Week 9: Lab 9 Multivariate Regression

Wenhao Jiang

Department of Sociology
New York University

November 4, 2022

Multivariate Regression

Multivariate Regression Basics

- We talked about bivariate regression in last week's lab
- For example, we use income as the main **dependent variable** and years of education as the main **independent variable** in a bivariate regression, assuming we are interested in the returns to education
- $y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i$

Multivariate Regression Basics

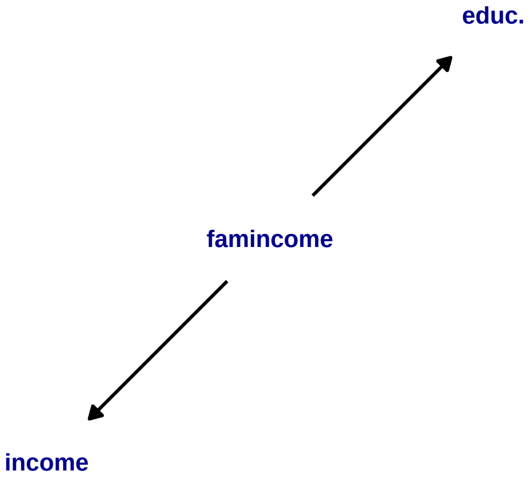
- We talked about bivariate regression in last week's lab
- For example, we use income as the main **dependent variable** and years of education as the main **independent variable** in a bivariate regression, assuming we are interested in the returns to education
- $y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i$
- We also noted that $\hat{\beta}_1$ may not capture the *causal* effect of education on income, as other factors including family background may affect both years of education and income
- In an extreme case, the positive association between education and income is completely driven by family income (family income -> more years of education & family income -> higher individual income)

Multivariate Regression Basics

- We talked about bivariate regression in last week's lab
- For example, we use income as the main **dependent variable** and years of education as the main **independent variable** in a bivariate regression, assuming we are interested in the returns to education
- $y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i$
- We also noted that $\hat{\beta}_1$ may not capture the *causal* effect of education on income, as other factors including family background may affect both years of education and income
- In an extreme case, the positive association between education and income is completely driven by family income (family income -> more years of education & family income -> higher individual income)
- This is called **confounding**, as typical cause of **spurious** correlation

Multivariate Regression Basics

- Confounding



Multivariate Regression Basics

- In most cases, **confounding** factors (e.g., family income) do not fully capture the original association of interest (e.g., the association between education and income)
- We therefore want to **control** for these confounding factors

Multivariate Regression Basics

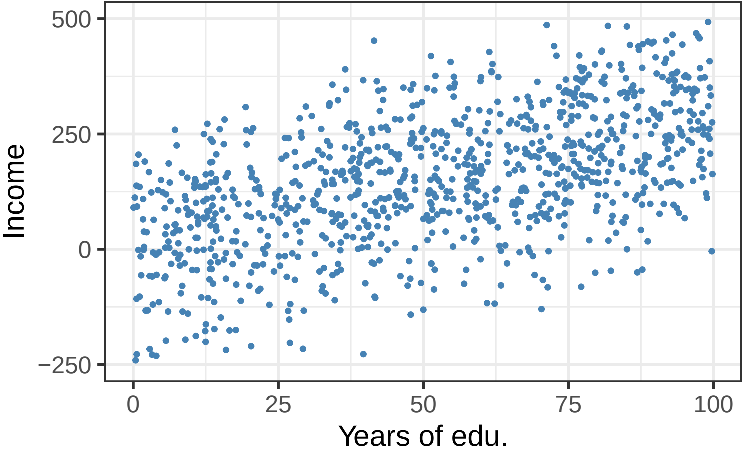
- In most cases, **confounding** factors (e.g., family income) do not fully capture the original association of interest (e.g., the association between education and income)
- We therefore want to **control** for these confounding factors
  - What does *control* mean? It means, conditioning on the same level of e.g. family income, what is the average asssociation between years of education and income

Multivariate Regression Basics

- In most cases, **confounding** factors (e.g., family income) do not fully capture the original association of interest (e.g., the association between education and income)
- We therefore want to **control** for these confounding factors
  - What does *control* mean? It means, conditioning on the same level of e.g. family income, what is the average asssociation between years of education and income
- Potential **confounding** effect is the main reason why we need to go beyond bivariate regression to multivariate regression

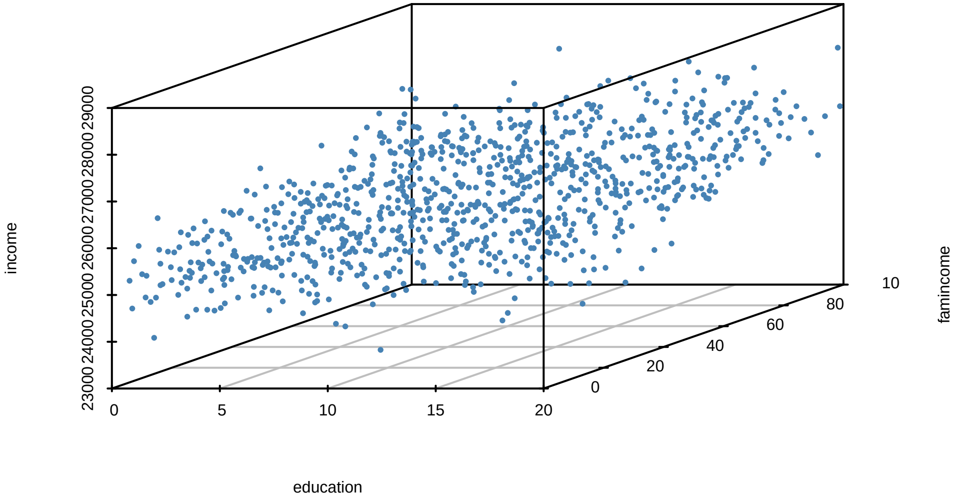
Multivariate Regression Basics

- We already know that a bivariate association, when plotted on a scatter plot, looks like



Multivariate Regression Basics

- Adding one more variable (i.e., two **independent** variables) would make the scatter plot look like:



Multivariate Regression Basics

- We see a positive association both between education and income, and between family income and income
- We want to estimate the association between education and income, conditioning on the same level of family income

Multivariate Regression Basics

- We see a positive association both between education and income, and between family income and income
- We want to estimate the association between education and income, conditioning on the same level of family income
- We cannot plot the points when there are three or more independent variables (i.e., we cannot imagine a 4-D or more-D case), but we can imagine the analogy

Multivariate Regression Basics

- We see a positive association both between education and income, and between family income and income
- We want to estimate the association between education and income, conditioning on the same level of family income
- We cannot plot the points when there are three or more independent variables (i.e., we cannot imagine a 4-D or more-D case), but we can imagine the analogy
- We use multivariate regression to estimate the associations
- $y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + e_i$

Exercise

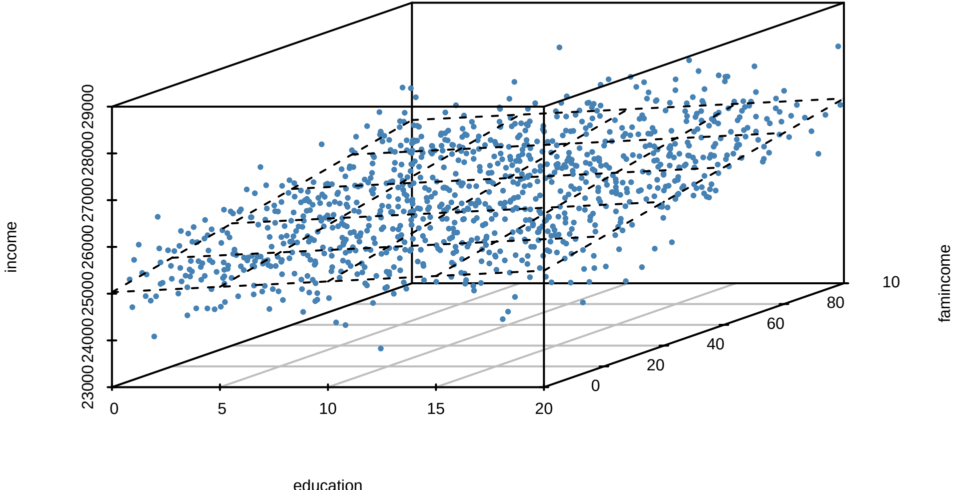
- True or False Statement
- Just as in the bivariate regression, we estimate $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ by minimizing $\sum_{i=1}^{n} e_i^2$, where $e_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i})$

Multivariate Regression Basics

- The estimate of a bivariate regression is a linear line, when using OLS (Ordinary Least Square)

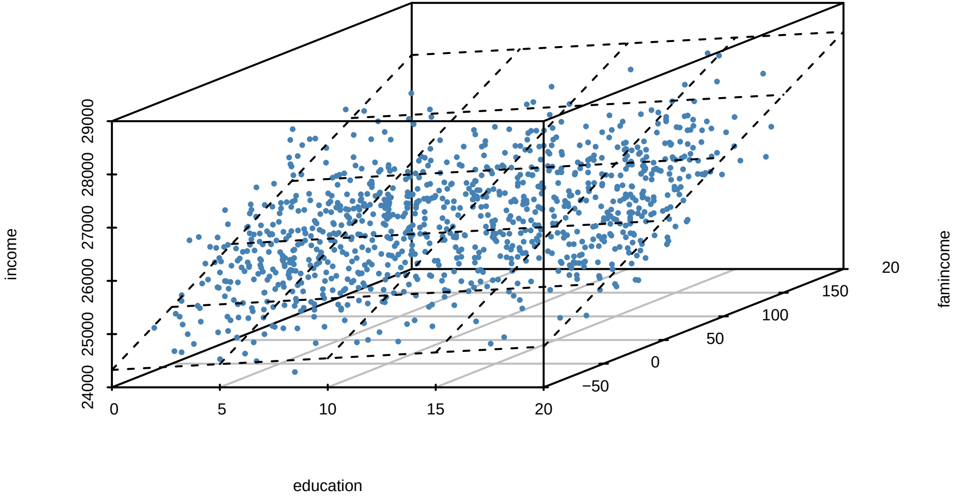
Multivariate Regression Basics

- The estimate of a bivariate regression is a linear line, when using OLS (Ordinary Least Square)
- The estimate of a trivariate regression is a plane, when using OLS



Multivariate Regression Basics

- The estimate of a trivariate regression is a plane, when using OLS



Multivariate Regression Basics

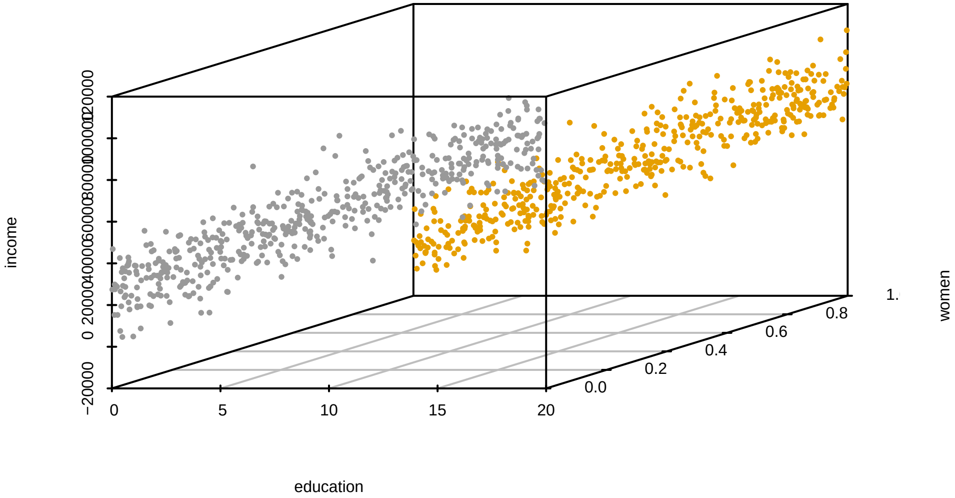
Table 1: The association between education and income

	Dependent variable:	
	income	
	(1)	(2)
education	39.80***	21.88***
	(3.93)	(2.80)
famincome		14.77***
		(0.46)
Constant	25,771.73***	25,065.81***
	(45.14)	(38.36)
Observations	1,000	1,000
Adjusted R^2	0.09	0.56
Note:	*p<0.1; **p<0.05; ***p<0.01	

Multivariate Regression Basics

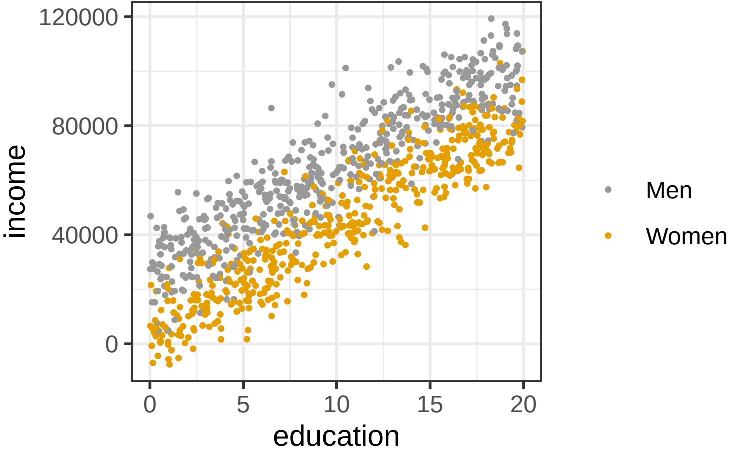
- A particular scenario is when the second independent variable is a binary (i.e., only two values available) variable (e.g., gender, although sociologically there may be more)
- $income_i = \hat{\beta}_0 + \hat{\beta}_1 educ_i + \hat{\beta}_2 women_i + e_i$
- Here $women_i = 1$ when the R is a woman, and $women_i = 0$ when the R is a man

Multivariate Regression Basics



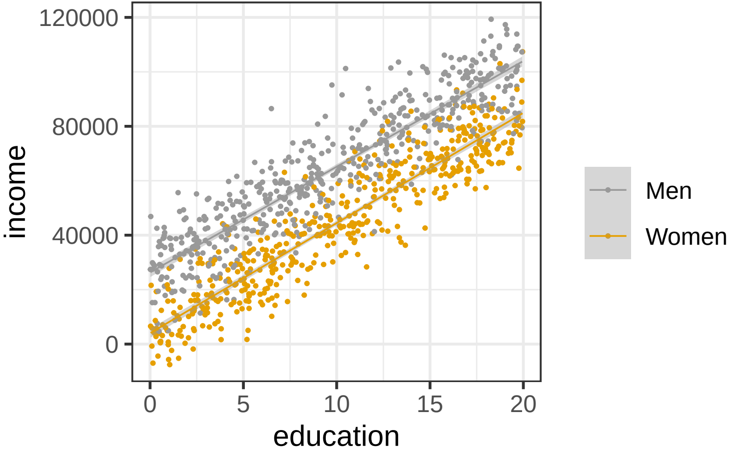
Multivariate Regression Basics

- In this particular scenario, we can also visualize the 3-D plot by a 2-D scatterplot



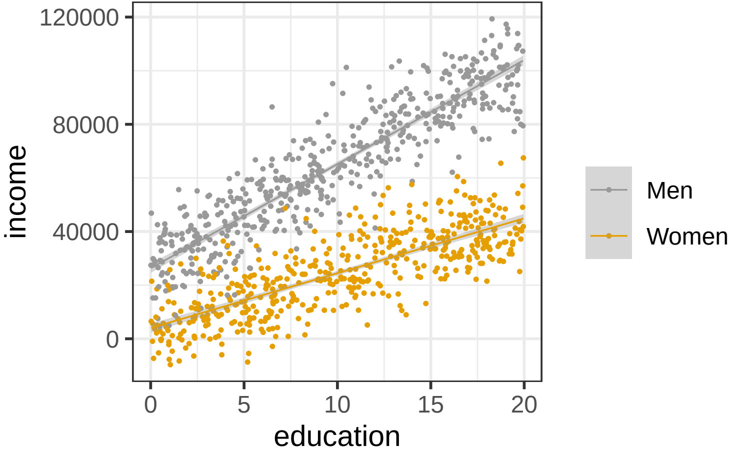
Multivariate Regression Basics

- Instead of fitting a multivariate regression, we can also fit the regression line to each gender
- Here, the slope for men and women are the same



Multivariate Regression Basics

- The slope for men and women can also differ



Multivariate Regression Intepretation

- Suppose in the above case, we find that
- $income_i = 5000 + 3000 educ_i - 10000 women_i + e_i$
- How do we intepret the results?

Multivariate Regression Intepretation

- Suppose in the above case, we find that
- $income_i = 5000 + 3000 educ_i - 10000 women_i + e_i$
- How do we intepret the results?
  - Conditioning on the same gender, an additional year of education is associated with $3000 more income. - Note that returns to education can differ by gender, so the estimate here is the average return across men and women.
  - For example, if the return for men is 4000 and 2000 for women, and the number of men and women is the same, we would still get $\hat{\beta}_1 = 3000$

Multivariate Regression Intepretation

- Suppose in the above case, we find that
- $income_i = 5000 + 3000 educ_i - 10000 women_i + e_i$
- How do we intepret the results?
  - Conditioning on the same gender, an additional year of education is associated with \$3000 more income. - Note that returns to education can differ by gender, so the estimate here is the average return across men and women.
  - For example, if the return for men is 4000 and 2000 for women, and the number of men and women is the same, we would still get $\hat{\beta}_1 = 3000$
  - Conditional on the same level of education, men on average earn \$10000 more income than women
  - Similarly, the gender pay gap can differ at each level of education, so the estimate here is the average gender gap across different levels of education

Multivariate Regression: Interaction

- When the slope for men and women differs (equivalently, the gender pay gap differs at each level of education), we do not observe it from the single equation: $income_i = 5000 + 3000 educ_i - 10000 women_i + e_i$
- We therefore want the **interaction** between gender and education, which literally means
- The effect of education on income depends on gender, and the effect of gender depends on education

Multivariate Regression: Interaction

- When the slope for men and women differs (equivalently, the gender pay gap differs at each level of education), we do not observe it from the single equation: $income_i = 5000 + 3000 educ_i - 10000 women_i + e_i$
- We therefore want the **interaction** between gender and education, which literally means
- The effect of education on income depends on gender, and the effect of gender depends on education
- How do we operationalize the interaction? We add a product of education and gender

Multivariate Regression: Interaction

- Without interaction: $income_i = \hat{\beta}_0 + \hat{\beta}_1 educ_i + \hat{\beta}_2 women_i + e_i$
- With interaction: $income_i = \hat{\beta}_0 + \hat{\beta}_1 educ_i + \hat{\beta}_2 women_i + \hat{\beta}_3 educ_i \times women_i + e_i$

Multivariate Regression: Interaction

- Without interaction: $income_i = \hat{\beta}_0 + \hat{\beta}_1 educ_i + \hat{\beta}_2 women_i + e_i$
- With interaction: $income_i = \hat{\beta}_0 + \hat{\beta}_1 educ_i + \hat{\beta}_2 women_i + \hat{\beta}_3 educ_i \times women_i + e_i$
- Without interaction: $\partial income_i / \partial educ_i = \hat{\beta}_1$
- Without interaction: $\partial income_i / \partial women_i = \hat{\beta}_2$

Multivariate Regression: Interaction

- Without interaction: $income_i = \hat{\beta}_0 + \hat{\beta}_1 educ_i + \hat{\beta}_2 women_i + e_i$
- With interaction: $income_i = \hat{\beta}_0 + \hat{\beta}_1 educ_i + \hat{\beta}_2 women_i + \hat{\beta}_3 educ_i \times women_i + e_i$
- Without interaction: $\partial income_i / \partial educ_i = \hat{\beta}_1$
- Without interaction: $\partial income_i / \partial women_i = \hat{\beta}_2$
- With interaction: $\partial income_i / \partial educ_i = \hat{\beta}_1 + \hat{\beta}_3 women_i$
- With interaction: $\partial income_i / \partial women_i = \hat{\beta}_2 + \hat{\beta}_3 educ_i$

Multivariate Regression: Interaction

- Without interaction: $income_i = \hat{\beta}_0 + \hat{\beta}_1 educ_i + \hat{\beta}_2 women_i + e_i$
- With interaction: $income_i = \hat{\beta}_0 + \hat{\beta}_1 educ_i + \hat{\beta}_2 women_i + \hat{\beta}_3 educ_i \times women_i + e_i$
- Without interaction: $\partial income_i / \partial educ_i = \hat{\beta}_1$
- Without interaction: $\partial income_i / \partial women_i = \hat{\beta}_2$
- With interaction: $\partial income_i / \partial educ_i = \hat{\beta}_1 + \hat{\beta}_3 women_i$
- With interaction: $\partial income_i / \partial women_i = \hat{\beta}_2 + \hat{\beta}_3 educ_i$
- When we look at the effect of education on income by gender, using an interaction is the same as regressing income on education by gender separately

Multivariate Regression: Interaction

- Without interaction: $income_i = \hat{\beta}_0 + \hat{\beta}_1 educ_i + \hat{\beta}_2 women_i + e_i$
- With interaction: $income_i = \hat{\beta}_0 + \hat{\beta}_1 educ_i + \hat{\beta}_2 women_i + \hat{\beta}_3 educ_i \times women_i + e_i$
- Without interaction: $\partial income_i / \partial educ_i = \hat{\beta}_1$
- Without interaction: $\partial income_i / \partial women_i = \hat{\beta}_2$
- With interaction: $\partial income_i / \partial educ_i = \hat{\beta}_1 + \hat{\beta}_3 women_i$
- With interaction: $\partial income_i / \partial women_i = \hat{\beta}_2 + \hat{\beta}_3 educ_i$
- When we look at the effect of education on income by gender, using an interaction is the same as regressing income on education by gender separately

Multivariate Regression: Interaction

- Do not confuse interaction with **mediation** and **confounding**. In the case of e.g. estimating the effect of gender on attitudes towards abortion, religious belief can never be a confounding factor. Why?

Multivariate Regression: Interaction

- Do not confuse interaction with **mediation** and **confounding**. In the case of e.g. estimating the effect of gender on attitudes towards abortion, religious belief can never be a confounding factor. Why?
- In the case of e.g. estimating the effect of gender on attitudes towards abortion, religious belief likely interacts with gender (which is never observed if the regression does not create the product), or mediates the effect of gender on abortion attitudes. (Go back to Page 4 of Mike's Review slides to see if you understand why the regression table indicates a mediation)

Multivariate Regression: Estimates

- Again, we estimate $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ by minimizing $\sum_{i=1}^{n} e_i^2$, where $e_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i})$
- In bivariate regression, $\hat{\beta}_1 = \frac{Cov(x_i, y_i)}{Var(x_1)}$ and $\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$

Multivariate Regression: Estimates

- Again, we estimate $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ by minimizing $\sum_{i=1}^{n} e_i^2$, where $e_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i})$
- In bivariate regression, $\hat{\beta}_1 = \frac{Cov(x_i, y_i)}{Var(x_1)}$ and $\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$
- In multivariate regression, vector $\hat{\boldsymbol{\beta}}$ is $(\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{Y}$. You will need a whole semester's Linear Algebra to understand the magic behind it (definitely not required in the course although it appears on the lecture notes)

Multivariate Regression: Estimates

- Again, we estimate $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ by minimizing $\sum_{i=1}^{n} e_i^2$, where $e_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i})$
- In bivariate regression, $\hat{\beta}_1 = \frac{Cov(x_i, y_i)}{Var(x_1)}$ and $\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$
- In multivariate regression, vector $\hat{\boldsymbol{\beta}}$ is $(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$. You will need a whole semester's Linear Algebra to understand the magic behind it (definitely not required in the course although it appears on the lecture notes)
- But as a little anatomy of multivariate regression $y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + ... + \hat{\beta}_k x_{ki} + e_i$
- $\hat{\beta}_k = \frac{Cov(\tilde{x}_{ki}, y_i)}{Var(\tilde{x}_{ki})}$, where $\tilde{x}_{ki}$ is the residual from a regression of x_{ki} on all other covariates (i.e., $x_{1i} ... x_{k-1i}$)

Multivariate Regression: Predicted Values

- We can fit/predict values from regression equation
- For example, in the regression, $income_i = 5000 + 3000 educ_i - 10000 women_i + e_i$, what is the predicted income for a man with 12 years of education?

Multivariate Regression: Predicted Values

- We can fit/predict values from regression equation
- For example, in the regression, $income_i = 5000 + 3000 educ_i - 10000 women_i + e_i$, what is the predicted income for a man with 12 years of education?
- What about the predicted income for a woman with 16 years of education?

Multivariate Regression: Predicted Values

- We can fit/predict values from regression equation
- For example, in the regression, $income_i = 5000 + 3000educ_i - 10000women_i + e_i$, what is the predicted income for a man with 12 years of education?
- What about the predicted income for a woman with 16 years of education?
- The predictions are related to the calculation of $R^2 = Var(\hat{y}_i)/Var(y_i)$, where $\hat{y}_i$ are the predicted values

Exercise

R Squared

- True or False statement
- R^2 cannot be greater or equal than 1

R Squared

- True or False statement
- R^2 cannot be greater or equal than 1
- When x_i and y_i are indpendent with each other, i.e., x_i has no predicting power at all, $R^2 = 0$

t-score

- True or False statement
- The formula and the calculation of the $t-$score is the same as the $z-$score

t-score

- True or False statement
- The formula and the calculation of the $t-$score is the same as the $z-$score
  - $t = \frac{point\ estimate - null\ hypothesis}{SE}$

t-score

- True or False statement
- The formula and the calculation of the $t-$score is the same as the $z-$score
  - $t = \frac{point\ estimate - null\ hypothesis}{SE}$
- True or False statement
- At the same Significance Level (e.g., $\alpha = 0.05$), the corresponding t-score is smaller than the $z-$score (e.g., 1.96)

t-score

- True or False statement
- The formula and the calculation of the $t-$score is the same as the $z-$score
  - $t = \frac{point\ estimate - null\ hypothesis}{SE}$
- True or False statement
- At the same Significance Level (e.g., $\alpha = 0.05$), the corresponding t-score is smaller than the $z-$score (e.g., 1.96)
- True or False statement
- In regression analysis, the null hypothesis is $\beta_k = 0$, and $t = \frac{\hat{\beta}_k - 0}{SE_{\hat{\beta}_k}}$

t-score and regression

- What is the null hypothesis for the slope of `women`?

Table 2: The association between education and income

	Dependent variable:	
	income	
	(1)	(2)
education	3,962.73*** (54.28)	3,990.69*** (74.12)
women	−711.46 (633.14)	906.82 (1,285.06)
education:women		−1,021.35*** (109.32)
Constant	25,554.96*** (685.84)	24,945.19*** (847.68)
Observations	1,000	1,000
Adjusted R^2	0.84	0.81
Note:		*p<0.1; **p<0.05; ***p<0.01

t-score and regression

t = (point est. - null)/SE = (529.95 - 0)/634.60 = 0.835 < 1.96 < 2.12

- What is the $t-$score for the slope of `women` in column 1?

Table 3: The association between education and income

	Dependent variable:	
	income	
	(1)	(2)
education	3,932.42*** (56.08)	3,884.01*** (75.58)
women	529.95 (634.60)	−2,317.80* (1,246.42)
education:women		−837.64*** (108.54)
Constant	25,197.44*** (742.07)	26,313.40*** (852.57)
Observations	1,000	1,000
Adjusted R^2	0.83	0.81
Note:		*p<0.1; **p<0.05; ***p<0.01

t-score and regression

- Do we reject the null hypothesis at $\alpha = 0.05$?

Table 4: The association between education and income

	Dependent variable:	
	income	
	(1)	(2)
education	3,962.73*** (54.28)	3,884.01*** (75.58)
women	−711.46 (633.14)	−2,317.80* (1,246.42)
education:women		−837.64*** (108.54)
Constant	25,554.96*** (685.84)	26,313.40*** (852.57)
Observations	1,000	1,000
Adjusted R^2	0.84	0.81
Note:		*p<0.1; **p<0.05; ***p<0.01

t-score and regression

income = 26313 + 3884 education - 2317.8 women - 837.6 education*women

d income / d education = 3884 - 837.6*women

d income / d women = -2317.8 - 837.64*education

- Does the association between education and income depend on gender?

Table 5: The association between education and income

	Dependent variable:	
	income	
	(1)	(2)
education	3,962.73*** (54.28)	3,884.01*** (75.58)
women	−711.46 (633.14)	−2,317.80* (1,246.42)
education:women		−837.64*** (108.54)
Constant	25,554.96*** (685.84)	26,313.40*** (852.57)
Observations	1,000	1,000
Adjusted R^2	0.84	0.81
Note:		*p<0.1; **p<0.05; ***p<0.01

Regression Estimates

- To study the association between education (x_i) and income (y_i), a researcher collected $n = 501$ individuals, finding that $\bar{x} = 15$, $\bar{y} = 45000$, $\frac{1}{500}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = 10000$, $\frac{1}{500}\sum_{i=1}^{n}(x_i - \bar{x})^2 = 2.5$, and $\frac{1}{500}\sum_{i=1}^{n}(y_i - \bar{y})^2 = 2500$. She wants to estimate the following regression $y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i$
- What is $\hat{\beta}_1$?
- What is $\hat{\beta}_0$?