Math 584 Exam 2 is on Tuesday, March. 23. You are allowed 9 sheets of notes and a calculator. CHECK FORMULAS!

37) **Know:** If $\boldsymbol{Z}_n \xrightarrow{D} \boldsymbol{Z}$, then $\boldsymbol{Z}$ is the limiting distribution of $\boldsymbol{Z}_n$ and does not depend on n (since $\boldsymbol{Z}$ is found by taking a limit as $n \to \infty$).

Often $\boldsymbol{Z}_n \xrightarrow{D} N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\boldsymbol{Z}_n \sim N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ behave similarly (compare 7) and 36)). A big difference is that the distribution on the RHS (right hand side) can depend on n for $\sim$ but not for $\xrightarrow{D}$.

38) **Know:** Often want a normal approximation where the RHS can depend on n. Write $\boldsymbol{Z}_n \sim AN_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for an approximate multivariate normal distribution where the RHS may depend on n. For the model in 35), if $\boldsymbol{\epsilon} \sim N_n(\boldsymbol{0}, \sigma^2 I)$, then
$\hat{\boldsymbol{\beta}} \sim N_p(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1})$. If the ϵ_i are iid with $E(\epsilon_i) = 0$ and $V(\epsilon_i) = \sigma^2$, use the multivariate normal approximation $\hat{\boldsymbol{\beta}} \sim AN_p(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1})$ or $\hat{\boldsymbol{\beta}} \sim AN_p(\boldsymbol{\beta}, MSE(\boldsymbol{X}^T\boldsymbol{X})^{-1})$. The RHS depends on n since the number of rows of $\boldsymbol{X}$ is n.

39) Suppose $\hat{\boldsymbol{\Sigma}}_n$ is positive definite and symmetric. If $\boldsymbol{W}_n \xrightarrow{D} N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\hat{\boldsymbol{\Sigma}}_n \xrightarrow{P} \boldsymbol{\Sigma}$, then $\boldsymbol{Z}_n = \hat{\boldsymbol{\Sigma}}_n^{-1/2}(\boldsymbol{W}_n - \boldsymbol{\mu}) \xrightarrow{D} N_k(\boldsymbol{0}, \boldsymbol{I})$, and $\boldsymbol{Z}_n^T\boldsymbol{Z}_n = (\boldsymbol{W}_n - \boldsymbol{\mu})^T\hat{\boldsymbol{\Sigma}}_n^{-1}(\boldsymbol{W}_n - \boldsymbol{\mu}) \xrightarrow{D} \chi^2_k$.

40) Let $\boldsymbol{x} = (1 \;\; \boldsymbol{u}^T)^T$ where $\boldsymbol{u}$ is the vector of nontrivial predictors. Let the sample mean and sample covariance matrix of the nontrivial predictors be $\overline{\boldsymbol{u}} = \dfrac{1}{n}\sum_{i=1}^n \boldsymbol{u}_i$ and $\boldsymbol{C}_{\boldsymbol{u}} = \dfrac{1}{n-1}\sum_{i=1}^n (\boldsymbol{u}_i - \overline{\boldsymbol{u}})(\boldsymbol{u}_i - \overline{\boldsymbol{u}})^T$. Let the ith *squared Mahalanobis distance* $\mathrm{MD}_\mathrm{i}^2 = (\boldsymbol{u}_\mathrm{i} - \overline{\boldsymbol{u}})^\mathrm{T}\boldsymbol{C}_{\boldsymbol{u}}^{-1}(\boldsymbol{u}_\mathrm{i} - \overline{\boldsymbol{u}})$. Then $h_i = \frac{1}{n-1}\mathrm{MD}_\mathrm{i}^2 + \frac{1}{\mathrm{n}}$. Then $\mathrm{MD}_\mathrm{i}^2 = \mathrm{d}^2$ is the equation of a hyperellipsoid. Points that lie on the hyperellipsoid all have $\mathrm{MD}_\mathrm{i}^2 = \mathrm{d}^2$. The MD_i^2 tend to be bounded in probability ($\mathrm{MD}_\mathrm{i}^2 \approx \chi^2_{\mathrm{p}-1}$ if the $\boldsymbol{u}_i$ are iid MVN). Hence $\max(h_1, ..., h_n) \xrightarrow{P} 0$ as $n \to \infty$ is considered to be a mild assumption.

41) Let $f(\boldsymbol{y}|\boldsymbol{\theta})$ be the joint pdf of $Y_1, ..., Y_n$. If $\boldsymbol{Y} = \boldsymbol{y}$ is observed, then **the likelihood function** $L(\boldsymbol{\theta}) = f(\boldsymbol{y}|\boldsymbol{\theta})$. Note: it is crucial to observe that the likelihood function is a function of $\boldsymbol{\theta}$ (and that $y_1, ..., y_n$ act as fixed constants).

42) For each sample point $\boldsymbol{y} = (y_1, ..., y_n)$, let $\hat{\boldsymbol{\theta}}(\boldsymbol{y})$ be a parameter value at which $L(\boldsymbol{\theta}|\boldsymbol{y})$ attains its maximum as a function of $\boldsymbol{\theta}$ with $\boldsymbol{y}$ held fixed. Then a maximum likelihood estimator (**MLE**) of the parameter $\boldsymbol{\theta}$ based on the sample $\boldsymbol{Y}$ is $\hat{\boldsymbol{\theta}}(\boldsymbol{Y})$.

Note: If the MLE $\hat{\boldsymbol{\theta}}$ exists, then $\hat{\boldsymbol{\theta}} \in \Theta$, the parameter space.

43) Know how to find the max and min of a function h that is continuous on an interval [a,b] and differentiable on (a, b). Solve $h'(x) \equiv 0$ and find the places where $h'(x)$ does not exist. These values are the **critical points**. Evaluate h at a, b, and the critical points. One of these values will be the min and one the max.

Assume h is continuous. Then a critical point θ_o is a local max of $h(\theta)$ if h is increasing for $\theta < \theta_o$ in a neighborhood of θ_o and if h is decreasing for $\theta > \theta_o$ in a neighborhood of θ_o. The first derivative test is often used.

If h is strictly concave ($\dfrac{d^2}{d\theta^2}h(\theta) < 0$ for all θ), then any local max of h is a global max.

Suppose $h'(\theta_o) = 0$. The 2nd derivative test states that if $\dfrac{d^2}{d\theta^2}h(\theta_o) < 0$, then θ_o is a local max.

If $h(\theta)$ is a continuous function on an interval with endpoints $a < b$ (not necessarily finite), and differentiable on (a, b) and if the **critical point is unique**, then the critical point is a **global maximum** if it is a local maximum (because otherwise there would be a local minimum and the critical point would not be unique). To show that $\hat{\theta}$ is the MLE (the global maximizer of $\log L(\theta)$), show that $\log L(\theta)$ is differentiable on (a, b) where Θ may contain the endpoints a and b. Then show that $\hat{\theta}$ is the unique solution to the equation $\frac{d}{d\theta} \log L(\theta) = 0$ and that the 2nd derivative evaluated at $\hat{\theta}$ is negative: $\frac{d^2}{d\theta^2} \log L(\theta)|_{\hat{\theta}} < 0$.

44) **Know:** In addition to differentiating the log likelihood, the MLE can sometime be found by directly maximization of the likelihood $L(\boldsymbol{\theta})$. For regression, $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma)$ or $(\boldsymbol{\beta}, \sigma^2)$. Can often fix σ and then show $\hat{\boldsymbol{\beta}}$ is the MLE by direct maximization. The the MLE $\hat{\sigma}$ or $\hat{\sigma}^2$ can be found by maximizing the log profile likelihood function $\log[L_p(\sigma, \hat{\boldsymbol{\beta}})]$ or $\log[L_p(\sigma^2, \hat{\boldsymbol{\beta}})]$ where $L_p(\sigma, \hat{\boldsymbol{\beta}}) = L(\sigma, \boldsymbol{\beta} = \hat{\boldsymbol{\beta}})$. See HW5 1 and Q4 1.

45) **Orthogonal regression**: let $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$ where $\boldsymbol{X}$ is full rank p and $\boldsymbol{X} = [\boldsymbol{v}_0 \ \boldsymbol{v}_1 \ ... \ \boldsymbol{v}_{p-1}]$. Suppose the columns of $\boldsymbol{X}$ are orthogonal so $\boldsymbol{v}_i^T \boldsymbol{v}_j = 0$ for $i \neq j$. Then $\boldsymbol{X}^T\boldsymbol{X} = diag(\boldsymbol{v}_i^T \boldsymbol{v}_i)$ and $(\boldsymbol{X}^T\boldsymbol{X})^{-1} = diag(1/(\boldsymbol{v}_i^T \boldsymbol{v}_i))$. Then $\hat{\beta}_j = \dfrac{\boldsymbol{v}_j^T \boldsymbol{y}}{\boldsymbol{v}_j^T \boldsymbol{v}_j}$ for $j = 0, 1, ..., p-1$. Also, the $\hat{\boldsymbol{\beta}}_j$ remain unchanged if columns of $\boldsymbol{X}$ other than $\boldsymbol{v}_j$ are deleted.

46)-51) are for the nonfull rank linear model.

46) **Know:** The **nonfull rank linear model:** suppose $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$ where $\boldsymbol{X}$ has rank $r < p$ and $\boldsymbol{X}$ is an $n \times p$ matrix.

i) $\boldsymbol{P}_{\boldsymbol{X}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T$ is the unique projection matrix on $C(\boldsymbol{X})$ and does not depend on the generalized inverse $(\boldsymbol{X}^T\boldsymbol{X})^-$. (Recall that projection matrices are symmetric and idempotent but singular unless $\boldsymbol{P}_{\boldsymbol{X}} = \boldsymbol{I}$. Also recall that $\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{X} = \boldsymbol{X}$, so $\boldsymbol{X}^T\boldsymbol{P}_{\boldsymbol{X}} = \boldsymbol{X}^T$.)

ii) $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{Y}$ does depend on $(\boldsymbol{X}^T\boldsymbol{X})^-$ and is not unique.

iii) $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{Y}$, $\boldsymbol{e} = \boldsymbol{Y} - \hat{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} = (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{Y}$ and $RSS = \boldsymbol{e}^T\boldsymbol{e}$ are unique and so do not depend on $(\boldsymbol{X}^T\boldsymbol{X})^-$.

iv) $\hat{\boldsymbol{\beta}}$ is a solution to the *normal equations*: $\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}^T\boldsymbol{Y}$.

v) It can be shown that $\text{rank}(\boldsymbol{P}_{\boldsymbol{X}}) = r$ and $\text{rank}(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}}) = n - r$.

vi) Let $\hat{\boldsymbol{\theta}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\theta} = \boldsymbol{X\theta}$. Suppose there exists a constant vector $\boldsymbol{c}$ such that $E(\boldsymbol{c}^T\hat{\boldsymbol{\theta}}) = \boldsymbol{c}^T\boldsymbol{\theta}$. Then among the class of linear unbiased estimators of $\boldsymbol{c}^T\boldsymbol{\theta}$, the least squares estimator $\boldsymbol{c}^T\hat{\boldsymbol{\theta}}$ is BLUE.

vii) If $\text{Cov}(\boldsymbol{Y}) = \text{Cov}(\boldsymbol{\epsilon}) = \sigma^2\boldsymbol{I}$, then $MSE = \dfrac{RSS}{n-r} = \dfrac{\boldsymbol{e}^T\boldsymbol{e}}{n-r}$ is an unbiased estimator of σ^2.

viii) Let the columns of $\boldsymbol{X}_1$ form a basis for $C(\boldsymbol{X})$. For example, take r linearly independent columns of $\boldsymbol{X}$ to form $\boldsymbol{X}_1$. Then $\boldsymbol{P}_{\boldsymbol{X}} = \boldsymbol{X}_1(\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^T$.

47) **Know:** Let $\boldsymbol{a}$ and $\boldsymbol{b}$ be constant vectors. Then $\boldsymbol{a}^T\boldsymbol{\beta}$ is **estimable** if there exists a linear unbiased estimator $\boldsymbol{b}^T\boldsymbol{Y}$ so $E(\boldsymbol{b}^T\boldsymbol{Y}) = \boldsymbol{a}^T\boldsymbol{\beta}$.

48) **Know:** The quantity $\boldsymbol{a}^T\boldsymbol{\beta}$ is estimable iff $\boldsymbol{a}^T = \boldsymbol{b}^T\boldsymbol{X}$ iff $\boldsymbol{a} = \boldsymbol{X}^T\boldsymbol{b}$ (for some

constant vector $\boldsymbol{b}$) iff $\boldsymbol{a} \in C(\boldsymbol{X}^T)$.

49) If $\boldsymbol{a}^T\boldsymbol{\beta}$ is estimable and a least squares estimator $\hat{\boldsymbol{\beta}}$ is any solution to the normal equations $\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}^T\boldsymbol{Y}$. Then $\boldsymbol{a}^T\boldsymbol{\beta}$ is unique and $\boldsymbol{a}^T\hat{\boldsymbol{\beta}}$ is the BLUE of $\boldsymbol{a}^T\boldsymbol{\beta}$.

50) The term "estimable" is misleading since there are nonestimable quantities $\boldsymbol{a}^T\boldsymbol{\beta}$ that can be estimated with biased or nonlinear estimators.

51) Estimable quantities tend to go with the nonfull rank linear model. Can avoid nonestimable functions by using a full rank model instead of a nonfull rank model (delete columns of $\boldsymbol{X}$ until it is full rank).

Back to the full rank linear model.

52) The **Gauss Markov theorem:** Let $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$ where $\boldsymbol{X}$ is full rank p, $E(\boldsymbol{\epsilon}) = \boldsymbol{0}$ and $\text{Cov}(\boldsymbol{\epsilon}) = \sigma^2\boldsymbol{I}$. Then $\boldsymbol{a}^T\hat{\boldsymbol{\beta}}$ is the BLUE for $\boldsymbol{a}^T\boldsymbol{\beta}$ for any constant $p \times 1$ vector $\boldsymbol{a}$.

(Also see 32 b).)

53) The *generalized least squares* (GLS) model is $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$ where $\boldsymbol{Y}$ is an $n \times 1$ vector of dependent variables, $\boldsymbol{X}$ has full rank p, $E(\boldsymbol{\epsilon}) = \boldsymbol{0}$, and $\text{Cov}(\boldsymbol{\epsilon}) = \sigma^2\boldsymbol{V}$ where $\boldsymbol{V}$ is a **known** $n \times n$ symmetric positive definite matrix. The least squares (LS) or ordinary least squares (OLS) model is the special case where $\boldsymbol{V} = \boldsymbol{I}$.

54) The *weighted least squares* (WLS) model with weights $w_1, ..., w_n$ is the special case of the GLS model where $\boldsymbol{V}$ is diagonal: $\boldsymbol{V} = \text{diag}(\text{v}_1, ..., \text{v}_\text{n})$ and $w_i = 1/v_i$.

55) The *feasible generalized least squares* (FGLS) model is the same as the GLS estimator except that $\boldsymbol{V} = \boldsymbol{V}(\boldsymbol{\theta})$ is a function of an unknown $q \times 1$ vector of parameters $\boldsymbol{\theta}$. Let the estimator of $\boldsymbol{V}$ be $\hat{\boldsymbol{V}} = \boldsymbol{V}(\hat{\boldsymbol{\theta}})$. The *feasible weighted least squares* (FWLS) estimator is the special case of the FGLS estimator where $\boldsymbol{V} = \boldsymbol{V}(\boldsymbol{\theta})$ is diagonal. Hence the estimated weights $\hat{w}_i = 1/\hat{v}_i = 1/v_i(\hat{\boldsymbol{\theta}})$.

56) The *GLS estimator* $\hat{\boldsymbol{\beta}}_{GLS} = (\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{Y}$. The fitted values are $\hat{\boldsymbol{Y}}_{GLS} = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{GLS}$.

The *WLS estimator* $\hat{\boldsymbol{\beta}}_{WLS} = (\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{Y}$. The fitted values are $\hat{\boldsymbol{Y}}_{WLS} = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{WLS}$.

Then the FGLS estimator $\hat{\boldsymbol{\beta}}_{FGLS} = (\boldsymbol{X}^T\hat{\boldsymbol{V}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\hat{\boldsymbol{V}}^{-1}\boldsymbol{Y}$. The fitted values are $\hat{\boldsymbol{Y}}_{FGLS} = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{FGLS}$. The FWLS estimator and fitted values will be denoted by $\hat{\boldsymbol{\beta}}_{FWLS}$ and $\hat{\boldsymbol{Y}}_{FWLS}$, respectively.

57) It can be shown that the GLS estimator minimizes the GLS criterion

$$Q_{GLS}(\boldsymbol{\eta}) = (\boldsymbol{Y} - \boldsymbol{X\eta})^T\boldsymbol{V}^{-1}(\boldsymbol{Y} - \boldsymbol{X\eta}).$$

Notice that the FGLS and FWLS estimators have $p + q + 1$ unknown parameters. These estimators can perform very poorly if $n < 10(p + q + 1)$.

58) There is a symmetric, nonsingular $n \times n$ matrix $\boldsymbol{R} = \boldsymbol{V}^{1/2}$ (the square root matrix of $\boldsymbol{V}$) such that $\boldsymbol{V} = \boldsymbol{RR}$. Let $\boldsymbol{Z} = \boldsymbol{R}^{-1}\boldsymbol{Y}$, $\boldsymbol{U} = \boldsymbol{R}^{-1}\boldsymbol{X}$ and $\boldsymbol{a} = \boldsymbol{R}^{-1}\boldsymbol{\epsilon}$. This method uses the spectral theorem (singular value decomposition).

59) **GLS as OLS Theorem**: a) $\boldsymbol{Z} = \boldsymbol{U\beta} + \boldsymbol{a}$ follows the OLS model since $E(\boldsymbol{a}) = \boldsymbol{0}$ and $\text{Cov}(\boldsymbol{a}) = \sigma^2\boldsymbol{I}_n$.

b) The GLS estimator $\hat{\boldsymbol{\beta}}_{GLS}$ can be obtained from the OLS regression (without an intercept) of $\boldsymbol{Z}$ on $\boldsymbol{U}$.

c) For WLS, $Y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \epsilon_i$. The corresponding OLS model $\boldsymbol{Z} = \boldsymbol{U}\boldsymbol{\beta} + \boldsymbol{a}$ is equivalent to $Z_i = \boldsymbol{u}_i^T \boldsymbol{\beta} + a_i$ for $i = 1, ..., n$ where $\boldsymbol{u}_i^T$ is the ith row of $\boldsymbol{U}$. Then $Z_i = \sqrt{w_i}\ Y_i$ and $\boldsymbol{u}_i = \sqrt{w_i}\ \boldsymbol{x}_i$. Hence $\hat{\boldsymbol{\beta}}_{WLS}$ can be obtained from the OLS regression (without an intercept) of $\boldsymbol{Z}_i = \sqrt{w_i}\ Y_i$ on $\boldsymbol{u}_i = \sqrt{w_i}\ \boldsymbol{x}_i$.

60) The FGLS estimator can also be found from the OLS regression (without an intercept) of $\boldsymbol{Z}$ on $\boldsymbol{U}$ where $\boldsymbol{V}(\hat{\boldsymbol{\theta}}) = \boldsymbol{R}\boldsymbol{R}$. Similarly the FWLS estimator can be found from the OLS regression (without an intercept) of $Z_i = \sqrt{\hat{w}_i} Y_i$ on $\boldsymbol{u}_i = \sqrt{\hat{w}_i} \boldsymbol{x}_i$. But now $\boldsymbol{U}$ is a random matrix instead of a constant matrix. Hence these estimators are highly nonlinear.

61) Under regularity conditions, the OLS estimator $\hat{\boldsymbol{\beta}}_{OLS}$ is a consistent estimator of $\boldsymbol{\beta}$ when the GLS model holds ($\text{Cov}(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{V}$), but $\hat{\boldsymbol{\beta}}_{GLS}$ should be used because it generally has higher efficiency.

Ch. 4 Hypothesis Testing

62) Let $\boldsymbol{A}$ be a $q \times p$ constant matrix with $\text{rank}(\boldsymbol{A}) = q$, let $\boldsymbol{c}$ be a $q \times 1$ constant vector, and consider testing $H_0 : \boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{c}$. If $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ where $\text{rank}(\boldsymbol{X}) = p$, $E(\boldsymbol{\epsilon}) = \boldsymbol{0}$ and $\text{Cov}(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{I}$, then $\hat{\boldsymbol{\beta}} \sim N_p(\boldsymbol{\beta}, \sigma^2 (\boldsymbol{X}^T \boldsymbol{X})^{-1})$, and $\boldsymbol{A}\hat{\boldsymbol{\beta}} - \boldsymbol{c} \sim N_q(\boldsymbol{A}\boldsymbol{\beta} - \boldsymbol{c}, \sigma^2 \boldsymbol{A}(\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{A}^T)$. If H_0 is true then $\boldsymbol{A}\hat{\boldsymbol{\beta}} - \boldsymbol{c} \sim N_q(\boldsymbol{0}, \sigma^2 \boldsymbol{A}(\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{A}^T)$, and

$$qF = \frac{1}{\sigma^2} (\boldsymbol{A}\hat{\boldsymbol{\beta}} - \boldsymbol{c})^T [\boldsymbol{A}(\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{A}^T]^{-1} (\boldsymbol{A}\hat{\boldsymbol{\beta}} - \boldsymbol{c}) \sim \chi^2_q.$$

63) If H_0 is true, then by the LS CLT, $qF \xrightarrow{D} \chi^2_q$ for a large class of zero mean error distributions.

64) **Know:** The partial F test, and its special cases the Anova F test and the Wald t test, use $\boldsymbol{c} = \boldsymbol{0}$. Let the **full model** use Y, $x_0 \equiv 1$, $x_1, ..., x_{p-1}$, and let the **reduced model** use Y, x_0, $x_{j_1}, ..., x_{j_k}$ where $\{j_1, ..., j_k\} \subset \{1, ..., p-1\}$. Here $0 \leq k < p - 1$, and if $k = 0$, then the model is $Y_i = \beta_0 + \epsilon_i$. Hence the full model is $Y_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_{p-1} x_{i,p-1} + \epsilon_i$, while the reduced model is $Y_i = \beta_0 + \beta_{j_1} x_{i,j_1} + \cdots + \beta_{j_k} x_{i,j_k} + \epsilon_i$. In matrix form, the full model is $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ and the reduced model is $\boldsymbol{Y} = \boldsymbol{X}_R \boldsymbol{\beta}_R + \boldsymbol{\epsilon}$ where the columns of $\boldsymbol{X}_R$ are a proper subset of the columns of $\boldsymbol{X}$. i) The **partial F test** has $H_0 : \beta_{j_{k+1}} = \cdots = \beta_{j_{p-1}} = 0$, or H_0 : the reduced model is good, or $H_0 : \boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{0}$ where $\boldsymbol{A}$ is a $p - k - 1 \times p$ matrix where the ith row of $\boldsymbol{A}$ has a 1 in the j_{k+i}th position and zeroes elsewhere. In particular, if $\beta_0, ..., \beta_k$ are the only β_i in the reduced model, then $\boldsymbol{A} = [\boldsymbol{0} \ \ \boldsymbol{I}_{p-k-1}]$ and $\boldsymbol{0}$ is a $(p - k - 1) \times (k + 1)$ matrix. Hence $q = p - k + 1 =$ number of predictors in the full model but not in the reduced model. ii) The **Anova F test** is the special case of the partial F test where the reduced model is $Y_i = \beta_0 + \epsilon_i$. Hence $H_0 : \beta_1 = \cdots = \beta_{p-1} = 0$, or H_0 : none of the nontrivial predictors $x_1, ..., x_{p-1}$ are needed in the linear model, or $H_0 : \boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{0}$ where $\boldsymbol{A} = [\boldsymbol{0} \ \ \boldsymbol{I}_{p-1}]$ and $\boldsymbol{0}$ is a $(p - 1) \times 1$ vector. Hence $q = p - 1$. iii) The **Wald t test** uses the reduced model that deletes the jth predictor from the full model. Hence $H_0 : \beta_j = 0$, or H_0 : the jth predictor x_j is not needed in the linear model given that the other predictors are in the model, or $H_0 : \boldsymbol{A}_j \boldsymbol{\beta} = 0$ where $\boldsymbol{A}_j = [0, ..., 0, 1, 0, ..., 0]$ is a $1 \times p$ row vector with a 1 in the $j + 1$ position for $j = 0, ..., p - 1$. Hence $q = 1$.

65) A way to get the test statistic F_R for the partial F test is to fit the full model and the reduced model. Let $RSS(F)$ be the RSS of the full model, and let $RSS(R)$ be

the RSS of the reduced model. Similarly, let $MSE(F)$ be the MSE of the full model. Let $df_R = n - k - 1$ and $df_F = n - p$ be the degrees of freedom for the reduced and full models. Then $F_R = \dfrac{RSS(R) - RSS(F)}{qMSE(F)}$ where $q = df_R - df_F = p - k - 1 =$ number of predictors in the full model but not in the reduced model.

66) If $X_n \sim F_{q,d_n}$ where the positive integer $d_n \to \infty$ as $n \to \infty$, then $qX_n \xrightarrow{D} \chi^2_q$.

67) A test with test statistic T_n is a *large sample right tail δ test* if the test rejects H_0 if $T_n > a_n$ and $P(T_n > a_n) = \delta_n \to \delta$ as $n \to \infty$ when H_0 is true. Typically want $\delta \le 0.1$ and the values $\delta = 0.05$ or $\delta = 0.01$ are common. (An analogy is a large sample $100(1-\delta)\%$ confidence interval or prediction interval.)

Suppose when H_0 is true, $T_n \xrightarrow{D} \chi^2_q$. Suppose $P(W \le \chi^2_q(1-\delta)) = 1 - \delta$ and $P(W > \chi^2_q(1-\delta)) = \delta$ where $W \sim \chi^2_q$. Suppose $P(W \le F_{q,d_n}(1-\delta)) = 1-\delta$ when $W \sim F_{q,d_n}$. Also write $\chi^2_q(1-\delta) = \chi^2_{q,1-\delta}$ and $F_{q,d_n}(1-\delta) = F_{q,d_n,1-\delta}$. Then a test that rejects H_0 if $T_n > \chi^2_q(1-\delta)$ is a large sample right tail δ test. Also, a test that rejects H_0 if $T_n/q > F_{q,d_n}(1-\delta)$ is a large sample right tail δ test if the positive integer $d_n \to \infty$ as $n \to \infty$.

Suppose when H_0 is true, $T_n \xrightarrow{D} N(0,1)$. Suppose $P(W > Z(1-\delta)) = \delta$ when $W \sim N(0,1)$, and $P(W > t_{d_n}(1-\delta)) = \delta$ when $W \sim t_{d_n}$. Then a test that rejects H_0 if $T_n > Z(1-\delta)$ is a large sample right tail δ test. Also, a test that rejects H_0 if $T_n > t_{d_n}(1-\delta)$ is a large sample right tail δ test if the positive integer $d_n \to \infty$ as $n \to \infty$.

68) Large sample t tests and intervals are used instead of Z tests and intervals since the t tests and intervals are more accurate for moderate n. Large sample F tests and intervals are used instead of χ_2 tests and intervals since the F tests and intervals are more accurate for moderate n.

69) **Partial F Test Theorem:** Suppose $H_0 : \boldsymbol{A\beta} = \boldsymbol{0}$ is true for the partial F test. Under the OLS full rank model, a)

$$F_R = \frac{1}{qMSE}(\boldsymbol{A}\hat{\boldsymbol{\beta}})^T[\boldsymbol{A}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{A}^T]^{-1}(\boldsymbol{A}\hat{\boldsymbol{\beta}}).$$

b) If $\boldsymbol{\epsilon} \sim N_n(\boldsymbol{0}, \sigma^2\boldsymbol{I})$, then $F_R \sim F_{q,n-p}$.

c) For a large class of zero mean error distributions $qF_R \xrightarrow{D} \chi^2_q$.

d) The partial F test that rejects $H_0 : \boldsymbol{A\beta} = \boldsymbol{0}$ if $F_R > F_{q,n-p}(1-\delta)$ is a large sample right tail δ test for the OLS model for a large class of zero mean error distributions.

70) Let $X \sim t_{n-p}$. Then $X^2 \sim F_{1,n-p}$. The two tail Wald t test for $H_0 : \beta_j = 0$ versus $H_1 : \beta_j \neq 0$ is equivalent to the corresponding right tailed F test since rejecting H_0 if $|X| > t_{n-p}(1-\delta)$ is equivalent to rejecting H_0 if $X^2 > F_{1,n-p}(1-\delta)$.

71) The **pvalue** of a test is the probability, assuming H_0 is true, of observing a test statistic as extreme as the test statistic T_n actually observed. For a right tail test, pvalue $= P_{H_0}$(of observing a test statistic $\ge T_n$). Under the OLS model where $F_R \sim F_{q,n-p}$ when H_0 is true (so the ϵ_i are iid $N(0,\sigma^2)$), the pvalue $= P(W > F_R)$ where $W \sim F_{q,n-p}$. In general can only estimate the pvalue. Let pval be the estimated pvalue. Then pval $= P(W > F_R)$ where $W \sim F_{q,n-p}$, and pval $\xrightarrow{P}$ pvalue an $n \to \infty$ for the large sample partial F test. The pvalues in output are usually actually pvals (estimated pvalues).

72) Often $n > 10p$ starts to give good results for the OLS output for error distributions not too far from $N(0,1)$.

73) Let $\boldsymbol{P}$ and $\boldsymbol{P}_1$ be the projection matrices on $\boldsymbol{X}$ and $\boldsymbol{X}_1$ where $\boldsymbol{X} = [\boldsymbol{X}_1 \;\; \boldsymbol{X}_2]$, $\boldsymbol{\beta}^T = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)$, the full model is $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$, $H_0 : \boldsymbol{\beta}_2 = \boldsymbol{0}$, $\boldsymbol{X}$ has full rank p, and $\boldsymbol{\epsilon} \sim N_n(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$. Assume H_0 holds. Then the reduced model is $\boldsymbol{Y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\epsilon}$. Also, i) $\dfrac{\boldsymbol{Y}^T(\boldsymbol{P} - \boldsymbol{P}_1)\boldsymbol{Y}}{\sigma^2} \sim \chi^2_q \perp\!\!\!\perp \dfrac{\boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{Y}}{\sigma^2} \sim \chi^2_{n-p}$ where $rank(\boldsymbol{I}-\boldsymbol{P}) = trace(\boldsymbol{I}-\boldsymbol{P}) = n-p$ and $q = rank(\boldsymbol{P} - \boldsymbol{P}_1) = trace(\boldsymbol{P} - \boldsymbol{P}_1) = p - d$ if $\boldsymbol{X}_1$ is $n \times d$. Note that q is the number of predictors in the full model that are not in the reduced model. Also, ii) $F_R = \dfrac{n-p}{q} \dfrac{\boldsymbol{Y}^T(\boldsymbol{P} - \boldsymbol{P}_1)\boldsymbol{Y}}{\boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{Y}} \sim F_{q,n-p}$. Also $W = \dfrac{X_1/d_1}{X_2/d_2} \sim F_{d_1,d_2}$ if $X_1 \sim \chi^2_{d_1} \perp\!\!\!\perp X_2 \sim \chi^2_{d_2}$.

74) (**Population OLS Coefficients**): Let $\boldsymbol{x}_i^T = (1 \;\; \boldsymbol{u}_i^T)$ where $\boldsymbol{u}_i$ is the vector of nontrivial predictors. Let $\dfrac{1}{n}\sum_{j=1}^n X_{jk} = \overline{X}_{ok} = \overline{u}_{ok}$ for $k = 1, ..., p-1$. The subscript "ok" means sum over the first subscript j. Let $\overline{\boldsymbol{u}} = (\overline{u}_{o,1}, ..., \overline{u}_{o,p-1})^T$ be the sample mean of the $\boldsymbol{u}_i$. Let $\boldsymbol{\beta}^T = (\beta_0 \;\; \boldsymbol{\beta}_S^T)$ where the slopes vector $\boldsymbol{\beta}_S = (\beta_1, ..., \beta_{p-1})^T$. Let the population covariance matrices $\text{Cov}(\boldsymbol{u}) = E[(\boldsymbol{u} - E(\boldsymbol{u}))(\boldsymbol{u} - E(\boldsymbol{u}))^T] = \boldsymbol{\Sigma}_{\boldsymbol{u}}$ and $\text{Cov}(\boldsymbol{u}, Y) = E[(\boldsymbol{u} - E(\boldsymbol{u}))(Y - E(Y))] = \boldsymbol{\Sigma}_{\boldsymbol{u}Y}$. Then the population coefficients from an OLS regression of Y on $\boldsymbol{u}$ (even if a linear model does not hold) are

$$\beta_0 = E(Y) - \boldsymbol{\beta}_S^T E(\boldsymbol{u}) \;\; \text{and} \;\; \boldsymbol{\beta}_{\text{S}} = \boldsymbol{\Sigma}_{\boldsymbol{u}}^{-1} \boldsymbol{\Sigma}_{\boldsymbol{u}\text{Y}}.$$

75) (**2nd way to compute $\hat{\boldsymbol{\beta}}$**): Let the sample covariance matrices be $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}} = \dfrac{1}{n-1}\sum_{i=1}^n (\boldsymbol{u}_i - \overline{\boldsymbol{u}})(\boldsymbol{u}_i - \overline{\boldsymbol{u}})^T$ and $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}Y} = \dfrac{1}{n-1}\sum_{i=1}^n (\boldsymbol{u}_i - \overline{\boldsymbol{u}})(Y_i - \overline{Y})$. Let the method of moments or maximum likelihood estimators be $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{u}} = \dfrac{1}{n}\sum_{i=1}^n (\boldsymbol{u}_i - \overline{\boldsymbol{u}})(\boldsymbol{u}_i - \overline{\boldsymbol{u}})^T$ and $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y} = \dfrac{1}{n}\sum_{i=1}^n (\boldsymbol{u}_i - \overline{\boldsymbol{x}})(Y_i - \overline{Y}) = \dfrac{1}{n}\sum_{i=1}^n \boldsymbol{u}_i Y_i - \overline{\boldsymbol{u}}\,\overline{Y}$. Suppose that $\boldsymbol{w}_i = (Y_i, \boldsymbol{u}_i^T)^T$ are iid random vectors such that σ^2_Y, $\boldsymbol{\Sigma}_{\boldsymbol{u}}^{-1}$ and $\boldsymbol{\Sigma}_{\boldsymbol{u}Y}$ exist. Then $\hat{\beta}_0 = \overline{Y} - \hat{\boldsymbol{\beta}}_S^T \overline{\boldsymbol{u}} \xrightarrow{P} \beta_0$ and

$$\hat{\boldsymbol{\beta}}_S = \frac{n}{n-1} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}}^{-1} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}Y} = \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{u}}^{-1} \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{u}Y} \xrightarrow{P} \boldsymbol{\beta}_S \;\; \text{as} \;\; \text{n} \to \infty.$$

It is important to note that this result is for iid $\boldsymbol{w}_i$ with second moments. Do not need a linear model $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$ to hold.

76) Result 75) can be shown, after algebra, using $\boldsymbol{X}^T\boldsymbol{Y} = \begin{pmatrix} n\overline{Y} \\ \boldsymbol{X}_1^T \boldsymbol{Y} \end{pmatrix} = \begin{pmatrix} n\overline{Y} \\ \sum_{i=1}^n \boldsymbol{u}_i Y_i \end{pmatrix}$

$$(\boldsymbol{X}^T\boldsymbol{X})^{-1} = \begin{pmatrix} \frac{1}{n} + \overline{\boldsymbol{u}}^T \boldsymbol{D}^{-1} \overline{\boldsymbol{u}} & -\overline{\boldsymbol{u}}^T \boldsymbol{D}^{-1} \\ -\boldsymbol{D}^{-1}\overline{\boldsymbol{u}} & \boldsymbol{D}^{-1} \end{pmatrix}$$

where the $(p-1) \times (p-1)$ matrix $\boldsymbol{D}^{-1} = [(n-1)\hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}}]^{-1} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}}^{-1}/(n-1)$.

77) **Generalized Cochran's Theorem**: Let $\boldsymbol{Y} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Let $\boldsymbol{A}_i = \boldsymbol{A}_i^T$ have rank r_i for $i = 1, ..., k$, and let $\boldsymbol{A} = \sum_{i=1}^k A_i = \boldsymbol{A}^T$ have rank r. Then $\boldsymbol{Y}^T \boldsymbol{A}_i \boldsymbol{Y} \sim \chi^2(r_i, \frac{1}{2}\boldsymbol{\mu}^T \boldsymbol{A}_i \boldsymbol{\mu})$, and the $\boldsymbol{Y}^T \boldsymbol{A}_i \boldsymbol{Y}$ are independent, and $\boldsymbol{Y}^T \boldsymbol{A} \boldsymbol{Y} \sim \chi^2(r, \frac{1}{2}\boldsymbol{\mu}^T \boldsymbol{A} \boldsymbol{\mu})$, iff
I) any 2 of a) $\boldsymbol{A}_i \boldsymbol{\Sigma}$ are idempotent $\forall i$,
b) $\boldsymbol{A}_i \boldsymbol{\Sigma} \boldsymbol{A}_j = \boldsymbol{0} \ \ \forall i < j$,
c) $\boldsymbol{A}\boldsymbol{\Sigma}$ is idempotent
are true; or II) c) is true and d) $r = \sum_{i=1}^k r_i$;
or III) c) is true and e) $\boldsymbol{A}_1 \boldsymbol{\Sigma}, .., \boldsymbol{A}_{k-1} \boldsymbol{\Sigma}$ are idempotent and $\boldsymbol{A}_k \boldsymbol{\Sigma} \geq 0$ is singular.

78) **Distribution of F_R under normality when H_0 may not hold**: Assume $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon} \sim N_n(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$. Let $\boldsymbol{X} = [\boldsymbol{X}_1 \ \ \boldsymbol{X}_2]$ be full rank, and let the reduced model $\boldsymbol{Y} = \boldsymbol{X}_1 \boldsymbol{\beta}_1 + \boldsymbol{\epsilon}$. Then

$$F_R = \frac{\boldsymbol{Y}^T(\boldsymbol{P} - \boldsymbol{P}_1)\boldsymbol{Y}/q}{\boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{Y}/(n-p)} \sim F\left(q, n-p, \frac{\boldsymbol{\beta}^T \boldsymbol{X}^T (\boldsymbol{P} - \boldsymbol{P}_1)\boldsymbol{X}\boldsymbol{\beta}}{2\sigma^2}\right)$$

where $F(d_1, d_2, \gamma)$ is a noncentral F distribution with d_1 and d_2 numerator and denominator degrees of freedom and noncentrality parameter γ. If $H_0 : \boldsymbol{\beta}_2 = \boldsymbol{0}$ is true, then $\gamma = 0$.

79) $Y \sim F(d_1, d_2) \sim F(d_1, d_2, 0)$. Let $X_1 \sim \chi^2(d_1, \gamma) \perp\!\!\!\perp X_2 \sim \chi^2(d_2, 0)$. Then $W = \dfrac{X_1/d_1}{X_2/d_2} \sim F(d_1, d_2, \gamma)$.

80) Suppose $Y \perp\!\!\!\perp \boldsymbol{u} | \boldsymbol{u}^T \boldsymbol{\beta}_U$, e.g. $Y_i = \beta_0 + \boldsymbol{u}_i^T \boldsymbol{\beta}_U + \epsilon_i$, or $Y_i = m(\boldsymbol{u}_i^T \boldsymbol{\beta}_U) + \epsilon_i$, or a GLM (generalized linear model). If the $\boldsymbol{u}_i$ are iid from an elliptically contoured distribution, then often the OLS estimator $\hat{\boldsymbol{\beta}}_S \xrightarrow{P} c\boldsymbol{\beta}_U$ for some constant $c \neq 0$.

81) Let $\boldsymbol{\beta}^T = (\beta_0 \ \ \boldsymbol{\beta}_U^T)$ and suppose the full model is $Y \perp\!\!\!\perp \boldsymbol{u} | (\beta_0 + \boldsymbol{u}^T \boldsymbol{\beta}_U)$. Consider testing $\boldsymbol{C}\boldsymbol{\beta}_U = \boldsymbol{0}$. Let the full model be $Y \perp\!\!\!\perp \boldsymbol{u} | (\beta_0 + \boldsymbol{u}_R^T \boldsymbol{\beta}_R + \boldsymbol{u}_O^T \boldsymbol{\beta}_O)$, and let the reduced model be $Y \perp\!\!\!\perp \boldsymbol{u} | (\beta_0 + \boldsymbol{u}_R^T \boldsymbol{\beta}_R)$ where $\boldsymbol{u}^T = (\boldsymbol{u}_R^T \ \ \boldsymbol{u}_O^T)$ and $\boldsymbol{u}_O$ denotes the terms outside of the reduced model. Notice that OLS ANOVA F test corresponds to Ho: $\boldsymbol{\beta}_U = \boldsymbol{0}$ and uses $\boldsymbol{L} = \boldsymbol{I}_{p-1}$. The tests for Ho: $\beta_i = 0$ use $\boldsymbol{L} = (0, ..., 0, 1, 0, ..., 0)$ where the 1 is in the ith position and are equivalent to the OLS t tests. The test Ho: $\boldsymbol{\beta}_O = \boldsymbol{0}$ uses $\boldsymbol{L} = [\boldsymbol{0} \ \boldsymbol{I}_j]$ if $\boldsymbol{\beta}_O$ is a $j \times 1$ vector.

82) Assume $Y \perp\!\!\!\perp \boldsymbol{u} | (\beta_0 + \boldsymbol{\beta}_U^T \boldsymbol{u})$, which is equivalent to $Y \perp\!\!\!\perp \boldsymbol{u} | \boldsymbol{\beta}_U^T \boldsymbol{u}$. Let the population OLS residual

$$v = Y - \beta_0 - \boldsymbol{\beta}_S^T \boldsymbol{u}$$

with

$$\tau^2 = E[(Y - \beta_0 - \boldsymbol{\beta}_S^T \boldsymbol{u})^2] = E(v^2),$$

and let the OLS residual be

$$r = Y - \hat{\beta}_0 - \hat{\boldsymbol{\beta}}_S^T \boldsymbol{u}. \tag{1}$$

Then under regularity conditions, results i) – iv) below hold.

i) Li and Duan (1989): The OLS slopes estimator $\boldsymbol{\beta}_S = c\boldsymbol{\beta}_U$ for some constant c.

ii) Li and Duan (1989) and Chen and Li (1998):

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_S - c\boldsymbol{\beta}_U) \xrightarrow{D} N_{p-1}(\boldsymbol{0}, \boldsymbol{C}_{OLS})$$

where

$$\boldsymbol{C}_{OLS} = \boldsymbol{\Sigma}_{\boldsymbol{u}}^{-1} E[(Y - \beta_0 - \boldsymbol{\beta}_S^T \boldsymbol{u})^2 (\boldsymbol{u} - E(\boldsymbol{u}))(\boldsymbol{u} - E(\boldsymbol{u}))^T] \boldsymbol{\Sigma}_{\boldsymbol{u}}^{-1}.$$

iii) Chen and Li (1998): Let $\boldsymbol{L}$ be a known full rank constant $k \times (p-1)$ matrix. If the null hypothesis Ho: $\boldsymbol{L}\boldsymbol{\beta}_U = \boldsymbol{0}$ is true, then

$$\sqrt{n}(\boldsymbol{L}\hat{\boldsymbol{\beta}}_S - c\boldsymbol{L}\boldsymbol{\beta}_U) = \sqrt{n}\boldsymbol{L}\hat{\boldsymbol{\beta}}_S \xrightarrow{D} N_k(\boldsymbol{0}, \boldsymbol{L}\boldsymbol{C}_{OLS}\boldsymbol{L}^T)$$

and

$$\boldsymbol{L}\boldsymbol{C}_{OLS}\boldsymbol{L}^T = \tau^2 \boldsymbol{L}\boldsymbol{\Sigma}_{\boldsymbol{u}}^{-1}\boldsymbol{L}^T.$$

To create test statistics, the estimator

$$\hat{\tau}^2 = \mathrm{MSE} = \frac{1}{\mathrm{n-p}} \sum_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{r}_{\mathrm{i}}^2 = \frac{1}{\mathrm{n-p}} \sum_{\mathrm{i=1}}^{\mathrm{n}} (\mathrm{Y_i} - \hat{\beta}_0 - \hat{\boldsymbol{\beta}}_{\mathrm{S}}^{\mathrm{T}} \boldsymbol{u}_{\mathrm{i}})^2$$

will be useful. The estimator $\hat{\boldsymbol{C}}_{OLS} =$

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}}^{-1} \left[\frac{1}{n} \sum_{i=1}^{n} [(Y_i - \hat{\beta}_0 - \hat{\boldsymbol{\beta}}_S^T \boldsymbol{u}_i)^2 (\boldsymbol{u}_i - \overline{\boldsymbol{u}})(\boldsymbol{u}_i - \overline{\boldsymbol{u}})^T] \right] \hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}}^{-1}$$

can also be useful. Notice that for general 1D regression models, the OLS MSE estimates τ^2 rather than the error variance σ^2.

iv) Result iii) suggests that a test statistic for $Ho: \boldsymbol{L}\boldsymbol{\beta}_U = \boldsymbol{0}$ is

$$W_{OLS} = n\hat{\boldsymbol{\beta}}_S^T \boldsymbol{L}^T [\boldsymbol{L}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}}^{-1}\boldsymbol{L}^T]^{-1} \boldsymbol{L}\hat{\boldsymbol{\beta}}_S / \hat{\tau}^2 \xrightarrow{D} \chi_k^2.$$

83) Under the conditions of 82), if $Ho: \boldsymbol{L}\boldsymbol{\beta}_U = \boldsymbol{0}$ is true, then the test statistic

$$F_R = \frac{n-1}{kn} W_{OLS} \xrightarrow{D} \chi_k^2 / k$$

as $n \to \infty$. This result means that the OLS partial F tests are large sample tests for a large class of nonlinear models where $Y \perp\!\!\!\perp \boldsymbol{u} | \boldsymbol{u}^T \boldsymbol{\beta}_U$.

84) The *AR(p)* time series model is $Y_t = \phi_0 + \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + \epsilon_t$. In matrix form, this model is $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ or

$$\begin{bmatrix} Y_{p+1} \\ Y_{p+2} \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & Y_p & Y_{p-1} & \dots & Y_1 \\ 1 & Y_{p+1} & Y_p & \dots & Y_2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & Y_{n-1} & Y_{n-2} & \dots & Y_{n-p} \end{bmatrix} \begin{bmatrix} \phi_0 \\ \phi_1 \\ \vdots \\ \phi_p \end{bmatrix} + \begin{bmatrix} \epsilon_{p+1} \\ \epsilon_{p+2} \\ \vdots \\ \epsilon_n \end{bmatrix}.$$

If the *AR(p)* model is stationary, then under regularity conditions, OLS partial F tests are large sample tests for this model.