David Youngberg
BSAD 210—Montgomery College

LECTURE 18: MULTIVARIABLE REGRESSIONS II

I. Truck Loading data
   a. Open **Data Set 6**; you'll see hypothetical data concerning different shipments.
   b. How long does it take to load a truck? It makes sense to argue that the more pallets that have to be loaded, the longer it'll take. It also makes sense to argue that the heavier the shipment, the more time it takes. A heavier shipment is harder to move around. I thus used both pallets and weight to predict the time, in minutes, to load the truck:

	Coefficients	*ndard Err*	*t Stat*	*P-value*
Intercept	41.1497	11.8028	3.48642	0.00128
Pallets	1.62564	0.87316	1.86179	0.07059
Weight	0.26514	0.29869	0.88766	0.38046

      i. That's odd—neither variable seems to matter. What's going on?

II. Multicollinearity
   a. If one or more pairs of our independent variables are highly correlated, we have multicollinearity, which violates one our regression assumptions.
      i. Multicollinearity doesn't require perfect correlation (if there's perfect correlation, Excel will drop one of the variables). All that's required is that it's "high."[1] For purposes of simplicity, we'll say that if the absolute value of the correlation coefficient is greater than 0.8, there's multicollinearity.
      ii. In this case, we clearly have multicollinearity between pallets and weight, as shipments with more pallets will naturally be heavier shipments. The correlation coefficient between pallets and weight is 0.925.
   b. Multicollinearity is a problem because the regression will try to get two variables to do the same job. It can easily render both variables statistically insignificant because you split the explanatory power of one variable into two and you can't tell which variable's doing the

[1] A more technical way to do this is Variance Inflation Factors (VIFs). This technique is beyond the scope of this course.

work. Is this shipment taking a long time to load because it's heavy, or because there are a lot of pallets to manage?

c. Imagine you're testing cupcake recipes with customers rating different types. Some of your recipes have lots of sugar and butter (type A), some have a moderate amount of each (type B), some have only a little of each (type C).
    i. The type A cupcakes will certainly be most liked, followed by type B, and then by type C.
    ii. But are the type A's liked because of the sugar or because of the butter? How important is each one? Can you lose some sugar and get the same level of enjoyment? You don't know because you have multicollinearity!
    iii. You would need cupcakes with low amounts of sugar but lots of butter and vice versa. You need to reduce the correlation between the two explanatory variables.

d. The easiest way to correct for multicollinearity is to drop one the offending independent variables (it takes two variables to have multicollinearity).
    i. In our shipment example, let's drop weight and only use pallets.

	Coefficients	*tandard Err*	*t Stat*	*P-value*
Intercept	46.5697	10.0726	4.62342	4.3E-05
Pallets	2.34225	0.33174	7.06042	2E-08

    ii. Pallets is now statistically significant.

e. Side note: take a look at the R^2 and adjusted R^2 for each regression.

	Pallets only	*Pallets and Weight*
R^2	0.567	0.576
adjusted R^2	0.556	0.554

    i. While R^2 is higher when you have both variables (no surprise there), adjusted R^2 makes it clear that the second variable didn't help—adjusted R^2 is *lower*.

f. Didn't we have multicollinearity with the housing data because larger houses tend to be located far away? No—we had a sufficient number of observations that didn't follow that pattern to isolate the effect of each variable. The correlation coefficient between size and distance was only 0.258, well away from our 0.8 threshold.