Biostat Midterm - 2017 - Answer Key

653

Below find results of the NMES data analysis. There are three regressions of $\log_e$(total medical expenditures + 1) on the indicator of whether the person has a major smoking caused diease (e.g. lung cancer, cardiovascular disease,…) (*mscd*=1-yes; 0 - no), the persons age in years (*lastage*), and whether sex is male (*male*=1) or female (*male*=0); male:mscd indicates an interaction between these two variables. Log transform was used to make the response variable more nearly Gaussian

Model A: lm(formula = **lte ~ mscd**)

```
              Estimate Std. Error  t value   Pr(>|t|)
(Intercept)   5.67474    0.02295   247.23   <2e-16 ***
mscd          2.34343    0.06689    35.03   <2e-16 ***
Residual standard error: 2.519 on 13646 degrees of freedom
Multiple R-squared:  0.08252, Adjusted R-squared:  0.08245
F-statistic:  1227 on 1 and 13646 DF,  p-value: < 2.2e-16
```

Model B: lm(formula = **lte ~ ns(lastage, 3) + male + mscd**)

```
                  Estimate Std. Error  t value   Pr(>|t|)
(Intercept)        5.20666    0.06540   79.62   <2e-16 ***
ns(lastage, 3)1    1.17810    0.09755   12.08   <2e-16 ***
ns(lastage, 3)2    1.89054    0.17374   10.88   <2e-16 ***
ns(lastage, 3)3    1.59383    0.13505   11.80   <2e-16 ***
male              -0.53132    0.04258  -12.48   <2e-16 ***
mscd               1.99756    0.06741   29.63   <2e-16 ***
Residual standard error: 2.454 on 13642 degrees of freedom
Multiple R-squared:  0.1291,  Adjusted R-squared:  0.1288
F-statistic: 404.6 on 5 and 13642 DF,  p-value: < 2.2e-16
```

Model C: lm(formula = **lte ~ ns(lastage, 3) + male + mscd + male:mscd**)

```
                  Estimate    Std. Error t value    Pr(>|t|)
(Intercept)        5.24695    0.06577   79.774   < 2e-16 ***
ns(lastage, 3)1    1.16425    0.09749   11.942   < 2e-16 ***
ns(lastage, 3)2    1.89056    0.17357   10.892   < 2e-16 ***
ns(lastage, 3)3    1.61531    0.13497   11.968   < 2e-16 ***
male              -0.61436    0.04533  -13.552   < 2e-16 ***
mscd               1.66245    0.09236   18.000   < 2e-16 ***
male:mscd          0.69207    0.13053    5.302 1.16e-07 ***
Residual standard error: 2.452 on 13641 degrees of freedom
Multiple R-squared:  0.1309,  Adjusted R-squared:  0.1305
F-statistic: 342.5 on 6 and 13641 DF,  p-value: < 2.2e-16
```

Stem + Leaf of Class Scores (out/65)

```
6↓ |
 • | 8
5↓ | 1 1 2 2 3 3 3 3 4
 • | 5 6 7 7 8 8 8 9
4↓ | 0 1 1 3 4
 • | 6 7 8
3↓ | 3 4 4
```

Median: 47

Q1-Q3: 41 52

Extremes: 33 58

1. By examining the results of Model A, one can reasonably conclude that (select all correct answers)

≈ $e^{5.67} = \$290.$

~~(a) the median expenditure for persons without an mscd is roughly \$5~~

(b) the median expenditure for persons without an mscd is roughly \$290

~~(c) the median expenditure for persons with an mscd is roughly \$7~~

$e^{5.67+2.34} = \$3011$

~~(d) the median expenditure for persons with an mscd is roughly \$10~~

(e) the median expenditure for persons with an mscd is roughly \$3,000

nearly everyone got correct

Most often missed question

2. By examining the results of Model A and assuming the Gaussian assumption for the residuals is a reasonable approximation, make an interval that will include about 95% of the annual medical expenditures for persons **without a mscd**

$\log(\$_i+1) \equiv Y_i \sim G(x_i\hat{\beta}, \sigma^2) \Rightarrow \approx 95\%$ in interval $x\hat{\beta} \pm 2\hat{\sigma}$

without MSCD, $x\hat{\beta} = 5.67$; $x\hat{\beta} \pm 2\hat{\sigma} = 5.67 \pm 5.04 = (0.63, 10.71)$

$\Rightarrow \$$ in $(\exp(0.63)-1, \exp(10.71)-1)$
$= (\$0, \$44{,}800)$

3. By comparing the results from Models A and B, one can reasonably conclude that (select the single best answer)

(a) neither *lastage* nor *male* improve the prediction because the R^2 values from the two models are within 0.05 of one another — not a direct measure of improvement in fit

(b) neither *lastage* nor *male* improve the prediction because the coefficient for *mscd* changes relatively little ; new predictors could improve predictions but be ⊥ MSCD

(c) neither *lastage* nor *male* improve the prediction of the obserbed expenditures because $F_{4,13642} = 184.1$

(d) *lastage* and/or *male* improve the prediction of the observed expenditures because $F_{4,13642} = 184.1$

(e) none of the above - never right choice

4. By comparing the results from Models A, B and C, one can reasonably conclude that (select the single best answer)

does change not so)

(a) *male* is neither a confounder nor an effect modifier of the effect of *mscd* on log expenditures because the residual standard deviation changes little across the three models

(b) *male* is a confounder of the effect of *mscd* on log expenditures because the *mscd* coefficient in Model C changes substantially from its value in Model B

Valid, but not numerate

(c) *male* modifies the effect of *msdc* on log expenditures because the male:mscd coefficient in Model C is statistically significant

$e^{.69} \approx 2.0$ numerate version of (c) is preferred.

(d) *male* modifies the effect of *msdc* on ~~log~~ expenditures because the mscd effect for men is nearly twice as large as for women, a statistically significant finding

(e) one can not separate the concept of confounding from effect modification with data transformed to the log scale

5. Below find box plots of log expenditure orderded by *mscd : sex : age category*. Viewing this plot and the results of Models A, B and C, one can conclude (select all correct answers)

same thing - see figure

(a) median exenditures increase with age for persons without a mscd but are roughly the same at all ages for persons with a mscd

(b) the effect of mscd on expenditures is modified by age

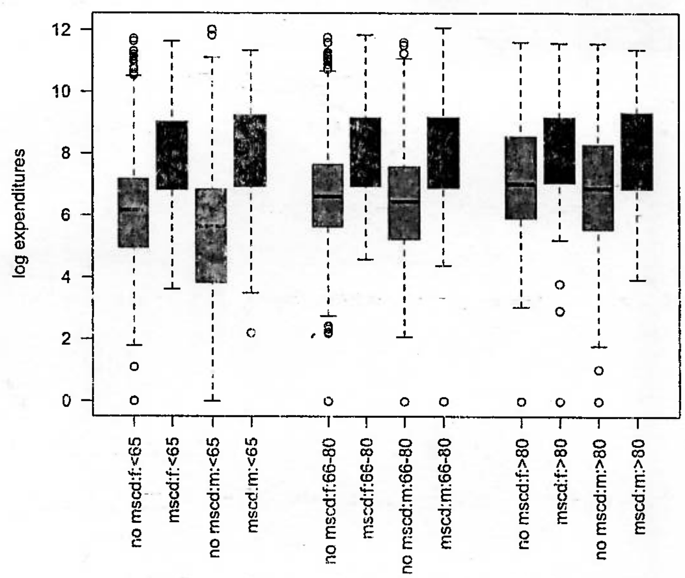
(c) median expenditures are higher for females than males of the same age among persons without an mscd but there is little or no sex difference for those with an mscd

(d) the effect of sex is modified by mscd [no effect for mscd+ ♀ > ♂ for mscd- male: mscd in C stat signif]

(e) the distribution of log expenditures is roughly symmetric with roughly constant interquartile range across the mscd: sex: age groups making the assumptions of the linear regression model more valid on this scale than on the original expenditure scale

yes

Log Expenditures by MSCD:Sex:Age Group



Below find 4 model estimates and a scatterplot of the data with the 4 fitted curves. For model C, the covariance matrix of the estimated regression coefficients is also provided.

Model A. `lm(formula = y ~ x)`

```
             Estimate  Std. Error t value Pr(>|t|)
(Intercept)  1.64743    0.15021  10.968  < 2e-16 ***
x            0.22489    0.03938   5.711 2.72e-08 ***
Residual standard error: 1.584 on 298 degrees of freedom
Multiple R-squared:  0.09865, Adjusted R-squared:  0.09563
F-statistic: 32.62 on 1 and 298 DF,  p-value: 2.717e-08
```

Model B. `lm(formula = y ~ x + x_sp3)`

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.18324    0.10861   10.89   <2e-16 ***
x            0.52622    0.03251   16.19   <2e-16 ***
x_sp3       -6.67539    0.37987  -17.57   <2e-16 ***
Residual standard error: 1.111 on 297 degrees of freedom
Multiple R-squared:  0.5581,  Adjusted R-squared:  0.5551
F-statistic: 187.6 on 2 and 297 DF,  p-value: < 2.2e-16
```

Model C. `lm(formula = y ~ x + x_sp1 + x_sp2 + x_sp3)`

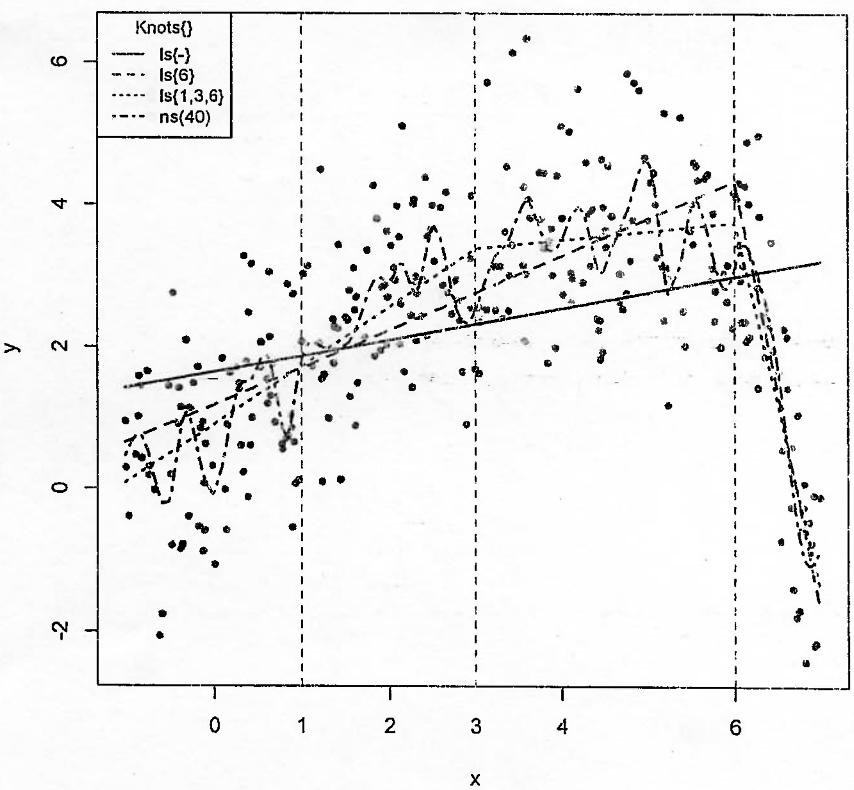
```
              Estimate Std. Error  t value Pr(>|t|)
(Intercept)  0.906853   0.115777    7.833 8.68e-14 ***
x            0.823758   0.166998    4.933 1.36e-06 ***
x_sp1        0.004347   0.264084    0.016 0.986879
x_sp2       -0.704500   0.191942   -3.670 0.000287 ***
x_sp3       -5.384054   0.438489  -12.279  < 2e-16 ***
Residual standard error: 1.061 on 295 degrees of freedomMultiple R-squared:
0.5995,     Adjusted R-squared:  0.5941
F-statistic: 110.4 on 4 and 295 DF,  p-value: < 2.2e-16
```

```
> summary(fit3)$cov.unscaled
              (Intercept)            x        x_sp1         x_sp2        x_sp3
(Intercept)  0.011905115 -0.006249419  0.002723164  0.004168002 -0.001447362
x           -0.006249419  0.024769184 -0.036316021  0.013648263 -0.004739435
x_sp1        0.002723164 -0.036316021  0.061940475 -0.033736304  0.018295008
x_sp2        0.004168002  0.013648263 -0.033736304  0.032721142 -0.036323644
x_sp3       -0.001447362 -0.004739435  0.018295008 -0.036323644  0.170768596
```

10 terms
4 vars
6 covs
for problem 8

Model D. `lm(formula = y ~ ns(x, 40))`

```
Residual standard error: 1.021 on 259 degrees of freedom
Multiple R-squared:  0.6748,  Adjusted R-squared:  0.6245
F-statistic: 13.43 on 40 and 259 DF,  p-value: < 2.2e-16
```



Most got correct

6. By examining the results of Model A and the figure above, one can reasonably conclude about the dependence of y on x that (select the single best answer)
(a) the mean y has a statistically significant positive linear dependence on x
(b) the mean y is estimated to increase 0.22±0.078 per unit increase in x } poor fit; poor choice of model
(c) the mean y is a non-linear function of x and a single slope is not an adequate summary of the relationship
(d) the mean y when x=0 is close to 0, then increases with x
(e) the residuals are not Gaussian making inferences about the mean incorrect } no inference asked for

7. In the space below, calculate the estimated rate of change in mean y per unit x above x=6 from Models B and C. Which slope is steeper?

B: $y \sim \beta_0+\beta_1 x+\beta_2(x-6)^+ \Rightarrow$ slope for $x>6$: $\beta_1+\beta_2 = 0.526 + -6.675 = -6.15$

C: $y \sim \beta_0+\beta_1 x+\beta_2(x-1)^+ +\beta_3(x-3)^+ +\beta_4(x-6)^+ \Rightarrow$ slope is $\beta_1+\beta_2+\beta_3+\beta_4 = .823+.004+.705-5.384 = -5.26$

B is steeper

8. In the space below, calculate a 95% confidence interval for the slope above 6 for Model C.

$$Var(\hat\beta_1+\hat\beta_2+\hat\beta_3+\hat\beta_4) = \sum_{j=1}^{4} Var(\hat\beta_j) + 2\sum_{j=1}^{3}\sum_{k=j+1}^{4} Cov(\hat\beta_j,\hat\beta_k) \quad (10 \text{ terms total})$$

$$= (.10248+.0619+.0327+.1707) + 2[-.0363+.0136-.00474-.0337+.0183-.0363] = .1318$$

95% CI: $\hat\beta \pm 1.96\sqrt{Var\hat\beta} = -5.26 \pm 1.96\sqrt{.1318} = (-5.97, -4.55)$

9. Test the null hypothesis that Model B ~~improves~~ does not improve the prediction of y relative to Model A. Calculate the test statistic and decide whether or not to reject the null that Models B does not improve upon A.

$F_{1,297} = [(1.584^2 \cdot 298 - 1.111^2 \cdot 297)/1] / 1.111^2 = 308.7$. Under null $F_{1,297} \sim \chi^2_1$. $Pr(\chi^2_1 > 308.7) \approx 0$.

Reject null!

Note $F_{1,297} = \left(\hat\beta / se\hat\beta\right)^2 = \left(\frac{-6.15}{\sqrt{.3799}}\right)^2 = (17.57)^2 \approx 308$

The mean squared errors (MSE) without cross-validation (equal to the squares of the residual standard deviations above) and the 10-fold cross-validated mean squared errors (CV-MSE) for the four models are as follows:

Model	A	B	C	D
MSE	2.51	1.23	1.13	1.04
CV-MSE	2.51	1.25	1.15	1.21

10. Which model will predict a new observation with the smallest mean squared error? (select the single best answer)

(a) A (b) B (c) C (d) D (e) no model has the smallest mean squared error

One with smallest cross-validated error

11. The difference between the MSE and the CV-MSE above increases across the 4 models A->D. This is because: (select the single best answer)

(a) the flexibility of the models increases from A to D providing more opportunity for optimization to capitalize on chance

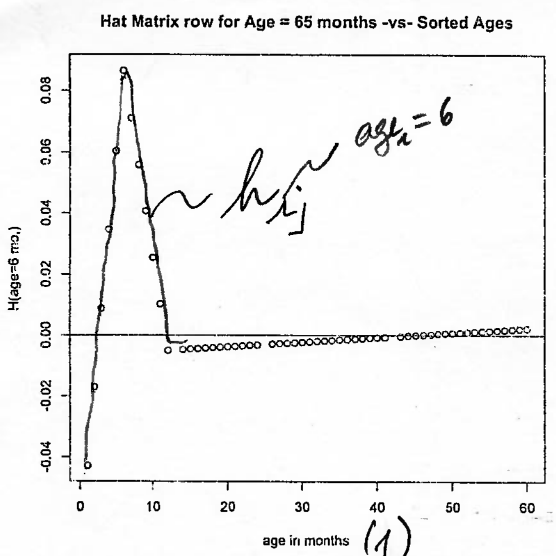
(b) the MSE values are decreasing

(c) the CV-MSE is 10-fold, not n-fold

(d) the 10% left out observations make it more difficult to estimate the bigger models

(e) CV-MSE is a biased estimator

From a regression of Nepal childrens' weight on {age, (age-6)+, and (age-12)+}, below find a plot of a row of the hat matrix H for which the corresponding age is 6 months.



$$\hat{Y}_{age=6} = \sum_{j=1}^{n} h_{ij} y_j$$

row where $age_i = 6$

12. The shape of the row in the plot tells us that: (select the single best answer)

(a) the predicted weight for a child of age 6 months largely depends on the weights of children in the sample with ages between 2 and 10 months of age

(b) the residual weight is smaller for ages near age 0 and again beyond age 12

(c) the predicted weight is more precise for ages beyond age 12

(d) the mean squared error is smaller for ages beyond age 12

(e) **(a) and (b)** ??

has nothing to do with residual or pred. value of weight

13. The correlation of predicted values at times 0 and 6 is: (select the single best answer)

(a) positive (b) negative (c) zero (d) positive, then negative $\boxed{\text{(e)}}$ (a) and (b) — what?

$Var\,\hat{Y} = \sigma^2 H$

$Cov(\hat{Y}_{age=6}, \hat{Y}_j) = h_{i: age=6,\, j} \Rightarrow Cov(\hat{Y}_{age=6}, \hat{Y}_{age=0}) \overset{?}{<} 0 \Rightarrow corr < 0.$