Exam 2 review. 12 sheets of notes and a calculator. Friday, Oct. 23.

**Know 1), 9), and 14) - 22) from exam 1 review.**

Types of problems.

24) A $p \times 1$ random vector $\boldsymbol{x}$ has an *elliptically contoured distribution,* if $\boldsymbol{x}$ has density

$$f(\boldsymbol{z}) = k_p |\boldsymbol{\Sigma}|^{-1/2} g[(\boldsymbol{z}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{z}-\boldsymbol{\mu})], \tag{1}$$

and we say $\boldsymbol{x}$ has an elliptically contoured $EC_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ distribution. If the second moments exist, then

$$E(\boldsymbol{x}) = \boldsymbol{\mu} \tag{2}$$

and

$$\mathrm{Cov}(\boldsymbol{x}) = c_x \boldsymbol{\Sigma} \tag{3}$$

for some constant $c_x > 0$.

25) The *population squared Mahalanobis distance*

$$U \equiv D^2 = D^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu}). \tag{4}$$

For elliptically contoured distributions, $U$ has pdf

$$h(u) = \frac{\pi^{p/2}}{\Gamma(p/2)} k_p u^{p/2-1} g(u). \tag{5}$$

$U \sim \chi^2_p$ if $\boldsymbol{x}$ has a multivariate normal $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution.

26) The classical estimator $(\overline{\boldsymbol{x}}, \boldsymbol{S})$ of multivariate location and dispersion is the sample mean and sample covariance matrix where

$$\overline{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \quad \text{and} \quad \boldsymbol{S} = \frac{1}{\mathrm{n}-1}\sum_{\mathrm{i}=1}^{\mathrm{n}} (\boldsymbol{x}_\mathrm{i} - \overline{\boldsymbol{x}})(\boldsymbol{x}_\mathrm{i} - \overline{\boldsymbol{x}})^\mathrm{T}. \tag{6}$$

27) The $n \times p$ data matrix

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

28) Let the $p \times 1$ column vector $T(\boldsymbol{W})$ be a multivariate location estimator, and let the $p \times p$ symmetric positive definite matrix $\boldsymbol{C}(\boldsymbol{W})$ be a dispersion estimator. Then the $i$th *squared sample Mahalanobis distance* is the scalar

$$D_i^2 = D_i^2(T(\boldsymbol{W}), \boldsymbol{C}(\boldsymbol{W})) = (\boldsymbol{x}_i - T(\boldsymbol{W}))^T \boldsymbol{C}^{-1}(\boldsymbol{W})(\boldsymbol{x}_i - T(\boldsymbol{W})) \tag{7}$$

for each observation $\boldsymbol{x}_i$. Notice that the Euclidean distance of $\boldsymbol{x}_i$ from the estimate of center $T(\boldsymbol{W})$ is $D_i(T(\boldsymbol{W}), \boldsymbol{I}_p)$. The classical Mahalanobis distance uses $(T, \boldsymbol{C}) = (\overline{\boldsymbol{x}}, \boldsymbol{S})$. Note that $D^2_{\boldsymbol{x}}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}) = (\boldsymbol{x}-\hat{\boldsymbol{\mu}})^T \hat{\boldsymbol{\Sigma}}^{-1} (\boldsymbol{x}-\hat{\boldsymbol{\mu}})$.

29) a) The $p \times p$ *population correlation matrix* $\text{Cor}(\boldsymbol{x}) = \boldsymbol{\rho}_{\boldsymbol{x}} = (\rho_{ij})$.

b) The *population covariance matrix* of $\boldsymbol{x}$ with $\boldsymbol{y}$ is $\text{Cov}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{\Sigma}_{\boldsymbol{x},\boldsymbol{y}} = E[(\boldsymbol{x} - E(\boldsymbol{x}))(\boldsymbol{y} - E(\boldsymbol{y}))^T]$.

30) The spectral decomposition of the symmetric matrix $\boldsymbol{A} = \sum_{i=1}^{p} \lambda_i \boldsymbol{e}_i \boldsymbol{e}_i^T = \lambda_1 \boldsymbol{e}_1 \boldsymbol{e}_1^T + \cdots + \lambda_p \boldsymbol{e}_p \boldsymbol{e}_p^T$.

31) The *generalized sample variance* $= |\boldsymbol{S}| = \det(\boldsymbol{S})$.

32) The hyperellipsoid $\{\boldsymbol{x} | D_{\boldsymbol{x}}^2 \leq h^2\} = \{\boldsymbol{x} : (\boldsymbol{x} - \overline{\boldsymbol{x}})^T \boldsymbol{S}^{-1} (\boldsymbol{x} - \overline{\boldsymbol{x}}) \leq h^2\}$ is centered at $\overline{\boldsymbol{x}}$ and has volume equal to

$$\frac{2\pi^{p/2}}{p\Gamma(p/2)} |\boldsymbol{S}|^{1/2} h^p.$$

Let $\boldsymbol{S}$ have eigenvalue eigenvector pairs $(\hat{\lambda}_i, \hat{\boldsymbol{e}}_i)$ where $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_p$. If $\overline{\boldsymbol{x}} = \boldsymbol{0}$, the axes are given by the eigenvectors $\hat{\boldsymbol{e}}_i$ where the half length in the direction of $\hat{\boldsymbol{e}}_i$ is $h\sqrt{\hat{\lambda}_i}$. Here $\hat{\boldsymbol{e}}_i^T \hat{\boldsymbol{e}}_j = 0$ for $i \neq j$ while $\hat{\boldsymbol{e}}_i^T \hat{\boldsymbol{e}}_i = 1$.

33) A **DD plot** is a plot of classical vs. robust Mahalanobis distances. The DD plot is used to check i) if the data is MVN (plotted points follow the identity line), ii) if the data is EC but not MVN (plotted points follow a line through the origin with slope $> 1$), iii) if the data is not EC (plotted points do not follow a line through the origin), iv) if multivariate outliers are present (e.g. some plotted points are far from the bulk of the data). v) The DD plot can be used to display the prediction regions of Chapter 4.

34) Many practical "robust estimators" generate a sequence of $K$ trial fits called *attractors*: $(T_1, \boldsymbol{C}_1), ..., (T_K, \boldsymbol{C}_K)$. Then the attractor $(T_A, \boldsymbol{C}_A)$ that minimizes some criterion is used to obtain the final estimator. One way to obtain attractors is to generate trial fits called *starts*, and then use the *concentration* technique. Let $(T_{-1,j}, \boldsymbol{C}_{-1,j})$ be the $j$th start and compute all $n$ Mahalanobis distances $D_i(T_{-1,j}, \boldsymbol{C}_{-1,j})$. At the next iteration, the classical estimator $(T_{0,j}, \boldsymbol{C}_{0,j})$ is computed from the $c_n \approx n/2$ cases corresponding to the smallest distances. This iteration can be continued for $k$ steps resulting in the sequence of estimators $(T_{-1,j}, \boldsymbol{C}_{-1,j}), (T_{0,j}, \boldsymbol{C}_{0,j}), ..., (T_{k,j}, \boldsymbol{C}_{k,j})$. Then $(T_{k,j}, \boldsymbol{C}_{k,j})$ is the $j$th attractor for $j = 1, ..., K$. Using $k = 10$ often works well, and the basic resampling algorithm is a special case $k = -1$ where the attractors are the starts.

35) The DGK estimator $(T_{DGK}, \boldsymbol{C}_{DGK})$ uses the classical estimator $(T_{-1,D}, \boldsymbol{C}_{-1,D}) = (\overline{\boldsymbol{x}}, \boldsymbol{S})$ as the only start.

36) The median ball (MB) estimator $(T_{MB}, \boldsymbol{C}_{MB})$ uses $(T_{-1,M}, \boldsymbol{C}_{-1,M}) = (\text{MED}(\boldsymbol{W}), \boldsymbol{I}_p)$ as the only start where $\text{MED}(\boldsymbol{W})$ is the coordinatewise median. Hence $(T_{0,M}, \boldsymbol{C}_{0,M})$ is the classical estimator applied to the "half set" of data closest to $\text{MED}(\boldsymbol{W})$ in Euclidean distance.

37) Elemental concentration algorithms use elemental starts: $(T_{-1,j}, \boldsymbol{C}_{-1,j}) = (\overline{\boldsymbol{x}}_j, \boldsymbol{S}_j)$ is the classical estimator applied to a randomly selected "elemental set" of $p + 1$ cases. If the $\boldsymbol{x}_i$ are iid with covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{x}}$, then the starts $(\overline{\boldsymbol{x}}_j, \boldsymbol{S}_j)$ are identically distributed with $E(\overline{\boldsymbol{x}}_j) = E(\boldsymbol{x}_i)$, $\text{Cov}(\overline{\boldsymbol{x}}_j) = \boldsymbol{\Sigma}_{\boldsymbol{x}}/(p+1)$, and $E(\boldsymbol{S}_j) = \boldsymbol{\Sigma}_{\boldsymbol{x}}$.

38) Let the "median ball" be the hypersphere containing the half set of data closest to $\text{MED}(\boldsymbol{W})$ in Euclidean distance. The two attractors for the FCH estimator are the MB and DGK estimators. Hence the two starts are $(\text{MED}(\boldsymbol{W}), \boldsymbol{I}_p)$ and $(\overline{\boldsymbol{x}}, \boldsymbol{S})$. The FCH estimator uses the MB estimator if the DGK location estimator $T_{DGK} = T_{k,D}$ is outside of the median ball, and the attractor with the smallest determinant, otherwise.

Let $(T_A, \boldsymbol{C}_A)$ be the attractor used. Then the estimator $(T_{FCH}, \boldsymbol{C}_{FCH})$ takes $T_{FCH} = T_A$ and

$$\boldsymbol{C}_{FCH} = \frac{\text{MED}(D_i^2(T_A, \boldsymbol{C}_A))}{\chi^2_{p,0.5}} \boldsymbol{C}_A \tag{8}$$

where $\chi^2_{p,0.5}$ is the 50th percentile of a chi–square distribution with $p$ degrees of freedom. The RFCH estimator uses two standard "reweight for efficiency steps" while the RMVN estimator uses a modified method for reweighting. The RMVN set $U$ and the RFCH set $V$ correspond to the $m_W \geq n/2$ cases used to compute RMVN or RFCH estimator where $W$ is $U$ or $V$.

39) For a large class of elliptically contoured distributions, FCH, RFCH, and RMVN are $\sqrt{n}$ consistent estimators of $(\boldsymbol{\mu}, c_i\boldsymbol{\Sigma})$ for $c_1, c_2, c_3 > 0$ where $c_i = 1$ for $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ data.

40) An estimator $(T, \boldsymbol{C})$ of multivariate location and dispersion (MLD), needs to estimate $p(p+3)/2$ unknown parameters when there are $p$ random variables. For $(\overline{\boldsymbol{x}}, \boldsymbol{S})$ we want $n \geq 10p$. We want $n \geq 20p$ for FCH, RFCH, or RMVN.

41) Brand name robust MLD estimators take too long to compute: F-brand name estimators that are not backed by breakdown or large sample theory are actually used. FMCD, F-MVE, F-S, F-MM, F-$\tau$, F-constrained-M and F-Stahel-Donoho are especially common. F-brand name estimators use a fixed number of starts.

42) The squared Euclidean distances of the $\boldsymbol{x}_i$ from the coordinatewise median is $D_i^2 = D_i^2(\text{MED}(\boldsymbol{W}), \boldsymbol{I}_p)$. Concentration type steps compute the weighted median $\text{MED}_j$: the coordinatewise median computed from the cases $\boldsymbol{x}_i$ with $D_i^2 \leq \text{MED}(D_i^2(\text{MED}_{j-1}, \boldsymbol{I}_p))$ where $\text{MED}_0 = \text{MED}(\boldsymbol{W})$. Often used $j = 0$ (no concentration type steps) or $j = 9$. Let $D_i = D_i(\text{MED}_j, \boldsymbol{I}_p)$. Let $W_i = 1$ if $D_i \leq \text{MED}(D_1, ..., D_n) + k\text{MAD}(D_1, ..., D_n)$ where $k \geq 0$ and $k = 5$ is the default choice. Let $W_i = 0$, otherwise.

43) Let the *covmb2 set* $B$ of at least $n/2$ cases correspond to the cases with weight $W_i = 1$. Then the *covmb2* estimator $(T, \boldsymbol{C})$ is the sample mean and sample covariance matrix applied to the cases in set $B$. Hence

$$T = \frac{\sum_{i=1}^n W_i \boldsymbol{x}_i}{\sum_{i=1}^n W_i} \text{ and } \boldsymbol{C} = \frac{\sum_{i=1}^n W_i (\boldsymbol{x}_i - T)(\boldsymbol{x}_i - T)^T}{\sum_{i=1}^n W_i \ - \ 1}.$$

The function `ddplot5` plots the Euclidean distances from the coordinatewise median versus the Euclidean distances from the `covmb2` location estimator. Typically the plotted points in this DD plot cluster about the identity line, and outliers appear in the upper right corner of the plot with a gap between the bulk of the data and the outliers.

**ch. 11**

44) Let $\boldsymbol{X}_n$ be a sequence of random vectors with joint cdfs $F_n(\boldsymbol{x})$ and let $\boldsymbol{X}$ be a random vector with joint cdf $F(\boldsymbol{x})$.

a) $\boldsymbol{X}_n$ **converges in distribution** to $\boldsymbol{X}$, written $\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{X}$, if $F_n(\boldsymbol{x}) \to F(\boldsymbol{x})$ as $n \to \infty$ for all points $\boldsymbol{x}$ at which $F(\boldsymbol{x})$ is continuous. The distribution of $\boldsymbol{X}$ is the **limiting distribution** or **asymptotic distribution** of $\boldsymbol{X}_n$. Note that $\boldsymbol{X}$ **does not depend** on $n$.

b) $\boldsymbol{X}_n$ **converges in probability** to $\boldsymbol{X}$, written $\boldsymbol{X}_n \xrightarrow{P} \boldsymbol{X}$, if for every $\epsilon > 0$, $P(\|\boldsymbol{X}_n - \boldsymbol{X}\| > \epsilon) \to 0$ as $n \to \infty$.

45) Multivariate Central Limit Theorem (MCLT): If $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ are iid $k \times 1$ random vectors with $E(\boldsymbol{X}) = \boldsymbol{\mu}$ and $\text{Cov}(\boldsymbol{X}) = \boldsymbol{\Sigma}_{\boldsymbol{x}}$, then

$$\sqrt{n}(\overline{\boldsymbol{X}}_n - \boldsymbol{\mu}) \xrightarrow{D} N_k(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{x}})$$

where the sample mean

$$\overline{\boldsymbol{X}}_n = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i.$$

46) Suppose $\sqrt{n}(T_n - \boldsymbol{\mu}) \xrightarrow{D} N_p(\boldsymbol{\theta}, \boldsymbol{\Sigma})$. Let $\boldsymbol{A}$ be a $q \times p$ constant matrix. Then $\boldsymbol{A}\sqrt{n}(T_n - \boldsymbol{\mu}) = \sqrt{n}(\boldsymbol{A}T_n - \boldsymbol{A}\boldsymbol{\mu}) \xrightarrow{D} N_q(\boldsymbol{A}\boldsymbol{\theta}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T)$.

47) Suppose $\boldsymbol{A}$ is a conformable constant matrix and $\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{X}$. Then $\boldsymbol{A}\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{A}\boldsymbol{X}$.

**ch. 4**

48) For $h > 0$, the hyperellipsoid $\{\boldsymbol{z} : (\boldsymbol{z} - T)^T \boldsymbol{C}^{-1}(\boldsymbol{z} - T) \leq h^2\} =$
$\{\boldsymbol{z} : D^2_{\boldsymbol{z}} \leq h^2\} = \{\boldsymbol{z} : D_{\boldsymbol{z}} \leq h\}$. A future observation (random vector) $\boldsymbol{x}_f$ is in this region if $D_{\boldsymbol{x}_f} \leq h$. A large sample $100(1-\delta)\%$ prediction region is a set $\mathcal{A}_n$ such that $P(\boldsymbol{x}_f \in \mathcal{A}_n)$ is eventually bounded below by $1 - \delta$ as $n \to \infty$ where $0 < \delta < 1$. A *large sample* $100(1-\delta)\%$ *confidence region* for a vector of parameters $\boldsymbol{\theta}$ is a set $\mathcal{A}_n$ such that $P(\boldsymbol{\theta} \in \mathcal{A}_n)$ is eventually bounded below by $1 - \delta$ as $n \to \infty$.

49) Let $q_n = \min(1 - \delta + 0.05, 1 - \delta + p/n)$ for $\delta > 0.1$ and $q_n =$
$\min(1 - \delta/2, 1 - \delta + 10\delta p/n)$, otherwise. If $q_n < 1 - \delta + 0.001$, set $q_n = 1 - \delta$. If $(T, \boldsymbol{C})$ is a consistent estimator of $(\boldsymbol{\mu}, d\boldsymbol{\Sigma})$, then $\{\boldsymbol{z} : D_{\boldsymbol{z}}(T, \boldsymbol{C}) \leq h\}$ is a large sample $100(1-\delta)\%$ prediction regions if $h = D_{(U_n)}$ where $D_{(U_n)}$ is the $100q_n$th sample quantile of the $D_i$. The large sample $100(1-\delta)\%$ nonparametric prediction region $\{\boldsymbol{z} : D^2_{\boldsymbol{z}}(\overline{\boldsymbol{x}}, \boldsymbol{S}) \leq D^2_{(U_n)}\}$ uses $(T, \boldsymbol{C}) = (\overline{\boldsymbol{x}}, \boldsymbol{S})$. We want $n \geq 10p$ for good coverage and $n \geq 50p$ for good volume. In the DD plot, cases with $MD \leq D_{(U_n)}$ are in the nonparametric prediction region. The cutoff is displayed by a vertical line by the function `ddplot4`. Cases in the semiparametric prediction region $\{\boldsymbol{z} : D^2_{\boldsymbol{z}}(T_{RMVN}, \boldsymbol{C}_{RMVN}) \leq D^2_{(U_n)}\}$ have $RD \leq D_{(U_n)} = D_{(U_n)}(T_{RMVN}, \boldsymbol{C}_{RMVN})$ displayed by a horizontal line in the DD plot. A horizontal line going to the identity line is at $RD = \sqrt{\chi^2_{p,1-\delta}}$.

50) Consider testing $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ versus $H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ where $\boldsymbol{\theta}_0$ is a known $g \times 1$ vector. Make a confidence region and reject $H_0$ if $\boldsymbol{\theta}_0$ is not in the confidence region. Let $q_B$ and $U_B$ be as in 49) with $n$ replaced by $B$ and $p$ replaced by $g$. Let $\overline{T}^*$ and $\boldsymbol{S}^*_T$ be the sample mean and sample covariance matrix of the bootstrap sample $T^*_1, ..., T^*_B$.
a) The prediction region method large sample $100(1-\delta)\%$ confidence region for $\boldsymbol{\theta}$ is $\{\boldsymbol{w} : (\boldsymbol{w} - \overline{T}^*)^T [\boldsymbol{S}^*_T]^{-1} (\boldsymbol{w} - \overline{T}^*) \leq D^2_{(U_B)}\} = \{\boldsymbol{w} : D^2_{\boldsymbol{w}}(\overline{T}^*, \boldsymbol{S}^*_T) \leq D^2_{(U_B)}\}$ where $D^2_{(U_B)}$ is computed from $D^2_i = (T^*_i - \overline{T}^*)^T [\boldsymbol{S}^*_T]^{-1} (T^*_i - \overline{T}^*)$ for $i = 1, ..., B$. Note that the corresponding test for $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ rejects $H_0$ if $(\overline{T}^* - \boldsymbol{\theta}_0)^T [\boldsymbol{S}^*_T]^{-1} (\overline{T}^* - \boldsymbol{\theta}_0) > D^2_{(U_B)}$. This procedure applies the nonparametric prediction region to the bootstrap sample.
b) The modified Bickel and Ren (2001) large sample $100(1-\delta)\%$ confidence region is $\{\boldsymbol{w} : (\boldsymbol{w} - T_n)^T [\boldsymbol{S}^*_T]^{-1} (\boldsymbol{w} - T_n) \leq D^2_{(U_B,T)}\} = \{\boldsymbol{w} : D^2_{\boldsymbol{w}}(T_n, \boldsymbol{S}^*_T) \leq D^2_{(U_B,T)}\}$ where the cutoff $D^2_{(U_B,T)}$ is the $100q_B$th sample quantile of the $D^2_i = (T^*_i - T_n)^T [\boldsymbol{S}^*_T]^{-1} (T^*_i - T_n)$.
c) The hybrid large sample $100(1-\delta)\%$ confidence region: $\{\boldsymbol{w} : (\boldsymbol{w} - T_n)^T [\boldsymbol{S}^*_T]^{-1} (\boldsymbol{w} - T_n) \leq D^2_{(U_B)}\} = \{\boldsymbol{w} : D^2_{\boldsymbol{w}}(T_n, \boldsymbol{S}^*_T) \leq D^2_{(U_B)}\}$.

If $g = 1$, confidence intervals can be computed without $\boldsymbol{S}_T^*$ or $D^2$ for a), b), and c).

51) **Theorem 4.1: Geometric Argument.** Suppose $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}$ with $E(\boldsymbol{u}) = \boldsymbol{0}$ and $Cov(\boldsymbol{u}) = \boldsymbol{\Sigma_u}$. Assume $T_1, ..., T_B$ are iid with nonsingular covariance matrix $\boldsymbol{\Sigma}_{T_n}$. Assume $(n\boldsymbol{S}_T)^{-1} \xrightarrow{P} \boldsymbol{\Sigma}_A^{-1}$. Then the large sample $100(1-\delta)\%$ prediction region $R_p = \{\boldsymbol{w} : D^2_{\boldsymbol{w}}(\overline{T}, \boldsymbol{S}_T) \leq D^2_{(U_B)}\}$ centered at $\overline{T}$ contains a future value of the statistic $T_f$ with probability $1-\delta_B \to 1-\delta$ as $B \to \infty$. Hence the region $R_c = \{\boldsymbol{w} : D^2_{\boldsymbol{w}}(T_n, \boldsymbol{S}_T) \leq D^2_{(U_B)}\}$ is a large sample $100(1-\delta)\%$ confidence region for $\boldsymbol{\theta}$.

52) Suppose $m$ independent large sample $100(1-\delta)\%$ prediction regions are made where $\boldsymbol{x}_1, ..., \boldsymbol{x}_n, \boldsymbol{x}_f$ are iid from the same distribution for each of the $m$ runs. Let $Y$ count the number of times $\boldsymbol{x}_f$ is in the prediction region. Then $Y \sim$ binomial $(m, 1-\delta_n)$ where $1-\delta_n$ is the true coverage. Simulation can be used to see if the true or actual coverage $1-\delta_n$ is close to the nominal coverage $1-\delta$. A prediction region with $1-\delta_n < 1-\delta$ is liberal and a region with $1-\delta_n > 1-\delta$ is conservative. It is better to be conservative by 3% than liberal by 3%. Parametric prediction regions tend to have large undercoverage and so are too liberal. Similar definitions are used for confidence regions.

**ch. 5**

Response = Y

Coefficient Estimates

| Label | Estimate | Std. Error | t-value | p-value |
|---|---|---|---|---|
| Constant | $\hat{\beta}_1$ | $se(\hat{\beta}_1)$ | $t_{o,1}$ | for Ho: $\beta_1 = 0$ |
| $x_2$ | $\hat{\beta}_2$ | $se(\hat{\beta}_2)$ | $t_{o,2} = \hat{\beta}_2/se(\hat{\beta}_2)$ | for Ho: $\beta_2 = 0$ |
| $\vdots$ | | | | |
| $x_p$ | $\hat{\beta}_p$ | $se(\hat{\beta}_p)$ | $t_{o,p} = \hat{\beta}_p/se(\hat{\beta}_p)$ | for Ho: $\beta_p = 0$ |

```
R Squared:                r^2
Sigma hat:                sqrt{MSE}
Number of cases:            n
Degrees of freedom:        n-p

Summary Analysis of Variance Table
```

| Source | df | SS | MS | F | p-value |
|---|---|---|---|---|---|
| Regression | p-1 | SSR | MSR | Fo=MSR/MSE | for Ho: |
| Residual | n-p | SSE | MSE | | $\beta_2 = \cdots = \beta_p = 0$ |

```
Response      = brnweight
Coefficient Estimates
Label      Estimate        Std. Error    t-value    p-value
Constant   99.8495         171.619         0.582     0.5612
size       0.220942        0.0357902       6.173     0.0000
sex        22.5491         11.2372         2.007     0.0458
breadth   -1.24638         1.51386        -0.823     0.4111
```

```
circum     1.02552          0.471868         2.173      0.0307

R Squared:                0.749755
Sigma hat:                 82.9175
Number of cases:              267
Degrees of freedom:           262

Summary Analysis of Variance Table
Source          df        SS             MS            F     p-value
Regression       4   5396942.       1349235.      196.24     0.0000
Residual       262   1801333.        6875.32
```

53) Know the meaning of the least squares multiple linear regression output. Shown above is an actual output and an output only using symbols.

54) The response variable is the variable that you want to predict. The predictor variables are the variables used to predict the response variable.

55) The MLR model is

$$Y_i = x_{i,1}\beta_1 + x_{i,2}\beta_2 + \cdots + x_{i,p}\beta_p + e_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i$$

for $i = 1, \ldots, n$. Here $n$ is the *sample size* and the random variable $e_i$ is the $i$th **error**. The constant variance MLR model assumes that the errors are iid with $E(e_i) = 0$ and $V(e_i) = \sigma^2 < \infty$. Assume that the errors are independent of the predictor variables $\boldsymbol{x}_i$. The *unimodal MLR model* also assumes that the $e_i$ are iid from a unimodal distribution that is not highly skewed. Usually $x_{i,1} \equiv 1$.

56) 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 dependent 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.

57) The OLS estimators are $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{OLS} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$ and $\hat{\sigma}^2 = MSE = \sum_{i=1}^n r_i^2/(n-p)$. Thus $\hat{\sigma} = \sqrt{MSE}$. The vector of *predicted* or *fitted values* $\hat{\boldsymbol{Y}} = \hat{\boldsymbol{Y}}_{OLS} = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{OLS} = \boldsymbol{H}\boldsymbol{Y}$ where the *hat matrix* $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$. The $i$th fitted value $\hat{Y}_i = \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}$. The $i$th residual $r_i = Y_i - \hat{Y}_i$ and the vector of residuals $\boldsymbol{r} = \boldsymbol{Y} - \hat{\boldsymbol{Y}} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$. The least squares regression equation for a model containing a constant is $\hat{Y} = \hat{\beta}_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$.

58) Always make the response plot of $\hat{Y}$ versus $Y$ and residual plot of $\hat{Y}$ versus $r$ for any MLR analysis. The response plot is used to visualize the MLR model, that is, to visualize the conditional distribution of $Y|\boldsymbol{x}^T\boldsymbol{\beta}$. If the unimodal MLR model of 55) is useful, then i) the plotted points in the response plot should scatter about the identity line with no other pattern, and ii) the plotted points in the residual plot should scatter about the $r = 0$ line with no other pattern. If either i) or ii) is violated, then the unimodal MLR model *is not sustained.* In other words, if the plotted points in the residual plot

show some type of dependency, e.g. increasing variance or a curved pattern, then the multiple linear regression model may be inadequate.

59) Use $x_f \leq \max h_i$ for valid predictions where $h_i = h_{ii}$ is the $i$th diagonal element of $\boldsymbol{H}$.

60) The classical 100 $(1-\delta)$% PI for $Y_f$ is $\hat{Y}_f \pm t_{n-p,1-\delta/2} se(pred)$, but should be replaced with the asymptotically optimal PI. Asymptotically, this PI finds the shorth$(c)$ interval $[r_{(s)}, r_{(s+c-1)}]$ of the residuals and uses $[\hat{Y}_f + r_{(s)}, \hat{Y}_f + r_{(s+c-1)}]$.

61) **OLS CLT (Least Squares Central Limit Theorem):** Consider the MLR model $Y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + e_i$ and assume that the zero mean errors are iid with $E(e_i) = 0$ and $V(e_i) = \sigma^2$. Also assume that $\max_i(h_1, ..., h_n) \to 0$ in probability as $n \to \infty$ and

$$\frac{\boldsymbol{X}^T \boldsymbol{X}}{n} \to \boldsymbol{W}^{-1}$$

as $n \to \infty$. Then the least squares (OLS) estimator $\hat{\boldsymbol{\beta}}$ satisfies

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \sigma^2 \ \boldsymbol{W}). \tag{9}$$

62) The response and residual plots are useful for detecting outliers. Beginners often label too many points as outliers. Mentally draw a box about the bulk of the data ignoring any outliers. Double the width of the box (about the identity line for the response plot and about the horizontal line $r = 0$ for the residual plot). Cases outside of this imaginary doubled box are potential outliers. Alternatively, visually estimate the standard deviation of the residuals in both plots. In the residual plot look for residuals that are more than 5 standard deviations from the $r = 0$ line.

The identity line can also pass through or near an outlier or a cluster of outliers. Then the outliers will be in the upper right or lower left of the response plot, and there will be a large gap between the cluster of outliers and the bulk of the data.

63) The outlying cases in 62) could be **good leverage points**. The response plot using the $\hat{\boldsymbol{\beta}}_B$ computed only from the bulk of the data is such that the identity line passes near or through the good leverage points, which have outlying $\boldsymbol{x}_i$. The infants in the Gladstone data are an example of good leverage points. Such cases should not be called outliers. *Masking* occurs if the analysis suggests that one or more outliers are good cases while *swamping* occurs if the analysis suggests that one or more good cases are outliers.

64) Given a response plot with highlighted cases corresponding to large Cook's distances, know that masking occurred for Cook's distances if all of the highlighted cases correspond to outliers but some of the outliers are not highlighted. See Figure 6.2. Swamping occurs for Cook's distances if some of the highlighted cases are good cases. See Figure 6.3 where the outliers are not highlighted but the "good leverage points" are highlighted. Masking and swamping also often occur with respect to residuals. Fitted values are often good for detecting outliers.

65) For MLR outlier detection, the **RR plot** is a scatterplot matrix of residuals from several MLR estimators. Adding $\hat{Y}_{OLS}$ to the top or bottom of the plot may be a good idea.. The **FF plot** is a scatterplot matrix of fitted values from several MLR estimators. Add the response $Y$ to the top or bottom of the plot to see the response

plots of each estimator and to detect outliers. If a marginal plot is a straight line, then the two estimators are fitting the data in roughly the same way.

66) Suppose $c = c_n \approx n/2$. The LMS($c$) criterion is

$$Q_{LMS}(\boldsymbol{b}) = r^2_{(c)}(\boldsymbol{b}) \tag{10}$$

where $r^2_{(1)} \leq \cdots \leq r^2_{(n)}$ are the ordered squared residuals, and the LTS($c$) criterion is

$$Q_{LTS}(\boldsymbol{b}) = \sum_{i=1}^{c} r^2_{(i)}(\boldsymbol{b}). \tag{11}$$

The LTA($c$) criterion is

$$Q_{LTA}(\boldsymbol{b}) = \sum_{i=1}^{c} |r(\boldsymbol{b})|_{(i)} \tag{12}$$

where $|r(\boldsymbol{b})|_{(i)}$ is the $i$th ordered absolute residual.

67) Three impractical high breakdown robust estimators are the least median of squares (LMS) estimator, the least trimmed sum of squares (LTS) estimator, and the least trimmed sum of absolute deviations (LTA) estimator. These estimators correspond to the $\hat{\boldsymbol{\beta}}_L \in \mathbb{R}^p$ that minimizes the corresponding criterion.

68) For multiple linear regression, an *elemental set* is a set of $p$ cases. The *elemental fit* from the $i$th elemental set $J_i$ is the OLS estimator $\hat{\boldsymbol{\beta}}_{J_i} = (\boldsymbol{X}^T_{J_i}\boldsymbol{X}_{J_i})^{-1}\boldsymbol{X}^T_{J_i}\boldsymbol{Y}_{J_i} = \boldsymbol{X}^{-1}_{J_i}\boldsymbol{Y}_{J_i}$ applied to the cases corresponding to the elemental set provided that the inverse of $\boldsymbol{X}_{J_i}$ exists.

69) A *start* is an initial trial fit and an *attractor* is the final fit generated by the algorithm from the start. Let $\boldsymbol{b}_{0,j}$ be the $j$th start and compute all $n$ residuals $r_i(\boldsymbol{b}_{0,j}) = Y_i - \boldsymbol{x}_i^T\boldsymbol{b}_{0,j}$. Let $\lfloor n/2 \rfloor \leq c_n \leq \lfloor n/2 \rfloor + \lfloor (p+1)/2 \rfloor$. i) For an *FLTS concentration algorithm*, at the next iteration, the OLS estimator $\boldsymbol{b}_{1,j}$ is computed from the $c_n \approx n/2$ cases corresponding to the smallest squared residuals $r_i^2(\boldsymbol{b}_{0,j})$. This iteration can be continued for $k$ steps resulting in the sequence of estimators $\boldsymbol{b}_{0,j}, \boldsymbol{b}_{1,j}, ..., \boldsymbol{b}_{k,j}$. The result of the iteration $\boldsymbol{b}_{k,j}$ is called the $j$th attractor where $j = 1, ..., K$. The final FLTS concentration algorithm estimator uses the attractor that minimizes the LTS criterion.

**ch. 6**

70) Let the $i$th case $\boldsymbol{w}_i = (Y_i, \boldsymbol{x}_i^T)^T$ where the continuous predictors from $\boldsymbol{x}_i$ are denoted by $\boldsymbol{u}_i$ for $i = 1, ..., n$. Now let $D$ be the RMVN set $U$, the RFCH set $V$, or the `covmb2` set $B$. Find $D$ by applying the MLD estimator to the $\boldsymbol{u}_i$, and then run the MLR method on the $m$ cases $\boldsymbol{w}_i$ corresponding to the set $D$ indices $i_1, ..., i_m$, where $m \geq n/2$. This estimator is the *MLD set MLR estimator*.

71) The Euclidean distance of the $i$th vector of predictors $\boldsymbol{x}_i$ from the $j$th vector of predictors $\boldsymbol{x}_j$ is

$$D_i(\boldsymbol{x}_j) = D_i(\boldsymbol{x}_j, \boldsymbol{I}_p) = \sqrt{(\boldsymbol{x}_i - \boldsymbol{x}_j)^T(\boldsymbol{x}_i - \boldsymbol{x}_j)}.$$

For a fixed $\boldsymbol{x}_j$ consider the ordered distances $D_{(1)}(\boldsymbol{x}_j), ..., D_{(n)}(\boldsymbol{x}_j)$. Next, let $\hat{\boldsymbol{\beta}}_j(\alpha)$ denote the OLS fit to the $\min(p + 3 + \lfloor \alpha n/100 \rfloor, n)$ cases with the smallest distances where the approximate percentage of cases used is $\alpha \in \{1, 2.5, 5, 10, 20, 33, 50\}$. (Here $\lfloor x \rfloor$ is the

greatest integer function so $\lfloor 7.7 \rfloor = 7$. The extra $p + 3$ cases are added so that OLS can be computed for small $n$ and $\alpha$.) This yields seven OLS fits corresponding to the cases with predictors closest to $\boldsymbol{x}_j$. A fixed number of $K$ cases are selected at random without replacement to use as the $\boldsymbol{x}_j$. Hence $7K$ OLS fits are generated. We use $K = 7$ as the default. A robust criterion $Q$ is used to evaluate the $7K$ fits and the OLS fit to all of the data. Hence $7K + 1$ OLS fits (attractors) are generated and the MBA estimator `mbareg` is the fit that minimizes the criterion $Q$ from 66).

72) Compute $(T, \boldsymbol{C})$ on the $\boldsymbol{x}_i$, perhaps using the RMVN estimator. Trim the $M\%$ of the cases with the largest Mahalanobis distances, and then compute the MLR estimator $\hat{\boldsymbol{\beta}}_M$ from the remaining cases. Use $M = 0$, 10, 20, 30, 40, 50, 60, 70, 80, and 90 to generate ten response plots of the fitted values $\hat{\boldsymbol{\beta}}_M^T \boldsymbol{x}_i$ versus $Y_i$ using all $n$ cases. (Fewer plots are used for small data sets if $\hat{\boldsymbol{\beta}}_M$ can not be computed for large $M$.) These plots are called "trimmed views." The `tvreg` trimmed views (TV) estimator $\hat{\boldsymbol{\beta}}_{T,n}$ corresponds to the trimmed view where the bulk of the plotted points follow the identity line with smallest variance function, ignoring any outliers.

73) For 71) and 72), the $K$ attractors ($K = 10$ for `tvreg`) are $\sqrt{n}$ consistent estimators of $\boldsymbol{\beta}$. Hence the `mbareg` and `tvreg` estimators are $\sqrt{n}$ consistent.

74) The `rmreg2` estimator is the OLS estimator computed from the cases in the RMVN set $U$ applied to $\boldsymbol{u}_i = (x_{i2}, ..., x_{ip}, Y_i)^T$ for $i = 1, ..., n$. Hence $\boldsymbol{u}_i$ is the $i$th case with $x_{i1} = 1$ deleted.

75) The `hbreg` estimator $\hat{\boldsymbol{\beta}}_H$ is defined as follows. Pick a constant $a > 1$ and set $\hat{\boldsymbol{\beta}}_H = \hat{\boldsymbol{\beta}}_C$. If $aQ_L(\hat{\boldsymbol{\beta}}_A) < Q_L(\hat{\boldsymbol{\beta}}_C)$, set $\hat{\boldsymbol{\beta}}_H = \hat{\boldsymbol{\beta}}_A$. If $aQ_L(\hat{\boldsymbol{\beta}}_B) < \min[Q_L(\hat{\boldsymbol{\beta}}_C), aQ_L(\hat{\boldsymbol{\beta}}_A)]$, set $\hat{\boldsymbol{\beta}}_H = \hat{\boldsymbol{\beta}}_B$. The default estimator uses $Q_L$ from 66), $a = 1.4$, $\hat{\boldsymbol{\beta}}_C = \hat{\boldsymbol{\beta}}_{OLS}$, $\hat{\boldsymbol{\beta}}_A$ the `mbareg` or `rmreg2` estimator, and $\hat{\boldsymbol{\beta}}_B$ is the FLTS concentration estimator that uses $(MED(n), 0, ..., 0)^T$ as the high breakdown start where $MED(n)$ is the sample median of the $Y_i$.