Take Away: Simple Regression

(Jing Li, Miami University)

```
> ad = "https://www.fsb.miamioh.edu/lij14/400_house.txt"
> data = read.table(url(ad), header=T)
> head(data)
  age area baths rprice
1  48 1660     1  60000
2  83 2612     2  40000
...
> attach(data)
> summary(lm(rprice~baths))
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    14511       4300   3.375  0.00083 ***
baths          29583       1746  16.944  < 2e-16 ***
Residual standard error: 24060 on 319 degrees of freedom
Multiple R-squared:  0.4737,    Adjusted R-squared:  0.472

> pvalue = 2*pnorm(-16.944)
> pvalue
[1] 2.131273e-64

> confint(lm(rprice~baths))
               2.5 %   97.5 %
(Intercept)  6051.42 22970.18
baths       26147.82 33017.53
```

1. We run simple regression to show how rprice and baths are related
2. The estimation is obtained by OLS, see eq (6), (11), (12) of slides
3. The estimated intercept is 14511, which is the average rprice for houses with 0 baths, see eq (4)
4. The estimated slope is 29583

(a) So increasing baths by one is associated or correlated with average rprice rising by 29583,see eq (5)

(b) So rprice and baths are positively related, or they move in the same direction

(c) So the OLS fitted line is upward sloping

5. R tests the default null hypothesis that baths does not matter $H_0 : \beta_1 = 0$

   (a) That hypothesis is rejected (so baths matters!) because t value 16.944 is greater than 1.96, see eq(18)

   (b) That hypothesis is rejected because the p value is less than 0.05

   (c) That hypothesis is rejected because 0 is outside the 95 percent confidence interval (26147.82, 33017.53).

   In short, we conclude that the slope coefficient is statistically significant at 5 percent level. The p value 2.131273e-64 is very close to zero. So the evidence is very strong

6. The formula for 95 percent confidence interval is

$$(\hat{\beta}_1 - 1.96 * se,\ \hat{\beta}_1 + 1.96 * se)$$

   where 1.96*se can be approximated by 2*se, called margin of error. The margin of error is due to sampling uncertainty. Bigger se means more uncertainty (estimate being less reliable), resulting in wider confidence interval. To reduce se we can increase sample size or increase variation in x.

7. R squared is 0.4737. So this simple regression explains 47 percent of variation of rprice. Because it is far from one, the predictive power of this regression is low

8. The prediction is computed by the formula below

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

   The prediction error or residual is computed as

$$\hat{u} = y - \hat{y}$$

   For instance, the first house has one bathroom. So its predicted price is 14511 + 29583(1). Its residual is 60000 − 14511 − 29583(1)