Lecture Notes 6

Econ B2000, Statistics and Introduction to Econometrics

Kevin R Foster, the City College of New York, CUNY

Fall 2014

(yeah, these notes start to get skimpy – you might want to, you know, actually go to class!)

CPS Data

We have been using various data sets; today we'll use another well-known data set, the Current Population Survey. This dataset has over 200,000 people; it is the basis for the US unemployment statistics. The Bureau of Labor Statistics (BLS) and Census work together to put together and maintain this data; every March a new group of people rotates in while the old group (who answered questions for the past year) rotate out. It is all publicly available: if some wacko thinks the government is fudging the unemployment statistics, they can go and re-calculate everything on their own to check. I've put the data file, `cps_mar2013.RData`, onto InYourClass along with `cps_mar2013_initial_recoding.r` (which you need not run) that created the file from the original data (and has the details of the coding) as well as `cps_1.R`.

Can run a basic linear regression to find what are principal determinants of wage/salary levels (looking at a subset of prime-age, fulltime, year-round workers):

```
model1 <- lm(WSAL_VAL ~ Age + female + AfAm + Asian + Amindian +
race_oth + Hispanic + educ_hs + educ_smcoll + educ_as +
educ_bach + educ_adv + married + divwidsep + union_m + veteran +
immigrant + immig2gen)
     coeftest(model1)
```

This gives an estimate of how important are various educational qualifications.

Statistical Significance

Statistical significance of coefficient estimates is the same when we look at individual coefficients but more complicated for multiple coefficients: we can ask whether a group of variables are jointly significant, which takes a more complicated test. We can even ask if all of the slope coefficients together are statistically significant.

For a univariate regression, if the single slope coefficient is statistically significant then the overall regression is as well (the F statistic is the square of the t-stat in that case).

The difference between the overall regression fit and the significance of any particular estimate is that a hypothesis test of one particular coefficient tests if that parameter is zero; is

$\beta_i = 0$? This uses the t-statistic $t = \frac{\hat{\beta}}{se\left(\hat{\beta}\right)}$ and compares it to a t distribution. The test of the regression significance tests if ALL of the slope coefficients are simultaneously zero; if $\beta_1 = \beta_2 = \beta_3 = \ldots = \beta_K = 0$. The latter is much more restrictive. (See Chapter 7 of Stock & Watson.)

It is often sensible to make joint tests of regression coefficients, for example with a group of dummy variables. If we have a set of dummies for education levels, it is strange to think of omitting just one or two; it is more reasonable to ask whether education measures (overall) are statistically significant. We might also want to know if individual coefficients are equal to each other (e.g. to ask if going to college, without getting any degree, is really different from the estimate for just a high school diploma.

To do this in R, there is a package, `linearHypothesis` (part of the package, `car`, Companion to Applied Regression, which is auto-loaded by `AER` package). But the commands shouldn't obscure the simple basic point: we evaluate variables based on how well they fit in the model.

To consider the question of whether a set of variables is statistically significant, we basically are just looking at how big is the error (the Sum of Squared Errors) with and without those variables. In general adding more variables to the model can never make the errors bigger (can never increase the Sum of Squared Errors) – basically this is a statement that the Marginal Benefit of more variables can never be negative. But profit maximization requires that we balance Marginal Benefit against Marginal Cost – what is the marginal cost of adding more variables? Statistical significance is one measure of profitability in this sense.

If adding new predictors makes the error "a lot" smaller, then those predictors are jointly statistically significant. The essence of statistical testing is just finding a good metric for "a lot".

Note that we can only properly make comparisons within models – it doesn't make much sense to look across models. If I have a model of the fraction of income spent on food, and another model of the level of income, it is difficulty to sensibly pose a question like, "in which model is education more important?" It would be like asking who scored more points per game, Shaq or Jeter? – you can ask the question but it's difficult to interpret in a sensible way.

But within a model we can make comparisons and many of them come down to asking, how much smaller are the errors? (Did the Sum of Squared Errors fall by a lot?) Sometimes it is easiest to just estimate the model twice, with or without the variables of interest, and look at how much the Sum of Squared Errors (from ANOVA in R) fell. But once you get some experience, you'll appreciate `linearHypothesis`.

Why do we always leave out a dummy variable? Multicollinearity. (See Chapter 6 of Stock & Watson.)

- OLS basic assumptions:
    - The conditional distribution of u_i given X_i has a mean of zero. This is a complicated way of saying something very basic: I have no additional information outside of the model, which would allow me to make better guesses. It can also be expressed as implying a zero correlation between X_i and u_i. We will work up to other methods that incorporate additional information.
    - The X and errors are i.i.d. This is often not precisely true; on the other hand it might be roughly right, and it gives us a place to start.
    - X and errors don't have values that are "too extreme." This is technical (about existence of fourth moments) and broadly true, whenever the X and Y data have a limit on the amount of variation, although there might be particular circumstances where it is questionable (sometimes in finance).
- So if these are true then the OLS are unbiased and consistent. So $E\left[\hat{\beta}_0\right] = \beta_0$ and $E\left[\hat{\beta}_1\right] = \beta_1$. The normal distribution, as the sample gets large, allows us to make hypothesis tests about the values of the betas. In particular, if you look back to the "eyeball" data at the beginning, you will recall that a zero value for the slope, β_1, is important. It implies no relationship between the variables. So we will commonly test the estimated values of β against a null hypothesis that they are zero.

Heteroskedasticity-consistent errors

You can choose to use heteroskedasticity-consistent errors as in the textbook.

The Stock and Watson textbook uses heteroskedasticity-consistent errors (sometimes called Eicker-Huber-White errors, after the various authors who figured out how to calculate them). Later you can additionally specify heteroskedasticity- and autocorrelation-consistent (HAC) errors, sometimes called Newey-West

Heteroskedasticity-Consistent Errors in R

These are HCerrors, in the "sandwich" package, which depends on "zoo" package; probably the easiest implementation is via the "lmtest" package. So install those 3.
On my Win7 machine, I find it easiest to download the packages, "Install from local zip file", [that way I don't need a wifi connection every time] then drop in these commands,

```
library("zoo")
library("sandwich")
library("lmtest")
```

For heteroskedasticity-consistent errors, use the coeftest() function but add the command, vcovHC. So from example of CPS data, use:

coeftest(model1,vcovHC)

The command `coeftest` will do a variety of coefficient tests; if you don't play with the defaults, you get the same standard errors as in the summary. If you use `vcovHC`, you get the heteroskedasticity-consistent standard errors. (Econometricians have worked their little butts off, coming up with variations on these, so there are HC0 through HC5 just in this package, don't worry for now about which one to use.)

If you compare the two sets of output, you should notice that the actual coefficient estimates are unchanged – it's the estimated standard errors that change. Then those changes propagate through, so the t-statistics and p-values also change. There is no generic result for whether the estimated standard errors are always bigger or smaller and even in the output from this simple case it goes both ways. However the standard errors often tend to be bigger with the heterskedasticity correction (which means that – test yourself! – the t-statistics are ___ [*bigger or smaller in absolute value?*] and p-values are ___ [*bigger or smaller?*]).

Nonlinear Regression

(more properly, **How to Jam Nonlinearities into a Linear Regression**)

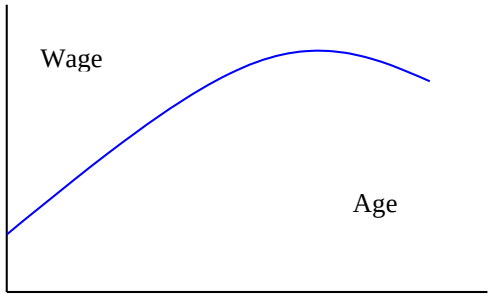
- $X, X^2, X^3, \ldots X^r$
- ln(X), ln(Y), both ln(Y) & ln(X)
- dummy variables
- interactions of dummies
- interactions of dummy/continuous
- interactions of continuous variables

There are many examples of, and reasons for, nonlinearity. In fact we can think that the most general case is nonlinearity and a linear functional form is just a convenient simplification which is sometimes useful. But sometimes the simplification has a high price. For example, my kids believe that age and height are closely related – which is true for their sample (i.e. mostly kids of a young age, for whom there is a tight relationship, plus 2 parents who are aged and tall). If my sample were all children then that might be a decent simplification; if my sample were adults then that's lousy.

The usual justification for a linear regression is that, for any differentiable function, the Taylor Theorem delivers a linear function as being a close approximation – but this is only within a neighborhood. We need to work to get a good approximation.

Nonlinear terms

We can return to our regression using CPS data. First, we might want to ask why our regression is linear. This is mostly convenience, and we can easily add non-linear terms such as Age^2, if we think that the typical age/wage profile looks like this:



So the regression would be:

$$Wage_i = \beta_0 + \beta_1 Age_i + \beta_2 Age_i^2 + \ldots + \varepsilon_i$$

(where the term "..." indicates "other stuff" that should be in the regression).

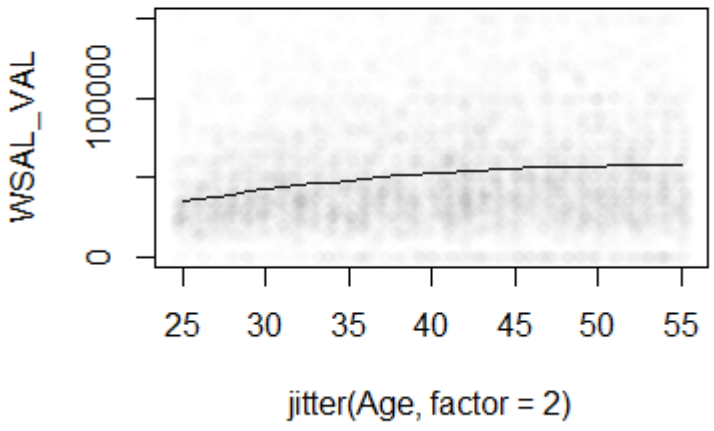
As we remember from calculus,

$$\frac{dWage}{dAge} = \beta_1 + \beta_2 \cdot 2 \cdot Age$$

so that the extra "boost" in wage from another birthday might fall as the person gets older, and even turn negative if the estimate of $\beta_2 < 0$ (a bit of algebra can solve for the top of the hill by finding the Age that sets $\frac{dWage}{dAge} = 0$).

We can add higher-order effects as well. Some labor econometricians argue for including Age^3 and Age^4 terms, which can trace out some complicated wage/age profiles. However we need to be careful of "overfitting" – adding more explanatory variables will never lower the R^2.

To show this in R, I will do a lot of plots – details in `cps_1.R`. (below)



Logarithms

Similarly can specify X or Y as ln(X) and/or ln(Y). But we've got to be careful: remember from math (or theory of insurance from Intermediate Micro) that E[ln(Y)] **IS NOT EQUAL TO** ln(E[Y]) ! In cases where we're regressing on wages, this means that the log of the average wage is not equal to the average log wage.

(Try it. Go ahead, I'll wait.)

When both X and Y are measured in logs then the coefficients have an easy economic interpretation. Recall from calculus that with $y = \ln(x)$ and $\frac{dy}{dx} = \frac{1}{x}$, so $dy = \frac{dx}{x} = \%\Delta x$ -- our usual friend, the percent change. So in a regression where both X and Y are in logarithms, then $\beta_j = \frac{\Delta y}{\Delta x} = \frac{\%\Delta y}{\%\Delta x}$ is the elasticity of Y with respect to X.

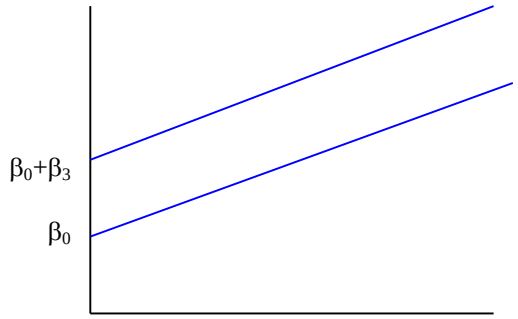
Also, if Y is in logs and D is a dummy variable, then the coefficient on the dummy variable is just the percent change when D switches from zero to one.

So the choice of whether to specify Y as levels or logs is equivalent to asking whether dummy variables are better specified as having a constant level effect (i.e. women make $10,000 less than men) or having a percent change effect (women make 25% less than men). As usual there may be no general answer that one or the other is always right!

Recall our discussion of dummy variables, that take values of just 0 or 1, which we'll represent as D_i. Since, unlike the continuous variable Age, D takes just two values, it represents a shift of the constant term. So the regression,

$$Wage_i = \beta_0 + \beta_1 Age_i + \beta_3 D_i + u_i$$

shows that people with D=0 have intercept of just β_0, while those with D=1 have intercept equal to $\beta_0 + \beta_3$. Graphically, this is:



We need not assume that the β_3 term is positive – if it were negative, it would just shift the line downward. We *do* however assume that the *rate* at which age increases wages is the same for both genders – the lines are parallel.

The equation could be also written as

$$Wage_i = \begin{cases} \beta_0 + \beta_1 Age_i + u_i & for\ D = 0 \\ \beta_0 + \beta_3 + \beta_1 Age_i + u_i & for\ D = 1 \cdot \end{cases}$$

Dummy Variables Interacting with Other Explanatory Variables

The assumption about parallel lines with the same slopes can be modified by adding interaction terms: define a variable as the product of the dummy times age, so the regression is

$$Wage_i = \beta_0 + \beta_1 Age_i + \beta_3 D_i + \beta_4 D_i Age_i + u_i$$

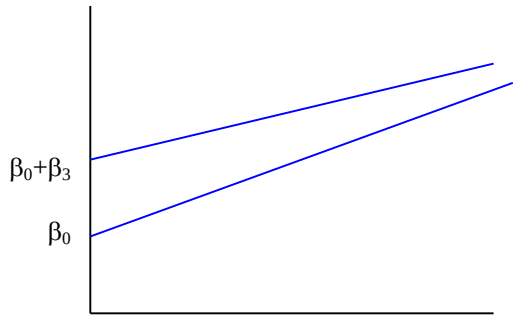
or

$$Wage_i = (\beta_0 + \beta_3 D_i) + (\beta_1 + \beta_4 D_i) Age_i + u_i$$

or

$$Wage_i = \begin{cases} \beta_0 + \beta_1 Age_i + u_i & for\ D = 0 \\ (\beta_0 + \beta_3) + (\beta_1 + \beta_4) Age_i + u_i & for\ D = 1 \end{cases}$$

so that, for those with D=0, as before $\frac{\Delta Wage}{\Delta Age} = \beta_1$ but for those with D=1, $\frac{\Delta Wage}{\Delta Age} = \beta_1 + \beta_4$. Graphically,

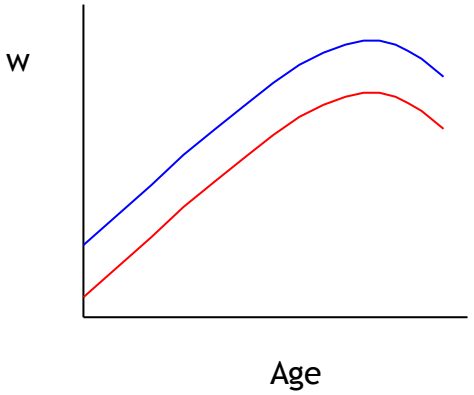


so now the intercepts and slopes are different.

So we might wonder if men and women have a similar wage-age profile. We could fit a number of possible specifications that are variations of our basic model that wage depends on age and age-squared. The first possible variation is simply that:

$$Wage_i = \beta_0 + \beta_1 Age_i + \beta_2 Age_i^2 + \beta_3 D_i + u_i ,$$

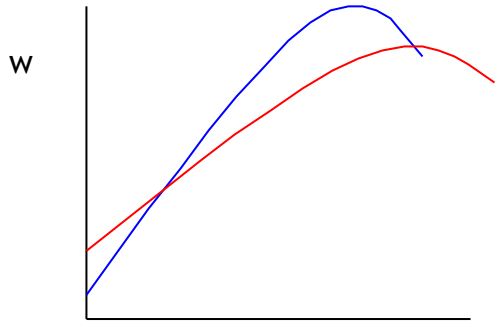
which allows the wage profile lines to have different intercept-values but otherwise to be parallel (the same hump point where wages have their maximum value), as shown by this graph:



The next variation would be to allow the lines to have different slopes as well as different intercepts:

$$Wage_i = \beta_0 + \beta_1 Age_i + \beta_2 Age_i^2$$
$$+\beta_3 D_i + \beta_4 D_i Age_i + \beta_5 D_i Age_i^2 + u_i$$

which allows the two groups to have different-shaped wage-age profiles, as in this graph:



(The wage-age profiles might intersect or they might not – it depends on the sample data.)

We can look at this alternately, that for those with D=0,

$$wage = \beta_0 + \beta_1 Age + \beta_2 Age^2$$
$$\frac{dWage}{dAge} = \beta_1 + 2\beta_2 Age$$

so the extreme value of Age (where $\frac{dWage}{dAge} = 0$) is $\frac{-\beta_1}{2\beta_2}$.

While for those with D=1,

$$Wage = \beta_0 + \beta_1 Age + \beta_2 Age^2 + \beta_3 + \beta_4 Age + \beta_5 Age^2$$
$$= (\beta_0 + \beta_3) + (\beta_1 + \beta_4) Age + (\beta_2 + \beta_5) Age^2$$
$$\frac{dWage}{dAge} = (\beta_1 + \beta_4) + 2(\beta_2 + \beta_5) Age$$

so the extreme value of Age (where $\frac{dWage}{dAge}=0$) is $\frac{-(\beta_1+\beta_3)}{2(\beta_2+\beta_4)}$. Or write the general value, for both cases, as $\frac{-(\beta_1+\beta_3 D)}{2(\beta_2+\beta_4 D)}$ where D is 0 or 1.

This specification, with a dummy variable multiplying each term: the constant and all the explanatory variables, is equivalent to running two separate regressions: one for men and one for women:

$D=0$

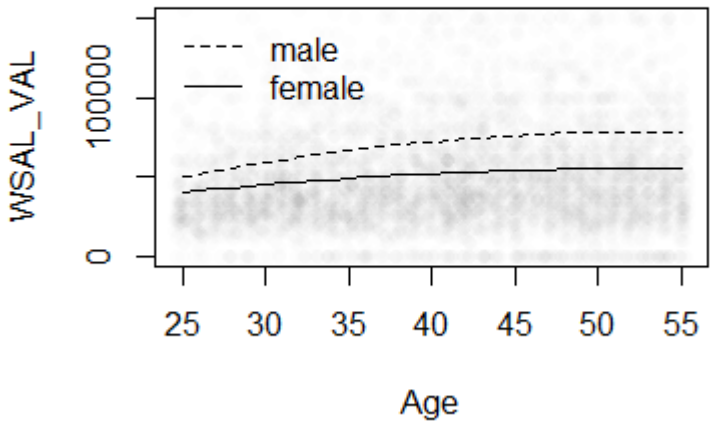
$$Wage_i=\beta_0^{male}+\beta_1^{male}Age_i+\beta_2^{male}Age_i^2+u_i$$

$D=1$

$$Wage_i=\beta_0^{female}+\beta_1^{female}Age_i+\beta_2^{female}Age_i^2+e_i$$

Where the new coefficients are related to the old by the identities: $\beta_0^{female}=\beta_0+\beta_3$, $\beta_1^{female}=\beta_1+\beta_4$, and $\beta_2^{female}=\beta_2+\beta_5$. Sometimes breaking up the regressions is easier, if there are large datasets and many interactions.

The plot for the CPS data is (code is below):



Testing if All the New Variable Coefficients are Zero

You're wondering how to tell if all of these new interactions are worthwhile. Simple: Hypothesis Testing! There are various formulas, some more complicated, but for the case of homoskedasticity the formula is relatively simple.

Why any formula at all – why not look at the t-tests individually? Because the individual t-tests are asking if each individual coefficient is zero, not if it is zero and others as well are also zero. That would be a stronger test.

To measure how much a group of variables contributes to the regression, we look at the residual values – how much is still unexplained, after the various models? And since this is OL**S**, we look at the **squared** residuals. R outputs the Sum of Squares for the Residuals in the ANOVA. We compare the sum of squares from the two models and see how much it has gone down with the extra variables. A big decrease indicates that the new variables are doing good work. And how do we know, how big is "big"? Compare it to some given distribution, in this case the F distribution. Basically we look at the percent change in the sum of squares, so something like:

$$F \approx \frac{SSR_0 - SSR_1}{SSR_1}$$

with the wavy equals sign to show that we're not quite done. Note that model 0 is the original model and model 1 is the model with the additional regressors, which will have a smaller residual (so this F can never be negative).

To get from approximately equal to an equals sign, we need to make it a bit like an elasticity – what is the percent change in the number of variables in the model? Suppose that we have N observations and that the original model has K variables, to which we're considering adding Q more observations. Then the original model has (N – K – 1) degrees of freedom [that "1" is for the constant term] while the new model has (N – K – Q – 1) degrees of freedom, so the difference is Q. So the percent change in degrees of freedom is $\frac{Q}{N-K-Q-1}$. Then the full formula for the F test is

$$F = \left(\frac{SSR_0 - SSR_1}{SSR_1}\right) \Big/ \left(\frac{Q}{N-K-Q-1}\right).$$

Which is, admittedly, fugly. But we know its distribution, it's F with (Q, N-K-Q-1) degrees of freedom – the F-distribution has 2 sets of degrees of freedom. Calculate that F, then use R to find pf(F, df1 =Q, df2 = (N-K-Q-1)) (or Excel to calculate FDIST(F,Q,N-K-Q-1)), to find a p-value for the test. If the p-value is less than 5%, reject the null hypothesis.

Usually you will have the computer spit out the results for you. In R, anova(model1, model2).

Multiple Dummy Variables

Multiple dummy variables, $D_{1,i}$, $D_{2,i}$, ...,$D_{J,i}$, operate on the same basic principle. Of course we can then have many further interactions! Suppose we have dummies for education and immigrant status. The coefficient on education would tell us how the typical person (whether immigrant or native) fares, while the coefficient on immigrant would tell us how the typical immigrant (whatever her education) fares. An interaction of "more than Bachelor's degree" with "Immigrant" would tell how the typical highly-educated immigrant would do beyond how the "typical immigrant" and "typical highly-educated" person would do (which might be different, for both ends of the education scale).

Many, Many Dummy Variables

Don't let the name fool you – you'd have to be a dummy not to use lots of dummy variables. For example regressions to explain people's wages might use dummy variables for the industry in which a person works. Regressions about financial data such as stock prices might include dummies for the days of the week and months of the year.

Dummies for industries are often denoted with labels like "two-digit" or "three-digit" or similar jargon. To understand this, you need to understand how the government classifies industries. A specific industry might get a 4-digit code where each digit makes a further more detailed classification. The first digit refers to the broad section of the economy, as goods pass from the first producers (farmers and miners, first digit zero) to manufacturers (1 in the first digit for non-durable manufacturers such as meat processing, 2 for durable manufacturing, 3 for higher-tech goods) to transportation, communications and utilities (4), to wholesale trade (5) then retail (6). The 7's begin with FIRE (Finance, Insurance, and Real Estate) then services in the later 7 and early 8 digits while the 9 is for governments. The second and third digits give more detail: e.g. 377 is for sawmills, 378 for plywood and engineered wood, 379 for prefabricated wood homes. Some data sets might give you 5-digit or even 6-digit information. These classifications date back to the 1930s and 1940s so some parts show their age: the ever-increasing number of computer parts go where plain "office supplies" used to be.

The CPS data distinguishes between "major industries" with 16 categories and "detailed industry" with about 50.

Creating 50 dummy variables could be tiresome so that's where R's "`factor`" data type comes in handy. Just add in a factor into your OLS model and let R take care of the rest. So toss in `A_DTIND` and `A_DTOCC`. So add these lines and fire away,

```
det_ind <- as.factor(A_DTIND)
det_occ <- as.factor(A_DTOCC)
```

In other models such as predictions of sales, the specification might include a time trend (as discussed earlier) plus dummy variables for days of the week or months of the year, to represent the typical sales for, say, "a Monday in June".

Why are we doing all of this? Because I want you to realize all of the choices that go into creating a regression or doing just about anything with data. There are a host of choices available to you. Some choices are rather conventional (for example, the education breakdown I used above) but you need to know the field in order to know what assumptions are common. Sometimes these commonplace assumptions conceal important information. You want to do enough experimentation to understand which of your choices are crucial to your results. Then you can begin to understand how people might analyze the exact same data but come to varying conclusions. If your results contradict someone else's, then you have to figure out what are the important assumptions that create the difference.

```
# cps_1.R
# looking at CPS 2013 data
# uses file from dataferret download of CPS March 2013 supplement,
downloaded June 12 2014
# accompanying lecture notes for KFoster class ECO B2000 in fall 2014 at
CCNY

rm(list = ls(all = TRUE))
setwd("C:\\Users\\Kevin\\Documents\\CCNY\\data for classes\\CPS_Mar2013")
load("cps_mar2013.RData")

attach(dat_CPSMar2013)
# use prime-age,fulltime, yearround workers
use_varb <- (Age >= 25) & (Age <= 55) & work_fullt & work_50wks
dat_use <- subset(dat_CPSMar2013,use_varb) # 47,550 out of 202,634 obs

detach(dat_CPSMar2013)

attach(dat_use) # just prime-age,fulltime, yearround workers

# always a good idea to get basic stats of all of the variables in your
regression to see if they make sense
summary(WSAL_VAL)
summary(Age)
summary(female)
summary(AfAm)
summary(Asian)
summary(Amindian)
summary(race_oth)
summary(Hispanic)
summary(educ_hs)
summary(educ_smcoll)
summary(educ_as)
summary(educ_bach)
summary(educ_adv)
summary(married)
summary(divwidsep)
summary(union_m)
summary(veteran)
summary(immigrant)
summary(immig2gen)

model1 <- lm(WSAL_VAL ~ Age + female + AfAm + Asian + Amindian + race_oth
             + Hispanic + educ_hs + educ_smcoll + educ_as + educ_bach +
educ_adv
             + married + divwidsep + union_m + veteran + immigrant +
immig2gen)

summary(model1)
coeftest(model1)
#sometimes log form is preferred
# dat_noZeroWage <- subset(dat_use,(WSAL_VAL > 0))
# model1a <- lm(log(WSAL_VAL) ~ Age + female + AfAm + Asian + Amindian +
race_oth
#               + Hispanic + educ_hs + educ_smcoll + educ_as + educ_bach +
educ_adv
```

```
#               + married + divwidsep + union_m + veteran + immigrant +
immig2gen, data = dat_noZeroWage)
# detach(dat_use)
# attach(dat_noZeroWage)
# log(mean(WSAL_VAL))
# mean(log(WSAL_VAL))
# detach(dat_noZeroWage)
# attach(dat_use)
# ^^ yes there are more elegant ways to do that, avoiding attach/detach -
find them!

# for heteroskedasticity consistent errors
require(sandwich)
require(lmtest)

coeftest(model1,vcovHC)

# jam nonlinear into linear regression
model2 <- lm(WSAL_VAL ~ Age + I(Age^2) + female + AfAm + Asian + Amindian +
race_oth
             + Hispanic + educ_hs + educ_smcoll + educ_as + educ_bach +
educ_adv
             + married + divwidsep + union_m + veteran + immigrant +
immig2gen)

model3 <- lm(WSAL_VAL ~ Age + I(Age^2) + female + I(female*Age) +
I(female*(Age^2)) + AfAm + Asian + Amindian + race_oth
             + Hispanic + educ_hs + educ_smcoll + educ_as + educ_bach +
educ_adv
             + married + divwidsep + union_m + veteran + immigrant +
immig2gen)
# could do this with "update" function instead
summary(model2)
summary(model3)
# the ANOVA function is flexible - can compare nested models
anova(model1,model2,model3)

# Applied Econometrics in R suggests also spline and kernel estimators, we
might get to that later

# subset in order to plot...
NNobs <- length(WSAL_VAL)
set.seed(12345) # just so you can replicate and get same "random" choices
graph_obs <- (runif(NNobs) < 0.1) # so something like 4000 obs
dat_graph <-subset(dat_use,graph_obs)

plot(WSAL_VAL ~ jitter(Age, factor = 2), pch = 16, col = rgb(0.5, 0.5, 0.5,
alpha = 0.02), data = dat_graph)
# ^^ that looks like crap since Wages are sooooooooo skew!  So try to find
ylim = c(0, ??)
plot(WSAL_VAL ~ jitter(Age, factor = 2), pch = 16, col = rgb(0.5, 0.5, 0.5,
alpha = 0.02), ylim = c(0,150000), data = dat_graph)

# to plot the predicted values might want to do something like,
lines(fitted.values(model2) ~ Age)
# but that will plot ALLLLL the values, which is 4500 too many and looks
awful
```

```
# so back to this,
to_be_predicted2 <- data.frame(Age = 25:55, female = 1, AfAm = 0, Asian = 0,
Amindian = 1, race_oth = 1,
                               Hispanic = 1, educ_hs = 0, educ_smcoll = 0,
educ_as = 0, educ_bach = 1, educ_adv = 0,
                              married = 0, divwidsep =0, union_m = 0,
veteran = 0, immigrant = 0, immig2gen = 1)
to_be_predicted2$yhat <- predict(model2, newdata = to_be_predicted2)

lines(yhat ~ Age, data = to_be_predicted2)

# now compare model3
to_be_predicted3m <- data.frame(Age = 25:55, female = 0, AfAm = 0, Asian =
0, Amindian = 1, race_oth = 1,
                                Hispanic = 1, educ_hs = 0, educ_smcoll = 0,
educ_as = 0, educ_bach = 1, educ_adv = 0,
                               married = 0, divwidsep =0, union_m = 0,
veteran = 0, immigrant = 0, immig2gen = 1)
to_be_predicted3m$yhat <- predict(model3, newdata = to_be_predicted3m)

to_be_predicted3f <- data.frame(Age = 25:55, female = 1, AfAm = 0, Asian =
0, Amindian = 1, race_oth = 1,
                                 Hispanic = 1, educ_hs = 0, educ_smcoll = 0,
educ_as = 0, educ_bach = 1, educ_adv = 0,
                                married = 0, divwidsep =0, union_m = 0,
veteran = 0, immigrant = 0, immig2gen = 1)
to_be_predicted3f$yhat <- predict(model3, newdata = to_be_predicted3f)

plot(WSAL_VAL ~ jitter(Age, factor = 2), pch = 16, col = rgb(0.5, 0.5, 0.5,
alpha = 0.02), ylim = c(0,150000), xlab = "Age", data = dat_graph)
lines(yhat ~ Age, data = to_be_predicted3f)
lines(yhat ~ Age, data = to_be_predicted3m, lty = 2)
legend("topleft", c("male", "female"), lty = c(2,1), bty = "n")

det_ind <- as.factor(A_DTIND)
det_occ <- as.factor(A_DTOCC)

model4 <- lm(WSAL_VAL ~ Age + I(Age^2) + female + AfAm + Asian + Amindian +
race_oth
             + Hispanic + educ_hs + educ_smcoll + educ_as + educ_bach +
educ_adv
             + married + divwidsep + union_m + veteran + immigrant +
immig2gen
             + det_ind + det_occ)
summary(model4)

# and always remember this part...
detach(dat_use)
```