Applied Statistics Comprehensive Exam

August 2018

Ph.D Day 2 - Exam

This comprehensive exam consists of 10 questions pertaining to two topics of your choice.

Before you start, Please make sure the topics are the one you have chosen.

1 This Ph.D level exam will run from **8:30 AM to 3:30 PM**.

2 Please label each page with your identification number.

DO NOT USE YOUR NAME OR BEAR NUMBER.

3 Please write only on one side of each page.

4 Please leave one inch margins on all sides of each page.

5 Please number all pages consecutively.

6 Please label the day number (Day 1 or Day 2) on each page.

7 Please begin each question on a new page, and number each question.

8 Please do not staple pages together.

9 No electronic devices, formula sheets, or other outside materials are permitted.

10 Statistical tables and paper will be provided.

11 Relax and good luck!

I have read and understand the rules of this exam.

Signature:_____ Date:_____

1. In paragraph forms answer the following questions regarding a Multiple Linear Regression Analysis.

   i. Describe piecewise linear regression models. Explain when/why they are used.

   ii. Describe the consequences of incorrect model specification.

   iii. Give two interpretations of VIF.

---

2. Consider the model given by

$$Y_i = \beta_0 + W_{i1}\gamma_1 + W_{i2}\gamma_2 + \epsilon_i,$$

   where $\epsilon_i \sim NID(0, \sigma^2)$, and

$$\begin{aligned} \sum_i W_{i1} &= \sum_i W_{i2} = \sum_i W_{i1}W_{i2} = 0 \\ \sum_i W_{i1}^2 &= 1+\rho, \quad \sum_i W_{i2}^2 = 1-\rho. \end{aligned}$$

   Consider the estimator

$$\widehat{\gamma}_{1(k_1)} = \frac{\sum_i W_{i1}Y_i}{\sum_i W_{i1}^2 + k_1}$$

   i. For $k_1 > 0$, show that $\widehat{\gamma}_{1(k_1)}$ is a biased estimate of γ_1.

   ii. Find the mean squared errors of $\widehat{\gamma}_{1(k_1)}$.

---

3. Consider the test for lack of fit for a multiple linear regression. Find $E(MS_{PE})$ and $E(MS_{LOF})$. Note that

$$MS_{PE} = \frac{1}{n-m}\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2$$

$$MS_{LOF} = \frac{1}{m-p}\sum_{i=1}^{m} n_i(\bar{y}_i - \hat{y}_i)^2$$

---

4. A data set was collected to model relationship between selling price to nine regressors. Using this data set, the attached SAS output has been compiled to test for the potential effects of the nine regressors on the selling price. Based on SAS output page 6, answer the following questions.

i. Test for significance of regression. What conclusions can you draw?

ii. Use t tests to assess the contribution of each regressor to the model. Discuss your findings.

iii. What is the contribution of lot size and living space to the model given that all of the other regressors are included?

iv. Is multicollinearity a potential problem in this model?

---

5. DUPLEX algorithm was used to split a data set on the gasoline mileage performance of 30 different automobiles into estimation and prediction sets. Based on SAS output page 7, answer the following questions.

   i. Evaluate the statistical properties of these data sets.
   [Hint: Use the relative volumes of the regions spanned by the two data sets.]

   ii. Fit a model involving x_1 and x_6 to the estimation data. Do the coefficients values from this model seem reasonable?

   iii. Use this model to predict the observations in the prediction data set. What is your evaluation of this model's predictive performance?

---

6. If the observation vectors $\boldsymbol{y_1}, \boldsymbol{y_2}, \cdots, \boldsymbol{y_n}$ is a random sample from $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The density function for $\boldsymbol{y} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is

$$\frac{1}{(\sqrt{2\pi})^p |\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{y}-\boldsymbol{\mu})}.$$

   Then

   i. show that $\sum_{i=1}^{n} (\boldsymbol{y_i} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{y_i} - \boldsymbol{\mu}) = tr(\boldsymbol{\Sigma}^{-1}[\boldsymbol{W} + n(\bar{\boldsymbol{y}} - \boldsymbol{\mu})(\bar{\boldsymbol{y}} - \boldsymbol{\mu})'])$, where $\boldsymbol{W} = \sum_{i=1}^{n} (\boldsymbol{y_i} - \bar{\boldsymbol{y}})(\boldsymbol{y_i} - \bar{\boldsymbol{y}})'$

   ii. show that the maximum likelihood estimator of $\boldsymbol{\mu}$ is $\widehat{\boldsymbol{\mu}} = \bar{\boldsymbol{y}}$;

   iii. show that the maximum likelihood estimator of $\boldsymbol{\Sigma}$ is $\widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\boldsymbol{W}$.

---

7. If the observation vectors $\boldsymbol{y_1}, \boldsymbol{y_2}, \cdots, \boldsymbol{y_n}$ is a random sample from $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma}$ unknown. The density function for $\boldsymbol{y} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is

$$\frac{1}{(\sqrt{2\pi})^p |\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{y}-\boldsymbol{\mu})}.$$

   Let $L(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ be the likelihood function for the sample. Then for H_0: $\boldsymbol{\mu} = \boldsymbol{\mu}_0$ versus H_1: $\boldsymbol{\mu} \neq \boldsymbol{\mu}_0$,

i. show that $\sum_{i=1}^{n}(\boldsymbol{y_i}-\boldsymbol{\mu_0})'\widehat{\boldsymbol{\Sigma}}_0^{-1}(\boldsymbol{y_i}-\boldsymbol{\mu_0}) = np$, where $\widehat{\boldsymbol{\Sigma}}_0 = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{y_i}-\boldsymbol{\mu_0})(\boldsymbol{y_i}-\boldsymbol{\mu_0})'$ that maximizes $L(\boldsymbol{\mu}_0, \boldsymbol{\Sigma})$ under H_0.

ii. show that the likelihood ratio

$$LR = \frac{\max\limits_{H_0} L(\boldsymbol{\mu}, \boldsymbol{\Sigma})}{\max\limits_{H_1} L(\boldsymbol{\mu}, \boldsymbol{\Sigma})}$$

leads to the test statistic $T^2 = (\bar{\boldsymbol{y}}-\boldsymbol{\mu_0})'(\frac{\boldsymbol{S}}{n})^{-1}(\bar{\boldsymbol{y}}-\boldsymbol{\mu_0})$, where $\boldsymbol{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{y_i}-\bar{\boldsymbol{y}})(\boldsymbol{y_i}-\bar{\boldsymbol{y}})'$.

iii. what is the distribution of T^2 that was obtained by Hotelling (1931), assuming H_0 is true and sampling is from $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$? What are the parameters the distribution of T^2 is indexed by?

---

8. In a one-way multivariate analysis of variance (MANOVA), we assume that a random sample of p-variate observations is available from each of k multivariate normal populations with equal covariance matrices $\boldsymbol{\Sigma}$. We define sample totals and means as follows:

$$\boldsymbol{y}_{i.} = \sum_{j=1}^{n}\boldsymbol{y}_{ij}, \qquad \boldsymbol{y}_{..} = \sum_{i=1}^{k}\sum_{j=1}^{n}\boldsymbol{y}_{ij},$$
$$\bar{\boldsymbol{y}}_{i.} = \frac{\boldsymbol{y}_{i.}}{n}, \qquad \bar{\boldsymbol{y}}_{..} = \frac{\boldsymbol{y}_{..}}{kn}.$$

To summarize variation in the data, we use “between" and “within" matrices $\boldsymbol{H}$ and $\boldsymbol{E}$, defined as

$$\boldsymbol{H} = n\sum_{i=1}^{k}(\bar{\boldsymbol{y}}_{i.}-\bar{\boldsymbol{y}}_{..})(\bar{\boldsymbol{y}}_{i.}-\bar{\boldsymbol{y}}_{..})'$$
$$\boldsymbol{E} = \sum_{i=1}^{k}\sum_{j=1}^{n}(\boldsymbol{y}_{ij}-\bar{\boldsymbol{y}}_{i.})(\boldsymbol{y}_{ij}-\bar{\boldsymbol{y}}_{i.})'$$

The likelihood ratio test of $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \cdots = \boldsymbol{\mu}_k$ is given by

$$\Lambda = \frac{|\boldsymbol{E}|}{|\boldsymbol{E}+\boldsymbol{H}|}$$

where p=the number of variables (dimension), ν_H=the hypothesis degrees of freedom and ν_E =the error degrees of freedom, which is known as Wilks' Λ. Show that Wilks' Λ can be expressed in terms of the eigenvalues $\lambda_1, \lambda_2, \cdots, \lambda_p$ of $\boldsymbol{E}^{-1}\boldsymbol{H}$

$$\Lambda = \prod_{i=1}^{s}\frac{1}{1+\lambda_i}$$

where the number of nonzero eigenvalues of $\boldsymbol{E}^{-1}\boldsymbol{H}$ is $s = min(p, \nu_H)$.

---

9. In a one-way multivariate analysis of variance (MANOVA), we assume that a random sample of p-variate observations is available from each of k multivariate normal populations with equal covariance matrices $\boldsymbol{\Sigma}$. Consider H_0: $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \cdots = \boldsymbol{\mu}_k$, the Lawley-Hotelling statistic (Lawley 1938, Hotelling 1951) defined as $U^{(s)} = \sum_{i=1}^{s} \lambda_i = tr(\boldsymbol{E^{-1}H})$ can be expressed as a linear combination of Hotelling T^2-statistics so that Lawley-Hotelling statistic is also known as Hotelling's generalized T^2-statistic, where

$$\boldsymbol{H} = n\sum_{i=1}^{k}(\bar{\boldsymbol{y}}_{i\cdot} - \bar{\boldsymbol{y}}_{\cdot\cdot})(\bar{\boldsymbol{y}}_{i\cdot} - \bar{\boldsymbol{y}}_{\cdot\cdot})',$$

$$\boldsymbol{E} = \sum_{i=1}^{k}\sum_{j=1}^{n}(\boldsymbol{y}_{ij} - \bar{\boldsymbol{y}}_{i\cdot})(\boldsymbol{y}_{ij} - \bar{\boldsymbol{y}}_{i\cdot})'.$$

   i. Show that $\boldsymbol{H} = n\sum_{i=1}^{k}(\bar{\boldsymbol{y}}_{i\cdot} - \boldsymbol{\mu})(\bar{\boldsymbol{y}}_{i\cdot} - \boldsymbol{\mu})' - kn(\bar{\boldsymbol{y}}_{\cdot\cdot} - \boldsymbol{\mu})(\bar{\boldsymbol{y}}_{\cdot\cdot} - \boldsymbol{\mu})'$, where $\boldsymbol{\mu}$ is the common value of $\boldsymbol{\mu_1}$, $\boldsymbol{\mu_2}$, $\cdots$, $\boldsymbol{\mu_k}$ under H_0.

   ii. Show that $U^{(s)} = \frac{n}{\nu_E}\sum_{i=1}^{k}(\bar{\boldsymbol{y}}_{i\cdot} - \boldsymbol{\mu})'\boldsymbol{S}_{pl}^{-1}(\bar{\boldsymbol{y}}_{i\cdot} - \boldsymbol{\mu}) - \frac{kn}{\nu_E}(\bar{\boldsymbol{y}}_{\cdot\cdot} - \boldsymbol{\mu})'\boldsymbol{S}_{pl}^{-1}(\bar{\boldsymbol{y}}_{\cdot\cdot} - \boldsymbol{\mu})$, where $\boldsymbol{S}_{pl} = \boldsymbol{E}/\nu_E$. Write the terms on the right side in terms of T^2-statistics.

---

10. If $\boldsymbol{y}_{ij}$, $i = 1, 2, \cdots, k$, $j = 1, 2, \cdots, n$, are independently observed from $N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$, the hypothesis matrix and error matrix for H_0: $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \cdots = \boldsymbol{\mu}_k$ are $\boldsymbol{H}$ and $\boldsymbol{E}$. Show that the maximum value of $\lambda = \frac{\boldsymbol{a'Ha}}{\boldsymbol{a'Ea}}$ and the vector $\boldsymbol{a}$ that produces the maximum are given by the largest eigenvalue λ_1 and the associated eigenvector of $\boldsymbol{E^{-1}H}$, respectively (hint: differentiate λ with respect to $\boldsymbol{a}$ and set the result equal to $\boldsymbol{0}$).

SAS output for Question 4

Model: model1

Dependent Variable: y sale price of the house (in thousands of dollars)

Number of Observations Read	24
Number of Observations Used	24

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	8	701.66438	87.70805	10.33	<.0001
Error	15	127.38187	8.49212		
Corrected Total	23	829.04625			

Root MSE	2.91412	**R-Square**	0.8464
Dependent Mean	34.61250	**Adj R-Sq**	0.7644
Coeff Var	8.41928		

Parameter Estimates							
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Variance Inflation**
Intercept	Intercept	1	12.75192	5.19667	2.45	0.0268	0
x1	taxes (in thousands of dollars)	1	1.72633	0.98816	1.75	0.1011	6.61889
x2	number of baths	1	8.08784	4.03447	2.00	0.0634	2.55561
x3	lot size(in thousands of square feet)	1	0.28738	0.45392	0.63	0.5362	2.15393
x4	living space(in thousands of square feet)	1	2.28954	4.27510	0.54	0.6001	3.77798
x5	number of garage stalls	1	2.20354	1.33560	1.65	0.1198	1.76578
x6	number of rooms	1	0.50740	2.06293	0.25	0.8090	9.02035
x7	number of bedrooms	1	-2.87189	2.82979	-1.01	0.3263	6.91500
x8	age of the home (in years)	1	-0.01681	0.06102	-0.28	0.7867	1.98792

Model: model2

Dependent Variable: y sale price of the house (in thousands of dollars)

Number of Observations Read	24
Number of Observations Used	24

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	6	693.77015	115.62836	14.53	<.0001
Error	17	135.27610	7.95742		
Corrected Total	23	829.04625			

Root MSE	2.82089	**R-Square**	0.8368
Dependent Mean	34.61250	**Adj R-Sq**	0.7792
Coeff Var	8.14992		

Parameter Estimates						
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	Intercept	1	12.94158	5.02453	2.58	0.0196
x1	taxes (in thousands of dollars)	1	2.14748	0.85729	2.50	0.0227
x2	number of baths	1	8.83544	3.57210	2.47	0.0242
x5	number of garage stalls	1	1.98550	1.27411	1.56	0.1376
x6	number of rooms	1	0.66117	1.98627	0.33	0.7433
x7	number of bedrooms	1	-2.71535	2.62127	-1.04	0.3148
x8	age of the home (in years)	1	-0.01859	0.05903	-0.31	0.7566

SAS output for Question 5

vol_est	vol_pred	ratio
0.546669	0.6313279	0.9531395

The CORR Procedure

3 Variables:	y	x1	x6

Simple Statistics							
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
y	15	20.47867	6.98018	307.18000	11.20000	36.50000	y
x1	15	275.04000	124.86596	4126	85.30000	500.00000	x1
x6	15	2.53333	1.12546	38.00000	1.00000	4.00000	x6

Pearson Correlation Coefficients, N = 15 Prob > |r| under H0: Rho=0			
	y	x1	x6
y y	1.00000	-0.85105 <.0001	-0.42115 0.1180
x1 x1	-0.85105 <.0001	1.00000	0.67330 0.0059
x6 x6	-0.42115 0.1180	0.67330 0.0059	1.00000

The REG Procedure
Model: PREDY
Dependent Variable: y y

Number of Observations Read	15
Number of Observations Used	15

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	522.83015	261.41508	19.69	0.0002
Error	12	159.29022	13.27418		
Corrected Total	14	682.12037			

Root MSE	3.64338	R-Square	0.7665
Dependent Mean	20.47867	Adj R-Sq	0.7276
Coeff Var	17.79108		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	32.07476	2.55105	12.57	<.0001
x1	x1	1	-0.05803	0.01055	-5.50	0.0001
x6	x6	1	1.72291	1.17016	1.47	0.1667

Obs	y	PREDY	residual
1	17.00	18.6556	-1.65561
2	18.25	15.1518	3.09824
3	21.47	20.3165	1.15350
4	30.40	29.8974	0.50261
5	16.50	18.6556	-2.15561
6	21.50	25.5973	-4.09731
7	19.70	18.8257	0.87428
8	14.89	13.4328	1.45716
9	16.41	17.0668	-0.65678
10	23.54	22.1155	1.42454
11	21.47	14.6295	6.84052
12	31.90	29.8974	2.00261
13	13.27	12.2722	0.99778
14	13.90	15.1518	-1.25176
15	13.77	18.0753	-4.30530

Applied Statistics Comprehensive Exam

August 2018

Ph.D Day 2 - Exam

This comprehensive exam consists of 10 questions pertaining to two topics of your choice.

Before you start, Please make sure the topics are the one you have chosen.

1 This Ph.D level exam will run from **8:30 AM to 3:30 PM**.

2 Please label each page with your identification number.

DO NOT USE YOUR NAME OR BEAR NUMBER.

3 Please write only on one side of each page.

4 Please leave one inch margins on all sides of each page.

5 Please number all pages consecutively.

6 Please label the day number (Day 1 or Day 2) on each page.

7 Please begin each question on a new page, and number each question.

8 Please do not staple pages together.

9 No electronic devices, formula sheets, or other outside materials are permitted.

10 Statistical tables and paper will be provided.

11 Relax and good luck!

I have read and understand the rules of this exam.

Signature:_____ Date:_____

1. Compare and contrast the Wilcoxon Signed Rank Test and the Wilcoxon Rank Sum Test. Give examples of both tests.

---

2. Describe the procedure behind the Binomial Test and the applications it can be used for.

---

3. What are the non-parametric alternative procedures for the following parametric tests?

   i. Repeated Measures ANOVA

   ii. Two Population Test for the Difference Between Two Means

   iii. Test for Slope in Simple Linear Regression

---

4. What are some alternative applications of the Sign Test in nonparametric statistics? (i.e. name several tests that employ the basic principle of the Sign Test to its procedure).

---

5. Name the parametric tests that the following nonparametric procedures replace.

   i. Mann-Whitney Test

   ii. RxC Median Test

   iii. Sign Test

---

6. The Education Commissioner of Colorado has hired you to estimate the average reading score of kindergartner students in the state of Colorado. Devise a sampling scheme (including how you would determine sample size) and explain your reasoning as to why you chose such a plan. Include costs in your plan.

---

7. Explain in detail how the Adaptive Cluster Sampling procedure works.

---

8. Compare and contrast Two-Stage Sampling and Double Sampling. Give examples of both.

---

9. Describe the conditions within the population when Stratified Sampling works best.

---

10. Explain how an auxiliary variable may be used to develop a Ratio Estimate of the population total.

Applied Statistics Comprehensive Exam

August 2018

Ph.D Day 2 - Exam

This comprehensive exam consists of 10 questions pertaining to two topics of your choice.

Before you start, Please make sure the topics are the one you have chosen.

1 This Ph.D level exam will run from **8:30 AM to 3:30 PM**.

2 Please label each page with your identification number.

DO NOT USE YOUR NAME OR BEAR NUMBER.

3 Please write only on one side of each page.

4 Please leave one inch margins on all sides of each page.

5 Please number all pages consecutively.

6 Please label the day number (Day 1 or Day 2) on each page.

7 Please begin each question on a new page, and number each question.

8 Please do not staple pages together.

9 No electronic devices, formula sheets, or other outside materials are permitted.

10 Statistical tables and paper will be provided.

11 Relax and good luck!

I have read and understand the rules of this exam.

Signature:_____ Date:_____

1. Compare and contrast the Wilcoxon Signed Rank Test and the Wilcoxon Rank Sum Test. Give examples of both tests.

---

2. Describe the procedure behind the Binomial Test and the applications it can be used for.

---

3. What are the non-parametric alternative procedures for the following parametric tests?

   i. Repeated Measures ANOVA

   ii. Two Population Test for the Difference Between Two Means

   iii. Test for Slope in Simple Linear Regression

---

4. What are some alternative applications of the Sign Test in nonparametric statistics? (i.e. name several tests that employ the basic principle of the Sign Test to its procedure).

---

5. Name the parametric tests that the following nonparametric procedures replace.

   i. Mann-Whitney Test

   ii. RxC Median Test

   iii. Sign Test

---

6. Explain how to set up the control limits of an x-bar and R chart in Phase I.

---

7. Compare and contrast the following charts. Give examples of each.

   i. p-chart

   ii. np-chart

   iii. u-chart

---

8. Discuss why multiple univariate x-bar charts are not used to follow p-variables simultaneously in a quality control process.

---

9. Compare and contrast the CUSUM and EWMA charts. Explain how each chart is set up and how the monitoring statistic is defined. Which chart has the better overall average run length performance, if so?

---

10. Discuss how to perform a gage R & R study.

Applied Statistics Comprehensive Exam

August 2018

Ph.D Day 2 - Exam

This comprehensive exam consists of 10 questions pertaining to two topics of your choice.

Before you start, Please make sure the topics are the one you have chosen.

1 This Ph.D level exam will run from **8:30 AM to 3:30 PM**.

2 Please label each page with your identification number.

DO NOT USE YOUR NAME OR BEAR NUMBER.

3 Please write only on one side of each page.

4 Please leave one inch margins on all sides of each page.

5 Please number all pages consecutively.

6 Please label the day number (Day 1 or Day 2) on each page.

7 Please begin each question on a new page, and number each question.

8 Please do not staple pages together.

9 No electronic devices, formula sheets, or other outside materials are permitted.

10 Statistical tables and paper will be provided.

11 Relax and good luck!

I have read and understand the rules of this exam.

Signature:_____ Date:_____

1. Briefly describe the Gibbs sampler in Bayesian analysis.

---

2. if $\pi_m(\theta)$, for $m = 1, \cdots, , M$, are conjugate prior densities for the sampling model $y|\theta$, show that the class of finite mixture prior densities given by

$$\pi(\theta) = \sum_{m=1}^{M} \lambda_i \pi_m(\theta)$$

is also a conjugate class, where the λ_m's are nonnegative weights that sum to 1.

---

3. Let $Y_1, Y_2, \cdots, Y_n$ be a random sample from $N(\mu, \sigma)$, assuming both μ and σ are unknown. Let $\theta = (\mu, \sigma)$.

   i. Find the Jeffreys' prior.

   ii. Find the posterior distribution of $\frac{\sqrt{n}(\mu - \bar{y})}{s}$, where s is the sample standard deviation.

   iii. Use part (b) to find a 95% HPD credible set for μ.

---

4. Suppose we have n independent observations from $Unif(0, \theta)$, $\theta > 0$.

   i. Find a conjugate prior distribution for θ.

   ii. Find the posterior mean and variance for θ.

---

5. A set of n counts $\boldsymbol{Y} = (Y_1, Y_2, \cdots, Y_n)$ are modeled as

$$P(Y_i = 0 | \gamma_i = 0, \pi, \lambda) = 1,$$

$$Y_i | (\gamma_i = 1, \pi, \lambda) \sim \text{Poisson}(\lambda), \quad \text{(independent across } i\text{)}$$

where $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \cdots, \gamma_n)$ is a set of n latent binary counts, $\pi \in [0, 1]$ and $\lambda > 0$ have the hierarchical prior:

$$\gamma_i | (\pi, \lambda) \sim iid \quad \text{Bernoulli}(\pi), i = 1, \cdots, n,$$

$$\pi | \lambda \sim \text{Beta}(c\lambda, 1), \quad \lambda \sim \text{Gamma}(a, b)$$

for some positive constants a, b and c

i. Show that the conditional prior of λ given π is $Gamma(a+1, b - c\log\pi)$.

ii. Write down the posterior conditional probability distributions of $\pi|(\gamma, \lambda, \boldsymbol{y})$, $\lambda|(\boldsymbol{\gamma}, \pi, \boldsymbol{y})$,and $\boldsymbol{\gamma}|(\pi, \lambda, \boldsymbol{y})$, where $\boldsymbol{y}$ is the given data on $\boldsymbol{Y}$. Answer in terms of conditional distributions with explicit formulas for their parameters and with appropriate use of conditional independence.

---

6. If the observation vectors $\boldsymbol{y_1}, \boldsymbol{y_2}, \cdots, \boldsymbol{y_n}$ is a random sample from $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The density function for $\boldsymbol{y} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is

$$\frac{1}{(\sqrt{2\pi})^p|\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})}.$$

Then

i. show that $\sum_{i=1}^n (\boldsymbol{y_i} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y_i} - \boldsymbol{\mu}) = tr(\boldsymbol{\Sigma}^{-1}[\boldsymbol{W} + n(\bar{\boldsymbol{y}} - \boldsymbol{\mu})(\bar{\boldsymbol{y}} - \boldsymbol{\mu})'])$, where $\boldsymbol{W} = \sum_{i=1}^n (\boldsymbol{y_i} - \bar{\boldsymbol{y}})(\boldsymbol{y_i} - \bar{\boldsymbol{y}})'$

ii. show that the maximum likelihood estimator of $\boldsymbol{\mu}$ is $\widehat{\boldsymbol{\mu}} = \bar{\boldsymbol{y}}$;

iii. show that the maximum likelihood estimator of $\boldsymbol{\Sigma}$ is $\widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\boldsymbol{W}$.

---

7. If the observation vectors $\boldsymbol{y_1}, \boldsymbol{y_2}, \cdots, \boldsymbol{y_n}$ is a random sample from $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma}$ unknown. The density function for $\boldsymbol{y} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is

$$\frac{1}{(\sqrt{2\pi})^p|\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})}.$$

Let $L(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ be the likelihood function for the sample. Then for H_0: $\boldsymbol{\mu} = \boldsymbol{\mu}_0$ versus H_1: $\boldsymbol{\mu} \neq \boldsymbol{\mu}_0$,

i. show that $\sum_{i=1}^n (\boldsymbol{y_i} - \boldsymbol{\mu_0})'\widehat{\boldsymbol{\Sigma}}_0^{-1}(\boldsymbol{y_i} - \boldsymbol{\mu_0}) = np$, where $\widehat{\boldsymbol{\Sigma}}_0 = \frac{1}{n}\sum_{i=1}^n (\boldsymbol{y_i} - \boldsymbol{\mu_0})(\boldsymbol{y_i} - \boldsymbol{\mu_0})'$ that maximizes $L(\boldsymbol{\mu}_0, \boldsymbol{\Sigma})$ under H_0.

ii. show that the likelihood ratio

$$LR = \frac{\max\limits_{H_0} L(\boldsymbol{\mu}, \boldsymbol{\Sigma})}{\max\limits_{H_1} L(\boldsymbol{\mu}, \boldsymbol{\Sigma})}$$

leads to the test statistic $T^2 = (\bar{\boldsymbol{y}} - \boldsymbol{\mu_0})'(\frac{\boldsymbol{S}}{n})^{-1}(\bar{\boldsymbol{y}} - \boldsymbol{\mu_0})$, where $\boldsymbol{S} = \frac{1}{n-1}\sum_{i=1}^n (\boldsymbol{y_i} - \bar{\boldsymbol{y}})(\boldsymbol{y_i} - \bar{\boldsymbol{y}})'$.

iii. what is the distribution of T^2 that was obtained by Hotelling (1931), assuming H_0 is true and sampling is from $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$? What are the parameters the distribution of T^2 is indexed by?

---

8. In a one-way multivariate analysis of variance (MANOVA), we assume that a random sample of p-variate observations is available from each of k multivariate normal populations with equal covariance matrices $\boldsymbol{\Sigma}$. We define sample totals and means as follows:

$$\boldsymbol{y}_{i.} = \sum_{j=1}^{n} \boldsymbol{y}_{ij}, \qquad \boldsymbol{y}_{..} = \sum_{i=1}^{k}\sum_{j=1}^{n} \boldsymbol{y}_{ij},$$
$$\bar{\boldsymbol{y}}_{i.} = \frac{\boldsymbol{y}_{i.}}{n}, \qquad \bar{\boldsymbol{y}}_{..} = \frac{\boldsymbol{y}_{..}}{kn}.$$

To summarize variation in the data, we use “between" and “within" matrices $\boldsymbol{H}$ and $\boldsymbol{E}$, defined as

$$\boldsymbol{H} = n\sum_{i=1}^{k}(\bar{\boldsymbol{y}}_{i.} - \bar{\boldsymbol{y}}_{..})(\bar{\boldsymbol{y}}_{i.} - \bar{\boldsymbol{y}}_{..})'$$
$$\boldsymbol{E} = \sum_{i=1}^{k}\sum_{j=1}^{n}(\boldsymbol{y}_{ij} - \bar{\boldsymbol{y}}_{i.})(\boldsymbol{y}_{ij} - \bar{\boldsymbol{y}}_{i.})'$$

The likelihood ratio test of $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \cdots = \boldsymbol{\mu}_k$ is given by

$$\Lambda = \frac{|\boldsymbol{E}|}{|\boldsymbol{E} + \boldsymbol{H}|}$$

where p=the number of variables (dimension), ν_H=the hypothesis degrees of freedom and ν_E =the error degrees of freedom, which is known as Wilks' Λ. Show that Wilks' Λ can be expressed in terms of the eigenvalues $\lambda_1, \lambda_2, \cdots, \lambda_p$ of $\boldsymbol{E}^{-1}\boldsymbol{H}$

$$\Lambda = \prod_{i=1}^{s} \frac{1}{1+\lambda_i}$$

where the number of nonzero eigenvalues of $\boldsymbol{E}^{-1}\boldsymbol{H}$ is $s = min(p, \nu_H)$.

---

9. In a one-way multivariate analysis of variance (MANOVA), we assume that a random sample of p-variate observations is available from each of k multivariate normal populations with equal covariance matrices $\boldsymbol{\Sigma}$. Consider H_0: $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \cdots = \boldsymbol{\mu}_k$, the Lawley-Hotelling statistic (Lawley 1938, Hotelling 1951) defined as $U^{(s)} = \sum_{i=1}^{s} \lambda_i = tr(\boldsymbol{E^{-1}H})$ can be expressed as a linear combination of Hotelling T^2-statistics so that Lawley-Hotelling statistic is also known as Hotelling's generalized T^2-statistic, where

$$\boldsymbol{H} = n\sum_{i=1}^{k}(\bar{\boldsymbol{y}}_{i.} - \bar{\boldsymbol{y}}_{..})(\bar{\boldsymbol{y}}_{i.} - \bar{\boldsymbol{y}}_{..})',$$
$$\boldsymbol{E} = \sum_{i=1}^{k}\sum_{j=1}^{n}(\boldsymbol{y}_{ij} - \bar{\boldsymbol{y}}_{i.})(\boldsymbol{y}_{ij} - \bar{\boldsymbol{y}}_{i.})'.$$

i. Show that $\boldsymbol{H} = n\sum_{i=1}^{k}(\bar{\boldsymbol{y}}_{i\cdot} - \boldsymbol{\mu})(\bar{\boldsymbol{y}}_{i\cdot} - \boldsymbol{\mu})' - kn(\bar{\boldsymbol{y}}_{\cdot\cdot} - \boldsymbol{\mu})(\bar{\boldsymbol{y}}_{\cdot\cdot} - \boldsymbol{\mu})'$, where $\boldsymbol{\mu}$ is the common value of $\boldsymbol{\mu_1}$, $\boldsymbol{\mu_2}$, $\cdots$, $\boldsymbol{\mu_k}$ under H_0.

ii. Show that $U^{(s)} = \frac{n}{\nu_E}\sum_{i=1}^{k}(\bar{\boldsymbol{y}}_{i\cdot} - \boldsymbol{\mu})'\boldsymbol{S}_{pl}^{-1}(\bar{\boldsymbol{y}}_{i\cdot} - \boldsymbol{\mu}) - \frac{kn}{\nu_E}(\bar{\boldsymbol{y}}_{\cdot\cdot} - \boldsymbol{\mu})'\boldsymbol{S}_{pl}^{-1}(\bar{\boldsymbol{y}}_{\cdot\cdot} - \boldsymbol{\mu})$, where $\boldsymbol{S}_{pl} = \boldsymbol{E}/\nu_E$. Write the terms on the right side in terms of T^2-statistics.

---

10. If $\boldsymbol{y}_{ij}$, $i = 1, 2, \cdots, k$, $j = 1, 2, \cdots, n$, are independently observed from $N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$, the hypothesis matrix and error matrix for H_0: $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \cdots = \boldsymbol{\mu}_k$ are $\boldsymbol{H}$ and $\boldsymbol{E}$. Show that the maximum value of $\lambda = \frac{\boldsymbol{a'Ha}}{\boldsymbol{a'Ea}}$ and the vector $\boldsymbol{a}$ that produces the maximum are given by the largest eigenvalue λ_1 and the associated eigenvector of $\boldsymbol{E^{-1}H}$, respectively (hint: differentiate λ with respect to $\boldsymbol{a}$ and set the result equal to $\boldsymbol{0}$).
