Week 11: Interpretation plus

Marcelo Coca Perraillon

University of Colorado
Anschutz Medical Campus

Health Services Research Methods I
HSMP 7607
2017

Updated notes are here: https://clas.ucdenver.edu/marcelo-perraillon/teaching/health-services-research-methods-i-hsmp-7607

Outline

- A bit of a patchwork class of topics I left behind
- The classic bias-variance trade off
- Overfitting models / saturated models / parsimony
- Stata global variables
- Notation, adding time and correlated errors (panel or longitudinal data)
- Presenting results: adjusting

Bias-variance trade off

- The bias-variance trade-off is a classic problem in statistics
- Many different versions appear in different contexts. The first introduction is usually in the context of the **mean square error**
- The mean square error is the average distance of the square of observed and predicted values:
- $MSE = \sum_{i=1}^{n} (\hat{Y} - Y_i)^2 / n$
- The MSE is a measure of performance in terms of prediction
- **Confusion alert**: We also use mean square error to estimate the explained variance in linear regression (they are related; more in a bit)

Mean square error

- Example of mean square error (MSE) and root mean square error (RMSE)
- Back to college grades example

```
  reg colgpa hsgpa male campus ACT mothcoll

      Source |       SS           df       MS      Number of obs   =       141
-------------+----------------------------------   F(5, 135)       =      6.07
       Model |  3.56184692         5  .712369384   Prob > F        =    0.0000
    Residual |  15.8442525       135  .117364834   R-squared       =    0.1835
-------------+----------------------------------   Adj R-squared   =    0.1533
       Total |  19.4060994       140  .138614996   Root MSE        =    .34259

------------------------------------------------------------------------------
      colgpa |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       hsgpa |   .4575161   .1010688     4.53   0.000      .257633    .6573991
        male |   .0054175   .0624735     0.09   0.931    -.1181358    .1289708
      campus |  -.0704102   .0776221    -0.91   0.366    -.2239228    .0831023
         ACT |   .0087551   .0113207     0.77   0.441    -.0136337    .0311439
    mothcoll |   .0331737   .0584164     0.57   0.571     -.082356    .1487035
       _cons |   1.279981   .3487628     3.67   0.000     .5902354    1.969726
------------------------------------------------------------------------------
```

Mean square error

- Calculating mean square error (MSE) and root mean square error (RMSE)

```
* Predict
predict gpahat

* Square error
gen mse = (gpahat - colgpa)^2

* Mean square error
sum mse

  Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         mse |        141    .1123706    .1356477   5.18e-06   .7398951

* Root mean square error
di sqrt(r(mean))
.33521722
```

Mean square error

- The smaller MSE or RMSE the better in terms of prediction
- Sensitive to large errors; there are other alternatives (**mean absolute error**, for example)
- More generally, the MSE of a **parameter** θ with an estimator $\hat{\theta}$ is
  $MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)^2]$
- After applying the rules of expectations and some algebra, MSE can be decomposed into:
  $MSE(\hat{\theta}) = var(\hat{\theta}) + Bias(\hat{\theta})^2$
- Recall that bias is $E[\hat{\theta} - \theta]$. If an estimator is unbiased, then bias $= 0$

Bias-variance trade off

- If the estimator is unbiased, then MSE is just the variance (hence the connection with OLS)
- Any estimator is a combination of both, bias and variance
- We know that our linear model estimated via OLS or MLE is BLUE; not only it's asymptotically unbiased but it also has smaller variance
- The intuition behind bias and variance of an estimator is often shown with a graph

Bias and variance in a graph

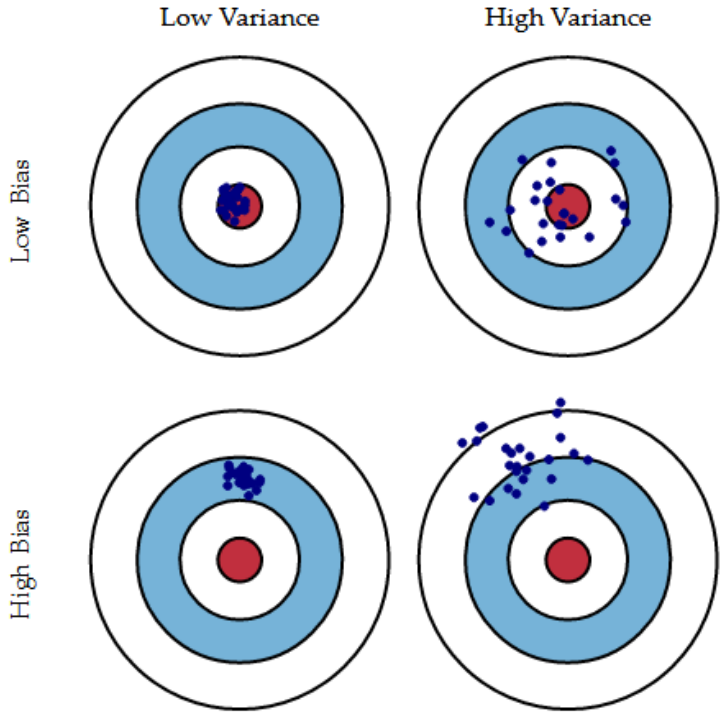


Fig. 1 Graphical illustration of bias and variance.

From
`http://scott.fortmann-roe.com/docs/BiasVariance.html`

Bias-variance trade off

- In many cases we need to choose between reducing bias at the cost of increasing variance or decreasing variance at the cost of increasing bias
- Choosing estimators involve this trade-off; sometimes the trade-off is clear
- In causal inference with confounding (endogeneity if you are an economist), for example, we prefer an instrumental variable approach because we obtain a less biased estimator of local treatment effects at the cost of a less efficient estimator (higher variance)
- In propensity scores, we compare "similar" units to reduce bias but increase the variance because we may restrict the sample to fewer units

Bias-variance trade off

- If we run a regression explaining wage as a function of age, sex, education, etc we could use it to predict or explain the average wage of those with higher education
- But we could also just use the data for those with higher education
- Using their data will reduce the bias because we are not including information on people with other levels of education
- On the other hand, we use less data so we increase the variance
- In regression discontinuity, using observations as close as possible to the cut-off point reduces bias but it also increases the variance because we restrict the sample to fewer observations

Over- and under-fitting models

- The other situation in which the concept emerges in when building models
- Remember that the R^2 will never go down if we add a variable in the model
- We will decrease the bias as measured by the prediction ability of the model, but we will also risk **over-fitting the data**
- If we pay too much attention to goodness of fit (R^2) we may run into problems very quickly
- Problem 1: collinearity. Problem 2: Model will predict well in one dataset but badly in another

Adding a lot of variables to a model

- Global macros in Stata are handy
- Useful for organizing sets covariates into logical groups (demographics, comordibity, SES, etc)
- Keeps code more readable

```
 * Set 1
global covars1 hsgpa male campus ACT mothcoll

* Set 2
global covars2 voluntr PC greek car bgfriend

* Set 3
global covars3 clubs skipped alcohol gradMI
```

Adding a lot of variables to a model

- Use the dollar sign to call the global macro

```
reg colgpa $covars1
scalar m1r2 = e(r2)
est sto m1

reg colgpa $covars1 $covars2
scalar m2r2 = e(r2)
est sto m2

reg colgpa $covars1 $covars2 $covars3
scalar m3r2 = e(r2)
est sto m3

di m1r2 "   " m2r2  "   "  m3r2
.18354265   .24846885   .34430613
```

Adding a lot of variables to a model

- Compare models (note the keep option to avoid displaying some of the coefficients)

```
. est table m1 m2 m3, se p stat(N r2 r2_a) keep($covars1)

----------------------------------------------------
    Variable |     m1           m2           m3
-------------+--------------------------------------
       hsgpa |  .45751607    .46457425    .40659732
             |  .10106884    .10244436    .09877931
             |     0.0000       0.0000       0.0001
        male |   .0054175    .03310817    .04603947
             |  .06247347    .06398082    .06484514
             |     0.9310       0.6057       0.4790
      campus | -.07041023   -.08905737   -.08976193
             |  .07762206    .07857388     .0781437
             |     0.3660       0.2591       0.2529
         ACT |  .00875514    .00560391    .01051044
             |  .01132067     .0111698    .01072888
             |     0.4407       0.6167       0.3291
    mothcoll |  .03317374   -.00254655   -.04668698
             |  .05841643    .05913929    .05854816
             |     0.5711       0.9657       0.4267
-------------+--------------------------------------
           N |        141          141          141
          r2 |  .18354265    .24846885    .34430613
        r2_a |  .15330349    .19065877    .27145126
----------------------------------------------------
                                    legend: b/se/p
```

Adding a lot of variables to model

- Nothing is obviously wrong, the model with more variables fits better using R^2 or R^2_a (as expected)
- Nothing is wrong with the usual collinearity stats (VIF) either (check by yourself)
- But do we really need to control for so many factors? Can we really expect that "holding other factors constant" will be true?
- Hard to give a clear answer; at least in this example, not a lot of theory to guide you (only intuition)
- If you are building a predictive model, however, **overfitting is always a potential problem**. The model might predict well in your dataset but not in another (that's why we often use a validation and estimation sample)
- Wooldridge writes that overffiting often "results from nervousness about potential biases that might arise by leaving out an important explanatory variable"

The purpose of modeling matters

- From Wooldridge, page 198

Consider the housing price example from... In that case, we wanted to test the rationality of housing price assessments. We regressed log(price) on **log(assess)**, log(lotsize), log(sqrft), and bdrms and tested whether the latter three variables had zero population coefficients while log(assess) had a coefficient of unity [assessments should be highly correlated to prices]. But what if we change the purpose of the analysis and estimate a hedonic price model, which allows us to obtain the marginal values of various housing attributes? Should we include log(assess) in the equation? The adjusted R-squared from the regression with log(assess) is .762, while the adjusted R-squared without it is .630. Based on goodness-of fit only, we should include log(assess). But this is incorrect if our goal is to determine the effects of lot size, square footage, and number of bedrooms on housing values. Including log(assess) in the equation amounts to holding one measure of value fixed and then asking how much an additional bedroom would change another measure of value. This makes no sense for valuing housing attributes. If we remember that different models serve different purposes, and we focus on the ceteris paribus interpretation of regression, then we will not include the wrong factors in a regression model.

Saturated models

- **Saturated** model has at least two meanings
- Precise definition: a saturated model has as many parameters as data values
- The predicted values from the saturated model will **fit the data perfectly**
- We can't estimate the variance

Saturated model

- A (useless) model that fits the data perfectly

```
reg colgpa $covars1 $covars2 $covars3 if _n < 12
      Source |       SS           df       MS      Number of obs   =        11
-------------+----------------------------------   F(10, 0)        =         .
       Model |  1.58909072        10  .158909072   Prob > F        =         .
    Residual |           0         0           .   R-squared       =    1.0000
-------------+----------------------------------   Adj R-squared   =         .
       Total |  1.58909072        10  .158909072   Root MSE        =         0

------------------------------------------------------------------------------
      colgpa |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       hsgpa |          0  (omitted)
        male |  -.6000001          .       .       .            .           .
      campus |   .0312501          .       .       .            .           .
         ACT |      .0125          .       .       .            .           .
    mothcoll |  -.8203124          .       .       .            .           .
     voluntr |          0  (omitted)
          PC |   .2984375          .       .       .            .           .
       greek |          0  (omitted)
         car |  -.4750001          .       .       .            .           .
    bgfriend |       .425          .       .       .            .           .
       clubs |   .5265624          .       .       .            .           .
     skipped |  -.1156251          .       .       .            .           .
     alcohol |     -.2875          .       .       .            .           .
      gradMI |          0  (omitted)
       _cons |   3.731251          .       .       .            .           .
------------------------------------------------------------------------------
```

Saturated model

- Perfect prediction for the **observations that were used to estimate the model**

```
predict hat2
list colgpa hat2 in 1/14
```

	colgpa	hat2
1.	3	3
2.	3.4	3.4
3.	3	3
4.	3.5	3.5
5.	3.6	3.6
6.	3	3
7.	2.7	2.7
8.	2.7	2.7
9.	2.7	2.7
10.	3.8	3.8
11.	2.8	2.8
12.	2.9	3.290625
13.	3	3.717188
14.	2.9	1.639062

Saturated model

- The second common meaning (less precise) is for models in which we use all the available variables we measured but we have many more observations than explanatory variables
- Sometimes people use the term saturated models for models that include a large set of variables and their interactions (less precise, too)
- But the technical meaning is reserved for models with as many parameters as observations

Big picture

- Not a good idea to include **too many variables** but hard to tell how many is too many
- Good idea to build models from small to large but not too large
- Good idea to think if "holding other variables constant" is plausible even if VIF or correlation among variables is not that high (**think conceptually**)
- Building models by first focusing on relationship of interest and then adding more complexity helps a lot
- It helps you figure out which variable is important – which variable changes p-values, coefficients, R^2, etc

Panel data (aka longitudinal data)

- We won't cover **panel data** (economics) or **longitudinal data** (statistics and the rest of the world)
- Suppose that we have the model

  $wage_i = \beta_0 + \beta_1 age_i + \beta_2 education_i + \beta_3 looks_i + \beta_4 female_i + \epsilon_i$
- With $\epsilon_i \sim N(0, \sigma^2)$ and i indexes an individual
- But now instead of measuring each person one time we have repeated measurements, say, every 3 years
- These type of models require different methods because now ϵ_i is **not independent**; it's still identically distributed but **not iid anymore**

Panel data (aka longitudinal data)

- Clearly, the error for person i in the first measurement is not independent from the error in the second measurement
- Remember that a key step in MLE was that we could multiply the likelihood because observations were independent
- In essence, longitudinal data analysis is about modeling the dependence among observations
- It's the same as **clustered** data (groups that are similar are correlated)
- Notation changes too

Panel data (aka longitudinal data)

- To make it clear that the model is different, we change the notation. For example,
- $wage_{it} = \beta_0 + \beta_1 age_{it} + \beta_2 education_{it} + \beta_3 looks_{it} + \beta_4 female_{it} + \epsilon_{it}$
- We read it as "wage for individual i at time t is a function age for individual i at time t..."
- If we are even more careful, we can drop the subscript t for female because (usually) sex doesn't not depend on time
- You will cover some panel data methods next semester. It's a fairly large field in statistics
- Mixed models (random effects), hierarchical models, GEE, Generalized Linear Mixed Models are examples of models for longitudinal data

Clustered data

- With clustered data, you can have something like

  $wage_{is} = \beta_0 + \beta_1 age_{is} + \beta_2 education_{is} + \beta_3 looks_{is} + \beta_4 female_{is} + \epsilon_{is}$
- We read it as wage for "individual i in state s is a function age for individual i in state s..."
- Of course, you could have three subscripts: "$wage_{its}$: wage for person i at time t in state s"
- The notation helps you understand the model (**don't be careless!**)
- **Take a class on longitudinal data analysis**

Adjusting

- Remember the dataset on income and age we used for the splines example
- We can show the predicted values of the regression between income and age very easily (only one predictor)
- But once we add education we get a strange looking graph

Adjusting

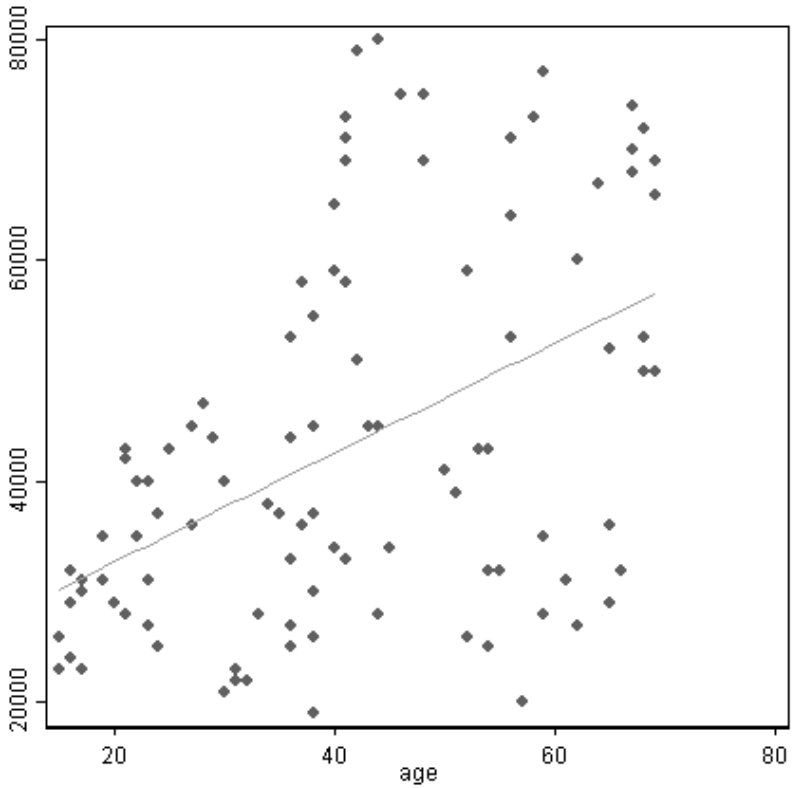
- Comparing two models

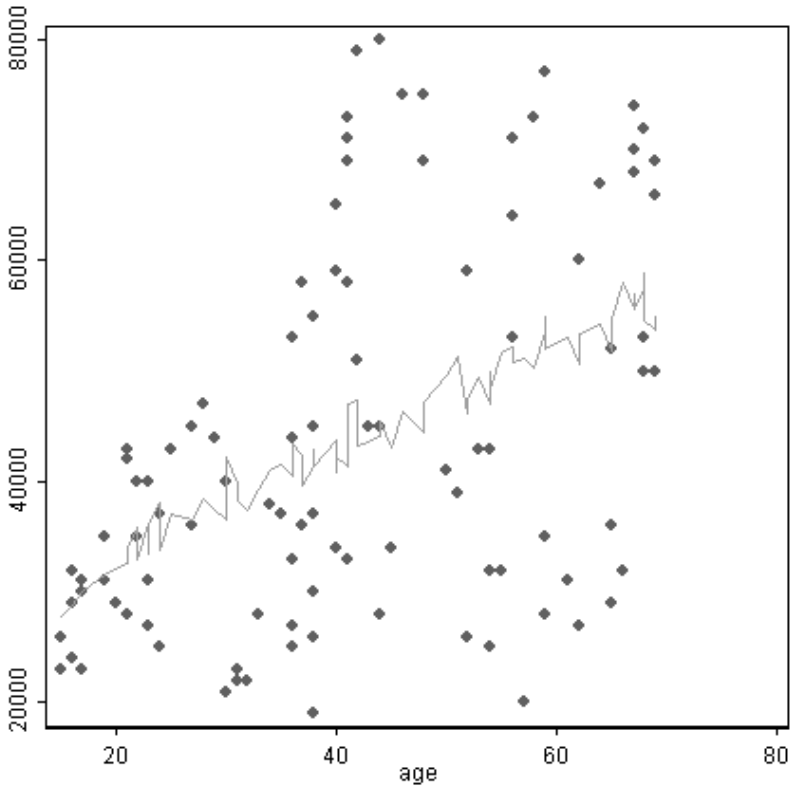
```
* Just age
reg income age
predict inchat
scatter income age || line inchat age, sort legend(off) saving(g1.gph, replace)
graph export g1.png, replace


* Age and education
reg income age educ
predict inchat1
scatter income age || line inchat1 age, sort legend(off) saving(g2.gph, replace)
graph export g2.png, replace

graph combine g1.gph g2.gph
graph export gc.png, replace
```

Comparing models





- In second graph, predictions depend on years of education, too

Adjusting

- We could instead present the relationship between age and income **holding education constant**
- There is no rule at which value you hold the other constant, typically at the mean
- If you have a 0/1 variable, you typically hold it at the mean, too, which is a proportion
- You could also compare the graphs at different values of education, which can be fairly useful
- We will do both

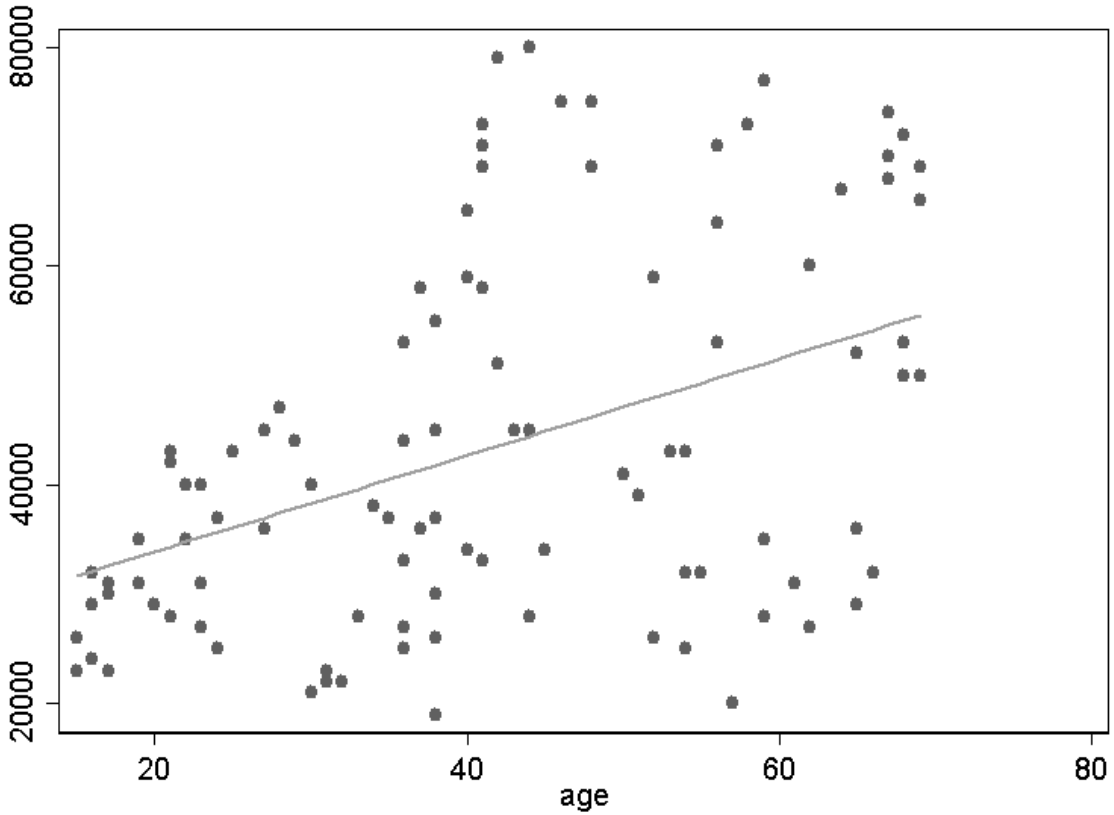
Adjusting at mean

- Hold education at its mean
- Note the **preserve** and **restore** options. If in interactive mode, run all the code from preserve to restore to avoid problems
- The idea is to change the data to make the prediction but then go back to the original dataset
- Note too that I **should have used the e(sample)** option (I knew there were no missings but better to add it)

```
qui sum educ
di r(mean)
14.58

preserve
   reg income age educ
   replace educ = 14.58
   predict inchat_m
   scatter income age || line inchat_m age, sort legend(off) ///
           saving(g_m.gph, replace)
   graph export gm.png, replace
restore
```

Adjusting at mean



- Of course, still a linear relationship between age and income but now education is held at 14.58 in all the predictions

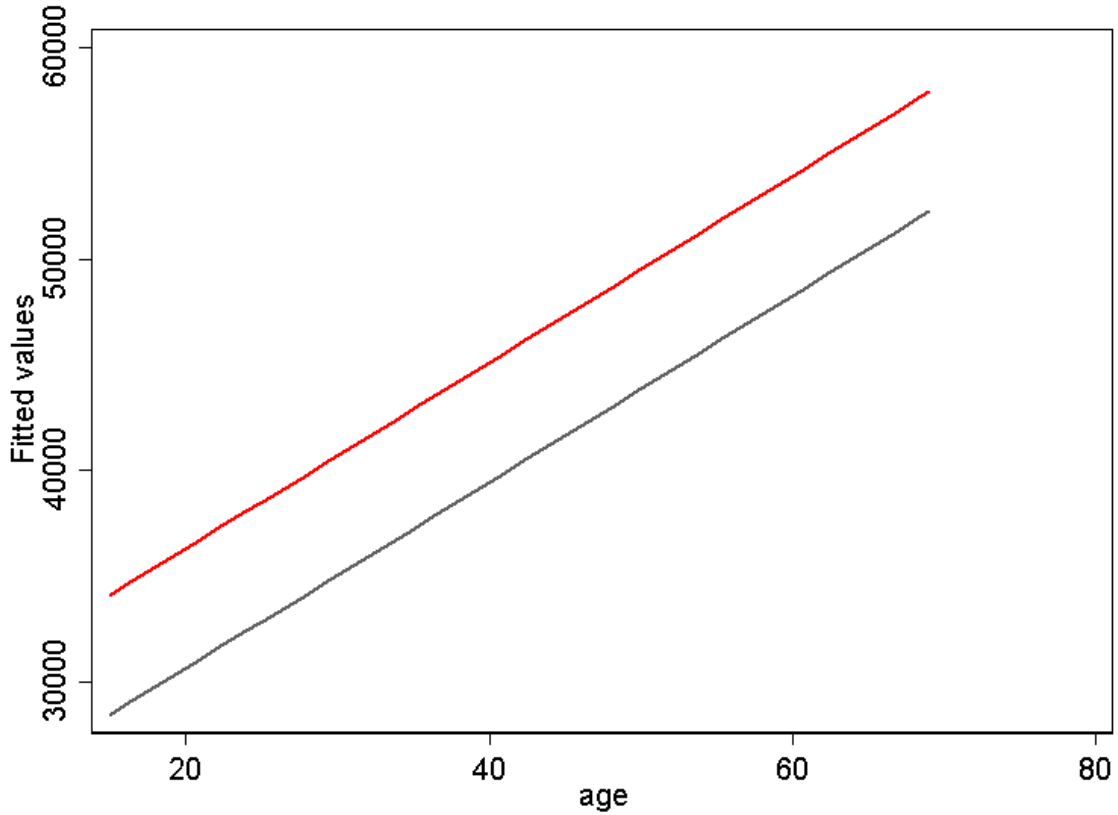
Adjusting at mean

- Compare models with education = 10 and education = 18 (less than high school and masters' degree)

```
preserve
   reg income age educ
   replace educ = 10
   predict inchat_10
   replace educ = 18
   predict inchat_18

   line inchat_10 age, sort || line inchat_18 age, sort color(red) ///
        saving(g_10_18.gph, replace) legend(off)
   graph export g10_18.png, replace
restore
```

Adjusting at mean



- As expected, more education, higher income. But why parallel? **We assumed so** in model

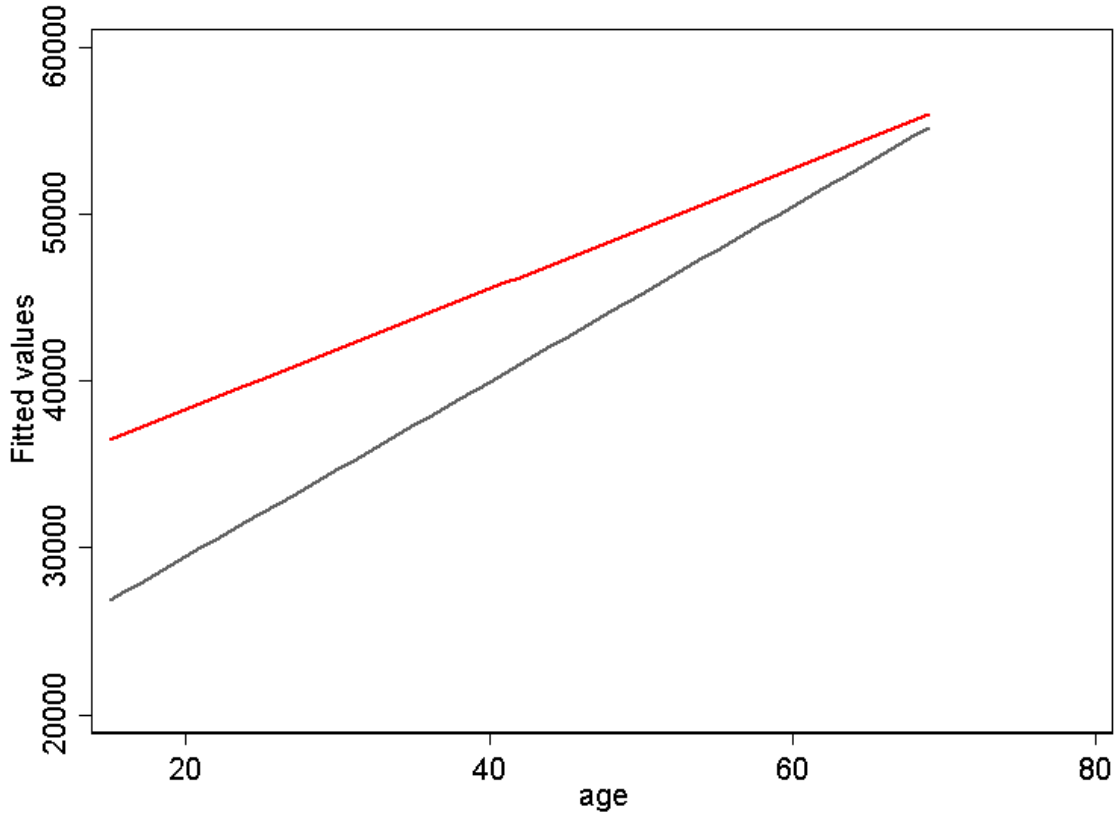
Adjusting at different values

- Assume interactions (so different slope and different intercept)

```
preserve
   reg income c.age##c.educ
   replace educ = 10
   predict inchat_101
   replace educ = 18
   predict inchat_181

   line inchat_101 age, sort || line inchat_181 age, sort color(red) ///
        saving(g_10_181.gph, replace) legend(off)
   graph export g10_181.png, replace
restore
```

Adjusting at different values, interacted



- More interesting. **But we just assumed so, too**. We would need to test and compare models to figure out which one fits the data better

Summary

- Bias and variance is a key concept in statistics that shows up in many forms
- **Make models as simple as possible but not simpler**
- Parsimony is a good thing with modeling
- Get used to presenting adjusted graphs; we will see more ways of presenting predictions next week