Midterm for STA 312

October 2, 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.

1. (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 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 ~ ., data = D)`

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

   (c) `lm( Y ~ I(X^3) + I(V^2), data = D)`

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

   (b) $f(x, v, z) = b_1 + b_2 x^{b_3}$

   (c) $f(x, v, z) = b_1 + b_2 x + b_3 xv$

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) -0.011000   0.011326  -1.048    0.297
x1           0.900000   0.011910  83.384   <2e-16 ***
x2          -0.010000   0.011095  -0.762    0.448
x1:x2        1.000000   0.011635  86.359   <2e-16 ***
---
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:  0.9929,    Adjusted R-squared:  0.9927
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.

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

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

   - ☐ There is no relationship between the predictors and the response.
   - ☐ The predictor x1 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 not correct.

5. Qualitative predictors in linear regression

   - ☐ Can be incorporated using dummy variables.
   - ☐ Can not be incorporated since the model would become nonlinear.
   - ☐ Can not be incorporated since qualitative predictors lead to a classification problem.

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

   - ☐ Can be incorporated by including a term of the form bx_1x_2.
   - ☐ 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 $b(x_1/x_2)$.
   - ☐ Can not be incorporated since the model would become nonlinear.

7. 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 regression model is correct.
   - ☐ Means that there is no relationship between the predictors and the response.

8. Including additional predictors in a linear regression model

   - ☐ Will typically increase the training error.
   - ☐ Will typically decrease the training error.
   - ☐ Will always increase the test error.
   - ☐ Will always increase the test error.

9. To find out if a linear regression model is correct

   - ☐ One should use R^2.
   - ☐ One should use a residual plot.
   - ☐ One should look at the F-test in **R**.

10. Assume that we run two regressions: k-nn with k selected by cross-validation and a linear regression. Which statements are true

- ☐ Linear regression will always give better test error.
- ☐ k-nn will always give the better test error.
- ☐ If the true relationship is not linear, k-nn will always give better test error.

11. Comparing 100-fold and 10-fold cross validation for choosing k in k-nn regression on a data set with $n = 100$

- ☐ 100-fold 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 involve the same amount of computation.

12. Using 5-fold cross validation with a data set of size $n = 100$ to select a tuning parameter

- ☐ Will always result the same tuning parameter since it does not involve any randomness.
- ☐ Might give different answers depending on the random splitting in 5-fold cross validation.
- ☐ Does not make sense since $n > 5$.

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

- ☐ Large values of k will always lead to large test error.
- ☐ Large values of k will always lead to small test error.
- ☐ Small values of k will always lead to large training error.

14. Assume that you observe data (x_i, y_i) with values $(1, 2), (2, 3), (3, 5)$. Compute the 1-nn estimator for $x = 5/4$.

15. (1.5 marks) Assume that you run k-nn regression on a data set with size n and that the data are generated from $y_i = f(x_i) + \varepsilon_i$ with one-dimensional x_i and $\varepsilon_i \sim N(0,1)$. Does there exist a regression function f for which choosing $k = n$ will lead to the best test error, independently of n? Justify your answer.

16. (1.5 marks) Assume that you observe n data points (x_i, y_i) with $x_i = i/n, i = 1, ..., n$ and $y_i = f(x_i) + \varepsilon_i$ where ε_i are iid $N(0,1)$ independent of x_i. What is the best possible MSE for predicting a new observation $y_0 = f(x_0) + \varepsilon_0$ you can hope to achieve in this setting by any regression method (you do not need to specify the method, n can be arbitrarily large)? Justify your answer.

17. The residual plot above was generated by running a linear regression of the form $f(x) = b_1 + b_2 x$. Residuals are plotted against x. The plot indicates that

- ☐ The errors have non-constant variance but the regression function f is correct.
- ☐ The errors have non-constant variance and the regression function f is wrong.
- ☐ The regression function f is correct and the errors have constant variance.
- ☐ The regression function f is wrong and the residuals have constant variance.

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

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