The Least Squares Assumptions for Multiple Regression (SW Section 6.5)

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + u_i, \quad i = 1,\ldots,n$$

1. The conditional distribution of u given the X's has mean zero, that is, $E(u|X_1 = x_1,\ldots, X_k = x_k) = 0$.
2. $(X_{1i},\ldots,X_{ki},Y_i)$, $i = 1,\ldots,n$, are i.i.d.
3. Large outliers are rare: $X_1,\ldots, X_k$, and Y have four moments:

   $E(X_{1i}^4) < \infty$,…, $E(X_{ki}^4) < \infty$, $E(Y_i^4) < \infty$.
4. There is no perfect multicollinearity.

Assumption #1: the conditional mean of *u* given the included *X*'s is zero.

$$E(u|X_1 = x_1, \ldots, X_k = x_k) = 0$$

- This has the same interpretation as in regression with a single regressor.
- If an omitted variable (1) belongs in the equation (so is in *u*) and (2) is correlated with an included *X*, then this condition fails
- Failure of this condition leads to omitted variable bias
- The solution – ***if possible*** – is to include the omitted variable in the regression.

Assumption #2: $(X_{1i},\ldots,X_{ki},Y_i)$, $i = 1,\ldots,n$, are i.i.d.

This is satisfied automatically if the data are collected by simple random sampling.

Assumption #3: large outliers are rare (finite fourth moments)

This is the same assumption as we had before for a single regressor. As in the case of a single regressor, OLS can be sensitive to large outliers, so you need to check your data (scatterplots!) to make sure there are no crazy values (typos or coding errors).

Assumption #4: There is no perfect multicollinearity

Perfect multicollinearity is when one of the regressors is an exact linear function of the other regressors.

Example: Suppose you accidentally include *STR* twice:

```
regress testscr str str, robust
Regression with robust standard errors                Number of obs =     420
                                                      F(  1,   418) =   19.26
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.0512
                                                      Root MSE      =  18.581
------------------------------------------------------------------------------
             |               Robust
    testscr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         str |  -2.279808   .5194892    -4.39   0.000    -3.300945   -1.258671
         str |  (dropped)
       _cons |    698.933   10.36436    67.44   0.000     678.5602    719.3057
------------------------------------------------------------------------------
```

Perfect multicollinearity is when one of the regressors is an exact linear function of the other regressors.

- In the previous regression, β_1 is the effect on *TestScore* of a unit change in *STR*, holding *STR* constant (???)
- We will return to perfect (and imperfect) multicollinearity shortly, with more examples…

With these least squares assumptions in hand, we now can derive the sampling dist'n of $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k$.

The Sampling Distribution of the OLS Estimator (SW Section 6.6)

Under the four Least Squares Assumptions,

- The exact (finite sample) distribution of $\hat{\beta}_1$ has mean β_1, var($\hat{\beta}_1$) is inversely proportional to n; so too for $\hat{\beta}_2$.
- Other than its mean and variance, the exact (finite-n) distribution of $\hat{\beta}_1$ is very complicated; but for large n…
- $\hat{\beta}_1$ is consistent: $\hat{\beta}_1 \xrightarrow{p} \beta_1$ (law of large numbers)
- $\dfrac{\hat{\beta}_1 - E(\hat{\beta}_1)}{\sqrt{\text{var}(\hat{\beta}_1)}}$ is approximately distributed $N(0,1)$ (CLT)
- So too for $\hat{\beta}_2, \ldots, \hat{\beta}_k$

Conceptually, there is nothing new here!

Multicollinearity, Perfect and Imperfect (SW Section 6.7)

Some more examples of perfect multicollinearity

- The example from earlier: you include *STR* twice.
- Second example: regress *TestScore* on a constant, *D*, and *B*, where: $D_i = 1$ if $STR \leq 20$, $= 0$ otherwise; $B_i = 1$ if $STR > 20$, $= 0$ otherwise, so $B_i = 1 - D_i$ and there is perfect multicollinearity
- Would there be perfect multicollinearity if the intercept (constant) were somehow dropped (that is, omitted or suppressed) in this regression?
- This example is a special case of…

The dummy variable trap

Suppose you have a set of multiple binary (dummy) variables, which are mutually exclusive and exhaustive – that is, there are multiple categories and every observation falls in one and only one category (Freshmen, Sophomores, Juniors, Seniors, Other). If you include all these dummy variables *and* a constant, you will have perfect multicollinearity – this is sometimes called ***the dummy variable trap***.

- *Why is there perfect multicollinearity here*?
- *Solutions to the dummy variable trap*:
    1. Omit one of the groups (e.g. Senior), or
    2. Omit the intercept
- *What are the implications of (1) or (2) for the interpretation of the coefficients?*

Perfect multicollinearity, ctd.

- Perfect multicollinearity usually reflects a mistake in the definitions of the regressors, or an oddity in the data
- If you have perfect multicollinearity, your statistical software will let you know – either by crashing or giving an error message or by "dropping" one of the variables arbitrarily
- The solution to perfect multicollinearity is to modify your list of regressors so that you no longer have perfect multicollinearity.

Imperfect multicollinearity

Imperfect and perfect multicollinearity are quite different despite the similarity of the names.

Imperfect multicollinearity occurs when two or more regressors are very highly correlated.

- Why this term? If two regressors are very highly correlated, then their scatterplot will pretty much look like a straight line – they are collinear – but unless the correlation is exactly ±1, that collinearity is imperfect.

Imperfect multicollinearity, ctd.

Imperfect multicollinearity implies that one or more of the regression coefficients will be imprecisely estimated.

- Intuition: the coefficient on X_1 is the effect of X_1 holding X_2 constant; but if X_1 and X_2 are highly correlated, there is very little variation in X_1 once X_2 is held constant – so the data are pretty much uninformative about what happens when X_1 changes but X_2 doesn't, so the variance of the OLS estimator of the coefficient on X_1 will be large.
- Imperfect multicollinearity (correctly) results in large standard errors for one or more of the OLS coefficients.
- The math? See SW, App. 6.2

Next topic: hypothesis tests and confidence intervals…

Question 6.5

$$\widehat{Price} = 119.2 + 0.485BDR + 23.4Bath + 0.156Hsize + 0.002Lsize + 0.090Age - 48.8Poor,$$

$$\bar{R}^2 = 0.72, \quad SER = 41.5, \quad n = 220$$

a) Homeowner converts part of an existing family room into bathroom. (What about converting a bedroom?)
b) Adds a new (100 sq. Ft.) bathroom
c) What is the loss in value of letting the house run-down?
d) Compute R^2