Midterm for STA 314

October 15, 2017

There can be zero to four correct answers to each multiple choice question. One point is assigned to a multiple choice question if and only if all boxes next to the correct answers to this question are checked and no box next to an incorrect answer to this question is checked. For all other questions, if not stated differently, one point is assigned if the task is completely and correctly fulfilled. No negative points are assigned for wrong answers. Throughout, we use the notation, models, estimators, etc. introduced in the lecture. Good luck!

Time: **90 minutes.**

Aids allowed: none.

Name:

Studet ID:

1. (0.5 marks each) *Assume that* `D` *is a data frame that contains 4 columns with names* `Y`, `X`, `V`, `W`. *For each of the following specifications, write down the regression function that corresponds to the* `lm` *call in* ***R***. Example:

   ```
   lm( Y ~ V, data = D)
   ```

   corresponds to $f(x, v, z) = b_1 + b_2 v$.

   (a) `lm( Y ~ . - W, data = D)`

   Solution $f(x, v, w) = b_1 + b_2 x + b_3 v$

   (b) `lm( Y ~ I(X^2) + V:X, data = D)`

   Solution $f(x, v, w) = b_1 + b_2 x^2 + b_3 vx$

   (c) `lm( Y ~ I(sin(X)) + I(V^3), data = D)`

   Solution $f(x, v, w) = b_1 + b_2 \sin(x) + b_3 v^3$

2. (0.5 marks each) *Assume that* `D` *is a data frame that contains 4 columns with names* `Y`, `X`, `V`, `Z`. *For each of the following regression functions, decide if they can be formulated as a linear regression in* ***R*** *(here, $b_1, ..., b_3$ are unknown). If yes, write the* ***R*** *call you would use.*

   (a) $f(x, v, z) = b_1 + b_2 z + b_3 \cos(x^2)$

   ```
   lm( Y ~ Z + I(cos(X^2)), data = D)
   ```

   (b) $f(x, v, z) = b_1 + b_2 \sin(z + b_3 v)$

   Not possible since this is a non-linear function of b_3.

   (c) $f(x, v, z) = b_1 + b_2 zv$

   ```
   lm( Y ~ V:Z, data = D)
   ```

3. *Running a linear regression in **R** and applying the summary function you get the following output*

```
Call:
lm(formula = y ˜ x1 + x2 + x1:x2)

Residuals:
     Min       1Q   Median       3Q      Max
-0.29881 -0.07232 -0.01262  0.07602  0.28263

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.100000   0.011326  -1.048    0.297
x1           0.900000   0.011910  83.384   <2e-16 ***
x2          -0.050000   0.011095  -0.762    0.448
x1:x2        0.100000   0.011635  0.359     0.672
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 0.111 on 96 degrees of freedom
Multiple R-squared:  0.7929,    Adjusted R-squared:  0.7927
F-statistic:  4487 on 3 and 96 DF,  p-value: < 2.2e-16
```

(a) (1 mark) Given the output above, what is your prediction for a new observation with predictor values `x1 = 1, x2 = 2`? You don’t need to simplify your answer, it is enough if you write down the correct formula.

Solution -1.1 + 0.9*1 - 0.05*2 + 0.1*1*2

(b) (0.5 marks) What is the value for R^2 in the model above?

Solution 0.7929

4. *Based on the **R** output in the previous problem, which of the following conclusions can you draw?*

   □ None of the predictors helps to predict the response in a linear model.

   X The predictor x2 can be dropped from the linear model since it does not help to predict the response in the presence of the second predictor.

   □ The regression function specified in the **R** input is correct.

5. (1 mark) *Assume that you observe data* (x_i, y_i) *with values* $(1,2),(2,3),(4,5),(0,2),(3,7)$. *Compute the 2-nn estimator for* $x = 0.5$.

   Solution $(2+2)/2$

6. *You run 12-fold cross-validation on a data set and obtain the following values for the cross-validated error and standard error for different values of k.*

k	1	2	3	4	5	6	7	8	9	10
cv(k)	12	7	6	2	3	4	1	6	15	30
$12^{-1/2}\widehat{se}(k)$	5	4	3	3	5	6	7	5	6	5

   (a) (0.5 marks) Which k would you select based on 12-fold cross-validation?

   Solution 7

   (b) (1 mark) Which k would you select based on the one standard error rule?

   Solution 8

7. *Which output will running the following code in **R** give?*

```
x = array(0,3)
for(k in 1:length(x)){
 x[k] = - k
}
which.min(x)
```

Solution 3

8. *Which output will running the following code in **R** give (you don't need to simplify the answer, writing down the correct formula is enough)?*

```
x = array(0,3)
for(k in 1:3){
 x[k] = k^2
}
sum(x)
```

Solution $1^2 + 2^2 + 3^2$

9. *Qualitative predictors in linear regression*

   X Can be incorporated using dummy variables.

   □ Can not be incorporated since the model would become nonlinear.

   □ Can only be incorporated if the qualitative predictor takes two different values.

10. *Interaction effects between the predictors x_1, x_2 in linear regression*

   □ Can be incorporated by including a term of the form $b(x_1 + x_2)$.

   □ Can be incorporated by including a term of the form $b \sin(x_1 + x_2)$.

   □ Can not be incorporated since the model would become nonlinear.

11. *A small value of R^2 (a value close to zero) in a linear regression model*

   □ Means that the regression model is incorrect.

   □ Means that the relationship between the predictors and the response is linear.

   □ Means that the relationship between the predictors and the response is non-linear.

   □ Means that there is no relationship between the predictors and the response.

12. *Including additional predictors in a linear regression model*

    □ Will always increase the test error.

    □ Will always decrease the test error.

    X Will never decrease R^2.

13. *Which of the following can help to find out if a linear regression model is correct?*

    □ A small value of R^2.

    X Looking for patterns in a residual plot.

    □ A small p-value of the F-test in **R**.

14. *Assume that you run two regressions: k-nn and a simple linear regression (i.e. the predictor is one-dimensional). Your data set contains* $n = 100$ *observations and all the values of* x_i *are different. Which statements are true?*

    □ Linear regression will always give better test error if k is selected by cross-validation.

    □ k-nn will always give the better test error (compared to linear regression) if k is selected by cross-validation.

    X Unless all data points lie on a line k-nn with $k = 1$ will give the smaller training error.

15. *Comparing leave-one-out and 10-fold cross validation for choosing k in k-nn regression on a data set with* $n = 500$

    X Leave-one-out cross validation will typically be computationally more expensive.

    □ 10-fold cross validation will typically be computationally more expensive.

    □ Both types of cross validation will always give the same answer.

16. *Using* 10*-fold cross validation with a data set of size* $n = 300$ *to select a tuning parameter*

    □ Will always result the same tuning parameter since it does not involve any randomness.

    X Might give different answers.

    □ Does not make sense since $n > 10$.

17. *Assume you have a data set with $n = 500$ observations. Which statements are true for k-nn regression?*

    □ Small values of k will always lead to large test error.

    □ Large values of k will always lead to small test error.

    □ Large values of k will always lead to large training error.

18. (1 mark) *You run 10-fold cross-validation on a data set and obtain the following values for the cross-validated error and standard error for different values of k.*

k	1	2	3	4	5	6	7	8	9	10
cv(k)	12	7	6	2	3	4	5	7	11	6
$10^{-1/2}\widehat{se}(k)$	5	4	3	X	5	6	7	5	6	5

Is it possible to find values of $10^{-1/2}\widehat{se}(k)$ in the missing spot (X) so that $k = 9$ is selected by 10-fold cross-validation with the one standard error rule? If yes write down such a value (any solution would be sufficient), if no justify why this is impossible.

Solution This is impossible since $cv(10) < cv(9)$. Reason: if $cv(4) + X < 11 = cv(9)$ $k = 9$ can not be selected and if $cv(4) + X \geq cv(9) = 11$ then also $cv(4) + X > cv(10)$ and $k = 10$ will be selected. So $k = 9$ will not be selected no matter what the value of X.

19. (1.5 marks) Consider a model of the form $y_i = f(x_i) + \varepsilon_i$ and assume that $\widehat{f}(x_0)$ is an estimator that you obtained on the training set. Consider a new observation $y_0 = f(x_0)+\varepsilon_0$ where x_0 is fixed and ε_0 is independent of $\widehat{f}(x_0)$ and satisfies $E[\varepsilon_0] = 0$. Derive an expression for $E[(y_0 - \widehat{f}(x_0))^2]$ in terms of $Var(\varepsilon_0)$, $Var(\widehat{f}(x_0))$, $E[\widehat{f}(x_0)] - f(x_0)$. Justify the steps in your derivation.

$$\begin{aligned}
E[(y_0 - \widehat{f}(x_0))^2] &= E[(f(x_0) - \widehat{f}(x_0) + \varepsilon_0)^2] \\
&= E[(f(x_0) - \widehat{f}(x_0))^2] + E[\varepsilon_0^2] + 2E[\varepsilon_0(f(x_0) - \widehat{f}(x_0))] \\
&= E[(f(x_0) - \widehat{f}(x_0))^2] + Var(\varepsilon_0) + 2E[\varepsilon_0]E[(f(x_0) - \widehat{f}(x_0))] \\
&= (f(x_0) - E[\widehat{f}(x_0)])^2 + Var(\widehat{f}(x_0)) + Var(\varepsilon_0)
\end{aligned}$$

Explanations: second line by linearity of E, third line since $E[\varepsilon_0] = 0$ and by independence of ε_0 and $f(x_0) - \hat{f}(x_0)$, fourth line since $E[\varepsilon_0] = 0$ and by definition of Var, last line by properties of Var and since $f(x_0)$ constant.

20. (1.5 marks) Write down the definition of RSS and TSS in a linear model with a one-dimensional predictor x. Prove that $TSS \geq RSS(\hat{b})$ where $\hat{b}$ is the least squares estimator (the estimator we discussed in class) in the linear model.

$TSS = \sum_{i=1}^n (y_i - \bar{y})^2$.

$RSS(\hat{b}) = \sum_{i=1}^n (y_i - x_i^T \hat{b})^2$

By definition $\hat{b} = \text{argmin}_b RSS(b)$, so

$$RSS(\hat{b}) = \min_b RSS(b) \leq RSS((\bar{y}, 0)^T) = \sum_{i=1}^n (y_i - x_i^T \hat{b})^2 = \sum_{i=1}^n (y_i - \bar{y})^2 = TSS$$