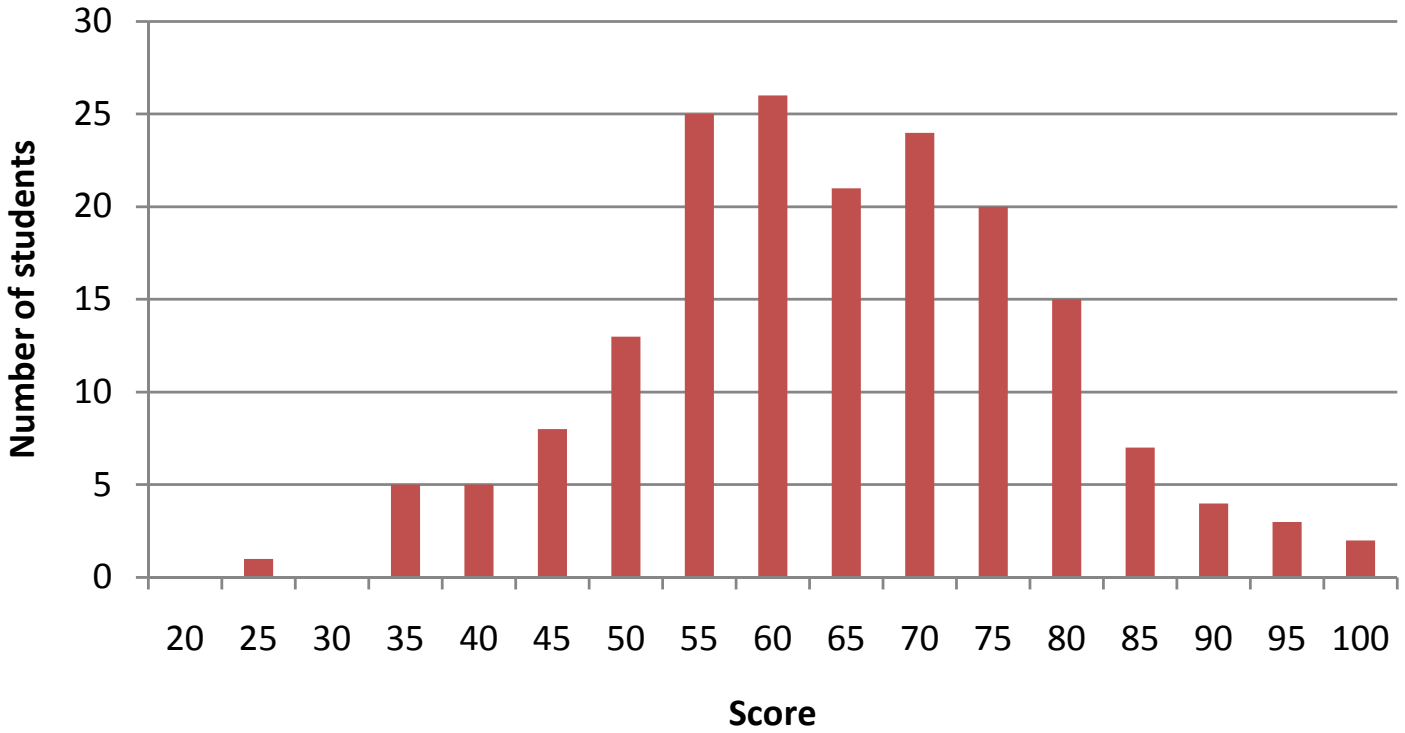
# Midterm 2 Grade Distribution



# Interaction Terms

Recall our basic setup using an interaction term from last class:

$$y_i = \beta_1 + \beta_2 x_i + \beta_3 D_i + \beta_4 x_i \cdot D_i + \varepsilon_i$$

$$E(y_i|D_i = 1) = (\beta_1 + \beta_3) + (\beta_2 + \beta_4)x_i$$

$$E(y_i|D_i = 0) = \beta_1 + \beta_2 x_i$$

$$E(y_i|D_i = 1) - E(y_i|D_i = 0) = \beta_3 + \beta_4 x_i$$

To Excel for an example with the basketball salary data for one big example with logs, polynomials, multiple dummies and an interaction term...

# Another Case of Interaction Terms

- Interaction terms are not limited to a dummy variable interacted with a continuous variable
- We can also have a continuous variable interacted with another continuous variable
- The idea and the steps are the same as last class, the interpretation is a just little more complicated

# Another Case of Interaction Terms

- Let's think about studying obesity, measured by the body mass index (bmi)
- If we think that obesity is a function of hours of exercise a week and calories consumed per day, we might try to predict bmi using the following equation:

$$\widehat{bmi}_i = b_1 + b_2 cal_i + b_3 hours_i$$

- More calories should increase bmi, more exercise should decrease bmi
- But calories will have a different effect for people who exercise a lot versus people who exercise very little

# Another Case of Interaction Terms

- If we think the effect of calories on bmi differs with the amount of exercise, we want to include an interaction term:

$$\widehat{bmi}_i = b_1 + b_2 cal_i + b_3 hours_i + b_4 cal_i \cdot hours_i$$

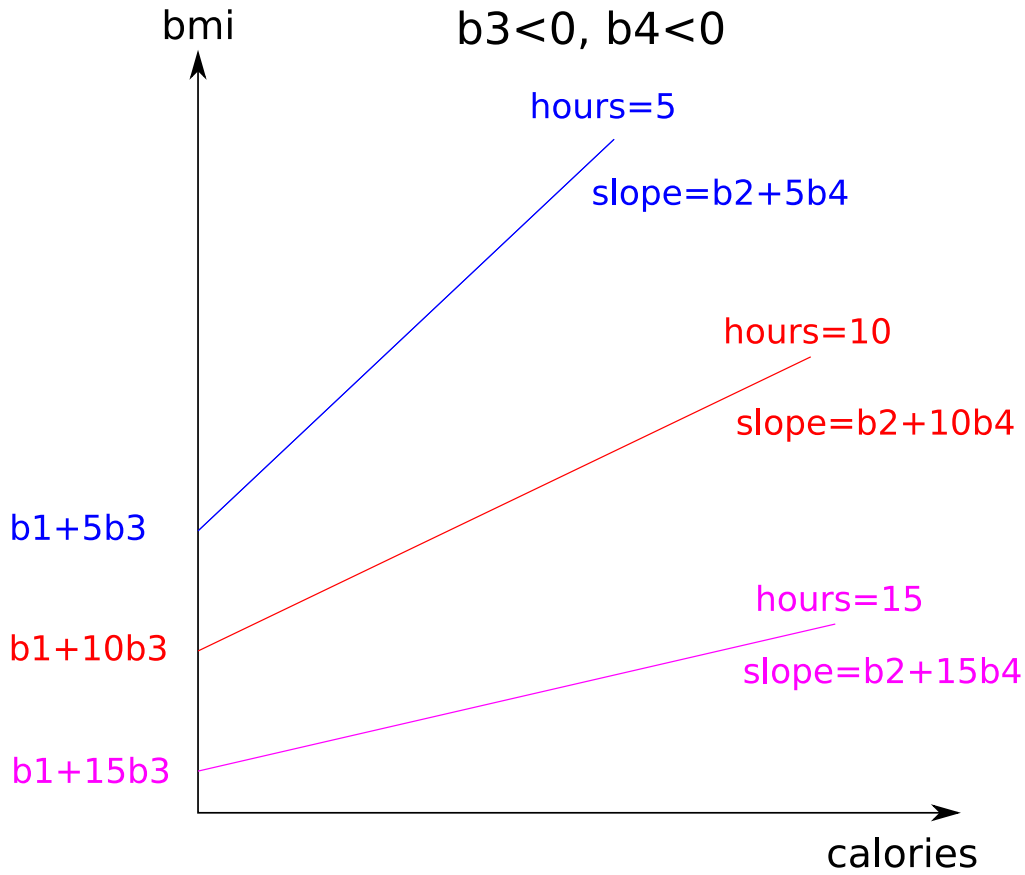
- How do we interpret this interaction term?
- It depends on whether we're most interested in the relationship between bmi and calories or the relationship between bmi and exercise

# Another Case of Interaction Terms

$$\widehat{bmi}_i = b_1 + b_2 cal_i + b_3 hours_i + b_4 cal_i \cdot hours_i$$

- If we care about the relationship between bmi and calories:
$$\frac{\Delta bmi}{\Delta cal} = b_2 + b_4 hours_i$$
- The change in bmi associated with a change in calories depends on the level of exercise
- Assuming $b_2$ is positive, if $b_4$ is positive the change in bmi with a change in calories will be greater for a person who exercises a lot compared to a person who exercises very little
- If $b_4$ is negative, the opposite is true

# Another Case of Interaction Terms



# Another Case of Interaction Terms

$$\widehat{bmi}_i = b_1 + b_2 cal_i + b_3 hours_i + b_4 cal_i \cdot hours_i$$

- If we care about the relationship between bmi and exercise:
$$\frac{\Delta bmi}{\Delta hours} = b_3 + b_4 cal_i$$
- The change in bmi associated with an increase in hours of exercise depends on the level of calories consumed
- If $b_4$ is positive, the change in bmi with an increase in hours of exercise will be greater for a person who eats a lot compared to a person who eats very little
- If $b_4$ is negative, the opposite is true

# Another Case of Interaction Terms

- Suppose we estimated the equation and came up with:

$$\widehat{bmi}_i = 30 + .05cal_i - 2hours_i - .01cal_i \cdot hours_i$$

- Suppose we want to say, "An increase of 100 calories a day is associated with _____ in bmi." To do this we need to pick a value for hours of exercise
- For example, an increase of 100 calories a day is associated with a 3 point increase in bmi for a person who exercises 2 hours a week $(.05 \cdot 100 - .01 \cdot 100 \cdot 2)$
- For what level of exercise will an increase in calories lead to no predicted change in bmi? 5 hours a week $(0 = .05\Delta cal_i - .01\Delta cal_i \cdot 5)$

# Model Misspecification

- We've spent a lot of time on interpreting coefficients and testing hyptheses
- However, everything we've done has been based on a rather strict set of assumptions
- When these assumptions are violated (which happens often), what happens to our results?
- We'll consider a few different ways in which are assumptions can be wrong: we chose the wrong model, errors are correlated with the regressors, errors have nonconstant variance and errors are correlated with each other

# Misspecified Models

- Recall that we assumed the population model was:

$$y = \beta_1 + \beta_2 x_2 + ... + \beta_k x_k + \varepsilon$$

- There are a few ways this model could be wrong
  - We may have omitted important variables
  - We may have included irrelevant variables
  - Relationships may not be linear

# Omitted Variable Bias: Motivation

- Let's think about what happened when we went from bivariate to multivariate regression
- The interpretation of coefficients changed slightly, with multivariate regression the coefficient on $x_j$ told us the change in $y$ with a change in $x_j$ *holding all of the other regressors constant*
- This means that the same variable in a bivariate regession may have a different coefficient when included in a multivariate regression (recall the basketball example from earlier in class)

# Omitted Variable Bias

- Suppose the true model is:

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

- If all our assumptions hold, regressing $y$ on $x_2$ and $x_3$ will get an unbiased estimate $b_2$ ($E(b_2) = \beta_2$)
- Suppose we regress $y$ on just $x_2$, getting:

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

- Will $E(\tilde{b_2}) = \beta_2$? Probably not.

# Omitted Variable Bias

- If $x_2$ is correlated with $x_3$, the coefficient $\widetilde{b_2}$ in the bivariate regression will be picking up the effects of both $x_2$ and of $x_3$
- How big is this effect? It depends on how strong the relationship between $x_2$ and $x_3$ is
- Suppose $x_3$ is related to $x_2$ by:

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

- If we aren't holding $x_3$ constant, a change in $x_2$ will have two effects on $y$:

$$E(\widetilde{b_2}) = \frac{\Delta y}{\Delta x_2} + \frac{\Delta y}{\Delta x_3}\frac{\Delta x_3}{\Delta x_2}$$

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

# Omitted Variable Bias

- So the expected value of $\tilde{b_2}$ is equal to $\beta_2$ plus another term that depends on the relationship between $x_2$ and the omitted variable as well as the omitted variable and the dependent varible
- As long as $\gamma_2$ isn't zero and $\beta_3$ isn't zero, $E(\tilde{b_2})$ won't equal $\beta_2$
- So $\tilde{b_2}$ is a *biased* estimator of the coefficient for $x_2$
- We refer to this as an *omitted variable bias*

# Omitted Variable Bias

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

- There will be an upward bias if $\beta_3 > 0$ and $\gamma_2 > 0$ or if $\beta_3 < 0$ and $\gamma_2 < 0$
- There will be a downward bias if $\beta_3 < 0$ and $\gamma_2 > 0$ or if $\beta_3 > 0$ and $\gamma_2 < 0$
- If $\gamma_2 = 0$, there will be no bias (but our model is incorrect)
- If $\beta_3 = 0$, there will be no bias (and $x_3$ shouldn't be in our model anyway)

# Dealing With Omitted Variable Bias

- What do we do about omitted variable bias?
- The easiest thing is to just include the omitted variable in our regression
- Often this isn't possible due to data limitations
- There are some more advanced techniques that may work (instrumental variables, natural experiments)
- If we can't add the omitted variable to the regression or use a fancy approach, one thing we can still do is try to sign the bias using economic intuition

# Example: Smeed's Law

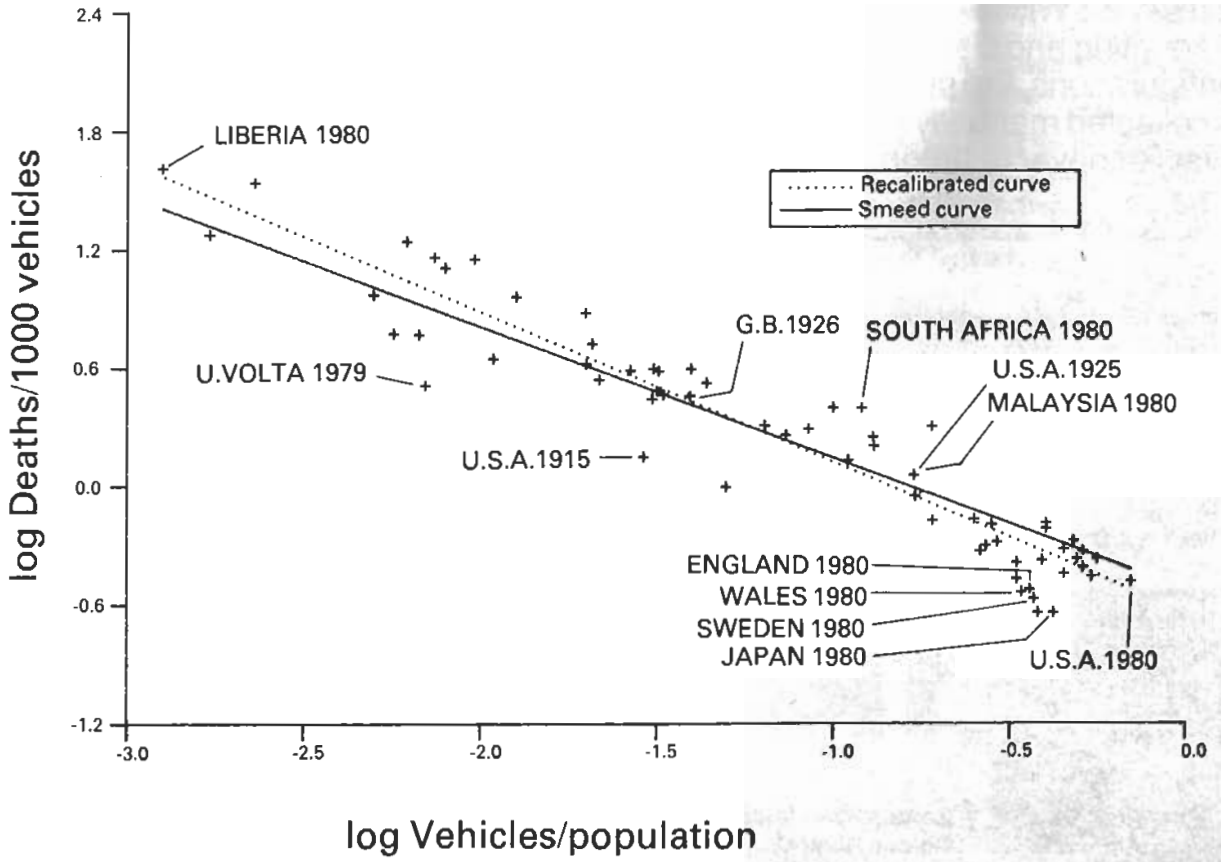



*Figure from John Adams (1987), "Smeed's Law: some further thoughts", Traffic Engineering and Control, 28 (2)*

# Example: Smeed's Law

- A regression of car accidents on the number of cars would give a negative coefficient ($\tilde{b_2} < 0$)
- But there may be a downward bias, why?
  - More cars mean slower speeds due to congestion ($\gamma_2 < 0$)
  - Slower speeds mean fewer accidents ($\beta_3 > 0$)
- If we could hold car speeds constant, more cars may very well lead to more accidents ($\beta_2 > 0$)

# Example: Returns to Education

- Economists have a really hard time coming up with good estimates of returns to education (the change in income associated with an increase in education)
- Why? There are always several important omitted variables
- One of the key ones is ability:
  - High ability people are more likely to go to school ($\gamma_2 > 0$)
  - High ability people will be better at their jobs and earn higher salaries ($\beta_3 > 0$)
  - Omitting ability will lead to an upward bias on the coefficient on education in a wage regression

# Example: Returns to Education

Table 3
Instrumenting schooling with month of birth dependent variable: Log annual income

| | (1) OLS | (2) IV Birthmonth | (3) IV Birthmonth × Birthyear |
|---|---|---|---|
| Years of education | 0.128*** | −0.099 | 0.079** |
| | [0.013] | [0.295] | [0.032] |
| Female | −0.601*** | −0.612*** | −0.602*** |
| | [0.051] | [0.069] | [0.057] |
| Relative position | | −0.035 | 0.000 |
| | | [0.090] | [0.072] |
| Birth year FE? | Yes | Yes | Yes |
| State FE? | Yes | Yes | Yes |
| *F*-test for excluded instruments | — | 0.65 | 554.89 |
| | | $P = 0.6605$ | $P = 0.000$ |
| Observations | 998 | 998 | 998 |
| *R*-squared | 0.22 | 0.21 | 0.22 |

*From Leigh and Ryan (2008), "Estimating returns to education using different natural experiment techniques", Economics of Education Review, 27(2)*

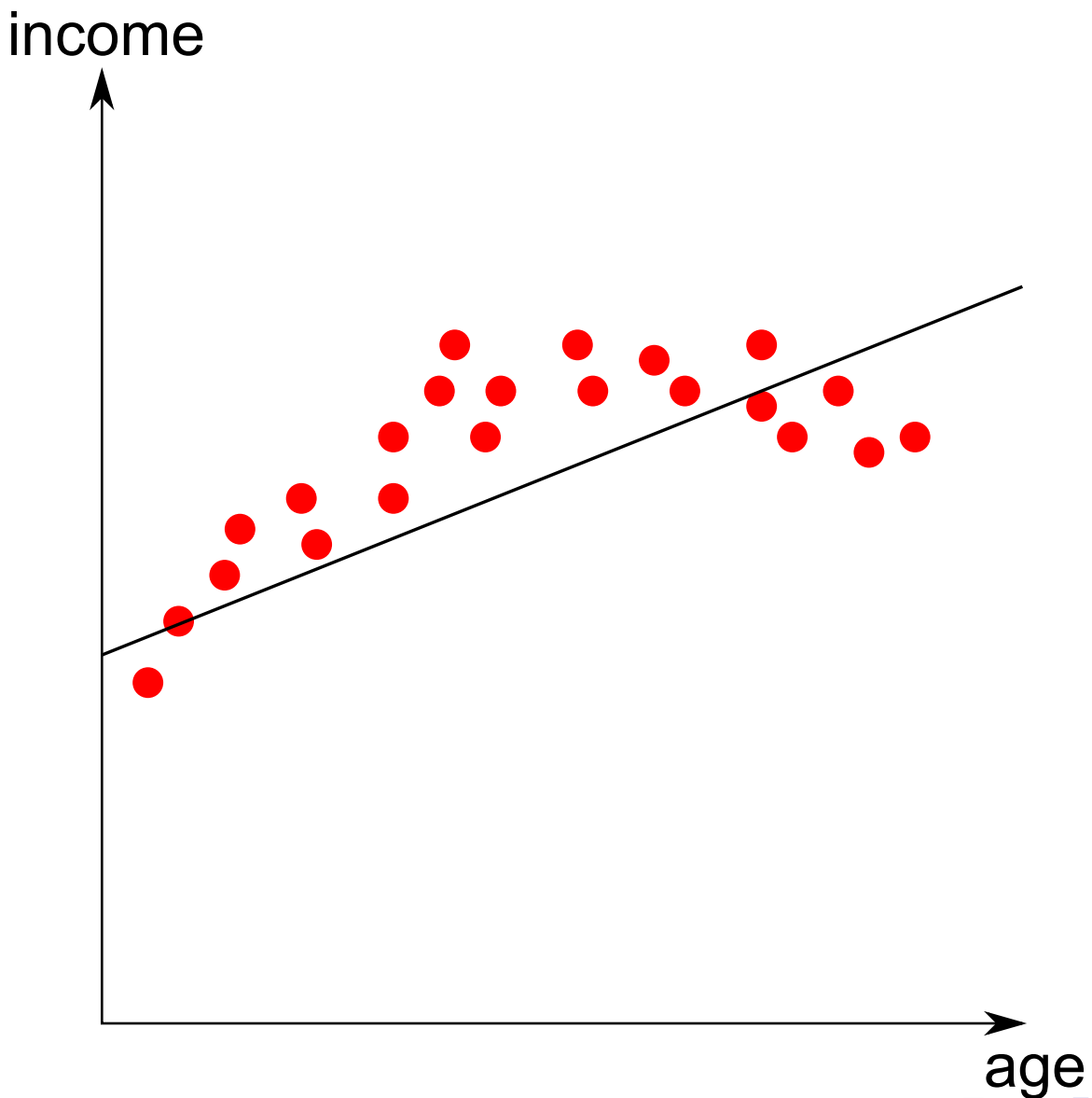
# Including Too Many Variables

- 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

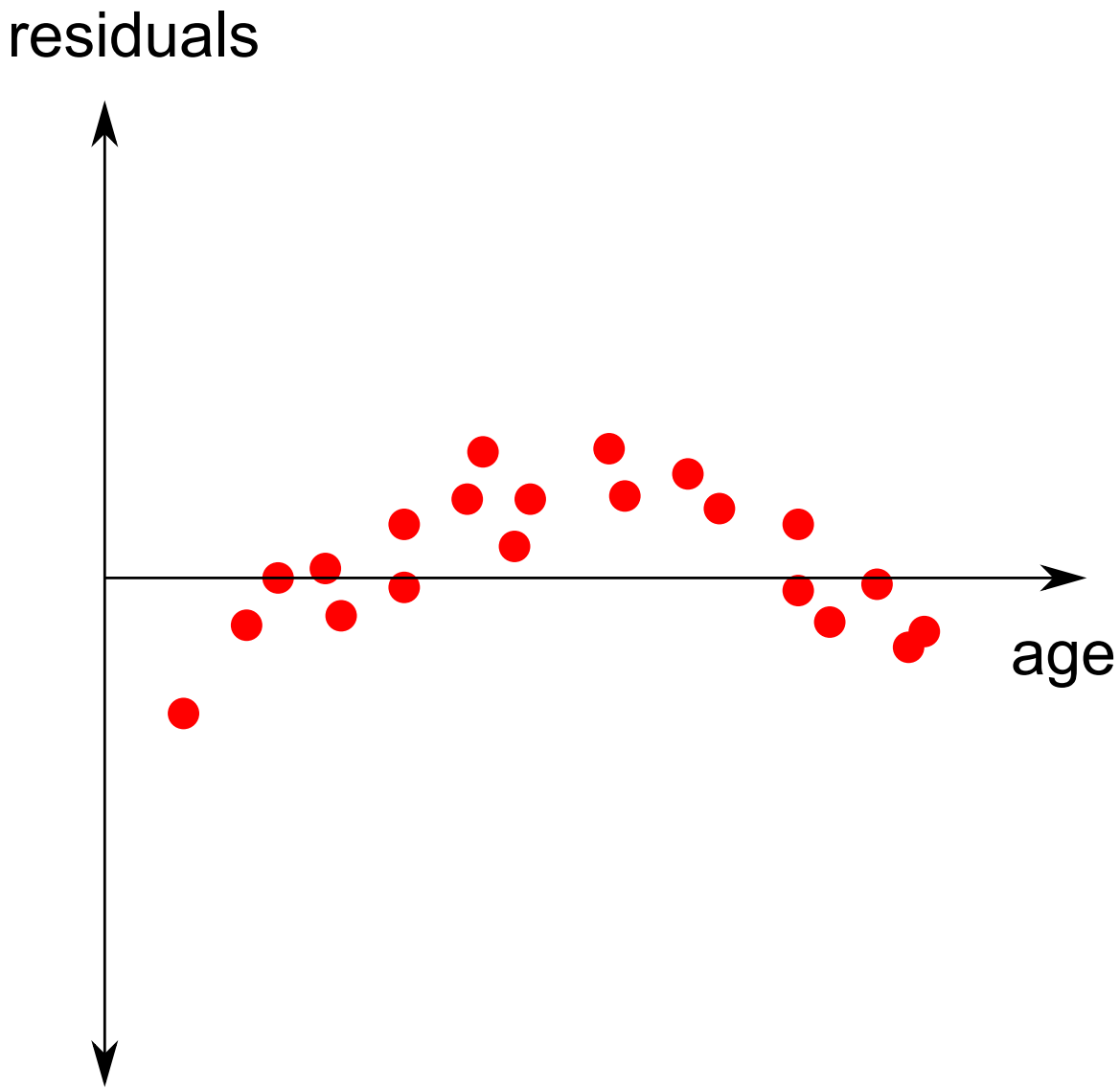
# Non-linear Relationships

- 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