Class 7-8: Multiple linear regression

Nic Rivers

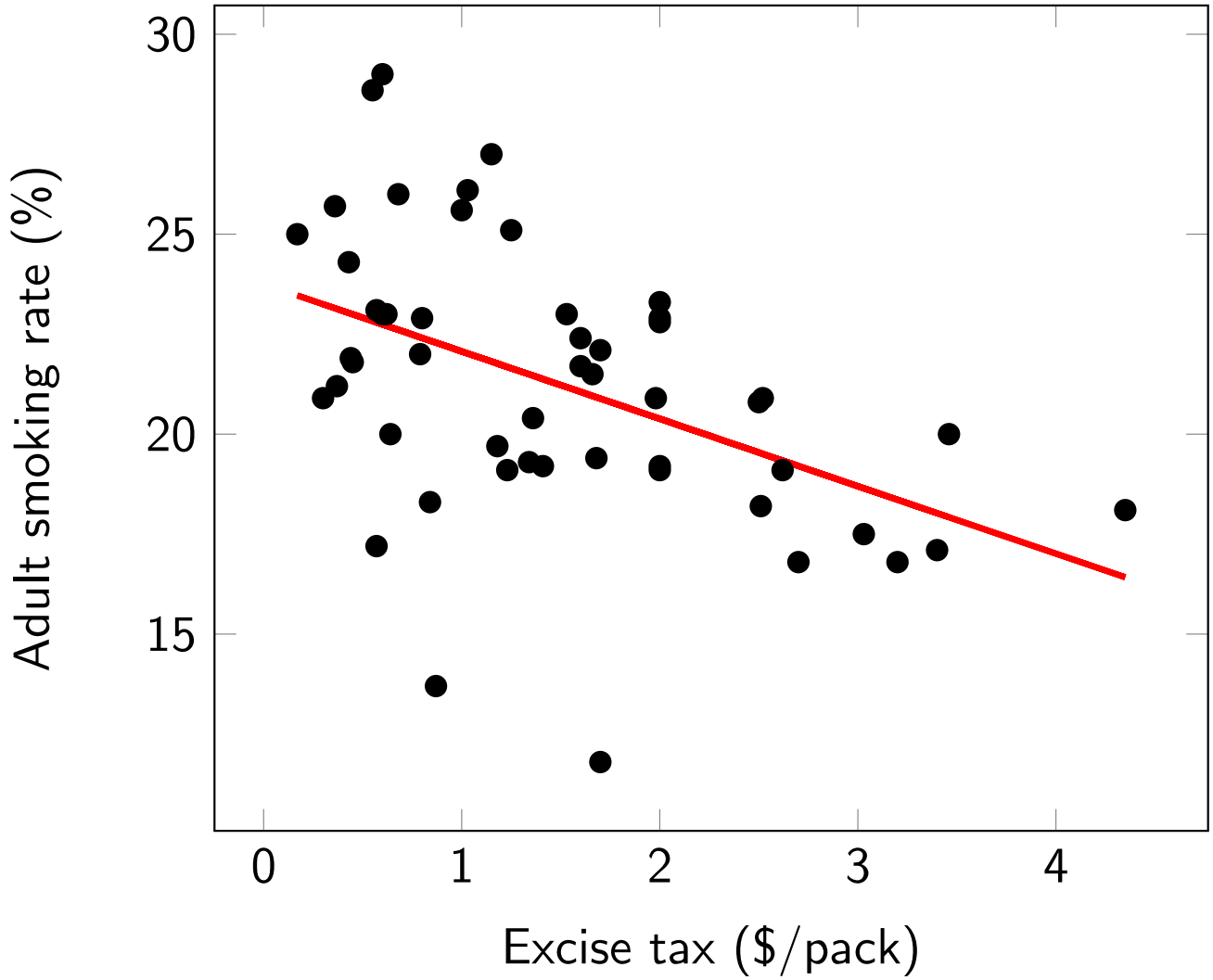
Advanced quantitative research methods, API6319
Fall 2019

Cigarettes and excise taxes: example

- We will examine the relationship between cigarette consumption and cigarette excise taxes
- Important public policy issue (e.g., revenue raising, 'sin' tax to discourage consumption, black-market concerns, etc.)
- We will use cross-sectional data from 50 states (plus DC), and estimate the following regression:

$$smoking\_rate = \beta_0 + \beta_1 cigarette\_tax + u$$

- "Estimating the regression" means that we are using the sample data to estimate the β_0 and β_1 parameters by choosing the parameters that best fit the data.

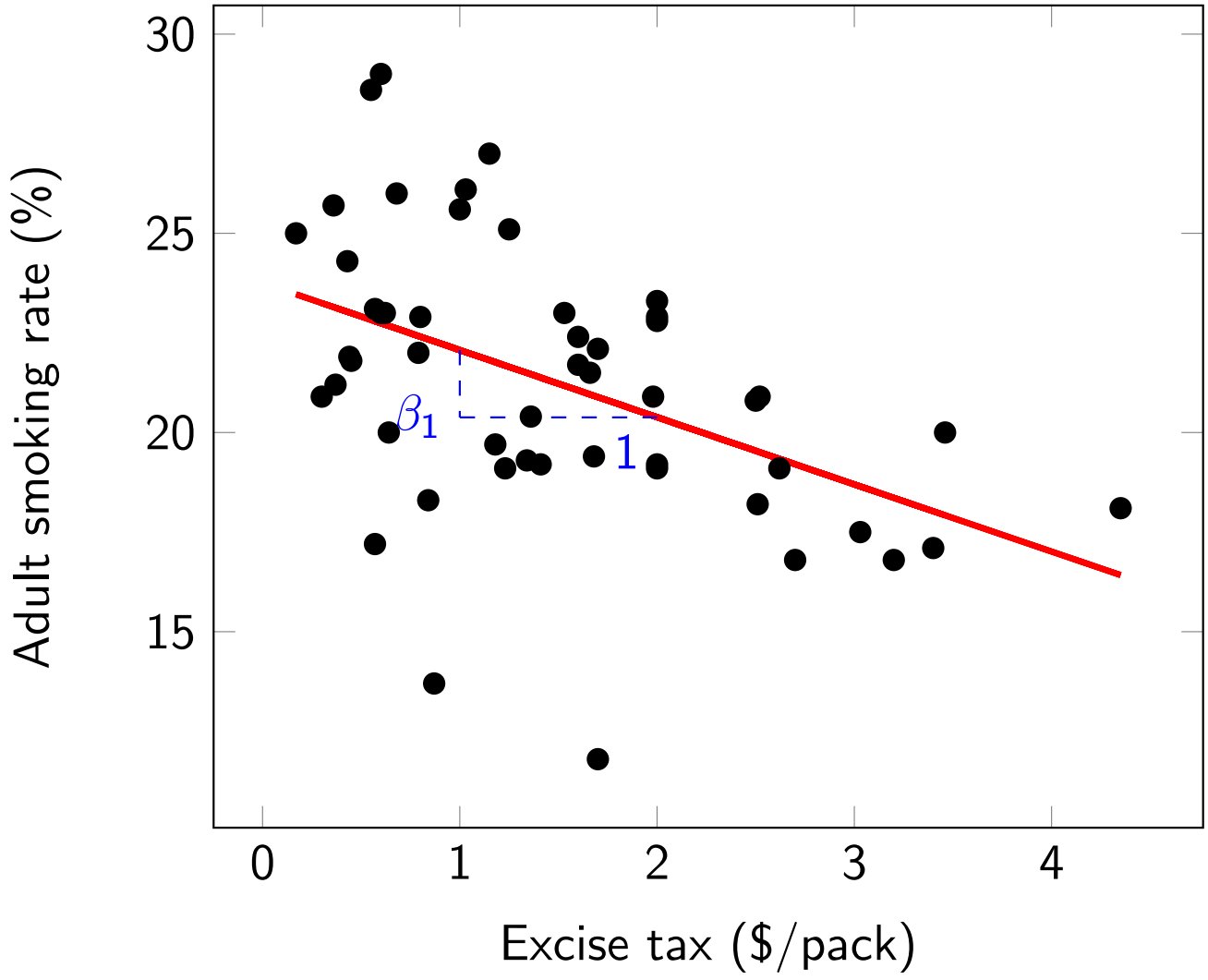

Adult smoking rate (%)
30
25
20
15
0
1
2
3
4
Excise tax ($/pack)

Smoking rate and taxes: regression results

Table: Regression results for cigarette consumption: US States

	Dependent variable:
	cigarette
excise	-1.7^{***}
	(0.5)
Constant	23.8^{***}
	(0.8)
Observations	50
R^2	0.2
Note:	*p<0.1; **p<0.05; ***p<0.01

- What does the *excise* coefficient mean? Values in brackets are standard errors. What do they mean?


Adult smoking rate (%)
30
25
20
15
0
1
2
3
4
Excise tax ($/pack)
β_1
1

Cigarettes and taxes: Interpreting the regression

- Our regression results lead to the following model:

$$\widehat{\text{smoking\_rate}} = 23.8 - 1.7 \times \text{excise\_taxes} + \text{other factors}$$

- An increase in cigarette taxes of \$1 per pack (one unit of x) is associated with a reduction in smoking rate by 1.7 units (percentage points)
- Our regression has uncovered a negative **association** between cigarette excise taxes and cigarette consumption. It predicts the average smoking rate in the sample at different levels of excise taxes.
- Thus we can say:

  "In our sample, when state excise taxes increase by \$1/pack, the average smoking rate decreased by 1.7 percentage points."

- Note that this is not a causal statement – we can make this claim whether or not we believe this is a causal relationship.

Cigarettes and excise taxes: causation?

- For public policy purposes, often we want to claim not just **association** but **causation**.
- Thus, we would like to be able to say:

  > *"Our data suggest that increasing cigarette excise taxes by \$1/pack* **while holding all other factors constant** *would result in a #?? reduction in cigarette consumption."*

- We can only make this claim based on our regression results if other factors are not systematically related to cigarette taxes in our sample (assumption #4)

$$\text{cigarette\_consumption} = \beta_0 + \beta_1 \times \text{excise\_taxes} + \text{other factors}$$

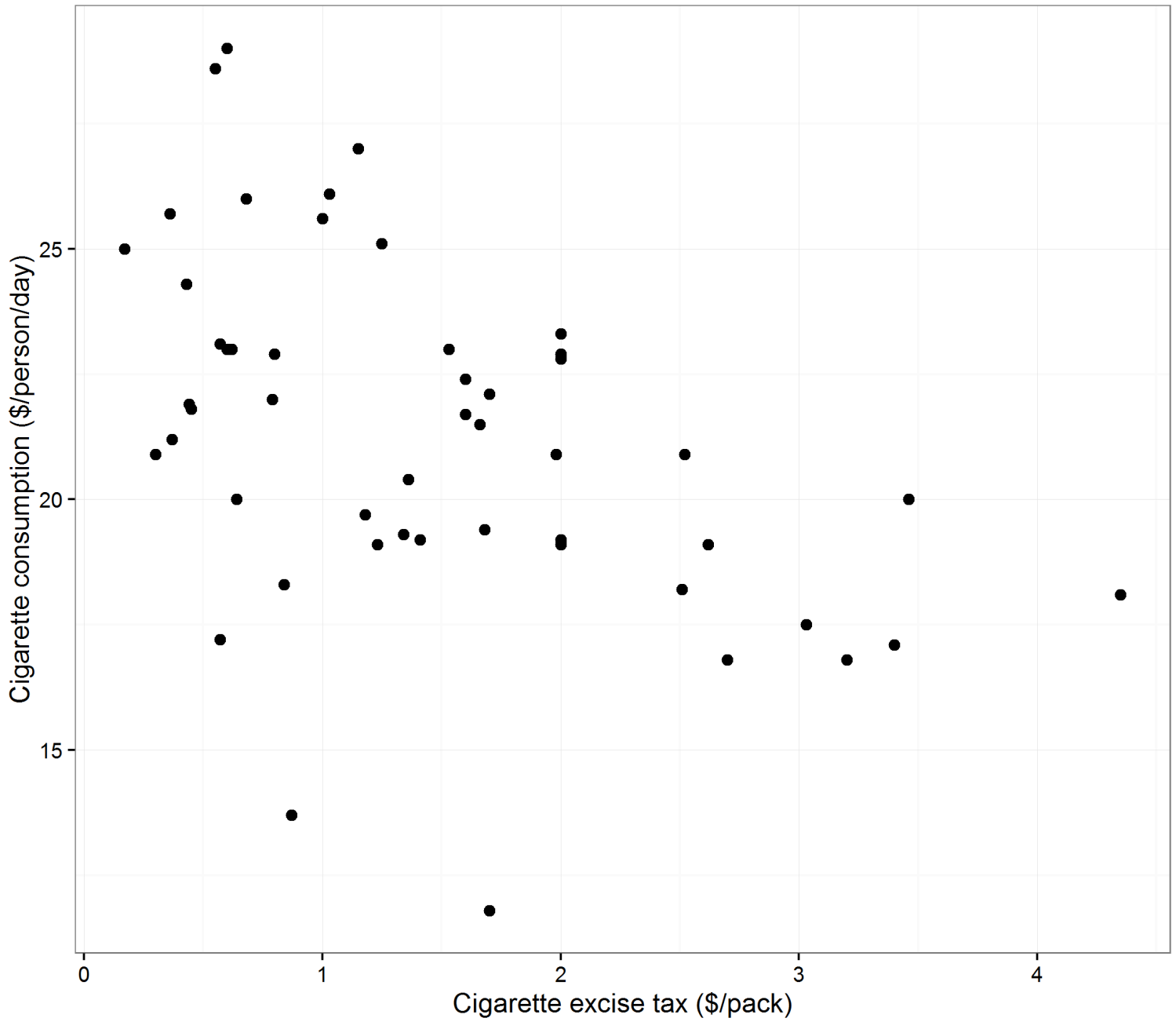
- If other factors are systematically related to cigarette taxes, then our regression result is reporting the effect of changing cigarette taxes and other factors, not just cigarette taxes. Our causal claim will therefore be inaccurate.

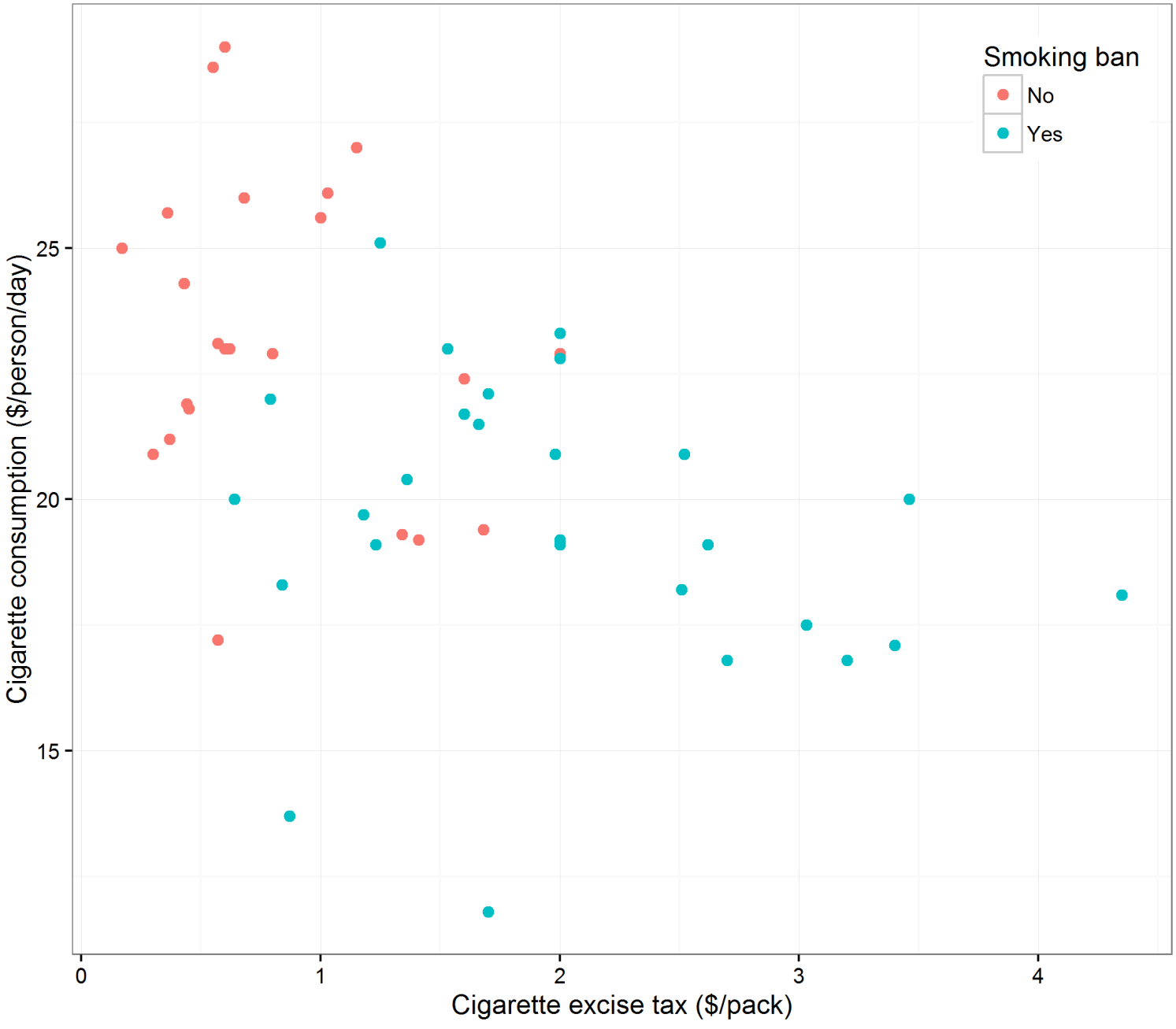
Cigarettes and taxes: other factors

- Excise taxes are not the only thing that affects cigarette consumption. What else affects cigarette consumption?

Cigarettes and taxes: other factors

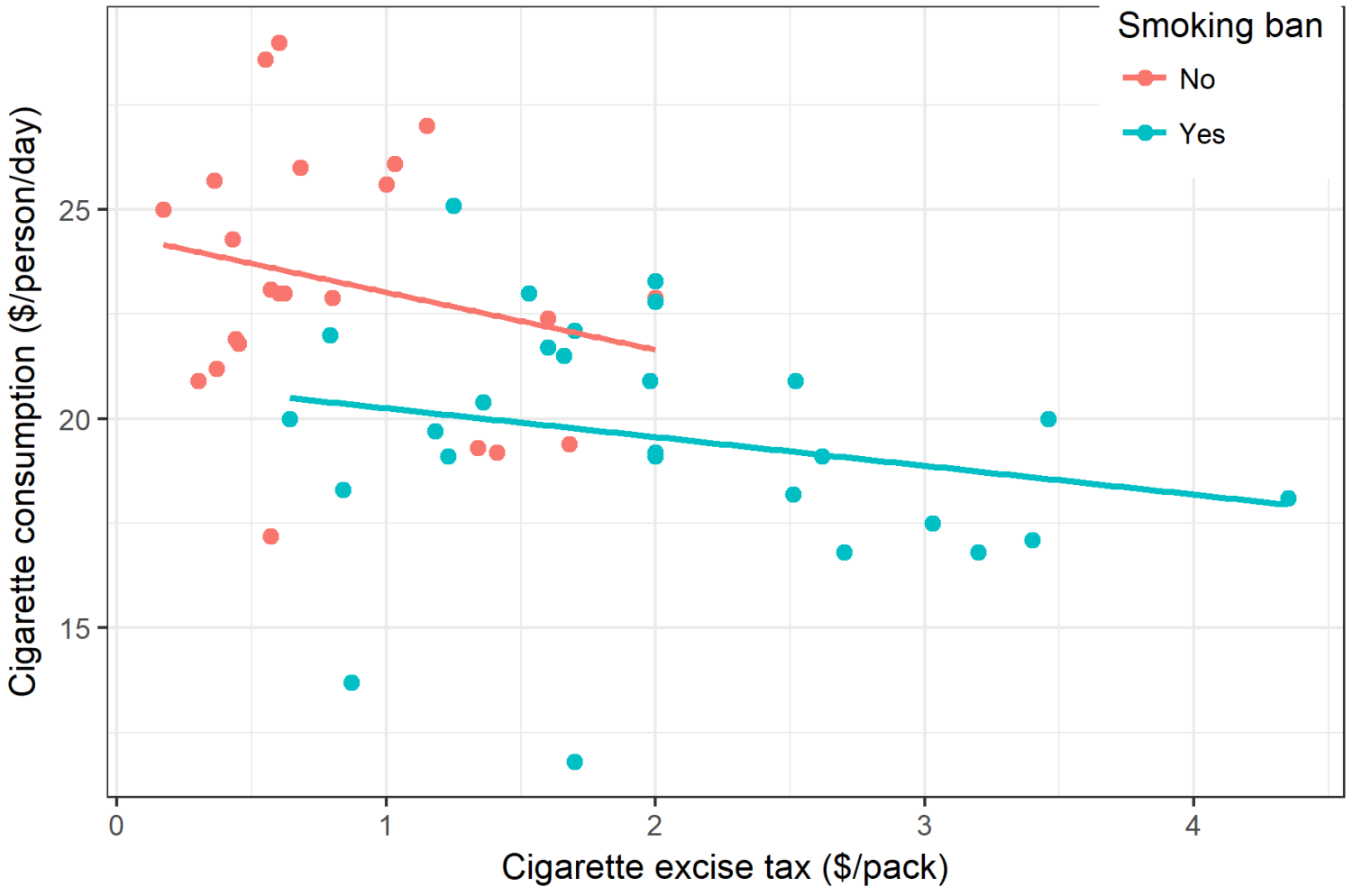
- Excise taxes are not the only thing that affects cigarette consumption. What else affects cigarette consumption?
- It is likely that *restaurant laws* are related to cigarette consumption: places with laws that prohibit smoking in public places like restaurants are likely to have lower cigarette consumption.
- It is also possible that states with strict smoking laws also have high cigarette excise taxes.
- (Many other factors likely also affect cigarette consumption.)
- Our regression of cigarette consumption on cigarette excise taxes is therefore likely to be **confounding** the effect of excise taxes and smoking laws.
- Our regression is attributing the effect of smoking bans (and possibly other factors too) to changes in excise taxes
- We can't make a causal claim about the effect of cigarette taxes on cigarette consumption holding all other factors constant, since these were not held constant in our sample
- (Note that we can still claim a negative association between cigarette taxes and cigarette consumption in our sample.)


Cigarette consumption ($/person/day)
25
20
15
0
1
2
3
4
Cigarette excise tax ($/pack)


Smoking ban
No
Yes
Cigarette consumption ($/person/day)
25
20
15
0
1
2
3
4
Cigarette excise tax ($/pack)

Cigarettes and excise taxes: dealing with confounding by statification

- One solution to confounding variables is sample stratification.
- Separate the sample into smaller sub-samples, which are relatively homogeneous in terms of 'other factors'
- Conduct separate regressions
- Initial (full sample) regression:
  $\widehat{\text{cigarette\_consumption}} = 23.8 - 1.7 \times \text{excise\_tax} + \text{other factors}$
- Subset of states with smoking bans only regression:
  $\widehat{\text{cigarette\_consumption}} = 20.9 - 0.7 \times \text{excise\_tax} + \text{other factors}$
- Subset of states without smoking bans only regression:
  $\widehat{\text{cigarette\_consumption}} = 24.4 - 1.4 \times \text{excise\_tax} + \text{other factors}$


Smoking ban
No
Yes
Cigarette consumption ($/person/day)
25
20
15
0
1
2
3
4
Cigarette excise tax ($/pack)

Cigarettes and excise taxes: dealing with confounding by multiple regression

- In simple regression, we control only for one independent variable
- In multiple regression, we control for more than one independent variable
- This means the other variables we control for are no longer included as 'other factors'
- Controlling for smoking bans in the cigarette regression:

$$\begin{aligned}\text{cigarette\_consumption} = \beta_0 + \beta_1 \times \text{excise\_taxes} \\ + \beta_2 \times \text{smoking\_ban} + \text{other factors}\end{aligned}$$

- In full sample, excise tax coefficient was "statistically significant", but not in sub-samples.

Cigarettes consumption: multiple regression

Table: Regression results for cigarette consumption: US States

	Dependent variable:	
	cigarette	
	(1)	(2)
excise	-1.7^{***}	-0.8
	(0.5)	(0.6)
ban		-2.7^{**}
		(1.1)
Constant	23.8^{***}	24.0^{***}
	(0.8)	(0.8)
Observations	50	50
R^2	0.2	0.3

Cigarette consumption and excise taxes: interpreting multiple regression

- In simple regression, we could say:

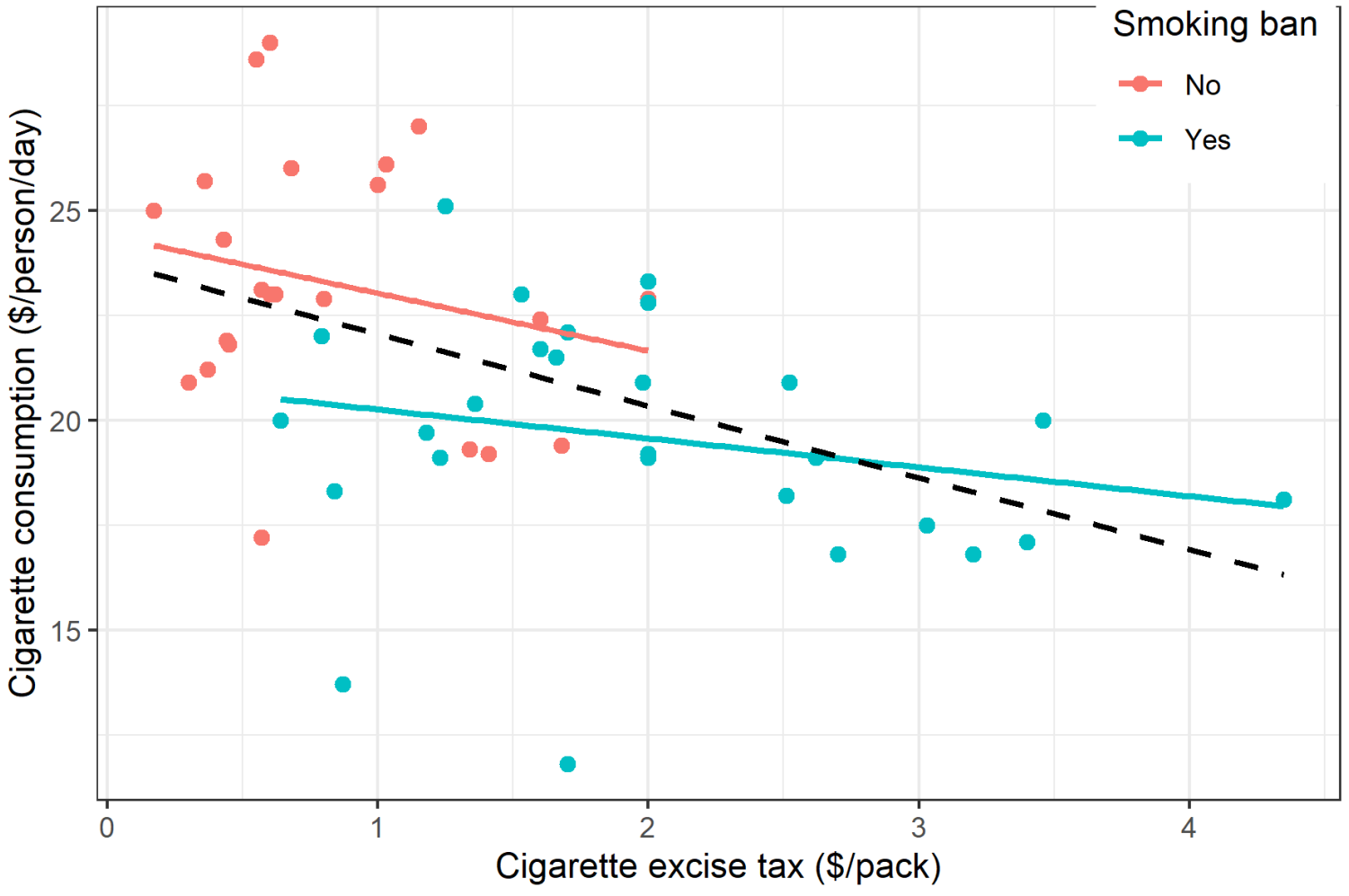
  "In our sample, increases in excise tax of $1/pack are associated with a statistically significant reduction in cigarette consumption of 1.7 cigarettes/person/day."

- In multiple regression, we could say:

  "In our sample, when smoking laws are held constant, increases in cigarette consumption of $1/pack are associated a reduction in cigarette consumption of 0.8 cigarettes/person/day, but the effect is not statistically significantly different from zero."

- This is still not a causal claim, but we can include other factors in the multiple regression (what other factors might be useful to include?). If we can include most other important factors, we can be more comfortable making causal claims.
- Note – we don't need to control for **all** other factors to believe a regression, but we do need to control for **confounding** factors.
- **Confounding factors** are other factors that systematically vary with X that also affect Y

Omitted confounders lead to biased coefficients



Controlling for other factors

- It is possible to control for more than two variables in a multiple regression.

	Dependent variable:			
	cigarette			
	(1)	(2)	(3)	(4)
excise	-1.7^{***}	-0.8	0.1	0.1
	(0.5)	(0.6)	(0.5)	(0.5)
income			-2.1^{***}	-2.1^{***}
			(0.5)	(0.5)
ban		-2.7^{**}	-2.3^{**}	-2.3^{**}
		(1.1)	(0.9)	(0.9)
no_high_school				0.01
				(0.1)
Constant	23.8^{***}	24.0^{***}	33.7^{***}	33.5^{***}
	(0.8)	(0.8)	(2.4)	(3.8)
Observations	50	50	50	50
R^2	0.2	0.3	0.5	0.5
Residual Std. Error	3.1 (df = 48)	2.9 (df = 47)	2.5 (df = 46)	2.5 (df = 45)

Note: *p<0.1; **p<0.05; ***p<0.01
Standard errors in parentheses

Why use multiple regression?

- Simple linear regression attempts to explain changes in a dependent variable as a function of an independent variable:

$$y = \beta_0 + \beta_1 x + u$$

- In order to make causal statements about the effect of x on y holding other factors constant, we are required to assume that 'other factors' (u) have mean zero for all values of x in the population regression line. (See assumption #4 from our first class on regression; same as fourth causal hurdle)
- This means that once we have controlled for x, 'other factors' do not vary systematically.
  - This was violated in the cigarette example; smoking laws was one of the 'other factors' that determined cigarette consumption, and smoking laws were stricter in states with high excise taxes.
  - In other words, `smoking laws` did not have mean of zero for different values of excise taxes.
  - Our simple regression produced misleading results about the causal impact of excise taxes on cigarette consumption.

The zero conditional mean assumption

- In many cases, the zero conditional mean assumption is difficult to justify.
- For example, consider the simple regression:

$$wage = \beta_0 + \beta_1 educ + u$$

  where *educ* captures the number of years of education.
- To interpret β_1 in a causal fashion, we are required to assume that on average, 'other factors' do not vary with education. Is this realistic?
- What else might influence *wage*? Are these other variables correlated with *educ*?

Why use multiple regression (con't)?

- Imagine that wage is also influenced by job experience, *exper*.
- If this is the case, is the simple regression model measuring the causal effect of education on wage?
- In the event that *exper* and *educ* are correlated, the simple regression model would not produce an unbiased estimate of β_1.
- But if we explicitly include *exper* in the regression function, we can determine the impact of *educ* on *wage*, holding *exper* constant. This is useful. The **multiple regression** would then be:

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + u$$

- Now we can make statements about the effect of education on wage, holding experience constant.
- It is possible to include more than two (i.e., many) explanatory variables.

Nomenclature for the multiple regression model

- The general multiple linear regression model can be written as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u$$

  where there are k **independent** variables, x_1 to x_k
- We continue to refer to β_0 as the **intercept**
- We continue to refer to u as the **error** or **disturbance**. It contains factors other than $x_1, x_2, \ldots, x_k$ that influence y.
- The β parameters capture the change in y associated with a one unit change in each of the x's, holding all other factors constant.
- In order to be able to make this type of "all other factors constant" statement, we must assume that 'other factors' have mean zero, given $x_1, x_2, \ldots, x_k$.
- Contrast this with the assumption that 'other factors' have mean zero given x_1 in the simple regression model. How does multiple regression relax the assumptions we need to make?

Multiple regression: an example

- The 1991 Canadian General Social Survey asked questions about health. Here, we aim to determine the factors that contribute to a person's self-reported health status:

  > *Compared to other people your age, how would you describe your state of health? Would you say it was.. Excellent [=1]? Very Good? Good? Fair? Poor [=5]?*[1]

- We estimate the regression:

$$health = \beta_0 + \beta_1 bmi + \beta_2 cigar + \beta_3 sleep + u$$

  where *bmi* measures body mass index, *cigar* measures daily cigarette consumption, and *sleep* measures hours of sleep per night.
- We find the following result:

$$\widehat{health} = 2.027 + 0.0255 bmi - 0.007 cigar - 0.0232 sleep$$

- Interpret these results. What key assumption is required for causality?

[1] Note that this is not a continuous variable, but I will use it as one in the following regression.

Using multiple regression results

- Results of health regression:

$$\widehat{health} = 2.027 + 0.0255bmi - 0.007cigar - 0.0232sleep$$

- We can use these results to make predictions. For example:
  - Holding cigarette consumption and sleep fixed, an increase in the bmi of ten units is associated with a 0.25 point worsening of health status.
  - A person with a bmi of 20 that doesn't smoke and sleeps 8 hours per night reports a health status of 2.35 on average. The same person with 6 hours of sleep reports a health status of 2.4 on average.

What variables to include in a multiple regression?

- Normally, we are interested in knowing the causal effect of X on Y. This is what would happen to Y if we could somehow change X, while somehow leaving everything unaffected.
- To try to estimate this causal effect, we run a regression of Y on X, while controlling for some other variables (let's call them Z).
- How do we decide what variables to include in Z?
  1. We should control for confounding variables.
  2. We should avoid controlling for outcomes or mediating variables.
  3. We should avoid (perfectly) collinear variables.

Omitted variable bias

Omitted variable bias results from failing to control for confounding variables

- What happens when we do not include a variable in the regression when it actually belongs in the regression ?
- Suppose that the population regression function is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

and that we estimate the simple regression model:

$$y = \tilde{\beta}_0 + \tilde{\beta}_1 x_1$$

(the $\tilde{\cdot}$ is used to indicate that these are estimates from the restricted model).

- Will $\tilde{\beta}_1$ be an unbiased estimate of β_1?
- The simple regression model will produce an unbiased estimate of β_1 only if:
  1. x_1 is uncorrelated with x_2, or
  2. x_2 has no effect on y

Omitted variable bias (con't)

- Imagine that $\beta_2 > 0$, and that x_1 and x_2 are positively correlated.
- Then, by leaving out x_2 from our regression, we will overestimate the effect of x_1 on y: our estimate $\tilde{\beta}_1$ will be upward-biased (positively biased). Our regression will attribute the entire effect of changes in x_1 and x_2 to changes in x_1 when calculating $\tilde{\beta}_1$.
- For other cases, we have:

	$Corr(x_1, x_2) > 0$	$Corr(x_1, x_2) < 0$
$\beta_2 > 0$	Upward bias	Downward bias
$\beta_2 < 0$	Downward bias	Upward bias

Omitted variable bias: an example

- Imagine that in the population, wage is determined by:

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + u$$

but that (lacking data, perhaps), we estimate:

$$wage = \alpha_0 + \alpha_1 educ + v$$

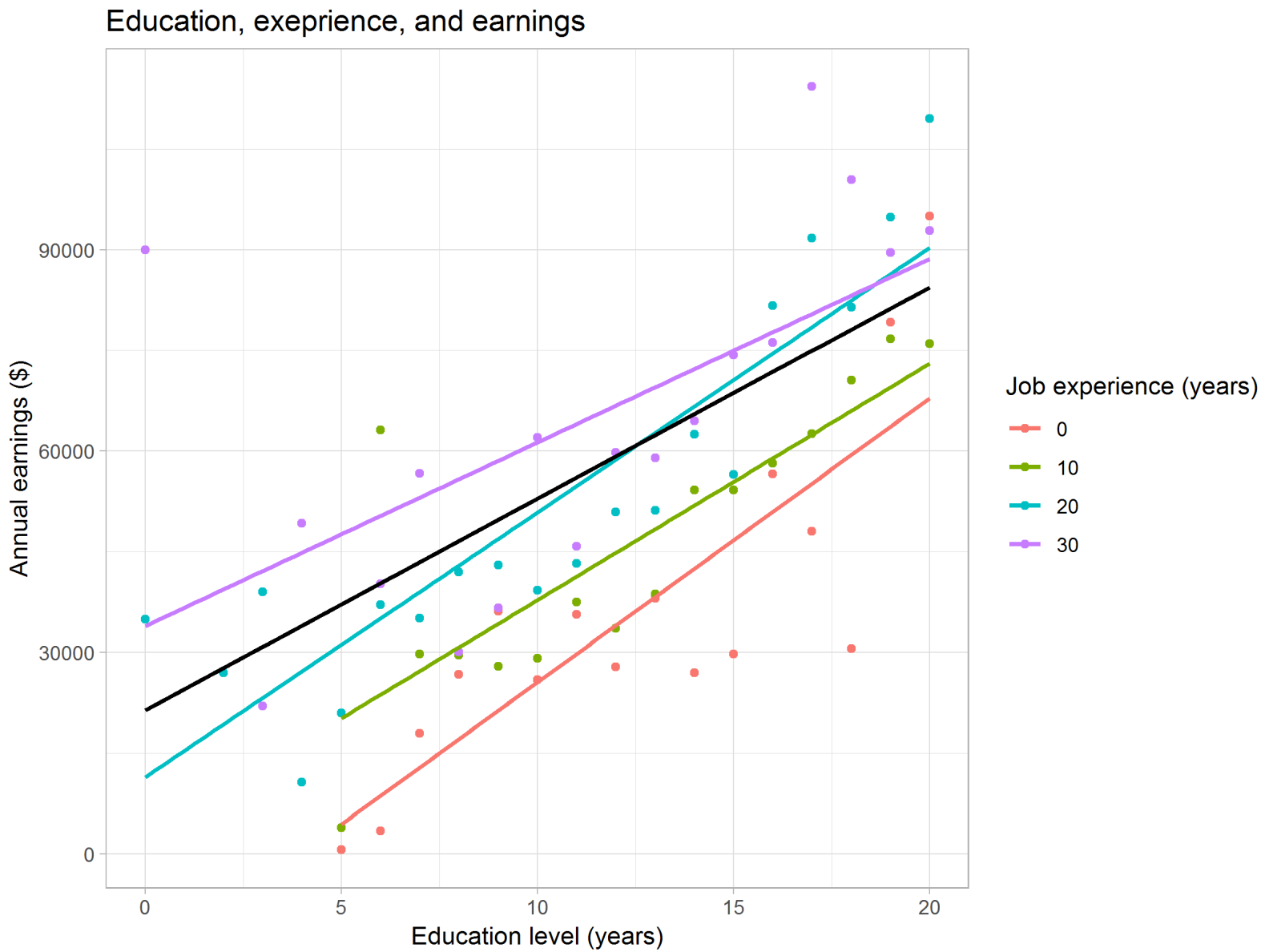
- Do you think α_1 is an unbiased estimator for β_1? What direction do you think the bias will be?
- Actual estimates for 40-50 year old men from the 2009 Survey of Labour and Income Dynamics are:

$$wage = -12509 + 5152 educ$$
$$wage = -42815 + 5470 educ + 1164 exper$$

- Note that these are a sample, so we can't conclude that there is bias in a single sample; bias is something that we would observe from multiple samples. However, here $\hat{Cor}(exper, educ) < 0$ and $\hat{\beta}_2 > 0$.

Multiple regression visualization

$wage = \beta_0 + \beta_1 educ + \beta_2 exper + u$



Omitted variable bias: Yes or No?

- Think about wage on education regression

$$wage = \alpha_0 + \alpha_1 educ + \text{other factors}$$

- In thinking about potential problematic variables, we are concerned with those that are both correlated with x and y. These will cause problems with estimating the causal effect of x by violating the zero conditional mean assumption.

	Correlated with x	**Uncorrelated with** x
Correlated with y	Experience **Problem**	Extrovert/introvert No problem
Uncorrelated with y	Like foreign travel No problem	Quantity of potatoes consumed No problem

- Only variables that are related to both the dependent and independent variable, and also left out of the regression, cause problems.

Observable and unobservable variables

- We would like to include in the regression (control for) variables that are systematically related to both x and y.
- If the variable is observable (measurable and measured), then we can in principle put it in the regression equation, and get rid of the omitted variable bias problem
- But lots of variables are unobservable
- For example, **ability** and **motivation** probably determine both education and income. These are difficult to measure. Thus they are left out of the equation. We need to be careful in interpreting the results of the regression.
- Choosing the appropriate variables to control for relies on theory and intuition (what should be considered in thinking about the dependent variable?)

Controlling for outcomes and mediators

- When we add a variable Z to a regression of Y on X, we get the effect of X on Y conditional on Z.
- Sometimes, this should lead us *not* to include certain Z's in a regression.
- In particular, we should avoid controlling for *mediators* and *outcomes*. These are "bad controls".
- Imagine you could do an experiment and randomly assign X. "Bad controls" are Z variables that change because X changes. Don't include these in a regression.

Mediators
$X \to Z \to Y$
example:
smoking rate $\to$
lung cancer rate $\to$ mortality rate

Outcomes
$X \to Y \to Z$
or $X \to Y$ and $X \to Z$
example:
education $\to$ wage and
education $\to$ white color job

Controlling for (perfectly) collinear variables

- Sometimes there are two variables that are very closely (or even perfectly) related. Including both in a regression will make it difficult to identify the effect of either.
- Perfect correlation:
  - When two variables are perfectly correlated, knowing one implies that you know the other. For example, imagine that I have two dummy variables: immigrant and citizen. If I know the value for immigrant, then I also know the value for citizen. Including both of these variables in a regression will cause R to report an error (NA) because they are perfectly collinear.
- Highly correlated:
  - When two variables are highly correlated (call them X_1 and X_2), knowing one implies that you have a good indication of the other. This makes it difficult to determine which X is responsible for changes in Y. You will get big standard errors.
  - For example, if I did a regression of trust (Y) on the violent crime rate (X_1), but controlled for the property crime rate (X_2) I would likely get large standard errors because of the high correlation between the two X variables.
  - This is less of a problem in really big data sets.

Multicollinearity

- If two independent variables are highly correlated, it is difficult to isolate the effect of each on the dependent variable. We say we have a problem with **multicollinearity**. For example,
  - $grade_i = \beta_0 + \beta_1 classsize_i + \beta_2 schoolexpend_i + \beta_3 income_i + \epsilon_i$
  - where $grade_i$ is average class grade/mark, $schoolexpend_i$ is average expenditures on schooling, and $income_i$ is average income in the region.
  - *schoolexpend* and *income* are likely to be highly correlated. As a result, it will be difficult to precisely determine whether school expenditures or income are driving changes in grades
  - If we are trying to obtain an estimate of β_1, this may not matter. If we are trying to obtain an estimate of β_2 or β_3, we may have a problem
- When two variables are highly correlated, the **variances** of the estimators are large; i.e., the effects are not precisely measured
- Solution to multicollinearity: collect more data!
- When two variables are **perfectly correlated**, the model cannot be estimated

What variables to include?

- Omitting an important confounding variable leads to omitted variable bias: it can cause us to make incorrect causal claims
- Including too many unimportant variables can lead to multicollinearity: it can inflate the standard errors on our parameter estimates, and reduce the precision of our coefficients
- Including "bad controls" can lead to bias.
- What to do?
- There is no one-size-fits-all rule
- Generally, we want a model that includes the main important variables, and excludes unimportant variables and bad controls.
- It is hard to know what these are, so it is normal for people to report results from a number of different specifications with different variables included/excluded, to show whether the main conclusions are **robust** to different model assumptions

Goodness of fit

- As with the simple regression model, we define **R-squared** (R^2) as a measure of goodness of fit.
- R-squared captures the percentage of variance in the dependent variable that is explained by independent variables.
- R-squared increases as more independent variables are added. It is calculated in the same way as for the simple regression model.
- In the previous example, $R^2 = 0.0123$. This means that body mass index, cigarette consumption, and sleep explain only a small part of the difference in self-reported health between individuals. (Presumably, things like disease and hospitalization explain a larger amount).
- A high (or low) R^2 does not suggest that the explanatory variables we include have no usefulness. It just suggests that knowing these three variables is not sufficient to predict an individual's self-reported health.

Population and sample regression functions

- The population regression function (PRF) describes the true average relationship that we are investigating:
  $E(y|x_1, \ldots, x_k) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$.
- The sample regression function (SRF) is an attempt to estimate the PRF.
- Because we estimate the SRF from a finite sample rather than the whole population, the SRF is an estimate of the PRF. We denote the fact that is an estimate by using a $\hat{\cdot}$ over the coefficients:
  $y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k + \hat{u}_i$.
- Given several assumptions, the SRF will be an **unbiased** estimator of the PRF. This means that on average, if we took a large number of samples, the coefficients from the SRF would equal the coefficients in the PRF. In a single sample, they may not be equal.

Assumptions required for SRF to be unbiased estimator of PRF

- The underlying PRF is linear in parameters: $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$.
- The sample upon which we are estimating the relationship is representative of the population (i.e., it is a random sample).
- There is variation in each of the x's, and there are no exact linear relationships among the x's (no perfect **multicollinearity**). (What does this mean?)
- The error u has expected value of zero for any value of explanatory variable: $E(u|x_1, \ldots, x_k) = 0$. (This is the **zero conditional mean** assumption that we encountered in simple linear regression.) The most obvious way for this to fail is if we omit a explanatory variable that is correlated with one of the x's and helps to explain y.
- For causality: the four hurdles.

Confidence and hypothesis testing

- We are estimating a population parameter (slope, intercept) with a sample of data
- Our estimate of the parameter from the sample is a random variable (with a different sample, we would obtain a different estimate)
- We can articulate the uncertainty in our estimate due to randomness in the sample in the same way as when we estimated a mean (and relying on central limit theorem in the same way)
- Exactly the same as for simple linear regression, except now we have more coefficients in the model
- We can construct confidence intervals in the same way, and conduct hypothesis tests in the same way

Hypothesis testing and multiple regression: An example

Regression of earnings in dollars on years of education and years of experience in 40-50 year old men. Data from Survey of Labour and Income Dynamics.

Table: Regression of earnings on education and experience

	Dependent variable:	
	earning	
	(1)	(2)
educ	5,368.18***	5,698.32***
	(279.02)	(277.93)
exper		1,190.43***
		(130.63)
Constant	−12,057.53***	−42,010.70***
	(3,899.43)	(5,062.95)
Observations	3,249	3,249
R^2	0.10	0.12
Note:	*p<0.1; **p<0.05; ***p<0.01	

- How should we interpret?

Qualitative information in regressions

- Our previous examples have focused on measuring the impact of a change in a quantitative variable on another quantitative variable
  - For example, we have measured the impact of changes in number of years of schooling on earnings
- However, in many cases, we would like to also include qualitative variables into analysis
  - For example, we might like to include a variable denoting the gender or race of an individual in a model of earnings, or we might like to include a variable denoting whether a country has been colonized in the past in a model of civil war

Dummy variables

- In cases where the qualitative information is binary (i.e., yes/no; male/female; big/small), then it is straightforward to create a **dummy variable** summarizing the qualitative information.
- A dummy variable is a variable that only takes on values of 0 or 1.
- For example, consider our model of wage determination:

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + u$$

- An additional relevant explanatory variable for wage might be a person's gender
- Gender is a binary qualitative variable: taking on values of *male* or *female*
- We can create a dummy variable to summarize this so that it can be included in the regression equation.
- For example, we could define a new variable *female* that takes on a value of 1 when the gender is female and a value of 0 when the gender is male

Regression with dummy variables

- Now consider the regression:

$$wage = \beta_0 + u$$

- What is β_0? (recall that regression aims to minimize (squared) residuals)

Regression with dummy variables

- Now consider the regression:

$$wage = \beta_0 + u$$

- What is β_0? (recall that regression aims to minimize (squared) residuals)
- This is just the mean wage. We can obtain it via regression, or just by asking for the mean wage.
- And consider the regression:

$$wage = \alpha_0 + \alpha_1 female + u$$

- What is α_0? What is α_1?

Regression with dummy variables

- Now consider the regression:

$$wage = \beta_0 + u$$

- What is β_0? (recall that regression aims to minimize (squared) residuals)
- This is just the mean wage. We can obtain it via regression, or just by asking for the mean wage.
- And consider the regression:

$$wage = \alpha_0 + \alpha_1 female + u$$

- What is α_0? What is α_1?
- α_0 is the mean wage when $female = 0$. $\alpha_0 + \alpha_1$ is the mean wage when $female = 1$.

Regression with dummy variables

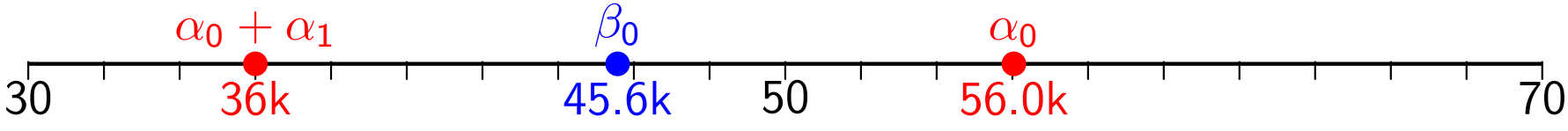
- Now consider the regression:

$$wage = \beta_0 + u$$

- What is β_0? (recall that regression aims to minimize (squared) residuals)
- This is just the mean wage. We can obtain it via regression, or just by asking for the mean wage.
- And consider the regression:

$$wage = \alpha_0 + \alpha_1 female + u$$

- What is α_0? What is α_1?
- α_0 is the mean wage when $female = 0$. $\alpha_0 + \alpha_1$ is the mean wage when $female = 1$.



Parallel to t-test

- Note that the coefficient α_1 captures the difference between mean male and female wages
- We have seen this before, when we used a t-test to compare mean wages across groups
- In the simple regression $wage = \alpha_0 + \alpha_1 female + u$, the t-value on α_1 will be exactly the same as in a two sample t-test

Including dummy variables in multiple regression

- Now we can estimate the regression with the dummy variable in the regression:

$$wage = \beta_0 + \beta_1 educ + \beta_2 female + u$$

- What does β_2 tell us?

Table: Regression of earnings on education and experience

	Dependent variable:		
	earning		
	(1)	(2)	(3)
educ	5,085.99***	5,215.03***	5,266.43***
	(172.18)	(166.86)	(163.16)
exper		1,318.12***	1,005.43***
		(61.39)	(62.49)
female			−16,930.15***
			(942.07)
Constant	−19,474.46***	−46,458.22***	−32,153.25***
	(2,414.88)	(2,655.15)	(2,715.10)
Observations	6,964	6,964	6,964
R^2	0.11	0.17	0.20
Note:			*p<0.1; **p<0.05; ***p<0.01

Including dummy variables in multiple regression

- Now we can estimate the regression with the dummy variable in the regression:

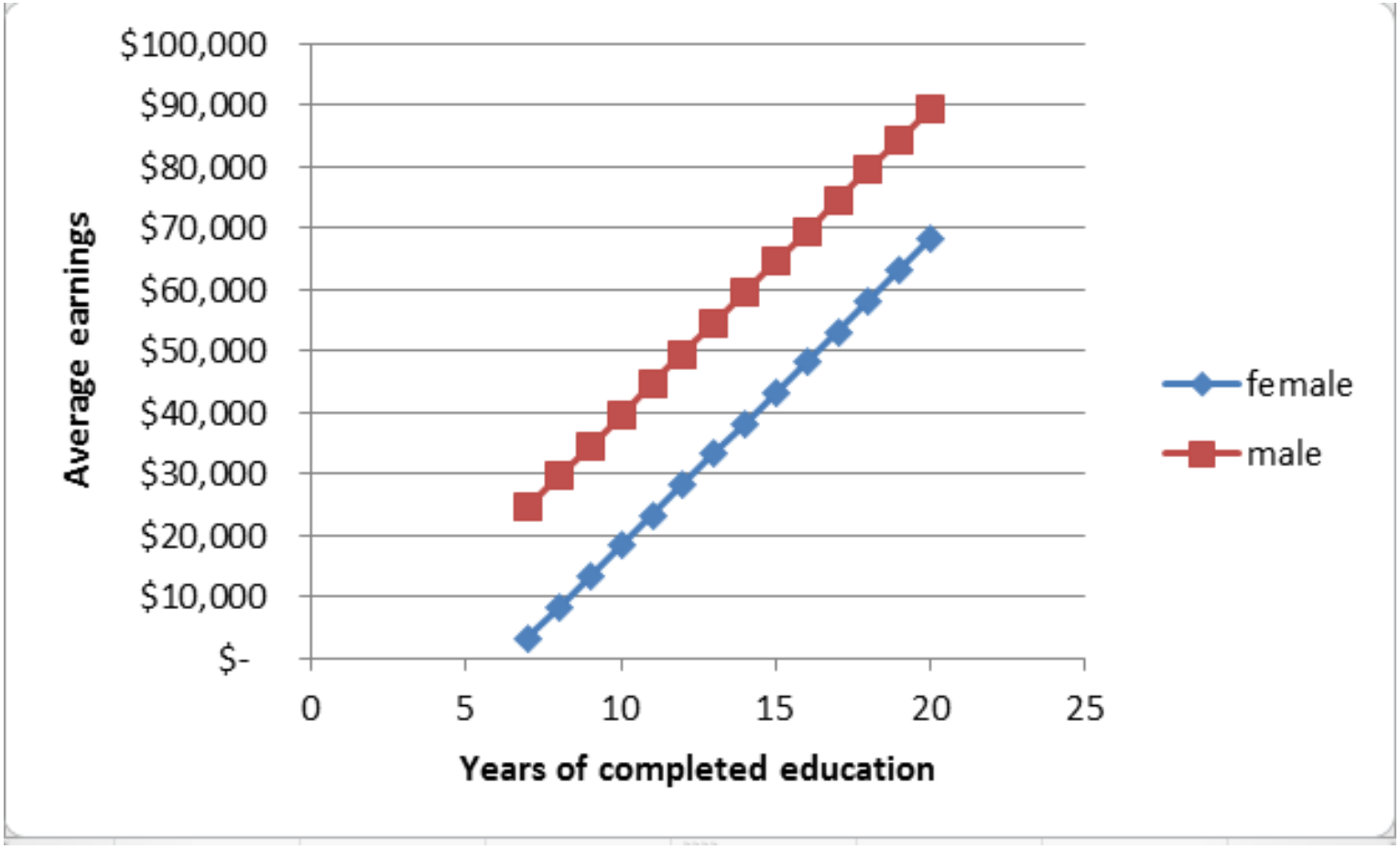
$$wage = \beta_0 + \beta_1 educ + \beta_2 female + u$$

- What does β_2 tell us?
  - A separate intercept for males and females

Table: Regression of earnings on education and experience

	Dependent variable:		
	earning		
	(1)	(2)	(3)
educ	5,085.99***	5,215.03***	5,266.43***
	(172.18)	(166.86)	(163.16)
exper		1,318.12***	1,005.43***
		(61.39)	(62.49)
female			−16,930.15***
			(942.07)
Constant	−19,474.46***	−46,458.22***	−32,153.25***
	(2,414.88)	(2,655.15)	(2,715.10)
Observations	6,964	6,964	6,964
R^2	0.11	0.17	0.20
Note:			*p<0.1; **p<0.05; ***p<0.01

Dummy variables in multiple regression



Coding categorical variables with many categories as dummies

- Our last example was of a binary variable that we coded as a dummy (male/female)
- We will also run into examples where there are more than two categories (small/medium/big classes; different cities in Canada; etc.)
- We can implement these as dummies as well:
  - Create a dummy for each category
  - Set it equal to one if the observation is in that category, and zero otherwise
  - Include one fewer dummy variables than there are categories in the variable.
  - Include these dummies in the regression

Linear regression and non-linearity

- We are conducting linear regression, but many of the relationships that we explore are likely non-linear. What to do?
- Transform variables, and then conduct linear regression
- Regression itself remains linear, but variables may be some transformation of initial variable
- Main types of transformations:
  - Squares
  - Logs

Log transformations

- The base-b logarithm of a number x is the value of y such that b^y equals x. So if $y = \log_b(x)$,

$$b^y = x$$

- The natural logarithm (ln, or log) uses a base of $e = 2.71828\ldots$
- A very convenient feature of the natural logarithm is that for small changes, the difference in the natural logarithm of x approximates a percent change:

$$\ln(x + \Delta x) - \ln(x) \approx \frac{\Delta x}{x}$$

Logarithmic regression

- This means that if we use the logarithm of x as the dependent variable in a regression, we can interpret the results as percent changes. So in the following regression:

$$\ln(y) = \beta_0 + \beta_1 x + u$$

  the estimate $\hat{\beta}_1$ tells us the percent change in y associated with a one-unit change in x
- And in the following regression:

$$\ln(y) = \alpha_0 + \alpha_1 \ln(x) + u$$

  the estimate $\hat{\alpha}_1$ tells us the percent change in y associated with a one-percent change in x
- This is the concept of an elasticity
- It is useful because it is dimensionless, and because many real-world relationships follow elasticity-style relationships

Cigarettes and consumption: multiple regression with logarithms

	Dependent variable:		
	cigarette		log(cigarette)
	(1)	(2)	(3)
excise	-1.71^{***}	0.08	
	(0.46)	(0.52)	
income		-2.13^{***}	
		(0.49)	
log(excise)			0.01
			(0.03)
log(income)			-0.58^{***}
			(0.13)
ban		-2.32^{**}	-0.11^{**}
		(0.91)	(0.05)
Constant	23.77^{***}	33.73^{***}	4.06^{***}
	(0.80)	(2.35)	(0.22)
Observations	50	50	50
R^2	0.22	0.52	0.49
Residual Std. Error	3.10 (df = 48)	2.50 (df = 46)	0.13 (df = 46)

Note: *p<0.1; **p<0.05; ***p<0.01
Standard errors in parentheses

Discrete dependent variable

- Sometimes, we want to run a regression with a discrete dependent variable.
- See the notes posted on the class website for addressing this situation.