Final Exam Details

- The final is Thursday, March 17 from 10:30am to 12:30pm in the regular lecture room
- The final is cumulative (multiple choice will be a roughly 50/50 split between material since the second midterm and old material, short answer will be focused on the new material)
- The old finals are a good guide to the format and length of the exam as well as the division of the exam between old and new material
- Office hours during exam week: Monday 2pm-4pm, Tuesday 10am-12pm, Wednesday 10am-12pm

Review: Omitted Variable Bias

True model:

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

Estimated model:

$$y = \tilde{b_1} + \tilde{b_2} x_2$$

Relationship between x_2 and x_3:

$$x_3 = \gamma_1 + \gamma_2 x_2 + \nu$$

Estimated slope coefficient:

$$E(\widetilde{b_2}) = \beta_2 + \beta_3 \gamma_2$$

Review: Omitted Variable Bias

$$E(\widetilde{b_2}) = \beta_2 + \beta_3\gamma_2$$

- So the estimated slope coefficient for x_2 will be equal to the true value (β_2) plus an additional term ($\beta_3\gamma_2$)
- As long as this additional term is nonzero, we have a biased estimate of the slope coefficient
- The sign of the bias depends on the sign of β_3 and the sign of γ_2

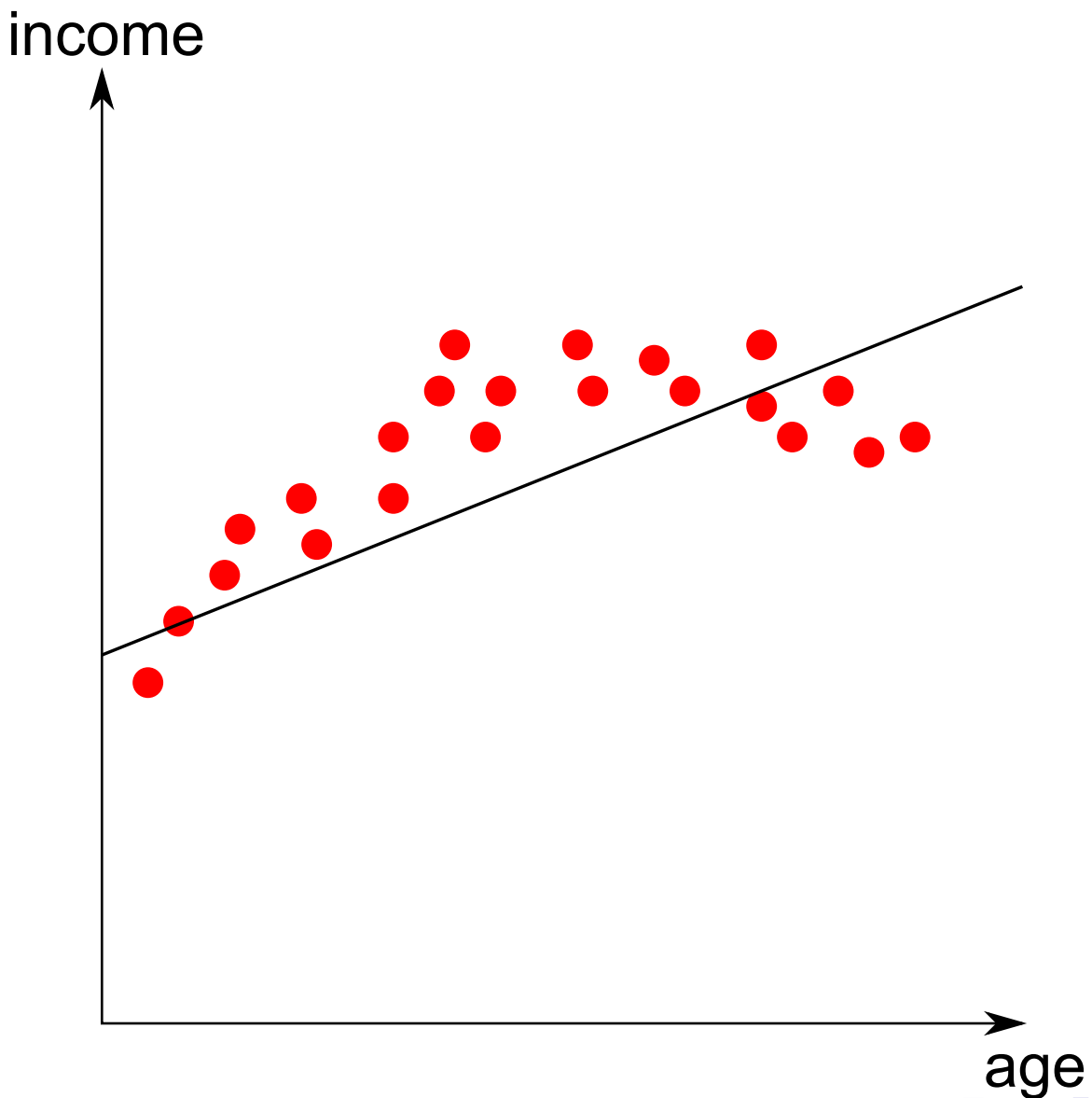
Including Too Many Variables

- We've seen that omitting important variables leads to big problems
- What if we include too many variables?
- It's not nearly as bad
- Our coefficients stay unbiased for the regressors that should be there but we lose some precision
- These problems are small compared to the problems of omitted variables, so it is best to error on the side of including questionable regressors

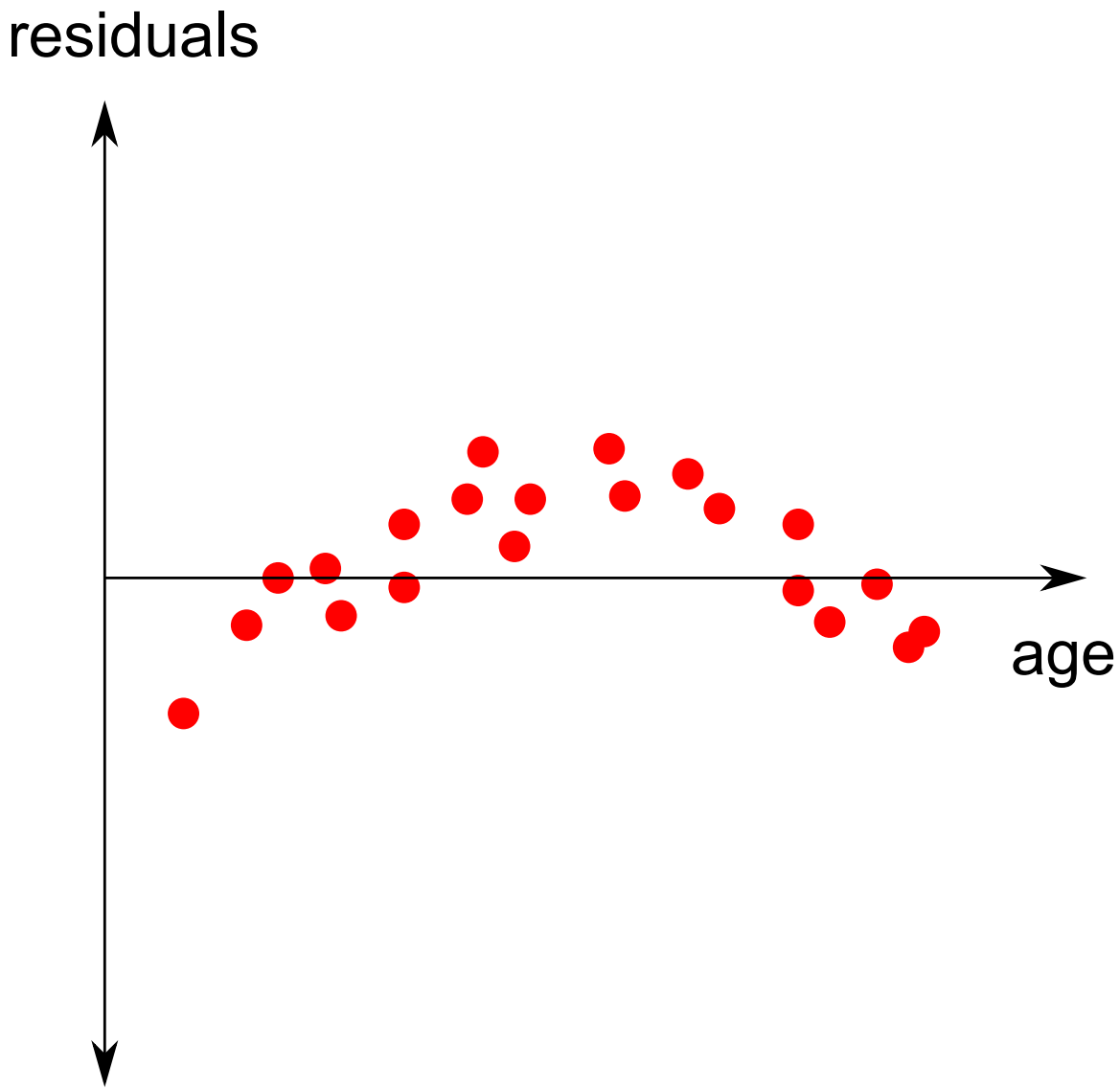
Non-linear Relationships

- We've covered the problems of including the wrong set of variables in our model
- The other way we can misspecify the model is by using the wrong functional form
- This is a problem we've already encountered and we solve it with data transformations
- One way we'll notice we have a problem is if we get distinct patterns in the residuals plotted against a regressor

Non-linear Relationships



Non-linear Relationships



Badly Behaved Errors

- We've just seen that one way we know that the model is misspecified is if a pattern shows up on a graph of the residuals and the regressor
- This leads us into a new set of problems: badly behaved error terms
- Several problems can pop up with the error terms:
  - Errors are correlated with the regressors
  - Errors have nonconstant variance
  - Errors are correlated with each other

Errors Correlated with the Regressors

- Recall the problem of omitted variable bias, we used the following (misspecified) regression model:

$$y = \tilde{b}_1 + \tilde{b}_2 x_2$$

- Because we omitted x_3 from the equation and x_3 was correlated with both y and x_2, our estimated coefficient $\tilde{b}_2$ was a biased estimator of β_2
- The error term included x_3 which was correlated with x_2
- This means that omitted variable bias is actually one case of when the error term is correlated with the regressors

Errors Correlated with the Regressors

- So omitted variable bias is one situation in which the errors are correlated with the regressor
- There are a couple of other important situations when we get errors correlated with the regressors and biased coefficients
- Two of the most common situations:
  - Measurement error in one of the regressors
  - Sample selection bias

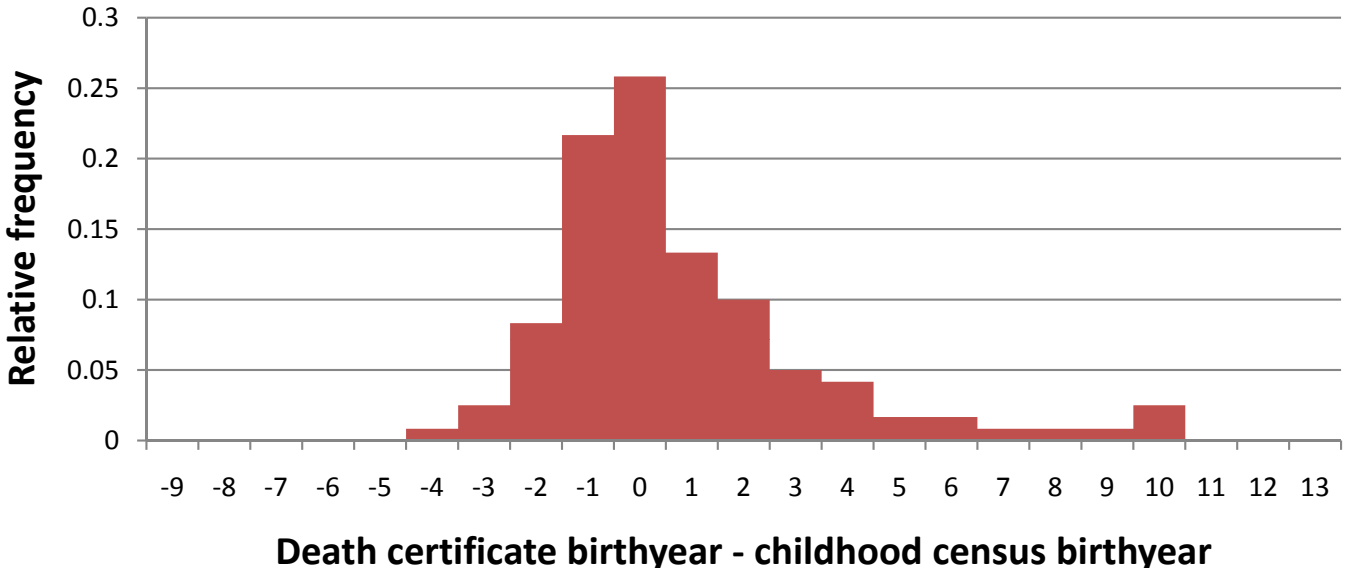
Measurement Error

SIPP misreporting by earnings category

	Percent within:			Difference ($SIPP - SER$):		
Earnings Category	5%	10%	25%	Mean	Median	Num. of Obs.
$<$ \$5,000	9.1	17.0	35.9	270.86	2,924.47	3,637
\$5–\$10,000	17.1	31.3	58.8	30.45	1,465.61	2,644
\$10–\$15,000	23.4	42.6	72.4	−283.43	815.17	2,392
\$15–\$20,000	26.6	48.3	78.2	−663.76	10.69	2,225
\$20–\$25,000	29.0	51.9	81.2	−1,064.16	−697.25	1,986
\$25–\$30,000	29.2	49.2	82.2	−1,715.47	−1,779.05	1,491
\$30–\$35,000	27.2	47.2	80.5	−2,313.97	−2,768.10	1,130
\$35–\$40,000	30.8	50.5	80.8	−2,741.55	−3,763.70	956
\$40–\$45,000	28.0	47.0	78.7	−3,560.01	−4,761.25	771
\$45–\$50,000	29.6	45.4	78.3	−4,078.39	−5,918.33	527
$>$ \$50,000	61.7	70.6	85.7	0	−5,274.04	1,758

Pedace, Roberto. "Using administrative records to assess earnings reporting error in the survey of income and program participation." Journal of Economic and Social Measurement, 26(3/4).

Measurement Error



Measurement Error

- Sometimes we can't measure things as precisely as we'd like
- So while we might be interested in:

$$y = \beta_1 + \beta_2 x + \varepsilon$$

we're actually regressing y on $\tilde{x} = x + \nu$ where ν is some random error

- Why is this a problem? It will lead to error terms negatively correlated with the regressor $\tilde{x}$:

$$y = \beta_1 + \beta_2(\tilde{x} - \nu) + \varepsilon$$

$$y = \beta_1 + \beta_2\tilde{x} + (-\beta_2\nu + \varepsilon)$$

- Regressing y on $\tilde{x}$ will give a biased estimate of β_2

Measurement Error

- With this kind of random measurement error, our estimate b_2 will actually be *biased toward zero*
- This can be shown with some math, but there is also simple intuition behind it
- The more random measurement error we have in x, the weaker the relationship between x and y will appear to be
- In the extreme case, if the measurement error is very large compared to the actual variation in x, differences in the observed values of x won't appear to have any relationship with y

Measurement Error

- What can we do about measurement error?
- Find a better way of measuring the variable we're interested in
- Take multiple measurements and average them
- If we can't improve the measurement, we need to consider the bias when we interpret the results

Measurement Error

- What if we have measurement error in y so we actually observe $\tilde{y} = y + \nu$?
- This is less of a problem:

$$y = \beta_1 + \beta_2 x + \varepsilon$$

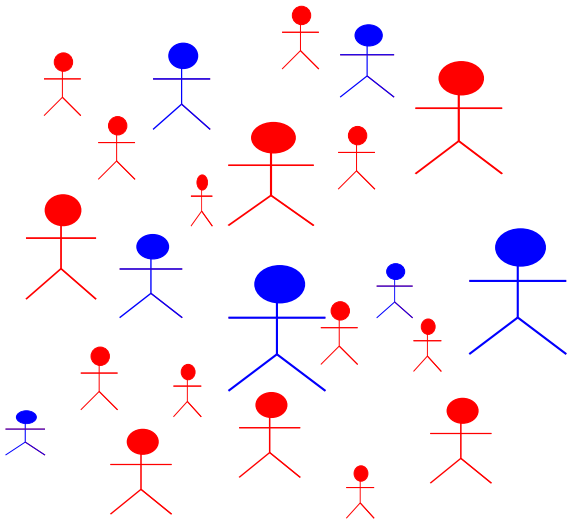
$$\tilde{y} - \nu = \beta_1 + \beta_2 x + \varepsilon$$

$$\tilde{y} = \beta_1 + \beta_2 x + (\varepsilon + \nu)$$

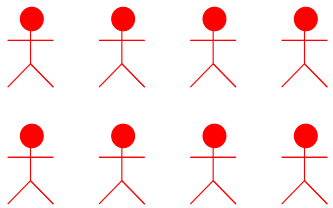
- If we regress $\tilde{y}$ on x, we can still get an unbiased estimate of β_2 since x is still uncorrelated with the error term
- We do lose precision (the variance of the error term is bigger)

Sample Selection Bias

Population



Sample



Sample Selection Bias

- If our chosen sample is not representative of the population, we can get a sample selection bias
- Often our observations look more similar to each other than a random selection from the population should
- This means that the observations in the sample may share unobserved characteristics that are correlated with our outcome variable and the regressors
- The coefficients we estimate will be capturing the relationship between y and x for the chosen sample and may not be unbiased estimates of the true population values

Sources of Sample Selection Bias

- We can get sample selection bias for a variety of reasons
- Our way of choosing the sample can favor one group over another (eg. phone survey vs internet survey)
- We often only get observations for a single geographical area or a single period in time
- Non-response on surveys or attrition is often not random

Sample Selection Bias: Non-random Samples

		Landline & cell, interviewed on:		
	Landline only	Landline	Cell	Cell only
18-29	11	12	17	47
30-49	21	37	41	36
50-64	27	31	29	12
65+	38	18	12	4
College grad	23	42	40	25
Some college	22	25	24	28
HS grad	40	28	29	35
Some HS	14	5	6	12
Renter	28	15	20	60

From Pew Research Center, "Costs and benefits of full dual-frame telephone survey designs", 2008

Sample Selection Bias: Dewey vs Truman, Bush vs Gore

Sample Selection Bias: Attrition

Table 1 Completion Rates by Demographic Variables

	Total		Complete		Incomplete	
Variable	N^a	%	n^b	% Total	n^c	% Total
Age						
13–16	8595	55	8079	94	516	6
17–21	7089	45	6443	91	646	9
Race						
Anglo	11,297	72	10,558	94	739	7
Black	440	3	395	90	45	10
Latino	2071	13	1832	88	239	12
Asian	1421	9	1326	93	95	7
Other	411	3	374	91	37	9
Sex						
Male	7206	46	6671	93	535	7
Female	8478	54	7851	93	627	7

Morrison et. al. "Tracking and follow-up of 16,915 adolescents: minimizing attrition bais." Controlled Clinical Trials, 18 (1997)

Sample Selection Bias: Attrition

Table 2 Success Rates for Individual Follow-Up Procedures

Procedure[a]	Used	% Total	Successful[b]	% Successful
Home phone	15,175	97	12,129	80
Alternate phone[c]	1392	9	436	31
Directory Assistance	2531	16	393	16
Orthodontic records	2243	14	984	44
Mail survey	1033	7	322	31
Mail (certified)[d]	935	6	288	31
Reverse directory	336	2	15	4
$10 incentive	749	5	96	13
$20 incentive	241	2	48	20

Morrison et. al. "Tracking and follow-up of 16,915 adolescents: minimizing attrition bais." Controlled Clinical Trials, 18 (1997)

Dealing with Sample Selection Bias

- First is simply figuring out whether you have a problem: think about why your sample may not be representative of the population, check some of the observable characteristics of your sample to see if the sample looks random
- Try to determine the direction of the bias resulting from any sample selection issues (use some economic intuition)
- You can try to gather more data using a better sampling strategy