Week 10

Inference for Linear Regression

ST 314
Introduction to Statistics for Engineers

Oregon State University

Estimating Parameters

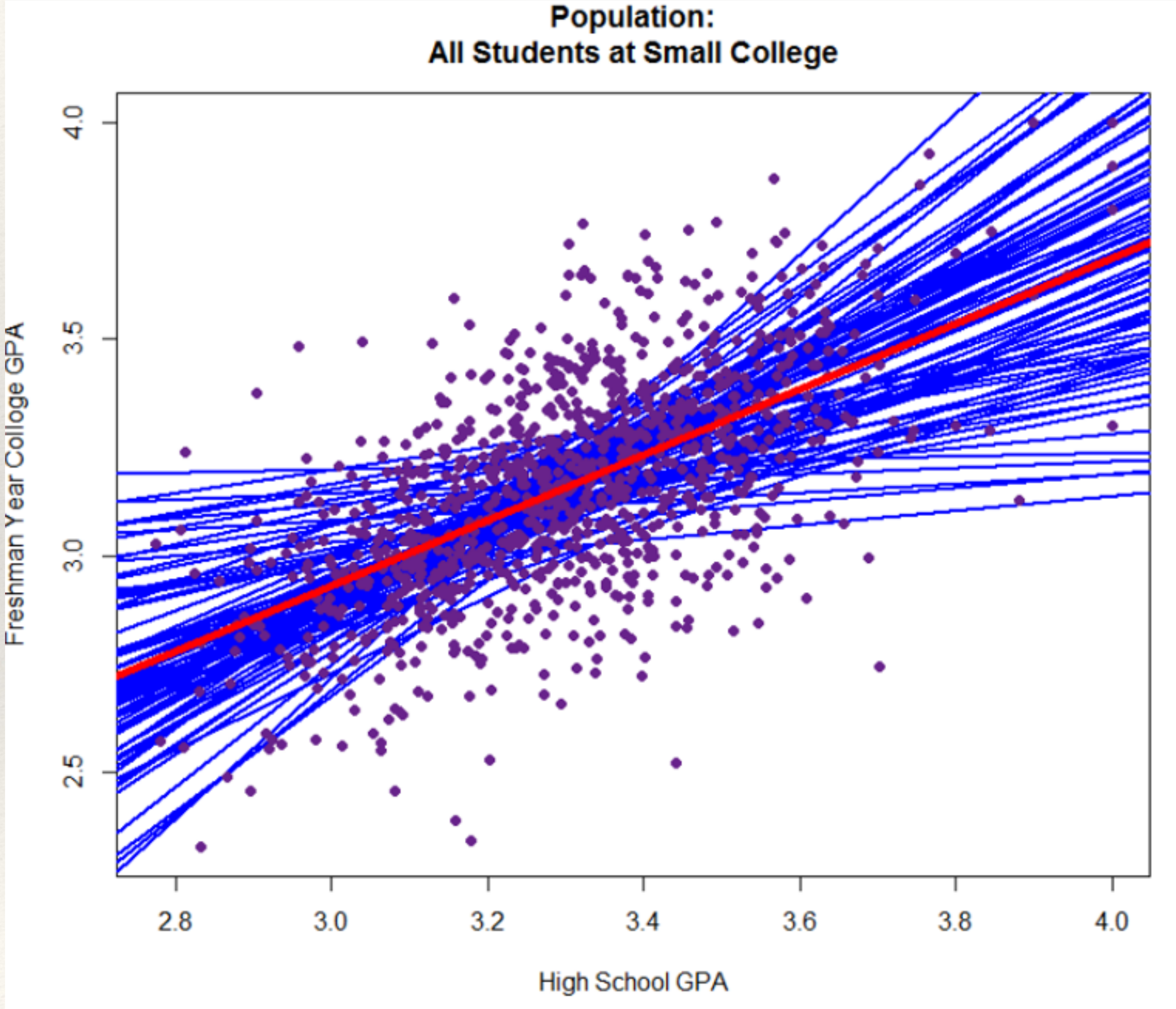
- The least squares regression line (LSRL) is based on sampled data. $\hat{y} = b_0 + b_1 x$ is the estimate for the true population regression line $\mu_y = \beta_0 + \beta_1 x$

  average response for given x value

- b_1 is the point estimate for $\underline{\beta_1}$.

- Random sampling implies that obtaining a different sample will obtain a different estimate for the LSRL.

Estimating Parameters



- b_0 and b_1 are random variables.

- b_0 and b_1 are unbiased point estimates for β_0 and β_1, respectively, and come with a certain amount of variability.

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

t Confidence Interval for β_1

- When to use: want to estimate the slope β_1 of the population regression equation
- Conditions required for inference:
  - sample must be representative of the population (achieved by random sampling)
  - Residuals indicated that the 4 regression conditions are met: linearity, normality in response, constant variance, independence
- Confidence interval for the slope of the population, β_1:

$$b_1 \pm t^*_{n-(k+1)} SE_{b_1}$$

sample size

# of explanatory variables in the model

FYI in SLR

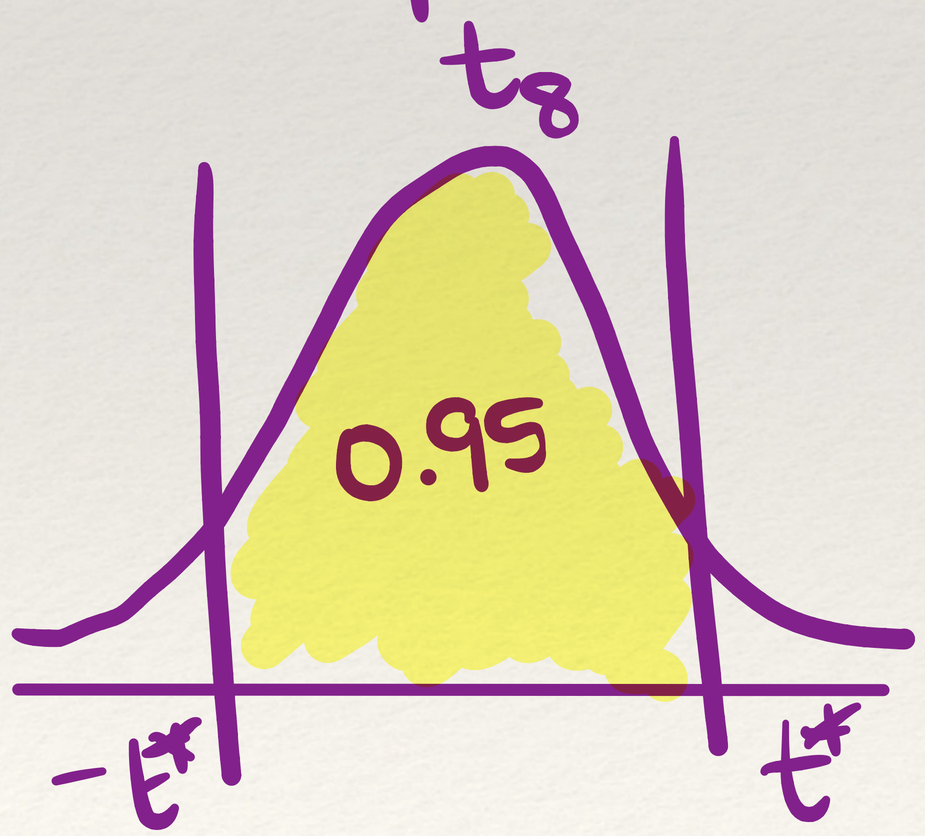
$$SE_{b_1} = \frac{s}{\sqrt{\Sigma(x_i - \bar{x})^2}}$$

t Confidence Interval for β_1 Example

From the random sample of 10 students, calculate and interpret the 95% confidence interval for the true change in College GPA given a change in High School GPA. Use the software output where $b_1 = 0.7758$ and $SE_{b_1} = 0.1937$. Assume all of the conditions for inference are met.

```
Call:
lm(formula = CollegeGPA ~ HSGPA)

Residuals:
     Min       1Q   Median       3Q      Max
-0.22364 -0.10636  0.01515  0.06045  0.35394

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.6861     0.6897   0.995  0.34902
HSGPA         0.7758     0.1937   4.006  0.00392 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1759 on 8 degrees of freedom
Multiple R-squared:  0.6673,    Adjusted R-squared:  0.6257
F-statistic: 16.05 on 1 and 8 DF,  p-value: 0.003918
```

$t^* = qt(0.975, 8) = 2.306$



$b_1 = 0.7758$

$t^* = 2.306$

$SE_{b_1} = 0.1937$

$$\underbrace{0.7758}_{\text{point estimate}} \pm \underbrace{2.306(0.1937)}_{\text{Margin of error}}$$

$= (0.329, 1.222)$

We are 95% confident that every one point increase in HS GPA, we expect college GPA to increase by 0.329 points to 1.222 points, on avg., with a point estimate of 0.7758 points.

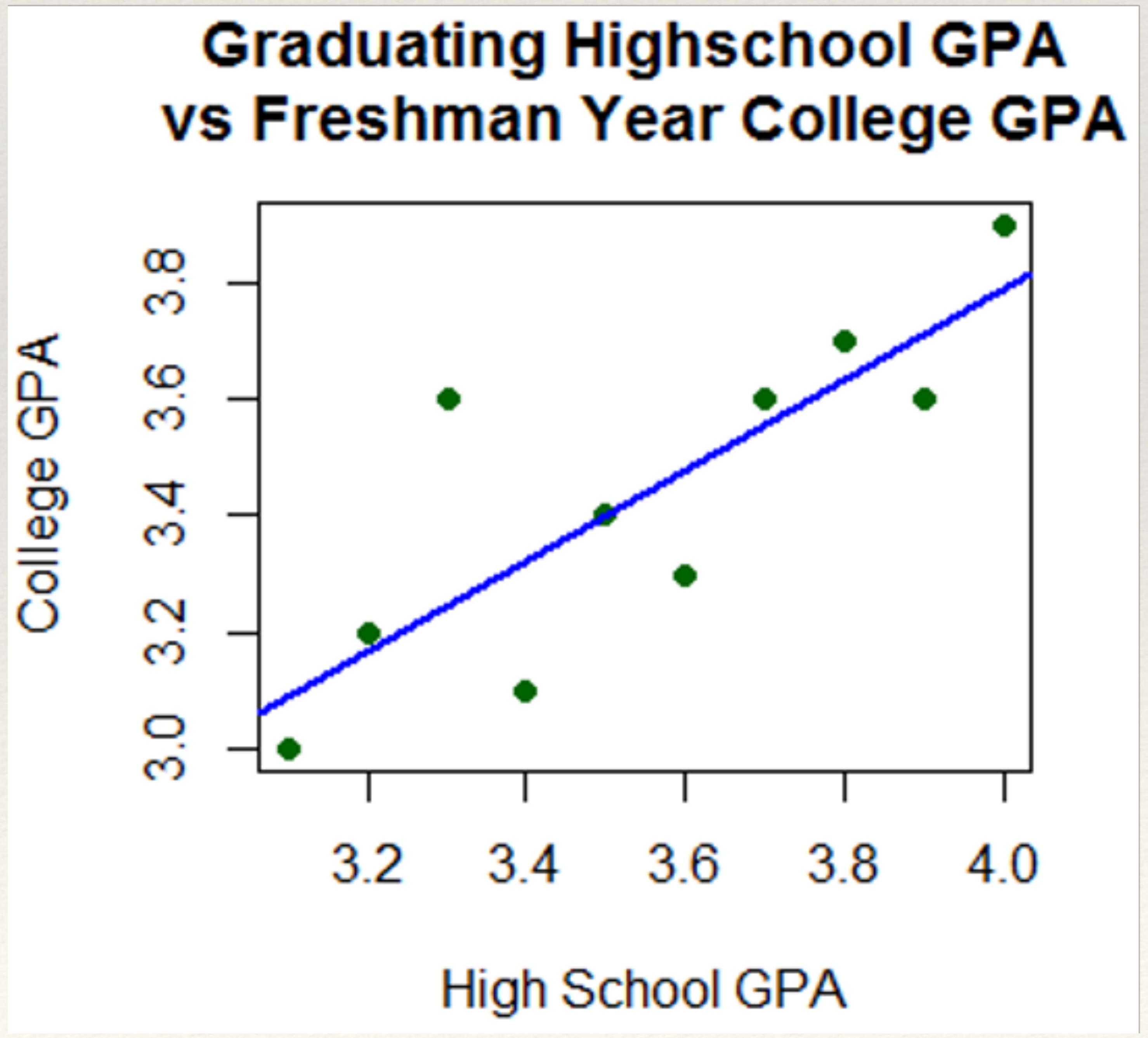
Testing for a Relationship

- It is common to test whether the relationship between two quantitative variables is **statistically significant**.

Do changes in the explanatory variable help to explain changes/variation in the response?

- Suppose a small college would like to know if their students high school GPA is a significant predictor of their freshman year college GPA. They take a random sample of 10 students. **How can they answer this question?**

t test on the slope, β_1



t Test for the Slope β_1

- When to use: want to test whether changes in the explanatory variable explain changes in the response
- Conditions required for inference:
  - Same as t confidence interval for β_1 conditions
  -
- Null and alternative hypotheses: If there was no relationship between x and y then the slope would be 0.

$H_0: \beta_1 = 0$

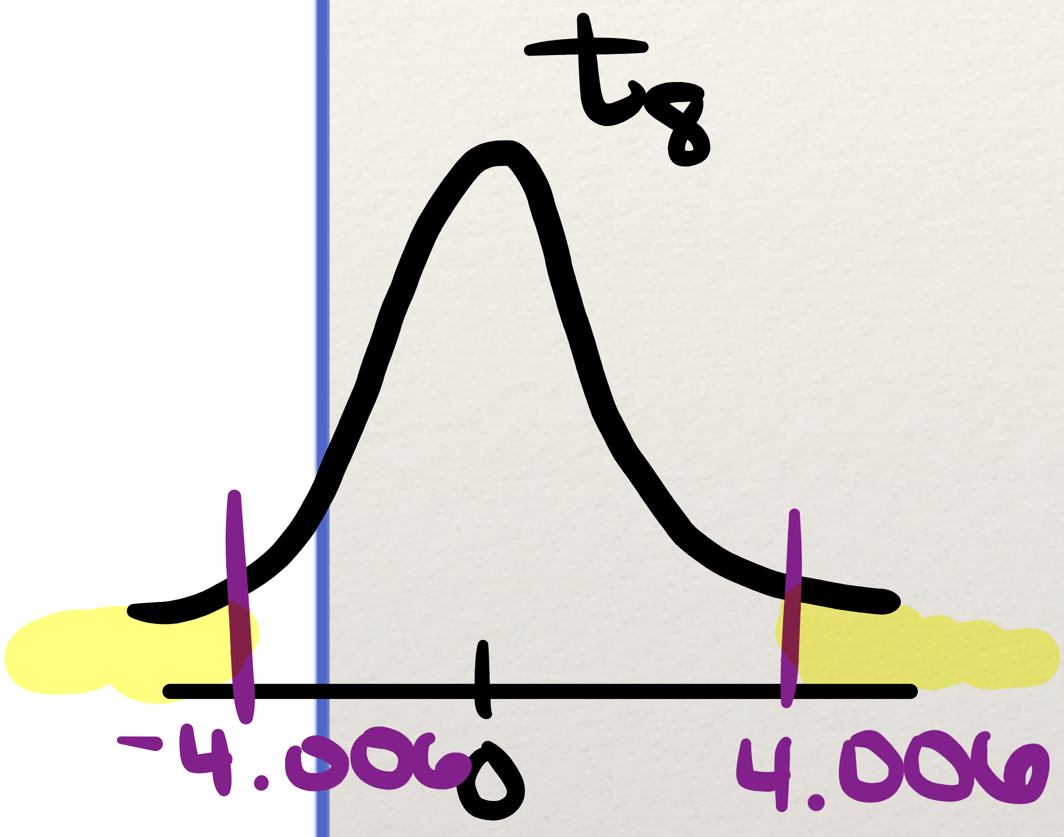
$H_A: \beta_1 \neq 0$

Test statistic: $t = \dfrac{b_1}{SE_{b_1}} \sim \underbrace{t_{n-(k+1)}}_{\text{Null dist'n}}$

Reading Software Output

$$t = \frac{0.7758}{0.1937} = 4.006$$

```
Call:
lm(formula = CollegeGPA ~ HSGPA)

Residuals:
     Min       1Q   Median       3Q      Max
-0.22364 -0.10636  0.01515  0.06045  0.35394

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.6861     0.6897   0.995  0.34902
HSGPA         0.7758     0.1937   4.006  0.00392 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1759 on 8 degrees of freedom
Multiple R-squared:  0.6673,    Adjusted R-squared:  0.6257
F-statistic: 16.05 on 1 and 8 DF,  p-value: 0.003918
```



Reading Statistical Software Output

	Estimate	Standard Error	t Statistic	p-value
Intercept	b_0	SE_{b_0}	b_0 / SE_{b_0}	p-value for t-test on the intercept
Explanatory Variable	b_1	SE_{b_1}	b_1 / SE_{b_1}	p-value for t-test on the slope

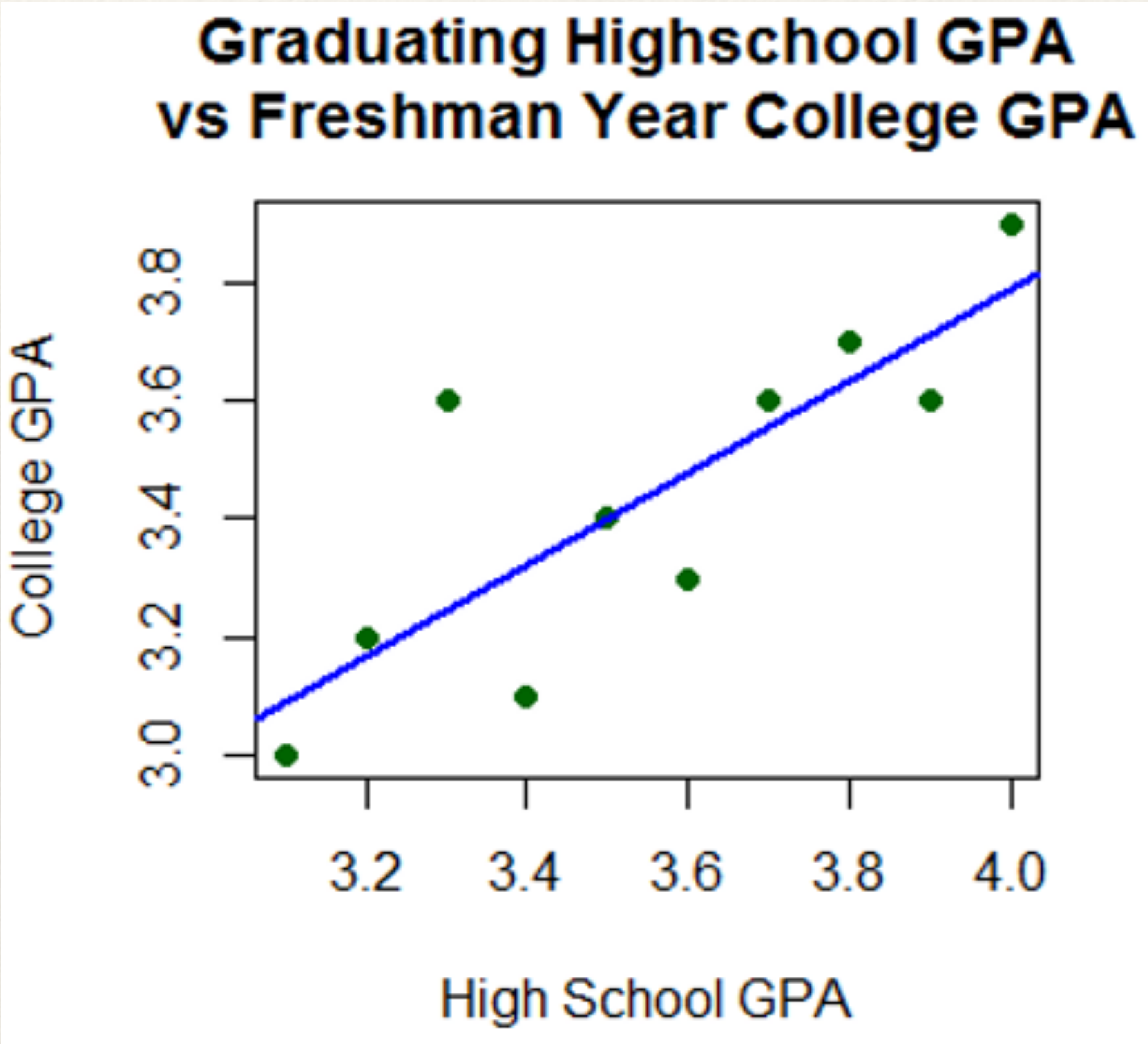
```
Call:
lm(formula = CollegeGPA ~ HSGPA)

Residuals:
     Min       1Q   Median       3Q      Max
-0.22364 -0.10636  0.01515  0.06045  0.35394

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.6861     0.6897   0.995  0.34902
HSGPA         0.7758     0.1937   4.006  0.00392 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1759 on 8 degrees of freedom
Multiple R-squared:  0.6673,     Adjusted R-squared:  0.6257
F-statistic: 16.05 on 1 and 8 DF,  p-value: 0.003918
```

Testing for a Linear Relationship Example

Use the software output to test whether high school GPA is a significant predictor of freshman year college GPA at the small college. The data is from a random sample of 10 students. Assume all of the conditions for inference are met. Use $\alpha = 0.05$.

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	0.6861	0.6897	0.995	0.34902	
HSGPA	0.7758	0.1937	4.006	0.00392	**

coefficient on HS GPA



Hypotheses: $H_0: \beta_1 = 0$

$H_A: \beta_1 \neq 0$

Test statistic and p-value:

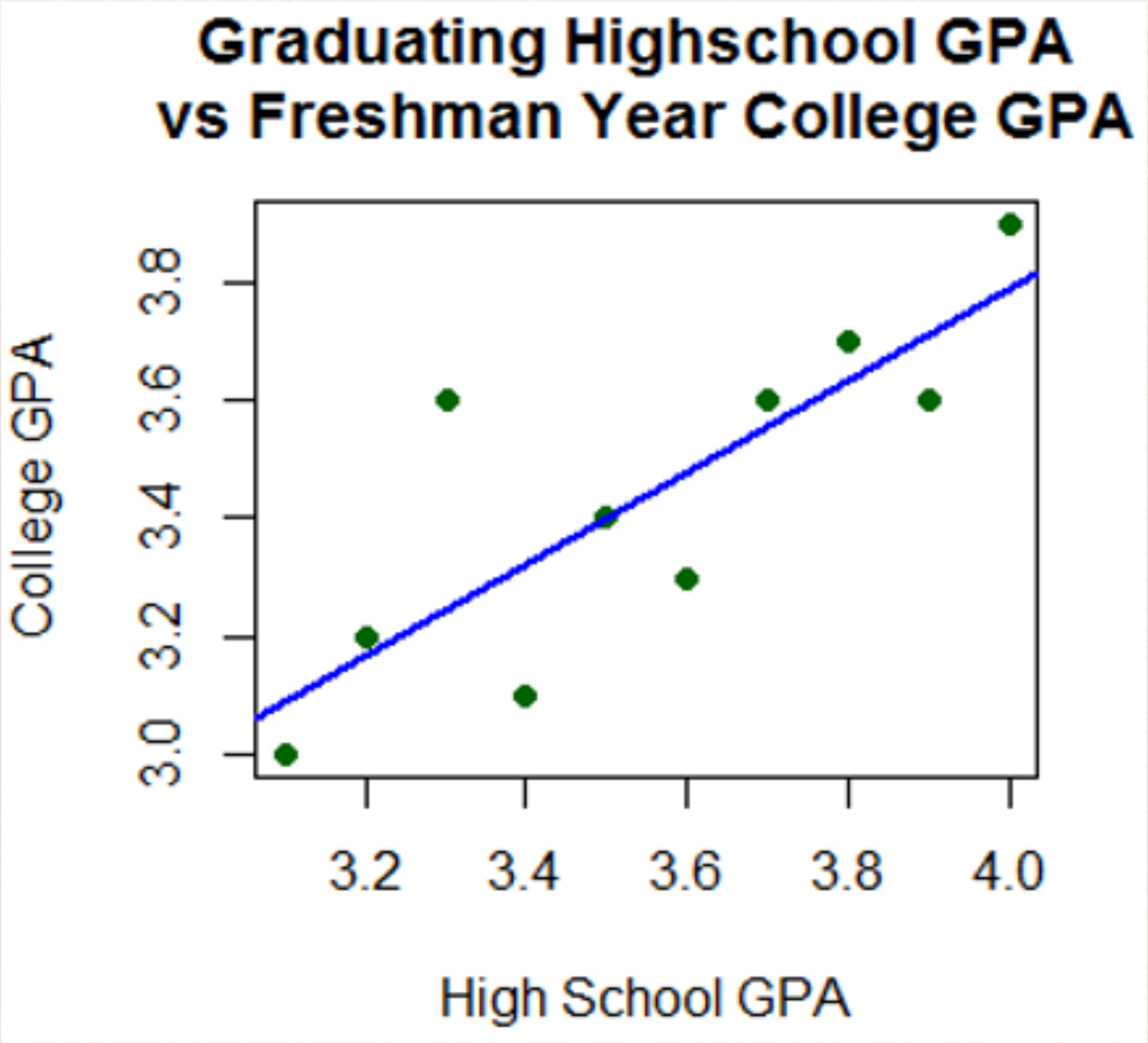
$t = 4.006$ p-value $= 0.00392$

~~Conclusion:~~ Interpretation of the p-value: The probability of observing a slope estimate greater than 0.7758 or less than -0.7758 when the true population slope is 0, is equal to 0.00392.

Testing for a Linear Relationship Example

Use the software output to test whether high school GPA is a significant predictor of freshman year college GPA at the small college. The data is from a random sample of 10 students. Assume all of the conditions for inference are met. Use $\alpha = 0.05$.

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	0.6861	0.6897	0.995	0.34902	
HSGPA	0.7758	0.1937	4.006	0.00392	**



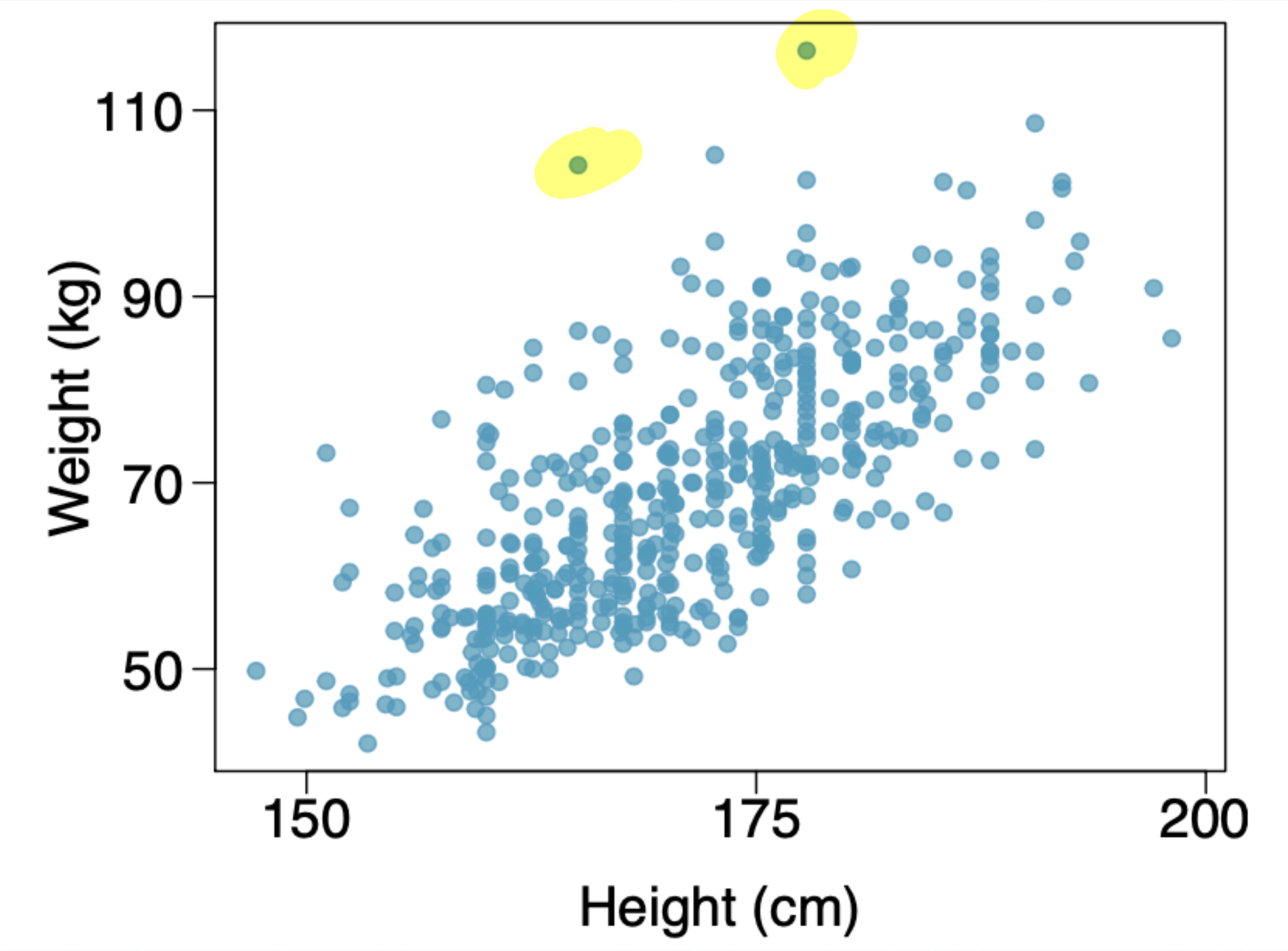
Hypotheses:

Test statistic and p-value:

Conclusion: There is **convincing** evidence to suggest that HS GPA is a significant predictor of college GPA at this small college. We reject the null hypothesis at the alpha = 0.05 significance level.

Regression Inference Example

The scatterplot and least squares summary below show the relation- ship between weight measured in kilograms and height measured in centimeters of 507 physically active individuals. **Describe the relationship between height and weight.**



Strong, positive, linear relationship between height and weight.

Regression Inference Example

The scatterplot and least squares summary below show the relation- ship between weight measured in kilograms and height measured in centimeters of 507 physically active individuals. **Write the equation of the regression line. Interpret the slope estimate.**

	Estimate	Std. Error	t value	P(>|t|)
Intercept	-105.01	7.54	-13.93	< 2e^-16
Height	1.02	0.04	25.5	< 2e^-16

$$\hat{y} = -105.01 + 1.02x$$

predicted weight

height

For each additional cm in height, the model predicts weight to increase by 1.02 kg.

Regression Inference Example

$n = 507$

Do the data provide strong evidence that an increase in height is associated with an increase in weight? Construct a 95% confidence interval for the slope parameter β_1. State the null and alternative hypotheses, report the p-value, and write a conclusion.

	Estimate	Std. Error	t value	P(>|t|)
Intercept	-105.01	7.54	-13.93	< 2e^-16
Height	1.02	0.04	25.5	< 2e^-16

$b_1 \pm t^* SE_{b_1}$

$t^* = qt(0.975, 505) = 1.965$

$1.02 \pm 1.965(0.04) = (0.941, 1.099)$

$H_0: \beta_1 = 0$

$H_A: \beta_1 \neq 0$

$t = 25.5$

p-value $< 2 \times 10^{-16}$

We are 95% confident that for each additional cm in height, we predict weight to increase by 0.941 kg to 1.099 kg, with a point estimate of 1.02 kg. There is convincing evidence that height is a significant predictor of weight. We reject the null hypothesis at any reasonable significance level.