ORIE 4740 Spring 2017

Homework 1

(Due March 29 by 12:30pm. **Note the non-standard date and time.**)

Instruction: Check the boxes next to the correct answers. There can be **zero to four** correct answers to each question. A question with m boxes is worth m points. One point is deduced each time a correct answer is not checked, or an incorrect answer is checked.

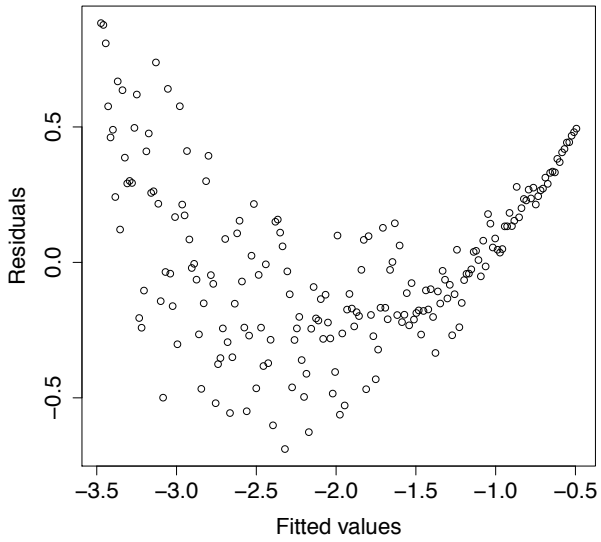
1. For linear regression, the residual plot below suggests that

   ☐ The errors have non-constant variance but the linear assumption is correct.

   ☐ The errors have non-constant variance and the linear assumption is wrong.

   ☐ The linear assumption is correct and the errors have constant variance.

   ☐ The linear assumption is wrong and the errors have constant variance.



2. Which of the following assumptions did we make when deriving the expression $\hat{\beta} = (X^\top X)^{-1} Xy$ for the least squares estimator for linear regression?

   ☐ The matrix $X^T X$ is invertible.

   ☐ The errors follow a normal (aka Gaussian) distribution.

   ☐ The distribution of the errors is symmetric around zero.

3. Nonlinearity between the response and a predictor x in regression

   ☐ May be handled by including a term of the form $\beta * x^2$.

   ☐ May be handled by including a term of the form $\beta * \sqrt{x}$.

   ☐ May be handled by including a term of the form $\beta * \log(x)$.

   ☐ Can never be handled since the model would become nonlinear.

4. Suppose the true model between the response y and a predictor x is $y = f(x)$. Which of the following models can be estimated using the least squares approach to linear regression?

- ☐ $f(x) = \beta_0 + \beta_1\sqrt{x}$
- ☐ $f(x) = \beta_0 + \beta_1 x \cdot \cos(x)$
- ☐ $f(x) = \beta_0 + \beta_1 x + \beta_2 \log(x + \beta_3)$

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

```
Call:
lm(formula = y ~ x + I(x^2))

Residuals:
     Min       1Q   Median       3Q      Max
-2.31384 -0.67054  0.01942  0.62198  2.35304

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.01030    0.06014  -0.171    0.864
x            1.02598    0.07512  13.658   <2e-16 ***
I(x^2)       0.07300    0.08409   0.868    0.387
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3079 on 97 degrees of freedom
Multiple R-squared:  0.8892,    Adjusted R-squared:  0.8881
F-statistic: 790.8 on 2 and 97 DF,  p-value: < 2.2e-16
```

Based on this output we can say that

- ☐ The predictor x can be dropped from the linear model since it does not help to predict the response in the presence of the second predictor.
- ☐ The predictor x^2 can be dropped from the linear model since it does not help to predict the response in the presence of the first predictor.
- ☐ Both the predictors x and x^2 can be dropped from the linear model since they have no relationship with the response.

6. Assume that we want to classify data points in one of $K = 4$ classes based on some predictor values. For a new data point with predictor value x_0, we use the K Nearest Neighbor algorithm to compute $\hat{p}_1(x_0), ..., \hat{p}_K(x_0)$, which denote estimators for the conditional class probabilities given the predictor value x_0 of class $1, ..., K$, respectively.

- ☐ If $\hat{p}_k(x_0) < 0.5$, then one should *never* classify the new data point to class k.
- ☐ Whenever $\hat{p}_k(x_0) = 1$, one should classify the new data point to class k.
- ☐ One should classify the new data point to class k if $\hat{p}_k(x_0)$ is the largest among $\hat{p}_1(x_0), ..., \hat{p}_K(x_0)$.
- ☐ One should classify the new data point to class k if $\hat{p}_k(x_0)$ is the smallest among $\hat{p}_1(x_0), ..., \hat{p}_K(x_0)$.

7. Comparing forward selection, backward selection and best subset selection (on the same data set)...

- ☐ Best subset selection can be computationally more expensive than forward selection.

☐ Best subset selection can be computationally less expensive than forward selection.

☐ Best subset selection and forward selection are computationally equally expensive.

☐ Forward selection and backward selection *always* lead to the same result.

8. Using 5-fold cross validation with a data set of size $n = 100$ to select the value of K in the K-Nearest Neighbor algorithm

   ☐ Will *always* result in the same K 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 is larger than the number of folds.

9. If we want to select a subset of variables in linear regression,

   ☐ It is *always* better to use adjusted R^2 than cross-validation

   ☐ Adjusted R^2 and cross-validation will *always* lead to a model with the same prediction error.

   ☐ We should choose the subset that leads to the lowest R^2.

10. Which of the following statements is/are true for best subset selection based on BIC?

    ☐ It will *sometimes* select a less flexible model than using adjusted R^2.

    ☐ It will *always* select a more flexible model than using adjusted R^2.

    ☐ It will *always* select the same model as using adjusted R^2.

11. In linear regression, if $p > n$, where p is the number of predictors and n is the number of training data points.

    ☐ We can compute the unique least squares solution using all predictors, although its accuracy will be low.

    ☐ We can select a subset of the predictors and compute the least squares solution using them.

    ☐ We can fit a linear model with zero error on the training data (that is, zero training error).

    ☐ Fitting a linear model using all p predictors will give good test error, since the model is flexible.

12. In linear regression, if we add more high order terms as predictors, the value of the R^2 statistic will typically

    ☐ increase.

    ☐ decrease.

    ☐ remain the same.

    ☐ first decrease and then increase.

13. Recall the bias-variance tradeoff. A more flexible model typically

    ☐ Has higher bias.

    ☐ Has lower bias.

    ☐ Has lower variance.

    ☐ Has lower test errors.

14. In the K-NN classification algorithm, using a larger K typically

- ☐ increases the bias.
- ☐ increases the variance.
- ☐ leads to a more flexible model.
- ☐ leads to a lower training error.

15. Adding interaction terms in linear regression

- ☐ increases the bias.
- ☐ increases the variance.
- ☐ leads to a more flexible model.
- ☐ leads to a lower training error.

16. Logistic regression

- ☐ has higher variance than K-NN.
- ☐ has lower variance than K-NN.
- ☐ has the same variance as K-NN.
- ☐ has better performance on test data than K-NN.

17. Which of the following is/are true about k-means clustering.

- ☐ It is a heuristic method that is not guaranteed to find the exact optimal solution (that minimizes the total within cluster variability).
- ☐ It is always better to use a larger value of k.
- ☐ If you run k-means with different initialization, you may get different clustering results.
- ☐ It alternates between computing the cluster centers and associating observations to clusters.