**Biostatistics 140.653: Applied Linear Regression**

**Third Term, 2017-18**

**February 22, 2018**

**10:35-11:35**

Midterm Test

**Instructions**: **This is a closed book test. Answer your favorite 10 out of 12 questions. Cross out the ones you choose not to answer. Do not consult with any other person or any materials beyond what is provided on these exam pages. If needed, you may use R to answer Questions 6 to 8, only.**

**Be brief, be numerate, don’t use statistical jargon where it is unnecessary.**

**Enjoy the hour as an opportunity to sit quietly, think critically, and assess your current knowledge about the course material. Good luck.**

By signing my name, I agree to abide by the Johns Hopkins University Bloomberg School of Public Health Academic Code:

Name (Print): \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Signature: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Below find a graphical display and the results of two regressions of the NMES dataset. The response is total medical expenditures; the predictors are: presence/absence of a major smoking cause (*mscd*=1/0), age in years (*lastage* in years), being below the poverty line or not (pov=1/0), and being male (*mal*e=1/0).**

**Model A**

Call: lm(formula = totalexp ~ mscd + ns(lastage, 3) + factor(pov) + factor(male), data = dsub\_cc)

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1154.27 564.65 2.044 0.04099 \*

mscd 6037.49 337.97 17.864 < 2e-16 \*\*\*

ns(lastage, 3)1 217.85 679.54 0.321 0.74854

ns(lastage, 3)2 4458.25 1364.81 3.267 0.00110 \*\*

ns(lastage, 3)3 3363.00 943.81 3.563 0.00037 \*\*\*

factor(pov)1 -397.97 328.94 -1.210 0.22640

factor(male)1 66.19 311.62 0.212 0.83181

Residual standard error: 10590 on 4795 degrees of freedom

F-statistic: 75.94 on 6 and 4795 DF, p-value: < 2.2e-16

**Model B**

Call: lm(formula = totalexp ~ mscd + **ns(lastage,3)\*factor(pov)\*factor(male),** data = dsub\_cc)

Estimate Std. Error t value

(Intercept) 1113.9 1288.8 0.864

mscd 6042.1 338.5 17.847

ns(lastage, 3)1 889.4 1413.3 0.629

ns(lastage, 3)2 4076.7 3201.0 1.274

ns(lastage, 3)3 706.5 1661.3 0.425

factor(pov)1 -212.8 1506.5 -0.141

factor(male)1 -1310.7 1972.7 -0.664

ns(lastage, 3)1:factor(pov)1 -1863.9 1803.2 -1.034

ns(lastage, 3)2:factor(pov)1 1556.3 3876.9 0.401

ns(lastage, 3)3:factor(pov)1 6237.4 2315.2 2.694

ns(lastage, 3)1:factor(male)1 1657.6 2361.4 0.702

ns(lastage, 3)2:factor(male)1 1236.3 5001.6 0.247

ns(lastage, 3)3:factor(male)1 1375.0 2958.0 0.465

factor(pov)1:factor(male)1 1432.6 2275.7 0.630

ns(lastage, 3)1:factor(pov)1:factor(male)1 -711.3 2873.1 -0.248

ns(lastage, 3)2:factor(pov)1:factor(male)1 -2638.8 5986.8 -0.441

ns(lastage, 3)3:factor(pov)1:factor(male)1 -5435.5 3959.5 -1.373

Residual standard error: 10590 on 4785 degrees of freedom

F-statistic: 29.29 on 16 and 4785 DF, p-value: < 2.2e-16

1. The regression in Model A can be used to quantify the evidence relevant to addressing what question? Be precise.

2. Suppose you want to ask the question of whether, among people who share the same mscd, poverty and gender status, average medical expenditures are higher for older relative to younger people. What regression model would you have to fit and how would you use its results with Model A results to test the null hypothesis that there is no age difference in the population?

3. A researcher’s goal is to estimate the difference in average medical expenditures between persons with a major smoking caused disease and “otherwise similar” persons without one. Using the results from Models A and/or B above, give your short summary of the evidence relevant to achieving her goal.

4. Do the iteractions among the variables age, poverty and gender confound the effect of having a major smoking caused disease on average medical expenditures? Use the results of Models A and/or B to address this question. Be quantitative in your one to two sentence answer.

5. Referring to the graph above and the results for Model B, what do you conclude about the effect of having a major smoking disease for women as opposed to men. Be quantitative.

**Below find the definitions of Y and X in the matrix form of the simple linear model (intercept and single X variable): Y=X + ) (note: = 1)**

**Y = ; X =**

**6. Using the method of maximum likelihood (least squares), evaluate , the estimator of and its variance**

**7. At X=1, evaluate the estimate of the mean Y and its standard error.**

**8. Which observation has the highest *leverage* (potential to influence the regression slope estimate) and why?**

**Below find 4 model (A-D) results and a scatterplot of the data with the 4 fitted curves. The splines x\_sp1, x\_sp2, x\_sp3 are defined to have knots at X=1,3 and 6 respectively. 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

**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

**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

> 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

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

Residual standard error: 1.021 on 259 degrees of freedom



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 |

9. In the space below, calculate a 95% confidence interval for the slope between X=3 and X=6 for Model C.

10. Test the null hypothesis that Model C 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 Model C does not improve upon A.

11. Which model gives the optimal prediction of a new Y value? Explain for a lay person what is meant by optimal?

12. Derive the joint distribution of the predicted values and residuals in the multiple linear regression model Y=X + )