QBUS3820: Practice Questions for Final Exam

1. Let

$$a = \begin{pmatrix} 1 \\ -1 \\ 2 \end{pmatrix}, \quad b = \begin{pmatrix} 2 \\ 1 \\ -2 \end{pmatrix}$$

   (a) Compute the inner product $a'b$

   (b) Compute the length of a

   (c) Compute the Eucledian distance between a and b

2. Let

$$X = \begin{pmatrix} 1 & 2 & 5 \\ 1 & -3 & 4 \\ 1 & 3 & 3 \end{pmatrix}$$

   (a) Compute $X'X$

   (b) Compute $tr(X'X)$

3. Suppose that you want to fit the model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

   where $\mathbb{E}(\epsilon_i) = \mu$ and $\mathbb{V}(\epsilon_i) = \sigma^2$. Suppose that $\mu \neq 0$ is known.

   How can you transform this model to the linear regression model where the mean of error terms is zero?

4. I were trying to predict personal income using the linear regression model based on two predictors `education` (years of education) and `children` (number of children). After fitting the linear regression model, I saw that the coefficient for `children` was significant but `education` was not. Hence, I concluded that education was not an important predictor of income and decided to predict the income based only on the number of children. Comment on this finding. What (if any) should be done instead?

5. Explain whether each following scenario is a classification or regression problem, and indicate whether we are most interested in inference (i.e. explaining the relationship between variables) or prediction or both.

   (a) We collect a data set on the top 500 firms. For each firm we record profit, number of employees, industry, etc and the CEO salary. We are interested in understanding which factors affect CEO salary.

(b) We are considering launching a new product and wish to know whether it will be a success or a failure. We collect data on 20 similar products that were previously launched. For each product we have recorded whether it was a success or failure, price charged for the product, marketing budget, competition price, and ten other variables.

(c) We are interested in the housing prices in Sydney in relation to variables such as location, land size, number of bedrooms, etc.

6. Suppose we are interested in explaining the starting salary after graduation (in thousands of dollars) using `GPA, IQ, Gender` (1 if female, 0 if male) and their interaction. The final model after doing variable selection is

$$\widehat{Y} = 50 + 20 \times \texttt{GPA} + 0.07 \times \texttt{IQ} + 35 \times \texttt{Gender} + 0.01 \times \texttt{GPA*IQ} - 10 \times \texttt{GPA*Gender}$$

(a) Which answer is correct, and why?

i. For a fixed value of IQ and GPA, males earn more on average than females.

ii. For a fixed value of IQ and GPA, females earn more on average than males.

iii. For a fixed value of IQ and GPA, males earn more on average than females provided that the GPA is high enough.

iv. For a fixed value of IQ and GPA, females earn more on average than males provided that the GPA is high enough.

(b) Predict the salary of a female with IQ of 110 and a GPA of 4.0

(c) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

7. Suppose we collect data for a group of master students in a statistics class with variables X_1 = hours studied, X_2 =undergrad GPA, and Y = receive an HD. We fit a logistic regression and produce estimated coefficient $\widehat{\beta}_0 = -6, \widehat{\beta}_1 = 0.05, \widehat{\beta}_2 = 1$.

(a) Estimate the probability that a student who studies for 40h and has an undergrad GPA of 3.5 gets an HD in the class.

(b) How many hours would the student in part (a) need to study to have a 50% chance of getting an HD in the class?

8. A study looked at 674 murder crime cases in Florida between 1976 and 1987. Table below gives the output from a logistic regression model where the response is death penalty verdict (1=Yes, 0=No), predictors are defendant's race (1=White,0=Black) and victim's race (1=White,0=Black).

Parameter	Estimate
Intercept	-3.5961
Defendant	-0.8678
Victim	2.4044

Which group among the four groups

(1) white defendant, white victim

(2) white defendant, black victim

(3) black defendant, white victim

(4) black defendant, black victim

is most likely to be sentenced a death penalty?

9. Each subject has two measurements X_1 and X_2. Given that the subject belongs to group 1, then $X = (X_1, X_2)$ follows a bivariate normal distribution with mean μ_1 and the covariance matrix Σ_1

$$\mu_1 = \begin{pmatrix} 157 \\ 45 \end{pmatrix}, \quad \Sigma_1 = \begin{pmatrix} 158 & 5 \\ 5 & 7 \end{pmatrix}, \quad \Sigma_1^{-1} = \begin{pmatrix} 0.088 & -0.063 \\ -0.063 & 0.188 \end{pmatrix}.$$

Given that the subject belongs to group 2, then $X = (X_1, X_2)$ follows a bivariate normal distribution with mean μ_2 and the covariance matrix Σ_2

$$\mu_2 = \begin{pmatrix} 168 \\ 56 \end{pmatrix}, \quad \Sigma_2 = \begin{pmatrix} 17 & 4 \\ 4 & 10 \end{pmatrix}, \quad \Sigma_2^{-1} = \begin{pmatrix} 0.065 & -0.026 \\ -0.026 & 0.110 \end{pmatrix}.$$

A subject has the overall probability of 0.47 belonging to group 1. Given a particular subject x with measurements $x = (167, 60)$

(a) compute the discriminant score $\delta_1(x)$ and $\delta_2(x)$

(b) which group would you classify this subject into?

10. Each subject has two measurements X_1 and X_2. Given that the subject belongs to group 1, then $X = (X_1, X_2)$ follows a bivariate normal distribution with mean μ_1 and the covariance matrix Σ. Given that the subject belongs to group 2, then $X = (X_1, X_2)$ follows a bivariate normal distribution with mean μ_2 and the covariance matrix Σ.

$$\mu_1 = \begin{pmatrix} 157 \\ 45 \end{pmatrix}, \quad \mu_2 = \begin{pmatrix} 168 \\ 56 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 158 & 5 \\ 5 & 7 \end{pmatrix}, \quad \Sigma^{-1} = \begin{pmatrix} 0.088 & -0.063 \\ -0.063 & 0.188 \end{pmatrix}.$$

A subject has the overall probability of 0.47 belonging to group 1. Given a particular subject x with measurements $x = (167, 60)$

(a) compute the discriminant score $\delta_1(x)$ and $\delta_2(x)$

(b) which group would you classify this subject into?

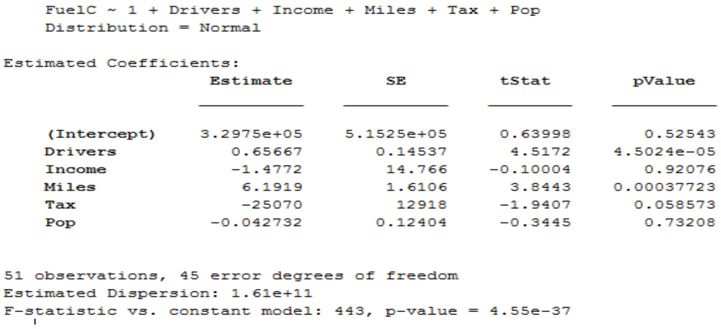
11. The US Federal Highway Administration wanted to understand how fuel consumption varies over the 50 states and the District of Columbia. It collected a dataset in 2001, which contains the following variables

FuelC	Gasoline consumption for road use (in 1000 gallons)
Drivers	Number of licensed drivers in the state
Income	Average personal income in 2000 (in \$1000)
Miles	Miles of Federal highway in the state
Tax	Gasoline state tax rate (cents per gallon)
Pop	2001 state population aged 16 and over

The correlation coefficients between the predictors are given in the following table

	Drivers	Income	Miles	Tax	Pop
Drivers	1.0000				
Income	0.2562	1.0000			
Miles	0.6686	-0.1352	1.0000		
Tax	-0.1654	-0.0107	-0.0645	1.0000	
Pop	0.9950	0.2651	0.6712	-0.1459	1.0000

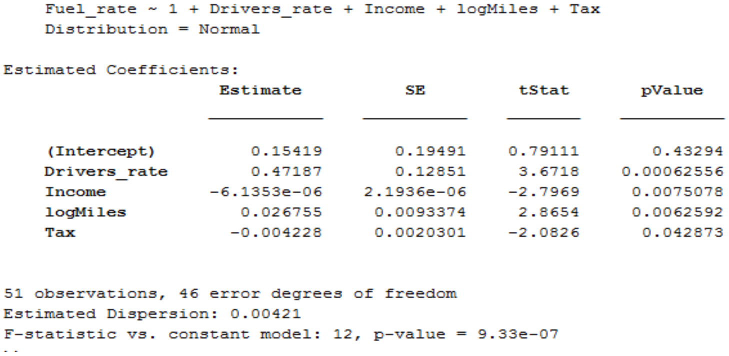
(a) I fit a linear regression model to explain `FuelC` using all the 5 predictors; see the output below. What would be wrong with this?



```
    FuelC ~ 1 + Drivers + Income + Miles + Tax + Pop
    Distribution = Normal

Estimated Coefficients:
                    Estimate        SE          tStat        pValue
                   __________    __________    ________    __________

    (Intercept)    3.2975e+05    5.1525e+05     0.63998       0.52543
    Drivers           0.65667       0.14537      4.5172    4.5024e-05
    Income            -1.4772        14.766    -0.10004       0.92076
    Miles              6.1919        1.6106      3.8443    0.00037723
    Tax                -25070         12918     -1.9407      0.058573
    Pop             -0.042732       0.12404     -0.3445       0.73208


51 observations, 45 error degrees of freedom
Estimated Dispersion: 1.61e+11
F-statistic vs. constant model: 443, p-value = 4.55e-37
```

(b) Both `Drivers` and `FuelC` are state totals, so these will be large in more populous states and smaller in less populous states. To eliminate the effect of state size, I define Drivers_rate = Drivers/Pop and Fuel_rate = FuelC/Pop. Aslo, to scale down the size of Miles, I use log(Miles).

Below is the output of regression of Fuel_rate on Drivers_rate, Income, log(Miles) and Tax



```
    Fuel_rate ~ 1 + Drivers_rate + Income + logMiles + Tax
    Distribution = Normal

Estimated Coefficients:
                      Estimate          SE         tStat        pValue
                     ___________    __________    _______    __________

    (Intercept)          0.15419       0.19491    0.79111       0.43294
    Drivers_rate         0.47187       0.12851     3.6718    0.00062556
    Income           -6.1353e-06    2.1936e-06    -2.7969     0.0075078
    logMiles            0.026755     0.0093374     2.8654     0.0062592
    Tax                -0.004228     0.0020301    -2.0826      0.042873


51 observations, 46 error degrees of freedom
Estimated Dispersion: 0.00421
F-statistic vs. constant model: 12, p-value = 9.33e-07
```

(i) Is this model more appropriate than the one in part (a)? Explain

(ii) It seems that the coefficient for Income is very small, approximately -0.0000061. Should I remove this predictor from the model? Justify your answer

(iii) Explain the effect of the predictors on the fuel consumption.

(iv) If Tax is increased by 1 cent per gallon, what would be the change in the personal fuel consumption?

12. True or false (explanation/discussion of the answer is needed)

(a) Logistic regression model can be estimated using the least squares method.

(b) Cubic spline regression is not suitable when the predictor is categorical

(c) Classification using logistic regression can be used when the predictors are numerical or categorical or both

(d) LASSO is a suitable method for variable selection when the number of potential predictors is large

(e) Best subset selection using AIC as the selection criterion is a suitable method for variable selection even when the number of potential predictors is large, because it's not expensive to compute the AIC value for each subset.

(f) Spline regression and neural networks are non-linear methods for regression and classification

13. (a) Using your own words, describe the cross-validation method for model selection

(b) Using your own words, describe how a suitable shrinkage parameter λ can be selected using the BIC-type criterion.

(c) What are the main differences between parametric methods and nonparametric methods?

14. We have an i.i.d sample $X_1, ..., X_n,\ n = 1000$, which are final scores of 1000 students. The range of the scores is between 0 and 100. Let $f(x)$ be the density function of scores. To construct a histogram estimate of $f(x)$, we divide the range into m equal bins. Let $S_n(x)$ be the total number of X_i that is smaller than or equal to x. Suppose that $S_n(40) = 160$, $S_n(50) = 343$ and $S_n(60) = 750$. Let $m = 10$.

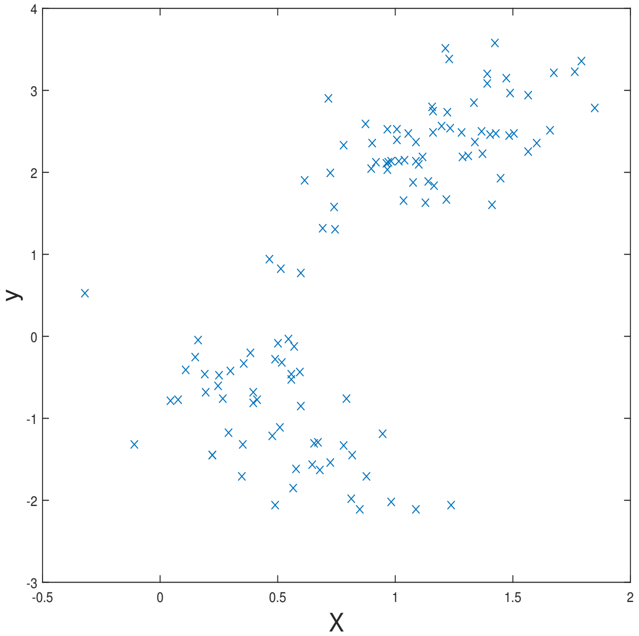
(a) What is the binwidth h? List down the bins.

(b) Find the histogram estimate of $f(x)$ at $x = 50$ and at $x = 55$.

15. Suppose that you are using LASSO for variable selection in logistic regression with $n = 100$ observations. Let $\ell(\beta)$ be the log-likelihood function. Given a shrinkage parameter λ, you solve for the Lasso estimate $\widehat{\beta}_\lambda$ of β. For a shrinkage parameter λ_1, your software gives $\ell(\widehat{\beta}_{\lambda_1}) = -201.4$ and 5 coefficients are estimated to be non-zero. For a shrinkage parameter λ_2, $\ell(\widehat{\beta}_{\lambda_2}) = -205.4$ and 3 coefficients are estimated to be non-zero. Compute BIC(λ_1) and BIC(λ_2). Which value of λ do you choose?

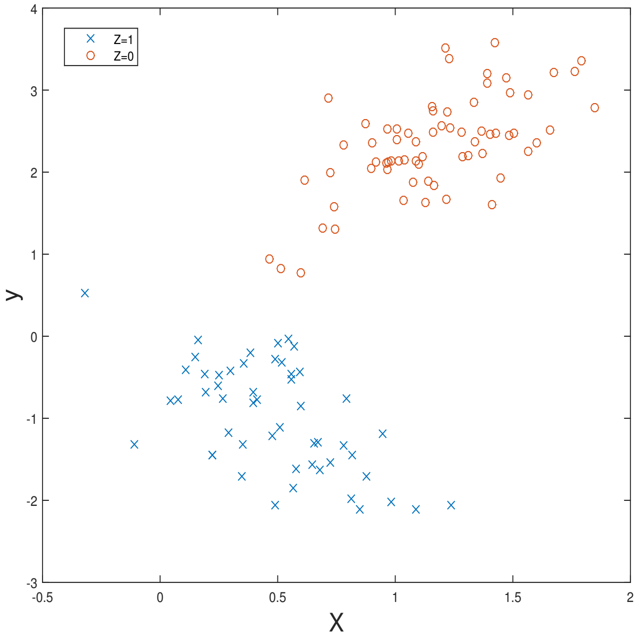
16. Suppose that you are using LASSO for variable selection in linear regression with $n = 100$ observations. Let $\ell(\beta) = \log(\|\boldsymbol{y} - \boldsymbol{X}\beta\|^2/n)$ with $\boldsymbol{y}$ the response vector and $\boldsymbol{X}$ the design matrix. Given a shrinkage parameter λ, let $\widehat{\beta}_\lambda$ be the Lasso estimate of β. For a shrinkage parameter λ_1, $\ell(\widehat{\beta}_{\lambda_1}) = -21.4$ and 25 coefficients are estimated to be non-zero. For a shrinkage parameter λ_2, $\ell(\widehat{\beta}_{\lambda_2}) = -19.7$ and 32 coefficients are estimated to be non-zero. Compute BIC(λ_1) and BIC(λ_2). Which value of λ do you choose?

17. (a) Consider the following scatter plot



Is it likely to be appropriate to fit the linear regression model to this dataset?

(b) Examining the dataset more closely reveals that there is a group membership Z associated with the predictor X: $Z_i = 1$ if X_i belongs to group 1 and $Z_i = 0$ if X_i belongs to group 2. The scatter plot below shows y v.s. X when the group membership Z is taken into account.

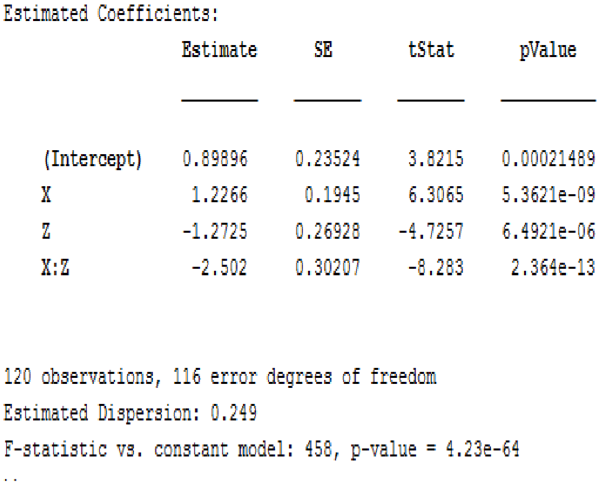


Does the plot suggest that X and Z have an interaction effect on y? Explain.

(c) Let's consider the linear regression model for Y with an interaction effect of X and Z.

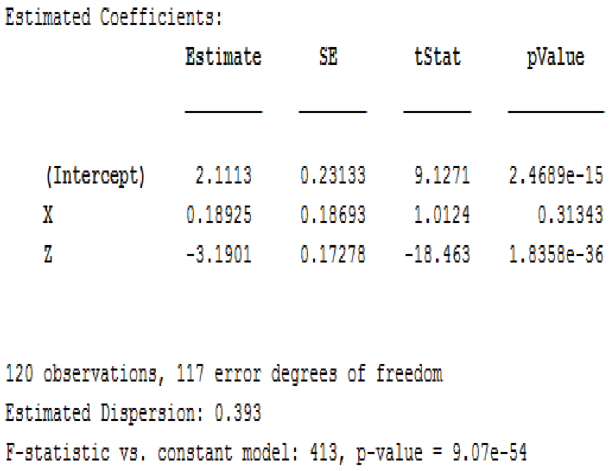
(i)

(d) Below is the output of the linear regression model with an interaction effect of X and Z



```
Estimated Coefficients:
                   Estimate       SE        tStat       pValue  
                   ________    _______     _______    __________

    (Intercept)    0.89896     0.23524      3.8215    0.00021489
    X               1.2266      0.1945      6.3065    5.3621e-09
    Z              -1.2725     0.26928     -4.7257    6.4921e-06
    X:Z             -2.502     0.30207      -8.283     2.364e-13


120 observations, 116 error degrees of freedom
Estimated Dispersion: 0.249
F-statistic vs. constant model: 458, p-value = 4.23e-64
..
```

(i) Is it important to add an interaction effect of X and Z into the model? Explain. Below is the regression output without the interaction term



```
Estimated Coefficients:
                   Estimate      SE        tStat       pValue   
                   ________    _______    _______    __________

    (Intercept)      2.1113    0.23133     9.1271    2.4689e-15
    X               0.18925    0.18693     1.0124       0.31343
    Z               -3.1901    0.17278    -18.463    1.8358e-36


120 observations, 117 error degrees of freedom
Estimated Dispersion: 0.393
F-statistic vs. constant model: 413, p-value = 9.07e-54
```

(ii) In the model with the interaction term, what are the estimated regression equations for the two groups?

(iii) In the model with the interaction term, what are the assumptions that have been made in this linear regression modelling?

(iv) Which estimation method has been used to obtain the regression output above?

Short answers

Q12

(a) False. Logistic regression model is estimated by MLE

(b) True. The way the basis functions are formed implicitly means that X must not be a categorical variable

(c) True. The model is still well defined no matter what the types of the predictors are.

(d) True. We don't have to browse through a large number of subsets to search for the best one.

(e) False. We still have to browse through a HUGE number of subsets to search for the best one.

(f) True. Non-linear effects such as X^2, X^3 are included in the models

Q13

(a) We first need a loss function. The dataset is randomly divided into K subsets. Consider a model M. For each subset, k, model M is fit to the rest $K-1$ subsets and the loss function is computed on the subset k. This gives us K prediction errors, and the cross-validated prediction error of model M is the average of these errors. The best model is the one that has the smallest CV prediction error.

(b) Create a range of values of λ from 0 to $\lambda_{\max}$. For each value of λ, compute the Lasso estimate and then compute BIC(λ). Select the λ that minimises these BIC values.

(c) Parametric approaches (PA) assume a functional form for the underlying model that generated the data, non-parametric approaches (NPA) don't. PA have a better convergence rate (if the assumed model is the true model) than NPA. NPA is more flexible and robust than PA.

Q14

(a) $h = 100/m = 10$. The bins are $[0, 10]$, $(10, 20]$,..., $(80, 90]$ and $(90, 100]$

(b) $f_n(50) = (S_n(50) - S_n(40))/(n \times h) = 0.0183$. $f_n(55) = (S_n(60) - S_n(50))/(n \times h) = 0.0407$.

Q15: BIC(λ_1) = 425.8259 and BIC(λ_2) = 424.6155. λ_2 is preferable.

Q16: BIC(λ_1) = -20.2478 and BIC(λ_2) = -18.2263. λ_2 is preferable.

Q17

(a) No

(b) Yes. It appears that the effect of X on Y depends on Z. When X increases, Y is likely to increase when $Z = 0$ and Y is likely to decrease when $Z = 1$

(c) $Y = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 X \times Z + \epsilon$.

(d-i) Yes, all the coefficients are significant at the significance level of 0.01. Furthermore, X is no longer significant when the interaction term is removed

(d-ii) For group 1, $Z = 0$, $\widehat{Y} = 0.9 + 1.23X$. For group 2, $Z = 1$, $\widehat{Y} = -0.37 - 1.27X$

(d-iii) Linearity assumption, zero error mean $\mathbb{E}(\epsilon) = 0$, homoscedasticity $\mathbb{V}(\epsilon)$ is a constant

(d-iv) Least squares method, as no assumption is made on the distribution of Y.