8/21/23:

Most common linear model:

- Repeat experiment n times
- Have p predictor variables and 1 response/output variable
- We label the data, predictors $(x_{i1}, x_{i2}, \ldots, x_{ip}; y_i)$, $i = 1, \ldots, n$
- Model given by $y_i = \beta_0 + \sum_{j=1} \beta_j x_{ij} + \varepsilon_i$, linear in terms of x_{ij}, must determine $p+1$ coefficients $\beta_0, \beta_1, \ldots, \beta_n$, ε_i takes into account randomness
    - Statistical assumption $\varepsilon_i \sim N(0, \sigma^2)$ $\forall i$
    - Assumption that all data fits linearly w/ normal random perturbations
- So, goal is to predict β_j for $j = 0, 1, \ldots, p$ and σ^2
- How? Ans 1: Maximum Liklihood Estimator, MLE

$$y_i \sim N\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}, \sigma^2\right), \quad i = 1, \ldots, n$$

y_i's independent & identically distributed, i.i.d.

Convert to liklihood function

$$L = \prod_{i=1}^{n} \frac{e^{-z_i^2/(2\sigma^2)}}{(2\pi\sigma^2)^{1/2}}, \quad z_i = y_i - \left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}\right)$$

where $L = L(\beta_0, \ldots, \beta_p, \sigma^2)$. We compute the parameters by setting

$$\frac{\partial L}{\partial(\beta_i)} = \frac{\partial L}{\partial(\sigma^2)} = 0, \quad i = 0, \ldots, p$$

- Matrix Formulation of model:

$$\underline{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \in \mathbb{R}^n, \quad \underline{\underline{X}} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \in \mathbb{R}^{n \times p+1},$$

$$\underline{\beta} = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_p \end{bmatrix} \in \mathbb{R}^{p+1}, \quad \underline{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix} \in \mathbb{R}^n$$

And model is

$$\underline{y} = \underline{\underline{X}}\,\underline{\beta} + \underline{\varepsilon}$$

with assumption $\underline{\varepsilon} \sim N_n(\underline{0}, \sigma^2 \underline{\underline{I}})$

so $\underline{y} \sim N_n(\underline{\underline{X}}\underline{\beta}, \sigma^2 \underline{\underline{I}})$

Multivariate Normal Distribution:

> $\underline{x} \sim N_n(\underline{\mu}, \underline{\underline{\Sigma}})$, $\underline{x}, \underline{\mu} \in \mathbb{R}^n$, $\underline{\underline{\Sigma}} \in \mathbb{R}^{n \times n}$

with $\underline{\underline{\Sigma}}$ symmetric and positive (semi-) definite.

> Means that $\underline{x}$ is a random vector with pdf

> $$f(\underline{x}) = \frac{\exp\left(\frac{-1}{2}(\underline{x}-\underline{\mu})^T \underline{\underline{\Sigma}}^{-1}(\underline{x}-\underline{\mu})\right)}{(2\pi)^{n/2} \left|\det \underline{\underline{\Sigma}}\right|^{1/2}}$$

> Use change of variables to show $\int f(\underline{x})\, d\underline{x} = 1$

> Properties:
- If $\underline{v} \in \mathbb{R}^n$, then $\underline{x} + \underline{v} \sim N_n(\underline{\mu} + \underline{v}, \underline{\underline{\Sigma}})$
- If $\underline{\underline{A}} \in \mathbb{R}^{m \times n}$, then $\underline{\underline{A}}\underline{x} \sim N_m(\underline{\underline{A}}\underline{\mu}, \underline{\underline{A}}\underline{\underline{\Sigma}}\underline{\underline{A}}^T)$

Necessary Linear Algebra:

> Properties of definite matrices

> Eigenvalue decomposition

> Cholesky decomposition

8/23/23:

Let $\underline{\underline{A}} \in \mathbb{R}^{m \times n}$

> The row-rank of $\underline{\underline{A}}$ is the # of linearly independent rows

> The column-rank — " ————————— columns

> Define $C(\underline{\underline{A}}) = \text{span}\{\text{cols of } \underline{\underline{A}}\}$, $R(\underline{\underline{A}}) = \text{span}\{\text{rows of } \underline{\underline{A}}\}$

Theorem: row-rank $= \dim(R(\underline{\underline{A}})) = \dim(C(\underline{\underline{A}})) =$ column-rank

> Nullspace $\mathcal{N}(\underline{\underline{A}}) = \{\underline{x} \mid \underline{\underline{A}}\underline{x} = \underline{0}\}$

Theorem: Rank Nullity: $\text{rank}(\underline{\underline{A}}) + \dim \mathcal{N}(\underline{\underline{A}}) = n$

Let $\underline{\underline{A}} \in \mathbb{R}^{m \times n}$ with rank $1 \le r \le \min(n,m)$. Then $(\underline{\underline{P}}, \underline{\underline{Q}})$ is a <u>rank factorization</u> of $\underline{\underline{A}}$ if $\underline{\underline{P}} \in \mathbb{R}^{m \times r}$, $\underline{\underline{Q}} \in \mathbb{R}^{r \times n}$, and $\underline{\underline{A}} = \underline{\underline{P}}\,\underline{\underline{Q}}$.

<u>Theorem:</u> Every matrix has a rank factorization.

<u>Pf:</u> If $\mathrm{rank}(\underline{\underline{A}}) = m$, $\underline{\underline{A}} = \underline{\underline{I}}\,\underline{\underline{A}}$, and if $\mathrm{rank}(\underline{\underline{A}}) = n$, then $\underline{\underline{A}} = \underline{\underline{A}}\,\underline{\underline{I}}$. Otherwise, partial SVD is $\underline{\underline{A}} = \underline{\underline{U}}\,\underline{\underline{\Sigma}}\,\underline{\underline{V}}^T$, $\underline{\underline{U}}$ is $m \times r$, $\underline{\underline{\Sigma}}$ is $r \times r$, $\underline{\underline{V}}^T$ is $r \times n$, so a factorization given by $(\underline{\underline{U}}\,\underline{\underline{\Sigma}})\underline{\underline{V}}^T$ or $\underline{\underline{U}}(\underline{\underline{\Sigma}}\,\underline{\underline{V}}^T)$.

A matrix $\underline{\underline{A}} \in \mathbb{R}^{n \times n}$ is <u>idempotent</u> if $\underline{\underline{A}}^2 = \underline{\underline{A}}$

<u>ex:</u> Projection matrices

<u>Theorem:</u> If $\underline{\underline{A}}$ is idempotent, then $\mathrm{rank}(\underline{\underline{A}}) = \mathrm{trace}(\underline{\underline{A}})$

<u>Pf:</u> Let $r = \mathrm{rank}(\underline{\underline{A}})$. By rank factorization, $\underline{\underline{A}} = \underline{\underline{P}}\,\underline{\underline{Q}}$, $\underline{\underline{P}}$ is $\overset{(\mathrm{rank}\ r)}{n \times r}$, $\underline{\underline{Q}}$ is $\overset{(\mathrm{rank}\ r)}{r \times n}$. Then $\underline{\underline{P}}\,\underline{\underline{Q}}\,\underline{\underline{P}}\,\underline{\underline{Q}} = \underline{\underline{P}}\,\underline{\underline{I}}_{r \times r}\underline{\underline{Q}}$. From this, $\underline{\underline{P}}(\underline{\underline{Q}}\,\underline{\underline{P}} - \underline{\underline{I}}_{r \times r})\underline{\underline{Q}} = \underline{\underline{0}}$
$\Rightarrow \underline{\underline{P}}^T\underline{\underline{P}}(\underline{\underline{Q}}\,\underline{\underline{P}} - \underline{\underline{I}}_{r \times r})\underline{\underline{Q}}\,\underline{\underline{Q}}^T = \underline{\underline{0}}$ with $\underline{\underline{P}}^T\underline{\underline{P}}$ and $\underline{\underline{Q}}\,\underline{\underline{Q}}^T$ invertible since $r \times r$, hence, $\underline{\underline{Q}}\,\underline{\underline{P}} = \underline{\underline{I}}_{r \times r}$. Thus, $\mathrm{trace}(\underline{\underline{Q}}\,\underline{\underline{P}}) = \mathrm{trace}(\underline{\underline{I}}_{r \times r}) = r$. And by cyclic property of trace, $\mathrm{trace}(\underline{\underline{A}}) = \mathrm{trace}(\underline{\underline{P}}\,\underline{\underline{Q}}) = \mathrm{trace}(\underline{\underline{Q}}\,\underline{\underline{P}}) = r$.

<u>Theorem:</u> If $\underline{\underline{A}}^2 = \underline{\underline{A}} \in \mathbb{R}^{n \times n}$, then $\mathrm{rank}(\underline{\underline{A}}) + \mathrm{rank}(\underline{\underline{I}} - \underline{\underline{A}}) = n$.

<u>Proof:</u> By last thm, $\mathrm{rank}(\underline{\underline{A}}) = \mathrm{trace}(\underline{\underline{A}})$. And, $(\underline{\underline{I}} - \underline{\underline{A}})^2 = (\underline{\underline{I}} - \underline{\underline{A}})$ so $\mathrm{rank}(\underline{\underline{I}} - \underline{\underline{A}}) = \mathrm{trace}(\underline{\underline{I}} - \underline{\underline{A}}) = n - \mathrm{trace}(\underline{\underline{A}}) = n - \mathrm{rank}(\underline{\underline{A}})$.

Let V be a vector space and $S \subseteq V$ a subspace. Then the <u>complement</u> of S is $S^0 := \{\underline{y} \in V : \langle \underline{y}, \underline{s} \rangle = 0 \ \forall\, \underline{s} \in S\}$.

> Fact: S^0 is also a subspace.

<u>Theorem:</u> Let $\{\underline{x}_i\}_{i=1}^k$ be an orthogonal basis for subspace $S \subseteq V$. This can be extended into an orthogonal basis $\{\underline{x}_i\}_{i=1}^n$ for V s.t. $\{\underline{x}_i\}_{i=k+1}^n$ is an orthogonal basis for S^0.

<u>Proposition:</u> For any $\underline{x} \in V$, there exists unique $y_1 \in S$ & $y_2 \in S^0$ s.t.

$$\underline{x} = \underline{y_1} + \underline{y_2}.$$

> $\underline{y_1} = \mathrm{Proj}_S(\underline{x})$, $\underline{y_2} = \mathrm{Proj}_{S^0}(\underline{x})$

> $\mathrm{Proj}_S : V \to S$ is a linear mapping

Linearity means $\mathrm{Proj}_S(\underline{x}+\underline{y}) = \mathrm{Proj}_S(\underline{x}) + \mathrm{Proj}_S(\underline{y})$ and $\exists\ \underline{\underline{P}}_S \in \mathbb{R}^{n\times n}$ s.t. $\mathrm{Proj}_S(\underline{x}) = \underline{\underline{P}}_S \underline{x}$

<u>Lemma:</u> $\underline{\underline{P}}_S$ is idempotent

<u>Lemma:</u> $\underline{\underline{P}}_S^T \underline{\underline{P}}_S = \underline{\underline{P}}_S$

<u>Proof:</u> Recall $(\underline{\underline{I}} - \underline{\underline{P}}_S)$ projects onto orthogonal space S°, so we have

$$0 = \langle \underline{\underline{P}}_S \underline{x}, (\underline{\underline{I}} - \underline{\underline{P}}_S)\underline{x} \rangle = \underline{x}^T(\underline{\underline{P}}_S^T - \underline{\underline{P}}_S^T \underline{\underline{P}}_S)\underline{x} \quad \forall\ \underline{x}, \text{ hence}$$

$$\underline{\underline{P}}_S^T = \underline{\underline{P}}_S^T \underline{\underline{P}}_S \Rightarrow \underline{\underline{P}}_S = (\underline{\underline{P}}_S^T \underline{\underline{P}}_S)^T = \underline{\underline{P}}_S^T \underline{\underline{P}}_S.$$

$\underline{\underline{A}} \in \mathbb{R}^{n\times n}$ is an <u>orthogonal matrix</u> if $\underline{\underline{A}}^T \underline{\underline{A}} = \underline{\underline{A}}\,\underline{\underline{A}}^T = \underline{\underline{I}}$. I.e., the columns of $\underline{\underline{A}}$ are <u>orthonormal</u> (and hence the rows too)

<u>Theorem:</u> $\underline{\underline{A}}$ is orthogonal $\Leftrightarrow \langle \underline{\underline{A}}\underline{x}, \underline{\underline{A}}\underline{y} \rangle = \langle \underline{x}, \underline{y} \rangle\ \forall\ \underline{x}, \underline{y}$

$\Leftrightarrow \|\underline{\underline{A}}\underline{x} - \underline{\underline{A}}\underline{y}\| = \|\underline{x} - \underline{y}\|\ \forall\ \underline{x}, \underline{y}$

<u>Proof:</u> $\langle \underline{x}, \underline{y} \rangle = \langle \underline{\underline{A}}^T \underline{\underline{A}}\underline{x}, \underline{y} \rangle = \langle \underline{\underline{A}}\underline{x}, \underline{\underline{A}}\underline{y} \rangle \Leftrightarrow \underline{\underline{A}}^T\underline{\underline{A}} = \underline{\underline{I}}$ so $\underline{\underline{A}}$ orthogonal.

And $\|\underline{\underline{A}}\underline{x} - \underline{\underline{A}}\underline{y}\|^2 = \langle \underline{\underline{A}}(\underline{x}-\underline{y}), \underline{\underline{A}}(\underline{x}-\underline{y}) \rangle = \langle \underline{x}-\underline{y}, \underline{x}-\underline{y} \rangle = \|\underline{x}-\underline{y}\|^2 \Leftrightarrow$ see above

<u>8/28/23:</u>

Let $\underline{\underline{A}} \in \mathbb{R}^{n\times n}$. The <u>determinant</u> of $\underline{\underline{A}}$ can be defined recursively. A <u>cofactor</u> of $\underline{\underline{A}}$ given by $\underline{\underline{A}}_i^j$ where it is an $n-1$ by $n-1$ matrix with row i and column j removed. Then

$$\det(\underline{\underline{A}}) = \sum_{k=1}^{n} (-1)^{i+k} a_{ik} \det(\underline{\underline{A}}_i^k) \quad \text{for any } 1 \le i \le n$$

$$= \sum_{k=1}^{n} (-1)^{k+j} a_{kj} \det(\underline{\underline{A}}_k^j) \quad \text{for any } 1 \le j \le n.$$

We can also define it as the unique map $\det: \mathbb{R}^{n\times n} \to \mathbb{R}$ satisfying

1) Fixing $n-1$ cols, it is linear in the last column
2) Exchanging two cols flips the sign of the determinant.
3) $I \mapsto 1$

Theorem: Letting S_n be the permutation group on $\{1, 2, \ldots, n\}$, we have that

$$\det(\underline{A}) = \sum_{\pi \in S_n} \operatorname{sgn}(\pi) \prod_{i=1}^{n} a_{i,\pi(i)}$$

where
$$\operatorname{sgn}(\pi) = \begin{cases} 1 & \text{if } \pi \text{ an even permutation} \\ -1 & \text{if } \pi \text{ odd} \end{cases}$$

> Can prove by 3 properties above

Theorem: $\det(\underline{A}\underline{B}) = \det(\underline{A})\det(\underline{B})$.

> Can prove by above form

Corrolary: If $\underline{A}$ is invertible, then $\det(\underline{A}^{-1}) = 1/\det(\underline{A})$.

Let $\underline{A} \in \mathbb{R}^{n \times n}$. The algebraic eigenvalues of $\underline{A}$ are given by roots to the polynomial $\det(\lambda \underline{I} - \underline{A})$, $\lambda \in \mathbb{C}$. The geometric eigenvalues are defined to be any $\lambda \in \mathbb{C}$ for which there exists $\underline{x} \neq \underline{0}$ s.t. $\underline{A}\underline{x} = \lambda \underline{x}$. The geometric multiplicity of λ is k if $\exists$ exactly k linearly independent, nonzero $\underline{x}_i$ s.t. $\underline{A}\underline{x}_i = \lambda \underline{x}_i$ for each $\underline{x}_i$. (Algebraic is mult. of root)

Fact: Geometric multiplicity $\leq$ Algebraic multiplicity (usually "=" here).

Theorem: If $\underline{A}\underline{x} = \lambda \underline{x}$, $\underline{x} \neq \underline{0}$, then $\underline{A}^k \underline{x} = \lambda^k \underline{x}$.

Theorem: If f a polynomial, and λ an eigenvalue of $\underline{A}$, then $f(\lambda)$ an eigenvalue of $f(\underline{A})$.

Corrolary: If $\underline{A}^2 = \underline{A}$ (idempotent), then the eigenvalues of $\underline{A}$ are either 0 or 1.

Proof: $\underline{A}\underline{x} = \lambda \underline{x} \Rightarrow \lambda \underline{x} = \underline{A}\underline{x} = \underline{A}^2 \underline{x} = \lambda^2 \underline{x} \Rightarrow \lambda = \lambda^2 \Rightarrow \lambda \in \{0, 1\}$

Theorem: If $\lambda_1, \ldots, \lambda_k$ are distinct eigenvalues of $\underline{A}$ with corresponding eigenvectors $\underline{x}_1, \ldots, \underline{x}_k$, then the eigenvectors are linearly independent.

Proof: Suppose not, let j be the smallest integer s.t. $\underline{x}_j = \sum_{i=1}^{j-1} \beta_i \underline{x}_i$. Applying $\underline{A}$, $\lambda_j \underline{x}_j = \underline{A}\underline{x}_j = \sum_{i=1}^{j-1} \beta_i \underline{A}\underline{x}_i = \sum_{i=1}^{j-1} \lambda_i \beta_i \underline{x}_i$. Now, we cannot have $\lambda_j = 0$ or else first $j-1$ eigenvectors linearly dependent. Dividing through by λ_j, so $\underline{x}_j = \sum_{i=1}^{j-1} \frac{\lambda_i}{\lambda_j} \beta_i \underline{x}_i$. But from $\underline{x}_j = \sum_{i=1}^{j-1} \beta_i \underline{x}_i$, must have $\frac{\lambda_i}{\lambda_j} \beta_i = \beta_i$, $i = 1, \ldots, j-1$, by linear independence. Since $\lambda_i \neq \lambda_j$ for $i = 1, \ldots, j-1$, must have $\beta_i = 0$ for $i = 1, \ldots, j$. But then $\underline{x}_j = \underline{0}$, a contradiction.

Let $\underline{\underline{A}} \in \mathbb{R}^{n \times n}$. The <u>diagonalization</u> of $\underline{\underline{A}}$ is a pair $(\underline{\underline{T}}, \underline{\underline{\Sigma}})$ of $n \times n$ matrices s.t.

1) $\underline{\underline{T}}$ is invertible
2) $\underline{\underline{\Lambda}}$ is diagonal
3) $\underline{\underline{A}} = \underline{\underline{T}}\,\underline{\underline{\Lambda}}\,\underline{\underline{T}}^{-1}$.

<u>Theorem:</u> Let $\underline{\underline{A}} \in \mathbb{R}^{n \times n}$ be invertible with n distinct eigenvalues, then there exists a diagonalization $\underline{\underline{A}} = \underline{\underline{T}}\,\underline{\underline{\Lambda}}\,\underline{\underline{T}}^{-1}$. Moreover, $\exists\ \underline{x} \neq \underline{0}$ s.t. $\underline{\underline{A}}\,\underline{x} = \lambda \underline{x}$ for every λ on the diagonal of $\underline{\underline{\Lambda}}$.

<u>Proof:</u> Let $\lambda_1, \dots, \lambda_n$ be such that $\underline{\underline{A}}\,\underline{x_i} = \lambda \underline{x_i}$ for $i = 1, \dots, n$ with $\underline{x_1}, \dots, \underline{x_n}$ linearly independent. Let $\underline{\underline{T}} = [\, \underline{x_1} \cdots \underline{x_n} \,]$. By linear independence, $\underline{\underline{T}}$ invertible. Also, $\underline{\underline{A}}\,\underline{\underline{T}} = [\, \underline{\underline{A}}\underline{x_1} \cdots \underline{\underline{A}}\underline{x_n} \,] = [\, \lambda_1 \underline{x_1} \cdots \lambda_n \underline{x_n} \,]$. And this equals $\underline{\underline{T}}\,\underline{\underline{\Lambda}} = [\, \underline{x_1} \cdots \underline{x_n} \,]\begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix}$. Hence, we have $\underline{\underline{A}}\,\underline{\underline{T}} = \underline{\underline{T}}\,\underline{\underline{\Lambda}} \Rightarrow \underline{\underline{A}} = \underline{\underline{T}}\,\underline{\underline{\Lambda}}\,\underline{\underline{T}}^{-1}$.

<u>Spectral Theorem:</u> Let $\underline{\underline{A}}$ be a symmetric $n \times n$ matrix. Then $\exists$ an <u>orthogonal</u> matrix $\underline{\underline{T}}$ s.t. $\underline{\underline{A}} = \underline{\underline{T}}\,\underline{\underline{\Lambda}}\,\underline{\underline{T}}^T$ where $\underline{\underline{\Lambda}}$ is diagonal with <u>real entries</u>.

<u>Theorem:</u> If $\underline{\underline{A}} \in \mathbb{C}^{n \times n}$, hermitian, then all its eigenvalues are real, consequently, all eigenvalues can be chosen to have real entries. (or $A \in \mathbb{R}^{n \times n}$ symmetric)

<u>Proof:</u> If $\underline{\underline{A}}\,\underline{x} = \lambda \underline{x}$, then $\underline{x}^* \underline{\underline{A}}\,\underline{x} = \lambda\, \underline{x}^* \underline{x}$. And taking conjugate, $\underline{x}^* \underline{\underline{A}}^* \underline{x} = \bar{\lambda}\, \underline{x}^* \underline{x}$. With $\underline{\underline{A}}$ real & hermitian, $\underline{\underline{A}} = \underline{\underline{A}}^*$, so $\lambda = \bar{\lambda} \Rightarrow \lambda \in \mathbb{R}$. In addition, suppose $\underline{x} = (\underline{a} + i\underline{b})$ an eigenvector. Then $\underline{\underline{A}}(\underline{a} + i\underline{b}) = \underline{\underline{A}}\,\underline{x} = \lambda \underline{x} = \lambda(\underline{a} + i\underline{b})$. Hence, $\underline{a}, \underline{b} \in \mathbb{R}^n$ are real eigenvectors for λ.

<u>8/30/23:</u>

<u>Theorem:</u> Let $\underline{\underline{A}}$ be real, symmetric. Then eigenvectors corresponding to distinct eigenvalues are orthogonal.

<u>Proof:</u> Let $\underline{\underline{A}}\,\underline{x_1} = \lambda_1 \underline{x_1}$ and $\underline{\underline{A}}\,\underline{x_2} = \lambda_2 \underline{x_2}$ for distinct λ_1, λ_2. Then $\lambda_1 \underline{x_2}^T \underline{x_1} = \underline{x_2}^T (\underline{\underline{A}}\,\underline{x_1}) = (\underline{x_2}^T \underline{\underline{A}}\,\underline{x_1})^T = \underline{x_1}^T \underline{\underline{A}}\,\underline{x_2} = \lambda_2 \underline{x_1}^T \underline{x_2} = \lambda_2 \underline{x_2}^T \underline{x_1}$. By $\lambda_1 \neq \lambda_2$, must have $\underline{x_2}^T \underline{x_1} = 0 \Rightarrow \underline{x_1} \perp \underline{x_2}$.

<u>Corrolary:</u> Let $\underline{\underline{A}}$ be real symmetric & λ be an eigenvalue of multiplicity d. Then we can choose d <u>orthonormal</u> eigenvectors for λ that is also orthogonal to all other eigenvectors of $\underline{\underline{A}}$.

The above theorems prove the Spectral Theorem.

Theorem: Let $\underline{\underline{A}} \in \mathbb{R}^{n\times n}$, and suppose $\exists$ a diagonalization exists $\underline{\underline{A}} = \underline{\underline{T}}\,\underline{\underline{\Lambda}}\,\underline{\underline{T}}^{-1}$. Then $\operatorname{rank}(\underline{\underline{A}}) = \operatorname{rank}(\underline{\underline{\Lambda}})$, number of nonzero diagonals.

Corollary: Let $\underline{\underline{A}} \in \mathbb{R}^{n\times n}$ with diagonalization $\underline{\underline{A}} = \underline{\underline{T}}\,\underline{\underline{\Lambda}}\,\underline{\underline{T}}^{-1}$ and $\operatorname{rank}(\underline{\underline{A}}) = n$. Then, $\underline{\underline{A}}^{-1} = (\underline{\underline{T}}\,\underline{\underline{\Lambda}}\,\underline{\underline{T}}^{-1})^{-1} = (\underline{\underline{T}}^{-1})^{-1}\underline{\underline{\Lambda}}^{-1}\underline{\underline{T}}^{-1} = \underline{\underline{T}}\,\underline{\underline{\Lambda}}^{-1}\underline{\underline{T}}^{-1}$.

Let $\underline{\underline{A}} \in \mathbb{R}^{n\times n}$ be symmetric. $\underline{\underline{A}}$ is <u>non-negative definite</u> if $\underline{x}^T\underline{\underline{A}}\,\underline{x} \geq 0 \;\forall\, \underline{x} \in \mathbb{R}^n$. $\underline{\underline{A}}$ is strictly <u>positive definite</u> if $\underline{x}^T\underline{\underline{A}}\,\underline{x} > 0 \;\forall\, \underline{x} \in \mathbb{R}^n \setminus \{\underline{0}\}$.

Note: $$\underline{x}^T\underline{\underline{A}}\,\underline{x} = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}x_i x_j,$$ a quadratic polynomial in $\underline{x}$.

We call $\underline{x}^T\underline{\underline{A}}\,\underline{y}$ a <u>quadratic form</u>.

Theorem: If $\underline{\underline{A}} \in \mathbb{R}^{n\times n}$ is symmetric, $\underline{\underline{A}}$ is nonnegative definite iff its eigenvalues are nonnegative. Same for positive definite.

Proof: $\underline{x}^T\underline{\underline{A}}\,\underline{x} = \underline{x}^T\underline{\underline{T}}^T\underline{\underline{\Lambda}}\,\underline{\underline{T}}\,\underline{x} = (\underline{\underline{T}}\,\underline{x})^T\underline{\underline{\Lambda}}(\underline{\underline{T}}\,\underline{x})$, and since $\underline{\underline{\Lambda}}$ diagonal, $\underline{\underline{T}}$ full rank, $\underline{x}^T\underline{\underline{A}}\,\underline{x} \geq 0 \Leftrightarrow \lambda_{ii} \geq 0$. Similarly, $\underline{x}^T\underline{\underline{A}}\,\underline{x} > 0 \Leftrightarrow \lambda_{ii} > 0$ and $\underline{x} \neq \underline{0}$.

Theorem: Let $\underline{\underline{A}}$ be nonnegative definite. $\det(\underline{\underline{A}}) = 0$ iff the smallest eigenvalue of $\underline{\underline{A}}$ is zero.

Proof: $\det(\underline{\underline{A}}) = \det(\underline{\underline{T}}\,\underline{\underline{\Lambda}}\,\underline{\underline{T}}^T) = \det(\underline{\underline{\Lambda}}) = \prod_{i=1}^{n}\lambda_i$.

Theorem: Let $\underline{\underline{A}} \in \mathbb{R}^{n\times n}$ be symmetric with all eigenvalues either 0 or 1. Then $\underline{\underline{A}}^2 = \underline{\underline{A}}$.

Proof: $\underline{\underline{A}} = \underline{\underline{T}}\,\underline{\underline{\Lambda}}\,\underline{\underline{T}}^T$. By $\lambda_i \in \{0,1\}$, $\underline{\underline{\Lambda}}^2 = \underline{\underline{\Lambda}}$. Hence,
$$\underline{\underline{A}}^2 = \underline{\underline{T}}\,\underline{\underline{\Lambda}}\,\underline{\underline{T}}^T\underline{\underline{T}}\,\underline{\underline{\Lambda}}\,\underline{\underline{T}}^T = \underline{\underline{T}}\,\underline{\underline{\Lambda}}^2\underline{\underline{T}}^T = \underline{\underline{T}}\,\underline{\underline{\Lambda}}\,\underline{\underline{T}}^T = \underline{\underline{A}}.$$

Theorem: Let $\underline{\underline{A}}$ be nonnegative definite and $\lambda_1, \dots, \lambda_n$ be the eigenvalues of $\underline{\underline{A}}$ in nonincreasing order: $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$. Then, the largest eigenvalue is $$\lambda_1 = \max_{\|\underline{x}\|=1} \underline{x}^T\underline{\underline{A}}\,\underline{x}.$$

Proof: By $\underline{A}$ nonnegative definite, $\max_{\|\underline{x}\|=1} \underline{x}^T \underline{A}\, \underline{x} = \max_{\|\underline{y}\|=1} \underline{y}^T \underline{\Lambda}\, \underline{y}$. This is easier as $\underline{y}^T \underline{\Lambda}\, \underline{y} = \sum_{i=1}^{n} \lambda_i y_i^2$. To maximize this, set $\underline{y} = (1,0,\dots,0)^T$, giving us $\underline{y}^T \underline{\Lambda}\, \underline{y} = \lambda_1$.

Theorem: Let $\underline{A}$ be positive definite. Then

$$\lambda_n \|\underline{x}\|^2 \le \underline{x}^T \underline{A}\, \underline{x} \le \lambda_1 \|\underline{x}\|^2$$

where $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_n$ are the eigenvalues of $\underline{A}$.

Theorem: Let $\underline{A}$ be nonnegative definite. Then, we have that $\underline{A} = \underline{T}\,\underline{\Lambda}\,\underline{T}^T \Rightarrow \underline{A} = (\underline{T}\,\underline{\Lambda}^{1/2}\underline{T}^T)^2$, so we can find the square root of $\underline{A}$ to be $\underline{T}\,\underline{\Lambda}^{1/2}\underline{T}^T$ ($\underline{\Lambda}^{1/2} = \begin{bmatrix} \lambda_1^{1/2} & & \\ & \ddots & \\ & & \lambda_n^{1/2} \end{bmatrix}$).

(in general, $\underline{A}^{1/2} = \underline{B}$ if $\underline{B}\,\underline{B}^T = \underline{A}$)

9/6/23:

The standard normal is characterized by the density function

$$\varphi(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}, \quad N(0,1).$$

How to show integrates to 1? Square it:

$$\left[\int_{-\infty}^{\infty} \varphi(t)\, dt\right]^2 = \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-(t^2+s^2)/2}\, ds\, dt \xrightarrow{\text{Polar coords}} = 1.$$

The cumulative density function is given by

$$\phi(t) = \int_{-\infty}^{t} \varphi(x)\, dx.$$

Let $N \sim N(0,1)$.

1) $E[N] = \int_{-\infty}^{\infty} t\, \varphi(t)\, dt \overset{\text{(anti-symmetric)}}{=} 0$

2) $E[N^2] = \int_{-\infty}^{\infty} t^2 \cdot \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, dt \qquad u = t \qquad dv = te^{-t^2/2}$

$$\overset{\text{IBP}}{=} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, dt \overset{\text{int. pdf}}{=} 1$$

3) $E[N^k] = 0$ for $k = 1, 3, 5, \dots$ by anti-symmetric

4) Now let k be even

$$E[N^k] = \int_{-\infty}^{\infty} t^k \cdot \frac{1}{\sqrt{2\pi}} e^{-t^2/2} \, dt$$

$$= \int_{-\infty}^{\infty} t^{k-1} \cdot \frac{1}{2\pi} \cdot t e^{-t^2/2} \, dt$$

$$\overset{IBP}{=} \int_{-\infty}^{\infty} (k-1) t^{k-2} \cdot \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$$

$$= (k-1) E[N^{k-2}].$$

$$= \prod_{i=0}^{k/2+1} (k-1-2i)$$

The <u>Gamma Distribution</u> with shape parameter $K > 0$ is

$$\Gamma(K) = \int_0^{\infty} x^{K-1} e^{-x} \, dx.$$

The corresponding density function is

$$f(x) = \begin{cases} \frac{1}{\Gamma(K)} x^{K-1} e^{-x}, & x > 0 \\ 0, & x \leq 0 \end{cases}$$

<u>Note:</u> The density function is <u>not</u> unique. Can change countable # of points and the cdf $F(x) = \int_{-\infty}^{x} f(t) \, dt$ doesn't change.

<u>Feller's Inequality</u>: $\frac{x}{x^2+1} \varphi(x) \leq 1 - \Phi(x) \leq \frac{1}{x} \varphi(x), \quad x > 0$

> Will be proved in HW.

The <u>moment generating function</u> of the standard normal is

$$m_N(t) = E[e^{Nt}] = \int_{-\infty}^{\infty} e^{tx} \varphi(x) \, dx$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{tx - x^2/2} \, dx = \frac{1}{\sqrt{2\pi}} e^{t^2/2} \int_{-\infty}^{\infty} e^{-(t-x)^2/2} = e^{t^2/2}.$$

The <u>characteristic function</u> is

$$C_N(t) = E[e^{itX}].$$

Can show the complex part dissapears,

$$C_N(t) = e^{-t^2/2}.$$

If $X_1, \dots, X_n$ are independent, then the properties hold

$$m_{X_1+X_2+\dots+X_n}(t) = \prod_{i=1}^{n} m_{X_i}(t)$$

and $$C_{X_1+X_2+\dots+X_n}(t) = \prod_{i=1}^{n} C_{X_i}(t).$$

Suppose $X_1, \dots, X_n$ iid $\exp(1)$, $f_{X_i}(x) = e^{-x} \cdot I\{x \geq 0\}$. We can show that $\sum_{i=1}^{n} X_i \sim \text{Gamma}(n)$. This will be homework.

> Recall timeless property of exponential distribution.

$$\mathbb{P}[X_i > a+h \mid X_i > a] = \mathbb{P}[X_i > h]$$

Let $N_1, \dots, N_r$ be iid standard normal random variables. Can show that $\boxed{\sum_{n=1}^{r} (N_n)^2 \sim \chi^2(r)}$. Will also be on HW, use mgf. And from this, $E[\chi^2] = \sum_{n=1}^{r} E[N_n^2] = r$.

<u>Lemma:</u> $X_1, \dots, X_n$ iid. Then $\text{var}(X_1 + \dots + X_n) = \text{var}(X_1) + \dots + \text{var}(X_n)$.

So $\text{var}(\chi^2) = r \cdot \text{var}(N_1^2) = r(E[N_1^4] - E[N_1^2]^2)$

$= r \cdot (3-1) = 2r.$

We also have the property

$$\frac{\chi^2(r) - r}{\sqrt{2r}} \xrightarrow[\text{distribution}]{n \to \infty \text{ in}} N \quad \text{i.e. in cdf.}$$

<u>9/11/23:</u>

Recall the chi-squared distribution $\chi^2(r) = N_1^2 + \dots + N_r^2$ where each N_i is iid, $N(0,1)$. And the <u>T-distribution</u> is given by

$$t(r) = \frac{N(0,1)}{\sqrt{\chi^2(r)/r}}$$

with $N(0,1)$, $\chi^2(r)$ independent. And the <u>F-distribution</u> is given by

$$F(r_1, r_2) = \frac{\chi_1^2(r_1)/r_1}{\chi_2^2(r_2)/r_2}$$

with each χ_i^2 independent.

And, we can define $N(\mu, \sigma^2) := \mu + \sigma N(0,1)$.

Note also that
$$\mu - \sigma N(0,1) - N(\mu, \sigma^2).$$

Let $\underline{X} = (X_1, \dots, X_d)^T$ be a d-dimensional random vector.
The joint distribution is $P[X_1 < t_1, X_2 < t_2, \dots, X_d < t_d) = F_{\underline{x}}(\underline{t})$.
If we can write
$$F_{\underline{x}}(\underline{t}) = \int_{-\infty}^{t_1}\int_{-\infty}^{t_2}\cdots\int_{-\infty}^{t_d} f_{\underline{x}}(u_1, \dots, u_d)\, du_d\, du_{d-1} \cdots du_1,$$

then $f_{\underline{x}}$ is the joint density function. (which implies $f_{\underline{x}}(\underline{x}) = \prod_{i=1}^{d} f_{X_i}(x_i)$)
We say $X_1, \dots, X_d$ are independent iff $F_{\underline{x}}(\underline{t}) = \prod_{i=1} P(X_i < t_i)$.
The multinomial $(\underline{p}, n)$, with $\sum p_i = 1$, simulates number of outcomes, $1, \dots, d$, with corresponding probabilities, $p_1, \dots, p_d$, over n samples.
We can define the covariance
$$\text{cov}(X_i, X_j) = E[(X_i - E[X_i])(X_j - E[X_j])].$$
$$=: \sigma_{ij}$$

We then define the correlation
$$\text{correlation}(X_i, X_j) = \frac{\text{cov}(X_i, X_j)}{\sqrt{\text{var}(X_i)\,\text{var}(X_j)}}.$$
Suppose we're given random $\underline{X} \in \mathbb{R}^d$, $\underline{Y} \in \mathbb{R}^m$. Then,
$$\text{cov}(\underline{X}, \underline{Y}) = E[(\underline{X} - E[\underline{X}])(\underline{Y} - E[\underline{Y}])^T] \in \mathbb{R}^{d \times m}.$$

This will satisfy $\text{cov}(\underline{X}, \underline{Y}) = \text{cov}(\underline{Y}, \underline{X})^T$.
The covariance matrix of $\underline{X}$ is given by
$$\underline{\underline{\Sigma}} = \text{cov}(\underline{X}, \underline{X}) \in \mathbb{R}^{d \times d}.$$

We know $\underline{\underline{\Sigma}}$ is symmetric, nonnegative definite. Also, suppose that $\underline{\underline{\Sigma}}$ is singular. If $d = 1$, this means X_1 is constant. For $d > 1$, it means $\exists\ \underline{x} \neq \underline{0}$ s.t. $\underline{x}^T \underline{\underline{\Sigma}}\, \underline{x} = 0$. Or that some column is a linear combination of others. This implies that for some i, there exists a linear combination $X_i = \sum_{j \neq i} \alpha_j X_j$.

The multivariate normal random vector $\underline{X} \in \mathbb{R}^d$ can be defined as

$$\underline{X} = \underline{\mu} + \underline{\underline{A}}\,\underline{N}, \quad \underline{N} = \begin{pmatrix} N_1 \\ \vdots \\ N_d \end{pmatrix}, \text{ each } N_i \sim N(0,1).$$

It's a linear combination of standard normal. From this definition, $E[\underline{X}] = \underline{\mu}$. And it has covariance

$$\mathrm{cov}(\underline{X}) = E[\underline{\underline{A}}\,\underline{N}(\underline{\underline{A}}\,\underline{N})^T] = E[\underline{\underline{A}}\,\underline{N}\,\underline{N}^T\underline{\underline{A}}^T]$$

$$= \underline{\underline{A}}\,E[\underline{N}^T\underline{N}]\,\underline{\underline{A}}^T \overset{*}{=} \underline{\underline{A}}\,\underline{\underline{A}}^T, \text{ so } \underline{\underline{\Sigma}} = \underline{\underline{A}}\,\underline{\underline{A}}^T$$

$$* \quad E[\underline{N}\,\underline{N}^T]_{ij} = E[N_i N_j] = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise} \end{cases} \quad *$$

So we have that $\underline{\underline{A}} = \underline{\underline{\Sigma}}^{1/2}$. Note $\underline{\underline{\Sigma}}^{1/2}$ is not unique, can be acted on by orthogonal matrices, rotations. $\underline{\underline{\Sigma}}^{1/2}$ can be unique if we require it to be upper-triangular.

The multivariate moment generating function is given by

$$m_{\underline{X}}(\underline{t}) = m_{\underline{X}}(t_1, \dots, t_d) = E[e^{\underline{t}^T\underline{X}}].$$

For now, assume $\underline{\mu} = \underline{0}$. $\underline{t}^T\underline{X} = \underline{t}^T\underline{\underline{A}}\,\underline{N}$ is univariate normal as it is a linear combination of linear combinations of N_i, $i = 1, \dots, d$. It has mean 0. $\mathrm{var}(\underline{t}^T\underline{\underline{A}}\,\underline{N}) = E[\underline{t}^T\underline{\underline{A}}\,\underline{N}\,\underline{t}^T\underline{\underline{A}}\,\underline{N}] = E[\underline{t}^T\underline{\underline{A}}\,\underline{N}\,\underline{N}^T\underline{\underline{A}}^T\underline{t}]$

$= \underline{t}^T\underline{\underline{A}}\,E[\underline{N}\,\underline{N}^T]\,\underline{\underline{A}}^T\underline{t} = \underline{t}^T\underline{\underline{A}}\,\underline{\underline{A}}^T\underline{t} = \underline{t}^T\underline{\underline{\Sigma}}\,\underline{t}$. So, can finally show

$$m_{\underline{X}}(\underline{t}) = e^{\frac{1}{2}\underline{t}^T\underline{\underline{\Sigma}}\,\underline{t}}. \quad \text{(see HW)}$$

From this, we can see all you need to uniquely define $\underline{X}$ is $\underline{\mu}, \underline{\underline{\Sigma}}$.

9/13/23:

Let (X, Y) be bivariate normal. As we discussed before, we can write $\begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} \sigma_1 & 0 \\ \alpha & \beta \end{pmatrix}\begin{pmatrix} N_1 \\ N_2 \end{pmatrix}$. Suppose $\mathrm{var}(X) = \sigma_1^2$, $\mathrm{var}(Y) = \sigma_2^2$, and $\mathrm{cov}(X, Y) = \xi$. Then $Y = \alpha N_1 + \beta N_2 \Rightarrow \sigma_2^2 = \mathrm{var}(Y) = \alpha^2 + \beta^2$. And, $\xi = \mathrm{cov}(X, Y) = E[(X-0)(Y-0)] = E[\sigma_1\alpha N_1^2 + \sigma_1\beta N_1 N_2] = \sigma_1\alpha_1$.

Let $\underline{X} \in \mathbb{R}^d$ with density $f_{\underline{X}}(\underline{x})$. Let $\underline{h}: \mathbb{R}^d \to \mathbb{R}^d$ be a one-to-one transformation. Then let $\underline{Y} = \underline{h}(\underline{X}) \in \mathbb{R}^d$. If $\underline{g} = \underline{h}^{-1}$, then $\underline{X} = \underline{g}(\underline{Y})$. Let $\underline{\underline{J}}(\underline{y}) \in \mathbb{R}^{d \times d}$ be the Jacobian of $\underline{g}(\underline{y})$. Then the density of $\underline{Y}$ is given by

$$f_{\underline{Y}}(\underline{y}) = f_{\underline{X}}(\underline{g}(\underline{y})) \left| \det \underline{\underline{J}}(\underline{y}) \right|.$$

Let $\underline{X} \sim N_d(\underline{0}, \underline{\underline{\Sigma}})$. Then we also have $\underline{X} = \underline{\underline{A}}\,\underline{N}$. For this map to be invertible, require $\underline{\underline{A}}$ nonsingular. This true iff $\underline{\underline{\Sigma}} = \underline{\underline{A}}^T \underline{\underline{A}}$ nonsingular. The inverse is then given by $\underline{N} = \underline{\underline{A}}^{-1} \underline{X}$. It's Jacobian is then $\underline{\underline{J}}(\underline{x}) = \underline{\underline{A}}^{-1}$. We also know that

$$f_{\underline{N}}(\underline{n}) = \prod_{i=1}^{d} \frac{1}{\sqrt{2\pi}} e^{-n_i^2/2} = \frac{1}{(2\pi)^{d/2}} e^{-\underline{n}^T \underline{n}/2}.$$

Hence, we can find that

$$f_{\underline{X}}(\underline{x}) = \frac{1}{(2\pi)^{d/2}} e^{-(\underline{\underline{A}}^{-1}\underline{x})^T (\underline{\underline{A}}^{-1}\underline{x})/2} \cdot \left| \det \underline{\underline{A}}^{-1} \right|$$

$$= \frac{1}{(2\pi)^{d/2}} e^{-\frac{1}{2} \underline{x}^T \underline{\underline{\Sigma}}^{-1} \underline{x}} \cdot \frac{1}{\det(\underline{\underline{\Sigma}})^{1/2}}$$

Last step uses $\det(\underline{\underline{A}}^{-1}) = \det(\underline{\underline{A}})^{-1} = \det(\underline{\underline{\Sigma}}^{1/2})^{-1} = \det(\underline{\underline{\Sigma}})^{-1/2}$.

Theorem: Let $\underline{X} \in \mathbb{R}^d$, $\underline{Y} \in \mathbb{R}^m$ be uncorrelated, $\operatorname{cov}(\underline{X}, \underline{Y}) = \underline{\underline{0}}$, such that $[\underline{X}\ \underline{Y}]^T \in \mathbb{R}^{d+m} \sim N_{d+m}(\underline{0}, \underline{\underline{\Sigma}})$. Then $\underline{X}$ and $\underline{Y}$ are independent. In addition, each must be normal random vectors.

note: Independence means you can separate the densities or mgfs.

Proof: Consider the covariance matrix of $[\underline{X}, \underline{Y}]^T$,

$$\underline{\underline{\Sigma}} = \begin{bmatrix} \underline{\underline{\Sigma}}_1 & \underline{\underline{0}} \\ \underline{\underline{0}} & \underline{\underline{\Sigma}}_2 \end{bmatrix}.$$

And, the mgf is given by (let $\underline{t} = [\underline{x}\ \underline{y}]^T$)

$$m_{\underline{X}+\underline{Y}}(\underline{t}) = \exp\left(\tfrac{1}{2} \underline{t}^T \underline{\underline{\Sigma}}\, \underline{t}\right)$$

$$= \exp\left(\tfrac{1}{2}\left(\underline{x}^T \underline{\underline{\Sigma}}_1 \underline{x} + \underline{y}^T \underline{\underline{\Sigma}}_2 \underline{y}\right)\right)$$

$$= \exp\left(\tfrac{1}{2} \underline{x}^T \underline{\underline{\Sigma}}_1 \underline{x}\right) \exp\left(\tfrac{1}{2} \underline{y}^T \underline{\underline{\Sigma}}_2 \underline{y}\right)$$

$$= m_{\underline{X}}(\underline{x})\, m_{\underline{Y}}(\underline{y})$$

showing that $\underline{X}$ and $\underline{Y}$ are independent. Proof is similar if use density.

Note, $\underline{\underline{\Sigma}}^{-1} = \begin{bmatrix} \underline{\underline{\Sigma}}_1^{-1} & \underline{\underline{0}} \\ \underline{\underline{0}} & \underline{\underline{\Sigma}}_2^{-1} \end{bmatrix}$. So can similarly split density function.

<u>Theorem:</u> Let $\underline{X}_d \sim N_d(\underline{0}, \underline{\underline{I}}_d)$. Let $\underline{\underline{Q}}$ be an orthogonal matrix. Then $\underline{Y} = \underline{\underline{Q}}\,\underline{X} \sim N_d(\underline{0}, \underline{\underline{I}}_d)$.

<u>Proof:</u> Immediately from defn of $\underline{Y}$, we know it is normally distributed. So, $E[\underline{Y}] = E[\underline{\underline{Q}}\,\underline{X}] = \underline{\underline{Q}}\,E[\underline{X}] = \underline{0}$. And finally,

$$\mathrm{cov}(\underline{Y}) = E[\underline{\underline{Q}}\,\underline{X}(\underline{\underline{Q}}\,\underline{X})^T] = E[\underline{\underline{Q}}\,\underline{X}\,\underline{X}^T\underline{\underline{Q}}^T]$$

$$= \underline{\underline{Q}}\,E[\underline{X}\,\underline{X}^T]\,\underline{\underline{Q}}^T = \underline{\underline{Q}}\,\underline{\underline{I}}_d\,\underline{\underline{Q}}^T = \underline{\underline{Q}}\,\underline{\underline{Q}}^T = \underline{\underline{I}}_d.$$

Hence, we know $\underline{Y} \sim N_d(\underline{0}, \underline{\underline{I}}_d)$.

<u>Ex:</u> $\underline{X} \sim N_d(\underline{0}, \underline{\underline{\Sigma}})$, $\underline{X}_1 = \underline{\underline{A}}\,\underline{X}$, $\underline{X}_2 = \underline{\underline{B}}\,\underline{X}$. When are $\underline{X}_1, \underline{X}_2$ independent? Consider $[\underline{X}_1\ \underline{X}_2]^T$, we know that this is normal as it is equal to $[\underline{\underline{A}}\ \underline{\underline{B}}]^T\underline{X}$. By previous theorem, X_1, X_2 independent iff they're uncorrelated. Well,

$$\mathrm{cov}(\underline{X}_1, \underline{X}_2) = E[\underline{X}_1\underline{X}_2^T] = E[\underline{\underline{A}}\,\underline{X}\,\underline{X}^T\underline{\underline{B}}^T]$$

$$= \underline{\underline{A}}\,E[\underline{X}\,\underline{X}^T]\,\underline{\underline{B}}^T = \underline{\underline{A}}\,\underline{\underline{\Sigma}}\,\underline{\underline{B}}^T.$$

Hence, $\underline{X}_1, \underline{X}_2$ independent iff $\underline{\underline{A}}\,\underline{\underline{\Sigma}}\,\underline{\underline{B}}^T = \underline{\underline{0}}$.

<u>Theorem:</u> Let $\underline{\underline{P}}$ be symmetric, idempotent, $\underline{N} \sim N_d(\underline{0}, \underline{\underline{I}})$. Define $\underline{X} = \underline{\underline{P}}\,\underline{N}$, $\underline{Y} = (\underline{\underline{I}} - \underline{\underline{P}})\underline{N}$. Then $\underline{X}$ and $\underline{Y}$ are independent.

<u>Proof:</u> By previous theorem, need to show uncorrelated.

$$\mathrm{cov}(\underline{X}, \underline{Y}) = E[\underline{\underline{P}}\,\underline{N}\,\underline{N}^T(\underline{\underline{I}} - \underline{\underline{P}})] = \underline{\underline{P}}\,\underline{\underline{I}}(\underline{\underline{I}} - \underline{\underline{P}}) = \underline{\underline{0}}.$$

Tests for Normality:

1) χ^2 – Test
2) Shapiro-Wiles
   > Assumes iid, can't be used for residuals
3) Jorgue-Berra
   > Checks the first 4 moments to see if matching normal
4) Many many more

9/18/23:

Let $\underline{Y} \sim N_d(\underline{0}, \underline{\underline{I}})$, then we know a few things.
First, if $\underline{\underline{A}}$ symmetric, $\underline{Y}^T \underline{\underline{A}} \underline{Y} \sim \sum_{i=1}^{d} d_i N_i^2$, $d_i \geq 0$ are the eigenvalues of $\underline{\underline{A}}$.

$$\underline{Y}^T \underline{\underline{A}} \underline{Y} = \underline{Y}^T \underline{\underline{T}}^T \underline{\underline{\Lambda}} \underline{\underline{T}} \underline{Y} = (\underline{\underline{T}} \underline{Y})^T \underline{\underline{\Lambda}} (\underline{\underline{T}} \underline{Y}) \Rightarrow \underline{\underline{T}} \underline{Y} \sim N_d(\underline{0}, \underline{\underline{I}}) \Rightarrow \underline{Y} \sim N_d(\underline{0}, \underline{\underline{I}})$$

Recall that idempotent matrices have eigenvalues of 0 or 1.

Theorem: Let $\underline{\underline{A}} \in \mathbb{R}^{d \times d}$ be symmetric. Then $\underline{\underline{A}}^2 = \underline{\underline{A}}$ & $\text{rank}(\underline{\underline{A}}) = r$ iff $\underline{\underline{A}}$ has r eigenvalues that are 1, $d-r$ that are 0.

Theorem: Let $\underline{\underline{A}} \in \mathbb{R}^{d \times d}$ be symmetric, $\underline{Y} \sim N_d(\underline{0}, \underline{\underline{I}})$. Then $\underline{Y}^T \underline{\underline{A}} \underline{Y} \sim \chi_r^2 \iff \underline{\underline{A}}^2 = \underline{\underline{A}}$, $\text{rank}(\underline{\underline{A}}) = r$.

Proof: $\Leftarrow$ is clear. Now suppose $\underline{Y}^T \underline{\underline{A}} \underline{Y} \sim \chi_r^2$, so $\underline{Y}^T \underline{\underline{A}} \underline{Y} = \sum_{i=1}^{r} N_i^2$. We then know that $m_{\underline{Y}^T \underline{\underline{A}} \underline{Y}}(t) = m_{\chi_r^2}(t) = (1-2t)^{-r/2}$ on a neighborhood of $t=0$. We know that $\underline{Y}^T \underline{\underline{A}} \underline{Y} \sim \sum_{i=1}^{d} d_i N_i^2$, so we also have $m_{\underline{Y}^T \underline{\underline{A}} \underline{Y}}(t) = (1-2d_i t)^{-d/2}$. From these forms, we must have $d_i = 1$ for r terms, and $d_i = 0$ otherwise.

Again, let $\underline{Y} \sim N_d(\underline{0}, \underline{\underline{I}})$, and let $\underline{\underline{P}} \in \mathbb{R}^{d \times d}$ be a projection matrix with rank r. Then, by $\underline{\underline{P}}^2 = \underline{\underline{P}}$, $(\underline{\underline{I}} - \underline{\underline{P}})^2 = \underline{\underline{I}} - \underline{\underline{P}}$, $\text{range}(\underline{\underline{P}}) = \ker(\underline{\underline{I}} - \underline{\underline{P}})$ (and visa-versa)

1) $\underline{Y}^T \underline{\underline{P}} \underline{Y} \sim \chi_r^2$
2) $\underline{Y}^T (\underline{\underline{I}} - \underline{\underline{P}}) \underline{Y} \sim \chi_{d-r}^2$
3) $\underline{Y}^T \underline{\underline{P}} \underline{Y}$ and $\underline{Y}^T (\underline{\underline{I}} - \underline{\underline{P}}) \underline{Y}$ are independent.

Theorem: Let $\underline{\underline{A}}, \underline{\underline{B}} \in \mathbb{R}^{d \times d}$ be symmetric, $\underline{Y} \sim N_d(\underline{0}, \underline{\underline{I}})$ such that $\underline{Y}^T \underline{\underline{A}} \underline{Y} \sim \chi^2_r$, $\underline{Y}^T \underline{\underline{B}} \underline{Y} \sim \chi^2_m$. Then $\underline{Y}^T \underline{\underline{A}} \underline{Y}$ and $\underline{Y}^T \underline{\underline{B}} \underline{Y}$ are independent iff $\underline{\underline{A}}\,\underline{\underline{B}} = \underline{\underline{0}}$.

Proof: By statement, $\underline{\underline{A}}$ and $\underline{\underline{B}}$ are idempotent. Suppose independence, then $\underline{Y}^T \underline{\underline{A}} \underline{Y} + \underline{Y}^T \underline{\underline{B}} \underline{Y} \sim \chi^2$. And, by factoring, this equals $\underline{Y}^T(\underline{\underline{A}} + \underline{\underline{B}})\underline{Y}$. Hence, $(\underline{\underline{A}} + \underline{\underline{B}})$ is symmetric, idempotent, has rank $r+m$. By idempotence, $\underline{\underline{A}} + \underline{\underline{B}} = (\underline{\underline{A}} + \underline{\underline{B}})^2 = \underline{\underline{A}}^2 + 2\underline{\underline{A}}\,\underline{\underline{B}} + \underline{\underline{B}}^2$
$= \underline{\underline{A}} + 2\underline{\underline{A}}\,\underline{\underline{B}} + \underline{\underline{B}} \Rightarrow \underline{\underline{A}}\,\underline{\underline{B}} = \underline{\underline{0}}$. For the reverse, suppose $\underline{\underline{A}}\,\underline{\underline{B}} = \underline{\underline{0}}$. From that, $(\underline{\underline{A}} + \underline{\underline{B}})$, $\underline{\underline{A}}$, $\underline{\underline{B}}$ idempotent, getting us to the result.

Estimation Theory:

Let $Y_1, Y_2, \dots, Y_n$ be iid random variables, suppose their distribution depends on some parameters $\underline{\theta}$. Suppose parameters $\underline{\theta_0}$ used to sample $Y_1, Y_2, \dots, Y_n$. How to approximate $\underline{\theta_0}$?

1) <u>Method of Moments</u>: Match first r moments where r is the number of unknowns
2) <u>Least Squares</u>: Given $E[Y] = g(\theta)$, minimize $\sum_{i=1}^{n} (Y_i - g(\underline{\theta}))^2$ by modifying the set of parameters
3) <u>Least Absolute Deviation</u>: Given median $g(\underline{\theta})$, minimize $\sum_{i=1}^{n} |Y_i - g(\underline{\theta})|$
   - Note: ② makes sense mathematically in computing derivatives is easy, but the mean is <u>not</u> robust to outliers. If data has outliers, ③ may work better
4) <u>Maximum Liklihood Estimator</u>:

<u>9/20/23:</u>

- Let each Y_i have pdf (or pmf) $f(y, \underline{\theta})$.
- Define $L(\theta) = \prod_{i=1}^{n} f(Y_i, \underline{\theta})$, wish to find optimal $\hat{\theta} := \underset{\theta}{\arg\sup}\, L(\theta)$
- Instead of working with <u>liklihood function</u>, L, work with <u>log-liklihood</u>, $\ell(\theta) := \log(L(\theta)) = \sum_{i=1}^{n} \log(f(Y_i, \underline{\theta}))$, nicer to differentiate, as long as $f \neq 0$, differentiable

Ex: Consider Unif$(0, \theta)$ w/ pdf $\frac{1}{\theta} \cdot I\{0 \le x \le \theta\}$. Then
$L(\theta) = \prod_{i=1}^{n} f(X_i, \theta) = \theta^{-n} \prod_{i=1}^{n} I\{0 \le X_i \le \theta\} = \theta^{-n} I\{0 \le \min(\underline{X}) \le \max(\underline{X}) \le \theta\}$.
And from this form, we can see that $\hat{\theta} = \max(\underline{X})$ because $L(\theta) = 0$ for $\theta < \max(\underline{X})$, and $L(\theta) = \theta^{-n}$ for $\theta \ge \max(\underline{X})$.

Theorem: $$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \xrightarrow[]{\mathcal{D}} N\left(0, \frac{1}{I(\theta_0)}\right)$$ (in distribution) if f smooth w.r.t. θ (not proven here)

where $I(\theta_0)$ is the Fisher information number. And no estimator can do "better" than this.

Note: MLE is relatively robust to slight changes or "mistakes". Say data is "almost" normally distributed.

Linear Models

Suppose given $(y_i, \underline{x}_i)_{i=1}^{n}$, $\underline{x}_i$ observation, y_i is the result. Basic linear model given by
$$y_i = \underline{x}_i^T \underline{\beta} + \varepsilon_i, \quad i = 1, \dots, n.$$

Let $\underline{Y} = [y_1 \cdots y_n]^T$, $\underline{\underline{X}} = [\underline{x}_1 \cdots \underline{x}_n]^T$, $\underline{\varepsilon} = [\varepsilon_1 \cdots \varepsilon_n]^T$. This can be written as $\underline{Y} = \underline{\underline{X}}\underline{\beta} + \underline{\varepsilon}$. Suppose $\underline{\underline{X}}$ is full rank, i.e., no observation a linear combination of another. Let $L(\underline{\beta}) = \sum_{i=1}^{n} (y_i - \underline{x}_i^T \underline{\beta})^2$ in which we wish to minimize w.r.t. $\underline{\beta}$.

$$L(\underline{\beta}) = \|\underline{Y} - \underline{\underline{X}}\underline{\beta}\|_2^2 = (\underline{Y} - \underline{\underline{X}}\underline{\beta})^T(\underline{Y} - \underline{\underline{X}}\underline{\beta})$$

$$= \underline{Y}^T\underline{Y} - 2\underline{\underline{X}}^T\underline{Y}\underline{\beta} + \underline{\beta}^T\underline{\underline{X}}^T\underline{\underline{X}}\underline{\beta}$$

$$\Rightarrow \frac{dL}{d\underline{\beta}} = -2\underline{\underline{X}}^T\underline{Y} + 2\underline{\underline{X}}^T\underline{\underline{X}}\underline{\beta} = \underline{0}$$

$$\Rightarrow \hat{\underline{\beta}} = (\underline{\underline{X}}^T\underline{\underline{X}})^{-1}\underline{\underline{X}}^T\underline{Y}$$

And, $$\frac{d^2L}{d\underline{\beta}^2} = 2\underline{\underline{X}}^T\underline{\underline{X}}$$ which is pos. definite.

9/25/23:

Recall, $\underline{Y} = \underline{\underline{X}}\underline{\beta} + \underline{\varepsilon}$, $\underline{\underline{X}} \in \mathbb{R}^{n\times d}$, $\text{rank}(\underline{\underline{X}}) = d$, $\underline{\varepsilon} \sim N_n(0, \sigma^2 \underline{\underline{I}})$, $\underline{\beta} \in \mathbb{R}^d$, $\underline{Y} \in \mathbb{R}^n$. And the least squares solution given by

$$\underline{\hat{\beta}} = (\underline{\underline{X}}^T\underline{\underline{X}})^{-1}\underline{\underline{X}}^T\underline{Y}.$$

Then, we define the fitted values as $\hat{\underline{Y}} = \underline{\underline{X}}\underline{\hat{\beta}}$. Now, assume there is some "true" $\underline{\beta_0}$. Then

$$E[\hat{\beta}] = E[(\underline{\underline{X}}^T\underline{\underline{X}})^{-1}\underline{\underline{X}}^T(\underline{\underline{X}}\underline{\beta_0} + \underline{\varepsilon})]$$

$$= \underline{\beta_0} + (\underline{\underline{X}}^T\underline{\underline{X}})^{-1}\underline{\underline{X}}^T E[\underline{\varepsilon}]$$

$$= \underline{\beta_0}.$$

And, $$\text{Cov}(\underline{\hat{\beta}}) = E[\underline{\hat{\beta}}\underline{\hat{\beta}}^T] - E[\underline{\hat{\beta}}]E[\underline{\hat{\beta}}^T]$$

$$= E[(\underline{\beta_0} + (\underline{\underline{X}}^T\underline{\underline{X}})^{-1}\underline{\underline{X}}^T\underline{\varepsilon})(\underline{\beta_0}^T + \underline{\varepsilon}^T\underline{\underline{X}}(\underline{\underline{X}}^T\underline{\underline{X}})^{-1})] - \underline{\beta_0}\underline{\beta_0}^T$$

$$\overset{\left(\substack{E[\underline{\varepsilon}]=\underline{0} \\ E[\underline{\varepsilon}\underline{\varepsilon}^T]=\sigma^2\underline{\underline{I}}}\right)}{=} (\underline{\underline{X}}^T\underline{\underline{X}})^{-1}\underline{\underline{X}}^T\sigma^2\underline{\underline{I}}\,\underline{\underline{X}}(\underline{\underline{X}}^T\underline{\underline{X}})^{-1} = \sigma^2(\underline{\underline{X}}^T\underline{\underline{X}})^{-1}$$

<u>Gauss-Markov Theorem:</u> (LSE is BLUE): The least squares estimate is the best linear unbiased estimator. I.e., assuming $y = \underline{\beta_0}^T\underline{x} + \varepsilon$ with $\varepsilon \sim N(0, \sigma^2)$, then given $\underline{\underline{X}}, \underline{Y}$, the least squares estimate given by $\underline{\hat{\beta}} = (\underline{\underline{X}}^T\underline{\underline{X}})^{-1}\underline{\underline{X}}^T\underline{Y}$ is the unbiased linear estimator of $\underline{\beta_0}$ with minimized variance.

<u>Proof:</u> Let $\underline{\alpha} = \underline{\underline{C}}\underline{Y}$ be another estimator. It can be written in the form $\underline{\underline{C}} = (\underline{\underline{X}}^T\underline{\underline{X}})^{-1}\underline{\underline{X}}^T + \underline{\underline{D}}$. Hence,

$$E[\underline{\alpha}] = E[((\underline{\underline{X}}^T\underline{\underline{X}})^{-1}\underline{\underline{X}}^T + \underline{\underline{D}})(\underline{\underline{X}}\underline{\beta_0} + \underline{\varepsilon})]$$

$$= E[(\underline{\underline{I}} + \underline{\underline{D}}\,\underline{\underline{X}})\underline{\beta_0}] + ((\underline{\underline{X}}^T\underline{\underline{X}})^{-1}\underline{\underline{X}} + \underline{\underline{D}})E[\underline{\varepsilon}]$$

$$= E[\underline{\underline{I}} + \underline{\underline{D}}\,\underline{\underline{X}}]\,\underline{\beta}_0$$

which equals $\underline{\beta}_0$ (unbiased) iff $\underline{\underline{D}}\,\underline{\underline{X}} = \underline{\underline{0}}$. Finally,

$$\text{cov}(\underline{\alpha}) = E[\underline{\underline{C}}\,\underline{Y}\,\underline{Y}^T\underline{\underline{C}}^T] - E[\underline{\underline{C}}\,\underline{Y}]\,E[\underline{Y}^T\underline{\underline{C}}^T]$$

$$= \underline{\underline{C}}\,E[\underline{Y}\,\underline{Y}^T]\,\underline{\underline{C}}^T - \underline{\underline{C}}\,E[\underline{Y}]\,E[\underline{Y}^T]\,\underline{\underline{C}}^T$$

$$= \underline{\underline{C}}\,\text{cov}(\underline{Y})\,\underline{\underline{C}}^T$$

$$= \underline{\underline{C}}\,\sigma^2\,\underline{\underline{I}}\,\underline{\underline{C}}^T$$

$$= \sigma^2\left((\underline{\underline{X}}^T\underline{\underline{X}})^{-1}\underline{\underline{X}}^T + \underline{\underline{D}}\right)\left(\underline{\underline{X}}\,(\underline{\underline{X}}^T\underline{\underline{X}})^{-1} + \underline{\underline{D}}^T\right)$$

$$\overset{(\underline{\underline{D}}\,\underline{\underline{X}} = \underline{\underline{0}})}{=} \sigma^2\left((\underline{\underline{X}}^T\underline{\underline{X}})^{-1} + \underline{\underline{D}}\,\underline{\underline{D}}^T\right)$$

$$\overset{(\text{see above})}{=} \text{cov}(\hat{\underline{\beta}}) + \sigma^2\,\underline{\underline{D}}^T\underline{\underline{D}}.$$

And since $\underline{\underline{D}}^T\underline{\underline{D}}$ is positive semidefinite, $\text{cov}(\underline{\alpha})$ exceeds $\text{cov}(\hat{\underline{\beta}})$ unless $\underline{\underline{D}} = \underline{\underline{0}}$ in which case $\underline{\alpha} = \hat{\underline{\beta}}$. This could also be proven in terms of a scalar quantity of interest, $\underline{a}^T\underline{\beta}_0$. We'd show that $E[\underline{a}^T\hat{\underline{\beta}}] = \underline{a}^T E[\hat{\underline{\beta}}] = \underline{a}^T\underline{\beta}_0$, and that for any other estimator, $\alpha = \underline{a}^T\underline{\underline{C}}\,\underline{Y}$, $E[\alpha] = \underline{a}^T\underline{\beta}_0$, and $\text{var}(\alpha) = \text{var}(\underline{a}^T\hat{\underline{\beta}}) + \sigma^2\underline{a}^T\underline{\underline{D}}^T\underline{\underline{D}}\,\underline{a} \geq 0$.

9/27/23:

Define the residuals $\hat{\underline{\varepsilon}} = \underline{Y} - \hat{\underline{Y}} = \underline{\underline{X}}\,\underline{\beta}_0 + \underline{\varepsilon} - \underline{\underline{X}}\,\hat{\underline{\beta}} = \underline{\underline{X}}\,\underline{\beta}_0 + \underline{\varepsilon} - \underline{\underline{X}}\,((\underline{\underline{X}}^T\underline{\underline{X}})^{-1}\underline{\underline{X}}^T(\underline{\underline{X}}\,\underline{\beta}_0 + \underline{\varepsilon})$ $= (\underline{\underline{I}} - \underline{\underline{X}}\,(\underline{\underline{X}}^T\underline{\underline{X}})^T\underline{\underline{X}}^T)\,\underline{\varepsilon} = (\underline{\underline{I}} - \underline{\underline{P}})\,\underline{\varepsilon}$ where $\underline{\underline{P}}$ is the projection matrix of $\underline{Y}$ onto $\hat{\underline{Y}}$. By projection properties, $(\underline{\underline{I}} - \underline{\underline{P}})$ is also a projector. Since $(\underline{\underline{I}} - \underline{\underline{P}})$ is idempotent, $\hat{\underline{\varepsilon}}$ also normal. $E[\hat{\underline{\varepsilon}}] = (\underline{\underline{I}} - \underline{\underline{P}})\,E[\underline{\varepsilon}] = \underline{0}$, and $\text{cov}(\hat{\underline{\varepsilon}}) = (\underline{\underline{I}} - \underline{\underline{P}})\,\sigma^2\underline{\underline{I}}\,(\underline{\underline{I}} - \underline{\underline{P}})^T = \sigma^2(\underline{\underline{I}} - \underline{\underline{P}})$ as $\underline{\underline{P}}$ is symmetric. Hence, $\hat{\underline{\varepsilon}} \sim N_n(\underline{0}, \sigma^2(\underline{\underline{I}} - \underline{\underline{P}}))$. And, $\text{rank}(\underline{\underline{X}}^T\underline{\underline{X}}) = d$, so $\text{rank}(\underline{\underline{P}}) = d$, hence, $\text{rank}(\underline{\underline{I}} - \underline{\underline{P}}) = n - d$, so $\hat{\underline{\varepsilon}}$ really only consists of $n-d$ independent normals.

We can compute the joint distribution of $\underline{\hat{\beta}}$ and $\underline{\hat{\varepsilon}}$ as they are both normal. We just need the covariance matrix.

$$\operatorname{cov}(\underline{\hat{\beta}}, \underline{\hat{\varepsilon}}) = E[\underline{\hat{\beta}}\,\underline{\hat{\varepsilon}}^T] - \underbrace{E[\underline{\hat{\beta}}]\,E[\underline{\hat{\varepsilon}}^T]}_{\underline{0}}$$

$$= E\left[\left(\underline{\beta_0} + (\underline{\underline{X}}^T\underline{\underline{X}})^{-1}\underline{\underline{X}}^T\underline{\varepsilon}\right)\underline{\varepsilon}^T(\underline{\underline{I}} - \underline{\underline{P}})^T\right]$$

$$= \underline{\underline{0}} + (\underline{\underline{X}}^T\underline{\underline{X}})^{-1}\underline{\underline{X}}^T\sigma^2\underline{\underline{I}}(\underline{\underline{I}} - \underline{\underline{P}})$$

$$= \sigma^2(\underline{\underline{X}}^T\underline{\underline{X}})^{-1}\underline{\underline{X}}^T(\underline{\underline{I}} - \underline{\underline{X}}(\underline{\underline{X}}^T\underline{\underline{X}})^{-1}\underline{\underline{X}}^T)$$

$$= \underline{\underline{0}} \Rightarrow \begin{bmatrix}\underline{\hat{\beta}} \\ \underline{\hat{\varepsilon}}\end{bmatrix} \sim N_{n+d}\left(\begin{bmatrix}\underline{\beta_0} \\ \underline{0}\end{bmatrix}, \left[\begin{array}{c:c} \sigma^2(\underline{\underline{X}}^T\underline{\underline{X}}) & \underline{\underline{0}} \\ \hdashline \underline{\underline{0}} & \sigma^2(\underline{\underline{I}} - \underline{\underline{P}})\end{array}\right]\right)$$

implying that $\underline{\hat{\beta}}$, $\underline{\hat{\varepsilon}}$ are independent. And, we know that $S^2 = \frac{1}{n-d}\sum_{i=1}^{d}\hat{\varepsilon}_i^2 = \frac{1}{n-d}\|\underline{\hat{\varepsilon}}\|^2$, which only depends on $\underline{\hat{\varepsilon}}$, so, S^2 and $\underline{\hat{\beta}}$ are independent.

How about $\underline{\hat{\varepsilon}}^T\underline{\hat{\varepsilon}}$?

$$\underline{\hat{\varepsilon}}^T\underline{\hat{\varepsilon}} = \underline{\varepsilon}^T(\underline{\underline{I}} - \underline{\underline{P}})^T(\underline{\underline{I}} - \underline{\underline{P}})\underline{\varepsilon} \overset{\text{(idempotent)}}{=} \underline{\varepsilon}^T(\underline{\underline{I}} - \underline{\underline{P}})\underline{\varepsilon}.$$

And we proved that since $(\underline{\underline{I}} - \underline{\underline{P}})$ is idempotent, rank $n-d$, that then $\frac{1}{\sigma}\underline{\varepsilon}^T(\underline{\underline{I}} - \underline{\underline{P}})\underline{\varepsilon}\frac{1}{\sigma} \sim \chi^2_{n-d} \Rightarrow \sigma^2\underline{\hat{\varepsilon}}^T\underline{\hat{\varepsilon}} \sim \chi^2_{n-d}$. So from $S^2 = \frac{1}{n-d}\underline{\hat{\varepsilon}}^T\underline{\hat{\varepsilon}}$, $\frac{n-d}{\sigma^2}S^2 \sim \chi^2_{n-d}$. And from that $E[S^2] = \frac{\sigma^2}{n-d}\cdot(n-d) = \sigma^2$ (unbiased).

We can also show that $\frac{1}{\sigma^2}(\underline{\hat{\beta}} - \underline{\beta_0})^T\underline{\underline{X}}^T\underline{\underline{X}}(\underline{\hat{\beta}} - \underline{\beta_0}) \sim \chi^2_d$ (from $\underline{Y} \sim N_d(\underline{0}, \underline{\underline{\Sigma}})$ $\Rightarrow \underline{Y}^T\underline{\underline{\Sigma}}^{-1}\underline{Y} \sim \chi^2_d$).

<u>Definition:</u> The <u>F-distribution</u> is given by

$$F(r_1, r_2) = \frac{\chi^2_1(r_1)/r_1}{\chi^2_2(r_2)/r_2} \quad \text{with } \chi^2_1 \perp \chi^2_2$$

Then, from definition

$$\frac{(\underline{\hat{\beta}} - \underline{\beta_0})^T\underline{\underline{X}}^T\underline{\underline{X}}(\underline{\hat{\beta}} - \underline{\beta_0})/d}{S^2} \sim F(d, n-d)$$

Now, let $\hat{\beta}_i$, $\beta_{0,i}$ denote the i'th entry of $\underline{\hat{\beta}}$, $\underline{\beta_0}$ respectively.

Then we expect $\hat{\beta}_i - \beta_{0,i}$ to be normal with mean 0. We know $\text{cov}(\hat{\underline{\beta}}) = \sigma^2(\underline{\underline{X}}^T\underline{\underline{X}})^{-1}$, so $\text{var}(\hat{\beta}_i) = \sigma^2((\underline{\underline{X}}^T\underline{\underline{X}})^{-1})_{ii} =: \sigma^2 a_{ii}$. So $(\hat{\beta}_i - \beta_{0,i}) \sim N(0, \sigma^2 a_{ii})$, then $\frac{1}{\sigma\sqrt{a_{ii}}}(\hat{\beta}_i - \beta_{0,i}) \sim N(0,1)$.

Definition: The <u>t-distribution</u> is given by

$$t(r) = \frac{N(0,1)}{\sqrt{\chi^2(r)/r}} \quad \text{(w/ independence)}$$

Then

$$\frac{\hat{\beta}_i - \beta_{0,i}}{\sqrt{a_{ii} S^2}} \sim t(n-d).$$

<u>10/2/23:</u>

Now, let $\varepsilon_i \sim N(0, \sigma^2)$ iid. From $y_i = \underline{x}_i^T \underline{\beta} + \varepsilon_i$, we have $y_i \sim N(\underline{x}_i^T\underline{\beta}, \sigma^2)$, so has density $(2\pi\sigma^2)^{-1/2}\exp\left(\frac{-1}{2}\left(\frac{t - \underline{x}_i^T\underline{\beta}}{\sigma}\right)^2\right)$.

The <u>liklihood method</u> assigns a liklihood function

$$L(\underline{\beta}, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-1}{2}\left(\frac{y_i - \underline{x}_i^T\underline{\beta}}{\sigma}\right)^2\right)$$

$$= (2\pi\sigma^2)^{-n/2} \exp\left(\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \underline{x}_i^T\underline{\beta})^2\right)$$

$$= (2\pi\sigma^2)^{-n/2} \exp\left(\frac{-1}{2\sigma^2}\|\underline{Y} - \underline{\underline{X}}\,\underline{\beta}\|_2^2\right).$$

Then, the log-liklihood function, $l = \log(L)$, is

$$l(\underline{\beta}, \sigma) = \frac{-n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\|\underline{Y} - \underline{\underline{X}}\,\underline{\beta}\|_2^2.$$

Try to maximize $l \Rightarrow$ maximize L.

$$\frac{\partial l}{\partial \underline{\beta}} = \frac{1}{\sigma^2}\underline{\underline{X}}^T(\underline{Y} - \underline{\underline{X}}\,\underline{\beta}) = \underline{0} \Rightarrow \hat{\underline{\beta}} = (\underline{\underline{X}}^T\underline{\underline{X}}^{-1})\underline{\underline{X}}^T\underline{Y}.$$

$$\frac{\partial l}{\partial \sigma^2} = \frac{-n}{2\sigma^2} + \frac{1}{2\sigma^4}\|\underline{Y} - \underline{\underline{X}}\,\underline{\beta}\|_2^2 = 0$$

$$\Rightarrow \hat{\sigma}^2 = \frac{1}{n}\|\underline{Y} - \underline{\underline{X}}\,\underline{\beta}\|_2^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^2.$$

And, recall $$S_n^2 = \frac{1}{n-d} \sum_{i=1}^{n} \hat{\varepsilon}_i^2$$
is an unbiased estimator for σ^2.

The <u>weighted least squares</u> method, instead of minimizing $\|\underline{Y} - \underline{\underline{X}}\,\underline{\beta}\|_2^2 = (\underline{Y} - \underline{\underline{X}}\,\underline{\beta})^T(\underline{Y} - \underline{\underline{X}}\,\underline{\beta})$, we try to minimize
$$S(\underline{\beta}) = (\underline{Y} - \underline{\underline{X}}\,\underline{\beta})^T \underline{\underline{W}} (\underline{Y} - \underline{\underline{X}}\,\underline{\beta}) = \|\underline{Y} - \underline{\underline{X}}\,\underline{\beta}\|_{\underline{\underline{W}}}^2$$
where $\underline{\underline{W}}$ is full rank, positive definite. If $\underline{\underline{W}}$ is diagonal, then this is $S(\underline{\beta}) = \sum_{i=1}^{n} w_i (y_i - \underline{x}_i^T \underline{\beta})^2$. Can show this is still unbiased. This can be used where the ε terms have different variances.
Differentiating, find $$\underline{\hat{\beta}} = (\underline{\underline{X}}^T \underline{\underline{W}}\,\underline{\underline{X}})^{-1} \underline{\underline{X}}^T \underline{\underline{W}}\,\underline{Y}.$$

<u>Ex:</u> Let $y_i = \underline{x}_i^T \underline{\beta} + \varepsilon_i$ for $i = 1, \dots, n$ with $\varepsilon_i \sim N(0, \sigma_i^2)$, independent, but <u>not</u> identically distributed. Then, should weight $w_i \propto \frac{1}{\sigma_i^2}$, so let $w_i = \frac{1}{i}$. This way, each $w_i(y_i - \underline{x}_i^T \underline{\beta})^2 = \left(\frac{y_i - \underline{x}_i^T \underline{\beta}}{\sqrt{w_i}}\right)^2 \sim (N(0,\sigma^2))^2$.
<u>Ex:</u> Similarly, if $\underline{\varepsilon} \sim N_n(\underline{0}, \underline{\underline{\Sigma}})$, then choose $\underline{\underline{W}} = \underline{\underline{\Sigma}}^{-1}$.

<u>10/4/23:</u>
In weighted LS, $\underline{\hat{\beta}} = (\underline{\underline{X}}^T \underline{\underline{V}}\,\underline{\underline{X}})^{-1} \underline{\underline{X}}^T \underline{Y}$. So
$$E[\underline{\hat{\beta}}] = (\underline{\underline{X}}^T \underline{\underline{V}}\,\underline{\underline{X}})^{-1} \underline{\underline{X}}^T \underline{\underline{V}}\, E[\underline{\underline{X}}\,\underline{\beta} + \underline{\varepsilon}] = \underline{\beta} + \underline{0} = \underline{\beta}.$$
And,
$$\operatorname{cov}(\underline{\hat{\beta}}) = \operatorname{cov}((\underline{\underline{X}}^T \underline{\underline{V}}\,\underline{\underline{X}})^{-1} \underline{\underline{X}}^T \underline{\underline{V}}\,\underline{\varepsilon})$$
$$= (\underline{\underline{X}}^T \underline{\underline{V}}\,\underline{\underline{X}})^{-1} \underline{\underline{X}}^T \underline{\underline{V}}\, E[\underline{\varepsilon}\,\underline{\varepsilon}^T]\, \underline{\underline{V}}^T \underline{\underline{X}} (\underline{\underline{X}}^T \underline{\underline{V}}\,\underline{\underline{X}})^{-1}$$
$$= (\underline{\underline{X}}^T \underline{\underline{V}}\,\underline{\underline{X}})^{-1} \underline{\underline{X}}^T \underline{\underline{V}}\, \underline{\underline{\Sigma}}\, \underline{\underline{V}}\, \underline{\underline{X}} (\underline{\underline{X}}^T \underline{\underline{V}}\,\underline{\underline{X}})^{-1}.$$
What is the distribution of $\underline{\hat{\beta}}$? Since $\underline{\varepsilon} \sim N_n(\underline{0}, \underline{\underline{\Sigma}})$,

$$\hat{\underline{\beta}} \sim N_d\left(\underline{\beta}, (\underline{\underline{X}}^T \underline{\underline{V}}\, \underline{\underline{X}})^{-1} \underline{\underline{V}}\, \underline{\underline{\Sigma}}\, \underline{\underline{V}} (\underline{\underline{X}}^T \underline{\underline{V}}\, \underline{\underline{X}})^{-1}\right).$$

The location & scale family is as follows. Suppose we have a random variable X with density $f_0(t)$, called the mother density. Then define the scale $\sigma > 0$, and location $\mu \in \mathbb{R}$. Then define $Y = \mu + \sigma X$. We can find the density of Y to be $f_Y(t) = \frac{1}{\sigma} f_0\left(\frac{t-\mu}{\sigma}\right)$.

Now, suppose we have the linear model $y_i = \underline{x_i}^T \underline{\beta} + \varepsilon_i$. Suppose $\varepsilon_i \sim$ location/scale. We know $\mu = E[y_i] = \underline{x_i}^T \underline{\beta}$. Then, if we don't know σ^2, variance of ε_i, generate liklihood function

$$L(\underline{\beta}, \sigma) = \prod_{i=1}^{n} \frac{1}{\sigma} f_0\left(\frac{y_i - x_i^T \beta}{\sigma}\right)$$

$$\Rightarrow \ell(\underline{\beta}, \sigma) = -n \log(\sigma) + \sum_{i=1}^{n} \log\left(f_0\left(\frac{y_i - x_i^T \beta}{\sigma}\right)\right)$$

We wish to maximize ℓ w.r.t. $\underline{\beta}$, σ. For shorthand, we define $\rho(x) = \log(f_0(x))$, and $e_i(\underline{b}) = t - \underline{x_i}^T \underline{b}$. Note that if $f_0 \sim N(0,1)$, $\rho(x) = \frac{1}{2}x^2$. The solution to this are called the M-estimators.

10/18/23:

In hypothesis testing, we have some $\underline{X} \sim f(\underline{x}, \underline{\theta})$, $\underline{\theta} \in \Theta$. We form the null hypothesis $\underline{\theta} \in \Theta_0$, versus the alternative hypothesis $\underline{\theta} \in \Theta_a$ with $\Theta_0 \cup \Theta_a \subseteq \Theta$ and $\Theta_0 \cap \Theta_a = \emptyset$. We wish to either reject the null, H_0, or fail to reject H_0. We form a rejection region, C. If $\underline{X} \in C$, we reject H_0, if $\underline{X} \notin C$, we do not reject H_0. The power function is defined as

$$\pi_C(\hat{\underline{\theta}}) = \mathbb{P}[\underline{X} \in C \mid \hat{\underline{\theta}} = \underline{\theta}].$$

Two types of errors possible
1) Type I Error: We reject H_0 when H_0 is correct
2) Type II Error: We fail to reject H_0 when H_0 is incorrect.

We would like to bound

$$\max_{\hat{\underline{\theta}} \in \Theta_0} \pi_C(\hat{\underline{\theta}}) = \max_{\hat{\underline{\theta}} \in \Theta_0} P[\underline{X} \in C \mid \hat{\underline{\theta}} = \underline{\theta}] \leq \alpha.$$

I.e., we would like to bound how likely we reject given the null is true, bounding type I errors.

Ex: $X_1, \dots, X_n$ iid $N(\mu, 1)$. Wish to test $H_0: \mu = \mu_0$ versus $H_a: \mu > \mu_0$. So $\Theta_0 = \{\mu_0\}$, $\Theta_a = \{x : x > \mu\}$, $\Theta = \mathbb{R}$. Now, must form a rejection region. Clearly, it will take the form of $C = \{x : x > c\}$, and see if $\bar{X} \in C$. We know $\bar{X} \sim N(\mu, \frac{1}{n})$. Under H_0, $\bar{X} \sim N(\mu_0, \frac{1}{n})$. We can compute the power function

$$\pi_c(\mu_0) = P[\bar{X} > c \mid \bar{X} \sim N(\mu_0, \tfrac{1}{n})]$$

$$= P[N(0,1) > \sqrt{n}(c - \mu_0)]$$

$$= 1 - \Phi(\sqrt{n}(c - \mu_0)) \overset{\text{set}}{=} \alpha$$

and then use a computer to approximate c. Note that then the probability of rejection given Ho is true is α, i.e., α is the probability of a type I error. Can see from this,

$$\lim_{\mu \to \infty} \pi_c(\mu) = 1 \quad \text{and} \quad \lim_{n \to \infty} \pi_c(\mu) \underset{(\,\&\; c > \mu)}{\text{given } \mu > \mu_0} = 1.$$

We could create another test for $\text{median}(\underline{X}) \geq a$. Which is better? Turns out the first is better, as $\bar{X}$ is the maximum liklihood estimator for μ.

Now, instead, let H_0 be $\mu \le \mu_0$, H_a be $\mu > \mu_0$.

want $\alpha = \max_{\mu \le \mu_0} \pi_c(\mu)$ which occurs at $\mu = \mu_0$, so get same analysis as before.

So $\alpha = \pi_c(\mu_0) = 1 - \Phi(\sqrt{n}(c - \mu_0))$ again.

10/23/23:

Given rejection regions C and D, we say C is better than D if $\pi_C(\underline{\theta}) \ge \pi_D(\underline{\theta})$ $\forall$ $\underline{\theta} \in \Theta_a$, i.e., for every possible observation under the alternative hypothesis, we are more (or equally) likely to reject.

Ex: $X_1, \ldots, X_n$ iid Poisson(λ). Wish to test $H_0: \lambda \ge \lambda_0$ vs $H_a: \lambda < \lambda_0$ with λ_0 given. We know $E[X_i] = \lambda$. $\bar{X}_n$ is a good estimator for $\lambda = E[X_i]$. So, intuitively define rejection as $\bar{X} \le c$. Then, for a size of α,

$$\alpha = \sup_{\lambda \ge \lambda_0} \pi_c(\lambda)$$

$$= \sup_{\lambda \ge \lambda_0} P[\bar{X} \le c \mid \lambda].$$

We know $\sum_{i=1}^{n} X_i \sim \text{Pois}(n\lambda)$, so $n\bar{X} \sim \text{Pois}(n\lambda)$. So

$$\alpha = \sup_{\lambda \ge \lambda_0} P[\text{Pois}(n\lambda) \le nc]$$

$$= P[\text{Pois}(n\lambda_0) \le nc]$$

$$= \sum_{x=0}^{\lfloor nc \rfloor} \frac{(n\lambda_0)^x e^{-n\lambda_0}}{x!}.$$

Ex: $X_1, \ldots, X_n$ iid $N(\mu, \sigma^2)$ with μ, σ^2 unknown, wish to test $H_0: \mu = \mu_0$ vs. $H_a: \mu \ne \mu_0$. We call σ^2 a <u>nusance parameter</u> as it doesn't appear in the null or alternative. We will reject H_0 when $|\bar{x} - \mu|$ is large. We don't know σ^2, so approximate it with S^2. So, we will reject when $\frac{|\bar{X} - \mu|}{S/\sqrt{n}} \ge c$. We know this is distributed as $t(n-1)$. So this hypothesis test becomes a t-test.

Neyman-Pearson Lemma: Given $\underline{X} \sim f(\underline{x};\underline{\theta})$, wish to test $H_0: \underline{\theta} = \underline{\theta}_0$ vs. $H_a: \underline{\theta} = \underline{\theta}_a$ with $\underline{\theta}_0 \neq \underline{\theta}_a$. Then, defining the rejection region

$$C = \left\{ \underline{x} : \frac{f(\underline{x};\underline{\theta}_0)}{f(\underline{x};\underline{\theta}_a)} \leq c \right\},$$

it is optimal in the sense that $\pi_c(\underline{\theta}_a) \geq \pi_{c^*}(\underline{\theta}_a)$ for any other rejection region C^* for a fixed size α with

$$\alpha = \pi_c(\underline{\theta}_0) = \pi_{c^*}(\underline{\theta}_0).$$

The quantity $f(\underline{x};\underline{\theta}_0)/f(\underline{x};\underline{\theta}_a)$ is the liklihood ratio.

The generalized liklihood method with $H_0: \underline{\theta} \in \Theta_0$, $H_a: \underline{\theta} \in \Theta_a$, calls to reject H_0 if

$$\frac{\sup_{\underline{\theta} \in \Theta_0} f(\underline{x};\underline{\theta})}{\sup_{\underline{\theta} \in \Theta_a} f(\underline{x};\underline{\theta})} \leq c.$$

Alternatively, can reject by the liklihood ratio

$$\lambda(\underline{x}) = \frac{\sup_{\underline{\theta} \in \Theta_0} f(\underline{x};\underline{\theta})}{\sup_{\underline{\theta} \in \Theta_0 \cup \Theta_a} f(\underline{x};\underline{\theta})} \leq c \ (\leq 1).$$

It is very tricky to solve for c in terms of a size α, so instead use the rule to reject if $-2\log(\lambda(x)) \geq a$. Under H_0, we have $-2\log(\lambda(x)) \sim \chi^2(r)$, $r = \dim(\Theta_0 \cup \Theta_a) - \dim(\Theta_0)$.

Ex: $X_1, \dots, X_n$ iid $N(\mu, 1)$. Wish to test $H_0: \mu \leq \mu_0$ vs $H_a: \mu > \mu_0$. The liklihood is given by

$$f(\underline{x};\mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x_i-\mu)^2} = (2\pi)^{-n/2} e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2}.$$

Now, to find the liklihood ratio, first we already know

$$\max_{\mu} f(\underline{x};\mu) = f(\underline{x}, \bar{x}).$$

Additionally, f has a unique maximizer, so

$$\max_{\mu \in \mu_0} f(\underline{x}; \mu) = \begin{cases} \bar{X} & \mu_0 \geq \bar{X} \\ \mu_0 & \text{otherwise} \end{cases} = \min(\mu_0, \bar{X}).$$

Hence, we can compute the liklihood ratio

$$\lambda(\underline{x}) = \frac{f(\underline{x}; \min(\mu_0, \bar{x}))}{f(\underline{x}; \bar{X})}$$

which equals one when $\bar{X} \leq \mu$. Recall, we reject when $\lambda(\underline{x})$ is small, i.e., $\bar{X}$ is large relative to μ_0. This makes sense with respect to definition of H_0, H_a. Going back to log,

$$-2\log(\lambda(\underline{x})) \sim \chi^2_r \quad \text{with } r = \dim(\Theta_0 \cup \Theta_a) - \dim(\Theta_0) = 1 - 1 = 0$$

so this method doesn't work.

10/25/23:

Consider the model $\underline{Y} = \underline{\underline{X}}\,\underline{\beta} + \underline{\varepsilon}$, $\underline{Y} \in \mathbb{R}^n$, $\underline{\underline{X}} \in \mathbb{R}^{n \times p}$ with rank p, $\underline{\beta} \in \mathbb{R}^p$, $\underline{\varepsilon} \sim N_n(\underline{0}, \sigma^2 \underline{\underline{I}})$. Let $\underline{\underline{A}} \in \mathbb{R}^{q \times p}$ with $\text{rank}(\underline{\underline{A}}) = q$, and $\underline{c} \in \mathbb{R}^q$. We wish to estimate $\underline{\beta}$ under the restriction $\underline{\underline{A}}\,\underline{\beta} = \underline{c}$. We introduce Lagrange multipliers, so we wish to minimize

$$g(\beta, \lambda) = \|\underline{Y} - \underline{\underline{X}}\,\underline{\beta}\|_2^2 + (\underline{\underline{A}}\,\underline{\beta} - \underline{c})^T \underline{\lambda}.$$

$$\frac{dg}{d\lambda}(\hat{\underline{\beta}}_H, \hat{\underline{\lambda}}_H) = \underline{\underline{A}}\,\hat{\underline{\beta}}_H - \underline{c} = \underline{0} \Rightarrow \underline{\underline{A}}\,\hat{\underline{\beta}}_H = \underline{c},$$

$$\frac{dg}{d\underline{\beta}}(\hat{\underline{\beta}}_H, \hat{\underline{\lambda}}_H) = -2\underline{\underline{X}}^T\underline{Y} + 2\underline{\underline{X}}^T\underline{\underline{X}}\,\hat{\underline{\beta}}_H + \underline{\underline{A}}^T\hat{\underline{\lambda}}_H = \underline{0}$$

$$\Rightarrow \hat{\underline{\beta}}_H = (\underline{\underline{X}}^T\underline{\underline{X}})^{-1}\underline{\underline{X}}^T\underline{Y} - \frac{1}{2}(\underline{\underline{X}}^T\underline{\underline{X}})^{-1}\underline{\underline{A}}^T\hat{\underline{\lambda}}_H$$

$$\Rightarrow \underline{c} = \underline{\underline{A}}\,\hat{\underline{\beta}}_H = \underline{\underline{A}}\,\hat{\underline{\beta}} - \frac{1}{2}\underline{\underline{A}}(\underline{\underline{X}}^T\underline{\underline{X}})^{-1}\underline{\underline{A}}^T\hat{\underline{\lambda}}_H$$

$$\Rightarrow \hat{\underline{\lambda}}_H = 2\left(\underline{\underline{A}}(\underline{\underline{X}}^T\underline{\underline{X}})^{-1}\underline{\underline{A}}^T\right)^{-1}(\underline{\underline{A}}\,\hat{\underline{\beta}} - \underline{c})$$

$$\Rightarrow \hat{\underline{\beta}}_H = \hat{\underline{\beta}} - (\underline{\underline{X}}^T\underline{\underline{X}})^{-1}\underline{\underline{A}}^T(\underline{\underline{A}}(\underline{\underline{X}}^T\underline{\underline{X}})^{-1}\underline{\underline{A}}^T)^{-1}(\underline{\underline{A}}\hat{\underline{\beta}} - \underline{c})$$

where $\hat{\underline{\beta}}$ is the standard least squares solution. Since this is a linear operator on $\hat{\underline{\beta}}$, we must have that $\hat{\underline{\beta}}_H$ is also normal. Now,

$$E[\hat{\underline{\beta}}_H] = \underline{\beta} - (\cdots\cdots)(\underline{\underline{A}}\underline{\beta} - \underline{c})$$

$$= \underline{\beta} \quad \text{if } \underline{\underline{A}}\underline{\beta} = \underline{c}.$$

Note if $\underline{\underline{A}}\underline{\beta} \neq \underline{c}$ (even though $\underline{\underline{A}}\hat{\underline{\beta}}_H = \underline{c}$), then this is not an unbiased estimator for $\underline{\beta}$. Next, we could solve for the covariance matrix of $\hat{\underline{\beta}}_H$, but it will be ugly.

An interesting question is to compare

$$RSS = \|\underline{Y} - \underline{\underline{X}}\hat{\underline{\beta}}\|_2^2 \quad \text{and} \quad RSS_H = \|\underline{Y} - \underline{\underline{X}}\hat{\underline{\beta}}_H\|_2^2.$$

It is clear that $RSS \leq RSS_H$ since the "H" problem is constrained.

10/30/23:

We previously showed that

$$\|\underline{Y} - \underline{\underline{X}}\underline{\beta}\|_2^2 = \|\underline{Y} - \underline{\underline{X}}\hat{\underline{\beta}}\|_2^2 + \|\underline{\underline{X}}(\hat{\underline{\beta}} - \underline{\beta})\|_2^2.$$

Setting $\underline{\beta} = \hat{\underline{\beta}}_H$,

$$\|\underline{Y} - \underline{\underline{X}}\hat{\underline{\beta}}_H\|_2^2 = \|\underline{Y} - \underline{\underline{X}}\hat{\underline{\beta}}\|_2^2 + \|\underline{\underline{X}}(\hat{\underline{\beta}} - \hat{\underline{\beta}}_H)\|_2^2$$

or, written in words:

$$RSS_H = RSS + \|\underline{\underline{X}}(\hat{\underline{\beta}} - \hat{\underline{\beta}}_H)\|_2^2.$$

Can derive that

$$\frac{(RSS_H - RSS)/q}{RSS/(n-p)} = \frac{(\underline{\underline{A}}\hat{\underline{\beta}} - \underline{c})(\underline{\underline{A}}(\underline{\underline{X}}^T\underline{\underline{X}})^{-1}\underline{\underline{A}})^{-1}(\underline{\underline{A}}\hat{\underline{\beta}} - \underline{c})/q}{s^2} \sim F(q, n-p).$$

$(q = \text{rank}(\underline{\underline{A}}))$

So, can use an F-test to investigate RSS_H vs RSS.

11/1/23:

Given ordinary linear model $\underline{Y} = \underline{\underline{X}}\,\underline{\beta} + \underline{\varepsilon}$, $\underline{\beta} \in \mathbb{R}^p$, $\underline{\underline{X}} \in \mathbb{R}^{n \times p}$ full rank (p), $\underline{Y} \in \mathbb{R}^n$, $\underline{\varepsilon} \sim N_n(\underline{0}, \sigma^2 \underline{\underline{I}})$. Then, given $\underline{\underline{A}} \in \mathbb{R}^{q \times p}$ rank (q), $\underline{c} \in \mathbb{R}^q$, and test $H_0 : \underline{\underline{A}}\,\underline{\beta} = \underline{c}$ vs H_a: H_0 not true.

Recall the liklihood function

$$f(\underline{\beta}, \sigma^2) = (2\pi\sigma^2)^{-n/2} e^{-\frac{1}{2\sigma^2}\|\underline{Y} - \underline{\underline{X}}\,\underline{\beta}\|_2^2}.$$

The liklihood ratio is then given by

$$\frac{\max_{\underline{\underline{A}}\,\underline{\beta} = \underline{c}} f(\underline{\beta}, \sigma^2)}{\max_{\underline{\beta}, \sigma^2} f(\underline{\beta}, \sigma^2)} = \frac{\max_{\underline{\underline{A}}\,\underline{\beta} = \underline{c}} f(\underline{\beta}, \sigma^2)}{f(\hat{\underline{\beta}}, \hat{\sigma}^2)}$$

where $\hat{\underline{\beta}} = (\underline{\underline{X}}^T \underline{\underline{X}})^{-1} \underline{\underline{X}}^T \underline{Y}$, $\hat{\sigma}^2 = \frac{1}{n}\|\underline{Y} - \underline{\underline{X}}\,\hat{\underline{\beta}}\|_2^2$. Now, using def of $\hat{\sigma}^2$,

$$f(\hat{\underline{\beta}}, \hat{\sigma}^2) = (2\pi\hat{\sigma}^2)^{-n/2} e^{-n/2}.$$

Now, we also solved for the constrained $\hat{\underline{\beta}}_H$, and $\hat{\sigma}_H^2 = \frac{1}{n}\|\underline{Y} - \underline{\underline{X}}\,\hat{\underline{\beta}}_H\|_2^2$, so

$$f(\hat{\underline{\beta}}_H, \hat{\sigma}_H^2) = (2\pi\hat{\sigma}_H^2) e^{-n/2},$$

hence,

$$\lambda_{LR} = \frac{\max_{\underline{\underline{A}}\,\underline{\beta} = \underline{c}} f(\underline{\beta}, \sigma^2)}{\max_{\underline{\beta}, \sigma^2} f(\underline{\beta}, \sigma^2)} = \left(\frac{\hat{\sigma}^2}{\hat{\sigma}_H^2}\right)^{n/2}.$$

We reject H_0 if $\lambda_{LR} = (\hat{\sigma}^2 / \hat{\sigma}_H^2)^{n/2}$ is small, i.e., $\hat{\sigma}_H^2 / \hat{\sigma}^2$ large, i.e., $\frac{\hat{\sigma}_H^2 - \hat{\sigma}^2}{\hat{\sigma}^2}$ large, i.e., $\frac{RSS_H - RSS}{RSS}$ large (we know $RSS_H \geq RSS$).

Ex: 2 samples: $U_i = \mu_1 + \varepsilon_i$, $i = 1, \dots, n_1$, $V_i = \mu_2 + \eta_i$, $i = 1, \dots, n_2$, with ε_i, η_i iid $N(\mu, \sigma^2)$. Wish to test $H_0 : \mu_1 = \mu_2$ against $H_a : \mu_1 \neq \mu_2$. Set up linear model

$$\underline{Y} = \begin{bmatrix} U_1 \\ \vdots \\ U_{n_1} \\ V_1 \\ \vdots \\ V_{n_2} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} + \underline{\varepsilon} = \underline{\underline{X}}\,\underline{\beta} + \underline{\varepsilon}.$$

And, wish to test if

$$\underline{\underline{A}}\,\underline{\beta} = \begin{bmatrix} 1 & -1 \end{bmatrix}\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} = 0 = c.$$

So, we first calculate that $\underline{\underline{X}}^T\underline{\underline{X}} = \begin{bmatrix} n_1 & 0 \\ 0 & n_2 \end{bmatrix}$, so $(\underline{\underline{X}}^T\underline{\underline{X}})^{-1} = \begin{bmatrix} 1/n_1 & 0 \\ 0 & 1/n_2 \end{bmatrix}$. Next, without restriction,

$$\begin{bmatrix} \hat{\mu}_1 \\ \hat{\mu}_2 \end{bmatrix} = \hat{\underline{\beta}} = (\underline{\underline{X}}^T\underline{\underline{X}})^{-1}\underline{\underline{X}}^T\underline{Y} = \begin{bmatrix} \bar{U} \\ \bar{V} \end{bmatrix}.$$

Next, define $\check{\mu} = \frac{1}{n_1+n_2}\left(\sum_{i=1}^{n_1} U_i + \sum_{i=1}^{n_2} V_i\right)$. From former work,

$$RSS_H - RSS = (\underline{\underline{A}}\hat{\underline{\beta}} - \underline{c})^T(\underline{\underline{A}}(\underline{\underline{X}}^T\underline{\underline{X}})^{-1}\underline{\underline{A}}^T)(\underline{\underline{A}}\hat{\underline{\beta}} - \underline{c})$$

$$= (\bar{U} - \bar{V})\left(\begin{bmatrix} 1 & -1 \end{bmatrix}\begin{bmatrix} 1/n_1 & 0 \\ 0 & 1/n_2 \end{bmatrix}\begin{bmatrix} 1 \\ -1 \end{bmatrix}\right)^{-1}(\bar{U} - \bar{V})$$

$$= (\bar{U} - \bar{V})^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{-1}.$$

Now, under H_0, $\bar{U} - \bar{V} \sim N\left(\mu_1 - \mu_2, \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}\right) = N\left(0, \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)$.
Hence, $\frac{1}{\sigma^2}(\bar{U} - \bar{V})^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{-1} \sim \chi^2_1$. Additionally, we know that $RSS = \|\underline{Y} - \underline{\underline{X}}\hat{\underline{\beta}}\|_2^2 = (n_1 + n_2)\hat{\sigma}^2$, so

$$\frac{RSS_H - RSS}{RSS} = \frac{(\bar{U} - \bar{V})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{-1}}{(n_1 + n_2)\hat{\sigma}^2}$$

test if this value is large.

If the previous problem is extended to 3 variables, U_i, V_i, Z_i, then it is more common to use ANOVA w/ 3 populations.

Now, consider the model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$. We can define the R^2 value as

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{\left(\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})\right)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

which can be interpreted as a correlation between y_i, x_i if each are random. Or, interpreted as a ratio of the model's (β_0, β_1's) impact to that of just the mean (β_0, β_1 not in model).

11/6/23:

Resampling:

1) Jackknife provides a simulation based method to estimate σ (unknown)

2) Bootstrap computes the density function of a statistic

Lemma: Let $X_1, \dots, X_n$ be iid r.v.s with cdf $F(x)$. Define the empirical density function,

$$F_n(t) = \frac{1}{n} \sum_{i=1}^{n} I[X_i \le t].$$

Then,

1) $F_n(t) \xrightarrow{P} F(t)$ for all t

2) $\sup_t |F_n(t) - F(t)| \xrightarrow{P} 0$

"Proof": The first statement falls out of the law of large numbers because $E[I[X_i \le t]] = 1 \cdot P[X_i \le t] = F(t)$. The second statement is the Fundamental Theorem of Statistics.

Now, from the above lemma, a few facts:

1) $n F_n(t) = \sum_{i=1}^{n} I[X_i \le t] \sim \text{Binomial}(n, F(t))$

2) $$\sqrt{n}\left(F_n(t) - F(t)\right) = \frac{n F_n(t) - n F(t)}{\sqrt{n}} \xrightarrow{D} N\left(0, F(t)(1-F(t))\right)$$

due to normalization & variance of Bernoulli(p) = Binom$(1,p)$ is $p(1-p)$.

3) $\sqrt{n} \sup_t |F_n(t) - F(t)| \xrightarrow{D}$ Kolmogorov-Smirnov

Lemma: Let $X_1, \dots, X_n$ iid with cdf F. Let X be a random variable where X is one of $X_1, \dots, X_n$ selected uniformly randomly. We know that (assuming each X_i not equal), $P[X = X_i \mid X_1, \dots, X_n] = \frac{1}{n}$, and $E[X \mid X_1, \dots, X_n] = \bar{X}$. Additionally, $P[X \le t \mid X_1, \dots, X_n] = F_n(t)$.

Back to model $\underline{Y} = \underline{X}\beta + \underline{\varepsilon}$, and test $H_0: \beta \in \Theta_0$ vs $H_a: \beta \notin \Theta_0$. Wish to find a statistic Δ_n s.t. $\Delta_n \xrightarrow{D} \xi$ under H_0, and $\Delta_n \xrightarrow{P} \infty$ under H_a.

$$\text{(cdf of } \Delta_n) \qquad \text{(cdf of } \xi)$$

Note, $\Delta_n \xrightarrow{D} \xi \Rightarrow G_n(t) \longrightarrow G(t)$ for all t except for on set of measure zero. And for a test of size α, wish to choose c_n s.t.

$$\alpha = 1 - G_n(c_n) = P[\xi \geq c_n] \approx P[\Delta_n \geq c_n].$$

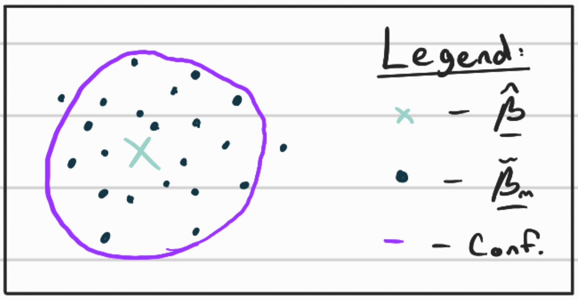
Bootstrap procedure: (Efron Bootstrap 1970's, Stanford):

1) Estimate $\hat{\beta}_H$ to be a constrained regression solution ($\hat{\underline{\beta}}_H \in \Theta_0$)
2) Generate residuals $\hat{\underline{\varepsilon}} = \underline{Y} - \underline{\underline{X}}\, \hat{\underline{\beta}}_H$. (without is negligible for n large)
3) Sample $i_1, \cdots, i_k$ from $\{1, \ldots, n\}$ with replacement
4) Compute $\check{\Delta}_{k,m}$ from $\hat{y}_{i_j} = \underline{x}^T \hat{\underline{\beta}}_H$, $j = 1, \ldots, k$, for $m = 1, \ldots, M$.
5) $\frac{1}{M} \sum_{m=1}^{M} I[\check{\Delta}_{k,m} \leq t] \approx G(t)$ (approximate cdf)

11/8/23:

Consider the same model $\underline{Y} = \underline{\underline{X}}\,\underline{\beta} + \underline{\varepsilon}$ with $\underline{\beta} \in \mathbb{R}^2$. We wish to find a confidence set with $1-\alpha$ coverage for $\underline{\beta}$. Let $\hat{\underline{\beta}}$ be the LSE for $\underline{\beta}$, and define $\hat{\underline{\varepsilon}} = \underline{Y} - \underline{\underline{X}}\,\hat{\underline{\beta}}$. Then, for $m = 1, \ldots, M$, sample from the data, and generate the bootstrapped LSE $\tilde{\underline{\beta}}_m$, $\tilde{\underline{\varepsilon}}_m$. Next, we create a cloud which contains $(1-\alpha)\%$ of the $\tilde{\underline{\beta}}_m$'s.

Note: This region is not unique, its shape / what is included or excluded is up to the statistician.



11/13/23:

Consider the same set-up as before. We can use Tikhonov regularization or ridge regularization, to generate a ridge estimator $\hat{\beta}_{ridge}$.

$$L(\underline{\beta}) = \| \underline{Y} - \underline{\underline{X}}\,\underline{\beta} \|_2^2 + \lambda \| \underline{\beta} \|_2^2$$

$$= \underline{Y}^T \underline{Y} - 2 \underline{\beta}^T \underline{\underline{X}}^T \underline{Y} + \underline{\beta}^T \underline{\underline{X}}^T \underline{\underline{X}}\, \underline{\beta} + \lambda \underline{\beta}^T \underline{\beta}$$

$$\Rightarrow \frac{dL(\underline{\beta})}{d\underline{\beta}} = -2\underline{\underline{X}}^T\underline{Y} + 2\underline{\underline{X}}^T\underline{\underline{X}}\,\underline{\beta} + 2\lambda\underline{\beta} = \underline{0}$$

$$\Rightarrow \underline{\hat{\beta}_{ridge}} = (\underline{\underline{X}}^T\underline{\underline{X}} + \lambda \underline{\underline{I}})^{-1}\underline{\underline{X}}^T\underline{Y}.$$

Note, the ridge estimator is biased, $E[\hat{\underline{\beta}}_{ridge}] \neq \underline{\beta}$ (for $\lambda \neq 0$).
We can rewrite it as

$$\underline{\hat{\beta}_{ridge}} = (\underline{\underline{X}}^T\underline{\underline{X}} + \lambda\underline{\underline{I}})^{-1}(\underline{\underline{X}}^T\underline{\underline{X}})\underline{\hat{\beta}} =: \underline{\underline{C}}\,\underline{\beta}.$$

So, we see $\quad \underline{\hat{\beta}_{ridge}} \sim N_d(\underline{\underline{C}}\,\underline{\beta}, \sigma^2\underline{\underline{C}}^T(\underline{\underline{X}}^T\underline{\underline{X}})^{-1}\underline{\underline{C}}).$

The <u>mean squared error, MSE,</u> is defined to be

$$MSE(\hat{\underline{\beta}}) = E\left[\|\hat{\underline{\beta}} - \underline{\beta}\|_2^2\right].$$

Can show that $MSE(\underline{\hat{\beta}_{ridge}}) < MSE(\hat{\underline{\beta}})$ for $0 < \lambda \ll 1$.

<u>Ex:</u> $\underline{Y} = \underline{\underline{X}}\,\underline{\beta} + \underline{\varepsilon}$, $H_0 : \underline{\beta} = \underline{\beta_0}$ vs $H_a: \underline{\beta} \neq \underline{\beta_0}$. We wish to find a test based on $\underline{\hat{\beta}_{ridge}}$.

We have that $\quad \dfrac{(\underline{\hat{\beta}_{ridge}} - \underline{\underline{C}}\,\underline{\beta_0})^T \underline{\underline{W}}(\underline{\hat{\beta}_{ridge}} - \underline{\underline{C}}\,\underline{\beta_0})/p}{S^2} \sim F(p, n-p)$

where $\quad S^2 = \dfrac{1}{n-p}\|\underline{Y} - \underline{\underline{X}}\hat{\underline{\beta}}\|_2^2, \quad \underline{\underline{W}} = (\underline{\underline{C}}^T(\underline{\underline{X}}^T\underline{\underline{X}})^{-1}\underline{\underline{C}})^{-1},$

so can do a two-sided test with this.

For a <u>lasso regularization,</u> we consider the loss function

$$L(\underline{\beta}) = \|\underline{Y} - \underline{\underline{X}}\,\underline{\beta}\|_2^2 + \lambda\|\underline{\beta}\|_1$$

where $\|\underline{\beta}\|_1 = \sum_{i=1}^{p}|\beta_i|$. Lasso shrinks small coordinates to zero while ridge shrinks all coordinates towards (but not to) zero. So lasso leads to a sparse model.

11/20/23:

Differentiating,

$$\frac{d\mathcal{L}}{d\underline{\beta}} = -2\underline{\underline{X}}^T\underline{Y} + 2\underline{\underline{X}}^T\underline{\underline{X}}\,\underline{\beta} + \lambda\,\mathrm{sign}(\underline{\beta}).$$

We then will assume the inputs are normalized, i.e., $\underline{\underline{X}}^T\underline{\underline{X}} = \underline{\underline{I}}$

$$\overset{\text{assump.}}{=} -2\underline{\underline{X}}^T\underline{Y} + 2\underline{\beta} + \lambda\,\mathrm{sign}(\underline{\beta}).$$

We can write the i^{th} component as

$$\sum_{j=1}^{n}(-2x_{ij}y_j) + 2\beta_i + \lambda\,\mathrm{sign}(\beta_i) = 0$$

$$\Rightarrow \quad \beta_i = \begin{cases} \sum_{j=1}^{n} x_{ij}y_j - \frac{\lambda}{2}, & \sum_{j=1}^{n} x_{ij}y_j - \frac{\lambda}{2} > 0 \\ \frac{\lambda}{2} - \sum_{j=1}^{n} x_{ij}y_j, & \frac{\lambda}{2} - \sum_{j=1}^{n} x_{ij}y_j < 0 \\ 0, & \text{otherwise} \end{cases}$$

The <u>elastic net</u> or <u>garotte</u> method mixes the above two:

$$\mathcal{L}(\underline{\beta}) = \|\underline{Y} - \underline{\underline{X}}\,\underline{\beta}\|_2^2 + \lambda\|\underline{\beta}\|_2^2 + \mu\|\underline{\beta}\|_1.$$