UCLA School of Education & Information Studies Department of Education

EDUC 231D
Advanced Quantitative Methods: Multilevel Analysis
Winter 2025

Regression Review

Lecture 2 Presentation Slides

January 7, 2025

Today's Topics

- The least squares regression model
- Predicted values, residuals, and residual variance
- Inference for parameter estimates
- Centering
- Hand-on R exercise

Least Squares Regression

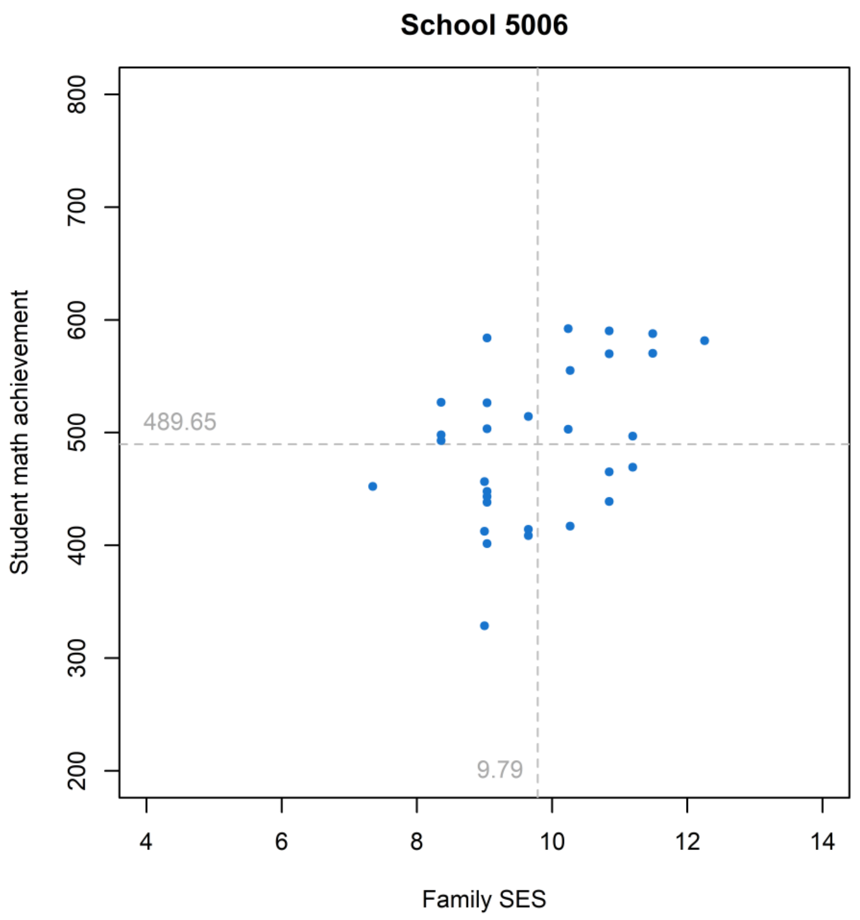
Estimating the relationship between two variables

- How is family SES related to math achievement in a school?
  - 2019 Grade 8 data from the TIMSS
  - 30 students in school 5006

Math score

Family SES

Variable	Mean	SD	Min	Median	Max
bsmmatxx	489.65	69.12	328.66	494.95	592.32
homeses	9.79	1.16	7.35	9.65	12.26

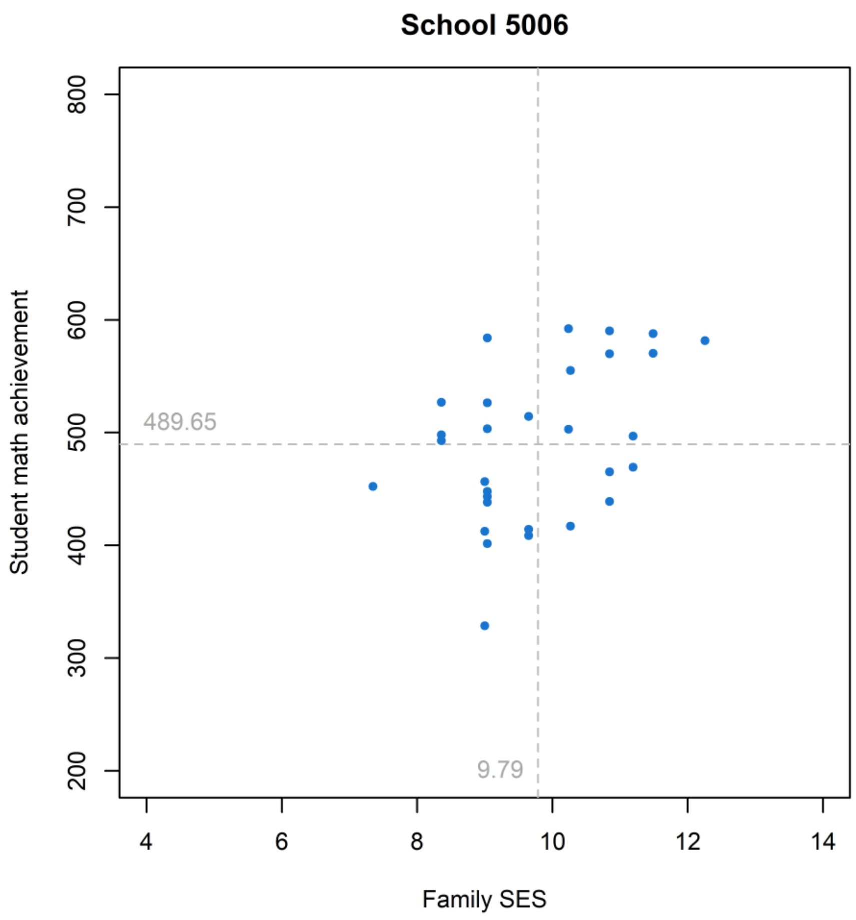


Estimating the relationship between two variables

- Estimate an ordinary least squares (OLS) linear regression model

$$Y_i = \beta_0 + \beta_1 X_i + r_i$$

$$r_i \sim N(0, \sigma^2)$$



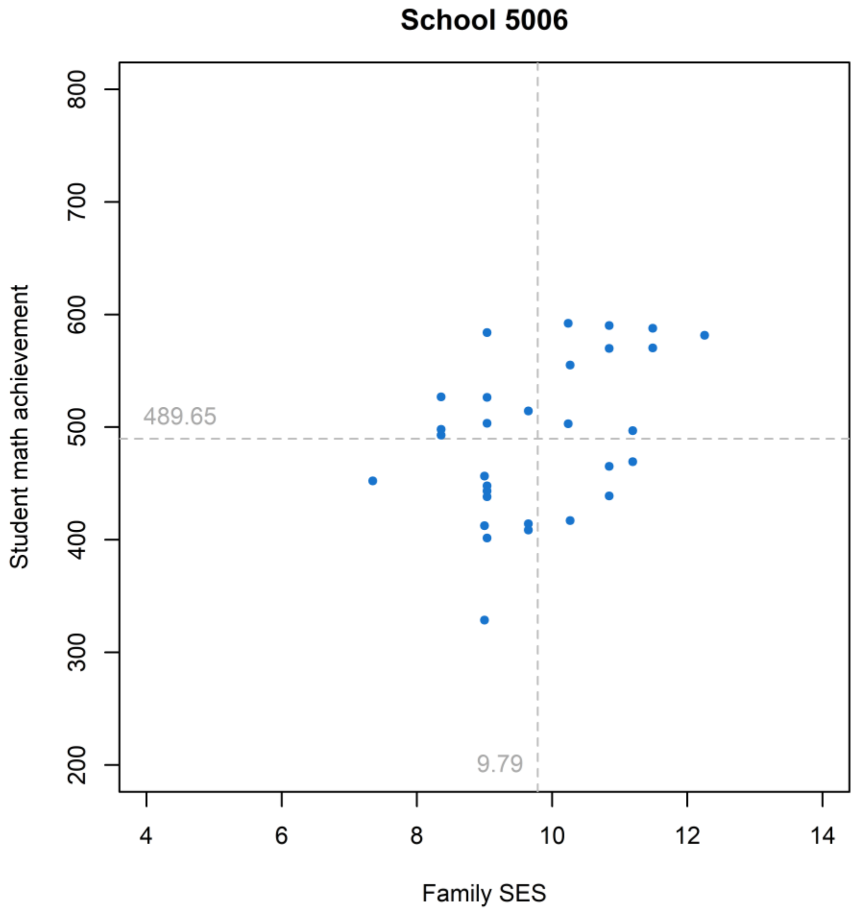
Estimating the relationship between two variables

- Estimate an ordinary least squares (OLS) linear regression model

$$Y_i = \beta_0 + \beta_1 X_i + r_i$$

Dependent variable:
Math score for student *i*

Independent variable:
Family SES for student *i*



Estimating the relationship between two variables

- Estimate an ordinary least squares (OLS) linear regression model

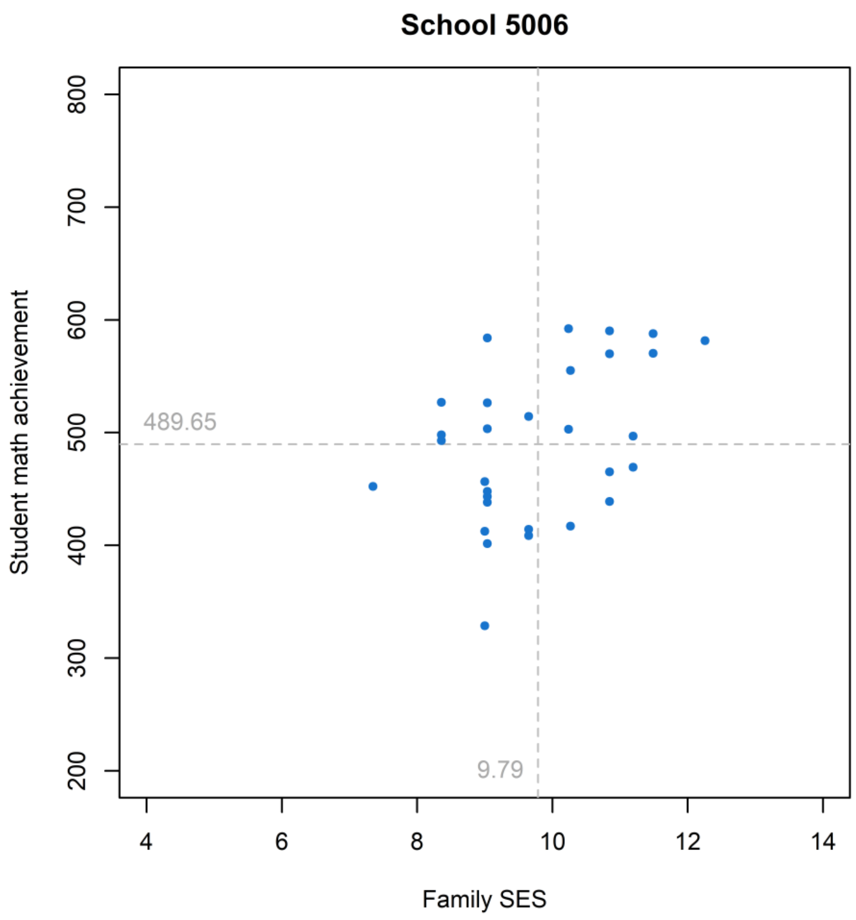
$$Y_i = \beta_0 + \beta_1 X_i + r_i$$

Math score for student *i*

Family SES for student *i*

The intercept:
expected math score when
family SES = 0

The slope:
Expected change in math
score when family SES
increases by 1 unit



Estimating the relationship between two variables

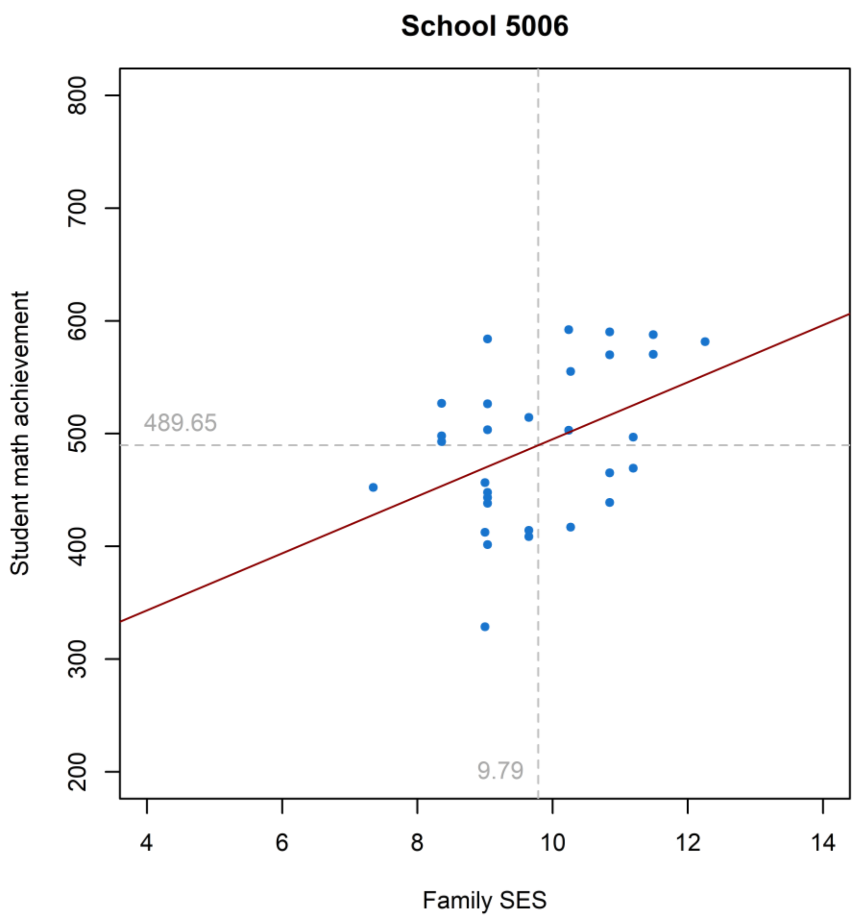
- Estimate an ordinary least squares (OLS) linear regression model

$$Y_i = \beta_0 + \beta_1 X_i + r_i$$

	Estimate	Standard Error	t value	Pr(>|t|)	
(Intercept)	242.095	100.322	2.413	0.0226	*
homeses	25.294	10.181	2.484	0.0192	*

*Signif. codes: 0 <= '***' < 0.001 < '**' < 0.01 < '*' < 0.05*

Residual standard error: 63.68 on 28 degrees of freedom
Multiple R-squared: 0.1806, Adjusted R-squared: 0.1513
F-statistic: 6.172 on 28 and 1 DF, p-value: 0.0192



Small group discussion

- In groups of 3-4, take 10 minutes to discuss …
  - What is the estimated value for β_0? Explain what that value means to somebody who's never taken a statistics class.
  - What is the estimated value for β_1? Explain what that value means to somebody who's never taken a statistics class.
  - What is the expected math score for a student with a family SES value of 9.00? What about a student with a family SES value of 9.79? What about a student with a family SES value of 11.00?
- Then share out with the whole class

Predicted Values and Residuals

Predicting values with the estimated model

- Estimated model:

$$\widehat{Y}_i = 242.095 + 25.294(X_i)$$

idschool	idstud	Observed Scores (Y) bsmmatxx	homeses	Predicted Scores ($\hat{Y}$) y_hat
5006	50060301	452.34	7.35	427.90
5006	50060303	447.92	9.04	470.65
5006	50060304	587.90	11.49	532.65
5006	50060305	555.24	10.27	501.74
5006	50060306	590.38	10.84	516.41
5006	50060307	526.52	9.04	470.65
5006	50060308	584.01	9.04	470.65
5006	50060311	569.97	10.84	516.41
5006	50060312	503.54	9.04	470.65
5006	50060313	570.42	11.49	532.65

Grade 8 Students in School 5006 (2019 TIMSS)

Residuals

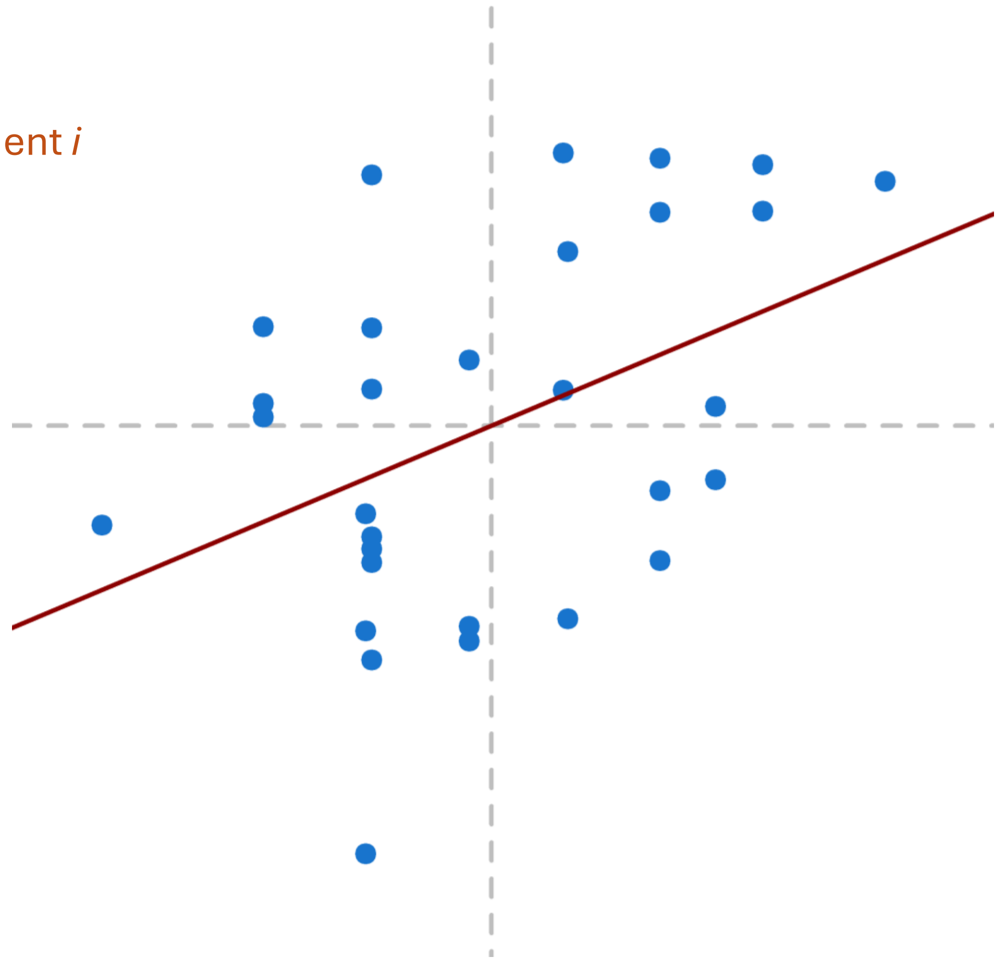
- Residual calculation:

The residual for student *i*

$$Y_i = \beta_0 + \beta_1 X_i + r_i$$

$$r_i = Y_i - \widehat{Y}_i$$

$$r_i = Y_i - 242.095 + 25.294(X_i)$$



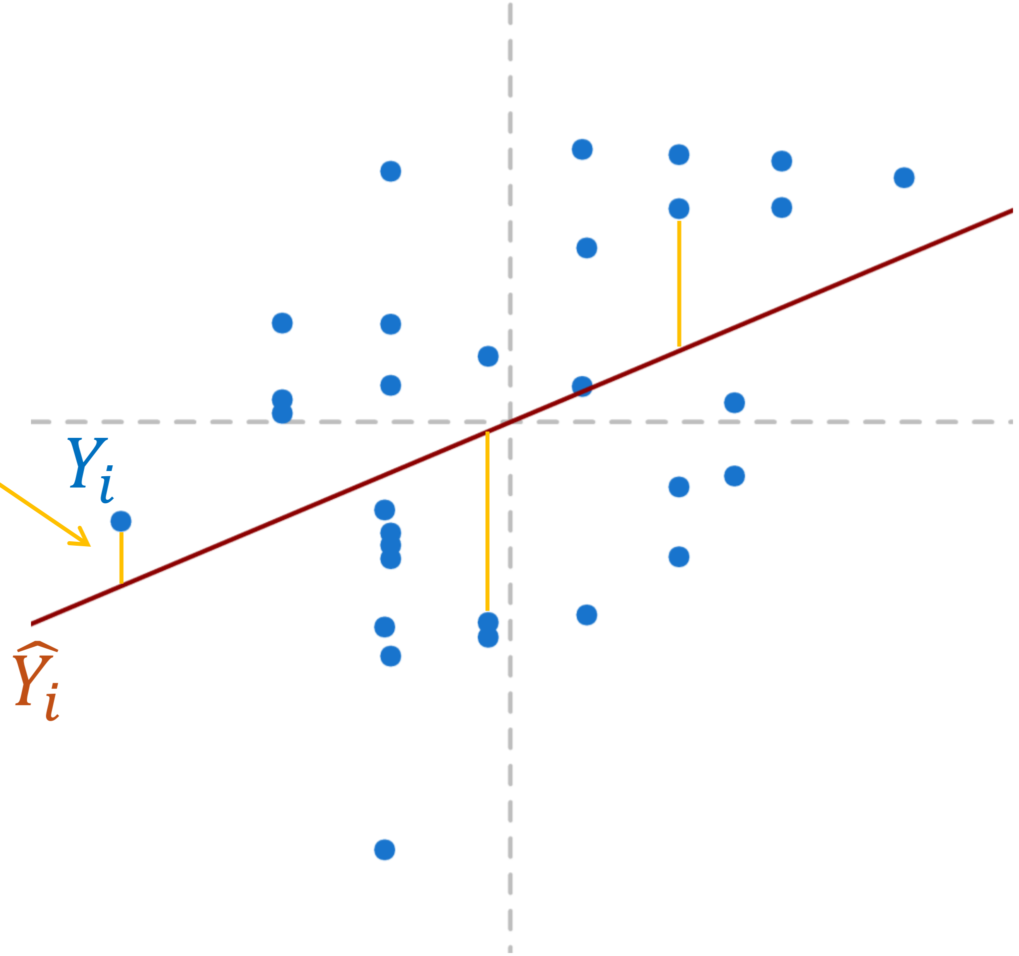
Residuals

- Residual calculation:

$$Y_i = \beta_0 + \beta_1 X_i + r_i$$

$$r_i = Y_i - \widehat{Y}_i$$

$$r_i = Y_i - 242.095 + 25.294(X_i)$$



Residuals

- Residual calculation:

$$Y_i = \beta_0 + \beta_1 X_i + r_i$$

$$r_i = Y_i - \widehat{Y}_i$$

$$r_i = Y_i - 242.095 + 25.294(X_i)$$

Residuals (r)

idschool	idstud	bsmmatxx	homeses	y_hat	r
5006	50060301	452.34	7.35	427.90	24.44
5006	50060303	447.92	9.04	470.65	-22.73
5006	50060304	587.90	11.49	532.65	55.25
5006	50060305	555.24	10.27	501.74	53.50
5006	50060306	590.38	10.84	516.41	73.97
5006	50060307	526.52	9.04	470.65	55.86
5006	50060308	584.01	9.04	470.65	113.36
5006	50060311	569.97	10.84	516.41	53.56
5006	50060312	503.54	9.04	470.65	32.89
5006	50060313	570.42	11.49	532.65	37.77

Grade 8 Students in School 5006 (2019 TIMSS)

Residuals

- Residual variance:

Variable	Mean	SD	Min	Max
bsmmatxx	489.65	69.12	328.66	592.32
y_hat	489.65	29.38	427.90	552.11
e	-0.00	62.57	-141.04	113.36

$$Y_i = \beta_0 + \beta_1 X_i + r_i$$

Assume residuals are normally distributed with mean = 0 and variance = σ^2

$$r_i \sim N(0, \sigma^2)$$

Residual variance:
How close the observed Y values are from the fitted model

- True population variance is unknow, must rely on the model-estimated variance

$$\hat{\sigma}^2 = \frac{\sum(Y_i - \hat{Y}_i)^2}{n-2} = (63.68)^2$$

Residuals

- Residual variance:

$$Y_i = \beta_0 + \beta_1 X_i + r_i$$

Assume residuals are normally distributed with mean = 0 and variance = σ^2

$$r_i \sim N(0, \sigma^2)$$

Variable	Mean	SD	Min	Max
bsmmatxx	489.65	69.12	328.66	592.32
y_hat	489.65	29.38	427.90	552.11
e	-0.00	62.57	-141.04	113.36

Note: the estimated residual variance is a little different than calculating the variance (or standard deviation) directly from the data.

Residual variance:
How close the observed Y values are from the fitted model

- True population variance is unknow, must rely on the model-estimated variance

$$\hat{\sigma}^2 = \frac{\sum (Y_i - \hat{Y}_i)^2}{n-2} = (63.68)^2$$

Residuals

- Total variance = explained variance + residual variance
  - $SS_{tot} = SS_{reg} + SS_{res}$
  - $\sum(Y_i - \bar{Y})^2 = \sum\left(\hat{Y}_i - \bar{Y}\right)^2 + \sum\left(Y_i - \hat{Y}_i\right)^2$

- Proportion of variance explained (R^2)
  - $\dfrac{SS_{reg}}{SS_{tot}} = 1 - \dfrac{SS_{res}}{SS_{tot}}$
  - $\dfrac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2} = 1 - \dfrac{\sum(Y_i - \hat{Y}_i)^2}{\sum(Y_i - \bar{Y})^2}$
  - $\dfrac{(29.38)^2}{(69.12)^2} = 0.1806$

	Estimate	Standard Error	t value	Pr(>|t|)	
(Intercept)	242.095	100.322	2.413	0.0226	*
homeses	25.294	10.181	2.484	0.0192	*

*Signif. codes: 0 <= '***' < 0.001 < '**' < 0.01 < '*' < 0.05*

Residual standard error: 63.68 on 28 degrees of freedom
Multiple R-squared: 0.1806, Adjusted R-squared: 0.1513
F-statistic: 6.172 on 28 and 1 DF, p-value: 0.0192

Inference for Parameter Estimates

Parameter Estimate: Slope

- Estimated model: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$

$$\hat{\beta}_1 = \frac{\sum(Y_i - \bar{Y})(X_i - \bar{X})}{\sum(X_i - \bar{X})^2}$$

	Estimate	Standard Error	t value	Pr(>|t|)	
(Intercept)	242.095	100.322	2.413	0.0226	*
homeses	$\boxed{25.294}$	10.181	2.484	0.0192	*

*Signif. codes: 0 <= '***' < 0.001 < '**' < 0.01 < '*' < 0.05*

Residual standard error: 63.68 on 28 degrees of freedom
Multiple R-squared: 0.1806, Adjusted R-squared: 0.1513
F-statistic: 6.172 on 28 and 1 DF, p-value: 0.0192

Standard Error: Slope

- Estimated model: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$
- $\hat{\beta}_1$ is based on sample data and is an estimate of the true slope (β_1)
- The standard error captures the uncertainty about $\hat{\beta}_1$ being β_1
  - As the sample size (n) increases, the SE decreases → a more precise estimate
  - As the variance of X increases, the SE decreases → a more precise estimate

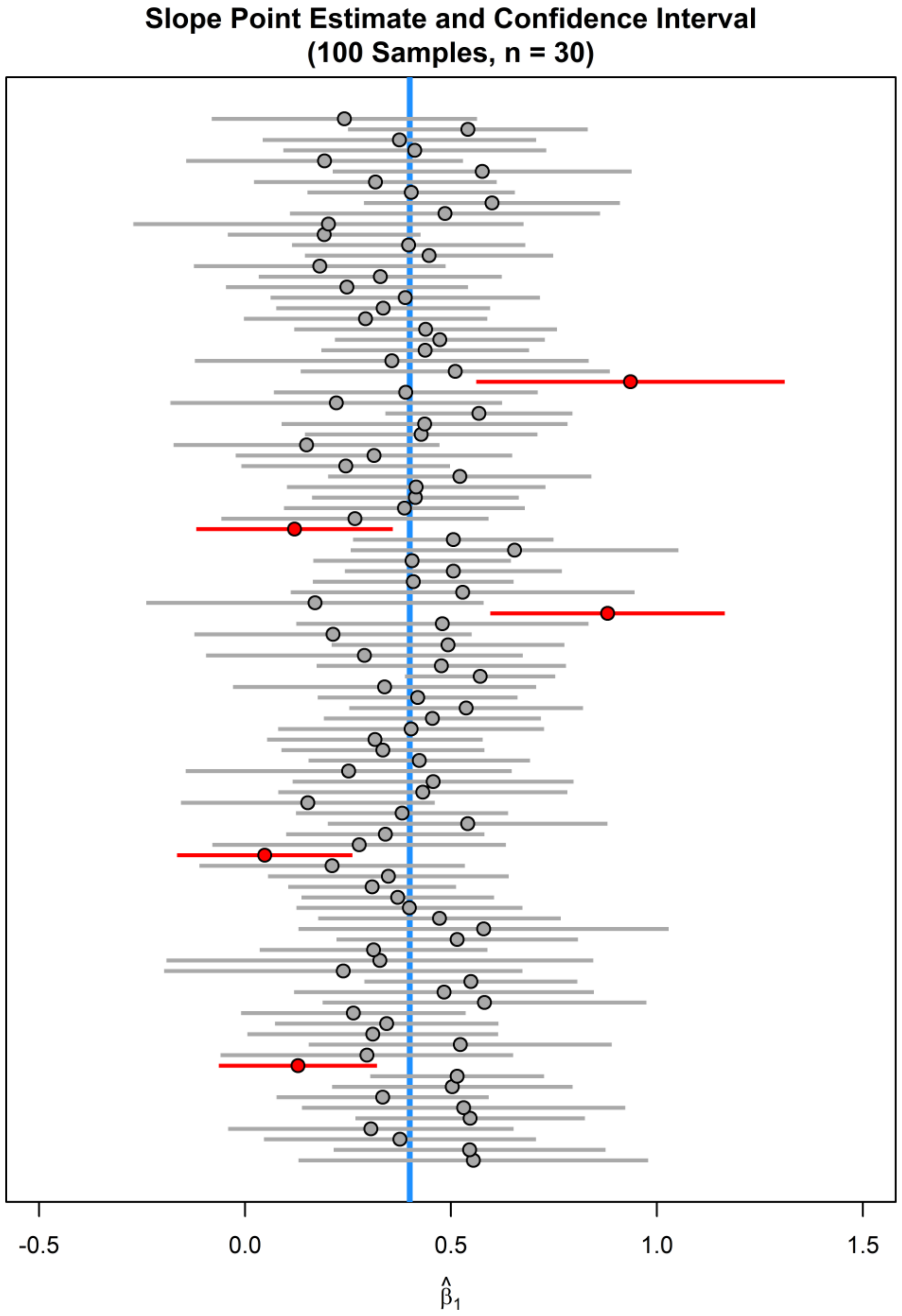
	Estimate	Standard Error	t value	Pr(>|t|)	
(Intercept)	242.095	100.322	2.413	0.0226	*
homeses	25.294	10.181	2.484	0.0192	*

*Signif. codes: 0 <= '***' < 0.001 < '**' < 0.01 < '*' < 0.05*

$$SE(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{\sum(X_i - \bar{X})^2}}$$

Confidence Intervals: Slope

- Can be more informative to express uncertainty using confidence intervals instead of standard errors
- Approximate 95% confidence interval for the slope:
  - Lower bound = $\hat{\beta}_1 - 2 \times SE(\hat{\beta}_1)$
  - Upper bound = $\hat{\beta}_1 + 2 \times SE(\hat{\beta}_1)$
- Under repeated sampling, the 95% confidence interval should contain the true population value 95% of the time



Parameter Estimate: Intercept

- Estimated model: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$

	Estimate	Standard Error	t value	Pr(>|t|)	
(Intercept)	242.095	100.322	2.413	0.0226	*
homeses	25.294	10.181	2.484	0.0192	*

*Signif. codes: 0 <= '***' < 0.001 < '**' < 0.01 < '*' < 0.05*

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}_i$$

Residual standard error: 63.68 on 28 degrees of freedom
Multiple R-squared: 0.1806, Adjusted R-squared: 0.1513
F-statistic: 6.172 on 28 and 1 DF, p-value: 0.0192

Standard Error: Intercept

- Estimated model: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$
- $\hat{\beta}_0$ is based on sample data and is an estimate of the true slope (β_0)
- The standard error captures the uncertainty about $\hat{\beta}_0$being β_0
  - As the sample size (n) increases, the SE decreases → a more precise estimate
  - As the variance of X increases, the SE decreases → a more precise estimate

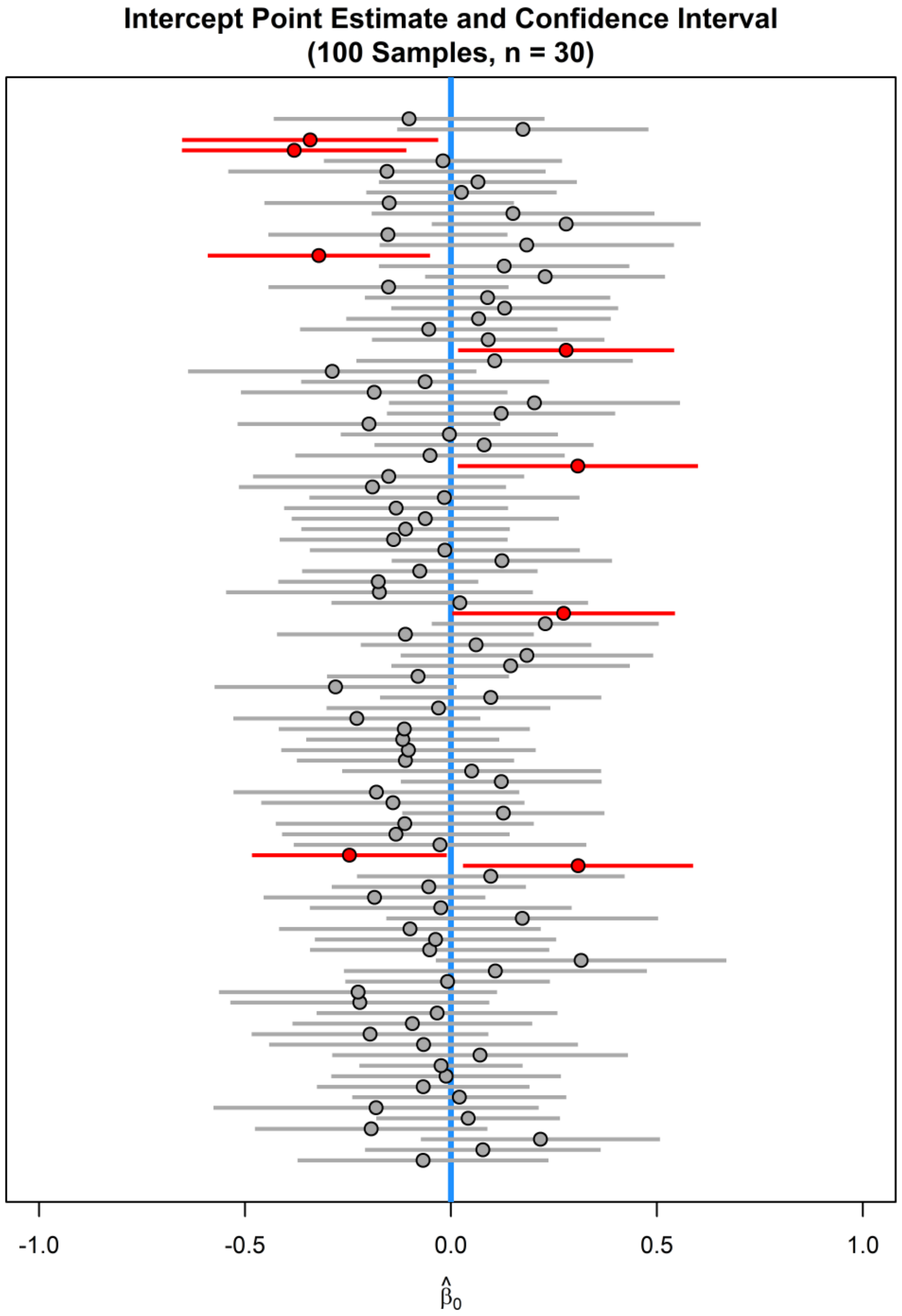
	Estimate	Standard Error	t value	Pr(>|t|)	
(Intercept)	242.095	100.322	2.413	0.0226	*
homeses	25.294	10.181	2.484	0.0192	*

*Signif. codes: 0 <= '***' < 0.001 < '**' < 0.01 < '*' < 0.05*

$$SE(\hat{\beta}_0) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{X}^2}{\sum(X_i - \bar{X})^2}}$$

Confidence Intervals: Intercept

- Can be more informative to express uncertainty using confidence intervals instead of standard errors
- Approximate 95% confidence interval for the intercept:
  - Lower bound = $\hat{\beta}_0 - 2 \times SE(\hat{\beta}_0)$
  - Upper bound = $\hat{\beta}_0 + 2 \times SE(\hat{\beta}_0)$
- Under repeated sampling, the 95% confidence interval should contain the true population value 95% of the time



A Note on *P*-values

	Estimate	Standard Error	t value	Pr(>|t|)	
(Intercept)	242.095	100.322	2.413	0.0226	*
homeses	25.294	10.181	2.484	0.0192	*

*Signif. codes: 0 <= '***' < 0.001 < '**' < 0.01 < '*' < 0.05*

- Be careful when interpreting *p*-values
  - *P*-values can indicate how incompatible the data are with a null hypothesis and the underlying model assumptions
  - *P*-values do <u>not</u> measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone
  - Scientific conclusions and business or policy decisions should not be based only on whether a *p*-value passes a specific threshold (e.g., $p < 0.05$)
  - Proper inference requires full reporting and transparency
  - A *p*-value, or statistical significance, does not measure the size of an effect or the importance of a result
  - By itself, a *p*-value does not provide a good measure of evidence regarding a model or hypothesis

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA Statement on *p* -Values: Context, Process, and Purpose. *The American Statistician*, *70*(2), 129–133. https://doi.org/10.1080/00031305.2016.1154108

Centering

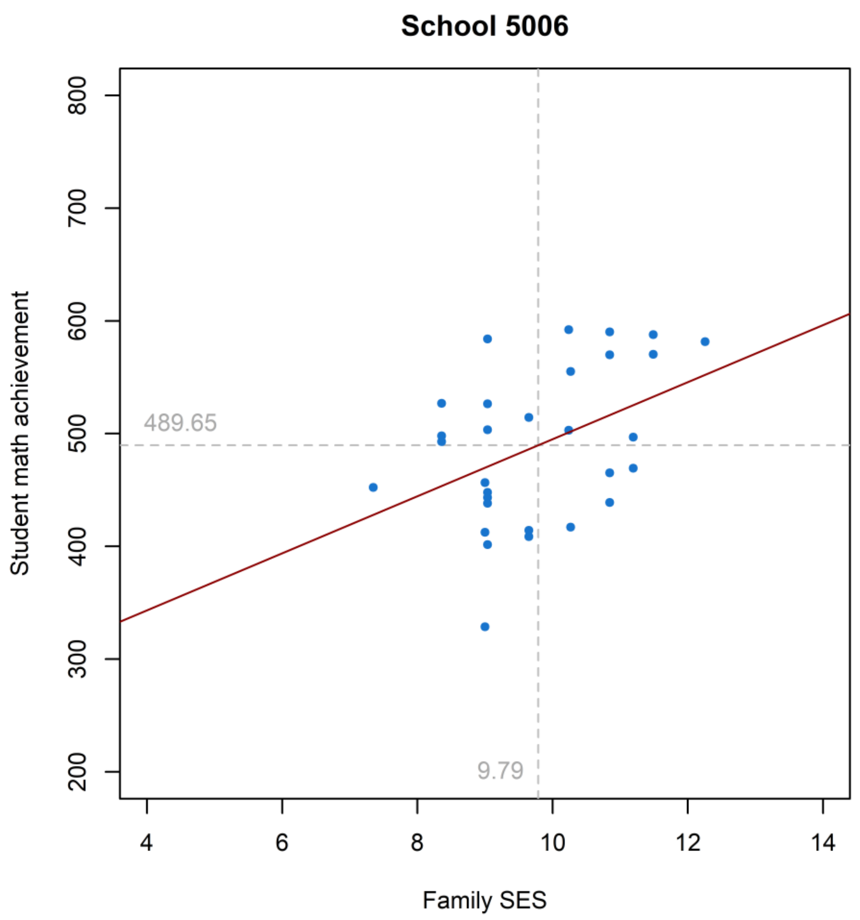
Centering

- Often helpful to rescale the independent variables (X_i) to help with the interpretation of results

$$Y_i = \beta_0 + \beta_1 X_i + r_i$$

	Estimate	Standard Error	t value	Pr(>|t|)	
(Intercept)	242.095	100.322	2.413	0.0226	*
homeses	25.294	10.181	2.484	0.0192	*

*Signif. codes: 0 <= '***' < 0.001 < '**' < 0.01 < '*' < 0.05*



Centering

- Mean centering is commonly used and particularly useful for interpretation

$$Y_i = \beta_0 + \beta_1 (X_i - \bar{X}_.) + r_i$$

Mean center by subtracting the mean value from each student's value

			SES Score (X_i)		Mean-Centered SES Score ($X_i - \bar{X}_.$)
idschool	idstud	bsmmatxx	homeses	meanses	homesesc
5006	50060301	452.34	7.35	9.79	-2.44
5006	50060303	447.92	9.04	9.79	-0.75
5006	50060304	587.90	11.49	9.79	1.70
5006	50060305	555.24	10.27	9.79	0.48
5006	50060306	590.38	10.84	9.79	1.06
5006	50060307	526.52	9.04	9.79	-0.75
5006	50060308	584.01	9.04	9.79	-0.75
5006	50060311	569.97	10.84	9.79	1.06
5006	50060312	503.54	9.04	9.79	-0.75
5006	50060313	570.42	11.49	9.79	1.70

Grade 8 Students in School 5006 (2019 TIMSS)

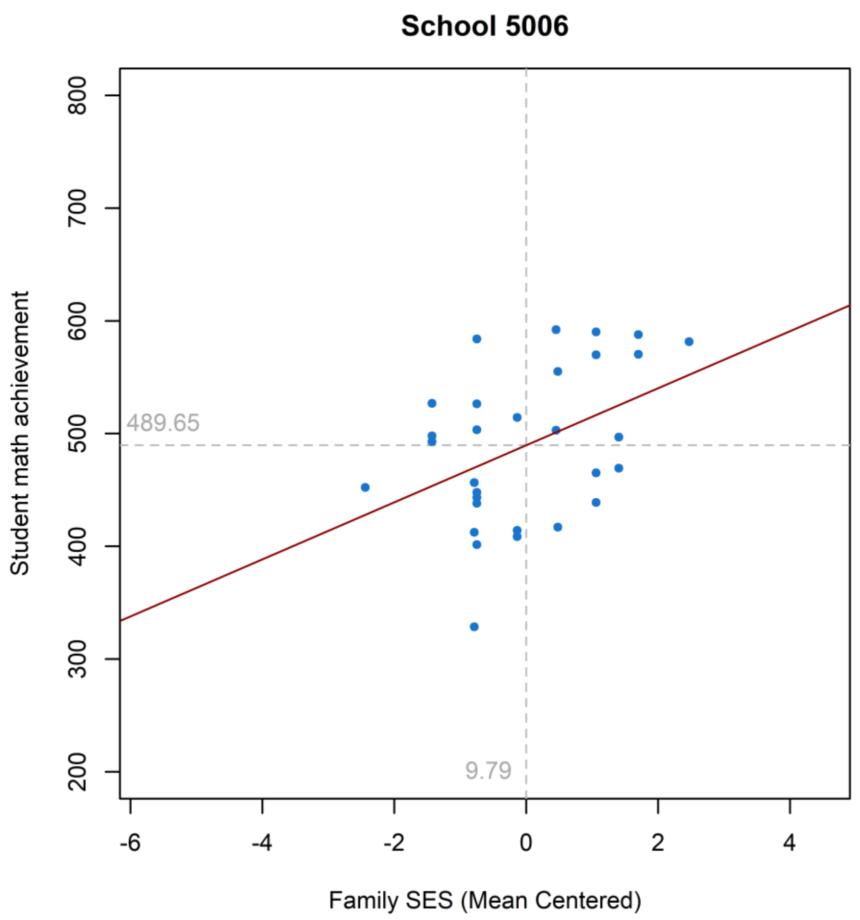
Centering

- Re-estimate the model using our centered SES variable

	Estimate	Standard Error	t value	Pr(>|t|)	
(Intercept)	489.648	11.626	42.116	0.0000	***
homesesc	25.294	10.181	2.484	0.0192	*

*Signif. codes: 0 <= '***' < 0.001 < '**' < 0.01 < '*' < 0.05*

Residual standard error: 63.68 on 28 degrees of freedom
Multiple R-squared: 0.1806, Adjusted R-squared: 0.1513
F-statistic: 6.172 on 28 and 1 DF, p-value: 0.0192



Small group exercise

- In groups of 3-4, take 20 minutes to conduct the following analysis of **School 5181** using R …
  - Calculate the mean math score and family SES value in School 5181
  - Center the family SES value on the school mean
  - Use a linear model to estimate the relationship between family SES and math achievement
- Discuss …
  - What is the estimated value for β_0? Explain what that value means to somebody who's never taken a statistics class.
  - What is the estimated value for β_1? Explain what that value means to somebody who's never taken a statistics class.
  - How do the regression estimates for School 5181 compare to the regression estimates for School 5006? Discuss the implications of the point estimates and standard errors

Small group exercise (results)

School 5006	Estimate	Standard Error	t value	Pr(>|t|)	
(Intercept)	489.648	11.626	42.116	0.0000	***
homesesc	25.294	10.181	2.484	0.0192	*

*Signif. codes: 0 <= '***' < 0.001 < '**' < 0.01 < '*' < 0.05*

Residual standard error: 63.68 on 28 degrees of freedom

Multiple R-squared: 0.1806, Adjusted R-squared: 0.1513

F-statistic: 6.172 on 28 and 1 DF, p-value: 0.0192

School 5181	Estimate	Standard Error	t value	Pr(>|t|)	
(Intercept)	425.338	9.324	45.619	0.0000	***
homesesc	16.333	6.150	2.656	0.0106	*

*Signif. codes: 0 <= '***' < 0.001 < '**' < 0.01 < '*' < 0.05*

Residual standard error: 66.58 on 49 degrees of freedom

Multiple R-squared: 0.1258, Adjusted R-squared: 0.108

F-statistic: 7.053 on 49 and 1 DF, p-value: 0.0106