MSiA401: Predictive Analtyics

1. (8 points) Suppose we estimate the coefficients in a linear regression by minimizing

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2$$

for a some $\lambda \geq 0$. For parts (a)–(d), indicate which of i. through v. is correct.

   (a) As we increase λ from 0, the training MSE will

      i. Increase initially, and then eventually start decreasing in an inverted U shape.

      ii. Decrease initially, and then eventually start increasing in a U shape.

      iii. Increase monotonically.

      iv. Decrease monotonically.

      v. Remain constant.

   (b) Repeat (a) for test MSE. *Answer: See JWHT, bottom of page 217. (a) iii; (b) ii; (c) iv; (d) iii.*

2. (12 points) This problem analyzes data from the "Wine Quality Data Set," which is a famous machine learning data set available from the UCI Machine Learning Repository. The data consist of 6497 wines, including 4898 whites and 1599 reds. The dependent variable is wine quality from experts on a 0–10 scale. The predictor variables come from a chemical analysis of the wines including the amount of chlorides, alcohol and free sulfur dioxide. The variable `red` is a dummy that equals 1 for red wines and 0 for white wines. Consider the following model:

```
Call: lm(quality ~ log(chlorides) + red * log(FreeSulfurDioxide) +
    log(alcohol), data = wine)

                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                -4.04908    0.22849 -17.721  < 2e-16 ***
log(chlorides)             -0.17800    0.03218  -5.531 3.31e-08 ***
red                         1.12062    0.10424  10.751  < 2e-16 ***
log(FreeSulfurDioxide)      0.34302    0.02029  16.903  < 2e-16 ***
log(alcohol)                3.49023    0.09697  35.994  < 2e-16 ***
red:log(FreeSulfurDioxide) -0.36078    0.03430 -10.519  < 2e-16 ***
```

   (a) (3 points) Write the estimated regression equation for white wines. Your answer should not involve any terms with the `red` variable. *Answer:* $\hat{y} = -4.04908 - 0.17800\log(\texttt{chlorides}) + 0.34302\log(\texttt{FreeSulferDioxide}) + 3.49023\log(\texttt{alcohol})$

(b) (3 points) Write the estimated regression equation for red wines. Your answer should not involve any terms with the red variable. *Answer:* $\hat{y} = (-4.04908 - 1.12062) - 0.17800 \log(\texttt{chlorides}) + (0.34302 - 0.36078) \log(\texttt{FreeSulferDioxide}) + 3.49023 \log(\texttt{alcohol})$

(c) (3 points) Test at the 5% level whether there is an interaction between the type of wine (red versus white) and log(FreeSulfurDioxide). To receive full credit, state the null, alternative, decision, and rationale. *Answer:* $H_0 : \beta_5 = 0$ *versus* $H_1 : \beta_5 \neq 0$, $P = 2 \times 10^{-16} < .05$ *so reject* H_0.

(d) (3 points) Sketch a graph showing the effect of log(FreeSulfurDioxide) on quality with separate lines for the type of wine (red versus white) (controlling for log(chlorides) and log(alcohol)). **Put your graph on the back of this page.**

3. (14 points) I have data from $n = 21$ Chrysler vehicles. The dependent variable is mpg (miles per gallon) and the predictors are: type (C=car, T=Truck); cid (cubic inches of engine displacement); and drv (drive, where F=front, R=rear and 4=4-wheel). I have regressed mpg on all of the predictor variables with the output below:

```
> fit = lm(formula = mpg ~ type + cid + drv, data = dat)
> summary(fit)
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 38.666726   1.362725  28.375 4.12e-15 ***
typeT       -5.010187   0.821092  -6.102 1.53e-05 ***
cid         -0.033781   0.004024  -8.395 2.95e-07 ***
drvF         4.688635   0.967162   4.848 0.000178 ***
drvR         2.901466   0.917772   3.161 0.006048 **

> drop1(fit)
Single term deletions
Model: mpg ~ type + cid + drv
       Df Sum of Sq     RSS    AIC
<none>              43.147 25.122
type    1   100.405 143.551 48.366
cid     1   190.035 233.181 58.553
drv     2    69.039 112.186 41.188
> qt(.975, 16:21)
[1] 2.119905 2.109816 2.100922 2.093024 2.085963 2.079614
```

(a) (2 points) What does the coefficient (-5.010187) in the typeT row of tell you? Give one sentence in English. *Answer: After controlling for displacement and drive, the gas miles of trucks is 5.01 mpg lower than cars, on the average.*

(b) (3 points) How could you construct a 95% confidence interval for cid? Hint: see the qt output in the output above. Generous partial credit if you get the t value wrong. *Answer:* $-0.033781 \pm 2.119905 \times 0.004024 = [-0.04231, -0.02525]$ *Deduct 1 point if the t value is wrong.*

(c) (3 points) When I called `drop1` I forgot to add the `test="F"` option. Compute the F statistic for the `drv` variable. *Answer:* $F = (69.039/2)/(43.147/16) = 12.801$

(d) (3 points) What null and alternative hypothesis does the F statistic in the previous part test? Define the parameters. *Answer: Let β_3 be the slope for `drv=F` and β_4 be the slope for `drv=R`. $H_0 : \beta_3 = \beta_4 = 0$ versus $H_1 : \beta_3 \neq 0$ and/or $\beta_4 \neq 0$.*

(e) (3 points) What is R^2 if the variance of `mpg` is 32.57062? *Answer:* $R^2 = 1 - 43.147/[32.57062(21-1)] = 0.9338$. *Give partial credit if they did not use $(21-1)$.*

4. (10 points) The salary of workers (in thousands of dollars) is regressed on the age of the worker, the number of years of experience, gender (female=1 for females and 0 for males), training level (takes values 1, 2, and 3, where larger values mean more training), and the interaction between gender and training levels. The output is below.

```
> summary(lm(salary/1000 ~ ageyrs + expyrs + female * trainlev, employee))

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      19.0255     5.3188   3.577 0.000662 ***
ageyrs            0.3221     0.1084   2.971 0.004151 **
expyrs            1.0942     0.2505   4.368 4.60e-05 ***
female          -19.6718     3.9203  -5.018 4.31e-06 ***
trainlev          5.1808     1.3387   3.870 0.000255 ***
female:trainlev   6.7540     2.1541   3.135 0.002576 **
> drop1(fit, test="F")
Single term deletions
                Df Sum of Sq    RSS    AIC F value    Pr(>F)
<none>                       2404.7 262.10
ageyrs           1    326.64 2731.3 269.14  8.8295  0.004151 **
expyrs           1    705.90 3110.6 278.37 19.0812 4.602e-05 ***
female:trainlev  1    363.69 2768.3 270.10  9.8309  0.002576 **
```

(a) (3 points) Is the interaction significant at the .05 level? State the appropriate null and alternative hypotheses, P-value, and decision. *Answer: $H_0 : \beta_5 = 0$ versus $H_1 : \beta_5 \neq 0$. Using either the t or F test we find $P = .002576 < .05$ so reject H_0 and conclude that the effect of training level on salary differs between males and females.*

(b) (5 points) Make an interaction plot showing training level on the horizontal axis and different lines for men and women. Write out the equations of the two lines. Hint: holding age and experience constant, let the intercept for males be the value a, then write the two equations in terms of a. Put answer on back of page. *Answer: For men,* $\hat{y} = a + 5.18\texttt{trainlev}$*. For women,* $\hat{y} = (a - 19.67) + (5.18 + 6.75)\texttt{trainlev} = (a - 19.67) + 11.93\texttt{trainlev}$*.*

(c) (2 points) For which value of training level do the lines intersect, i.e., female salary equals male salary? *Answer:* $19.6718/6.7540 \approx 2.913$

5. (This is probably more difficult than I will give you on the test, but it is an interesting problem to discuss with your friends.) To test the effect of the first Trump-Biden Presidential debate in 2020, researchers picked 1000 names at random from the Chicago telephone directory. An attempt was made to telephone all 1000 one day before the debate and ask their voting intentions. Only 512 were actually questioned; 488 either were not at home or refused to participate. One day after the debate, 387 of the 512 were reached and again asked their voting intentions. In addition, they were asked if they had seen the broadcast. The before-after change in voting preferences of the viewers and non-viewers were compared.

   (a) Draw a diagram for this experiment with X's and O's.

   (b) What is the causal factor?

   (c) What is the criterion variable?

   (d) Discuss threats to external validity.

   (e) Discuss threats to internal validity.

   Answer: (a) Before-after with control quasi design. (b) Exposure to debate. (c) Vote preference (Trump or Biden). (d) The sample is from Chicago phone book (poor design!). The phone book does not even represent Chicago, let alone the whole country. This is a bad sampling frame. (e) This is a fairly strong design and is robust to basic threats. For example, if something else (history) were to happen between the calls that affects candidate preference (e.g., a scandal is discovered), the control group give some protection. Likewise, if there is a selection bias the pre-measures would give some protection. Breaking this design would require a more complicated set of events. For example, suppose that those who watch the debate tend to consume a lot of news, and those who don't watch the debate also don't watch news. Suppose further than something else happens (e.g., scandal). Those who do not watch the debate would be less likely to hear about the scandal, and the scandal would be confounded with the treatment (debate) for those who watch.

3 In 2012 one of the largest firms of stock brokers in the United States noted that a certain type of trading business was growing extremely rapidly. This growth had occurred even though the firm had never had an active training program in options for its account executives. The top management of the company believed that by instituting a comprehensive training program in this new type of trading for its account executives, their sales performance could be further improved. With this goal in mind,

the firm's training department put together a comprehensive training program in the trading utilizing a number of leading business and academic experts. The 50 account executives in the New York area who had done the greatest volume of this type of trading in the first six months of 2012 were selected to participate in the course. Before exposing further large groups of account executives to the new program, the firm decided to monitor the performance of the employees who had participated in the course. In the first six months of 2013 the average volume of this type of trades for these account executives rose by 5% over the comparable period in 2012. The average volume of this type of trade for the account executives in the New York area who had not participated in the program was 29% higher in the first six months of 2013 as compared to the comparable period in 2012.

(a) Draw a diagram for this experiment with X's and O's.

(b) What is the causal factor?

(c) What is the criterion variable?

(d) Discuss threats to external validity.

(e) Discuss threats to internal validity.

Answer: (a) Before-after with control quasi design. (b) Training. (c) Options volume. (d) The sample was selected from NY area, which may not represent other regions. (e) Regression effect is a concern because assignment to treatment made based on a pre-measure. Those who did not get training may have had more untapped potential because they were not doing this type of trading before.