Recitation Three Notes

Topic One: Clarification of Standard Error and Standard Deviation.

Popul: $SD(X_i) = \sqrt{Var(X_i)} = \sqrt{E[(X_i - E(X_i))^2]} = \sigma$

Sample: $\widehat{SD}(X_i) = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \frac{1}{n}\sum_{i=1}^{n} X_i)^2} = s$

("estimated standard deviation/ sample standard deviation")

Now, Standard Error is simply the standard deviation of the random variable $\bar{X}$. i.e. it is a measure of the std. dev. of the population of possible estimates.

population:

$$SE(\bar{X}) = \sqrt{Var(\bar{X})} = \sqrt{Var(\frac{1}{n}\sum_{i=1}^{n} X_i)}$$

$$= \sqrt{\frac{1}{n^2} Var(\sum_{i=1}^{n} X_i)} = \sqrt{\frac{1}{n} Var(X_i)} = \sqrt{\frac{1}{n}}\, SD(X_i)$$

sample:

$$\widehat{SE}(\bar{X}) = \sqrt{\frac{1}{n}}\, \widehat{SD}(X_i) = \sqrt{\frac{1}{n}}\, s$$

In short SE is a type of SD. If it helps, think of SE as the measure of the precision of the sample estimate: as the variance of the population of possible estimates decreases, precision increases. $SD(X_i)$ is a measure of the variability of the underlying variable that is being estimated.

Regression Notes

- Accompanies R code

My method of analysis

Step One:

Write out your "true" model

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

Here, Y_i is the yellowness of person i's teeth.

X_i is the number of cigarettes they smoke.

u_i represents all other determinants of what makes people's teeth yellow, could include genes, amount of tea/coffee they drink...

Step Two: Assess your "true" model. Are the OLS assumptions satisfied?

1) is $E[u_i | X_i] = 0$?

- That is, if X_i changes, does $E[u_i | X_i]$ change?
- In the above example, if people smoke more, do we expect them on average (in the population) to drink more coffee, have yellow teeth genes?
- $E[u_i | X_i] \Rightarrow Corr(u_i, X_i)$

often, we just talk about Corr, but really they say different things.

2) (X_i, Y_i) iid

- Randomized experiment is sufficient.
- (X_i, Y_i) should not effect (X_{i+1}, Y_{i+1})

In this example, we should not sample a group of friends as if one smokes a lot, others likely smoke a lot as well.

3) $\mathbb{E}[Y_i^4] < \infty$, $\mathbb{E}[X_i^4] < \infty$

Is it impossible for a value to be ∞? If yes, this is sufficient.

These assumptions guarantee that $\mathbb{E}[\hat{\beta}_1] = \beta_1$ and $\hat{\beta}_1 \xrightarrow{d} N(\beta_1, (SE(\hat{\beta}_1))^2)$

Interpretations of betas:

$\beta_0 = \mathbb{E}[Y_i | X_i = 0]$. In our example: expected yellowness of teeth of those who do not smoke.

$\beta_1 = \Delta \mathbb{E}[Y_i | X_i]$ due to $\Delta X_i = 1$. In our example: increase in expected yellowness of teeth by smoking one more cigarette.

Now that we have established our "true" model, and that $\hat{\beta}_1$ is a "good" estimate of β_1, we turn to that estimation.

To discuss from summary(model):

	coef	se	t	P
(int)	$\hat{\beta}_0$	$SE(\hat{\beta}_0)$	$\frac{\hat{\beta}_0 - 0}{SE(\hat{\beta}_0)}$	$P(\lvert T \rvert > Z_{.975})$
("X")	$\hat{\beta}_1$	$SE(\hat{\beta}_1)$	$\frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$	$P(\lvert T \rvert > Z_{.975})$

How to make confidence intervals?

$$\hat{\beta}_1 \pm SE(\hat{\beta}_1) Z_{1-\alpha/2}$$

We can pull all of these values (besides $Z_{1-\alpha/2}$) from the 'coef test' function in R. The coef test function gives you the table above.

Finally, our final topic is a discussion of R^2:

$$R^2 = \frac{ESS}{TSS}$$

$$= \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{\hat{y}})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

where $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

$$\bar{\hat{y}} = \frac{1}{n} \sum_{i=1}^{n} \hat{y}_i$$

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

$$= \frac{\sum_{i=1}^{n} \left(\hat{\beta}_0 + \hat{\beta}_1 x_i - (\hat{\beta}_0 + \hat{\beta}_1 \bar{x})\right)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \qquad \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

from here, you can transform it into a function of $\hat{\beta}_1$ and variances.

Interpretation: R^2 is a measure of the share of variation in y_i that is explained by the covariates.