Take Away: Multiple Regression

(Jing Li, Miami University)

1. We wonder whether it makes economic sense to add a bathroom to a house. A contractor can provide a quote of cost of building a new bathroom. The benefit of building a new bathroom can be obtained from regression analysis. An idea is good only when benefit exceeds cost.

2. The starting point is a simple regression

```
> summary(lm(rprice~baths))$coef
             Estimate Std. Error   t value     Pr(>|t|)
(Intercept) 14510.80   4299.715  3.374829 8.297909e-04
baths       29582.67   1745.859 16.944481 2.225878e-46
```

   The concern is that the estimated slope (in red) 29582.67 may be biased as this regression ignores the fact that a house with one more bathroom tends to be larger, and a larger house on average is more expensive. In short, this simple regression mixes together bathroom effect and area effect. The actual causal effect of baths may not be 29582.67

3. Since we have area data, we can control for it (in green) by running a multiple regression

```
> summary(lm(rprice~baths+area))$coef
               Estimate  Std. Error   t value     Pr(>|t|)
(Intercept)  1504.68646 4294.234292 0.3503969 7.262728e-01
baths       18602.52007 2142.532922 8.6824897 2.057660e-16
area           18.36728    2.375508 7.7319372 1.406600e-13
```

   The new estimate of benefit of having one more bathroom is 18602.52007 (in red). This estimate is free of confounding effect of area, according to FW theorem. In other words, this estimate disentangles the direct effect of baths from the indirect effect of area.

4. This new estimate can be interpreted as—increasing bathroom by one is associated with average rprice going up by 18602.52007, after the effect of area has been netted out. We do not use word "cause" because this regression fails to control for factors such as age and school district.

5. The reported t value 8.6824897 (in blue) is concerned with the hypothesis that baths does not matter. We reject it since 8.6824897 exceeds 1.96. The p value (in yellow) being less than 0.05 leads to the same conclusion. We conclude that baths is statistically significant after area has been controlled for.

6. Area is omitted variable (confounder, lurking variable) in simple regression. It is called <u>control variable</u> in the multiple regression.

7. As we add more and more control variables, our estimate of β_1 gets closer and closer to true value. For instance, we can also control for age

```
> summary(lm(rprice~baths+area+age))$coef
              Estimate  Std. Error   t value     Pr(>|t|)
(Intercept)  9959.9862 4484.126479  2.221165 2.704600e-02
baths       13641.5988 2298.852799  5.934090 7.740657e-09
area           21.5844    2.383401  9.056134 1.415320e-17
age          -201.3679   40.770001 -4.939118 1.272098e-06
```

   Now the new estimate of benefit of having one more bathroom is only 13641.5988, but is still statistically significant since t value 5.934090 is still larger than 1.96

8. Now we can test a joint hypothesis that the above regression is useless

```
> summary(lm(rprice~baths+area+age))
Estimate Std. Error t value Pr(>|t|)
(Intercept)  9959.986   4484.126   2.221    0.027 *
baths       13641.599   2298.853   5.934 7.74e-09 ***
area           21.584      2.383   9.056  < 2e-16 ***
age          -201.368     40.770  -4.939 1.27e-06 ***

Residual standard error: 21340 on 317 degrees of freedom
Multiple R-squared:  0.5886,    Adjusted R-squared:  0.5847
F-statistic: 151.2 on 3 and 317 DF,  p-value: < 2.2e-16
```

   The F test (in red) for overall significance is 151.2, with p value being less than 0.05. So we conclude this regression is useful.

9. The R squared 0.5886 (in blue) tells us that about 58 percent of variation of rprice can be explained by baths, area and age. The unexplained part is about 42 percent!