Exam 3 review. 7 sheets of notes and a calculator. Wednesday, May 1.

Types of problems.

See the discriminant analysis problems from exam 2 review, especially 75)-84).

86) For PCA, a *biplot* is a plot of the first principal component versus the second principal component. The plotted points are $\hat{\boldsymbol{e}}_j^T \boldsymbol{x}_i$ for $j = 1, 2$ where the classical biplot uses $i = 1, ..., n$ and the robust plot uses cases in the RMVN set U. Let $\hat{\boldsymbol{e}}_j = (\hat{e}_{1j}, \hat{e}_{2j}, ..., \hat{e}_{pj})^T$. Then $\hat{e}_{kj}$ is called the *loading* of the kth variable on the jth principal component. An arrow with the kth variable name is the vector from the origin $(0, 0)^T$ to the loadings $(\hat{e}_{k1}, \hat{e}_{k2})^T$. So if the arrow is in the first quadrant, both loadings are positive, etc. If the arrow is long to the right but short down, then the loading with the first principal component is large and positive while the loading with the second principal component is small and negative. Be able to interpret the classical and robust biplots, as in HW8 B and Q8.

87) The one sample Hotelling's T^2 test is used to test $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ versus $H_A : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$. The test rejects H_0 if $T_H^2 = n(\overline{\boldsymbol{x}} - \boldsymbol{\mu}_0)^T \boldsymbol{S}^{-1} (\overline{\boldsymbol{x}} - \boldsymbol{\mu}_0) > \dfrac{(n-1)p}{n-p} F_{p,n-p,1-\alpha}$ where $P(Y \leq F_{p,d,\alpha}) = \alpha$ if $Y \sim F_{p,d}$.

If a multivariate location estimator T satisfies $\sqrt{n}(T - \boldsymbol{\mu}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{D})$, then a competing test rejects H_0 if $T_C^2 = n(T - \boldsymbol{\mu}_0)^T \hat{\boldsymbol{D}}^{-1} (T - \boldsymbol{\mu}_0) > \dfrac{(n-1)p}{n-p} F_{p,n-p,1-\alpha}$ if H_0 holds and $\hat{\boldsymbol{D}}$ is a consistent estimator of $\boldsymbol{D}$. The scaled F cutoff can be used since $T_C^2 \xrightarrow{D} \chi_p^2$ if H_0 holds, and $\dfrac{(n-1)p}{n-p} F_{p,n-p,1-\alpha} \to \chi_{p,1-\alpha}^2$ as $n \to \infty$.

88) Let pval be an estimate of the pvalue. As a benchmark for hypothesis testing, use $\alpha = 0.05$ if α is not given.

89) Typically use $T_C^2 = T_H^2$ in the following 4 step **one sample Hotelling's T_C^2 test**.
i) State the hypotheses $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0 \quad H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$.
ii) Find the test statistic $T_C^2 = n(T - \boldsymbol{\mu}_0)^T \hat{\boldsymbol{D}}^{-1} (T - \boldsymbol{\mu}_0)$.
iii) Find pval =

$$P\left(\frac{n-p}{(n-1)p} \; T_C^2 < F_{p,n-p}\right).$$

iv) State whether you fail to reject H_0 or reject H_0. If you reject H_0 then conclude that $\boldsymbol{\mu} \neq \boldsymbol{\mu}_0$ while if you fail to reject H_0 conclude that the population mean $\boldsymbol{\mu} = \boldsymbol{\mu}_0$ or that there is not enough evidence to conclude that $\boldsymbol{\mu} \neq \boldsymbol{\mu}_0$. Reject H_0 if pval $< \alpha$ and fail to reject H_0 if pval $\geq \alpha$.

90) The multivariate matched pairs test is used when there are $k = 2$ treatments applied to the same n cases with the same p variables used for each treatment. Let $\boldsymbol{y}_i$ be the p variables measured for treatment 1 and $\boldsymbol{z}_i$ be the p variables measured for treatment 2. Let $\boldsymbol{x}_i = \boldsymbol{y}_i - \boldsymbol{z}_i$. Let $\boldsymbol{\mu} = E(\boldsymbol{x}) = E(\boldsymbol{y}) - E(\boldsymbol{z})$. Want to test if $\boldsymbol{\mu} = \boldsymbol{0}$, so $E(\boldsymbol{y}) = E(\boldsymbol{z})$. The test can also be used if $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ are matched (highly dependent) in some way. For example if identical twins are in the study, $\boldsymbol{x}_i$ and $\boldsymbol{y}_i$ could be the measurements on each twin. Let $(\overline{\boldsymbol{x}}, \boldsymbol{S}_x)$ be the sample mean and covariance matrix of

the $\boldsymbol{x}_i$.

91) The **large sample multivariate matched pairs test** has 4 steps.
i) State the hypotheses $H_0 : \boldsymbol{\mu} = \mathbf{0}$ $\quad H_1 : \boldsymbol{\mu} \neq \mathbf{0}$.
ii) Find the test statistic $T_M^2 = n\overline{\boldsymbol{x}}^T \boldsymbol{S}_x^{-1} \overline{\boldsymbol{x}}$.
iii) Find pval =

$$P\left(\frac{n-p}{(n-1)p}\ T_M^2 < F_{p,n-p}\right).$$

iv) State whether you fail to reject H_0 or reject H_0. If you reject H_0 then conclude that $\boldsymbol{\mu} \neq \mathbf{0}$ while if you fail to reject H_0 conclude that the population mean $\boldsymbol{\mu} = \mathbf{0}$ or that there is not enough evidence to conclude that $\boldsymbol{\mu} \neq \mathbf{0}$. Reject H_0 if pval $< \alpha$ and fail to reject H_0 if pval $\geq \alpha$. Give a nontechnical sentence if possible.

92) Repeated measurements = longitudinal data analysis. Take p measurements on the same unit, often the same measurement, eg blood pressure, at several time periods. The variables are X_1, ..., X_p where often X_k is the measurement at the kth time period. The $E(\boldsymbol{x}) = (\mu_1, ..., \mu_p)^T = (\mu + \tau_1, ..., \mu + \tau_p)^T$. Let $y_{ij} = x_{ij} - x_{i,j+1}$ for $i = 1, ..., n$ and $j = 1, ..., p-1$. Then $\overline{\boldsymbol{y}} = (\overline{x}_1 - \overline{x}_2, \overline{x}_2 - \overline{x}_3, ..., \overline{x}_{p-1} - \overline{x}_p)^T$. If $\boldsymbol{\mu}_Y = E(\boldsymbol{y}_i)$, then $\boldsymbol{\mu}_Y = \mathbf{0}$ is equivalent to $\mu_1 = \cdots = \mu_p$ where $E(X_k) = \mu_k$. Let $\boldsymbol{S}_y$ be the sample covariance matrix of the $\boldsymbol{y}_i$.

93) The **large sample repeated measurements test** has 4 steps.
i) State the hypotheses $H_0 : \boldsymbol{\mu}_y = \mathbf{0}$ $\quad H_1 : \boldsymbol{\mu}_y \neq \mathbf{0}$.
ii) Find the test statistic $T_R^2 = n\overline{\boldsymbol{y}}^T \boldsymbol{S}_y^{-1} \overline{\boldsymbol{y}}$.
iii) Find pval =

$$P\left(\frac{n-p+1}{(n-1)(p-1)}\ T_R^2 < F_{p-1,n-p+1}\right).$$

iv) State whether you fail to reject H_0 or reject H_0. If you reject H_0 then conclude that $\boldsymbol{\mu}_y \neq \mathbf{0}$ while if you fail to reject H_0 conclude that the population mean $\boldsymbol{\mu}_y = \mathbf{0}$ or that there is not enough evidence to conclude that $\boldsymbol{\mu}_y \neq \mathbf{0}$. Reject H_0 if pval $< \alpha$ and fail to reject H_0 if pval $\geq \alpha$. Give a nontechnical sentence, if possible.

94) The F tables give left tail area and the pval is a right tail area. Table 15.5 gives $F_{k,d,0.95}$. If $\alpha = 0.05$ and $\dfrac{n-p}{(n-1)p}\ T_C^2 < F_{k,d,0.95}$, then fail to reject H_0. If $\dfrac{n-p}{(n-1)p}\ T_C^2 \geq F_{k,d,0.95}$ then reject H_0.

a) For the one sample Hotelling's T_C^2 test, and the matched pairs T_M^2 test, $k = p$ and $d = n - p$. See HW8 D) and Q8.

b) For the repeated measures T_R^2 test, $k = p - 1$ and $d = n - p + 1$. See HW8 E) and Q8.

95) If $n > 10p$, the tests in 89), 91) and 93) are robust to nonnormality. For the one sample Hotelling's T_C^2 test and the repeated measurements test, make a DD plot. For the multivariate matched pairs test, make a DD plot of the $\boldsymbol{x}_i$, of the $\boldsymbol{y}_i$ and of the $\boldsymbol{z}_i$.

96) Suppose there are two independent random samples $X_{1,1}, ..., X_{n_1,1}$ and $X_{1,2}, ..., X_{n_2,2}$ from populations with mean and covariance matrices $(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_{\boldsymbol{x}_i})$ for $i = 1, 2$ where the $\boldsymbol{\mu}_i$ are $p \times 1$ vectors. Let $d_n = \min(n_1 - p, n_2 - p)$. The

large sample two sample Hotelling's T_0^2 test is a 4 step test:
i) State the hypotheses $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 \quad H_1 : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$.
ii) Find the test statistic $t_0 = T_0^2/p$.
iii) Find pval $= P(t_0 < F_{p,d_n})$.
iv) State whether you fail to reject H_0 or reject H_0. If you reject H_0 then conclude that the population means are not equal while if you fail to reject H_0 conclude that the population means are equal or that there is not enough evidence to conclude that the population means differ. Reject H_0 if pval $< \alpha$ and fail to reject H_0 if pval $\geq \alpha$. Give a nontechnical sentence if possible.

97) Tests for covariance matrices are very nonrobust to nonnormality. Let a plot of x versus y have x on the horizontal axis and y on the vertical axis. A good diagnostic is to use the DD plot. So a diagnostic for $H_0 : \boldsymbol{\Sigma}_{\boldsymbol{x}} = \boldsymbol{\Sigma}_0$ is to plot $D_i(\overline{\boldsymbol{x}}, \boldsymbol{S})$ versus $D_i(\overline{\boldsymbol{x}}, \boldsymbol{\Sigma}_0)$ for $i = 1, ..., n$. If $n > 10p$ and H_0 is true, then the plotted points in the DD plot should cluster tightly about the identity line.

98) A test for sphericity is a test of $H_0 : \boldsymbol{\Sigma}_{\boldsymbol{x}} = d\boldsymbol{I}_p$ for some unknown constant $d > 0$. As a diagnostic, make a "DD plot" of $D_i^2(\overline{\boldsymbol{x}}, \boldsymbol{S})$ versus $D_i^2(\overline{\boldsymbol{x}}, \boldsymbol{I}_p)$. If $n > 10p$ and H_0 is true, then the plotted points in the "DD plot" should cluster tightly about the line through the origin with slope d.

99) Now suppose there are k samples, and want to test $H_0 : \boldsymbol{\Sigma}_{\boldsymbol{x}_1} = \cdots = \boldsymbol{\Sigma}_{\boldsymbol{x}_k}$, that is, all k populations have the same covariance matrix. As a diagnostic, make a DD plot of $D_i(\overline{\boldsymbol{x}}_j, \boldsymbol{S}_j)$ versus $D_i(\overline{\boldsymbol{x}}_j, \boldsymbol{S}_{pool})$ for $j = 1, ..., k$ and $i = 1, ..., n_i$.

100) The **multivariate linear model** $\boldsymbol{y}_i = \boldsymbol{B}^T\boldsymbol{x}_i + \boldsymbol{\epsilon}_i$ for $i = 1, ..., n$ has $m \geq 2$ response variables $Y_1, ..., Y_m$ and p predictor variables $X_1, X_2, ..., X_p$. The ith case is $(\boldsymbol{x}_i^T, \boldsymbol{y}_i^T) = (x_{i1}, x_{i2}, ..., x_{ip}, Y_{i1}, ..., Y_{im})$. If a constant $x_{i1} = 1$ is in the model, then x_{i1} could be omitted from the case. The model is written in matrix form as $\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{E}$. The model has $E(\boldsymbol{\epsilon}_k) = \boldsymbol{0}$ and $\text{Cov}(\boldsymbol{\epsilon}_k) = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} = ((\sigma_{ij}))$ for $k = 1, ..., n$. Also $E(\boldsymbol{e}_i) = \boldsymbol{0}$ while $\text{Cov}(\boldsymbol{e}_i, \boldsymbol{e}_j) = \sigma_{ij}\boldsymbol{I}_n$ for $i, j = 1, ..., m$. Then $\boldsymbol{B}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ are unknown matrices of parameters to be estimated, and $E(\boldsymbol{Z}) = \boldsymbol{X}\boldsymbol{B}$ while $E(Y_{ij}) = \boldsymbol{x}_i^T\boldsymbol{\beta}_j$.

The data matrix $\boldsymbol{W} = [\boldsymbol{X} \;\; \boldsymbol{Y}]$ except usually the first column $\boldsymbol{1}$ of $\boldsymbol{X}$ is omitted if $X_1 = 1$. The $n \times m$ matrix

$$\boldsymbol{Z} = \begin{bmatrix} Y_{1,1} & Y_{1,2} & \dots & Y_{1,m} \\ Y_{2,1} & Y_{2,2} & \dots & Y_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{n,1} & Y_{n,2} & \dots & Y_{n,m} \end{bmatrix} = \begin{bmatrix} \boldsymbol{Y}_1 & \boldsymbol{Y}_2 & \dots & \boldsymbol{Y}_m \end{bmatrix} = \begin{bmatrix} \boldsymbol{y}_1^T \\ \vdots \\ \boldsymbol{y}_n^T \end{bmatrix}.$$

The $n \times p$ matrix

$$\boldsymbol{X} = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,p} \\ x_{2,1} & x_{2,2} & \dots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \dots & x_{n,p} \end{bmatrix} = \begin{bmatrix} \boldsymbol{v}_1 & \boldsymbol{v}_2 & \dots & \boldsymbol{v}_p \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_n^T \end{bmatrix}$$

where often $\boldsymbol{v}_1 = \boldsymbol{1}$.

The $p \times m$ matrix

$$\boldsymbol{B} = \begin{bmatrix} \beta_{1,1} & \beta_{1,2} & \cdots & \beta_{1,m} \\ \beta_{2,1} & \beta_{2,2} & \cdots & \beta_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{p,1} & \beta_{p,2} & \cdots & \beta_{p,m} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\beta}_1 & \boldsymbol{\beta}_2 & \cdots & \boldsymbol{\beta}_m \end{bmatrix}.$$

The $n \times m$ matrix

$$\boldsymbol{E} = \begin{bmatrix} \epsilon_{1,1} & \epsilon_{1,2} & \cdots & \epsilon_{1,m} \\ \epsilon_{2,1} & \epsilon_{2,2} & \cdots & \epsilon_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ \epsilon_{n,1} & \epsilon_{n,2} & \cdots & \epsilon_{n,m} \end{bmatrix} = \begin{bmatrix} \boldsymbol{e}_1 & \boldsymbol{e}_2 & \cdots & \boldsymbol{e}_m \end{bmatrix} = \begin{bmatrix} \boldsymbol{\epsilon}_1^T \\ \vdots \\ \boldsymbol{\epsilon}_n^T \end{bmatrix}.$$

Warning: The $\boldsymbol{e}_i$ are error vectors, not orthonormal eigenvectors.

101) The univariate linear model is $Y_i = x_{i,1}\beta_1 + x_{i,2}\beta_2 + \cdots + x_{i,p}\beta_p + e_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i = \boldsymbol{\beta}^T\boldsymbol{x}_i + e_i$ for $i = 1, \ldots, n$. In matrix notation, these n equations become $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$, where $\boldsymbol{Y}$ is an $n \times 1$ vector of response variables, $\boldsymbol{X}$ is an $n \times p$ matrix of predictors, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown coefficients, and $\boldsymbol{e}$ is an $n \times 1$ vector of unknown errors.

102) Each response variable in a multivariate linear model follows a univariate linear model $\boldsymbol{Y}_j = \boldsymbol{X}\boldsymbol{\beta}_j + \boldsymbol{e}_j$ for $j = 1, ..., m$ where it is assumed that $E(\boldsymbol{e}_j) = \boldsymbol{0}$ and $\text{Cov}(\boldsymbol{e}_j) = \sigma_{jj}\boldsymbol{I}_n$.

103) The one way MANOVA model is a generalization of the Hotelling's T^2 test from 2 groups to $p \geq 2$ groups, assumed to have different means but a common covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$. Want to test $H_0 : \boldsymbol{\mu}_1 = \cdots = \boldsymbol{\mu}_p$. This model is a multivariate linear model so there are m response variables $Y_1, ..., Y_m$ measured for each group. Each Y_i follows a one way ANOVA model for $i = 1, ..., m$.

104) For the one way MANOVA model, make a DD plot of the residuals $\hat{\boldsymbol{\epsilon}}_i$ where $i = 1, ..., n$. Use the plot to check whether the $\boldsymbol{\epsilon}_i$ follow a multivariate normal distribution or some other elliptically contoured distribution. Want $n > 10p$.

105) For the one way MANOVA model, write the data as Y_{ijk} where $i = 1, ..., p$ and $j = 1, ..., n_i$. So k corresponds to the kth variable Y_k for $k = 1, ..., m$. Then $\hat{Y}_{ijk} = \hat{\mu}_{ik} = \overline{Y}_{i0k}$ for $i = 1, ..., p$. So for the kth variable, mean $\mu_{1k}, ..., \mu_{pk}$ are of interest. The residuals are $r_{ijk} = Y_{ijk} - \hat{Y}_{ijk}$. For each variable Y_k make a response plot of $\overline{Y}_{i0k}$ versus Y_{ijk} and a residual plot of $\overline{Y}_{i0k}$ versus r_{ijk}. Both plots will consist of p dot plots of n_k cases located at the $\overline{Y}_{i0k}$. The dot plots should follow the identity line in the response plot and the horizontal $r = 0$ line in the residual plot for each of the m response variables $Y_1, ..., Y_m$. For each variable Y_k, let R_{ik} be the range of the ith dot plot. If each $n_i \geq 5$, want $\max(R_{1k}, ..., R_{pk}) \leq 2\min(R_{1k}, ..., R_{pk})$. The one way MANOVA model may be reasonable if the m response and residual plots satisfy the above graphical checks.

106) The four steps of the one way MANOVA test follow.
i) State the hypotheses $H_0 : \boldsymbol{\mu}_1 = \cdots = \boldsymbol{\mu}_p$ and H_1 : not H_0.
ii) Get t_0 from output.
iii) Get pval from output.
iv) State whether you reject H_0 or fail to reject H_0. If pval $\leq \alpha$, reject H_0 and conclude that not all of the p treatment means are equal. If pval $> \alpha$, fail to reject H_0 and conclude that all p treatment means are equal or that there is not enough evidence to conclude that not all of the p treatment means are equal. Give a nontechnical sentence as the conclusion, if possible.

107) The one way MANOVA test assumes that $\boldsymbol{\Sigma}_{\boldsymbol{x}_1} = \cdots = \boldsymbol{\Sigma}_{\boldsymbol{x}_p}$, but has some resistance to this assumption. See point 105).

108) Know how to use randomization to assign units to treatment groups with the *R/Splus* function `sample` that is used to draw a random permutation of $\{1, 2, ..., n\}$. If the units are $a_1, ..., a_9$ and the `sample(9)` command gives 6 7 9 5 1 4 2 8 3, then a_6, a_7 and a_9 are assigned treatment 1, a_5, a_1 and a_4 are assigned treatment 2, and a_2, a_8 and a_3 are assigned treatment 3.

109) Factor analysis is use to write $\hat{\boldsymbol{\Sigma}} \approx \hat{\boldsymbol{L}}\hat{\boldsymbol{L}}^T + \hat{\boldsymbol{\Psi}} = \hat{\boldsymbol{\Sigma}}_F$. Factor analysis clusters variables into groups called factors and suggests that the $m < p$ factors explain the dispersion more simply than $X_1, ..., X_p$. $\hat{\boldsymbol{L}} = [\boldsymbol{L}_1, ..., \boldsymbol{L}_m]$ is the matrix of factor loadings.

110) Factor analysis output is a lot like PCA output, but replace PC1, ..., PCp by Factor 1, ..., Factor m:

Factor 1	Factor 2	$\cdots$	Factor m
$\hat{\boldsymbol{L}}_1$	$\hat{\boldsymbol{L}}_2$	$\cdots$	$\hat{\boldsymbol{L}}_m$

111) To try to explain Factor j, look at entries in $\hat{\boldsymbol{L}}_j$ that are large in magnitude and ignore entries close to zero. Sometimes only one entry is large. Sometimes all of the large entries have approximately the same size and sign, then the Factor is interpreted as an average of these entrees. If all of the large entries have approximately the same size but different signs then the Factor is interpreted as the sum of the variables with the positive sign − the sum of the variables with a minus sign. Thus if exactly two entries are of similar large magnitude but of different sign, the Factor is interpreted as a difference of the two entrees. If there are $k \geq 2$ large entrees that differ in magnitude, then the Factor is interpreted as a linear combination of the corresponding variables.

112) The proportion of variance explained and cumulative proportion of variance explained are interpreted as for PCA. Use the k factor model if the proportion of the variance explained by the first k Factors is larger than some percentage such as 50%, 60%, 70%, 80% or 90%.

113) For a k factor model, want the degrees of freedom $d \geq 0$ where
$d = 0.5(p-k)^2 - 0.5(p+k)$.

114) If the 1 factor model is not adequate, R will give a test for whether a k factor model is sufficient. A k factor model with $pval < 0.05$ is not sufficient: more factors are needed. A k factor model with $pval > 0.05$ is sufficient.

115) Let $\hat{\boldsymbol{\Gamma}}$ be an orthogonal matrix. The $\hat{\boldsymbol{L}}_\Gamma\hat{\boldsymbol{L}}_\Gamma^T = \hat{\boldsymbol{L}}\hat{\boldsymbol{\Gamma}}\hat{\boldsymbol{\Gamma}}^T\hat{\boldsymbol{L}}^T = \hat{\boldsymbol{L}}\hat{\boldsymbol{L}}^T$. The varimax and promax rotations seek $\hat{\boldsymbol{\Gamma}}$ such that $\hat{\boldsymbol{L}}_\Gamma = \hat{\boldsymbol{L}}\hat{\boldsymbol{\Gamma}}$ has loadings that are easier to interpret than the loadings of $\hat{\boldsymbol{L}}$. The promax rotation attempts to produce loading with a lot of zeroes.

116) The multivariate linear regression model is a special case of the multivariate linear model where at least one predictor variable X_j is continuous. The MANOVA model is a multivariate linear model where all of the predictors are categorical variables so the X_j are coded and are often indicator variables.

117) The **multivariate linear regression model** $\boldsymbol{y}_i = \boldsymbol{B}^T\boldsymbol{x}_i + \boldsymbol{\epsilon}_i$ for $i = 1, ..., n$ has $m \geq 2$ response variables $Y_1, ..., Y_m$ and p predictor variables $X_1, X_2, ..., X_p$. The ith case is $(\boldsymbol{x}_i^T, \boldsymbol{y}_i^T) = (x_{i1}, x_{i2}, ..., x_{ip}, Y_{i1}, ..., Y_{im})$. The constant $x_{i1} = 1$ is in the model, and is often omitted from the case and the data matrix. The model is written in matrix form as $\boldsymbol{Z} = \boldsymbol{XB} + \boldsymbol{E}$. The model has $E(\boldsymbol{\epsilon}_k) = \boldsymbol{0}$ and $\text{Cov}(\boldsymbol{\epsilon}_k) = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} = ((\sigma_{ij}))$ for $k = 1, ..., n$. Also $E(\boldsymbol{e}_i) = \boldsymbol{0}$ while $\text{Cov}(\boldsymbol{e}_i, \boldsymbol{e}_j) = \sigma_{ij}\boldsymbol{I}_n$ for $i, j = 1, ..., m$. Then $\boldsymbol{B}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ are unknown matrices of parameters to be estimated, and $E(\boldsymbol{Z}) = \boldsymbol{XB}$ while $E(Y_{ij}) = \boldsymbol{x}_i^T\boldsymbol{\beta}_j$.

118) Each response variable in a multivariate linear regression model follows a univariate linear regression model $\boldsymbol{Y}_j = \boldsymbol{X}\boldsymbol{\beta}_j + \boldsymbol{e}_j$ for $j = 1, ..., m$ where it is assumed that $E(\boldsymbol{e}_j) = \boldsymbol{0}$ and $\text{Cov}(\boldsymbol{e}_j) = \sigma_{jj}\boldsymbol{I}_n$.

119) For each variable Y_k make a response plot of $\hat{Y}_{ik}$ versus Y_{ik} and a residual plot of $\hat{Y}_{ik}$ versus $r_{ik} = Y_{ik} - \hat{Y}_{ik}$. If the multivariate linear regression model is appropriate, then the plotted points should cluster about the identity line in each of the m response plots. If outliers are present or if the plot is not linear, then the current model or data need to be changed or corrected. If the model is good, then the each of the m residual plots should be ellipsoidal with no trend and should be centered about the $r = 0$ line. There should not be any pattern in the residual plot: as a narrow vertical strip is moved from left to right, the behavior of the residuals within the strip should show little change. Outliers and patterns such as curvature or a fan shaped plot are bad.

120) Make a scatterplot matrix of $Y_1, ..., Y_m$ and of the continuous predictors. Use power transformations to remove strong nonlinearities.

121) Consider testing $\boldsymbol{LB} = \boldsymbol{0}$ where $\boldsymbol{L}$ is a $r \times p$ full rank matrix. Let $\boldsymbol{W}_e = \hat{\boldsymbol{E}}^T\hat{\boldsymbol{E}}$ and $\boldsymbol{W}_e/(n-p) = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}$. Let $\boldsymbol{H} = \hat{\boldsymbol{B}}^T\boldsymbol{L}^T[\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T]^{-1}\boldsymbol{L}\hat{\boldsymbol{B}}$. Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$ be the ordered eigenvalues of $\boldsymbol{W}_e^{-1}\boldsymbol{H}$. Then there are four commonly used test statistics.

The Wilk's Λ statistic is $\Lambda(\boldsymbol{L}) = |(\boldsymbol{H} + \boldsymbol{W}_e)^{-1}\boldsymbol{W}_e| = |\boldsymbol{W}_e^{-1}\boldsymbol{H} + \boldsymbol{I}|^{-1} = \prod_{i=1}^{m}(1+\lambda_i)^{-1}$.

The Pillai's trace statistic is $V(\boldsymbol{L}) = tr[(\boldsymbol{H} + \boldsymbol{W}_e)^{-1}\boldsymbol{H}] = \sum_{i=1}^{m}\frac{\lambda_i}{1+\lambda_i}$.

The Hotelling-Lawley trace statistic is $U(\boldsymbol{L}) = tr[\boldsymbol{W}_e^{-1}\boldsymbol{H}] = \sum_{i=1}^{m}\lambda_i =$

$$\frac{1}{n-p}[vec(\boldsymbol{L}\hat{\boldsymbol{B}})]^T[\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} \otimes (\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T)^{-1}][vec(\boldsymbol{L}\hat{\boldsymbol{B}})].$$

The Roy's maximum root statistic is $\lambda_{max}(\boldsymbol{L}) = \lambda_1$.

122) Under regularity conditions, $-[n-p+1-0.5(m-r+3)]\log(\Lambda(\boldsymbol{L})) \xrightarrow{D} \chi^2_{rm}$, $(n-p)V(\boldsymbol{L}) \xrightarrow{D} \chi^2_{rm}$, $\quad (n-p)U(\boldsymbol{L}) \xrightarrow{D} \chi^2_{rm}$, and if $h = \max(r, m)$,

$$\frac{n-p-h+r}{h}\lambda_{max}(\boldsymbol{L}) \approx F(h, n-p-h+r).$$

The Hotelling Lawley statistic is robust against nonnormality.

123) For the Wilk's Lambda test,
$pval = P\left(\frac{-[n-p+1-0.5(m-r+3)]}{rm} \ \log(\Lambda(\boldsymbol{L})) < F_{rm,n-rm}\right)$.

For the Pillai's trace test, $pval = P\left(\frac{n-p}{rm} \ V(\boldsymbol{L}) < F_{rm,n-rm}\right)$.

For the Hotelling Lawley trace test, $pval = P\left(\frac{n-p}{rm} \ U(\boldsymbol{L}) < F_{rm,n-rm}\right)$.

The above three tests are large sample tests, P(reject $H_0|H_0$ is true) $\rightarrow \alpha$ as $n \rightarrow \infty$, under regularity conditions.

For the Roy's largest root test, use
$pval = P\left(\frac{n-p-h+r}{h} \ \lambda_{max}(\boldsymbol{L}) < F_{h,n-p-h+r}\right)$.

The F statistic is an upper bound on the F statistic that provides a lower bound on the nominal level of significance, α, under regularity conditions.

124) The 4 step MANOVA F test of hypotheses uses $\boldsymbol{L} = [\boldsymbol{0} \ \ \boldsymbol{I}_{p-1}]$:
i) State the hypotheses H_0: the nontrivial predictors are not needed in the mreg model
H_1: at least one of the nontrivial predictors is needed
ii) Find the test statistic F_o from output.
iii) Find the pval from output.
iv) If pval $< \alpha$, reject H_0. If pval $\geq \alpha$, fail to reject H_0. If H_0 is rejected, conclude that there is a mreg relationship between the response variables $Y_1, ..., Y_m$ and the predictors X_2, ..., X_p. If you fail to reject H_0, conclude that there is a not a mreg relationship between $Y_1, ..., Y_m$ and the predictors X_2, ..., X_p. (Get the variable names from the story problem.)

125) The 4 step F_j test of hypotheses uses $\boldsymbol{L}_j = [0, ..., 0, 1, 0, ..., 0]$ where the 1 is in the jth position. Let $\boldsymbol{b}_j^T$ be the jth row of $\boldsymbol{B}$. i) State the hypotheses H_0 :
$bb_j^T = \boldsymbol{0} \quad H_1 : \ \ \boldsymbol{b}_j^T \neq \boldsymbol{0}$
ii) Find the test statistic F_j from output.
iii) Find pval from output.
iv) If pval $< \alpha$, reject H_0. If pval $\geq \alpha$, fail to reject H_0. Give a nontechnical sentence restating your conclusion in terms of the story problem. If H_0 is rejected, then conclude that X_j is needed in the mreg model for $Y_1, ..., Y_m$ given that the other predictors are in the model. If you fail to reject H_0, then conclude that X_j is not needed in the mreg model for $Y_1, ..., Y_m$ given that the other predictors are in the model. (Get the variable names from the story problem.)

126) The 4 step **MANOVA partial F test** of hypotheses has a full model using all of the variables and a reduced model where r of the variables are deleted. The ith row of $\boldsymbol{L}$ has a 1 in the position corresponding to the ith variable to be deleted. Omitting the jth variable corresponds to the F_j test while omitting variables $X_2, ..., X_p$ corresponds to the MANOVA F test.
i) State the hypotheses H_0: the reduced model is good $\quad H_1$: use the full model.
ii) Find the test statistic F_R from output.
iii) Find the pval from output.
iv) If pval $< \alpha$, reject H_0 and conclude that the full model should be used.
If pval $\geq \alpha$, fail to reject H_0 and conclude that the reduced model is good.

127) The 4 step MANOVA F test should reject H_0 if the response and residual plots look good, n is large enough and at least one response plot does not look like the corresponding residual plot. A response plot for Y_j will look like a residual plot if the identity line appears almost horizontal, hence the range of $\hat{Y}_j$ is small.

128) The *mpack* function `mltreg` produces the m response and residual plots, gives $\hat{\boldsymbol{B}}$, $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}$, the MANOVA partial F test statistic and pval corresponding to the reduced model that leaves out the variables given by indices (so X_2 and X_4 in the output below with $F = 0.77$ and pval $= 0.614$), F_j and the pval for the F_j test for variables 1, 2, ..., p (where $p = 4$ in the output below so $F_2 = 1.51$ with pval $= 0.284$) and F_0 and pval for the MANOVA F test (in the output below $F_0 = 3.15$ and pval$= 0.06$). The command `out <- mltreg(x,y,indices=c(2))` would produce a MANOVA partial F test corresponding to the F_2 test while the command `out <- mltreg(x,y,indices=c(2,3,4))` would produce a MANOVA partial F test corresponding to the MANOVA F test for a data set with $p = 4$ predictor variables. The Hotelling Lawley trace statistic is used in the tests.

```
out <- mltreg(x,y,indices=c(2,4))

$Bhat
            [,1]        [,2]        [,3]
[1,] 47.96841291 623.2817463 179.8867890
[2,]  0.07884384   0.7276600  -0.5378649
[3,] -1.45584256 -17.3872206   0.2337900
[4,] -0.01895002   0.1393189  -0.3885967

$Covhat
          [,1]      [,2]     [,3]
[1,]  21.91591  123.2557  132.339
[2,] 123.25566 2619.4996 2145.780
[3,] 132.33902 2145.7797 2954.082

$partial
      partialF      Pval
[1,] 0.7703294 0.6141573

$Ftable
             Fj      pvals
[1,] 6.30355375 0.01677169
[2,] 1.51013090 0.28449166
[3,] 5.61329324 0.02279833
[4,] 0.06482555 0.97701447

$MANOVA
      MANOVAF       pval
[1,] 3.150118 0.06038742
```

129) Given $\hat{\boldsymbol{B}} = [\hat{\boldsymbol{\beta}}_1 \;\; \hat{\boldsymbol{\beta}}_2 \;\; \cdots \;\; \hat{\boldsymbol{\beta}}_m]$ and $\boldsymbol{x}_f$, find $\hat{\boldsymbol{y}}_f = (\hat{y}_1, ..., \hat{y}_m)^T$ where $\hat{y}_i = \hat{\boldsymbol{\beta}}_i^T \boldsymbol{x}_f$.

130) $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}} = \dfrac{\hat{\boldsymbol{E}}^T \hat{\boldsymbol{E}}}{n-p} = \dfrac{1}{n-p} \sum_{i=1}^{n} \hat{\boldsymbol{\epsilon}}_i \hat{\boldsymbol{\epsilon}}_i^T$ while the sample covariance matrix of the residuals is $\boldsymbol{S}_r = \dfrac{n-p}{n-1} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}} = \dfrac{\hat{\boldsymbol{E}}^T \hat{\boldsymbol{E}}}{n-1}$. Both $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}$ and $\boldsymbol{S}_r$ are $\sqrt{n}$ consistent estimators of $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ for a large class of error distributions for $\boldsymbol{\epsilon}_i$.

131) The $100(1-\alpha)\%$ nonparametric prediction region for $\boldsymbol{y}_f$ given $\boldsymbol{x}_f$ is the nonparametric prediction region from § 5.2 applied to $\hat{\boldsymbol{z}}_i = \hat{\boldsymbol{y}}_f + \hat{\boldsymbol{\epsilon}}_i = \hat{\boldsymbol{B}}^T \boldsymbol{x}_f + \hat{\boldsymbol{\epsilon}}_i$ for $i = 1, ..., n$. This takes the data cloud of the n residual vectors $\hat{\boldsymbol{\epsilon}}_i$ and centers the cloud at $\hat{\boldsymbol{y}}_f$. Let

$$D_i^2(\hat{\boldsymbol{y}}_f, \boldsymbol{S}_r) = (\hat{\boldsymbol{z}}_i - \hat{\boldsymbol{y}}_f)^T \boldsymbol{S}_r^{-1} (\hat{\boldsymbol{z}}_i - \hat{\boldsymbol{y}}_f)$$

for $i = 1, ..., n$. Let $q_n = \min(1 - \alpha + 0.05, 1 - \alpha + m/n)$ for $\alpha > 0.1$ and

$$q_n = \min(1 - \alpha/2, 1 - \alpha + 10\alpha m/n), \quad \text{otherwise.}$$

If $q_n < 1 - \alpha + 0.001$, set $q_n = 1 - \alpha$. Let $0 < \alpha < 1$ and $h = D_{(U_n)}$ where $D_{(U_n)}$ is the q_nth sample quantile of the D_i. The $100(1-\alpha)\%$ nonparametric prediction region for $\boldsymbol{y}_f$ is

$$\{\boldsymbol{z} : (\boldsymbol{z} - \hat{\boldsymbol{y}}_f)^T \boldsymbol{S}_r^{-1} (\boldsymbol{z} - \hat{\boldsymbol{y}}_f) \leq D_{(U_n)}^2\} = \{\boldsymbol{z} : D_{\boldsymbol{z}}(\hat{\boldsymbol{y}}_f, \boldsymbol{S}_r) \leq D_{(U_n)}\}.$$

a) Consider the n prediction regions for the data where $(\boldsymbol{y}_{f,i}, \boldsymbol{x}_{f,i}) = (\boldsymbol{y}_i, \boldsymbol{x}_i)$ for $i = 1, ..., n$. If the order statistic $D_{(U_n)}$ is unique, then U_n of the n prediction regions contain $\boldsymbol{y}_i$ where $U_n/n \to 1 - \alpha$ as $n \to \infty$.

b) If $(\hat{\boldsymbol{y}}_f, \boldsymbol{S}_r)$ is a consistent estimator of $(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$ then the nonparamtric prediction region is a large sample $100(1-\alpha)\%$ prediction region for $\boldsymbol{y}_f$.

c) If $(\hat{\boldsymbol{y}}_f, \boldsymbol{S}_r)$ is a consistent estimator of $(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$, and the $\boldsymbol{\epsilon}_i$ come from an elliptically contoured distribution such that the highest density region is $\{\boldsymbol{z} : D_{\boldsymbol{z}}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}) \leq D_{1-\alpha}\}$, then the nonparametric prediction region is asymptotically optimal.

132) On the DD plot for the residuals, the cases to the left of the vertical line correspond to cases that would have $\boldsymbol{y}_f = \boldsymbol{y}_i$ in the nonparametric prediction region if $\boldsymbol{x}_f = \boldsymbol{x}_i$ while the cases to the right of the line would not have $\boldsymbol{y}_f = \boldsymbol{y}_i$ in the nonparametric prediction region.

133) The DD plot for the residuals is interpreted almost exactly as a DD plot for iid multivariate data is interpreted. Plotted points clustering about the identity line suggests that the $\boldsymbol{\epsilon}_i$ may be iid from a multivariate normal distribution while plotted points that lie above the identity line but cluster about a line through the origin with slope greater than 1 suggests that the $\boldsymbol{\epsilon}_i$ may be iid from an elliptically contoured distribution that is not MVN. The semiparametric and parametric MVN prediction regions correspond to horizontal lines on the DD plot. Robust distances have not been shown to be consistent estimators of the population distances, but are useful for a graphical diagnostic.

134) A robust multivariate linear regression method replaces least squares with the hbreg estimator. The probability that the robust estimator equals the least squares estimator goes to 1 as $n \to \infty$ for a large class of error distributions. Hence the hypothesis

tests and nonparametric prediction regions for the classical method can be applied to the robust method. The entries of $\hat{\boldsymbol{B}}$ are hard to drive to $\pm\infty$ for the robust estimator, and the residuals corresponding to outliers are often large. Since the residuals are used to compute $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}$, the tests of hypothesis based on the robust estimator are not robust to the presence of outliers. But the robust estimator and classical estimator tend to give different response and residual plots and test statistics when outliers are present.

135) For factor analysis, variables given nonzero loadings by promax are important for the factor. See Quiz 10 and homework 10.

Emphasis on quiz 7-11 and homework 7-11. Sections covered: Olive (2012) skim ch.4 with emphasis on p. 62, DGK, MB, FCH, RFCH and RMVN estimators, DD plot. From $\oint$ 5.1, Def. 5.1, Applications 5.1 and 5.2. Ch. 6, Ch. 8-12.

Johnson and Wichern (1988): $\oint$ 5.3, 5.5; 6.4, ch. 7, 8, 9, 11

Final: Tuesday, May 7, 8:00-10:00.

Cumulative: 14 sheets of notes and a calculator.

Projects are also due Tuesday, May. 7, by 2:50, but you may turn your project in earlier.