# Post-Week 9 Notes

Two notes following week 9 (our intro to regressions).

**First**, a couple students asked whether you need to center and scale (standardize) data to make it work with linear regression. I think I had a brain freeze and gave an inconsistent answer, so let me clarify here: no, you don't. Even if you're doing multiple regression, and the different right-side variables are on wildly different scales.

To get some intuition for why this is the case, and why wildly different scales won't distort regression results the same way that outliers do, think about a graph again. When we do a linear regression, each variable is its own dimension, and so each coefficient is optimized to minimize the squared residuals *along its dimension*. For example, if you are trying to test whether distance from a university is associated with rental prices, you might run a regression that has, say, number of miles from nearest university and median income of the census tract on the right side and rent on the left side (this is kind of a crappy regression, because you'd want lots of other things that might confound the relationship here, like number of bedrooms, and also because as usual you'd probably want to take a close look at the income variable over its range and make sure it doesn't need to be transformed, but let's run with it for purposes of illustration). Each point in the ultimate regression plane will be located in four-dimensional space: it'll have a value for miles, income, and rent. Let's say rent is on the y axis, income (in thousands/year) is on the x axis, and miles are on the z axis, and represent your regression model as $y = \alpha \cdot x + \beta \cdot z + \epsilon$.

Now let's imagine that we've fit a regression, and we've found coefficients for it, hypothetically, $y = -100 \cdot x + 50 \cdot z + \epsilon$. In other words, for every additional mile away from a university, rent goes down by a hundred bucks, and for every thousand dollars/year median income in the census tract it goes up by fifty bucks. Then imagine that you go back into your dataset and change it to be hundreds of dollars of income/year (i.e., multiply every income by 10). Effectively, what this amounts to is scaling the z axis in the regression, so the plane would stay the same shape, but the points for income would change, they'd all be divided by 10, and you'd end up with the same model with different coefficients. But that would've affect the coefficients along the other dimensions.

Another good way to see this, actually, is just to play with some data yourself. Go and get the data from my rule of law book, for example and do a regression with, say, property rights, political pluralism, and GDP on the right side, note the coefficients and p-values, and then standardize the variables (or just multiply them by something, or add something to them, or do anything else you can dream of). See what kinds of transformations change the t-scores and p-values and what don't. (They'll all change the coefficients.)

For example, run the following code:

```
import pandas as pd
import numpy as np
import statsmodels.formula.api as sm
df = pd.read_csv("http://rulelaw.net/downloads/rol-scores.csv")
df["gdp"] = df["2012GDP"] # just to fix a glitch with columns starting with numbers
model1 = sm.ols(formula='pol_plur ~ assoc_org + free_expr + gdp', data=df).fit()
print(model1.summary())

df["bigorg"] = df["assoc_org"] * 100

model2 = sm.ols(formula='pol_plur ~ bigorg + free_expr + gdp', data=df).fit()
print(model2.summary())

df["shiftbigorg"] = df["bigorg"] + 50000
model3 = sm.ols(formula='pol_plur ~ shiftbigorg + free_expr + gdp', data=df).fit()
print(model3.summary())
```

What you'll notice is that the statistically meaningful stuff, i.e., the t-scores and p-values on the right-side variables, r-squared, etc., don't actually change under these transformations. (If you want to test your intuition further, try to figure out why the coefficient on the intercept and the coefficient on the variable we're messing with, each of which doesn't much matter, each change on exactly one of those...)

That being said, standardizing doesn't hurt, and it can help with interpreting coefficients by having everything be on the same scale for a sorta apples-to-apples comparison. Also, when you're messing around with interaction terms (which I haven't taught you yet, but I will, planning to discuss in part 2 of our regressions material around week 12...) or polynomial regressions there are sometimes meaningful reasons to do it. See more discussion here.

**Second**, with our reorganization/slow-down of the syllabus in response to mid-semester-feedback requests, we won't get to as much of the problems with models material as I was expecting before pset 3 is due. Accordingly, problem 2.3 might be more difficult than I'd intended. So if you'd like, instead of doing 2.3, you can repeat 2.2 with a different (but plausible) method of analysis, like a different regression model or hypothesis test.