" or "we expect" to indicate that the observations aren't exactly on the line.

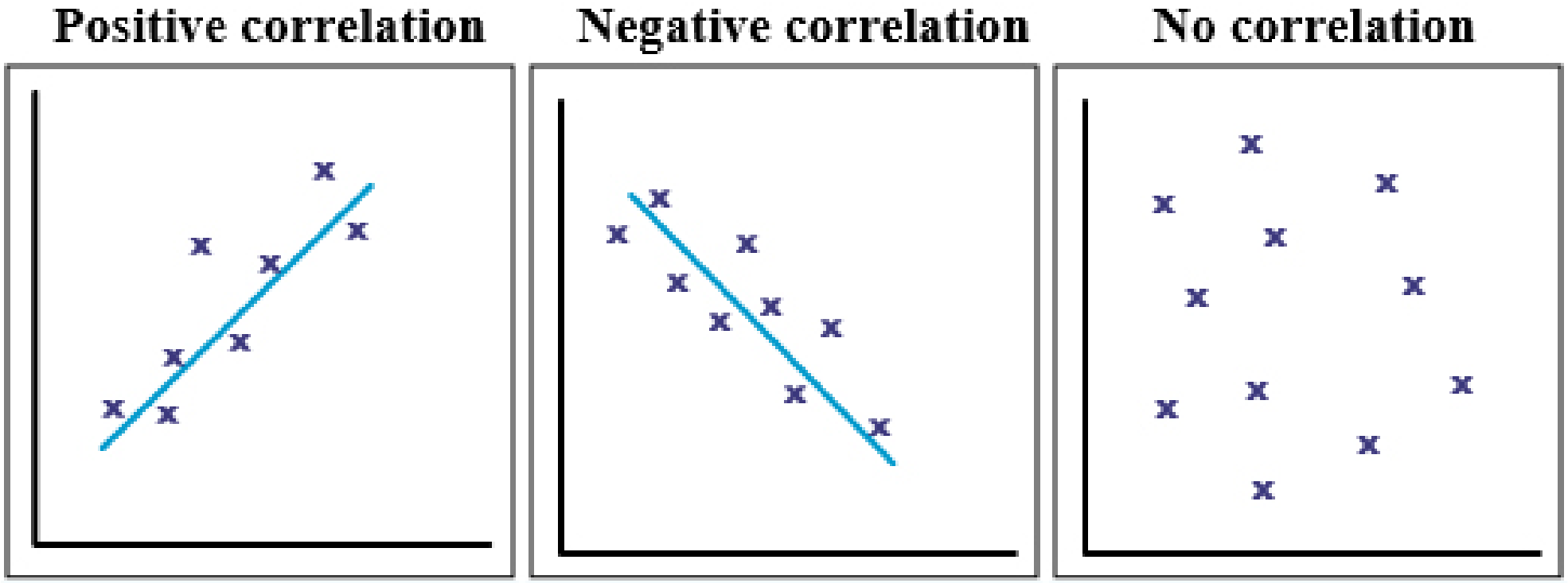
Slope (β_1)

- In regression, our interest is usually in the slope
  - Slope characterizes the relationship between the predictor and outcome variables

- Interpretation of slope:
  - "On average, every 1 unit increase in (predictor variable) is associated with a (slope) unit increase in (outcome variable)."
  - "For every 1 unit increase in (predictor variable), we expect a (slope) unit increase in (outcome variable)."

Important: Avoid causation language, and must say either "on average" or "we expect" to indicate that the observations aren't exactly on the line.

Slope (β_1)



$\beta_1 > 0$ $\beta_1 < 0$ $\beta_1 = 0$

Low Birth Weight Data

- Information on 100 low birth weight infants born in two teaching hospitals in Boston, Massachusetts

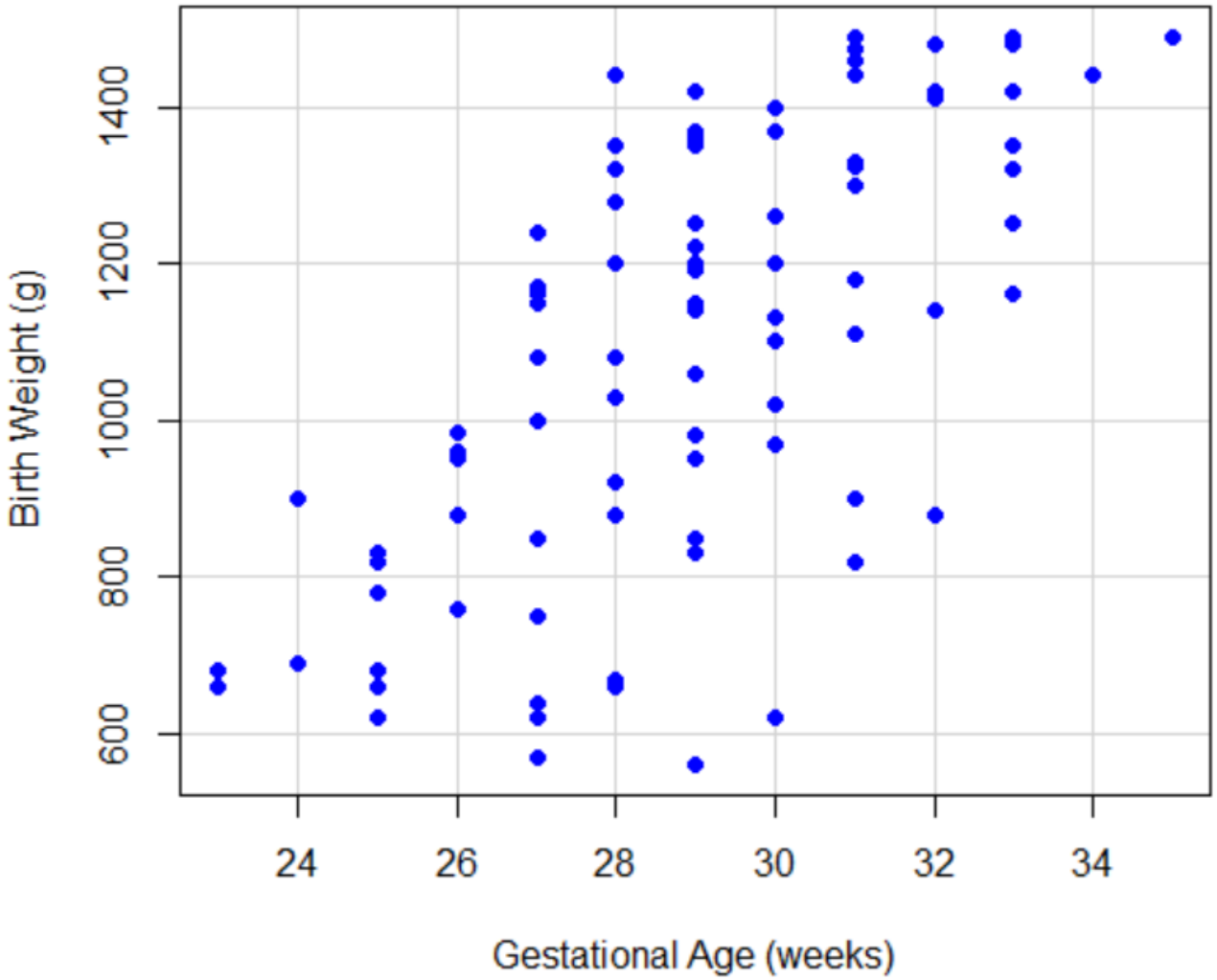
Variable	Description
gestage	Gestational age at time of birth (weeks)
momage	Mother's age (years)
birthwt	Birth weight of the baby (g)
length	Length of the baby (cm)
headcirc	Baby's head circumference (cm)
apgar	Apgar score (integers, min=0, max=10). This is a scoring system used for assessing the clinical status of a newborn. 7 or higher is generally considered normal, 4-6 is low, and 3 or below is critically low.

Find the dataset (lowbwt.xlsx) and the full data dictionary (lowbwt Data Dictionary.pdf) in the Data Module on the Canvas site

Example: Gestational Age/Birth Weight

- Fit a linear regression model to assess the association between gestational age and birth weight in low birth weight infants.
  - Interpret the slope and intercept.
  - What is the predicted birth weight for an infant with a gestational age of 28 weeks?

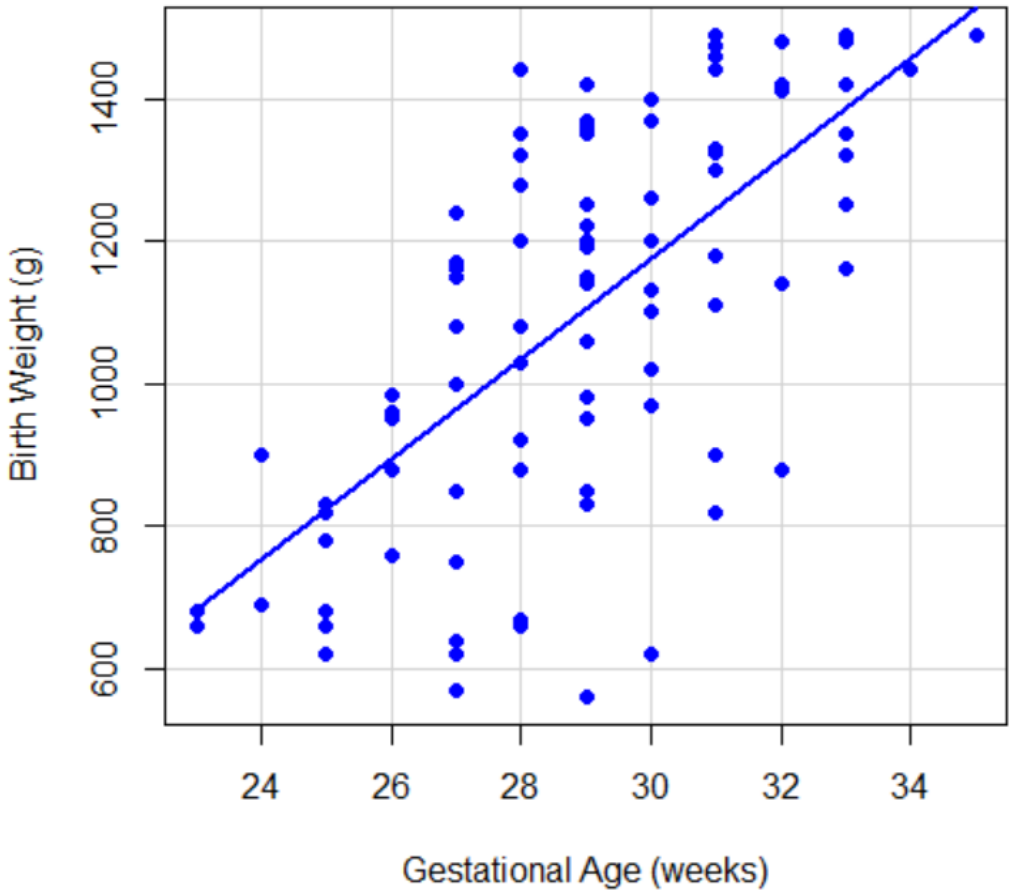
Example: Gestational Age/Birth Weight



Example: Gestational Age/Birth Weight

Linear regression model relating gestational age and birth weight:

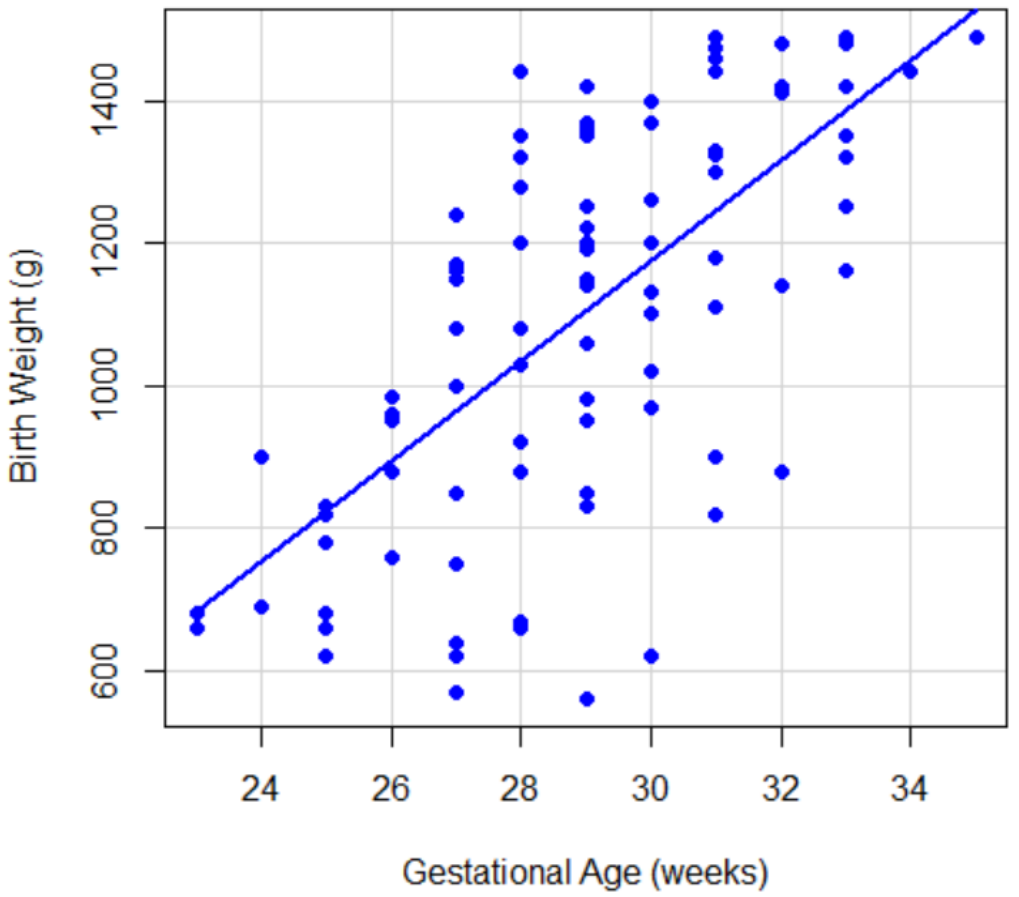
```
Call:
lm(formula = birthwt ~ gestage, data = lowbwt)

Residuals:
   Min     1Q Median     3Q    Max
-556.9 -136.0   32.8  149.3  403.7

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -932.404    234.488  -3.976 0.000134 ***
gestage       70.310      8.086   8.695 8.15e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 203.9 on 98 degrees of freedom
Multiple R-squared:  0.4355, Adjusted R-squared:  0.4298
F-statistic: 75.61 on 1 and 98 DF,  p-value: 8.149e-14
```



Example: Gestational Age/Birth Weight

- Interpretation of intercept:
  - When gestational age is 0 weeks, birth weight is -932.4 grams, on average.
- Interpretation of slope:
  - On average, every 1 week increase in gestational age is associated with a 70.3 grams increase in birth weight.

$$\hat{y} = -932.4 + 70.3x$$



Example: Gestational Age/Birth Weight

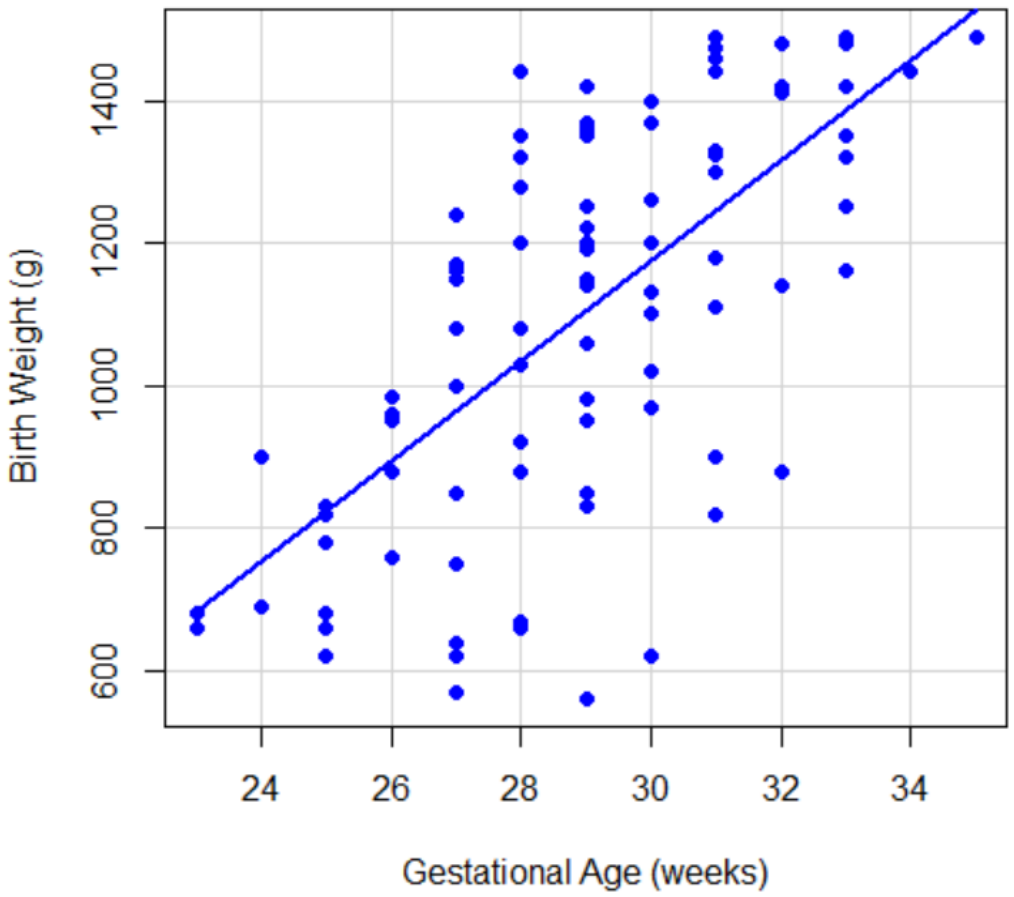
- Predicted birth weight for an infant with a gestational age of 28 weeks:

$$\hat{y} = -932.4 + 70.3x$$

$$\hat{y} = -932.4 + 70.3(28)$$

$$\hat{y} = 1036 \text{ grams}$$

$$\hat{y} = -932.4 + 70.3x$$



Estimation and Testing

- Goal: Estimate the association between two variables in the population (β_1) and determine if the association is significant.
  - Best estimate of the true slope is the slope in the sample ($\hat{\beta}_1$)
  - Can calculate confidence intervals for the slope
    - "We are 95% confident that the true slope relating _____ to _____ is between _____ and _____."
  - Slope test: hypothesis test to show if association is significant

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

Reject H_0 when p-value ≤ α
Fail to reject H_0 when p-value > α

P-value will be exactly the same as you would get if you did a correlation test!

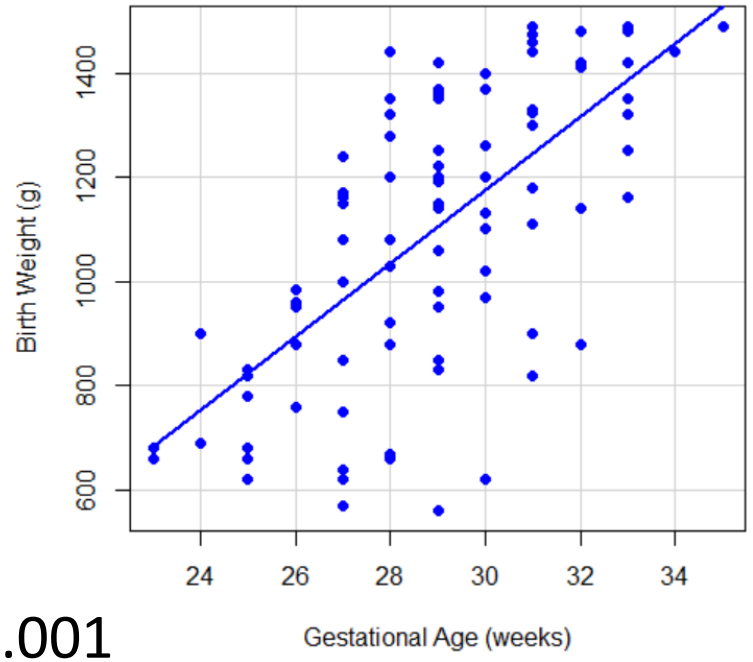
Example: Gestational Age/Birth Weight

- Is there a significant association between gestational age and birth weight in low birth weight infants?

Example: Gestational Age/Birth Weight

Linear regression model relating gestational age and birth weight:

```
Call:
lm(formula = birthwt ~ gestage, data = lowbwt)

Residuals:
   Min     1Q Median     3Q    Max
-556.9 -136.0   32.8  149.3  403.7

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -932.404    234.488  -3.976 0.000134 ***
gestage       70.310      8.086   8.695 8.15e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 203.9 on 98 degrees of freedom
Multiple R-squared:  0.4355, Adjusted R-squared:  0.4298
F-statistic: 75.61 on 1 and 98 DF,  p-value: 8.149e-14
```



p-value < 0.001

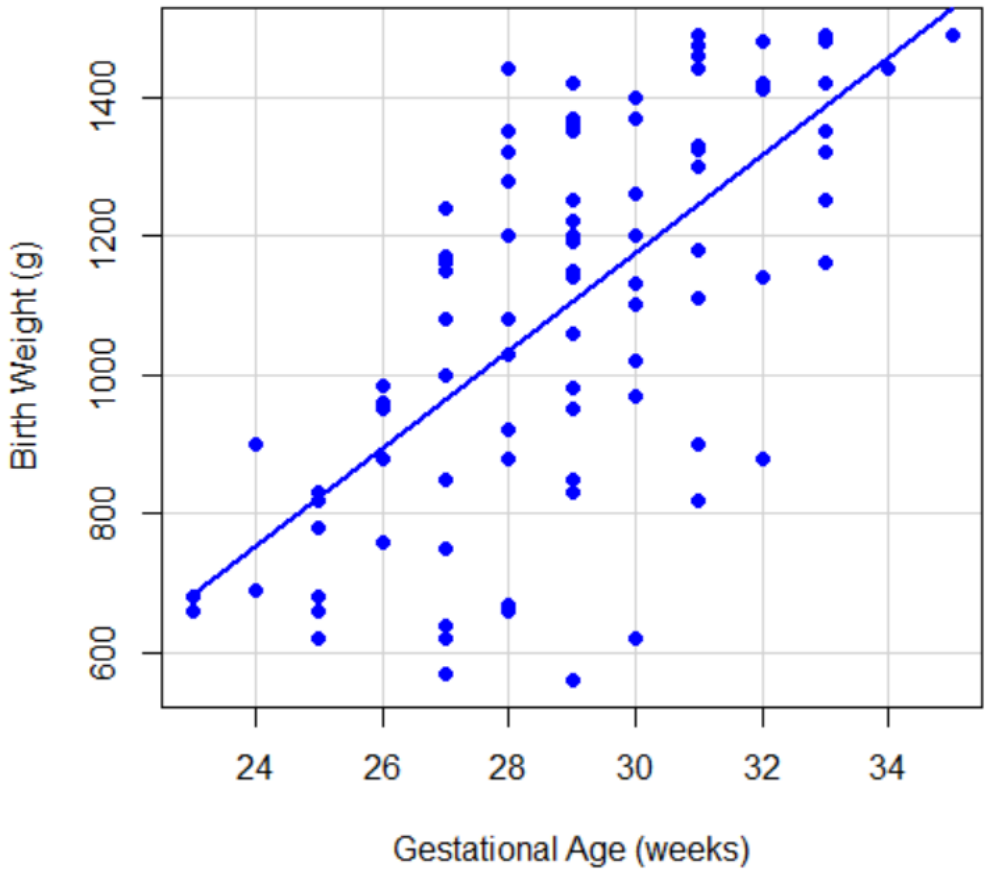
Since the p-value is less than 0.05, we reject H_0 and conclude that there is sufficient evidence to suggest that there is an association between gestational age and birth weight in low birth weight infants.

Example: Gestational Age/Birth Weight

95% confidence intervals for the coefficients from the linear regression model relating gestational age and birth weight:

	Estimate	2.5 %	97.5 %
(Intercept)	-932.40386	-1397.73841	-467.06931
gestage	70.30993	54.26382	86.35604

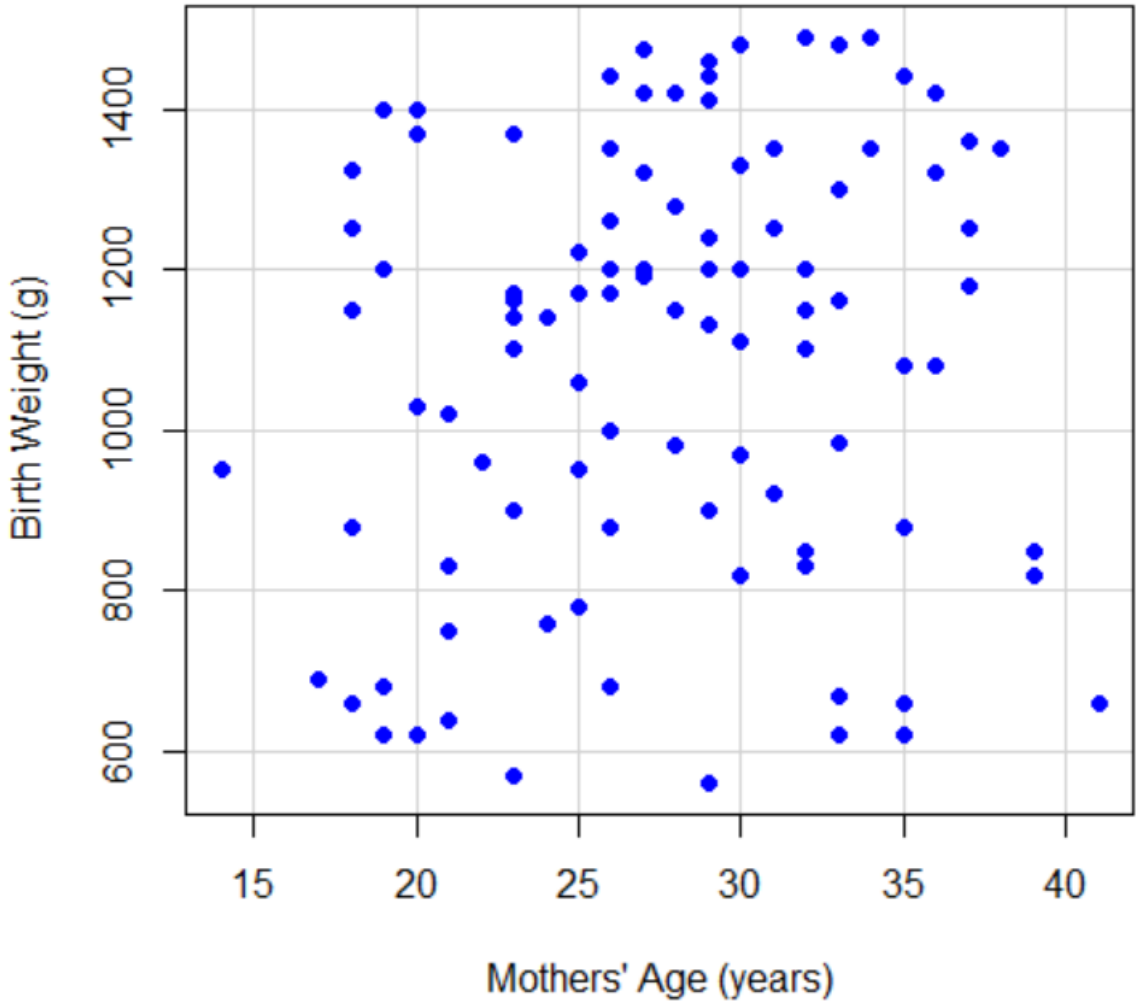
We are 95% confident that the true slope relating gestational age to birth weight is between 54.3 grams/week and 86.4 grams/week.



Example: Mothers' Age/Birth Weight

- Fit a linear regression model to estimate the association between the age of the mother and the birth weight of the infant in low birth weight infants.
- Is there a significant association between mothers' age and infants' birth weight in low birth weight infants?

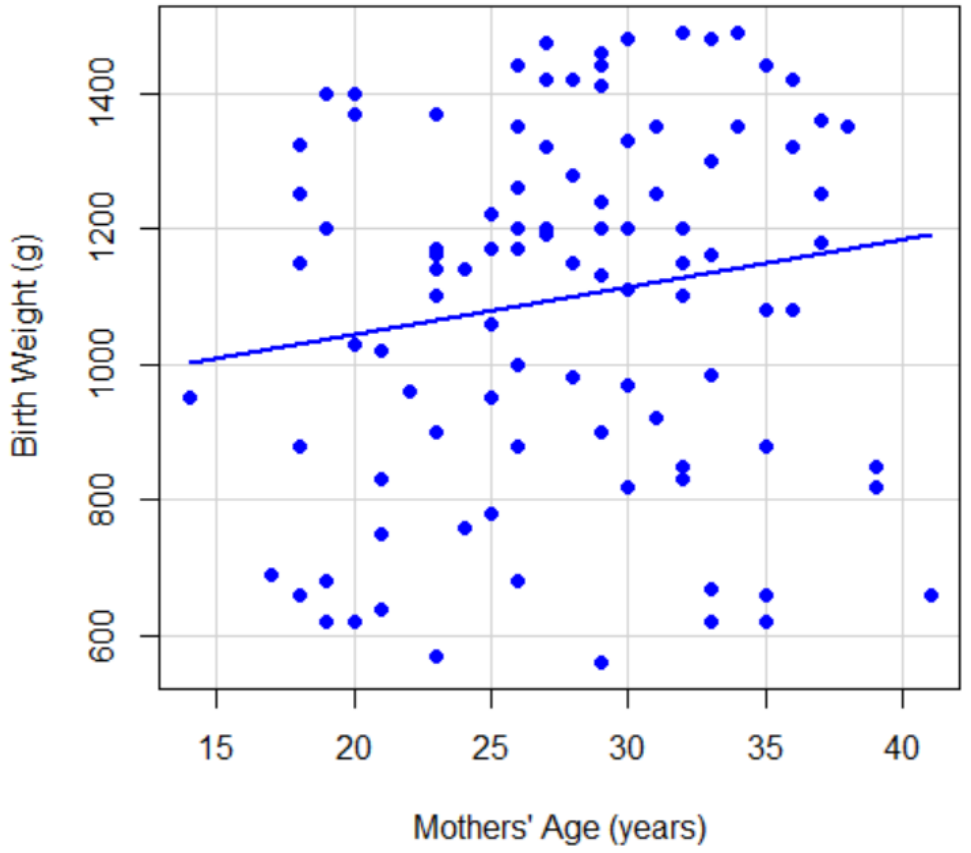
Example: Mothers' Age/Birth Weight



Example: Mothers' Age/Birth Weight

Linear regression model relating mothers' age and infants' birth weight:

```
Call:
lm(formula = birthwt ~ momage, data = lowbwt)

Residuals:
    Min      1Q  Median      3Q     Max
-547.71 -207.01   69.27  209.39  381.24

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  905.421    127.735   7.088 2.11e-10 ***
momage         6.975      4.504   1.549    0.125
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 268.1 on 98 degrees of freedom
Multiple R-squared:  0.02389, Adjusted R-squared:  0.01393
F-statistic: 2.399 on 1 and 98 DF,  p-value: 0.1247
```



$$\hat{y} = 905.42 + 6.98x$$

Example: Mothers' Age/Birth Weight

Linear regression model relating mothers' age and infants' birth weight:

```
Call:
lm(formula = birthwt ~ momage, data = lowbwt)

Residuals:
    Min      1Q  Median      3Q     Max
-547.71 -207.01   69.27  209.39  381.24

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  905.421    127.735   7.088 2.11e-10 ***
momage         6.975      4.504   1.549    0.125
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 268.1 on 98 degrees of freedom
Multiple R-squared:  0.02389, Adjusted R-squared:  0.01393
F-statistic: 2.399 on 1 and 98 DF,  p-value: 0.1247
```

$\hat{\beta}_1 = 6.98$ p-value = 0.125

On average, every 1 year increase in mothers' age is associated with a 6.98 grams increase in infants' birth weight.

Since the p-value is greater than 0.05, we fail to reject the null hypothesis and conclude that there is not sufficient evidence to say that there is a significant association between mothers' age and infants' birth weight.

Assumptions of Linear Regression

- **Independence** of the observations
- **Linearity** of the relationship between the two variables
- **Constant variance** of the residuals
- **Normality** of the residuals

Stay tuned for regression diagnostics!

Important Points

- How the regression line is determined and how it is characterized (slope and intercept)
- Interpretations of slope and intercept
- Concepts of fitted values and residuals
- Making predictions from a regression model
- Hypothesis test for the slope (set up and interpretation)