

# Homework Solution (#6)

Additional: Analysis of Urban Mortality data

## Chapter 12. Nonlinear and Multiple Regression

1.

$$a). \rho(\text{Mortality, Shs}) = -\sqrt{\frac{46358}{228389}} = -.4505$$

$$\rho(\text{Shs, Resid. Mort//SH}) = 0$$

b). To create best possible multiple regression prediction equation with Shs and one more predictor variable, we will choose %WhiteCollar since it has the biggest absolute correlation with Residual.

2.

$$\rho(\text{Shs, \%WhiteCollar}) = .1659$$

-- It is making sense since as the sound housing in an area goes up, so should the price of the house and the proportion of the white-collar;

$$\rho(\text{Shs, p<3}) = -.4386$$

-- It is making sense since as sound housing goes up, the proportion of less income people will go down;

$$\rho(\text{Shs, pop/house}) = -.4228$$

-- It is making sense since as sound housing goes up, the income of family will go up and less people have to live in each room.

3.

$$a). \text{Rsquare} = 55118/228398 = .241$$

$$b). \text{Root Mean Square Error} = \sqrt{3040} = 55.14$$

$$c). \text{Mean of Response} = 940.3 \text{ (from JMP summary on page 8)}$$

$$d). \text{Shs -- t Ratio} = -4.445/1.576 = -2.82$$

$$e). \text{Shs -- Prob>|t|} = P(|t_{57}| > 2.82) = .006$$

$$f). \text{pop/house -- Std Error} = 74.12/1.697 = 43.68$$

$$g). \text{pop/house -- t Ratio} = \sqrt{2.88} = 1.697$$

$$h). \text{pop/house -- Prob>|t|} = P(|t_{57}| > 1.697) = .0951$$

$$i). \text{Shs -- F Ratio} = 24194/3040 = 7.96$$

$$j). \text{Shs -- Prob > F} = P(F_{1,57} > 7.96) = .006$$

$$k). \text{pop/house -- Sum of Squares} = 55118 - 46358 = 8760$$

$$l). \text{pop/house -- F Ratio} = 8760/3040 = 2.88$$

$$m). \text{pop/house -- Prob>F} = P(F_{1,57} > 2.88) = .0951$$

$$n). \text{Model -- DF} = 2$$

$$o). \text{Model -- Sum of Squares} = 228398 - 173280 = 55118$$

$$p). \text{Model -- Mean Square} = 55118/2 = 27559$$

- q). Model – F Ratio =  $27559/3040 = 9.07$   
 r). Error – DF = 57  
 s). Error – Mean Square =  $173280/57 = 3040$   
 t). C total – DF = 59  
 u). C total – Sum of Squares = 228398 (from Summary 1)

4.

Model:  $y = \beta_0 + \beta_1 \text{Shs} + \beta_2 \text{Pop/House} + \varepsilon$

e).  $H_0: \beta_1 = 0$  v.s.  $H_a: \beta_1 \neq 0$

p-value =  $.006 < .05$ , so reject  $H_0$ .

h).  $H_0: \beta_2 = 0$  v.s.  $H_a: \beta_2 \neq 0$

p-value =  $.0951 > .05$ , so don't reject  $H_0$ .

j).  $H_0: \beta_1 = 0$  v.s.  $H_a: \beta_1 \neq 0$

p-value =  $.006 < .05$ , so reject  $H_0$ .

m).  $H_0: \beta_2 = 0$  v.s.  $H_a: \beta_2 \neq 0$

p-value =  $.0951 > .05$ , so don't reject  $H_0$ .

q).  $H_0: \beta_1 = \beta_2 = 0$  v.s.  $H_a: \beta_1 \neq 0$  or  $\beta_2 \neq 0$

p-value =  $.0004 < .05$ , so reject  $H_0$ .

5.

i). Mortality =  $1058 - 4.445 * 78 + 74.12 * 3.1 = 941.1$

ii). Residual =  $1000 - 941.1 = 58.9$

iii). Suppose  $\varepsilon \sim N(0, \sigma^2)$ , then  $\hat{\sigma}^2 = 55.14$ , so  $P(\varepsilon_i > 58.9) = .1450$

iv). 90% prediction interval =  $(941.1 - t_{.05, 57} * 56.34, 941.1 + t_{.05, 57} * 56.34)$   
 =  $(846.9, 1035.3)$  (Since  $SD(\hat{y}) = 56.34$ .)

6.

99% confidence interval for  $\beta_1 = -4.445 \pm t_{.005, 57} * 1.576 = (-8.67, -.224)$

7.

$\beta_2 = 74.12$ , so it is correct for the claim of “If the value of SHs in a community is held constant then decreasing the pop/house by 0.5 people will reduce the mortality rate by roughly 35 per 100,000. Consequently, reducing the pop/house while holding Shs constant increases life expectancy.”

8.

From Residual plot, point B has small residual while point A and C have large residuals. From leverage plots, point A and C should be labeled as high leverage points. High leverage point means that if the point is excluded, the estimated coefficient will change a lot.

9.

If A is deleted from the data set,  $\hat{\beta}_1$  will be bigger, tends to zero. While  $\hat{\beta}_2$  will be smaller, tends to zero also.

10.

If B is deleted from the data set,  $\hat{\beta}_1$  and  $\hat{\beta}_2$  won't be changed much.

11.

It seems that Log(HC) will be more helpful as an additional variable. Looking at plot Mortality by HC, D is a high influential point! A small change of D will change the coefficient's estimate a lot. While looking at Mortality by LogHC, the data are distributed more evenly and the estimates of coefficients will be more robust and plausible.

12.

With the additional multiple regression analysis, we would like to choose SHs, % white collar, and Log(HC) as predict variables. By this way, we are having most effective, parsimonious prediction model for mortality.