STAT 571 Exam #2: Open book & notes. Calculators are ok, but no cell phones, iPads, Kindles, Nooks, etc.

1. Which of the following are linear subspace in $\mathbb{R}^2$? (Circle the letters for the ones that are.) The vectors **a** and **b** are fixed elements of $\mathbb{R}^2$. The vectors are represented as 1×2 row vectors.

(a) span{(1,1)}

Answer: Yes, a span is always a linear subspace.

(b) $\{\mathbf{x} \in \mathbb{R}^2 \mid x_1 > 0\}$

Answer: No, can't have half a space be a linear subspace.

(c) $\{\mathbf{x} \in \mathbb{R}^2 \mid \mathbf{x}\mathbf{a}' = 1\}$

Answer: No, this one doesn't include **0**.

(d) $\{c\mathbf{b} \mid c \in \mathbb{R}\}$

Answer: Yes, this is the same as a span.

(e) $\{c\mathbf{b} \mid c \in \mathbb{R}\} \cap \{\mathbf{x} \in \mathbb{R}^2 \mid \mathbf{x}\mathbf{a}' = 0\}$

Answer: Yes, since an intersection of two linear subspace is also a linear subspace. In this case, since each of the subspaces is a line through 0, their intersection is 0, unless the two subspaces are the same, in which case the intersection is that subspace.

(f) $\{\mathbf{x} \in \mathbb{R}^2 \mid \mathbf{x}\mathbf{a}' = 0\}$

Answer: Yes, this is the line through **0** given by $a_1x_1 + a_2x_2 = 0$.

2. (a) Use Gram-Schmidt on the columns of the matrix below:

$$\begin{pmatrix} 1 & 3 & 0 \\ 1 & 0 & 2 \\ 1 & 0 & -2 \end{pmatrix}$$

Answer: First, dot $(1,1,1)'$ out of $(3,0,0)'$:

$$\mathbf{d}_{2\cdot 1} = \mathbf{d}_2 - \frac{\mathbf{d}_1\mathbf{d}_2'}{\mathbf{d}_1\mathbf{d}_1'}\mathbf{d}_1 = \begin{pmatrix} 3 \\ 0 \\ 0 \end{pmatrix} - \frac{(1,1,1)(3,0,0)'}{(1,1,1)(1,1,1)'}\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 3 \\ 0 \\ 0 \end{pmatrix} - \frac{3}{3}\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 2 \\ -1 \\ -1 \end{pmatrix}.$$

You can stop now, since $(1,1,1)'$, $(2,-1,-1)$, and $(0,2,-2)$ are mutually orthogonal. Or you can proceed, but dotting $(1,1,1)'$ out of $(0,2,-2)'$ gives you back $(0,2,-2)'$, then dotting $(2,-1,-1)'$ out of $(0,2,-2)'$ gives you back $(0,2,-2)'$.

(b) Find an orthogonal basis for the span of the columns of the following matrix:

$$\begin{pmatrix} 1 & -3 & 5 \\ 1 & 0 & 0 \\ 1 & 3 & -5 \end{pmatrix}$$

Answer: You can use Gram-Schmidt again. The **1** vector is orthogonal to the other two, so dotting it out doesn't change anything. Then note that the second and third vectors are proportional, so if you dot the second out of the third, you end up with **0**. Throw that out, and your orthogonal basis is just the first two vectors.

3. Consider the multivariate regression model,

$$\mathbf{Y} = \mathbf{x}\boldsymbol{\beta} + \mathbf{R}, \quad \mathbf{R} \sim N(\mathbf{0}, \mathbf{I}_n \otimes \boldsymbol{\Sigma}_R),$$

where $\mathbf{Y}$ is $n \times q$, and $\mathbf{x}$ is $n \times n$ and invertible.

(a) Show that the least squares estimate of $\boldsymbol{\beta}$ is

$$\widehat{\boldsymbol{\beta}} = \mathbf{x}^{-1}\mathbf{Y}.$$

(You can start with the usual formula for the estimator.)

Answer: There's no $\mathbf{z}$, so

$$\widehat{\boldsymbol{\beta}} = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{Y}.$$

Since $\mathbf{x}$ itself is invertible,

$$(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}' = \mathbf{x}^{-1}(\mathbf{x}')^{-1}\mathbf{x}' = \mathbf{x}^{-1},$$

the last two matrices canceling.

(b) Find $Cov[\widehat{\boldsymbol{\beta}}]$.

Answer: From the usual formula, without $\mathbf{z}$ (or with $\mathbf{z} = \mathbf{I}_q$),

$$Cov[\widehat{\boldsymbol{\beta}}] = C_{\mathbf{x}} \otimes \boldsymbol{\Sigma}_z = (\mathbf{x}'\mathbf{x})^{-1} \otimes \boldsymbol{\Sigma}_R.$$

(c) Find $\mathbf{P}_{\mathbf{x}}$ and $\mathbf{Q}_{\mathbf{x}}$ explicitly. (All the entries in these matrices are integers. What are these integers?) What are $\widehat{\mathbf{Y}}$ and $\widehat{\mathbf{R}}$?

Answer: Using the formula, then writing out the inverses:

$$\mathbf{P}_{\mathbf{x}} = \mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}' = \mathbf{x}\mathbf{x}^{-1}(\mathbf{x}')^{-1}\mathbf{x}' = \mathbf{I}_n.$$

Then

$$\mathbf{Q}_{\mathbf{x}} = \mathbf{I}_n - \mathbf{P}_{\mathbf{x}} = \mathbf{I}_n - \mathbf{I}_n = 0.$$

So $\widehat{\mathbf{Y}} = \mathbf{P}_{\mathbf{x}}\mathbf{Y} = \mathbf{Y}$ and $\widehat{\mathbf{R}} = \mathbf{0}$. Or, use $\widehat{\mathbf{Y}} = \mathbf{x}\widehat{\boldsymbol{\beta}} = \mathbf{x}\mathbf{x}^{-1}\mathbf{Y} = \mathbf{Y}$.

4. In the prostaglandin data set, measurements were taken at six four-hour intervals over the course of a day for 10 individuals. The measurements are prostaglandin contents in their urine. The data matrix $\mathbf{Y}$ is then 10×6. We assume the same sine/cosine curve is fit to the measurements for each man, so that $\mathbf{x} = \mathbf{1}_{10}$, and

$$\mathbf{z} = \begin{pmatrix} 1 & \cos(2\pi(1/6)) & \sin(2\pi(1/6)) \\ 1 & \cos(2\pi(2/6)) & \sin(2\pi(2/6)) \\ 1 & \cos(2\pi(3/6)) & \sin(2\pi(3/6)) \\ 1 & \cos(2\pi(4/6)) & \sin(2\pi(4/6)) \\ 1 & \cos(2\pi(5/6)) & \sin(2\pi(5/6)) \\ 1 & \cos(2\pi(6/6)) & \sin(2\pi(6/6)) \end{pmatrix}.$$

The model is the both-sides model,

$$\mathbf{Y} = \mathbf{x}\boldsymbol{\beta}\mathbf{z}' + \mathbf{R}, \quad \mathbf{R} \sim N(\mathbf{0}, \mathbf{I}_n \otimes \boldsymbol{\Sigma}_R).$$

The least squares estimate of $\boldsymbol{\beta}$ is

$$\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1 \;\; \widehat{\beta}_2 \;\; \widehat{\beta}_3) = (188.5 \;\; -53.18 \;\; 4.186),$$

and the estimate of $\boldsymbol{\Sigma}_{\mathbf{z}}$ is

$$\widehat{\boldsymbol{\Sigma}}_{\mathbf{z}} = \begin{pmatrix} 2076 & -454 & 273 \\ -454 & 1283 & 235 \\ 273 & 235 & 281 \end{pmatrix}.$$

(a) Show that $\mathbf{C_x} = \frac{1}{10}$.

Answer: $\mathbf{C_x} = (\mathbf{x}'\mathbf{x})^{-1}$, and $\mathbf{x} = \mathbf{1}_{10}$, so

$$\mathbf{C} - \mathbf{x} = (\mathbf{1}_{10}'\mathbf{1}_{10})^{-1} = (10)^{-1} = \frac{1}{10}.$$

(b) Find the estimated standard error of $\widehat{\beta}_2$. Is the $\widehat{\beta}_2$ statistically significant?

Answer: Here,

$$Cov[\widehat{\boldsymbol{\beta}}] = (\mathbf{x}'\mathbf{x})^{-1} \otimes \boldsymbol{\Sigma}_{\mathbf{z}} = \frac{1}{10}\boldsymbol{\Sigma}_{\mathbf{z}},$$

so that variances of $\widehat{\beta}_j$ is the j^{th} diagonal of that, $\sigma_{\mathbf{z}jj}/10$. The estimates of the variances of the three $\widehat{\beta}_j$'s are, respectively, $207.6, 128.3,$ and 28.1. Square-root those to get the se's:

$$se(\widehat{\beta}_2) = \sqrt{128.3} = 11.33, \; se(\widehat{\beta}_3) = \sqrt{28.1} = 5.30.$$

The t's are

$$\widehat{\beta}_2/se(\widehat{\beta}_2) = -53.18/11.33 = -4.69 \;\text{ and }\; \widehat{\beta}_3/se(\widehat{\beta}_3) = 4.186/28.1 = 0.79.$$

So $\widehat{\beta}_2$ is statistically significant, and $\widehat{\beta}_3$ is not. (You had to do only one.)

(c) Consider testing the null hypothesis that $\beta_2 = \beta_3 = 0$. The $T^2 = 30.44$. Find the F version of the statistic. What are its degrees of freedom? Do you reject the null hypothesis? (You can use 4 as the cutoff point.)

Answer: You need the various dimensions. So $\mathbf{x}$ is $n \times p = 10 \times 1$, $\mathbf{Y}$ is $n \times q = 10 \times 6$, and $\boldsymbol{\beta}^* = (\beta_2 \;\; \beta_3)$ is $p^* \times l^* = 1 \times 2$. This means that $\nu = n - p = 9$, so that

$$F = \frac{\nu - l^* + 1}{\nu p^* l^*}\, T^2 = \frac{9 - 2 + 1}{9 \cdot 1 \cdot 2}\, T^2 = \frac{4}{9}\,(30.44) = 13.53.$$

Under the null, the F is $F_{p^*l^*, \nu - l^* + 1} = F_{2,8}$, so the $df = (2, 8)$. You can reject the null.

5. Consider the mouth-size data, where the $\mathbf{x} = (\mathbf{1}_n, \mathbf{g})$, where $\mathbf{g}$ is the vector with 1's for the girls and 0's for the boys, and the $\mathbf{z}$ has the constant, linear, and quadratic orthogonal polynomial vectors for the growth curves. The $\mathbf{Y}$ is 27×4, so

$$\mathbf{Y} = \mathbf{x}\beta\mathbf{z}' + \mathbf{R} = \begin{pmatrix} \mathbf{1}_{11} & \mathbf{1}_{11} \\ \mathbf{1}_{16} & \mathbf{0}_{16} \end{pmatrix} \begin{pmatrix} \beta_1 & \beta_2 & \beta_3 \\ \delta_1 & \delta_2 & \delta_3 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 & 1 \\ -3 & -1 & 1 & 3 \\ 1 & -1 & -1 & 1 \end{pmatrix} + \mathbf{R}.$$

(a) Three submodels are considered in the table below. The table has the patterns, deviances, and BIC-based estimates of the probabilities of the models. You are to fill in the dimensions, and the BIC for the first model. (You don't have to fill in the other two BIC's.)

Pattern	$\begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}$	$\begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \end{pmatrix}$	$\begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}$
Deviance	228	221	218
Dimension			
BIC		***	***
$\widehat{Prob}$	11.86	75.59	12.55

Answer: The dimensions are the number of parameters in the $\mathbf{\Sigma}$, which is $4 \times 5/2 = 10$, plus the non-zero β_{ij}'s, which is 3, 4, and 6 for the three models. Since the number of $\mathbf{\Sigma}$ parameters is the same for the three models, you could also just use the $3, 4, 6$. Then the BIC = $deviance + \log(n)(dimension)$, where $n = 27$. The complete table is

Pattern	$\begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}$	$\begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \end{pmatrix}$	$\begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}$
Deviance	228	221	218
Dimension	13	14	16
BIC	270.85	267.14	270.73
$\widehat{Prob}$	11.86	75.59	12.55

(b) From the table in part (a), find the estimated probability that in the true model, the boys' and girls' curves are not parallel.

Answer: The second row in $\boldsymbol{\beta}$ gives the differences between the boys and girls for the intercept, slope, and quadratic terms, respectively. Thus the curves are not parallel if δ_2 or δ_3 are not 0. So add the probabilities of the second and third models, to get 75.59+12.55 = 88.14%.

Find the estimated probability that the curves for both the boys and girls are actually straight lines.

Answer: They are both straight lines as long as there are no quadratic terms, so you leave out the third model: 100 - 12.55 = 87.45%.

6. The South African Heart Disease Data contains various health measurements on $n = 462$ men. We start by focussing on four variables: two blood measurements, blood pressure and bad cholesterol, and two heaviness measurements, BMI and fat percentage.

(a) The covariance matrix is estimated from the multivariate regression model which has the constant and the age variable in the design matrix. What are the degrees of freedom for the estimated covariance matrix from this model?

Answer: $n = 462$, and there are $p = 2$ vectors in the $\mathbf{x}$-matrix (the one vector and the age vector), so $\nu = n - p = 460$.

(b) Consider testing the null hypothesis that the blood measurements are independent of the heaviness measurements. The log determinants of the estimated covariance matrices are given in the table:

Variables →	All four	Just the blood measurements	Just the heaviness measurements
Log determinant →	12.8	7.23	5.68

What is the value of the asymptotic χ^2 test statistic? What are the degrees of freedom for the χ^2? Do you accept or reject the null hypothesis?

Answer: The statistic is

$$\nu(\log(|\widehat{\boldsymbol{\Sigma}}_{11}|) + \log(|\widehat{\boldsymbol{\Sigma}}_{22}|) - \log(|\widehat{\boldsymbol{\Sigma}}|)) = 460 \times (7.23 + 5.68 - 12.8) = 50.6.$$

The degrees of freedom in the χ^2 is the number of parameters set to 0 in the null hypothesis. Since we are testing the independence of two sets of two variables, the df = 4. (Or you can use the dimension of the alternative, $4 \times 5/2 = 10$, minus the dimension of the null, which has two $2 \times 3/2$'s, i.e., 6, so df = 10-6=4.) The statistic is way higher than the χ^2 in the table (9.49), so reject the null hypothesis.

(c) Now consider the conditional independence of the blood measurements and heaviness measurements, **conditional** on the two lifestyle measurements, alcohol and tobacco. The estimated conditional covariance matrix now has how many degrees of freedom?

Answer: We are conditioning on two variables, so we lose two df, to get $460 - 2 = 458$.

What are the degrees of freedom for the χ^2 statistic now?

Answer: Still 4, the number of conditional covariances set to 0.

The statistic here is 66.2. Do you accept the null hypothesis that the blood measurements and heaviness measurements are conditionally independent given the lifestyle measurements?

Reject again.

7. The skulls data has 4 measurements on each of 150 skulls. The skulls were collected over five time periods. In this question, the analysis will be based on the residuals from the model that fits the constant and the linear term in time period. We wish to consider the one-factor model and the model that the four variables are mutually independent.

(a) The table below has the deviances for the one-factor model and the unrestricted model, basing the deviances on the **correlation** matrices. Fill in the deviance for the first model, and the dimensions for all three models.

Model	Mutual Independence	One-factor	Unrestricted
Deviance		-11.53	-14.26
Dimension			
BIC	19.99	28.44	35.71

Answer: Under mutual independence, the correlation matrix is $\mathbf{I}_4$, so $\log(|\mathbf{I}_4|) = \log(1) = 0$. The dimension for mutual independence is 4, since the $\boldsymbol{\Sigma}$ is diagonal. For the one-factor model, the dimension is $pq - \binom{p}{2} + q$. Since $p = 1$ factor and $q = 4$, we get 4+4=8. ($\binom{1}{2} = 0$.) For the unrestricted model, the dimension is that for the $\boldsymbol{\Sigma}$, which is $4 \times 5/2 = 10$. The table:

Model	Mutual Independence	One-factor	Unrestricted
Deviance	0	-11.53	-14.26
Dimension	4	8	10
BIC	19.99	28.44	35.71

(b) Based on the BIC's, which model is best? Do either of the other models have a substantial probability (say over 10%)?

Answer: BIC-wise, the best model is that of mutual independence. Neither of the other two models has much of a probability, since they are at least 9 BIC points worse. Calculating, the probabilities are 98.52%, 1.44%, and 0.04%.

(c) Does the mutual independence model fit at the 5% level? Does the one-factor model fit?

Answer: To test the fit of a model, take the difference between the deviances of the model and the unrestricted model, and compare that to a χ^2 with df equal to the difference in the dimensions. So for mutual independence, the statistic is 14.26 on $10 - 4 = 6$ df. The $\chi^2_{6,0.05} = 12.59$, so we reject: The model doesn't fit.

For the one-factor model, the statistic is $14.26 - 11.53 = 2.73$, on $10 - 8 = 2$ df. That is not significant, so the one-factor model fits ok.