Multivariate Regression

Econ 2560, Fall 2023

Prof. Josh Abel

(Chapters 6, 7.1)

Introduction

- So far, we have estimated means conditional on 1 X variable
  - $E[Y|X] \rightarrow E[Y|X_1]$
    - Linear
    - Nonlinear, including higher-order polynomials
    - Non-parametric
- Now will consider conditioning on multiple variables
  - $E[Y|X] \rightarrow E[Y|X_1, X_2, ..., X_K]$

Motivation

- Why use multiple X variables?
  - Because you can!
    - Useful to have more information when estimating a mean
  - May strengthen the causal interpretation of coefficients on individual variables

Quick discussion of causality

- Have not discussed causation yet
  - Will discuss in much more detail later
- For now, suffice to say that this probably does not give the right causal effect:
  - $E[\text{Earnings}|\text{Education}] = \beta_0^U + \beta_1^U \cdot \text{Education}$
- Why not?

Quick discussion of causality

- Have not discussed causation yet
  - Will discuss in much more detail later
- For now, suffice to say that this probably does not give the right causal effect:
  - $E[\text{Earnings}|\text{Education}] = \beta_0^U + \beta_1^U \cdot \text{Education}$
- Why not?
  - Maybe more "connected" people get more schooling and have better access to jobs
  - Maybe "smarter" people find it easier to advance in school and excel at work
  - In either case, people with more education will have higher earnings, even without any causal relationship

A multivariate regression function

$$E[\text{Earnings}_i|X_i] = \beta_0^M + \beta_1^M \cdot \text{Education}_i + \beta_2^M \cdot \text{AFQT}_i$$

A multivariate regression function

$$E[\text{Earnings}_i|X_i] = \beta_0^M + \beta_1^M \cdot \text{Education}_i + \beta_2^M \cdot \text{AFQT}_i$$

- OLS chooses $\hat{\beta}_0^M$, $\hat{\beta}_1^M$, and $\hat{\beta}_2^M$ such that:

$$E[\hat{u}_i \cdot \text{Education}_i] = 0$$
$$E[\hat{u}_i \cdot \text{AFQT}_i] = 0$$
$$E[\hat{u}_i] = 0$$

where $\hat{u}_i = Y_i - [\hat{\beta}_0^M + \hat{\beta}_1^M \cdot \text{Education}_i + \hat{\beta}_2^M \cdot \text{AFQT}_i]$

Interpreting a multivariate regression function

$$E[\text{Earnings}_i|X_i] = \beta_0^M + \beta_1^M \cdot \text{Education}_i + \beta_2^M \cdot \text{AFQT}_i$$

- Let's suppose AFQT is a perfect measure of how "smart" someone is
  - Not true...
- How can we interpret β_1^M?

"Holding constant" interpretation

$$E[\text{Earnings}_i|X_i] = \beta_0^M + \beta_1^M \cdot \text{Education}_i + \beta_2^M \cdot \text{AFQT}_i$$

- If AFQT stays constant but Education increases by 1 year, how much does expected Earnings increase?

"Holding constant" interpretation

$$E[\text{Earnings}_i|X_i] = \beta_0^M + \beta_1^M \cdot \text{Education}_i + \beta_2^M \cdot \text{AFQT}_i$$

- If AFQT stays constant but Education increases by 1 year, how much does expected Earnings increase?
- β_1^M
  - β_1^M is the (predictive) effect of Education "holding AFQT constant"

"Holding constant" interpretation

$$E[\text{Earnings}_i|X_i] = \beta_0^M + \beta_1^M \cdot \text{Education}_i + \beta_2^M \cdot \text{AFQT}_i$$

- If AFQT stays constant but Education increases by 1 year, how much does expected Earnings increase?
- β_1^M
  - β_1^M is the (predictive) effect of Education "holding AFQT constant"
- Can think of a partial derivative from calculus:

$$\frac{\partial E[\text{Earnings}_i|\text{Education}_i, \text{AFQT}_i]}{\partial \text{Education}_i} = \beta_1^M$$

"Holding constant" interpretation (2)

$$E[\text{Earnings}_i|X_i] = \beta_0^M + \beta_1^M \cdot \text{Education}_i + \beta_2^M \cdot \text{AFQT}_i$$

- Can we think of β_1^M as being the causal effect of education now?

"Holding constant" interpretation (2)

$$E[\text{Earnings}_i|X_i] = \beta_0^M + \beta_1^M \cdot \text{Education}_i + \beta_2^M \cdot \text{AFQT}_i$$

- Can we think of β_1^M as being the causal effect of education now?
- Probably not
  - Education might have associations with factors other than AFQT that drive β_1^M
- Still, this is cleaner than the univariate case
  - At least β_1^M is (mostly) not being driven by AFQT

"Residual regression" interpretation

$$E[\text{Earnings}_i|\text{Education}_i, \text{AFQT}_i] = \hat{\beta}_0^M + \hat{\beta}_1^M \cdot \text{Education}_i + \hat{\beta}_2^M \cdot \text{AFQT}_i$$

"Residual regression" interpretation

$$E[\text{Earnings}_i|\text{Education}_i, \text{AFQT}_i] = \hat{\beta}_0^M + \hat{\beta}_1^M \cdot \text{Education}_i + \hat{\beta}_2^M \cdot \text{AFQT}_i$$

- Now consider the auxiliary regression of one regressor on another:

$$E[\text{Education}_i|\text{AFQT}_i] = \hat{\alpha}_0 + \hat{\alpha}_1 \cdot \text{AFQT}_i,$$

and let $\hat{u}_i^{X_1}$ be the residual from this regression.

"Residual regression" interpretation

$$E[\text{Earnings}_i|\text{Education}_i, \text{AFQT}_i] = \hat{\beta}_0^M + \hat{\beta}_1^M \cdot \text{Education}_i + \hat{\beta}_2^M \cdot \text{AFQT}_i$$

- Now consider the auxiliary regression of one regressor on another:

$$E[\text{Education}_i|\text{AFQT}_i] = \hat{\alpha}_0 + \hat{\alpha}_1 \cdot \text{AFQT}_i,$$

  and let $\hat{u}_i^{X_1}$ be the residual from this regression.
- You can (but won't have to) show the following:

$$E[\text{Earnings}_i|\hat{u}_i^{X_1}] = \kappa_0 + \hat{\beta}_1^M \cdot \hat{u}_i^{X_1}$$

"Residual regression" interpretation (2)

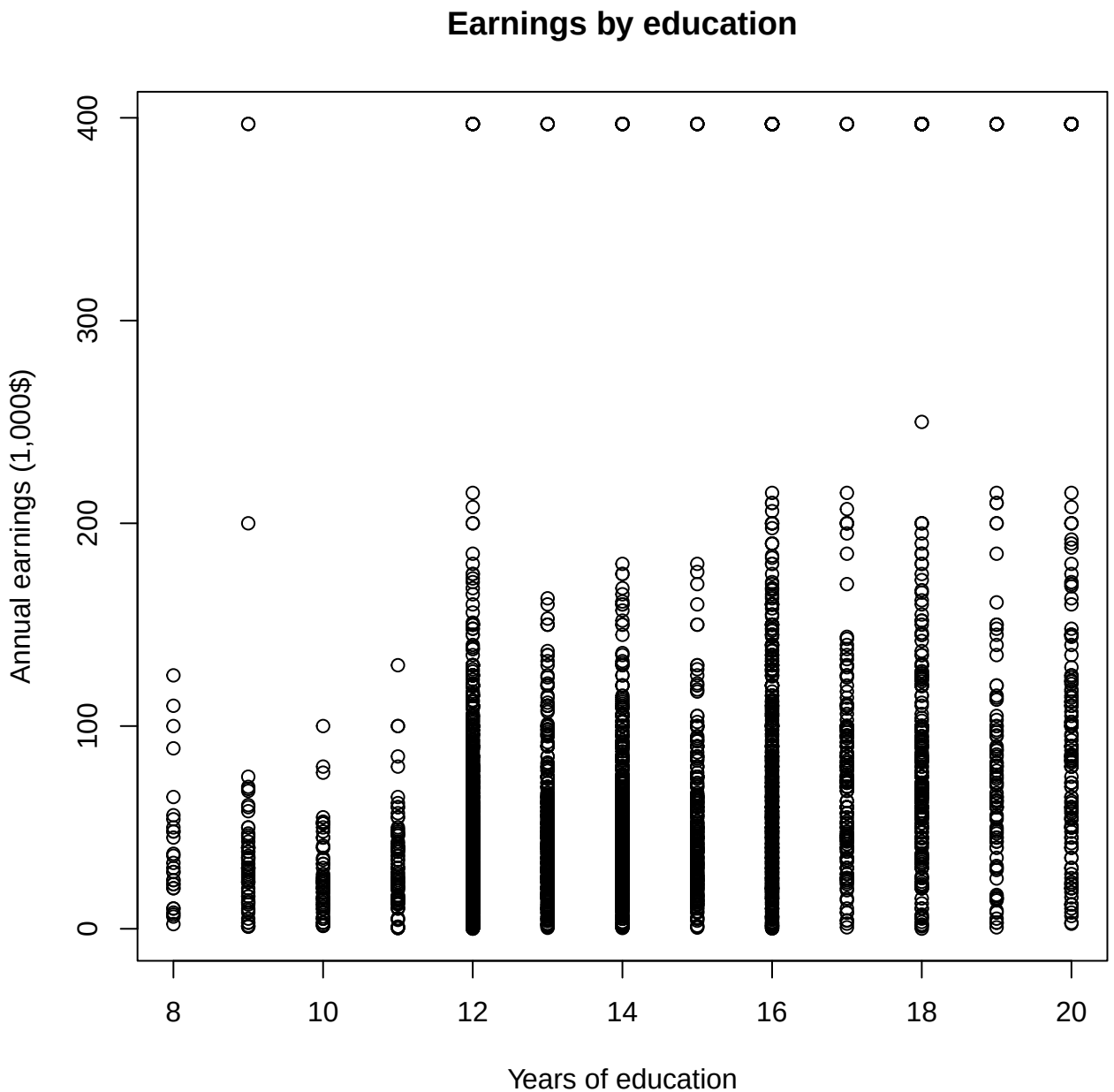
$$E[Y_i|X_i] = \hat{\beta}_0^M + \hat{\beta}_1^M \cdot X_{1i} + \hat{\beta}_2^M \cdot X_{2i}$$

- In words: β_1^M (or any other β^M) from a multivariate regression can be estimated as follows:
  - "Residualize" X_1 on all other regressors (call it $\hat{u}_i^{X_1}$)
  - Regress Y on $\hat{u}_i^{X_1}$: coefficient will be same β_1^M from above equation
- **Key interpretation: β_1^M measures the effect on Y of the portion of X_1 that cannot be explained by other variables**
  - β_1^M is "identified off of" the "residual variation" of X_1
  - Similarly, β_2^M is identified off of the residual variation of X_2

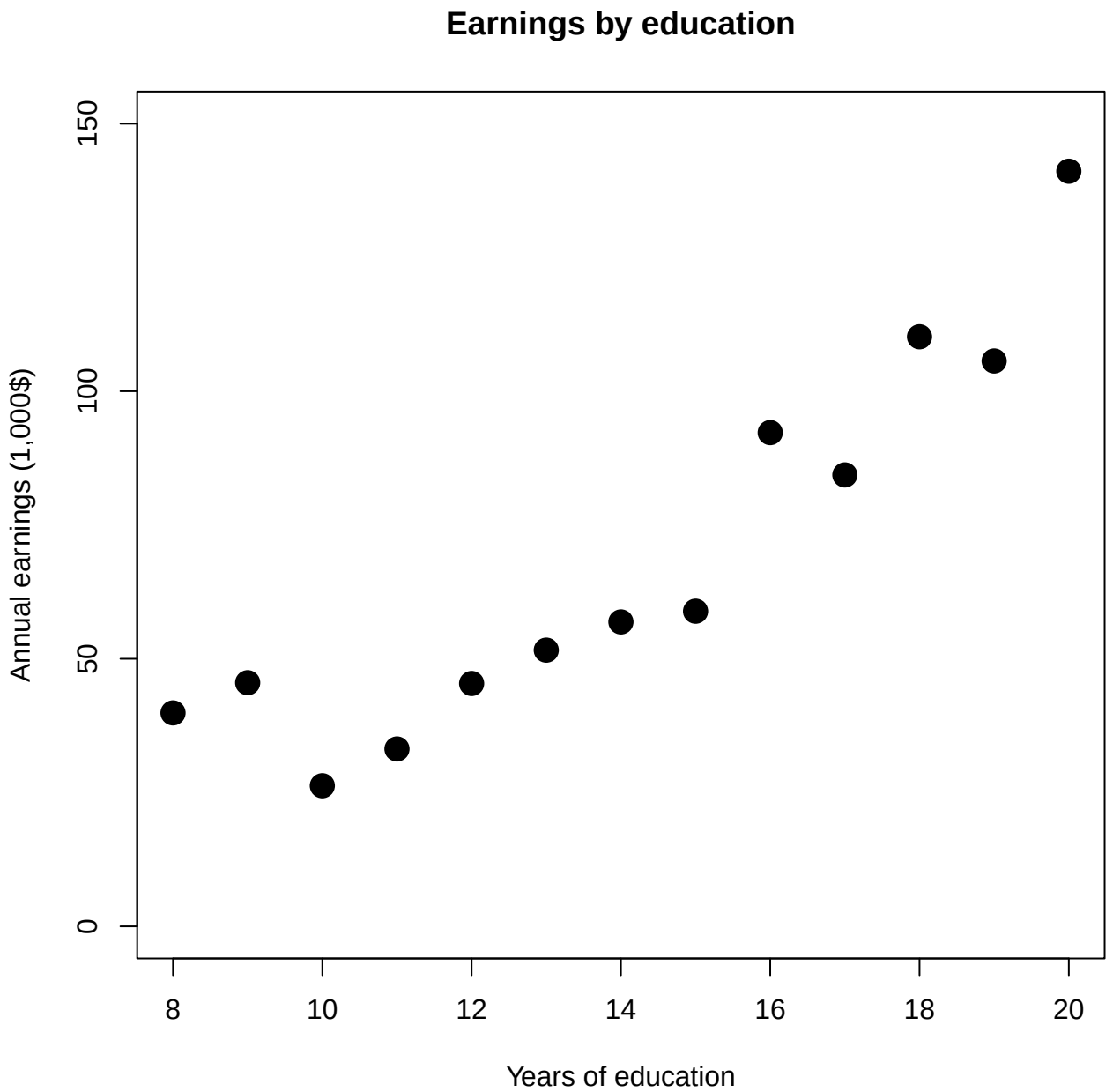
Income and education regressions

Outcome: Annual earnings (1,000$s)		
	Univariate	Multivariate
Constant	-75.7	-53.2
Education	10.1	6.7
AFQT		0.5

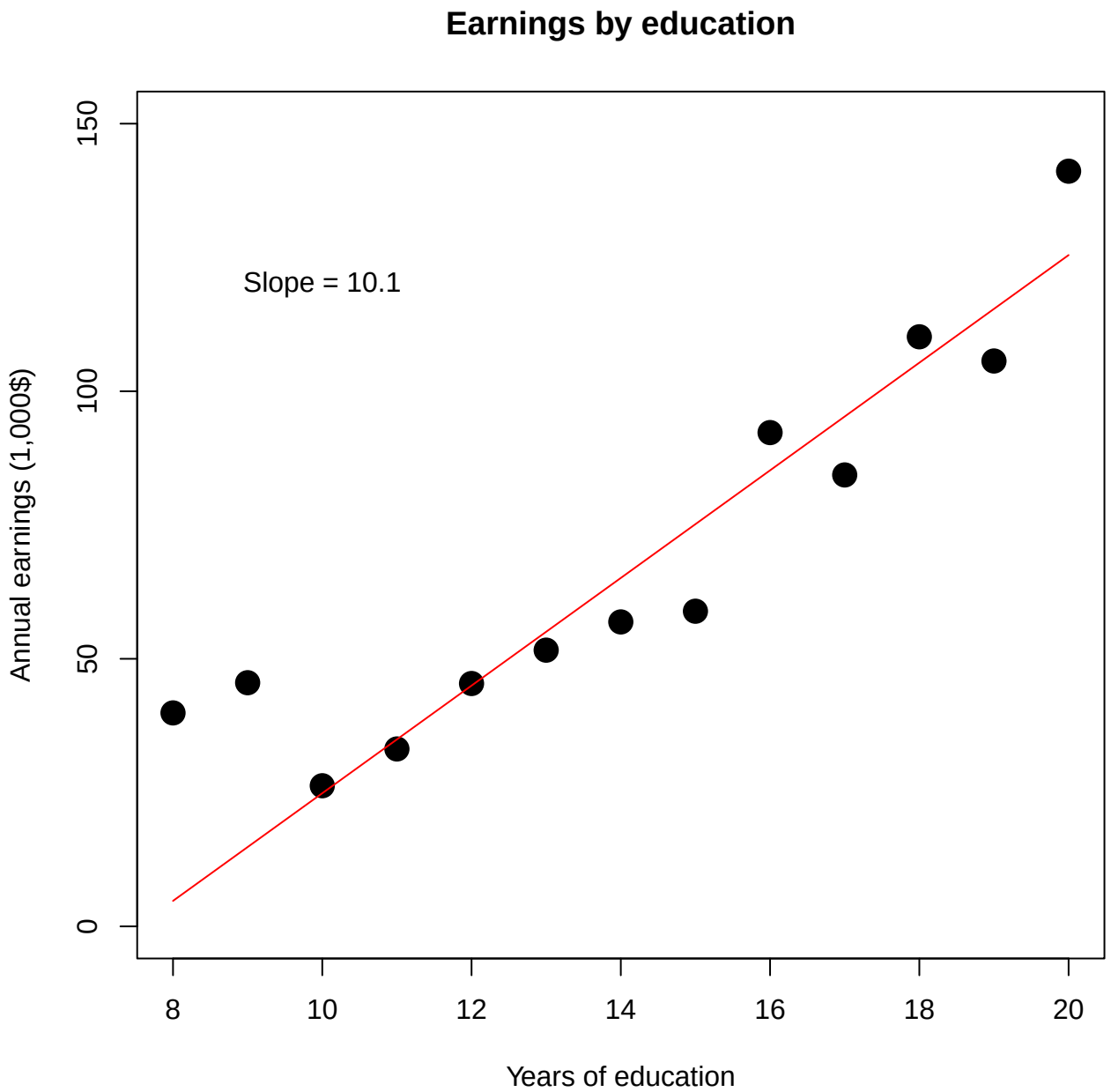
Income and education



Income and education



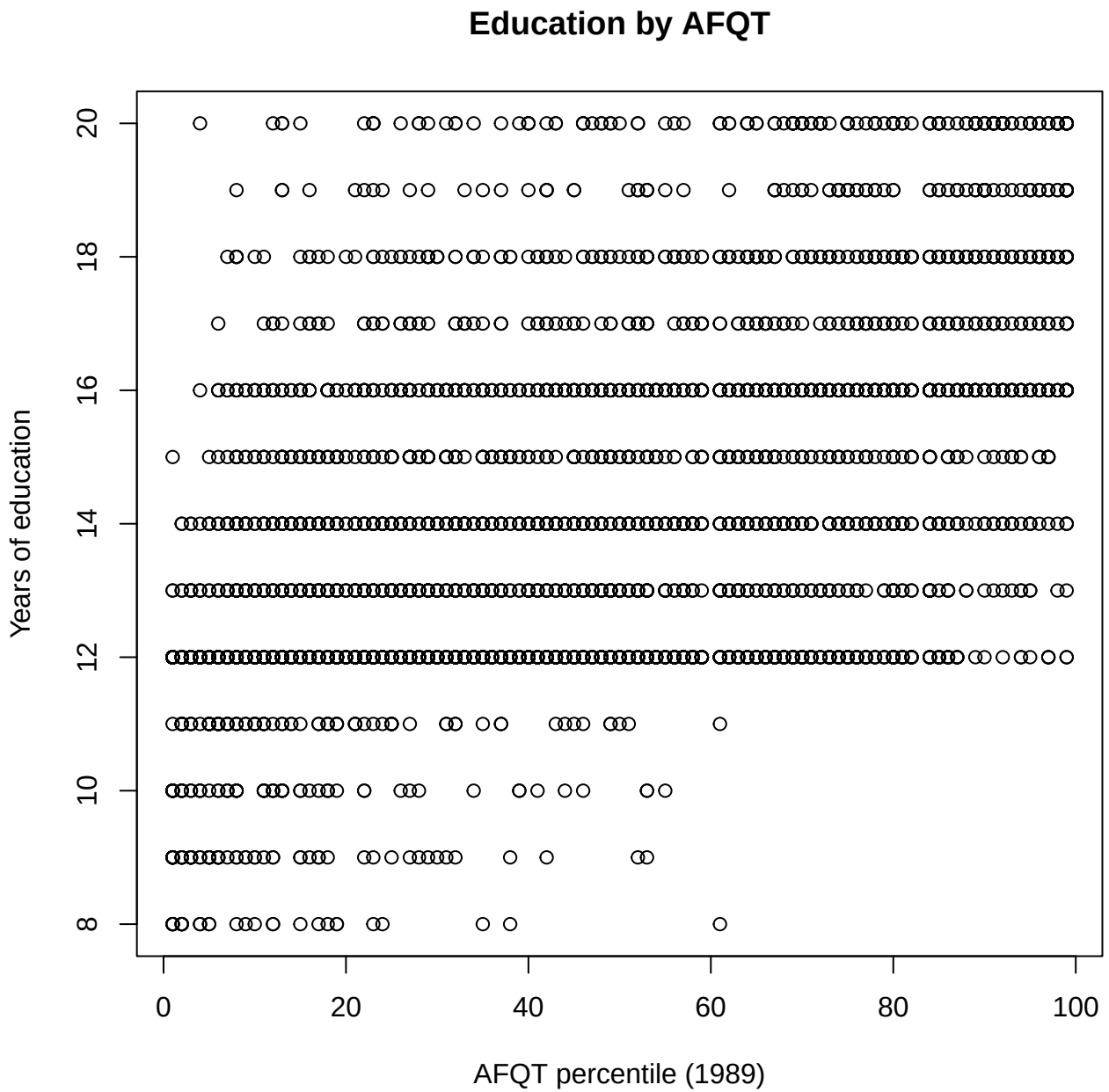
Income and education



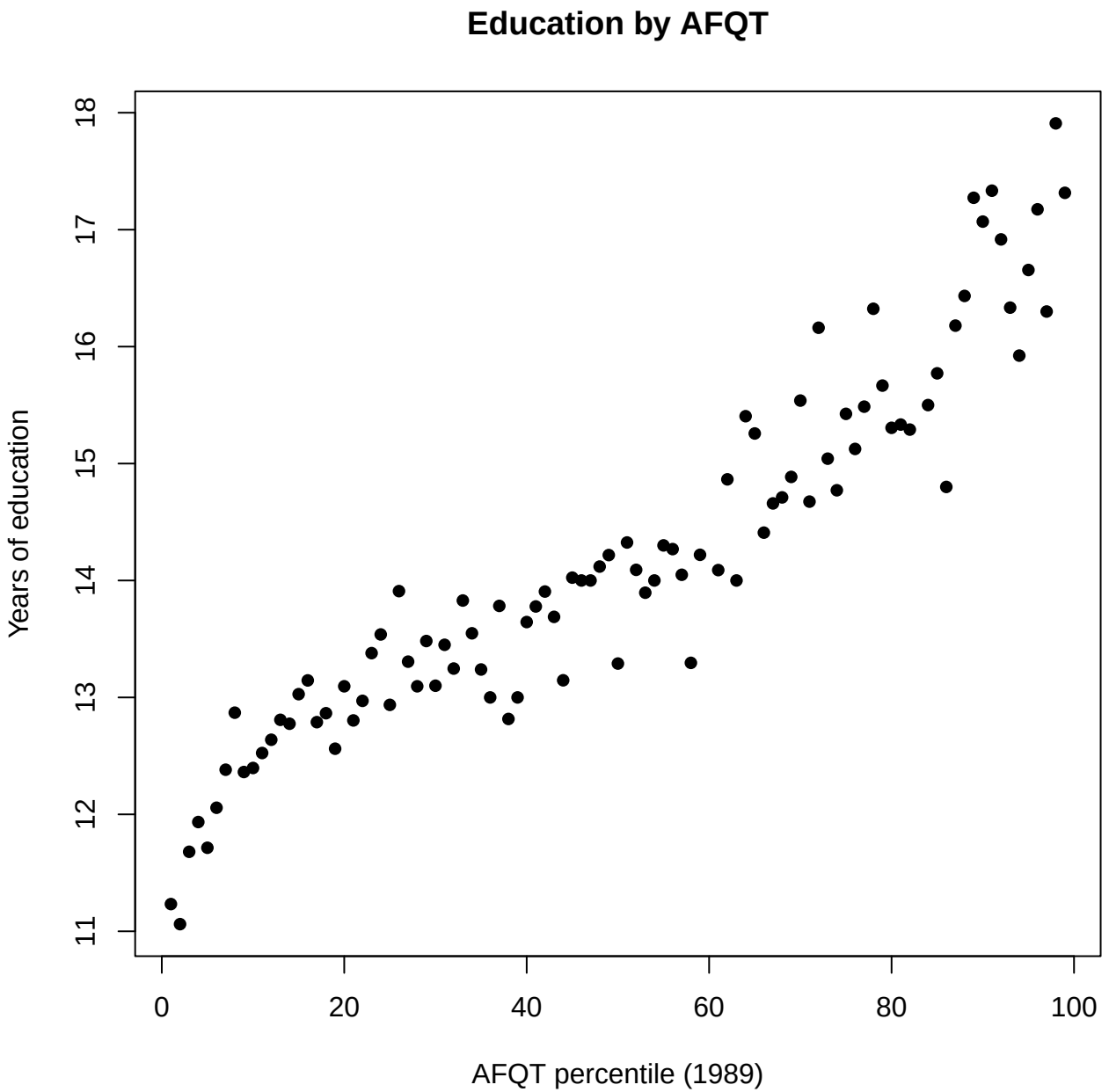
Income and education regressions

Outcome: Annual earnings (1,000$s)		
	Univariate	Multivariate
Constant	-75.7	-53.2
Education	10.1	6.7
AFQT		0.5

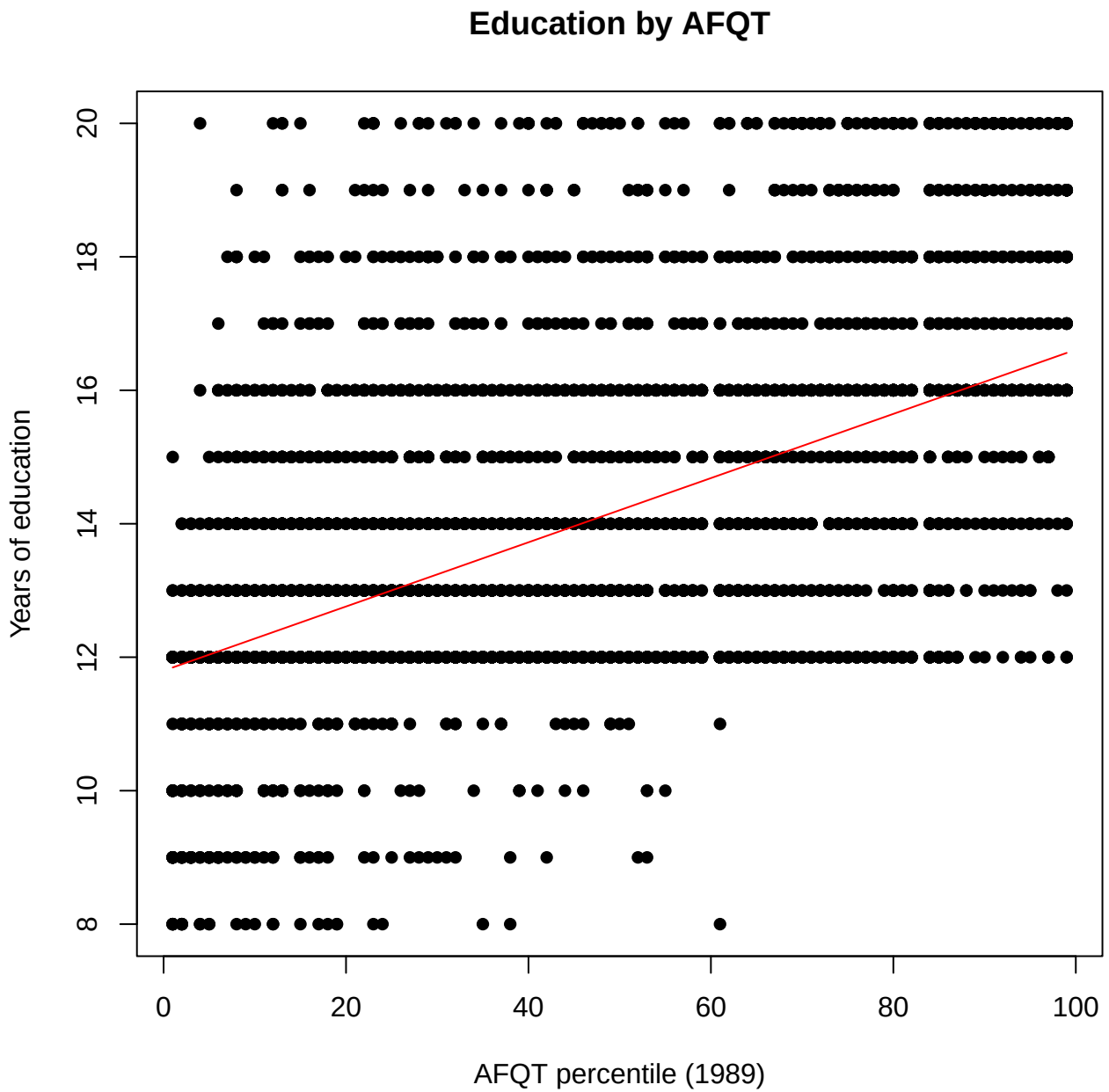
Education by AFQT



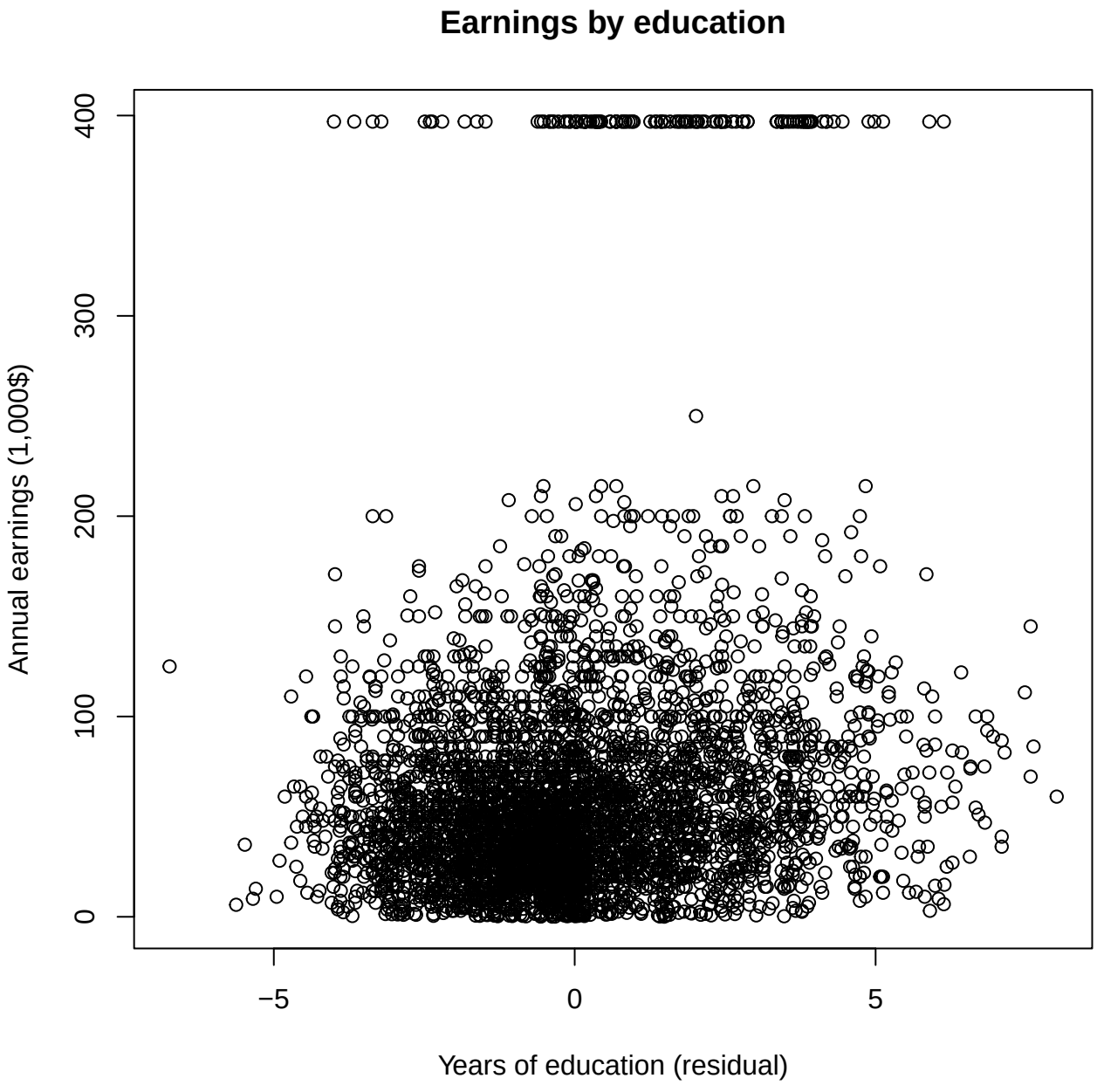
Education by AFQT



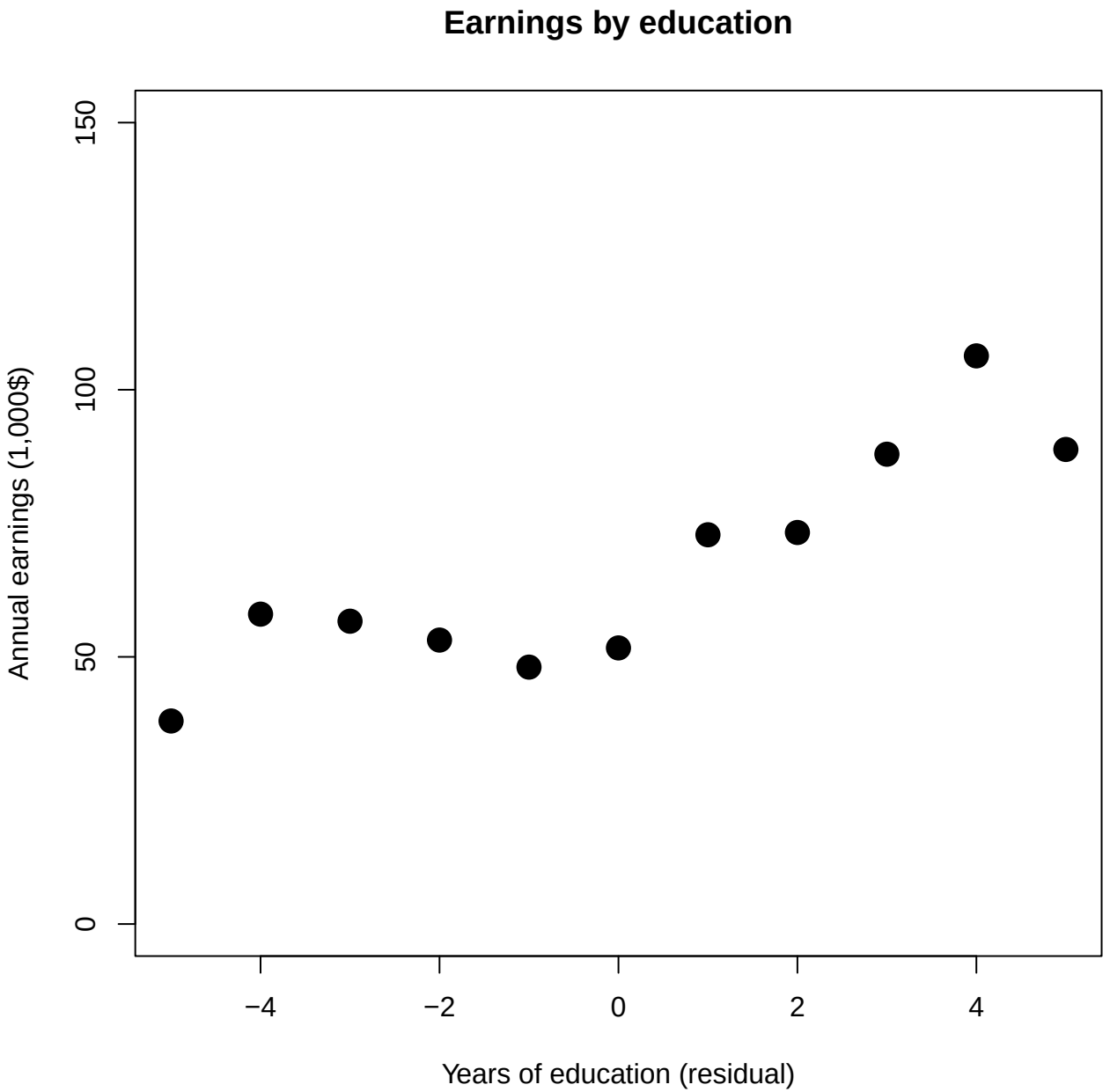
Education by AFQT



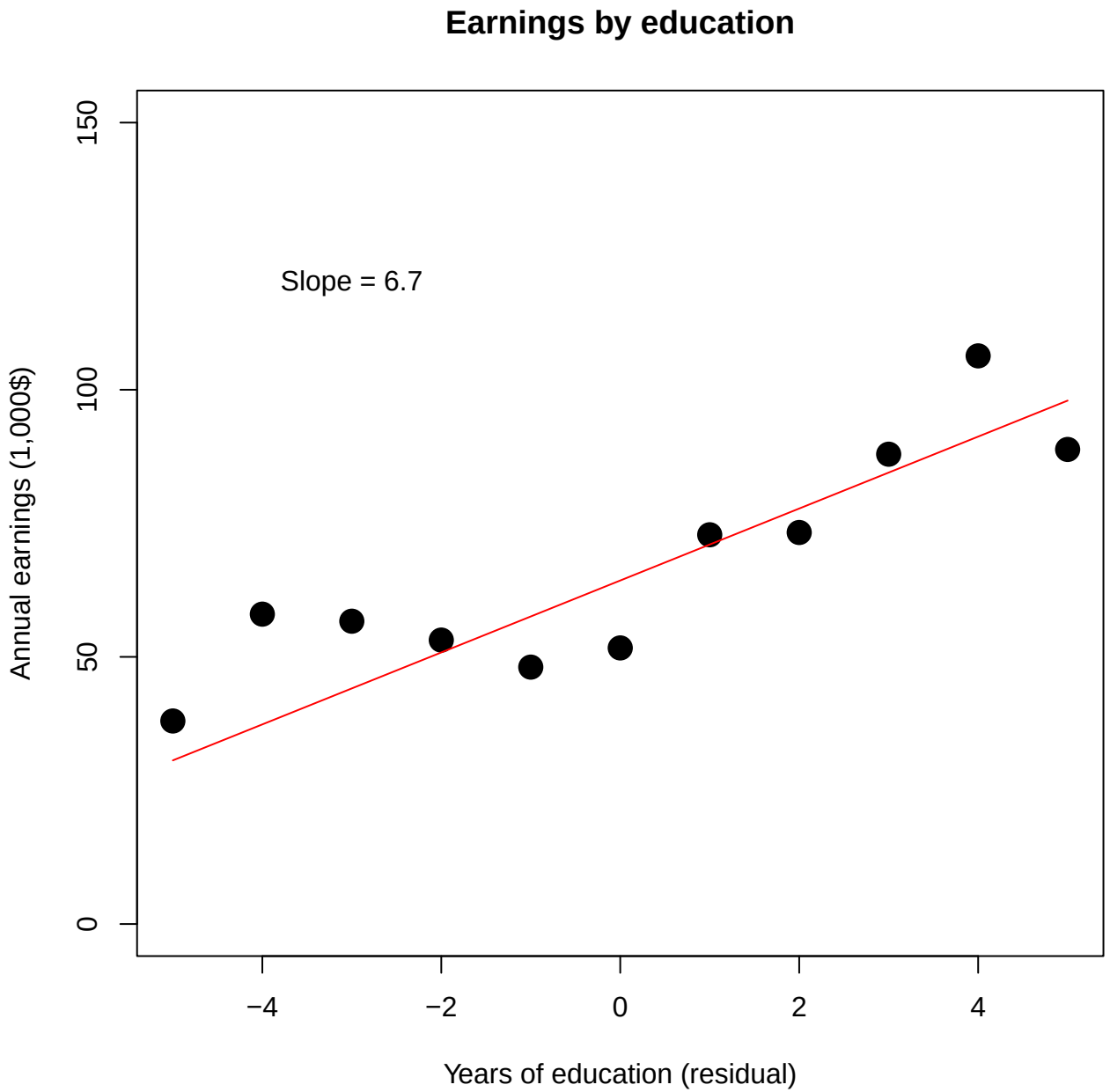
Residual regression



Residual regression



Residual regression



Income and education regressions

Outcome: Annual earnings (1,000$s)		
	Univariate	Multivariate
Constant	-75.7	-53.2
Education	10.1	6.7
AFQT		0.5

Omitted Variable "Bias"

- Frequently, we don't have data on every variable we'd like
- For instance, suppose we don't observe AFQT for this regression:

$$E[\text{Income}_i|\text{Education}_i, \text{AFQT}_i] = \beta_0^M + \beta_1^M \cdot \text{Education}_i + \beta_2^M \cdot \text{AFQT}_i$$

- We then have no choice but to estimate this equation:

$$E[\text{Income}_i|\text{Education}_i] = \beta_0^U + \beta_1^U \cdot \text{Education}_i$$

- Want to consider whether $\beta_1^U = \beta_1^M$.
  - Seems unlikely: β_1^U is estimated with all variation in Education while β_1^M only uses *residual variation*
  - It turns we can be more precise about this.

Omitted Variable "Bias" (2)

$$E[\text{Income}_i|\text{Education}_i, \text{AFQT}_i] = \beta_0^M + \beta_1^M \cdot \text{Education}_i + \beta_2^M \cdot \text{AFQT}_i$$

$$E[\text{Income}_i|\text{Education}_i] = \beta_0^U + \beta_1^U \cdot \text{Education}_i$$

- Consider the following auxiliary regression:

$$E[\text{AFQT}_i|\text{Education}_i] = \delta_0 + \delta_1 \cdot \text{Education}_i$$

Omitted Variable "Bias" (2)

$$E[\text{Income}_i|\text{Education}_i, \text{AFQT}_i] = \beta_0^M + \beta_1^M \cdot \text{Education}_i + \beta_2^M \cdot \text{AFQT}_i$$

$$E[\text{Income}_i|\text{Education}_i] = \beta_0^U + \beta_1^U \cdot \text{Education}_i$$

- Consider the following auxiliary regression:

$$E[\text{AFQT}_i|\text{Education}_i] = \delta_0 + \delta_1 \cdot \text{Education}_i$$

- It then follows that:

$$\begin{aligned} E[\text{Income}_i|\text{Education}_i] &= \beta_0^M + \beta_1^M \cdot \text{Educaton}_i + \beta_2^M \cdot E[\text{AFQT}_i|\textit{Education}_i]) \\ &= \beta_0^M + \beta_1^M \cdot \text{Education}_i + \beta_2^M \cdot (\delta_0 + \delta_1 \cdot \text{Education}_i) \\ &= (\beta_0^M + \beta_2^M \cdot \delta_0) + (\beta_1^M + \beta_2^M \cdot \delta_1) \cdot \text{Education}_i \end{aligned}$$

- So $\beta_1^U = \beta_1^M + \beta_2^M \cdot \delta_1 \neq \beta_1^M$ (typically)

Omitted Variable "Bias" results

- $\beta_1^U = \beta_1^M + \beta_2^M \cdot \delta_1$
- "Bias" from omitting AFQT $(\beta_1^U - \beta_1^M)$ is $\delta_1 \cdot \beta_2^M$.
- Breakdown:

	AFQT-Education relation is positive ($\delta_1 > 0$)	AFQT-Education relation is negative ($\delta_1 < 0$)
AFQT increases income ($\beta_2^M > 0$)		
AFQT decreases income ($\beta_2^M < 0$)		

Omitted Variable "Bias" results

- $\beta_1^U = \beta_1^M + \beta_2^M \cdot \delta_1$
- "Bias" from omitting AFQT $(\beta_1^U - \beta_1^M)$ is $\delta_1 \cdot \beta_2^M$.
- Breakdown:

	AFQT-Education relation is positive ($\delta_1 > 0$)	AFQT-Education relation is negative ($\delta_1 < 0$)
AFQT increases income ($\beta_2^M > 0$)	$\beta_1^U > \beta_1^M$	$\beta_1^U < \beta_1^M$
AFQT decreases income ($\beta_2^M < 0$)	$\beta_1^U < \beta_1^M$	$\beta_1^U > \beta_1^M$

Omitted Variable "Bias" example

Outcome:	Earnings		AFQT
Constant	-75.68	-53.17	-41.83
Education	10.06	6.73	6.17
AFQT		0.54	

$$10.06 = 6.73 + 0.54 \cdot 6.17$$

$$\beta_1^U = \beta_1^M + \beta_2^M \cdot \delta_1$$

Omitted Variable "Bias" discussion

- OV "B" result is extremely useful in practice!
  - Even though you don't observe β_2^M or δ_1, you can sometimes use the OV "B" logic to get a sense of whether your "bias" is positive or negative
- "Bias" is a loaded word
  - Sometimes, the correct regression is the one that omits a particular variable!

"OVB for SEs"

- Adding/removing a variable from a regression doesn't just change the point estimates of the other coefficients
  - It changes the standard errors, too!
- In case of homoskedasticity, there are two countervailing effects:

"OVB for SEs"

- Adding/removing a variable from a regression doesn't just change the point estimates of the other coefficients
  - It changes the standard errors, too!
- In case of homoskedasticity, there are two countervailing effects:
  - An additional variable reduces residuals and therefore shrinks SEs
  - An additional variable reduces the amount of residual variation in X and increases SEs
- Overall impact on SEs is ambiguous: depends on the strength of those two effects
- With heteroskedasticy, this stark tradeoff is not technically guaranteed, but in practice it still typically holds

"OVB for SEs" results

	New X is very correlated with old X	New X is not very correlated with old X
New X is a weak predictor of Y	New X will increase SE on old X	Ambiguous
New X is a strong predictor of Y	Ambiguous	New X will decrease SE on old X

"OVB for SEs" example

Outcome: Δ GDP			
Δ G	-0.87	-0.96	-0.83
(SE)	(0.89)	(1.04)	(0.82)
Lagged Δ G		0.29	
		(0.76)	
Lagged Δ GDP			-0.14
			(0.30)
Variance of $\hat{u}$	20,322	20,287	19,927
Residual variation of Δ G	442	401	442

Note: constant term not shown

- Tip: Adding the lagged outcome variable is often a good trick – good explanatory power, often not as correlated with other regressors

"OVB for SEs" takeaways

- Takeaway: When considering controls, don't just think about bias – SEs matter, too!
- A control that addresses bias might blow up the SE and make the regression useless (if the Xs are highly correlated)
- A control variable that doesn't address bias might reduce the SE and be super-useful

Multicollinearity

- Consider the following regression model:

$$\text{Income}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot \text{Age}_i + \hat{\beta}_2 \cdot \text{CurrentYear}_i + \hat{\beta}_3 \cdot \text{BirthYear}_i + \hat{u}_i$$

  - Modeling income as a linear trend of age, time, and birth cohort
- Note that $\text{Age}_i = \text{CurrentYear}_i$ - BirthYear_i

Multicollinearity

- Consider the following regression model:

$$\text{Income}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot \text{Age}_i + \hat{\beta}_2 \cdot \text{CurrentYear}_i + \hat{\beta}_3 \cdot \text{BirthYear}_i + \hat{u}_i$$

  - Modeling income as a linear trend of age, time, and birth cohort
- Note that $\text{Age}_i = \text{CurrentYear}_i$ - BirthYear_i
- This model cannot be estimated!
  - In a model with **perfect "multicollinarity"** – one regressor is a linear combination of other regressors – the coefficients are not uniquely identified

Multicollinearity example

```
. regress incwage year age birthyear, robust
note: year omitted because of collinearity

Linear regression                               Number of obs     =    127,981
                                                F(2, 127978)      =    1302.51
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0240
                                                Root MSE          =      49443
```

incwage	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
year	0	(omitted)				
age	2243.377	55.70913	40.27	0.000	2134.188	2352.566
birthyear	1058.089	21.87837	48.36	0.000	1015.208	1100.97
_cons	-2131519	44244.88	-48.18	0.000	-2218239	-2044800

Intuition for multicollinearity

$$Income_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot \text{Age}_i + \hat{\beta}_2 \cdot \text{CurrentYear}_i + \hat{\beta}_3 \cdot \text{BirthYear}_i + \hat{u}_i$$

- Interpretation 1: holding constant
  - You can't look at a change in Age while keeping CurrentYear and BirthYear constant
  - If Age changes, one of the others must as well!
- Interpretation 2: residual regression
  - A regression of Age on CurrentYear and BirthYear will explain Age perfectly; all residuals will be 0
  - Therefore, there is no residual variation to identify the effect of Age on Income $\rightarrow$ SE will be ∞

Intuition for Multicollinearity

```
. regress year age birthyear
```

Source	SS	df	MS	Number of obs	=	557,953
				F(2, 557950)	=	.
Model	19775554.6	2	9887777.3	Prob > F	=	.
Residual	0	557,950	0	R-squared	=	1.0000
				Adj R-squared	=	1.0000
Total	19775554.6	557,952	35.4431109	Root MSE	=	0

year	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
age	1	.	.	.	.	.
birthyear	1	.	.	.	.	.
_cons	-6.90e-09	.	.	.	.	.

Intuition for Multicollinearity

```
. summarize year_residual, detail

                        Residuals
-------------------------------------------------------------
      Percentiles      Smallest
 1%    -1.34e-11      -1.34e-11
 5%    -1.32e-11      -1.34e-11
10%    -1.32e-11      -1.34e-11       Obs                 557,953
25%    -9.78e-12      -1.34e-11       Sum of wgt.         557,953

50%    -6.59e-12                      Mean               2.07e-14
                        Largest       Std. dev.          2.03e-11
75%    -6.14e-12       6.55e-11
90%     3.14e-11       6.55e-11       Variance           4.12e-22
95%     4.82e-11       6.55e-11       Skewness           2.041255
99%     6.53e-11       6.55e-11       Kurtosis           5.950411
```

Intuition for multicollinearity (2)

$$\text{Income}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot \text{Age}_i + \hat{\beta}_2 \cdot \text{CurrentYear}_i + \hat{\beta}_3 \cdot \text{BirthYear}_i + \hat{u}_i$$

$$\text{Age}_i = \text{CurrentYear}_i - \text{BirthYear}_i$$

- Regression equation can be rewritten as:

$$\text{Income}_i = \hat{\beta}_0 + \underbrace{(\hat{\beta}_2 + \hat{\beta}_1)}_{3} \cdot \text{CurrentYear}_i + \underbrace{(\hat{\beta}_3 - \hat{\beta}_1)}_{2} \cdot \text{BirthYear}_i + \hat{u}_i$$

Intuition for multicollinearity (2)

$$\text{Income}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot \text{Age}_i + \hat{\beta}_2 \cdot \text{CurrentYear}_i + \hat{\beta}_3 \cdot \text{BirthYear}_i + \hat{u}_i$$

$$\text{Age}_i = \text{CurrentYear}_i - \text{BirthYear}_i$$

- Regression equation can be rewritten as:

$$\text{Income}_i = \hat{\beta}_0 + \underbrace{(\hat{\beta}_2 + \hat{\beta}_1)}_{3} \cdot \text{CurrentYear}_i + \underbrace{(\hat{\beta}_3 - \hat{\beta}_1)}_{2} \cdot \text{BirthYear}_i + \hat{u}_i$$

- Each of these 2 sets of parameters are consistent with the results above:
  - $\hat{\beta}_1 = 1$, $\hat{\beta}_2 = 2$, $\hat{\beta}_3 = 3$
  - $\hat{\beta}_1 = 2$, $\hat{\beta}_2 = 1$, $\hat{\beta}_3 = 4$
- I.e. we don't know how to pick the single solution
- The model is asking something nonsensical, and so OLS fails

Multicollinearity in practice

- Your software will tell you if you have perfect multicollinearity
  - You can "fix" it by removing one of the variables
    - This often comes up with "indicator variables" which we'll discuss soon
  - Before you proceed, you should pause to think through whether your model makes sense...

Multicollinearity in practice

- Your software will tell you if you have perfect multicollinearity
  - You can "fix" it by removing one of the variables
    - This often comes up with "indicator variables" which we'll discuss soon
  - Before you proceed, you should pause to think through whether your model makes sense...
- Subtler issues arise with high-but-imperfect multicollinearity

Multicollinearity in practice

- Your software will tell you if you have perfect multicollinearity
  - You can "fix" it by removing one of the variables
    - This often comes up with "indicator variables" which we'll discuss soon
  - Before you proceed, you should pause to think through whether your model makes sense...
- Subtler issues arise with high-but-imperfect multicollinearity
- Suppose your regressors are Age, YearsEducation, YearsWorking

Multicollinearity in practice

- Your software will tell you if you have perfect multicollinearity
  - You can "fix" it by removing one of the variables
    - This often comes up with "indicator variables" which we'll discuss soon
  - Before you proceed, you should pause to think through whether your model makes sense...
- Subtler issues arise with high-but-imperfect multicollinearity
- Suppose your regressors are Age, YearsEducation, YearsWorking
  - There is a very tight relationship between these 3 variables, though not perfect because some people take gap years, maternity leave, unemployment, etc.
  - There will be little residual variation and likely high SEs!
  - May want to consider dropping a variable