

AP Statistics Extensions

Multiple Regression

Spring 2026

Matthew J. Madison
mjmadison@uga.edu

Multiple Regression II

Agenda

- Multiple regression
 - Review of multiple linear regression
 - Review Lab Assignment #1
- Multiple regression with categorical variables
 - Examples in R Studio
 - Lab Assignment #2

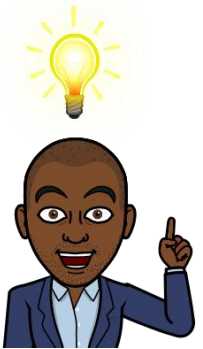


Multiple Regression

- Response = AP Statistics final exam (0 – 100 scale)
- Predictors: study time per week

$$\widehat{score}_i = 50 + 13(hours_i)$$

- $R^2 = .32$
- 32% of the variance in final exam scores is explained by hours studied
 - 68% is unexplained
- Let's add more predictors to explain more variance



Multiple Regression

- Multiple regression goal is to predict a continuous DV with **multiple predictor variables**:

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + e_i$$

- Assumptions:
 - $e_i \sim N(0, \sigma^2)$
 - Homoscedasticity
 - Linearity
 - Independence
- Other considerations
 - Multicollinearity
 - Outliers and influential observations

Multiple Regression

- Overall goal: explain variance in the DV **with multiple predictors**
- **Interpret coefficients** for continuous predictors
- Intercept (β_0): the predicted y when **all** predictors are zero
- Slope (β_i): the predicted change in y given a 1-unit change in x_i , **holding other variables constant**

Multiple Regression

- Response = AP Statistics final exam (0 – 100 scale)
- Predictors: study time per week, algebra EOC score

$$\widehat{score}_i = 26 + 12(hours_i) + 0.3(algebra_i)$$

- Intercept interpretation: if you study 0 hours per week **and** made a 0 on the algebra EOC, predicted score is 26
- Slope interpretation: for every 1 additional hour studied per week, predicted score increases by 12 points, **holding other variables constant**
- Slope interpretation: for every 1 additional point on the algebra EOC, predicted score increases by 0.3 points, **holding other variables constant**

Multiple Regression

- **Model Evaluation**
- R^2 : proportion of variance explained by all the predictors

- Predictors: study time per week, algebra EOC score

$$\widehat{score}_i = 26 + 12(hours_i) + 0.3(algebra_i)$$

- R^2 was .32 (68% unexplained)
- With the added predictor, R^2 increases to .48
 - Still 52% unexplained, but better than 68%

Multiple Regression

- Predictors: study time per week, algebra EOC score

$$\widehat{score}_i = 26 + 12(hours_i) + 0.3(algebra_i)$$

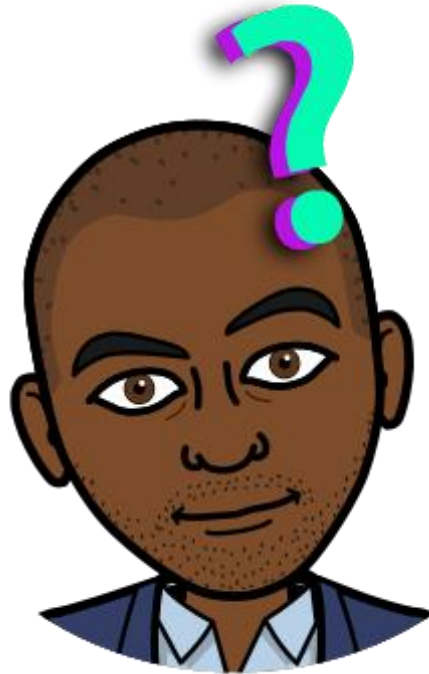
- For someone that studies for 4 hours per week, and made a 65 on the Algebra EOC, their predicted score is:

$$\widehat{score}_i = 26 + 12(4) + 0.3(65) = \mathbf{93.5}$$

Multiple Regression

- **Model Evaluation**
- Significance of all predictors
 - A p-value $< .05$ for a predictor indicates that it is a significant predictor
 - A p-value $> .05$ for a predictor indicates that it does not predict the response variable
- Model *parsimony*
 - Want a model that predicts well, but is efficient and doesn't have redundant or wasted space
 - Drop insignificant predictors until the model only has significant predictors

Questions



Lab Assignment #1

- Open software instructions
- Data: predicting freshman year college GPA
 - *fygpa*: first year college GPA
 - *hsgpa*: high school GPA
 - *sat*: SAT total score
 - *numap*: number of AP courses taken in high school
 - *study*: self-reported hours studied per week

Multiple Regression II

Multiple Regression

- In previous examples, our predictors were continuous
 - Hours studied
 - Algebra EOC score
 - SAT score
 - GPA
- But what if we had categorical predictors?
 - Algebra course grade (5 categories: A – F)
 - If you took statistics before (2 categories: yes or no)
 - Demographics (gender, race/ethnicity)

Multiple Regression

- For categorical predictors, we have to recode our data
 - Dummy coding, reference coding, baseline coding
- Suppose our predictor of AP stat exam grade was Algebra course grade
 - A
 - B - C
 - D - F
- Then we are going to create two dummy variables and set one group to baseline (D-F)
 - In general, with k categories create $k-1$ dummy variables

Multiple Regression

- Then we are going to create two dummy variables and set one group to baseline (D-F)
 - In general, with k categories create $k-1$ dummy variables

ID	Stat	Alg	gradeA	gradeBC
1	78	A	1	0
2	92	C	0	1
3	71	B	0	1
4	64	D	0	0
5	84	D	0	0
6	88	B	0	1
7	81	F	0	0

Parameter Interpretations

- Regression model:

$$\widehat{score} = b_0 + b_1(\text{grade}A) + b_2(\text{grade}BC)$$

- Intercept (b_0): average score for the reference group (D-F)
- Slope (b_1): increase in score for the coded group (A) relative to the reference group (D-F)
- Slope (b_2): increase in score for the coded group (B-C) relative to the reference group (D-F)

Parameter Interpretations

- Regression model:

$$\widehat{score} = 66 + 22(gradeA) + 14(gradeBC)$$

- Intercept: average score for students with D or F in Algebra was 66
- Slope (*gradeA*): average score for students with A in Algebra was 88 ($66 + 22$)
- Slope (*gradeBC*): average score for students with BC in Algebra was 80 ($66 + 14$)
 - Check significance of slopes for group comparisons

Example in R Studio

- Create a new R script
 - Read in the statscores.csv data file

Lab Assignment #2

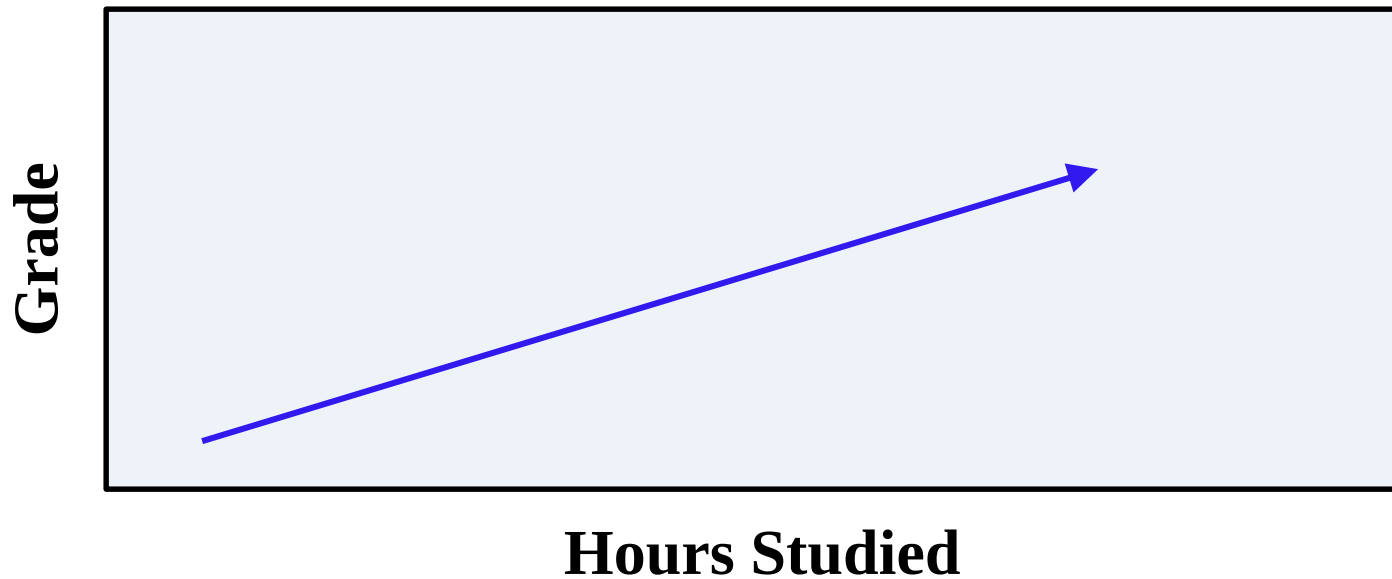
- Open software instructions
- Data: predicting social media views
 - ***views***: views received by a post
 - ***platform***: social media platform

Future Directions

- We have seen continuous predictors
- We have modeled categorical predictors
- What if we wanted to model both?

Future Directions

- What if we wanted to model both?
- Consider the time you study (continuous predictor)
 - Effect should be positive



Future Directions

- But would the effect of study hours be the same for everyone?
 - Consider different types of studying (reading, practice problems, flashcards)

