Math 484 Exam 2 is on Wednesday, Oct. 26 and covers ch. 2, 3.1, 3.2, 3.3, 3.4, homeworks 1-7 and quizzes 1-7. You are allowed 10 sheets of notes and a calculator. Any needed tables will be provided. CHECK FORMULAS: YOU ARE RESPONSIBLE FOR ANY ERRORS ON THIS HANDOUT!

Everything from Exam 1 is fair game.

Types of problems likely to appear on Exam 2:

1) – 13) on Exam 1 review.

Given a new or future vector of predictors $\boldsymbol{x}_f = (1, x_{f,2}, ..., x_{f,p})^T$, let a new or future observation Y_f be independent of $Y_1, ..., Y_n$.

14) Be able to find $\hat{Y}_f = \boldsymbol{x}_f^T \hat{\boldsymbol{\beta}}$, the point estimator of i) Y_f given $\boldsymbol{x}_f$ and of ii) $E(Y_f|\boldsymbol{x}_f)$.

iii) The 100 $(1-\delta)$% CI for E(Y_f) $= \boldsymbol{x}_f^T \boldsymbol{\beta} = E(\hat{Y}_f)$ is $\hat{Y}_f \pm t_{n-p,1-\delta/2} se(\hat{Y}_f)$. Generally $se(\hat{Y}_f)$ will come from output.

iv) The 100 $(1-\delta)$% prediction interval (PI) for Y_f is $\hat{Y}_f \pm t_{n-p,1-\delta/2} se(pred)$. Generally $se(pred)$ will come from output. Note that Y_f is a random variable not a parameter. See Q4 2, HW4 A, B. Use $z_{1-\delta/2}$ for $t_{n-p,1-\delta/2}$ if $df = n - p > 30$.

The simple linear regression (SLR) model is $Y = \beta_1 + \beta_2 x + e$ is a special case of the MLR model with $p = 2$.

15) From a story problem be able to determine which variable is the response variable and which variable is the explanatory variable. Given two sample means $\overline{x}$ and $\overline{y}$, two sample standard deviations s_x and s_y and the sample correlation $\hat{\rho} \equiv \hat{\rho}(x, y)$, be able to find the LS line $\hat{Y} = \hat{\beta}_1 + \hat{\beta}_2 x$ where $\hat{\beta}_2 = \hat{\rho} s_y / s_x$ and $\hat{\beta}_1 = \overline{y} - \hat{\beta}_2 \overline{x}$. See Q4 1, and HW4 C).

16) Given $\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})$, $\sum_{i=1}^n (x_i - \overline{x})^2$, $\overline{x}$, and $\overline{y}$, find the least squares line $\hat{Y} = \hat{\beta}_1 + \hat{\beta}_2 x$ where the slope

$$\hat{\beta}_2 = \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^n (x_i - \overline{x})^2}$$

and the intercept $\hat{\beta}_1 = \overline{y} - \hat{\beta}_2 \overline{x}$. See HW4 D). Note that $\hat{\beta}_2 = \widehat{Cov}(x, y)/s_x^2$.

17) Given a small data set, find the least squares line by finding the sums needed in 16). You need the x_i, y_i and a table with headers x_i, y_i, $x_i - \overline{x}$, $y_i - \overline{y}$, $(x_i - \overline{x})(y_i - \overline{y})$, and $(x_i - \overline{x})^2$. See Q4 3, HW4 E).

18) Given $\hat{\beta}_1$ and $\hat{\beta}_2$, be able to add the LS line to a scatterplot of x vs Y. Simply find $\hat{Y}$ for 2 values of x, plot the 2 points and draw the line determined by the 2 points.

19) Suppose the least squares criterion is $Q(\eta) = \sum_{i=1} (Y_i - a_i - \eta x_i)^2$ where the constants a_i are known.

i) Then the MLR model is $Y_i = a_i + \beta x_i + e_i$ and $E(Y_i) = a_i + \beta x_i$. See HW4 F), HW5 Aa), Q5.

ii) Then find the least squares estimator $\hat{\beta}$ of β by differentiating $Q(\eta)$ with respect to η, setting the derivative to zero, solving for η, and calling the solution $\hat{\beta}$. To show that $\hat{\beta}$ is actually the LS estimator, show that the 2nd derivative of Q wrt η is greater than zero for all values of η.

By the **chain rule,**

$$\frac{dQ}{d\eta} = \sum_{i=1}^n 2(Y_i - a_i - \eta x_i)(-x_i) = -2 \sum_{i=1}^n \ x_i(Y_i - a_i - \eta x_i)$$

$$= -2[\sum_{i=1}^n \ x_i(Y_i - a_i) - \eta \sum_{i=1}^n x_i^2] = 2\eta \sum_{i=1}^n x_i^2 - 2 \sum_{i=1}^n \ x_i(Y_i - a_i) \overset{set}{=} 0.$$

Or

$$\eta \ [2 \ \sum_{i=1}^n x_i^2] = 2 \sum_{i=1}^n \ x_i(Y_i - a_i)$$

or

$$\hat{\eta} = \hat{\beta} = \frac{\sum_{i=1}^n \ x_i(Y_i - a_i)}{\sum_{i=1}^n x_i^2}.$$

Now

$$\frac{d^2Q}{d\eta^2} = 2 \sum_{i=1}^n x_i^2 > 0.$$

If $x_i \equiv 1$, then

$$\frac{dQ}{d\eta} = 2\eta \sum_{i=1}^n 1 - 2 \sum_{i=1}^n (Y_i - a_i) \overset{set}{=} 0.$$

Or

$$\eta \ [2n] = 2 \sum_{i=1}^n (Y_i - a_i)$$

or

$$\hat{\eta} = \hat{\beta} = \frac{\sum_{i=1}^n (Y_i - a_i)}{n}$$

and

$$\frac{d^2Q}{d\eta^2} = 2n > 0.$$

In particular, if the model is $Y_i = \beta_1 + \beta_2 x_i + e_i$ then $Q(\eta_1, \eta_2) = \sum_{i=1}^n (Y_i - \eta_1 - \eta_2 x_i)^2$. If the unknown parameter is β_1, then $\eta_2 = \beta_2$ and by the **chain rule,**

$$\frac{dQ}{d\eta_1} = -2 \sum_{i=1}^n (Y_i - \eta_1 - \beta_2 x_i), \quad n\hat{\beta}_1 = \sum_{i=1}^n (Y_i - \beta_2 x_i)$$

and

$$\frac{d^2Q}{d\eta_1^2} = 2n > 0.$$

If the unknown parameter is β_2, then $\eta_1 = \beta_1$, and by the **chain rule,**

$$\frac{dQ}{d\eta_2} = -2 \sum_{i=1}^n x_i(Y_i - \beta_1 - \eta_2 x_i), \quad \hat{\beta}_2 \sum_{i=1}^n x_i^2 = \sum_{i=1}^n \ x_i(Y_i - \beta_1)$$

and

$$\frac{d^2Q}{d\eta_2^2} = 2 \sum_{i=1}^n x_i^2 > 0.$$

See HW4 Fb), HW5 Aa) and Q5.

20) An **added variable plot** is used to give information about the test $Ho : \beta_i = 0$. The points in the plot cluster about a line with slope $= \hat{\beta}_i$. If there is strong trend then x_i is needed in the MLR for Y given that the other predictors $x_2, ..., x_{i-1}, x_{i+1}, ..., x_p$ are in the model. If there is almost no trend, then x_i may not be needed in the MLR for Y given that the other predictors $x_2, ..., x_{i-1}, x_{i+1}, ..., x_p$ are in the model. If the zero line with 0 slope and 0 intercept and the OLS line are added to the plot, the variable is probably needed if it is clear that the two lines intersect at the origin. The variable is probably not needed if the two lines nearly coincide near the origin in that you can not clearly tell that they intersect at the origin. Know how to use the plot. See HW4 G.

21) Given a least squares estimator $\hat{\beta}_j = \sum_{i=1}^n k_i Y_i$ and the least squares model $Y_i = \beta_1 + \beta_2 x_i + e_i$ where the e_i are iid with $E(e_i) = 0$ and $\text{V}(e_i) = \sigma^2$, and possibly one of β_1 or β_2 is known, be able to find
i) $E(\hat{\beta}_j) = \sum_{i=1}^n k_i E(Y_i) = \sum_{i=1}^n k_i(\beta_1 + \beta_2 x_i)$. Typically the LS estimator $\hat{\beta}_j$ will be an unbiased estimator for the parameter β_j that it is estimating.
ii) $\text{V}(\hat{\beta}_j) = \sum_{i=1}^n k_i^2 V(Y_i) = \sigma^2 \sum_{i=1}^n k_i^2$.

Be able to simplify $\sum_{i=1}^n k_i(\beta_1 + \beta_2 x_i)$ and $\sum_{i=1}^n k_i^2$. See HW5 Abc.

22) Other residual plots are also useful. Plot x_j vs r for each predictor variable j in the model and for any potential predictors w_j not in the model. Plot the time order versus r_i if the time order is known. Again, trends and outliers suggest that the model could be improved. An ellipsoidal plot with no trend suggests that the MLR model is good. A parabolic plot suggests adding x_j^2 or w_j and w_j^2 to the MLR model.

23) The **FF plot** of $\hat{Y}_{i,I}$ vs. $\hat{Y}_i$ and the **RR plot** of $r_{i,I}$ vs. r_i can be used to check whether a candidate submodel or reduced model I is good. The submodel is good if the plotted points in the FF and RR plots cluster tightly about the identity line. In the RR plot, the OLS line and identity line can be added to the plot as visual aids. It should be difficult to see that the OLS and identity lines intersect at the origin (the OLS line is the identity line in the FF plot). If the FF plot looks good but the RR plot does not, the submodel may be good if the main goal of the analysis is to predict Y. The two plots are also useful for examining the reduced model in the partial F test. Note that if the candidate model seems to be good, the usual MLR checks should still be made. In particular, the response plot and residual plot (of $\hat{Y}_{i,I}$ vs. $r_{i,I}$) need to be made for the submodel.

24) The plot of the residuals $Y_i - \overline{Y}$ vs. r_i is useful for the Anova F test of $Ho : \beta_2 = \cdots = \beta_p = 0$ vs. Ha: not Ho. If Ho is true, then the plotted points in this special case of the RR plot should cluster tightly about the identity line.

25) The *no intercept MLR model*, also known as *regression through the origin*, is still $Y = \boldsymbol{x}^T\boldsymbol{\beta} + \boldsymbol{e}$, but there is no intercept in the model (no constant $x_1 \equiv 1$). The residual and response plots, $\hat{Y} = \boldsymbol{x}^T\hat{\boldsymbol{\beta}}$, the t tests for Ho $\beta_i = 0$, the CI for β_i, and the partial F test are nearly the same as the usual MLR model. See HW5 Bdef.

26) The 4 step **no intercept Anova F test** is i) Ho $\boldsymbol{\beta} = \mathbf{0}$ Ha $\boldsymbol{\beta} \neq \mathbf{0}$
ii) Fo = MSM/MSE is usually given in output
iii) pval = $P(F_{p,n-p} > Fo)$ is usually given in output
iv) If pval $< \delta$, reject Ho and conclude that there is an MLR relationship between Y and the predictors x_1, ..., x_p. If pval $\geq \delta$, fail to reject Ho, and conclude that there is a not a MLR relationship between Y and the predictors x_1, ..., x_p. See HW5 Bc.

27) A **scatterplot** of x vs. y is used to visualize the conditional distribution of $y|x$. A **scatterplot matrix** is an array of scatterplots. It is used to examine the marginal relationships of the predictors and response. Variable names are on the diagonal. Plots above and below the variable W have W on the horizontal axis. Plots to the left and right of the variable W have W on the vertical axis. It is often useful to transform predictors if strong nonlinearities are apparent in the scatterplot matrix. Be able to tell if the marginal relationships are linear or nonlinear. See HW5 Cbd.

28) Suppose you have a scatterplot of two variables $x_1^{\lambda_1}$ versus $x_2^{\lambda_2}$, $x_1, x_2 > 0$ and that the plotted points follow a nonlinear one to one function. If $\lambda = 0$ use the log transformation $\log(x_i)$. Consider the **ladder of powers** $-1, -0.5, -1/3, 0, 1/3, 0.5$, and 1. **Ladder rule:** To spread small values of the variable, make λ_i smaller. To spread large values of the variable, make λ_i larger. Be able to use the Ladder Rule. See HW5 Ce.

29) A power transformation is $z = t_\lambda(w)$ where z=Y or $z = x_j$ and $t_\lambda(w) = w^\lambda$ for $\lambda \neq 0$ while $t_0(w) = \log(w)$.

i) Suppose that all values of the variable w to be transformed are positive. The **log rule** says use $\log(w)$ if $\max(w_i)/\min(w_i) > 10$. This rule often works wonders on the data and the log transformation is the most used (modified) power transformation. If the variable w can take on the value of 0, use $\log(w + c)$ where c is a small constant like 1, 1/2, or 3/8. Be able to tell which variables in a scatterplot matrix satisfy the log rule. See HW5 Cd, Db.

ii) The **unit rule** says that if x_i and y have the same units, then use the same transformation of x_i and y.

iii) The **cube root rule** says that if w is a volume measurement, then cube root transformation $w^{1/3}$ may be useful.

Consider the ladder of powers given in point 28). No transformation ($\lambda = 1$) is best, then the log transformation, then the square root transformation. Theory, if available, should be used to select a transformation. Frequently more than one transformation will work. For example if y = weight, X_1 = volume = $X_2 * X_3 * X_4$, then y vs. $X_1^{1/3}$ or $\log(y)$ vs. $\log(X_1) = \log(X_2) + \log(X_3) + \log(X_4)$ may both work. Also if y is linearly related with X_2, X_3, X_4 and these three variables all have length units mm, say, then the units of X_1 are $(mm)^3$. Hence the units of $X_1^{1/3}$ are mm.

30) There are also several guidelines for building a MLR model. Suppose that variable Z is of interest and variables $W_1, ..., W_r$ have been collected along with Z. Make a scatterplot matrix of $W_1, ..., W_r$ and Z. (If r is large, several matrices may need to be made. Each one should include Z.) Remove or correct any gross outliers. It is often a good idea to transform the W_i to **remove any strong nonlinearities from the predictors**.

31) Given a scatterplot matrix, be able to tell whether no transformation or the log rule applies. See Q6 1, HW 5 Cd, Da.

32) Given a plot of x versus Y, be able to use the ladder rule to decide between two transformations, one decreasing λ, eg $\log(Y)$, and one increasing λ, eg Y^2. A variant might have a plot of $\sqrt{x}$ versus $\sqrt{Y}$. Then choose between Y and $\log(Y)$ or between x and $\log(x)$. See Q7.

Assume that **all** of the values of the "response" Z_i are **positive**. A *power transformation* has the form $Y = t_\lambda(Z) = Z^\lambda$ for $\lambda \neq 0$ and $Y = t_0(Z) = \log(Z)$ for $\lambda = 0$ where

$$\lambda \in \Lambda_L = \{-1, -1/2, -1/3, 0, 1/3, 1/2, 1\}.$$

The *modified power transformation family*

$$t_\lambda(Z_i) \equiv Z_i^{(\lambda)} = \frac{Z_i^\lambda - 1}{\lambda} \tag{1}$$

for $\lambda \neq 0$ and $Z_i^{(0)} = \log(Z_i)$ where $\lambda \in \Lambda_L$.

A graphical method for response transformations computes the "fitted values" $\hat{W}_i = \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}_\lambda$ from the multiple linear regression model using $W_i = t_\lambda(Z_i)$ as the "response." A *transformation plot* is a plot of $\hat{W}$ versus W with the identity line added as a visual aid. and is made for each of the seven values of $\lambda \in \Lambda_L$. The plotted points follow the identity line in a (roughly) evenly populated band if the iid error MLR model is reasonable for $Y = W$ and $\boldsymbol{x}$. Often TZHAT or YHAT is on the horizontal axis and $Y = t(Z)$ on the vertical axis.

33) Given several transformation plots or several response plots (with $Y = t(Z)$ or $t(Z)$ on the vertical axis), be able to find the response transformation $Y = t(Z)$ corresponding to a plot that looks like a good MLR response plot. Q6, HW6 B, C.

Suppose that the explanatory variables have the form $x_1, ..., x_k$, $x_{jj} = x_j^2$, $x_{ij} = x_i x_j$, $x_{123} = x_1 x_2 x_3$, et cetera. Then the variables $x_1, ..., x_k$ are *main effects.* A product of two or more different main effects is an *interaction.* A variable such as x_1^2 or x_7^3 is a *power.* An $x_1 x_2$ interaction will sometimes also be denoted as $x_1 : x_2$ or $x_1 * x_2$.

A *factor* (with c levels $a_1, ..., a_c$) is incorporated into the MLR model by using $c-1$ indicator variables $x_{Wi} = 1$ if $W = a_i$ and $x_{Wi} = 0$ otherwise, where one of the levels a_i is omitted, eg, use $i = 1, ..., c-1$. The degrees of freedom of the $c-1$ indicator variables is $c-1$.

For **variable selection**, the model $Y = \boldsymbol{x}^T \boldsymbol{\beta} + e$ that uses all of the predictors is called the *full model.* A model $Y = \boldsymbol{x}_I^T \boldsymbol{\beta}_I + e$ that only uses a subset $\boldsymbol{x}_I$ of the predictors is called a *submodel.* The **full model is always a submodel**. The full model has $SP = \boldsymbol{x}^T \boldsymbol{\beta}$ and the submodel has $SP = \boldsymbol{x}_I^T \boldsymbol{\beta}_I$.

Either include all of the indicator variables for a factor in the model or exclude all of them. If the model contains powers or interactions, also include all main effects in the model.

After selecting a submodel I, make the response and residual plots for the full model and the submodel. Make the RR plot of $r_{I,i}$ versus r_i and the FF plot of $\hat{Y}_{I,i}$ versus Y_i. The submodel is good if the plotted points in the FF and RR plots cluster tightly about the identity line. In the RR plot, the OLS line and identity line can be added to the plot as visual aids. It should be difficult to see that the OLS and identity lines intersect at the origin, so the two lines should nearly coincide at the origin. If the FF plot looks good but the RR plot does not, the submodel may be good if the main goal of the analysis is for prediction.

Let I_{min} correspond to the submodel with the smallest C_p. Find the submodel I_I with the fewest number of predictors such that $C_p(I_I) \leq C_p(I_{min}) + 1$. Then I_I is the initial submodel that should be examined. It is possible that $I_I = I_{min}$ or that I_I is the full model. Models I with fewer predictors than I_I such that $C_p(I) \leq C_p(I_{min}) + 4$ are interesting and should also be examined. Models I with k predictors, including a constant and with fewer predictors than I_I such that $C_p(I_{min}) + 4 < C_p(I) \leq \min(2k, p)$ should be checked.

Assume that the full model has good response and residual plots and than $n > 5p$. Let subset I have k predictors, including a constant. The following rules of thumb may be useful, but may not all hold simultaneously. Do not use more predictors than model I_I to avoid overfitting. Then the submodel I is good if
i) the response and residual plots for the submodel looks like the response and residual plots for the full model.
ii) corr(ESP,ESP(I)) = $\text{corr}(\hat{\text{Y}}, \hat{\text{Y}}_{\text{I}}) \geq 0.95$.
iii) The plotted points in the FF plot cluster tightly about the identity line.
iv) Want the p-value ≥ 0.01 for the partial F test that uses I as the reduced model.
v) Want $k \leq n/10$.
vi) The plotted points in the RR plot cluster tightly about the identity line.
vii) Want $R^2(I)$ close to $R^2(full)$ (recall that $R^2(I) \leq R^2(full)$ since adding predictors to I does not decrease $R^2(I)$).
viii) Want $C_p(I_{min}) \leq C_p(I) \leq \min(2k, p)$ with no big jumps in C_p (the increase should be less than four) as variables are deleted.
ix) Want hardly any predictors with p-values > 0.05.
x) Want few predictors with p-values between 0.01 and 0.05.

34) Suppose you are A) given output from forward selection or backward elimination, or B) given a table with several models $L_1, L_2, ..., L_k$ where L_1 is the full model and model I_{min} and perhaps I_I are included. Then be able to find I_{min}, I_I and models with fewer predictors than I_I such that $C_p(I) \leq C_p(I_{min}) + 4$. Know that models with more predictors than I_I and with $C_p(I) > 2k$ should not be used. Q7, HW6 Ebc, HW7 A (min AIC model instead of min C_p model), B, C, D.

Forward selection Step 1) $k = 1$: Start with a constant $w_1 = x_1$. Step 2) $k = 2$: Compute C_p for all models with $k = 2$ containing a constant and a single predictor x_i. Keep the predictor $w_2 = x_j$, say, that minimizes C_p.
Step 3) $k = 3$: Fit all models with $k = 3$ that contain w_1 and w_2. Keep the predictor w_3 that minimizes C_p. ...
Step j) $k = j$: Fit all models with $k = j$ that contains $w_1, w_2, ..., w_{j-1}$. Keep the predictor w_j that minimizes C_p. ...
Step p): Fit the full model.

Backward elimination: All models contain a constant $= u_1$. Step 0) $k = p$: Start

with the full model that contains $x_1, ..., x_p$. We will also say that the full model contains $u_1, ..., u_p$ where $u_1 = x_1$ but u_i need not equal x_i for $i > 1$.
Step 1) $k = p - 1$: Fit each model with $k = p - 1$ predictors including a constant. Delete the predictor u_p, say, that corresponds to the model with the smallest C_p. Keep $u_1, ..., u_{p-1}$.
Step 2) $k = p - 2$: Fit each model with $p - 2$ predictors including a constant. Delete the predictor u_{p-1} corresponding to the smallest C_p. Keep $u_1, ..., u_{p-2}$. ...
Step j) $k = p - j$: fit each model with $p - j$ predictors including a constant. Delete the predictor u_{p-j+1} corresponding to the smallest C_p. Keep $u_1, ..., u_{p-j}$. ...
Step $p - 2$) $k = 2$. The current model contains u_1, u_2 and u_3. Fit the model u_1, u_2 and the model u_1, u_3. Assume that model u_1, u_2 minimizes C_p. Then delete u_3 and keep u_1 and u_2.

35) Consider intervals that contain c cases $(Y_{(1)}, Y_{(c)}), (Y_{(2)}, Y_{(c+1)}), ..., (Y_{(n-c+1)}, Y_{(n)})$. Compute $Y_{(c)} - Y_{(1)}, Y_{(c+1)} - Y_{(2)}, ..., Y_{(n)} - Y_{(n-c+1)}$. Then the estimator shorth(c) $= (Y_{(d)}, Y_{(d+c-1)})$ is the interval with the shortest length. A large sample $100(1-\delta)\%$ prediction interval (PI) $[L_n, U_n]$ is such that $P(Y_f \in [L_n, U_n]) \to 1 - \delta$ as $n \to \infty$. The shorth(c) interval is a large sample $100(1-\delta)\%$ PI if $c/n \to 1 - \delta$ as $n \to \infty$ that often has the asymptotically shortest length.

Most of the rest of the material, 36)–41), will not be tested on Exam 2.

36) Let $W_1, ..., W_n$ be iid random variables from a distribution with cdf F, mean μ and variance σ^2. Let $w_1, ..., w_n$ be the observed values of the X_i. The distribution of the RV D is the *empirical distribution* if D is a discrete RV with the following pmf.

w	w_1	w_2	$\cdots$	w_n
$P(D = w)$	$1/n$	$1/n$	$\cdots$	$1/n$

Then $E(D) = \overline{w} = \frac{1}{n}\sum_{i=1}^n w_i$ and $V(D) = \hat{\sigma}^2_E = \frac{1}{n}\sum_{i=1}^n (w_i - \overline{w})^2$. If the w_i are not distinct, then let k_j = number of $w_i = w_j$, then $P(D = w_j) = k/n$, but this just combines columns in the above table that have $w_i = w_j$. The cdf of D is the empirical cdf F_n. As a statistic (random variable), the empirical cdf $F_n(x) = \frac{1}{n}\sum_{i=1}^n I(W_i \leq x) = \frac{\text{number of } \mathrm{W_i} \leq \mathrm{x}}{n}$. The observed value of the statistic (empirical cdf) is $F_n(x) = \frac{1}{n}\sum_{i=1}^n I(w_i \leq x) = \frac{\text{number of } \mathrm{w_i} \leq \mathrm{x}}{n}$, a nondecreasing step function that can be plotted. Here the indicator random variable $Z_i = I(W_i \leq x) = 1$ if $W_i \leq x$ and $Z_i = I(W_i \leq x) = 0$ if $W_i > x$. Hence the Z_i are iid Bernoulli($q = F(x)$) RVs. Fix x. By the CLT, $\sqrt{n}(F_n(x) - F(x)) \xrightarrow{D} N(0, F(x)(1 - F(x)))$.

37) The *residual bootstrap* computes the least squares estimator and obtains the n residuals and fitted values $r_1, ..., r_n$ and $\hat{Y}_1, ..., \hat{Y}_n$. Then a sample of size n is selected with replacement from the residuals resulting in $r^*_{11}, ..., r^*_{n1}$. Hence the empirical distribution of the residuals is used. Then a vector $\boldsymbol{Y}^*_1 = (Y^*_{11}, ..., Y^*_{n1})^T$ is formed where $Y^*_{i1} = \hat{Y}_i + r^*_{i1}$. Then $\boldsymbol{Y}^*_1$ is regressed on $\boldsymbol{X}$ resulting in the estimator $\hat{\boldsymbol{\beta}}^*_1$. This process is repeated B times resulting in the estimators $\hat{\boldsymbol{\beta}}^*_1, ..., \hat{\boldsymbol{\beta}}^*_B$. This method should have $n \geq 10p$ so that the residuals r_i are close to the errors e_i.

38) If the $\boldsymbol{x}_i = (Y_i, \boldsymbol{x}_i^T)^T$ are iid observations from some population, then a sample of

size n can be drawn with replacement from $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$. Then the response and predictor variables can be formed into vector $\boldsymbol{Y}_1^*$ and design matrix $\boldsymbol{X}_1^*$. Then $\boldsymbol{Y}_1^*$ is regressed on $\boldsymbol{X}_1^*$ resulting in the estimator $\hat{\boldsymbol{\beta}}_1^*$. This process is repeated B times resulting in the estimators $\hat{\boldsymbol{\beta}}_1^*, ..., \hat{\boldsymbol{\beta}}_B^*$. If the $\boldsymbol{x}_i$ are the rows of a matrix $\boldsymbol{X}$, then this *nonparametric bootstrap* uses the empirical distribution of the $\boldsymbol{x}_i$.

39) Suppose $\boldsymbol{w}_1, ..., \boldsymbol{w}_n$ are iid $p \times 1$ random vectors with mean $\boldsymbol{\mu}$ and nonsingular covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{w}}$. Let a future test observation $\boldsymbol{w}_f$ be independent of the $\boldsymbol{w}_i$ but from the same distribution. Let $(\overline{\boldsymbol{w}}, \boldsymbol{S})$ be the sample mean and sample covariance matrix where

$$\overline{\boldsymbol{w}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{w}_i \quad \text{and} \quad \boldsymbol{S} = \boldsymbol{S}_{\boldsymbol{w}} = \frac{1}{\mathrm{n}-1}\sum_{\mathrm{i}=1}^{\mathrm{n}} (\boldsymbol{w}_{\mathrm{i}} - \overline{\boldsymbol{w}})(\boldsymbol{w}_{\mathrm{i}} - \overline{\boldsymbol{w}})^{\mathrm{T}}. \tag{2}$$

Then the ith *squared sample Mahalanobis distance* is the scalar

$$D^2_{\boldsymbol{w}} = D^2_{\boldsymbol{w}}(\overline{\boldsymbol{w}}, \boldsymbol{S}) = (\boldsymbol{w} - \overline{\boldsymbol{w}})^T \boldsymbol{S}^{-1} (\boldsymbol{w} - \overline{\boldsymbol{w}}). \tag{3}$$

Let $D_i^2 = D^2_{\boldsymbol{w}_i}$ for each observation $\boldsymbol{w}_i$. Let $D_{(c)}$ be the cth order statistic of $D_1, ..., D_n$. Consider the hyperellipsoid

$$\mathcal{A}_n = \{\boldsymbol{w} : D^2_{\boldsymbol{w}}(\overline{\boldsymbol{w}}, \boldsymbol{S}) \le D^2_{(c)}\} = \{\boldsymbol{w} : D_{\boldsymbol{w}}(\overline{\boldsymbol{w}}, \boldsymbol{S}) \le D_{(c)}\}. \tag{4}$$

If n is large, we can use $c = k_n = \lceil n(1-\delta) \rceil$. If n is not large, using $c = U_n$ where U_n decreases to k_n, can improve small sample performance. Then (4) is a large sample $100(1-\delta)\%$ prediction region for a large class of distributions, although regions with smaller volumes may exist.

40) Applying (4) to a bootstrap sample of $T_1^*, ..., T_B^*$ results in a large sample $100(1-\delta)\%$ confidence region for $\boldsymbol{\mu}$, where one sufficient condition is $\sqrt{n}(T_n - \boldsymbol{\mu}) \xrightarrow{D} N_r(\boldsymbol{0}, \boldsymbol{\Sigma}_T)$. If $r = 1$, applying the shorth(c) interval to $T_1^*, ..., T_B^*$ results in a large sample $100(1-\delta)\%$ confidence interval for μ

41) We can also apply a shorth(c_n) estimator to the residuals, getting an interval $[r_{(d)}, r_{(d+c_n-1)}]$ where c_n decreases to $\lceil n(1-\delta) \rceil$. Then a large sample $100(1-\delta)\%$ prediction interval (PI) for Y_F is $[\hat{Y}_F + a_n r_{(d)}, \hat{Y}_F + a_n r_{(d+c_n-1)}]$ where $a_n \ge 1$ and $a_n \to 1$. This PI can work if the iid errors $e_1, ..., e_n$ are from an unknown distribution, and after variable selection if $n \ge 10d$ where $\hat{\boldsymbol{\beta}}_I$ is $d \times 1$.