Seth Temple
sdtemple@uw.edu

Notes

1 Lectures

This section records key facts presented in lectures in roughly chronological order.

Singular value decomposition. Let $\mathbf{X} \in \mathbb{R}^{n \times p}$. We can write $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T$, where

- $\mathbf{U}$ is an orthogonal $n \times n$ matrix
- $\mathbf{V}$ is an orthogonal $p \times p$ matrix
- $D_{ij} = 0$ for all $i \neq j$ and in non-decreasing order $D_{ii} \geq 0$ for all $i \leq \min(n, p)$.

Some facts about SVDs are

- A singular value decomposition is unique up to the signs of columns of $\mathbf{U}$ and $\mathbf{V}$
- All matrices have SVDs whereas only symmetric matrices have spectral decompositions
- We can construct compact SVDs.

Subspace. A *subspace* is contained in a larger vector space and is a vector space itself. *Vector spaces* are closed under addition and scalar multiplication. An *orthogonal complement* of a subspace of a vector space is the set of all vectors in the vector space orthogonal to every vector in the subspace. We can decompose $\mathbf{Y} = \mathbf{Y}_{\mathcal{V}} + \mathbf{Y}_{\mathcal{V}^\perp}$. $\hat{\mathbf{Y}} \in \mathbf{Y}_{\mathcal{V}}$ and $\hat{\mathbf{e}} \in \mathbf{Y}_{\mathcal{V}^\perp}$.

Generalized inverse. Let $\mathbf{F} \in \mathbb{R}^{n \times p}$. Then generalized inverse $\mathbf{F}^-$ satisfies $\mathbf{F}\mathbf{F}^-\mathbf{F} = \mathbf{F}$.

- Every matrix has a generalized inverse.
- A matrix can have more than 1 generalized inverse.
- The inverse of an invertible matrix is unique and is a generalized inverse.

Pseudoinverse. For any matrix $\mathbf{F}$, $\exists$ a unique Moore-Penrose inverse $\mathbf{F}^+$ satisfying

- $\mathbf{F}^+$ is a generalized inverse of $\mathbf{F}$
- $\mathbf{F}$ is a generalized inverse of $\mathbf{F}^+$
- $\mathbf{F}\mathbf{F}^+$ and $\mathbf{F}^+\mathbf{F}$ are symmetric

This pseudoinverse is often implemented in computer programs.

Estimability. Consider model $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$ where $\mathbb{E}[\varepsilon|\mathbf{X}] = \mathbf{0}$. $a^T\beta$ is estimable if a is in the row space of $\mathbf{X}$.

- For $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-}\mathbf{X}^T\mathbf{Y}$, $a^T\hat{\beta}$ is unbiased estimator of $a^T\beta$. If $\text{Var}(\varepsilon|\mathbf{X}) = \sigma^2\mathbf{I}_n$, then $\text{Var}(a^T\hat{\beta}|\mathbf{X}) = \sigma^2 a^T(\mathbf{X}^T\mathbf{X})^{-}a$ (exercise 8).
- $a^T\hat{\beta}$ is BLUE if $a^T\beta$ is estimable (Gauss-Markov theorem).
- There are connections to identifiability, defined as $\theta \neq \theta_0 \implies f_\theta \neq f_{\theta_0}$.

Rank deficiency.

- Reduce to full rank.
  - Best. Easiest. Most common.
  - If $\mathbf{X} = \begin{bmatrix}\mathbf{Z}_1 & \mathbf{Z}_2\end{bmatrix}$, columns of $\mathbf{Z}_1$ are linearly independent, and columns of $\mathbf{Z}_2$ are linear combinations of columns of $\mathbf{Z}_1$, then $\hat{\beta} = \begin{bmatrix}(\mathbf{Z}_1^T\mathbf{Z}_1)^{-1}\mathbf{Z}_1^T\mathbf{Y} \\ \mathbf{0}\end{bmatrix}$.
- Use a generalized inverse ($\hat{\beta}$ still satisfies normal equations).
- Impose identifiability constraints.
  - $\mathbf{H}\beta = \mathbf{0}_s$ is an identifiability constraint if
    1. The rows of $\mathbf{H}$ are linearly independent of $\mathbf{X}$
    2. $\text{rank}\left(\begin{bmatrix}\mathbf{X} \\ \mathbf{H}\end{bmatrix}\right) = p$.
  - $\text{rank}(\mathbf{H}) = p - \text{rank}(\mathbf{X})$.
  - $\hat{\beta} = (\mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T\mathbf{Z}$, where $\mathbf{W} = \begin{bmatrix}\mathbf{X} \\ \mathbf{H}\end{bmatrix}$, $\mathbf{Z} = \begin{bmatrix}\mathbf{Y} \\ \mathbf{0}\end{bmatrix}$, and $\mathbf{H}$ corresponds to an identifiability constraint, is a unique solution to constrained least squares.

Consistency. The Gauss-Markov theorem is a result that holds for finite samples. We now discuss under which conditions we have asymptotically (weakly) consistent $\hat{\beta}$.

- An estimator $\hat{\theta}$ is consistent for θ if

$$\lim(P(|\hat{\theta} - \theta| < \varepsilon)) = 1,$$

  or, equivalently,

$$\lim(P(|\hat{\theta} - \theta| \geq \varepsilon)) = 0.$$

  Note that $|\hat{\theta} - \theta|$ is a random quantity and $P(\cdot)$ is a deterministic quantity.

- We often argue consistency using Chebyshev's inequality:

$$P\left(\frac{|X-\mu|}{\sigma} \geq \varepsilon\right) \leq \frac{\sigma^2}{\varepsilon^2},$$

  where X is a random variable with $\mathbb{E}[X] = \mu$ and $\sigma^2 < \infty$, and this inequality holds for any $\varepsilon > 0$.

- $\lim a_n = a$ if for all $\varepsilon > 0$ there exists m such that, for all $n > m$,

$$|a_n - a| < \varepsilon.$$

- Suppose we have a linear model with a full rank design matrix. If $\lambda_{\min}(\mathbf{X}'\mathbf{X}) \to \infty$, then $\hat{\beta} \xrightarrow{p} \beta$.

Correlated errors.

- Time series, spatially correlated, and longitudinal datasets have correlated observations.
- Random effects describe a class of models where the parameters themselves have a distribution. Examples include land plots and technical replicates.
- Fixed effects describe a class of models where the parameters are fixed, but unknown. Examples include experiments with levels, e.g. apply different fertilizer treatments.
- Mixed models refer to models with both fixed and random effects.
- We apply transforms to work with an uncorrelated covariance matrix.
- For $\mathbf{C} \in \mathbb{R}^{n\times n}$, if $\mathbf{C}$ is positive (semi-)definite, then $\exists$ a positive (semi-)definite symmetric square root denoted $\mathbf{C}^{1/2}$. (We may have to be careful describing the diagonalization for rank-deficient $\mathbf{C}$.)
- $\hat{\beta}_G = (\mathbf{X}^T\Sigma^{-1}\mathbf{X})^{-1}\mathbf{X}^T\Sigma^{-1}\mathbf{Y}$ when $\mathbf{X}^T\Sigma^{-1}\mathbf{X}$ is full rank is the least squares solution to

$$\arg\min_{\beta}(\mathbf{Y}-\mathbf{X}\beta)^T\Sigma^{-1}(\mathbf{Y}-\mathbf{X}\beta)$$

Central limit theorems.

- Weighted averages are often normally distributed.
- Levy CLT. Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be a iid random vectors in $\mathbb{R}^p$.

$$\sqrt{n}(\bar{\mathbf{X}}_n - \mu) \xrightarrow{d} N_p(\mathbf{0}, \Sigma)$$

- Lindeberg-Feller CLT. Let $X_1, \dots X_n$ be independent random variables with zero mean and possibly different variances. Label $S_n = \sum_{i=1}^n X_i$ and $\sigma^2_{(n)} = \sum_{i=1}^n \sigma_i^2$. Then $S_n/\sigma_{(n)} \xrightarrow{d} N(0,1)$ and $\max\{\sigma_i^2/\sigma_{(n)^2}\} \to 0$ iff the Lindeberg condition holds.

- Lindeberg condition. For all $\varepsilon > 0$

$$\frac{1}{\sigma^2_{(n)}} \sum_{i=1}^n \mathbb{E}[X_i^2 \mathbf{1}_{|X_i| \geq \varepsilon \sigma_{(n)}}] \to 0$$

  We usually use $\Leftarrow$ of the LF-CLT, showing that the Lindeberg condition holds and concluding $S_n/\sigma_{(n)} \xrightarrow{d} N(0,1)$.

- Dominated convergence theorem. If $f_n \to f$ pointwise and $|f_n(x)| \leq g(x)$ for all n and $\int g < \infty$, then $\int f_n \to \int f$. This statement of the theorem is a corollary to DCT in Shorack (2017).

- Cramér-Wold device. $\mathbf{X}_n \in \mathbb{R}^d$ satisfies $\mathbf{X}_n \xrightarrow{d} \mathbf{X}_0$ iff $a^T \mathbf{X}_n \xrightarrow{d} a^T \mathbf{X}_0$ for all $a \in \mathbb{R}^d$. We get a nice corollary for $\mathbf{X}_0 \sim N_d(\mathbf{0}, \mathbf{I}_d)$ for all $a \in \mathbb{R}^d$ such that $a^T a = 1$.

- Asymptotic normality of $\hat{\beta}$. Suppose we have our LM setup and full rank $\mathbf{X}$ for all n. $\max\{X_k^T (\mathbf{X}^T \mathbf{X})^{-1} X_k\} \to 0$ implies

$$(\mathbf{X}^T \mathbf{X})^{1/2} (\hat{\beta} - \beta) \xrightarrow{d} N_p(0, \sigma^2 \mathbf{I}_p)$$

  (Observe above that we consider the maximum *leverage*.)

- Mann-Wald. If g is a continuous function, then $Z_n \xrightarrow{p} Z$ implies $g(Z_n) \xrightarrow{p} g(Z)$ and $Z_n \xrightarrow{d} Z$ implies $g(Z_n) \xrightarrow{d} g(Z)$

Hypothesis testing.

- For multivariate rejection regions, statisticians may disagree on which rejection region to use (min volume ellipsoid, min diameter sphere, or box constraints). This motivates finding a 1-dimensional test statistic.

- Consider $\mathbf{Z} \sim N_n(\mu, \Sigma)$ with $\text{rank}(\Sigma) = n$. Then

$$Q = (\mathbf{Z} - \mu)^T \Sigma^{-1} (\mathbf{Z} - \mu) \sim \chi^2_n$$

- Suppose we have a linear model with full rank design matrix and some regularity conditions are satisfied. Then, under $H_0 : \mathbf{A}\beta = c$,

$$\frac{(\mathbf{A}\hat{\beta} - c)^T (\mathbf{A}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{A}^T)^{-1} (\mathbf{A}\hat{\beta} - c)}{\sigma^2} \xrightarrow{d} \chi^2_k$$

- If we have normal errors and σ^2 is known, then a χ^2 test is exact (correct for finite n)
- If we have normal errors and σ^2 is unknown, then a F-test is exact
  - Suppose we have a linear model with normal errors and full rank design matrix. Then, under $H_0 : \mathbf{A}\beta = c$,

$$F = \frac{(\mathbf{A}\hat{\beta} - c)^T(\mathbf{A}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{A}^T)^{-1}(\mathbf{A}\hat{\beta} - c) \div k}{s^2} \xrightarrow{d} F_{k,n-p}$$

  - Under $H_0 : \beta_i = 0$,

$$\frac{\hat{\beta}_i^2}{s^2(\mathbf{X}^T\mathbf{X})_{ii}^{-1}} \xrightarrow{d} F_{1,n-p}$$

    and, equivalently, due to the relationship between t- and F-distributions,

$$\frac{\hat{\beta}_i}{s\sqrt{(\mathbf{X}^T\mathbf{X})_{ii}^{-1}}} = \frac{\hat{\beta}_i}{s.e.(\hat{\beta}_i)} \xrightarrow{d} t_{n-p}$$

  - Another framing: let $RSS_{H_0} = (\mathbf{Y} - \mathbf{X}\hat{\beta}_{H_0})^T(\mathbf{Y} - \mathbf{X}\hat{\beta}_{H_0})$ under the null hypothesis restrictions and $RSS = (\mathbf{Y} - \mathbf{X}\hat{\beta})^T(\mathbf{Y} - \mathbf{X}\hat{\beta})$ under no restrictions. We can write

$$(\mathbf{A}\hat{\beta} - c)^T(\mathbf{A}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{A}^T)^{-1}(\mathbf{A}\hat{\beta} - c) = RSS_{H_0} - RSS$$

    Therefore, we derive the same asymptotic distribution

$$\frac{(RSS_{H_0} - RSS) \div k}{RSS \div (n-p)} \xrightarrow{d} F_{k,n-p}$$

- If the errors are not normal, the F-test is asymptotically the same as the χ^2 test (up to constant multiplier). We achieve this result via a fact that $k \times F \xrightarrow{d} \chi_k^2$
- Suppose we have a linear model with normal errors and full rank design matrix. Then

$$\frac{(n-p)s^2}{\sigma^2} \sim \chi_{n-p}^2$$

  where $s^2 = (\mathbf{Y} - \mathbf{X}\hat{\beta})^T(\mathbf{Y} - \mathbf{X}\hat{\beta})/(n-p))$.
- s^2 and $\hat{\beta}$ are independent, using
  - $\mathbf{Z} \sim N_n$ if and only if $a^T\mathbf{Z} \sim N_1$ for all non-zero vectors a
- If $U \sim \chi_m^2$ and $V \sim \chi_n^2$, then

$$\frac{U/m}{V/n} \sim F_{m,n}$$

Heteroscedasticity. (Homo)heteroscedasticity means that the variance of $\mathbf{Y}$ does (not) depend on $\mathbf{X}$. Heteroscedasticity is the more reasonable assumption to make, but this complicates the math. The most common approach to assume heteroscedasticty is via a weight matrix $\mathbf{W}$. Let $Y = \mathbf{X}\beta + \mathbf{W}\varepsilon$.

- $\text{Var}(\hat{\beta}) = ((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}^2\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1})^{-1}$.
  - The true variance is often larger than model-based variances under the homoscedastic assumption
- Using Cramér-Wold device, LF-CLT, and DCT, and assuming the max leverage with respect to $\mathbf{WX}$ goes to zero, we achieve

$$(\text{Var}(\hat{\beta}))^{-1/2}(\hat{\beta} - \beta) \xrightarrow{d} N(\mathbf{0}, \mathbf{I}_p)$$

- Ignoring heteroscedasticity can result in too small of confidence intervals, misleading inference, etc.
  - Variance-stabilizing transformations are often used to handle this
- Huber-White sandwich estimation is often used to determine unknown $\mathbf{W}$

Experimental Design. Orthogonal designs are nice because estimates for $\hat{\beta}_i$ do not change when we include a new orthogonal covariate and the variance of $\hat{\beta}_i$ is minimized (optimal). An orthogonal design is one where the covariates in the design matrix $\mathbf{X}$ are orthogonal. Under such an assumption,

- $\hat{\beta}_i = \frac{\mathbf{x}_i^T\mathbf{Y}}{\mathbf{x}_i^T\mathbf{x}_i}$
- $\text{Var}(\hat{\beta}_i) = \frac{\sigma^2}{\mathbf{x}_i^T\mathbf{x}_i}$ which is the variance bound!
- Amy suggested an orthogonal design to a collaborator for experiment on gene expression of regenerative worms
- We may desire to add another observation to the experiment that is $\mathbf{X}_{n+1} = c \cdot \mathbf{v}_{\min}$ where $\mathbf{v}_{\min}$ is the eigenvector corresponding to the smallest eigenvalue, if we have the resources (and can play god)

Blocking

- Including relevant covariates in the model
- (often) under the control of the experimenter

2 Exercises

This section records the facts presented the in-class exercises in chronological order.

1. Any solution $\hat{\beta}$ to $\arg\min_{\beta} (\mathbf{Y} - \mathbf{X}\beta)^T(\mathbf{Y} - \mathbf{X}\beta)$ satisfies that $\mathbf{X}^T\mathbf{X}\hat{\beta} = \mathbf{X}^T\mathbf{Y}$.
2. Let $\mathbf{A} \in \mathbb{R}^{s\times s}$, $\text{rank}(\mathbf{A}) = s$, and $\mathbf{B} \in \mathbb{R}^{s\times t}$. Then, $\text{rank}(\mathbf{AB}) = \text{rank}(\mathbf{B})$.
3. (a) The columns of $\mathbf{U}$ in the SVD of $\mathbf{X}$ are the eigenvectors of $\mathbf{X}\mathbf{X}^T$.
   (b) The columns of $\mathbf{V}$ in the SVD of $\mathbf{X}$ are the eigenvectors of $\mathbf{X}^T\mathbf{X}$.
   (c) The diagonal elements of $\mathbf{D}$ in the SVD of $\mathbf{X}$ are the square roots of the eigenvalues of $\mathbf{X}^T\mathbf{X}$ and $\mathbf{X}\mathbf{X}^T$.
4. (a) $\text{rank}(\mathbf{X}'\mathbf{X}) = \text{rank}(\mathbf{X})$. (Full rank $\mathbf{X}$ is a sufficient condition for LSE to be unique.)
   (b) If $\text{rank}(\mathbf{X}) = p \leq n$, then $\mathbf{X}'\mathbf{X}$ is positive definite. (Full rank $\mathbf{X}$ is sufficient condition for SSE to be strictly convex.)
5. Let $\mathbf{P_X}$ be the projection matrix onto $\mathbf{X}$ where $\mathbf{X} \in \mathbb{R}^{n\times p}$.
   (a) $\mathbf{P_X}$ can be written $\mathbf{U}\mathbf{A}\mathbf{U}'$ using SVD.
   (b) $\mathbf{P_X}$ has eigenvalue 1 of multiplicity p and eigenvalue 0 of multiplicity $n - p$.
   (c) $\text{rank}(\mathbf{P_X}) = p$.
6. Every matrix has a generalized inverse.
7. If $\mathbf{G}$ and $\mathbf{H}$ are generalized inverses of $\mathbf{X}'\mathbf{X}$, then $\mathbf{XGX}' = \mathbf{XHX}'$.
8. For $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$ and $\varepsilon|\mathbf{X} \sim (\mathbf{0}, \sigma^2\mathbf{I}_n)$, if $a^T\beta$ is estimable, then $\text{var}(a^T\hat{\beta}|\mathbf{X}) = \sigma^2 a^T(\mathbf{X}^T\mathbf{X})^- a$ where $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T\mathbf{Y}$.
9. Gauss-Markov theorem for full rank $\mathbf{X}$. $a^T\hat{\beta}$ is unique UMVUE for $a^T\beta$.
10. Using Chebyshev's inequality, we show that
$$P(|Y_n - \mu| \geq \delta) \leq \frac{\sigma_n^2}{\delta^2}$$
    where $Y_1, \ldots, Y_n$ is a sequence of random variables with indexed variances and common expectation. If $\lim \sigma_n^2 = 0$, then $Y_n \xrightarrow{p} \mu$. We use this exercise to say that, if our estimator's variance goes to zero as the sample gets asymptotically large, then the estimator is asymptotically (weakly) consistent for μ.
11. Suppose $\mathbf{Y} \sim (\mathbf{X}\beta, \Sigma)$ where Σ is full rank. Then $\Sigma^{-1/2}(\mathbf{Y} - \mathbf{X}\beta) \sim (\mathbf{0}_n, \mathbf{I}_n)$

12. If we have full rank $\mathbf{X}$ and Σ, the OLS and GLS estimates are both unbiased estimators of β. They often have different variances. In this case, the Gauss Markov theorem gives that $a^T\hat{\beta}_G$ is BLUE for $a^T\beta$.

13. Reflect on when least squares are normally distributed

14. We have our usual OLS setup with full rank $\mathbf{X}$ and the max leverage converging to 0. Under $H_0 : \mathbf{A}\beta = c$ and the rank of $\mathbf{A}$ is k,

$$\frac{(\mathbf{A}\hat{\beta} - c)^T(\mathbf{A}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{A}^T)^{-1}(\mathbf{A}\hat{\beta} - c)}{\sigma^2} \xrightarrow{d} \chi^2_k$$

15. The setup is the same as above, except we have normal errors and a finite sample. Instead,

$$\frac{(\mathbf{A}\hat{\beta} - c)^T(\mathbf{A}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{A}^T)^{-1}(\mathbf{A}\hat{\beta} - c)}{\sigma^2} \sim \chi^2_k$$

    That is, χ^2 is an exact test.

16. We prefer orthogonal designs

17. Some derivations on the way to the asymptotic distribution for ordinary least squares in the heteroscedastic case

3 Homeworks

This section records the facts presented in homeworks in roughly chronological order.

1. For any matrix $\mathbf{A}$, $\mathbf{A}\mathbf{A}' = \mathbf{0}$ implies $\mathbf{A} = 0$.

2. Projection matrices.

   (a) For any matrix $\mathbf{A}$, $\mathbf{P_A} = \mathbf{A}(\mathbf{A}'\mathbf{A})^{-}\mathbf{A}'$ is a projection matrix onto $\mathcal{C}(\mathbf{A})$.

   (b) $\mathbf{P_A}\mathbf{A} = \mathbf{A}$.

   (c) $\text{rank}(\mathbf{P_A}) = \text{rank}(\mathbf{A})$.

3. Given two OLS estimates of β, $\mathbf{X}\hat{\beta}_1 = \mathbf{X}\hat{\beta}_2$.

4. Consider models $\mathbb{E}[\mathbf{Y}|\mathbf{X}] = \mathbf{1}\alpha_0 + \mathbf{W}\alpha$ and $\mathbb{E}[\mathbf{Y}|\mathbf{X}] = \mathbf{1}\beta_0 + \mathbf{X}\beta$. Suppose $\mathbf{W}$, a column centered version of design matrix $\mathbf{X}$, has full rank $p < n$. Then least squares estimates of α and β are unique and $\hat{\alpha} = \hat{\beta}$.

5. Let $\mathbf{P}$ be a $n \times n$ projection matrix and $\mathbf{R}$ be a $n \times n$ orthogonal matrix.

   - $\mathbf{P}$ is positive semidefinite.
   - If $\text{rank}(\mathbf{P}) = r$, then $\mathbf{P}$ has eigenvalue 1 with multiplicity r and eigenvalue 0 with multiplicity $n - r$.
   - $\mathbf{R}$ has real eigenvalues ± 1.

6. The (unique) least squares estimate is unbiased when the design matrix is full rank.

7. In simple linear regression, $\hat{\beta}_0$ and $\hat{\beta}_1$ are uncorrelated if and only if $\bar{x} = 0$.

8. (Seber and Lee page 64.) Rank-deficient $\mathbf{X}$ implies that a least squares estimator cannot be unbiased for β. Moreover, a least squares estimate is of the form $\mathbf{C}\mathbf{Y}_n$ where $\mathbf{C} \in \mathbb{R}^{p \times n}$ and $\mathbf{X}^T\mathbf{X}\mathbf{C} = \mathbf{X}^T$.

9. The sum of the leverages equals the rank of the design matrix. Moreover, leverages lie in between 0 and 1 inclusive.

10. There are more ways to show that the Lindeberg condition holds besides just using the dominated convergence theorem. Sometimes inequalities like Hölder's and Markov's can be useful.

11. For $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$ and $\varepsilon \sim (0, \sigma^2)$,

$$s^2 = \frac{(\mathbf{Y} - \mathbf{X}\hat{\beta})^T(\mathbf{Y} - \mathbf{X}\hat{\beta})}{n - p} \xrightarrow{p} \sigma^2$$

4 Potpourri

- Suppose $\mathbf{AX'X} - \mathbf{BX'X} = \mathbf{0}$. Then $\mathbf{AX'} = \mathbf{BX'}$.
- $\text{trace}(\mathbf{P}) = \text{rank}(\mathbf{P})$ for any projection matrix $\mathbf{P}$.
- Expected value of the residuals is $\mathbf{0}$.
- For our standard LM setup, $\frac{1}{n-\text{rank}(\mathbf{X})}(\mathbf{Y} - \mathbf{X}\hat{\beta})^T(\mathbf{Y} - \mathbf{X}\hat{\beta})$ is unbiased estimator of $\hat{\sigma}^2$.
- The only full rank projection matrix is the identity matrix.
- $\mathbb{E}[\mathbf{Z}^T\mathbf{AZ}] = \text{trace}(\mathbf{A}\text{Var}(\mathbf{Z})) + \mathbb{E}[\mathbf{Z}]^T\mathbf{A}\mathbb{E}[\mathbf{Z}]$.
- If $Y \sim N(\mathbf{X}\beta, \sigma^2 I_n)$, then
  - $\hat{\beta}$ is the MLE for β
  - $\hat{\beta}$ is unbiased for β
  - $\hat{\beta}$ is efficient, i.e. achieves CR lower bound
  - F-test is UMP level α test
- Hölder's inequality. For $p, q > 1$ and $\frac{1}{p} + \frac{1}{q} = 1$,

$$\mathbb{E}[|XY|] \leq \mathbb{E}[|X|^p]^{\frac{1}{p}} \times \mathbb{E}[|Y|^q]^{\frac{1}{q}}$$

- Cauchy-Schwarz inequality.

$$\begin{aligned}\mathbb{E}[XY]^2 &\leq \mathbb{E}[|XY|]^2 \\ &\leq \mathbb{E}[|X|^2] \times \mathbb{E}[|Y|^2] \\ &= \mathbb{E}[X^2] \times \mathbb{E}[Y^2]\end{aligned}$$

- Markov inequality. $\mu(|X| > \lambda) \leq \frac{\mathbb{E}[|X|^r]}{\lambda^r}$ for all $\lambda > 0$. This inequality provides upper bounds on probabilities.