Kevin Su
14 Oct 2021

STA 206 Homework 2

1. True or False / Explain

(a) Under the same confidence level, the prediction interval of a new observation is always wider than the confidence interval for the corresponding mean response.

True. Prediction Intervals include potential irreducible error, so one extra unit of estimated variance is added to the confidence interval in producing the prediction interval.

(b) When estimating the mean response corresponding to X_h, the further X_h is from the sample mean $\bar{X}$, the wider the confidence interval for the mean response tends to be.

True. The standard error of the estimated mean response has an $(X_h - \bar{X})^2$ in its numerator. The further from the sample mean X_h is, the greater this term.

(c) If all observations Y_i fall on one straight line, then the coefficient of determination is $R^2 = 1$.

True. The regression sum of squares is equal to the total sum of squares. As $R^2 = SSR/SSTO$, then $R^2 = 1$.

(d) A large R^2 means the fitted regression line is a good fit of the data, while a small R^2 means that the predictor and response are not related.

Not quite true. Can possibly have nonlinear relationship while maintaining high R^2 despite linear fit being unsuitable. A small R^2 can still hold meaningful relation between variables (esp if t-test rejects H_0)

(cont'd)

1. ⓒ The regression sum of squares (SSR) tends to be large if the estimated regression slope is large in magnitude or if the dispersion of the predictor is large.

True. A larger regression slope indicates a larger regression sum of squares (explains more variation in data)

2. Under SLR model:

ⓒ Derive $E[\hat{\beta}_1^2]$

$\longrightarrow \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$

Minimize $\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$ by taking partial derive wrt $\hat{\beta}_1$.

$$\frac{\partial}{\partial \hat{\beta}_1} \sum_{i=1}^{n} (Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i))^2 = 0$$

$$\sum_{i=1}^{n} -2X_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0$$

$\longleftarrow \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$

$$\sum_{i=1}^{n} X_i (Y_i - \bar{Y} + \hat{\beta}_1 (\bar{X}) - \hat{\beta}_1 X_i) = 0$$

$$\sum_{i=1}^{n} \left[X_i (Y_i - \bar{Y}) + \hat{\beta}_1 X_i (\bar{X} - X_i) \right] = 0$$

$$\sum_{i=1}^{n} X_i (Y_i - \bar{Y}) = \sum_{i=1}^{n} \hat{\beta}_1 X_i (X_i - \bar{X})$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_i (Y_i - \bar{Y})}{\sum_{i=1}^{n} X_i (X_i - \bar{X})}$$

2 ② (Cont'd)

$$\hat{\beta}_1^2 = \left(\frac{\sum_{i=1}^{n} x_i (Y_i - \bar{Y})}{\sum_{i=1}^{n} x_i (x_i - \bar{x})} \right)^2$$

$$= \frac{\left(\sum_{i=1}^{n} (x_i Y_i - x_i \bar{Y}) \right)^2}{\left(\sum_{i=1}^{n} (x_i^2 - x_i \bar{x}) \right)^2}$$

$$= \frac{\left(\sum_{i=1}^{n} x_i Y_i - \sum_{i=1}^{n} x_i \bar{Y} \right)^2}{\left(\sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i \bar{x} \right)^2}$$

$$E[\hat{\beta}_1^2] = \frac{\left(nE[XY] - n\bar{X}\bar{Y} \right)^2}{\left(nE[X^2] - n\bar{X}^2 \right)^2}$$

$$= \frac{\left(E[XY] - E[X]E[Y] \right)^2}{\left(E[X^2] - (E[X])^2 \right)^2}$$

$$E[\hat{\beta}_1^2] = \left(\frac{Cov(X,Y)}{Var(X)} \right)^2$$

2. (b) (cont'd) Show that the regression sum of squares

$$SSR = \hat{\beta}_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2$$

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

$$\Rightarrow \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = \bar{y} + \hat{\beta}_1 (x_i - \bar{x})$$

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

$$= \sum_{i=1}^{n} \left(\hat{\beta}_1 (x_i - \bar{x}) \right)^2$$

$$\boxed{SSR = \hat{\beta}_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

(c)

$$E[SSR] = E\left[\hat{\beta}_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 \right]$$

$$= \left[\frac{(Cov(X,Y))^2}{(Var(X))^2} \cdot n Var(X) \right]$$

$$= n \frac{(E[XY] - E[X]E[Y])^2}{E[X^2] - E[X]^2}$$

3. Under the simple linear regression model, show that the residuals e_i's are uncorrelated with the LS estimators $\hat{\beta}_0$ and $\hat{\beta}_1$, i.e.

$$\text{Cov}(e_i, \hat{\beta}_0) = 0 \qquad \text{Cov}(e_i, \hat{\beta}_1) = 0$$

for $i = 1, 2, \cdots, n$.

$$\text{Cov}(e_i, \hat{\beta}_0) = E[(e_i - \bar{e})(\hat{\beta}_0 - E[\hat{\beta}_0])] \overset{?}{=} 0$$

$$= E[e_i \hat{\beta}_0] - E[e_i] E[\hat{\beta}_0]$$

$$= E[e_i \hat{\beta}_0] - 0$$

$$= \hat{\beta}_0 E[e_i] \qquad = 0$$

Similarly,

$$\text{Cov}(e_i, \hat{\beta}_1) = E[(e_i - \bar{e})(\hat{\beta}_1 - E[\hat{\beta}_1])]$$

$$= E[e_i \hat{\beta}_1] - E[e_i] E[\hat{\beta}_1]$$

$$= E[e_i \hat{\beta}_1] \qquad 0$$

$$= \sum_{i=1}^{n} e_i \frac{\sum_{i=1}^{n} x_i (Y_i - \bar{Y})}{\sum (x_i - \bar{x})^2}$$

$$0 \qquad = 0$$

4. Under the Normal Error Model: Show SSE is independent with LS estimators $\hat{\beta}_0$ and $\hat{\beta}_1$.

Hint: (Use the fact that, if 2 sets of random variables, say $(Z_1, \cdots Z_s)$ and $(W_1, \cdots, W_t)$, are independent with each other, then their functions, say $f(Z_1, \cdots, Z_s)$ and $g(W_1, \cdots, W_t)$, are independent.

ε_i's are iid $N(0, \sigma^2)$ r.v.

$\hat{\beta}_0 \sim N(\beta_0, \sigma^2\{\hat{\beta}_0\})$

$\hat{\beta}_1 \sim N(\beta_1, \sigma^2\{\hat{\beta}_1\})$

$$SSE := \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

Y_i's come from linear combinations of Normal distributions.

5. Using SLR model, derive $Var(\hat{Y}_h)$ where

$$\hat{Y}_h = \hat{\beta}_0 + \hat{\beta}_1 X_h$$

is the estimator of the mean response $\beta_0 + \beta_1 X_h$

$$Var(\hat{Y}_h) = Var\left[\hat{\beta}_0 + \hat{\beta}_1 X_h\right]$$

$$= Var\left[\bar{Y} + \hat{\beta}_1 (X_h - \bar{X})\right]$$

$$= Var\left[\hat{\beta}_1 (X_h - \bar{X})\right] + Var\left[\bar{Y}\right]$$

$$= (X_h - \bar{X})^2 Var(\hat{\beta}_1) + Var\left[\bar{Y}\right]$$

$$= \sigma^2 \left[\frac{(X_h - \bar{X})^2}{\sum_i^n (X_i - \bar{X})^2}\right] + Var\left[\bar{Y}\right]$$

$$\rightarrow Var(\bar{Y}) = \frac{\sigma^2}{n}$$

$$Var(\hat{Y}_h) = \sigma^2 \left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\right]$$

6. $\sum_{i=1}^{84} X_i = 6602$ $\quad \bar{X} = 78.59524$

$\sum Y_i = 597341$ $\quad \bar{Y} = 7111.20238$

$\sum X_i^2 = 522048$

$\sum Y_i^2 = 4796548849$

$\sum X_i Y_i = 46400230$

(a) An increase in HS graduation rates seems to be associated with a decrease in crime rates. Data seems to be concentrated in the ~75–85% graduate range.

(b) $$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{Cov(XY)}{Var(X)}$$

$$= \frac{E[XY] - E[X]E[Y]}{E[X^2] - E[X]^2}$$

$$= \frac{552383.69 - (78.59524)(7111.20238)}{6215.4524 - 6177.21176}$$

$$\hat{\beta}_1 = -170.577$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = 7111.20238 + (170.577)(78.59524)$$

$$\hat{\beta}_0 = 20517.73$$

$$\hat{Y}_h = \hat{\beta}_0 + \hat{\beta}_1 X_h = 20517.73 - 170.577X$$

Every percent increase in HS graduation rate is associated with a $\frac{170.577}{100k \text{ people}}$ decrease in crime rate.

6. ⓒ Calculate SSE and MSE. df(SSE?)

$df(SSE) = n-2 = 82$

$$SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = SST - SSR$$

$$= \sum_{i=1}^{n}(Y_i - \bar{Y})^2 - \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$$

$$= n(E[Y^2] - E[Y]^2) - \sum_{i=1}^{n}\left(\bar{Y} + \hat{\beta}_1(X_i - \bar{X}) - \bar{Y}\right)^2$$

$$= n(E[Y^2] - E[Y]^2) - \sum_{i=1}^{n}\hat{\beta}_1^2(X_i - \bar{X})^2$$

$$= 84(E[Y^2] - E[Y]^2) - 29096.51(84)(E[X^2] - E[X]^2)$$

$$= 548736108.7 - 93464186.2$$

$SSE = 455271922.5$

$$MSE = \frac{SSE}{n-2} = 5552096.62$$

6 (d)

$$s\{\hat{\beta}_0\} = \sqrt{MSE\left[\frac{1}{n} + \frac{\bar{X}^2}{\sum(X_i - \bar{X})^2}\right]}$$

$$s\{\hat{\beta}_1\} = \sqrt{MSE\left[\frac{1}{\sum(X_i - \bar{X})^2}\right]}$$

$$\Rightarrow MSE = 5552096.62$$

$$\frac{1}{\sum(X_i - \bar{X})^2} = \frac{1}{n(E[X^2] - E[X]^2)} = \frac{1}{84(38.24093)}$$

$$s\{\hat{\beta}_1\} = \sqrt{\frac{MSE}{84(38.24093)}} = 41.5743$$

$$s\{\hat{\beta}_0\} = \sqrt{MSE\left[\frac{1}{84} + \frac{\bar{X}^2}{84(38.24093)}\right]} = 3277.638$$

6 (c) $H_0: \beta_1 = 0$
$H_1: \beta_1 \neq 0$

$$T^* = \frac{\hat{\beta}_1 - 0}{s\{\hat{\beta}_1\}}$$

Under H_0
Reject if $|T^*| > t(.995, df = 82)$
if $|T^*| > 2.6371$

$$= \frac{-170.577}{41.5743}$$

$$T^* = -4.08907$$

As $|T^*| \approx 4.08 > 2.6371$, we reject the null at the .01 significance level. There is likely a linear association between crime rate and percentage of high school graduates.

(d) $\hat{\beta}_0$ is an unbiased estimator for β_0.

$df(\hat{\beta}_0) = 1$

The 99% confidence interval for β_0 is

$$\hat{\beta}_0 \pm t(.995, 82)\, s\{\hat{\beta}_0\}$$

$$20517.73 \pm 3277.638(2.6371)$$

$$20517.73 \pm (8643.46)$$

We are 99% confident the intercept for the estimator lies between 20517.73 ± 8643.46

6. (a) Construct a 95% CI for mean crime rate for counties with 85% HS graduates

$$\hat{Y}_h = \bar{Y} + \hat{\beta}_1(X_h - \bar{X})$$

$$= 7111.20238 - 170.577(X_h - 78.59524)$$

$$\hat{Y}_{85} = 20517.74 - 170.577(85)$$

$$\hat{Y}_{85} = 6018.6976$$

$$s\{\hat{Y}_{85}\} = \sqrt{MSE\left[\frac{1}{84} + \frac{(85-\bar{X})^2}{\sum(X_i-\bar{X})^2}\right]}$$

$$= 2453.939$$

95% CI of $E[\hat{Y}_{85}]$ is

$$6018.6976 \pm t(.975, 82)\, 2453.939$$

$$6018.6976 \pm (1.9893)\, 2453.939$$

$$6018.6976 \pm 4881.62$$

We are 95% confident the mean crime rate for counties with 85% HS graduation rate lies within 6018.6976 ± 4881.62 per 100K residents.

6. (h)

$$s\{Pred_{Y_{BS}}\} = \sqrt{MSE\left[\frac{8s}{84} + \frac{(8s-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right]}$$

The predicted crime rate of county A is

$$\hat{Y}_{BS} = 6018.6976$$

$$s\{Pred_{Y_{BS}}\} = 3402.045$$

95% Prediction interval

$$6018.6976 \pm 1.9893\,(3402.045)$$

$$6018.6976 \pm 6767.69$$

We are 95% confident the crime rate of county A will lie within 6018.6976 ± 6767.69.

The prediction interval is always wider than the confidence interval for the mean.

6. (a) The only assumptions used are the normality assumptions for errors/residuals. It may perhaps be useful to consider that a negative crime rate may be nonsense.

7. Perform ANOVA on the "Crime Rate and Education" data.

(a) Calculate sum of squares:

$SST = n\,Var(Y) = n(E[Y^2] - E[Y]^2) = (6532572.7)(84)$

$SSE = 455271922$ (From Problem 6)

$SSR = 93464185.6 = SST - SSE$

$df(SSR) = 1$ $\quad SST = 548736107$

$df(SSE) = 82$ $\quad SSE = 455271922$

$df(SST) = 83$ $\quad SSR = 93464185$

(b)

$MSE = SSE/n-2 = 5552096.61$

$MSR = SSR/df(SSR) = SSR/1 = 93464185$

(c)

	Sum of Squares	df	Mean Squares	F^*
Regression	93464185	1	93464185	16.834
Error	455271922	82	5552096.6	
Total	548736107	83		

7. (d) $H_0: \beta_1 = 0$
$H_1: \beta_1 \neq 0$

$$F^* = \frac{MSR}{MSE} = 16.834$$

Null distribution: $F^* \underset{H_0}{\sim} F_{df=1,82}$

Decision Rule: Reject H_0 if $F^* > F_{(.99, df=1,82)} = 6.95442$

As $F^* = 16.834 > 6.95442$, we reject H_0 and there likely exists a linear association between crime rate and percentage of HS graduates.

(e) The F^* statistic is approximately (due to rounding errors) equal to the T^* statistic from Problem 6.

(f) $$R^2 = \frac{SSR}{SST} = 0.170326$$

(g) $$r = \frac{Cov(XY)}{S_X S_Y} = \hat{\beta}_1 \frac{S_X}{S_Y} = -170.577 \left(\frac{2571.24}{6.22106}\right) = -.4127$$

$$S_Y = \sqrt{\frac{n}{n-1} 6532572.71} = 2571.24$$

$$S_X = \sqrt{\frac{n}{n-1} 38.24093} = 6.22106$$

$\Rightarrow r^2 = .170328 \approx R^2$. r^2 is equal to R^2 under the simple linear regression model with one predictor.