STAT 2131
Orly Olbum

Homework 4

1. $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$

$\hat{\beta} = (\hat{\beta_0}, \hat{\beta_1})$, OLS estimates of (β_0, β_1)

(a) Derive distribution for $\hat{\beta}$ (why 2d?)

we know $E(Y_i) = \beta_0 + \beta_1 X_i$ and $Var(Y_i) = \sigma^2$

Find $E(\hat{\beta_1})$ & $var(\hat{\beta_1})$, $E(\hat{\beta_0})$ & $var(\hat{\beta_0})$

we have also shown that

$$\hat{\beta_1} = \frac{\Sigma(X_i - \bar{X})(Y_i - \bar{Y})}{\Sigma(X_i - \bar{X})^2}$$

so $E(\hat{\beta_1}) = \sum \frac{X_i - \bar{X}}{\Sigma(X_i - \bar{X})^2} \cdot (\beta_0 + \beta_1 X_i)$

$$= \frac{\beta_0}{\Sigma(X_i - \bar{X})^2} \sum (X_i - \bar{X}) + \frac{\beta_1}{\Sigma(X_i - \bar{X})^2} \sum (X_i - \bar{X})^2 X_i$$

$$= \frac{\beta_1}{\Sigma(X_i - \bar{X})^2} \sum (X_i - \bar{X})^2 X_i$$

$$= \frac{\beta_1}{\Sigma(X_i - \bar{X})^2} \left[\Sigma X_i^2 - \bar{X} \Sigma X_i\right] = \frac{\beta_1}{\Sigma(X_i - \bar{X})^2} \left[\Sigma X_i^2 - n\bar{X}\right]$$

$$= \frac{\beta_1}{\Sigma(X_i - \bar{X})^2} \sum (X_i - \bar{X})^2 = \beta_1$$

Say $A = \Sigma_{i=1}^n a_i Y_i$ with $E(Y_i) = \mu_i$ and $var(Y_i) = \sigma^2$

$var(\hat{\beta_1}) = \Sigma_{i=1}^n a_i^2 var(Y_i) = \Sigma a_i^2 \sigma^2 = \sigma^2 \Sigma \left(\frac{X_i - \bar{X}}{\Sigma(X_i - \bar{X})^2}\right)^2$

$$= \left[\frac{1}{\Sigma(X_i - \bar{X})^2}\right]^2 \sigma^2 \sum (X_i - \bar{X})^2 = \sigma^2 \left(\frac{1}{\Sigma(X_i - \bar{X})^2}\right)$$

And because of ϵ_i assumptions,

$$\hat{\beta_1} \sim N\left(\beta_1, \sigma^2 \left(\frac{1}{\Sigma(X_i - \bar{X})^2}\right)\right)$$

$\hat{\beta_0} = \bar{Y} - \hat{\beta_1} \bar{X}$

$E(\hat{\beta_0}) = E(\bar{Y} - \hat{\beta_1} \bar{X}) = E(\bar{Y}) - E(\hat{\beta_1} \bar{X}) = \beta_0 + \beta_1 \bar{X} - \beta_1 \bar{X} = \beta_0$

Say $U = \Sigma_{i=1}^n a_i Y_i$ and $V = \Sigma_{i=1}^n d_i Y_i$, a_i & d_i are constants

$var(\hat{\beta_0}) = Var(U) + Var(V) - 2cov(U, V)$

$$= \frac{\sigma^2}{n} + \frac{\bar{X}^2 \sigma^2}{\Sigma(X_i - \bar{X})^2} = \sigma^2 \left[\frac{1}{n} + \frac{\bar{X}^2}{\Sigma(X_i - \bar{X})^2}\right]$$

and also following assumptions,

$$\hat{\beta_0} \sim N\left(\beta_0, \sigma^2 \left[\frac{1}{n} + \frac{\bar{X}^2}{\Sigma_{i=1}^n (X_i - \bar{X})^2}\right]\right)$$

Distribution of $\hat{\beta}$ is two-dimensional because by CLT, joint sampling distribution of $\hat{\beta} = \{\hat{\beta_0}, \hat{\beta_1}\}$ and the expectation (mean) of the estimators is unbiased, relying only on those two estimators.

$\rightarrow$

ⓑ From ⓐ, we find the distribution of $\hat{Y}$ to be as follows:

$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i = \bar{Y} + \hat{\beta}_1 (X_i - \bar{X})$

$E(\hat{Y}_i) = E(\bar{Y}) + (X_i - \bar{X}) E(\hat{\beta}_1) = \beta_0 + \beta_1 \bar{X} + \beta_1 (X_i - \bar{X}) = \beta_0 + \beta_1 X_i$

and $var(\hat{Y}_i) = var(\bar{Y}) + (X_i - \bar{X})^2 var(\hat{\beta}_1) + 2(X_i - \bar{Y}) cov(\bar{Y}, \hat{\beta}_1)$

$= \frac{\sigma^2}{n} + \frac{(X_i - \bar{X})\sigma^2}{\Sigma (X_i - \bar{X})^2} + 0 = \sigma^2 \left[\frac{1}{n} + \frac{(X_i - \bar{X})^2}{\Sigma (X_i - \bar{X})^2} \right]$

and $\hat{Y}_i \sim N\left(\beta_0 + \beta_1 X_i, \sigma^2 \left[\frac{1}{n} + \frac{(X_i - \bar{X})^2}{\Sigma (X_i - \bar{X})^2} \right]\right)$

To find 99% CI for $E(Y_{n+1} | X_{n+1})$ with known σ^2, we use the formula

$$\hat{Y}_h \pm t(1 - \alpha/2; n-2) \, s\{\hat{Y}_h\}$$

where $t(1 - \alpha/2; n-2)$ is t-stat for $\alpha = .01$ and n-2 degrees of freedom and $s\{\hat{Y}_h\}$ is $\sqrt{s^2\{\hat{Y}_h\}}$

Giving us

$$(\hat{\beta}_0 + \hat{\beta}_1 X_{n+1}) \pm t(1 - \alpha/2; n-2) \cdot \sqrt{\sigma^2 \left(\frac{1}{n} + \frac{(X_i - \bar{X})^2}{\Sigma (X_i - \bar{X})^2} \right)}$$

Use z-score since sigma^2 is known

ⓒ If σ^2 is unknown and replaced with $\{\sigma^{2(OLS)}\}$, our interval in ⓑ should not change. We showed in a previous homework that $\{\sigma^{2(OLS)}\}$ is an unbiased estimator for σ^2, so our interval should remain the same.

(d) A 99% CI for Y_{n+1} will be modeled after a prediction interval, since we are looking for the ~~next~~ value. We know σ^2. Our interval will take the form

$$\hat{Y}_{n+1} \pm t(1-\alpha/2; n-2)\, s\{\hat{Y}_{pred}\}$$

where $s^2\{\hat{Y}_{pred}\} = s^2\left[1 + \frac{1}{n} + \frac{(X_{n+1} - \bar{X})^2}{\Sigma(X_i - \bar{X})^2}\right]$

and we know σ^2, so:

$$\hat{Y}_{n+1} \pm t(1-\alpha/2; n-2) \sqrt{\sigma^2\left[1 + \frac{1}{n} + \frac{(X_{n+1} - \bar{X})^2}{\Sigma(X_i - \bar{X})^2}\right]}$$

z-score

This interval will be wider (term) than (b), by the nature of prediction we are less certain for a new value outside of our data (observed | fitted).