Lecture Notes 7

Econ B2000, Statistics and Introduction to Econometrics

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

Fall 2014

Nonlinear Regression

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

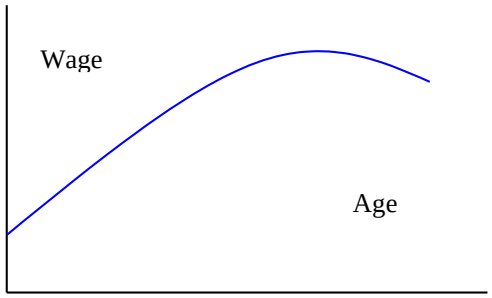
- $X, X^2, X^3, \ldots X^r$
- ln(X), ln(Y), both ln(Y) & ln(X)
- dummy variables
- interactions of dummies
- interactions of dummy/continuous
- interactions of continuous variables

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).

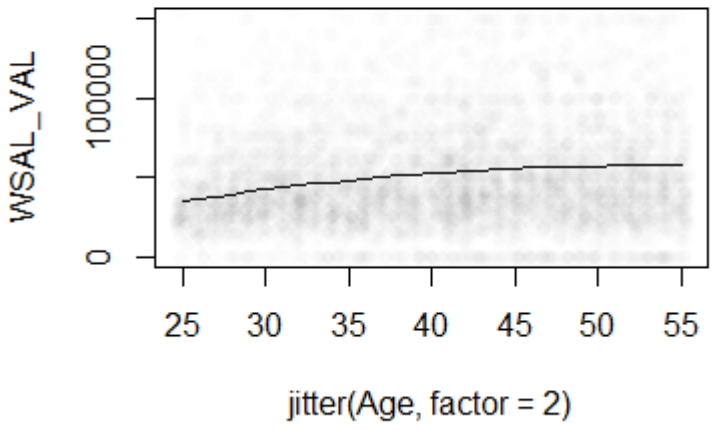
As we remember from calculus,

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

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).

(You also need to figure out how to work with observations where Y=0 since ln(0) doesn't give good results. Dropping those observations might be OK or might not, it depends.)

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.

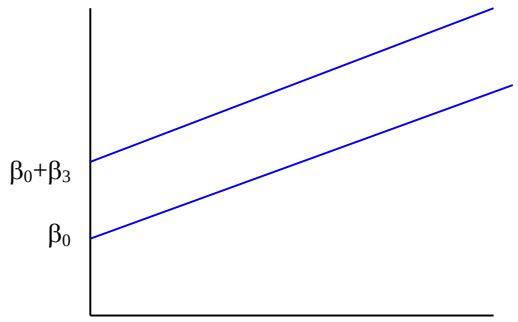
So the choice of whether to specify Y as levels or logs is equivalent to asking whether dummy variables are better specified as having a constant level effect (i.e. women make

$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

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

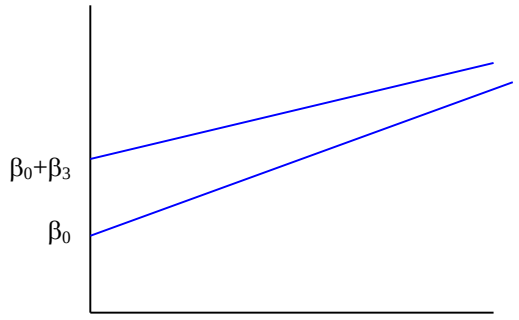
or

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

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,

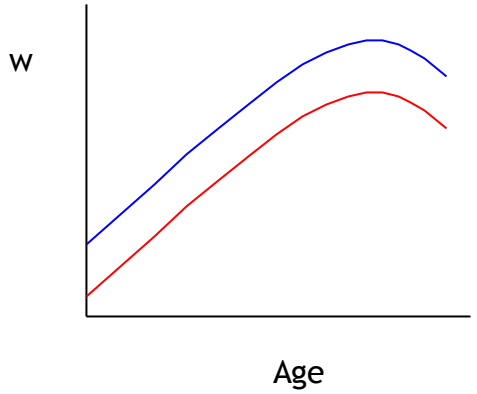


so now the intercepts and slopes are different.

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:

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

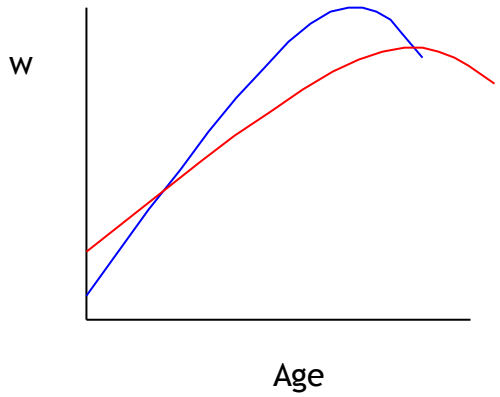
which allows the wage profile lines to have different intercept-values but otherwise to be parallel (the same hump point where wages have their maximum value), as shown by this graph:



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:

$D=0$

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

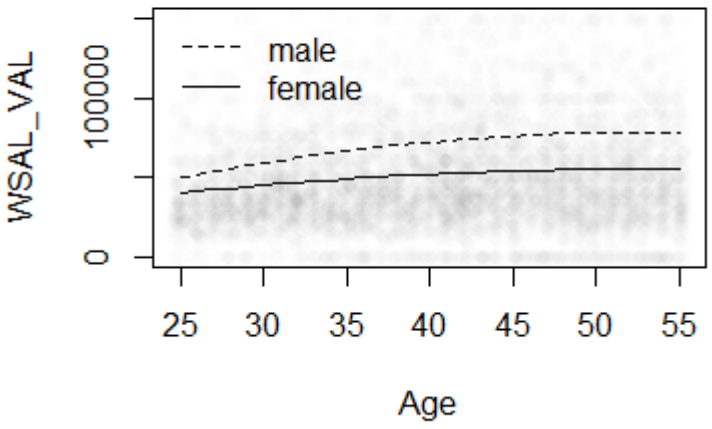
$D=1$

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

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.

Note that it would be very weird (and difficult to justify) to have an interaction of the dummy with the Age term but not with Age-squared or vice versa. Why would we want to assume that, say, men and women have different linear effects but the same squared effect?

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 perhaps similar enough to elasticity formulas to seem vaguely reassuring. 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)` or else `linearHypothesis()` as we did before.

Don't be a dummy about Dummy Variables

It's important to think about the implicit restrictions imposed by the dummy specification – e.g. just putting in a dummy for "high school diploma or above" implicitly assumes that there are two groups, each relatively homogenous. So a regression of wage on just a dummy for high-school diploma assumes that there are two groups: those with a diploma and those without (many of whom have more than a high school degree) – and that each of these groups is relatively homogenous. In many cases the data might be too coarse to estimate fine distinctions: some datasets distinguish between people with a high school diploma and those with a GED while other data lump together those categories. (Many New Yorkers would distinguish which high school!) Every model makes certain assumptions but you want to consider them.

It might be wise to pack the education dummies into a factor and use that factor in R rather than playing around choosing to put in some but not all. This also takes care of automatically dropping one of the dummies (to use it as comparison). Consider these examples:

```
model1wrong <- lm(WSAL_VAL ~ educ_hs + educ_smcoll + educ_as +
educ_bach + educ_adv,  data = dat_use)
summary(model1wrong)
model2wrong <- lm(WSAL_VAL ~ educ_nohs + educ_hs + educ_smcoll +
educ_as + educ_bach + educ_adv,  data = dat_use)
summary(model2wrong)
model3wrong <- lm(WSAL_VAL ~ educ_hs + educ_bach,  data =
dat_use)
```

```
summary(model3wrong)
```

In general it is better to use underlying continuous variables if you have them (e.g. for sports, net points scored rather than win/loss) – this is the basic intuition that there is no need to throw out information. On the other hand this imposes assumptions about linearity which might be inappropriate. For example,

```
model_continuousAge <- lm(WSAL_VAL ~ Age,  data = dat_use)
summary(model_continuousAge)
Age_factr <- cut(dat_use$Age,breaks=25:55)
model_factrAge <-lm(WSAL_VAL ~ Age_factr,  data = dat_use)
summary(model_factrAge)
plot(coef(model_factrAge))
```

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

Often it is sensible 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.

Panel Data

A panel of data contains repeated observations of a single economic unit over time. This might be a survey like the CPS where the same person is surveyed each month to investigate changes in their labor market status. There are medical panels that have given annual exams to the same people for decades. Publicly-traded firms that file their annual reports can provide a panel of data: revenue and sales for many years at many different firms. Sometimes data covers larger blocks such as states in the US or, if we're looking at macroeconomic development, even countries over time.

Other data sets are just cross-sectional, like the March CPS that we've used. If we put together a series of cross-sectional samples that don't follow the same people (so we use the March 2012, 2011, and 2010 CPS samples) then we have a pooled sample. A long stream of data on a single unit is a time series (for example US Industrial Production or the daily returns on a single stock).

In panel data we want to distinguish time from unit effects. Suppose that you are analyzing sales data for a large company's many stores. You want to figure out which stores are well-managed. You know that there are macro trends: some years are good and some are

rough, so you don't want to indiscriminately reward everybody in good years (when they just got lucky) and punish them in bad years (when they got unlucky). There are also location effects: a store with a good location will get more traffic and sell more, regardless. So you might consider subtracting the average sales of a particular location away from current sales, to look at deviations from its usual. After doing this for all of the stores, you could subtract off the average deviation at a particular time, too, to account for year effects (if everyone outperforms their usual sales by 10% then it might just indicate a good economy). You would be left with a store's "unusual" sales – better or worse than what would have been predicted for a given store location in that given year.

A regression takes this even further to use all of our usual "prediction" variables in the list of X, and combine these with time and unit fixed effects.

Now the notation begins. Let the t-subscript index time; let j index the unit. So any observations of y and x must be at a particular date and unit; we have $y_{t,j}$ and then the k x-variables are each $x^k_{t,j}$ (the superscript for which of the x-variables). So the regression equation is

$$y_{t,j} = \alpha_j + \gamma_t + \beta_1 x^1_{t,j} + \beta_2 x^2_{t,j} + \ldots + \beta_{K-1} x^{K-1}_{t,j} + \beta_K x^K_{t,j} + e_{t,j},$$

where α_j (alpha) is the fixed effect for each unit j, γ_t (gamma) is the time effect, and then the error is unique to each unit at each time.

This is actually easy to implement, even though the notation might look formidable. Just create a dummy variable for each time period and another dummy for each unit and put the whole slew of dummies into the regression.

So, to take a tiny example, suppose you have 8 store locations over 10 years, 1999-2008. You have data on sales (Y) and advertising spending (X) and want to look at the relationship between this simple X and Y. So the data look like this:

$X_{1999,1}$	$X_{1999,2}$	$X_{1999,3}$	$X_{1999,4}$	$X_{1999,5}$	$X_{1999,6}$	$X_{1999,7}$	$X_{1999,8}$
$X_{2000,1}$	$X_{2000,2}$	$X_{2000,3}$	$X_{2000,4}$	$X_{2000,5}$	$X_{2000,6}$	$X_{2000,7}$	$X_{2000,8}$
$X_{2001,1}$	$X_{2001,2}$	$X_{2001,3}$	$X_{2001,4}$	$X_{2001,5}$	$X_{2001,6}$	$X_{2001,7}$	$X_{2001,8}$
$X_{2002,1}$	$X_{2002,2}$	$X_{2002,3}$	$X_{2002,4}$	$X_{2002,5}$	$X_{2002,6}$	$X_{2002,7}$	$X_{2002,8}$
$X_{2003,1}$	$X_{2003,2}$	$X_{2003,3}$	$X_{2003,4}$	$X_{2003,5}$	$X_{2003,6}$	$X_{2003,7}$	$X_{2003,8}$
$X_{2004,1}$	$X_{2004,2}$	$X_{2004,3}$	$X_{2004,4}$	$X_{2004,5}$	$X_{2004,6}$	$X_{2004,7}$	$X_{2004,8}$
$X_{2005,1}$	$X_{2005,2}$	$X_{2005,3}$	$X_{2005,4}$	$X_{2005,5}$	$X_{2005,6}$	$X_{2005,7}$	$X_{2005,8}$
$X_{2006,1}$	$X_{2006,2}$	$X_{2006,3}$	$X_{2006,4}$	$X_{2006,5}$	$X_{2006,6}$	$X_{2006,7}$	$X_{2006,8}$
$X_{2007,1}$	$X_{2007,2}$	$X_{2007,3}$	$X_{2007,4}$	$X_{2007,5}$	$X_{2007,6}$	$X_{2007,7}$	$X_{2007,8}$
$X_{2008,1}$	$X_{2008,2}$	$X_{2008,3}$	$X_{2008,4}$	$X_{2008,5}$	$X_{2008,6}$	$X_{2008,7}$	$X_{2008,8}$

and similarly for the Y-variables. To do the regression, create 9 time dummy variables: D2000, D2001, D2002, D2003, D2004, D2005, D2006, D2007, and D2008. Then create 7 unit

dummies, D2, D3, D4, D5, D6, D7, and D8. Then regress the Y on X and these 16 dummy variables.

Then the interpretation of the coefficient on the X variable is the amount by which an increase in X, above its usual value for that unit and above the usual amount for a given year, would increase Y.

One drawback of this type of estimation is that it is not very useful for forecasting, either to try to figure out the sales at some new location or what will be sales overall next year – since we don't know either the new location's fixed effect (the coefficient on D9 or its alpha) or we don't know next year's dummy coefficient (on D2009 or its gamma).

We also cannot put in a variable that varies only on one dimension – for example, we can't add any other information about store location that doesn't vary over time, like its distance from the other stores or other location information. All of that variation is swept up in the firm-level fixed effect. Similarly we can't include macro data that doesn't vary across firm locations like US GDP since all of that variation is collected into the time dummies.

You can get much fancier; there is a whole econometric literature on panel data estimation methods. But simple fixed effects, put into the same OLS regression that we've become accustomed to, can actually get you far.

Multi-Level Modeling

After Fixed Effects, we can generalize to Multi-Level Modeling (much of my explanation is based on the excellent book, *Data Analysis Using Regression and Multilevel/Hierarchical Models*, by Andrew Gelman & Jennifer Hill). From the wage regressions based on CPS data that we were using, we can consider adding information about the person's occupation (the data gives a rough grouping of people into about 20 occupations). You've probably done a version of this regression in your head, if you've ever read someone's job title and tried to figure out how much she makes.

There are a few ways to use the occupation data. One way is to ignore it, to not use it – which is what we were doing when we left it out of the regression. Everyone started from the same value. Gelman & Hill call this the "pooling" estimator since it pools everyone together. Another way would be to put in fixed effects for each occupation, letting each vary as needed – every occupation has a different intercept term, starting from a different value. This is "no-pooling." This puts no constraints at all on what the intercepts might be – some high, some low, some way far afield. A multilevel model imposes a model on how those intercepts vary: usually that they have a normal distribution with a central mean and variance. The math to define the estimator gets a bit more complicated, but we let the computer worry about that. But it's basically a weighted average of the "pooled" and "no-pooled" estimates, where the number of people reporting being in that particular group give the weights. So groups with a lot of members get nearly that "no-pooled" estimate, while a group with few members would be estimated to be like the larger group.

So in this example, the pooling case has wages of person i in industry j explained as $w_{i,j} = \alpha + \beta X_{i,j} + e_{i,j}$ (where the X includes all the rest of the variables, lumped together). The no-pooling case has $w_{i,j} = \alpha_j + \beta X_{i,j} + e_{i,j}$ so the intercept varies by industry, j. The multilevel case has $w_{i,j} = \alpha_0 + \alpha_{[j]} + \beta X_{i,j} + e_{i,j}$ but $\alpha_{[j]} \sim N(0, \sigma_\alpha)$.

With just a single level (like Occupation) this doesn't seem like a big thing, but if we want to define a lot of levels (Occupation, Industry, State or even City) then this gets more important. Instead of estimating a separate parameter for each level, we can estimate just overall parameters – and levels with only a small number of observations will be partially pooled.

Once we decide we want to do such a thing, the remaining question is, "how?" With R it's easy, just `lmer()` instead of `lm()`.

```
modelmm1 <- lmer(WSAL_VAL ~ as.factor(A_DTOCC) + (1 | as.factor(A_DTOCC)), dat_use)
summary(modelmm1)

modelmm2 <- lmer(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
               + as.factor(A_DTOCC) + (1 | as.factor(A_DTOCC)), dat_use)
summary(modelmm2)
```

In these cases we can compute the Intra-Class Correlation (ICC) which is the ratio of the variance in the groups (σ_α) to the total variance, so $\frac{\sigma_\alpha}{\sigma_\alpha + \sigma_\epsilon}$. Kind of like R^2, this goes from zero to one and is graded on a curve. It tells how important the within-group variation is, relative to the total variation.

Of course the next step would be to expand these coefficient estimates to be for slope as well as intercept – something like $w_{i,j} = \alpha_0 + \alpha_{[j]} + \left(\beta_0 + \beta_{[j]}\right) X_{i,j} + e_{i,j}$. Multilevel modeling is a growing trend within econometrics.

```
# 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)
```

```
# cps_2.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)

# create a single index variable (factor) from education dummies
# educ_indx <- as.factor(educ_nohs + 2*educ_hs + 3*educ_smcoll + 4*educ_as +
5*educ_bach + 6*educ_adv)
# levels(educ_indx)[1] <- "No HS"
# levels(educ_indx)[2] <- "HS"
# levels(educ_indx)[3] <- "Some Coll"
# levels(educ_indx)[4] <- "AS"
# levels(educ_indx)[5] <- "Bach"
# levels(educ_indx)[6] <- "Adv Deg"
# levels(educ_indx)

attach(dat_use) # just prime-age,fulltime, yearround workers
# will look at some info by industry so look how wage varies by ind:
by(WSAL_VAL, A_DTOCC, summary)
plot(as.factor(female) ~ A_DTOCC)

detach(dat_use)
# A_DTOCC values:
# 1 'Management occupations'
# 2 'Business and financial operations occupations'
# 3 'Computer and mathematical science occupations'
# 4 'Architecture and engineering occupations'
# 5 'Life, physical, and social service occupations'
# 6 'Community and social service occupations'
# 7 'Legal occupations'
# 8 'Education, training, and library occupations'
# 9 'Arts, design, entertainment, sports, and media occupations'
# 10 'Healthcare practitioner and technical occupations'
# 11 'Healthcare support occupations'
# 12 'Protective service occupations'
# 13 'Food preparation and serving related occupations'
# 14 'Building and grounds cleaning and maintenance occupations'
# 15 'Personal care and service occupations'
# 16 'Sales and related occupations'
# 17 'Office and administrative support occupations'
# 18 'Farming, fishing, and forestry occupations'
# 19 'Construction and extraction occupations'
# 20 'Installation, maintenance, and repair occupations'
# 21 'Production occupations'
# 22 'Transportation and material moving occupations'
# 23 'Armed Forces'



# for heteroskedasticity consistent errors
```

```
require(sandwich)
require(lmtest)


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,  data = dat_use)
summary(model1)
coeftest(model1,vcovHC)

# can do it wrong...
model1wrong <- lm(WSAL_VAL ~ educ_hs + educ_smcoll + educ_as + educ_bach +
educ_adv,  data = dat_use)
summary(model1wrong)
model2wrong <- lm(WSAL_VAL ~ educ_nohs + educ_hs + educ_smcoll + educ_as +
educ_bach + educ_adv,  data = dat_use)
summary(model2wrong)
model3wrong <- lm(WSAL_VAL ~ educ_hs + educ_bach,  data = dat_use)
summary(model3wrong)
# model1 leaves out varbs;
# model2 creates perfect multicollinearity with too many dummies;
# model3 has too few dummies

# example with Age
model_continuousAge <- lm(WSAL_VAL ~ Age,  data = dat_use)
summary(model_continuousAge)
Age_factr <- cut(dat_use$Age,breaks=25:55)
model_factrAge <-lm(WSAL_VAL ~ Age_factr,  data = dat_use)
summary(model_factrAge)
plot(coef(model_factrAge))

model2 <- 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
             + as.factor(A_DTOCC), data = dat_use)
summary(model2)
coeftest(model2,vcovHC)

require(lme4)
# next use multilevel based on industry A_DTOCC
modelmm1 <- lmer(WSAL_VAL ~ as.factor(A_DTOCC) + (1 | as.factor(A_DTOCC)),
dat_use)
summary(modelmm1)

modelmm2 <- lmer(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
              + as.factor(A_DTOCC) + (1 | as.factor(A_DTOCC)), dat_use)
summary(modelmm2)
```