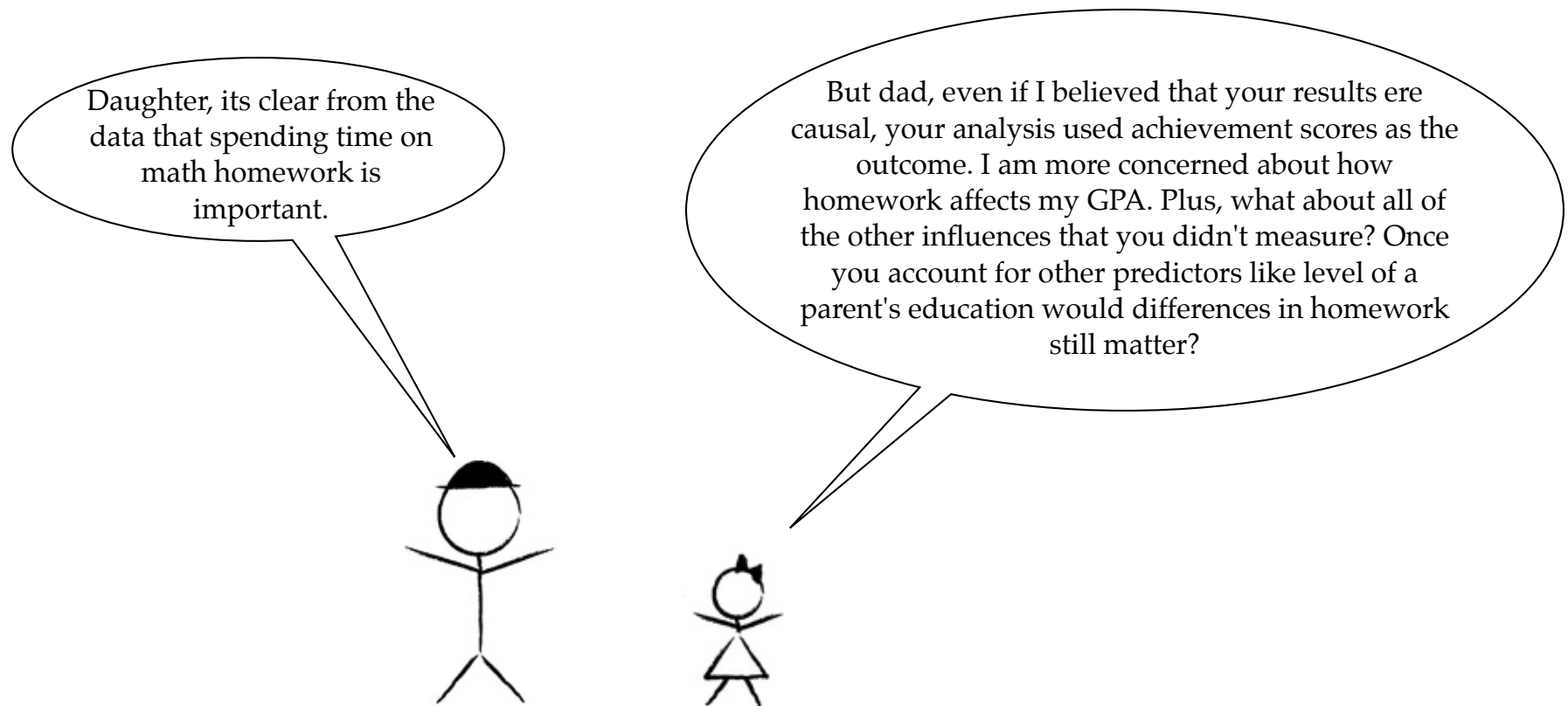


Multiple Regression



# Back to the Drawing Board

Does variation in the amount of time spent on homework explain variation in GPA? Even after accounting for differences in parent education levels?



## Data

- Eighth grade students' overall GPA (in all subjects) on a 100-pt scale.
- Average time on homework spent per week, in hours, on all subjects
- Level of education of the students' parents, in years of schooling for the parent that has the higher level of education.

## Read in Data and Load Libraries

```
# Load the data (homework-education-gpa.csv)
> multReg = read.csv("EPSY-8262/data/homework-education-gpa.csv")

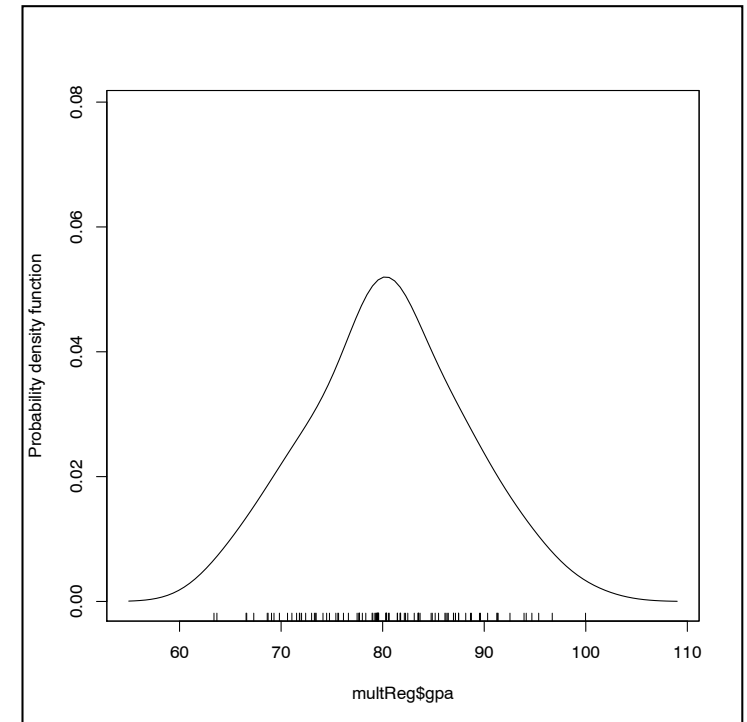
# Load libraries
> library(ggplot2)
> library(psych)
> library(sm)

> head(multReg)
```

	gpa	parentEd	homework
1	78	13	2
2	79	14	6
3	79	13	1
4	89	13	5
5	82	16	3
6	77	13	4

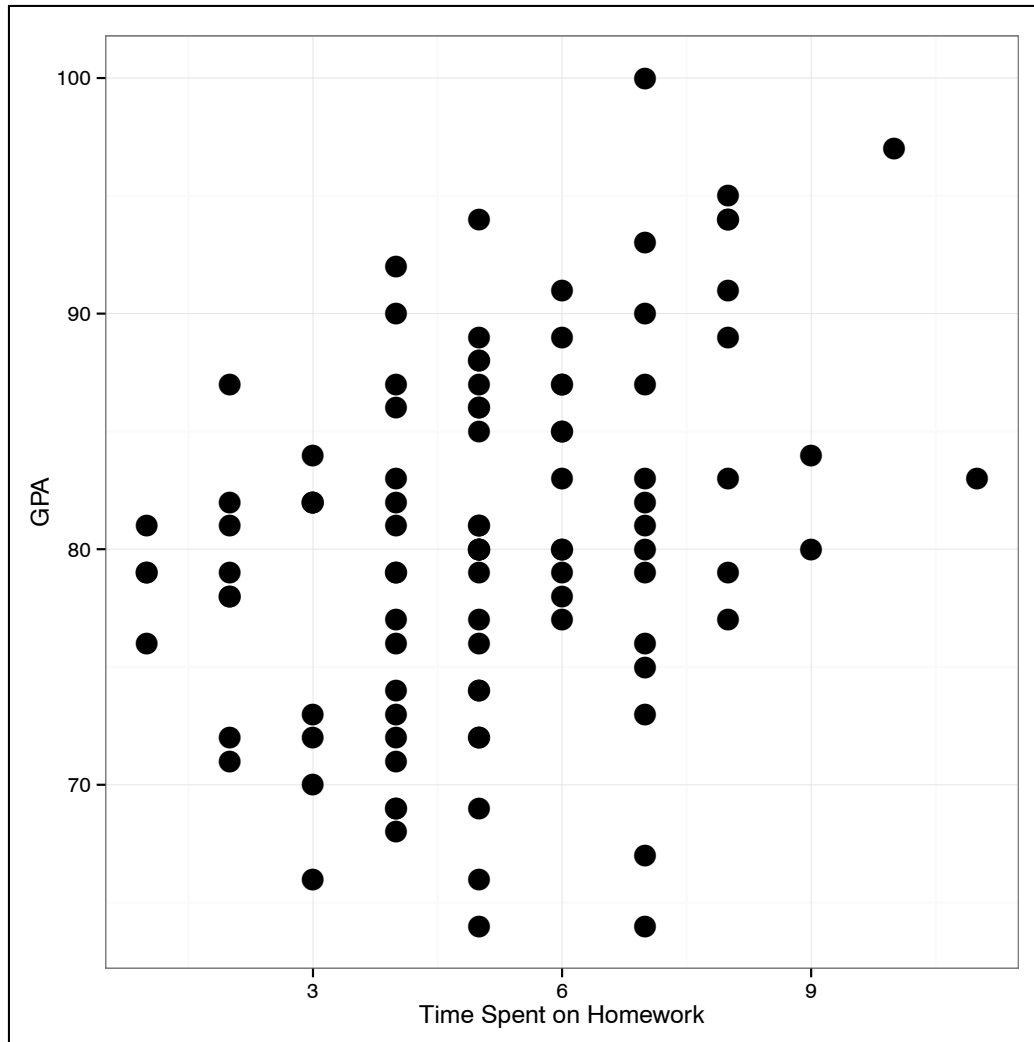
# Examine the Outcome and Predictors

The marginal distribution of the GPAs (outcome) is unimodal with a mean of 80.4. There is variation in these GPAs ( $SD = 7.6$ ).



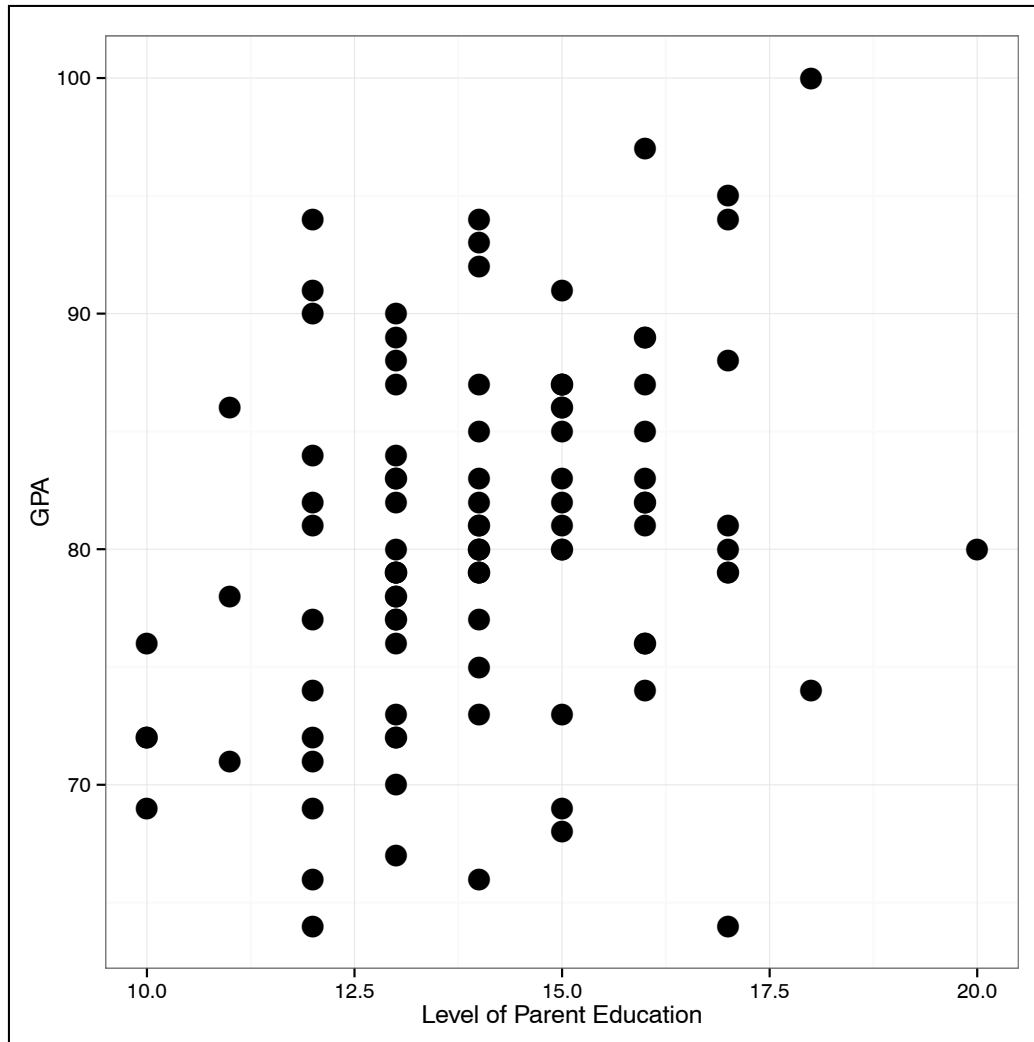
Variable	<i>M</i>	<i>SD</i>
GPA (100-point scale)	80.4	7.6
Time spent on homework (in hours)	5.1	2.1
Level of parent education (in years)	14.0	1.9

# Examining the Distribution of GPAs Conditioned on Time Spent on Homework



The plot suggests a positive, moderate, linear relationship ( $r = 0.33$ ) between time spent on homework and GPA. There do not look to be any outlying observations in the plot.

# Examining the Distribution of GPAs Conditioned on Parent Education



The plot suggests a positive, moderate, linear relationship ( $r = 0.29$ ) between level of parent education and GPA. There do not look to be any outlying observations in the plot.

# Fitting the Multiple Regression Model Using R

```
> lm.a = lm(gpa ~ homework + parentEd, data = multReg)
> summary(lm.a)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	63.2270	5.2398	12.067	< 2e-16	***
homework	0.9878	0.3609	2.737	0.00737	**
parentEd	0.8706	0.3842	2.266	0.02568	*

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.092 on 97 degrees of freedom

Multiple R-squared: 0.1521, Adjusted R-squared: 0.1346

F-statistic: 8.697 on 2 and 97 DF, p-value: 0.0003357

Note that the model-level and parameter-level tests and output is different—the  $p$ -value for the model is different than all of the parameter-level  $p$ -values.



# Model-Level Inference

Residual standard error: 7.092 on 97 degrees of freedom  
Multiple R-squared: 0.1521, Adjusted R-squared: 0.1346  
F-statistic: 8.697 on 2 and 97 DF, p-value: 0.0003357

$$H_0 : \beta_0 = \beta_1 = \beta_2 = \dots = \beta_k$$

The regression is statistically reliable,  $F(2, 97) = 8.70, p < 0.001$ . This suggests that *at least* one of the regression parameters is probably not 0.

$$H_0 : \rho^2 = 0$$

The regression is statistically reliable,  $F(2, 97) = 8.70, p < 0.001$ . This suggests that the model probably explains some of the variation in GPAs, in the population.

The model explains 15.2% of the variation in GPAs (in the sample).

# Regression Coefficients

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	63.2270	5.2398	12.067	< 2e-16	***
homework	0.9878	0.3609	2.737	0.00737	**
parentEd	0.8706	0.3842	2.266	0.02568	*

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

$$\hat{GPA} = 63.23 + 0.99(\text{homework}) + 0.87(\text{parentEd})$$

The slopes (of which there are now more than one) are referred to as *partial regression slopes* or *partial effects*.

## Interpretation of the Intercept

$$\hat{\text{GPA}} = 63.23 + 0.99(\text{homework}) + 0.87(\text{parentEd})$$

The intercept, 63.23, is the average (predicted) GPA for all students who spend 0 hours per week on homework **and** whose parent with the highest level of education is 0 years of schooling.

**Danger:** This is prediction falls outside the range of the data we used to fit the model given that our lowest homework value in the data was 1 and our lowest parentEd value in the data was 10.

## Interpretation of the Partial Effects

$$\hat{GPA} = 63.23 + 0.99(\text{homework}) + 0.87(\text{parentEd})$$

Each one-hour difference in the time students spend on homework is associated with a 0.99-unit difference in GPA...**controlling for** differences in parent education.

Each one-year difference in the level of parent education is associated with a 0.87-unit difference in GPA...**controlling for** differences in the time students spend on homework.

# Parameter-Level Inference

## Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	63.2270	5.2398	12.067	< 2e-16	***
homework	0.9878	0.3609	2.737	0.00737	**
parentEd	0.8706	0.3842	2.266	0.02568	*

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

The test for the intercept,  $t(97) = 12.07$ ,  $p < 0.001$ , is statistically reliable.  
This suggests that the population intercept is probably not 0.

$$H_0 : \beta_0 = 0$$

$$H_0 : \beta_1 = 0$$

$$H_0 : \beta_2 = 0$$


The test for the partial effect of time spent on homework,  $t(97) = 2.74$ ,  $p = 0.007$ , is statistically reliable. This suggests that in the population, differences in homework explain a statistically significant amount variation in GPA, **after accounting for** the variation explained by differences in parent education.

The test for the partial effect of parent education,  $t(97) = 2.27$ ,  $p = 0.026$ , is statistically reliable. This suggests that in the population, differences in parent education explain a statistically significant amount variation in GPA, **after accounting for** the variation explained by differences in time spent of homework.

If a partial effect is statistically reliable it means that predictor is statistically important in explaining variation in the outcome above and beyond what the other predictors in the model explain.

# Multiple Regression Model

The multiple regression model says that a case's outcome ( $Y$ ) is a function of two or more predictors ( $X_1, X_2, \dots, X_k$ ) and some amount of error.

$$Y = \beta_0 + \beta_1(X_1) + \beta_2(X_2) + \dots + \beta_k(X_k) + \epsilon$$


$$\hat{Y} = \beta_0 + \beta_1(X_1) + \beta_2(X_2) + \dots + \beta_k(X_k)$$

We estimate the regression coefficients using the sample data to get the observed regression equation,

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1(X_1) + \hat{\beta}_2(X_2) + \dots + \hat{\beta}_k(X_k)$$

## Which Predictor has a Stronger Effect on GPA?

$$\hat{GPA} = 63.23 + 0.99(\text{homework}) + 0.87(\text{parentEd})$$

Based on the values for the estimated regression coefficients, you might suggest that time spent on homework has the bigger effect on GPA...

...This is true, but we **cannot** make that judgment by looking at the size of the unscaled regression coefficients. The two predictors are measured using different scales and remember that the magnitude of a regression coefficient is influenced by the unit of measurement.

To compare the relative influence of the predictors in a model, we need to examine the **standardized regression coefficients**, or the **beta weights**.

This time, we will use the `scale()` function to create the z-scores for each variable. Also, rather than include the z-scores as additional predictors in the data frame, we will create them directly in the model.

```
> lm.b = lm(scale(gpa) ~ scale(homework) + scale(parentEd),  
  data = multReg)
```

```
> summary(lm.b)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	2.343e-17	9.303e-02	0.000	1.00000
scale(homework)	2.663e-01	9.730e-02	2.737	0.00737 **
scale(parentEd)	2.205e-01	9.730e-02	2.266	0.02568 *

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9303 on 97 degrees of freedom

Multiple R-squared: 0.1521, Adjusted R-squared: 0.1346

F-statistic: 8.697 on 2 and 97 DF, p-value: 0.0003357

$$\hat{z}_{\text{gpa}} = 0 + 0.266(z_{\text{homework}}) + 0.221(z_{\text{parentEd}})$$



$$\hat{z}_{\text{gpa}} = 0 + 0.266(z_{\text{homework}}) + 0.221(z_{\text{parentEd}})$$

Each one-standard deviation difference in the time students spend on homework is associated with a 0.266-standard deviation difference in GPA...**controlling** for differences in parent education.

Each one-standard deviation difference in the level of parent education is associated with a 0.221-standard deviation difference in GPA...**controlling** for differences in the time students spend on homework.

Since the two predictors are now on the same scale, they can be compared. Time spent on homework has more influence on GPAs than parent education. A good question might be whether this difference in the effects is statistically reliable...we will not worry about that here, but you can test that if it was of interest.

# Using the Unstandardized vs. the Standardized Coefficients

Both types of coefficients can be useful to applied researchers, but perhaps for different parts of the interpretational process.

## *Rules of Thumb for When to Interpret $b$ vs. $\beta$ -Weights*

### **Interpret $b$ (unstandardized coefficients)**

When variables are measured in a meaningful metric

To develop intervention or policy implications

To compare effects *across different* studies or samples

### **Interpret $\beta$ -weights (standardized coefficients)**

When the variables are not measured in a meaningful metric

To compare the relative effects of predictors *in the same* sample

### **Policy Decisions/Interventions**

Advice for the school board is probably more interpretable if you use the unstandardized coefficients. e.g., What are the effects on GPA if we increase the amount of homework by 5 hours a week? (Note: This assumes the metric for the variables is meaningful...)

### **Comparing Across Studies**

In different studies the variable you want to compare will often have a different distribution. For example, it is likely that the mean and SD for time spent on homework will be different across the studies (even if it is measured in hours/week in all the studies you are comparing). These differences in the distribution affect the magnitude of the  $\beta$ -Weights, not the  $b$ 's. Because of this it is better to interpret the unstandardized coefficients if your goal is to compare effects across studies.