Multiple Regression —Chapters 3 and 4 of Wooldridge's textbook

(Jing Li, Miami University)

Big Picture

In this lecture you will learn

1. Omitted variable bias and spurious regression
2. How multiple regression mitigates those two issues
3. Frisch-Waugh (FW) Theorem, which implies the interpretation of coefficient in multiple regression
4. F test for a joint hypothesis
5. Multicollinearity

Prediction vs Causality

Regression can be used for two purposes. Consider a simple regression

$$y = \beta_0 + \beta_1 x + u$$

1. ISA291 (Applied Regression Analysis) focuses on predicting y, and R^2 is important
2. ECO311 focuses on β_1, which measure the causal (ceteris paribus) effect of x on y. To see this, holding u constant, it follows that

   $$\Delta y = \beta_1 \Delta x \qquad (\texttt{Causal Interpretation}) \tag{1}$$

   In words, β_1 measures the change in y when x changes by one unit, holding constant other factors. Therefore, β_1 measures causal effect (or β_1 has causal interpretation)
3. The main goal of ECO311 is seeking an unbiased estimate of β_1

Simple regression and omitted variable bias (OVB)

1. Suppose the truth is that y depends on both x and w, and other factors

$$y = \beta_0 + \beta_1 x + \beta_2 w + u \qquad (\texttt{Multiple Regression}) \tag{2}$$

2. Usually w is unobserved (no data) so we have to run a simple regression that ignores w

$$y = \beta_0 + \beta_1 x + e \qquad (\texttt{Simple Regression}) \tag{3}$$

where the error term e in simple regression contains w :

$$e \equiv \beta_2 w + u \tag{4}$$

3. It follows that the estimated $\hat{\beta}_1$ in simple regression converges to

$$\hat{\beta}_1 = \frac{S_{x,y}}{S_x^2} \to \frac{cov(x,y)}{var(x)} = \frac{cov(x, \beta_0 + \beta_1 x + \beta_2 w + u)}{var(x)} = \beta_1 + \frac{\beta_2 cov(x,w)}{var(x)} \tag{5}$$

where we apply the law of large number to sample covariance and sample variance, and we use the fact that $cov(x, \beta_0) = 0$, $cov(x,x) = var(x)$, and we assume $cov(x,u) = 0$

OVB

1. Result (5) is important—it shows that in general $\hat{\beta}_1$ from the simple regression does not converge to the true value β_1. Instead we just proved that

$$\hat{\beta}_1 = \beta_1 + OVB \tag{6}$$

   where the omitted variable bias (OVB) is defined as

$$OVB \equiv \frac{\beta_2 cov(x,w)}{var(x)} \tag{7}$$

2. In general OVB is present (nonzero) unless $\beta_2 = 0$ or $cov(x,w) = 0$
3. A variable w is called omitted variable if it satisfies $\beta_2 \neq 0$ and $cov(x,w) \neq 0$
4. The consequence of omitting w is that $\hat{\beta}_1$ is a biased estimate: $\hat{\beta}_1 \neq \beta_1$, or equivalently, β_1 becomes unidentified since OVB involves unknown w

$$\beta_1 = \hat{\beta}_1 - OVB = \texttt{unkown} \tag{8}$$

5. Intuitively, $\hat{\beta}_1$ is a biased estimate of β_1 because it also captures β_2, see (5)

Exogeneity and Endogeneity

1. A regressor is exogenous if it is uncorrelated the error term

$$cov(x, \texttt{error term}) = 0 \Rightarrow x \texttt{ is exogenous} \tag{9}$$

   Otherwise, the regressor is endogenous

$$cov(x, \texttt{error term}) \neq 0 \Rightarrow x \texttt{ is endogenous} \tag{10}$$

2. Bias arises whenever the regressor is endogenous

3. Because of omitted variables, a regressor in simple regression is most likely to be endogenous. After explicitly controlling for (at least some) omitted variables, a regressor in multiple regression is more likely to be exogenous

4. Put differently, estimate from multiple regression is more likely to be unbiased.

Example 1

Let y be number of shark attacks, x be sales of ice cream, and w is temperature

1. we believe $\beta_2 \neq 0$ since higher temperature leads to more people visiting beach
2. we also believe $cov(x,w) \neq 0$ since people buy more ice cream when it gets hotter
3. As a result, temperature is an omitted variable when we run the simple regression

$$\texttt{shark attack} = \beta_0 + \beta_1 \texttt{ice cream} + u \qquad (\texttt{Bad Regression})$$

   where the coefficient of `ice cream` is biased

$$\hat{\beta}_1 \neq \beta_1$$

4. In this case, because $\beta_2 > 0$ and $cov(x,w) > 0$, we have positive OVB and overestimation

$$\hat{\beta}_1 = \beta_1 + positive\ OVB \Rightarrow \hat{\beta}_1 > \beta_1$$

Example 2—Spurious Causality

1. Spurious causality is an extreme case of OVB where x has no causal effect on y, i.e.,

$$\beta_1 = 0.$$

However, due to an omitted variable, $\hat{\beta}_1$ is nonzero:

$$\hat{\beta}_1 = \beta_1 + OVB = 0 + OVB \neq 0$$

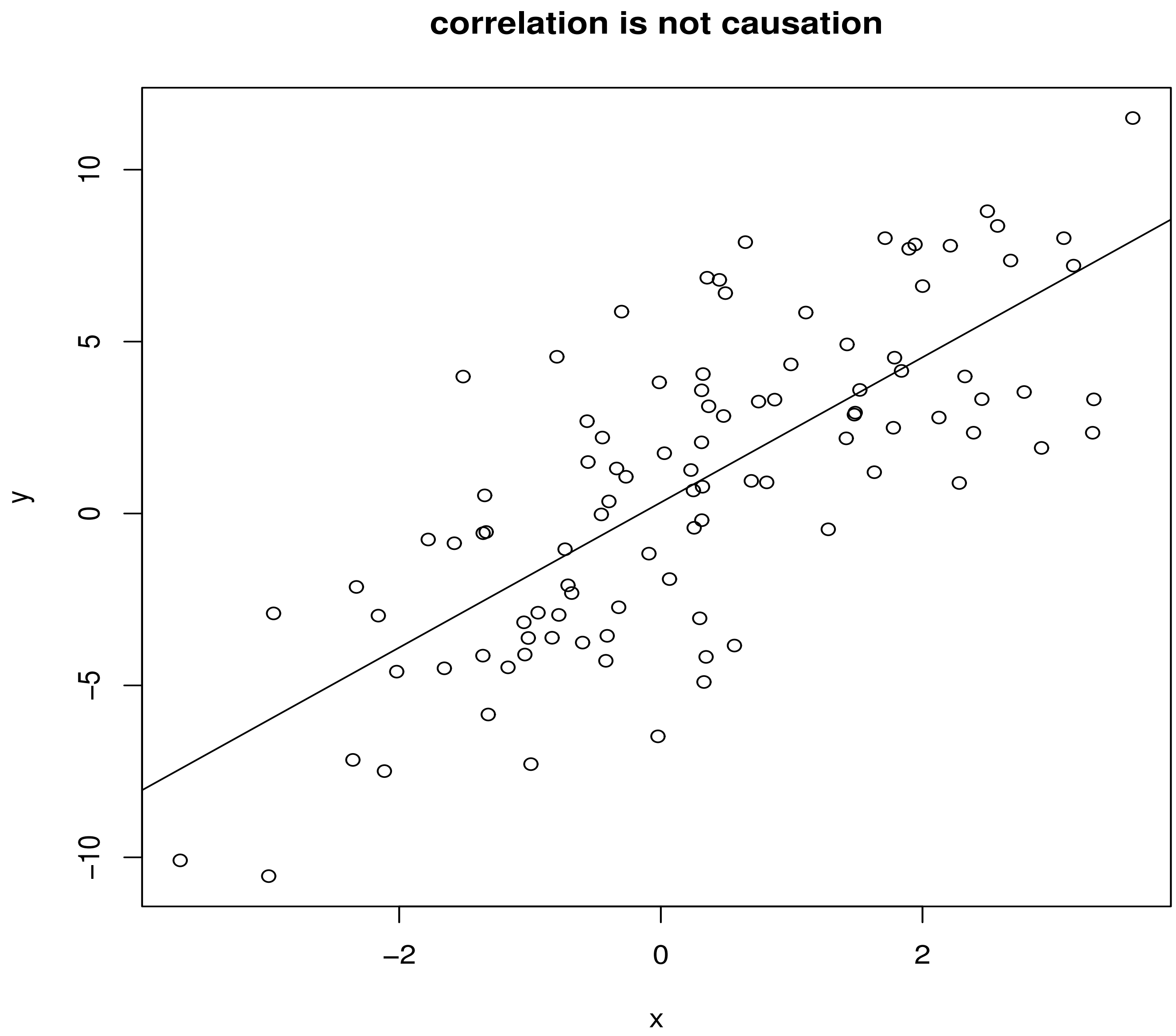
In words, the spurious causality is a bad situation in which because of the omitted variable, a simple regression mistakenly indicates a non-existing causal effect

2. For instance, maybe going to expensive private schools such as Stanford has zero effect on future earning $\beta_1 = 0$. However, super smart kids are more likely to attend Stanford and make big money in the future. In this case OVB is positive (why?), and $\hat{\beta}_1 = 0 + OVB > 0$, which falsely implies that attending Stanford helps make big money

Spurious Causality—Monte Carlo Simulation

```
> set.seed(12345)
> n = 100
> # w is omitted variable. By construction, x does not cause y. w does
> w = rnorm(n)
> x = w + rnorm(n)
> y = 4*w + rnorm(n)
> # Scatterplot just shows correlation, not causation
> plot(y~x, main = "correlation is not causation")
> abline(lm(y~x))
```

Misleading scatter plot (showing correlation other than causation)



Remarks

1. By construction we know x has no causal effect on y, but w has
2. However, the scatter plot and simple regression line seem to indicate that x affects y (or y and x are related)
3. The scatter plot is misleading—it only shows correlation; it does not indicate causality
4. In this example, x and y are positively correlated because both are positively correlated with the omitted variable (lurking variable, confounder) w
5. Lesson: a scatter plot or a simple regression can rarely show causality since it fails to account for other factors
6. Lesson: always asking "***Are there other factors***?"

Example 2—continued

```
> # Simple regression is misleading: it indicates spurious causality
> summary(lm(y~x))$coef
             Estimate Std. Error   t value     Pr(>|t|)
(Intercept) 0.3216882  0.3109865  1.034412 3.034893e-01
x           2.1102790  0.1943846 10.856203 1.694367e-18
> # spurious causality is gone in multiple regression that controls for w
> summary(lm(y~x+w))$coef
               Estimate Std. Error    t value     Pr(>|t|)
(Intercept) -0.03484038 0.09566585 -0.3641883 7.165101e-01
x           -0.11386087 0.09332419 -1.2200575 2.254011e-01
w            4.08848979 0.13237606 30.8854165 5.502261e-52
```

Remarks

1. $\hat{\beta}_1$ in simple regression is 2.110, significantly different from the true value 0 (t statistic is 10.856). We reject the truth of no causality, and this fact is indicative of spurious causality (moreover, we may reject truth due to type-I-error, see optional reading TIE)

2. After we <u>control</u> for w in a multiple regression, the t statistic of $\hat{\beta}_1$ is -1.220. Now we cannot reject the truth that $\beta_1 = 0$. In other words, multiple regression correctly indicates that x does not matter.

3. The coefficient of w is 4.088, close to the true value 4

4. Lesson: multiple regression can <u>mitigate</u>, in this case totally eliminates, spurious causality

5. Lesson: ***Avoid simple regression; Try to control for other factors in a multiple regression***

Example 3

```
> ad = "https://www.fsb.miamioh.edu/lij14/400_house.txt"
> data = read.table(url(ad), header=T)
> attach(data)
> summary(lm(rprice~baths))$coef
            Estimate Std. Error   t value     Pr(>|t|)
(Intercept) 14510.80   4299.715  3.374829 8.297909e-04
baths       29582.67   1745.859 16.944481 2.225878e-46
> summary(lm(rprice~baths+area))$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
> cor(baths, area)
[1] 0.6628154
```

Remarks

1. We use house data to illustrate OVB
2. The coefficient of baths is 29582.67 in the simple regression
3. The omitted variable w can be area. Because area affects house price positively $\beta_2 > 0$, and because area and bathroom number are positively correlated $cov(x, w) > 0$, there is positive OVB. In other words 29582.67 overestimates the true causal effect
4. After we include (control for) area in the multiple regression, the coefficient of baths decreases to 18602.
5. The estimated β_2 is 18.3675, greater than 0; the correlation between area and baths is shown to be positive as expected

In Class Exercise

We still want to quantify how baths affects rprice. Now focus on the omitted variable of age

1. what is the sign of β_2?
2. what is the sign of $cov(x,w)$?
3. what is the sign of OVB?
4. how to use R to verify those guesses?

Frisch-Waugh (FW) Theorem

Consider a multiple regression that involves two regressors

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

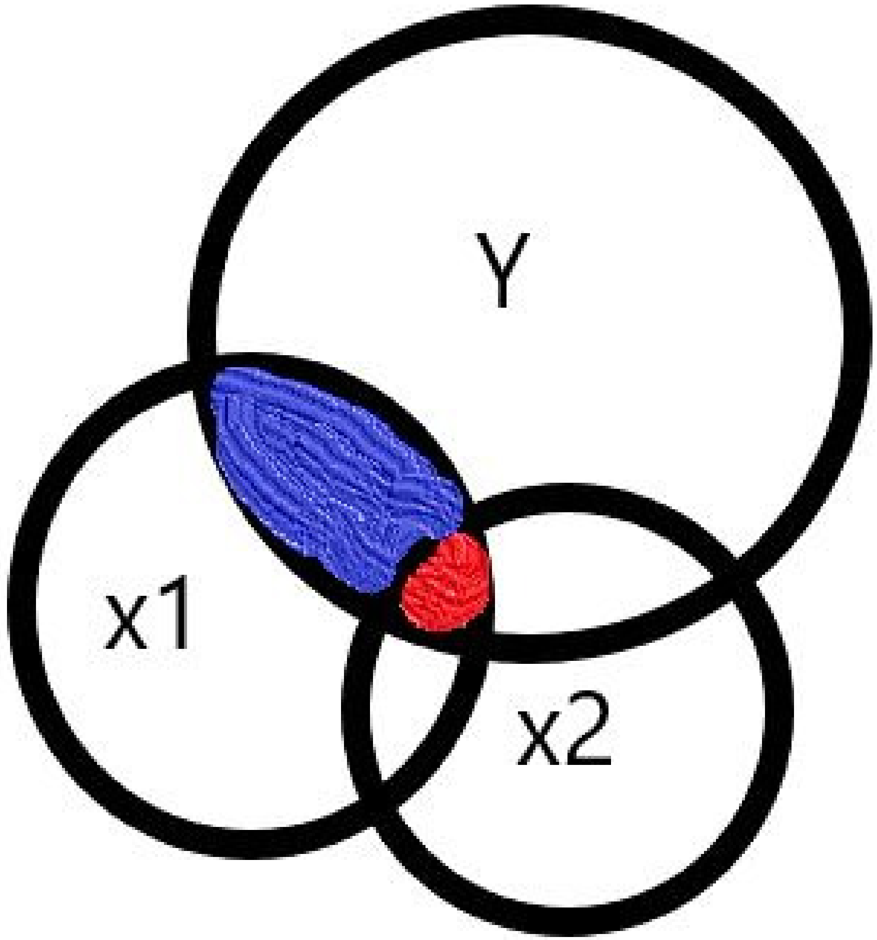
The FW theorem states that $\hat{\beta}_1$ in the multiple regression can be obtained in a two-step procedure:

1. Step I: regress x_1 onto x_2, and save the residual (denoted by $\hat{r}$)

2. Step II: regress y onto $\hat{r}$

Because the residual $\hat{r}$ captures the part of x_1 that <u>cannot</u> be explained by x_2, the coefficient β_1 in multiple regression measures the <u>direct</u> effect of x_1 on y after the effect of x_2 has been <u>netted out</u>. This is how we interpret coefficient in multiple regression
Another common interpretation is, β_1 measures the effect of x_1 on y after holding x_2 constant

Venn diagram for FW Theorem



Remarks

1. We use three circles to represent y, x_1 and x_2
2. A simple regression captures the overlapping part between y and x_1. In the Venn diagram it is the red area plus blue area (looking like a fish or croach)
3. Step I regression removes the red fish head and we are left with blue fish body (part of x_1 that has nothing to do with x_2, the residual $\hat{r}$)
4. Multiple regression only uses the blue fish body $\hat{r}$ to measure the direct effect of x_1 on y. The <u>indirect</u> effect (red fish head) has been netted out by the multiple regression

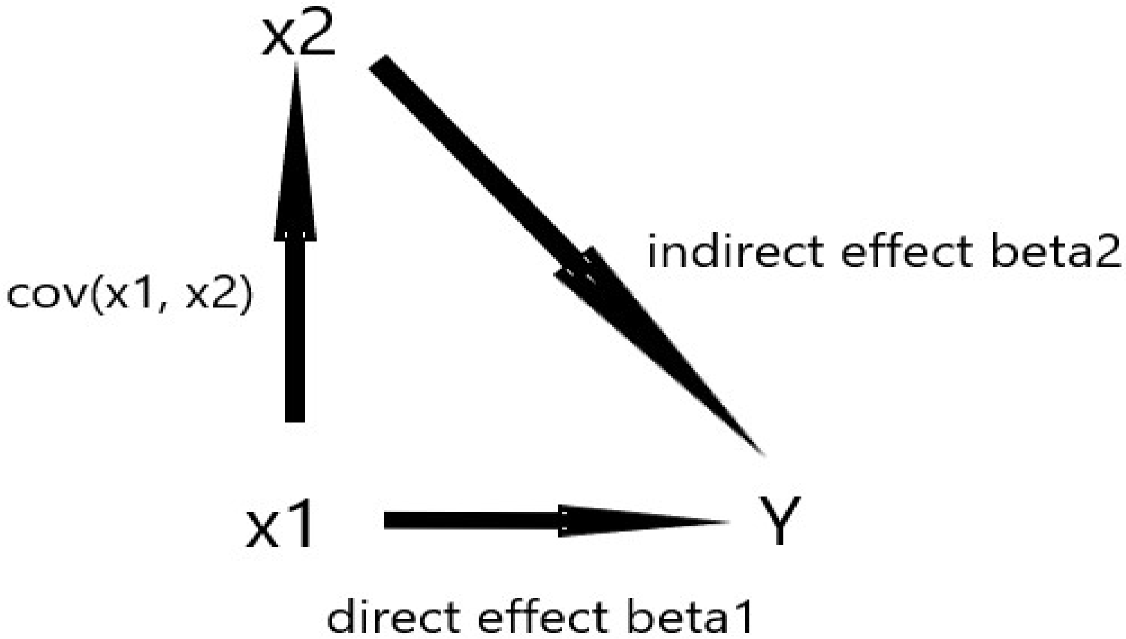
Example 4

```
> summary(lm(rprice~baths+area))$coef
               Estimate  Std. Error   t value     Pr(>|t|)
(Intercept)  1504.68646 4294.234292 0.3503969 7.262728e-01
baths       18602.52007 2142.532922 8.6824897 2.057660e-16
area           18.36728    2.375508 7.7319372 1.406600e-13
> m_first = lm(baths~area)
> rhat = resid(m_first)
> m_second = lm(rprice~rhat)
> summary(m_second)$coef
            Estimate Std. Error  t value      Pr(>|t|)
(Intercept) 83721.36   1751.490 47.80008 1.745125e-147
rhat        18602.52   3040.477  6.11829  2.768572e-09
```

Remarks

1. We use house data to verify FW theorem
2. We let x_1 be baths, x_2 be area. We run the first step regression and use function resid to obtain the residual as rhat
3. In step two we regress rprice onto rhat, and obtain the estimated $\hat{\beta}_1 = 18602.52$, which is the same as the reported coefficient from the multiple regression
4. 18602.52 measures the effect of baths on rprice, after the effect of area has been netted out
5. Alternatively we can say 18602.52 measures the effect of baths on rprice, after holding area constant
6. Exercise: please apply FW theorem, and specify the two-step procedure to obtain $\hat{\beta}_2 = 18.36728$. How to interpret $\hat{\beta}_2 = 18.36728$?

Direct and Indirect Effects



Multiple Regression Estimates Direct Effect

1. Recall the OVB formula

$$\hat{\beta}_1^{simpleregression} \to \beta_1 + \frac{\beta_2 cov(x_1, x_2)}{var(x_1)}$$

2. The estimate of simple regression captures both direct and indirect effects, so is biased
3. By netting out indirect effect, multiple regression only captures direct effect, so is unbiased.

F Test

1. Consider a regression with three regressors

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u \qquad (\texttt{Unrestricted Model}) \tag{11}$$

2. We want to test x_1 and x_2 are jointly insignificant (neither matters)

$$H_0 : \beta_1 = 0, \beta_2 = 0 \tag{12}$$

3. An F test can be used for this kind of joint hypothesis. We need to run a restricted model that imposes the restriction of (12)

$$y = \beta_0 + \beta_3 x_3 + u \qquad (\texttt{Restricted Model}) \tag{13}$$

   and compute the F statistic as

$$F = \frac{(R^2_{\texttt{unrestricted}} - R^2_{\texttt{restricted}})/q}{(1 - R^2_{\texttt{unrestricted}})/(n-k-1)} \tag{14}$$

   where R^2 is R squared, q is the number of restrictions (number of equal signs in H_0), n is sample size, k is the number of regressors in the unrestricted model

Example 5

```
> m = lm(rprice~baths+area+age)
> library("car")
> linearHypothesis(m, c("baths=0", "area=0"))
Linear hypothesis test
Hypothesis:
baths = 0
area = 0

Model 1: restricted model
Model 2: rprice ~ baths + area + age
  Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
1    319 3.1234e+11
2    317 1.4438e+11  2 1.6795e+11 184.37  < 2.2e-16 ***

> m_u = summary(lm(rprice~baths+area+age))
> m_r = summary(lm(rprice~age))
```

```
> f_test = ((m_u$r.squared-m_r$r.squared)/2)/((1-m_u$r.squared)/(321-3-1))
> f_test
[1] 184.3727

> m_u$fstatistic
   value    numdf    dendf
151.2046   3.0000 317.0000
```

Remarks

1. After we run the multiple regression, function **linearHypothesis** in **car** package can be used to obtain the F statistic = 184.37. The null hypothesis is that neither baths nor area matters for rprice

2. The p-value is less than 2.2e-16, so is less than 0.05, so the null hypothesis is rejected, meaning that either baths or area or both matter for rprice

3. To see how to get the F statistic, we then run the restricted regression without baths and area, save the R squared. Finally we compute F statistic explicitly using formula (14). For this problem, $q = 2, k = 3$

4. R by default reports the F test for the null hypothesis that the model is useless or all regressors have zero coefficients (called the test for <u>overall significance</u>). In this example, the F statistic is 151.20 and p-value is less than 0.05. So the model is useful because at least one of the regressor matters.

Multicollinearity

We can use FW theorem to show the variance of $\hat{\beta}_1$ in the multiple regression is

$$var(\hat{\beta}_1) = \frac{\sigma^2}{\sum \hat{r}_i^2} = \frac{\sigma^2}{(n-1)S_{x1}^2(1 - R^2_{Step\ I\ regession})} \tag{15}$$

The last equality is due to facts that (i) $\hat{r}$ is residual and $\sum \hat{r}_i^2$ is RSS of the first step regression; (ii) $R^2 = 1 - RSS/TSS$ implies $RSS = TSS(1 - R^2)$; and (iii) $S^2 = TSS/(n-1)$ implies that $TSS = (n-1)S^2$. In order to <u>increase</u> t statistic (decrease standard error and variance) we can consider

1. increase sample size n
2. increase variation of x_1
3. use other regressors that are weakly correlated with x_1. Multicollinearity arises when other regressors are strongly correlated with x_1 (i.e., $R^2_{Step\ I\ regession}$ is close to one)

Intuitively, blue fish body in Venn diagram is very small if multicollinearity is present. The consequence of multicollinearity is big standard error and small t value.

Example 6

```
> set.seed(12345)
> bedroom = baths + 0.1*rnorm(321)
> summary(lm(rprice~baths+bedroom))$coef
             Estimate Std. Error   t value     Pr(>|t|)
(Intercept) 14349.97   4306.971 3.3318009 0.0009645333
baths       19211.37  13176.703 1.4579803 0.1458328431
bedroom     10395.33  13090.658 0.7941031 0.4277275937
> summary(lm(rprice~baths+bedroom))$fstatistic
   value    numdf    dendf
143.7068   2.0000 318.0000
> summary(lm(baths~bedroom))$r.squared
[1] 0.9824245
```

Remarks

1. To illustrate multicollinearity we generate the number of bedrooms, which by construction is strongly correlated with the number of bathrooms

2. The t values of bedrooms and bathrooms are both less than 1.96, this is the first flag for multicollinearity

3. However, the F value = 143.7068 is significant, that is the second flag for multicollinearity

4. The R squared from the step I regression is 0.9824245, very close to one. That is the cause of multicollinearity

5. Lesson is, because bathroom and bedroom are highly correlated, one of them is <u>redundant</u>. We better just use one of them, not both, as regressor.

Mini Project

Please use GaltonFamilies data in the HistData package.

1. Run a simple regression that predicts childHeight using father's height
2. Do you believe mother's height can be an omitted variable? If so, is OVB positive or negative?
3. Run a multiple regression to verify your guess
4. Interpret the coefficient of father's height in that multiple regression
5. Do you see sign of multicollinearity?