305a

June 26, 2025

1 Basic testing

1. Paired t-test Observations A_i, B_i paired, then $Z_i = A_i - B_i \overset{iid}{\sim} N(\mu, \sigma^2)$, test $\mathcal{H}_0 : \mu = 0$.

$$t = \frac{\overline{z}}{s_z/\sqrt{n}} \sim t_{n-1} \quad \text{under null} \quad \text{where} \quad s_z^2 = \frac{1}{n-1}\sum_{i=1}^{n}(z_i - \overline{z})^2.$$

2. Non parametric test, eg sign test, permutation test

$$\mathcal{H}_0 \text{ is symmetric about } 0.$$

3. Unpaired

$$t = \frac{\overline{x} - \overline{y}}{s\sqrt{1/n_1 + 1/n_2}} \quad \text{where } s^2 = \frac{1}{n_1 + n_2 - 2}\sum_{i=1}^{n}(x_i - \overline{x})^2 + (y_i - \overline{y})^2$$

2 Least Squares Basics

$$\begin{aligned}
\hat{\beta} &= \arg\min \|y - X\beta\|^2 && (1)\\
\iff X\hat{\beta} &= \mathrm{Proj}_{\mathrm{range}(X)}(y) && (2)\\
\iff y - X\hat{\beta} &\perp \mathrm{range}(X) && (3)\\
\iff \langle Xv, y - \hat{y}\rangle &= 0 \quad \forall v && (4)\\
\iff X^T(y - \hat{y}) &= 0 && (5)\\
\iff X^T X\hat{\beta} &= X^T y. && (6)
\end{aligned}$$

If X is not full rank, then $\hat{\beta}$ is not necessarily unique, but $\hat{y}$ is.
Unbiased:

$$\mathbf{E}\hat{\beta} = \mathbf{E}(X^T X)^{-1}X^T y = \beta$$

Covariance:

$$\mathbf{Cov}\hat{\beta} = \sigma^2(X^T X)^{-1}X^T X(X^T X)^{-1} = \sigma^2(X^T X)^{-1}$$

2.1 Hat Matrix

Facts:

1. range X = range H
2. range$(I - H)$ = range $H^{\perp}$ = ker H (projection onto orthogonal complement)
3. $H(I - H) = 0$ (orthogonality)
4. $X \perp I - H$, ie $X^T(I - H) = (I - H)X = 0$

3 OLS Residuals and Canonical Change of Basis

Uses:

1. Estimate σ
2. Assess adequacy of model (homoskedasticity, distribution)

$$\hat{\sigma}^2 = \frac{\|y - \hat{y}\|^2}{n - 2},$$

1. Columns of X are orthogonal to residuals
2. predictions vector is orthogonal to the residuals.
3. If intercept, then residuals sum to 0
4.

3.1 Canonical Change of Basis

4 Distributions and testing

Fact 1 (Independence of $\hat{\beta}$ and $\hat{\sigma}^2$)
Independence of $\hat{\beta}$ and $\hat{\sigma}^2$:

Proof.

$$\begin{pmatrix} \hat{\beta} \\ Y - \hat{Y} \end{pmatrix} = \begin{pmatrix} (X^T X)^{-1} X^T \\ I_n - H \end{pmatrix} Y \sim \mathcal{N}((\beta, 0)^T, \begin{pmatrix} \sigma^2 (X^T X)^{-1} & 0 \\ 0 & \sigma^2 (I - H) \end{pmatrix})$$

$\square$

Fact 2 (Distribution of $\hat{\sigma}^2$)

$$\hat{\sigma}^2 = \frac{1}{n}\|y - \hat{y}\|^2 \sim \frac{\sigma^2}{n}\chi^2_{n-p}.$$

Related to canonical change of basis.

Fact 3 (F-Test)
F-test. Let full model have $p = p_1 + p_2$ predictors and reduced model have p_1 predictors (Eg, $p_1 = 1$, $p_2 = p - 1$ in the case when reduced is just intercept).

$$\frac{\|\hat{y} - \hat{y}_{reduced}\|^2/p_2}{RSS/(n - p_1 - p_2)} \sim F_{p_2, n-p_1-p_2}$$

Note that $F_{1,n-p} = t^2_{n-p}$ and equivalent to below

Fact 4 (T-test)

$$\frac{\hat{\beta}_i - \beta_i}{\widehat{SE}(\hat{\beta}_i)} \sim t_{n-p}, \quad \text{where use estimate } \hat{\sigma}^2 = \frac{1}{n-p}\|y - \hat{y}\|^2,$$

so in testing $\beta_j = 0$:

$$\frac{\hat{\beta}_j}{\sqrt{[(X^T X)^{-1}]_{jj} RSS/(n-p)}}.$$

More generally, for testing $\mathcal{H}_0 : v^T \beta = 0$:

$$\frac{v^T \hat{\beta}_j}{\sqrt{[v^T (X^T X)^{-1}] v RSS/(n-p)}} \sim t_{n-p}.$$

Definition 5 (R^2)

$$R^2 = 1 - \frac{\|y - \hat{y}\|^2}{\|y - \overline{y}\|^2}.$$

5 Singular X

If rank$(X) < p$, then $\hat{y}$ is unique but $\hat{\beta}$ **is not**- take a vector in the null space of X and add to $\hat{\beta}$.
Ways to cope:

1. Restrict to **estimable** functions of β
2. Introduce **side-conditions** on β
3. Reparametrize (equivalent to 1

5.1 Estimability

Eg if neither $\hat{\beta}_1$ nor $\hat{\beta}_2$ are uniquely determined, but $\hat{\beta}_2 - \hat{\beta}_1$ **is**. Q: For $c \in \mathbb{R}^p$, for which c is $c\hat{\beta}$ uniquely determined?
Want:

$$1) X\beta' = X\beta'' \implies c\beta' = c\beta''$$

equivalently, find c s.t.

$$X\beta = 0 \implies c\beta = 0$$

Definition 6 (Estimability)
$c\beta$ **estimable** if any of the following:

1. $c\hat{\beta}$ uniquely determined by $\hat{y} = X\hat{\beta}$ (even if $\hat{\beta}$ not unique
2. $X\beta = 0 \implies c\beta = 0$
3. $c \in \text{row}(X)$
4. $\exists\, a \in \mathbb{R}^n$ such that $a^T X = c$
5. $\exists$ linear unbiased estimate of $c\beta$, ie $\exists a \in \mathbb{R}^n$ such that $\mathbf{E}_\beta(a^T y) = c\beta$

Theorem 7 (Gauss Markov)
Every estimable $c\beta$ has a **unique**, unbiased, linear estimate which has minimum variance within this class. The estimate is $c\hat{\beta}$, where $\hat{\beta}$ is any OLS estimate.
(Assumes homoskedastic noise)

Proof. (Sketch). Fix c estimable. Exists some $a \in \mathbb{R}^n$ such that $\mathbf{E}a^T y = c\beta$. Then write $a = a^* + (a - a^*)$ where $a^* = \text{Proj}_{\text{range}\, X}(a)$. Then $(a^*)^T X = a^T X = c$, show minimum variance:

$$\mathbf{Var}(a^T y) = a^T \mathbf{Cov}\, y a = \sigma^2 \|a\|^2 = \mathbf{Var}[(a^*)^T y] + \sigma^2 \|a - a^*\|^2$$

then show $(a^*)^T y = c\beta$ □

Fact 8 (Gauss Markov Assumptions)
Gauss Markov requires no distributional assumptions, just first and second moments of errors.

5.2 Side Conditions

We don't want to just remove one of the features even if its linearly dependent, because then would change interpretation; eg, treatment effects. Instead of imposing $\beta_1 = 0$ (removing a feature), impose something like $\beta_1 + \beta_2 + \beta_3 = 0$.

Fact 9 (Estimability and Side Conditions)
Side conditions must be in terms of ***non***-estimable functions; ie, constrain the thing we can't estimate uniquely

Let $H \in \mathbb{R}^{s \times p}$ set of side conditions, ie require $H\beta = 0$.

Theorem 10 (Side Conditions, ie when $\exists$ unique $\hat{\beta}$ satisfying conditions)
$X\hat{\beta} = \hat{y}$, $H\beta = 0$ has exactly one solution for any $\hat{y} \in \text{range}\, X$ iff:

1. $\text{row}\, H \bigcap \text{row}\, X = \emptyset$ (side conditions not in row space, ie $\vec{h}_i\beta$ is not estimable
2. $\text{row}\, H \oplus \text{row}\, X = \mathbb{R}^p$, ie rank $\begin{pmatrix} X \\ H \end{pmatrix} = p$

Ie: enough conditions to span the space (part 2) , but not too many (part 1), so that we can still solve $\hat{y} = X\hat{\beta}$

Corollary 11 (Side conditions make components of β estimable)
If we satisfy the above, then uniqueness from the side conditions β^h has estimable components, where $\beta = \beta^h + \beta^x$ where $\beta^h \in (\text{row } H)^{\perp}$, $\beta^x \in (\text{row } X)^{\perp}$.
??

One idea for incorporating constraints is let C_0 be constraints such that $C_0\beta = 0$, with $C_0 \in \mathbb{R}^{s\times p}$ then let

$$C = \begin{pmatrix} C_0 \\ C_1 \end{pmatrix},$$

so that $C \in \mathbb{R}^{p\times p}$ full rank. Write

$$X\beta = XC^{-1}C\beta = X_1^*\beta_1^* \quad \text{where } \beta_1^* = C_1\beta \in \mathbb{R}^s$$

We can **transform** a rank-deficient $X \in \mathbb{R}^{n\times p}$ to a full rank $X_1^* \in \mathbb{R}^{n\times s}$, so that all the components of β_1^* are estimable.

6 Least Squares Computations

First assume X full rank. Find a Q orthogonal ($\{x_1, \ldots, x_p\} \mapsto \{q_1, \ldots, q_p\}$ ONB, then form Q based on completing the ONB of $\mathbb{R}^n$). such that:

$$Q^TX = R = \begin{pmatrix} \tilde{R}_{p\times p} \\ 0_{n-p\times p} \end{pmatrix}$$

Then with $Q^Ty = y^*$,

$$\|y - X\beta\|^2 = \|Q^Ty - Q^TX\beta\|^2 = \|y_1^* - \tilde{R}\beta\|^2 + \|y_2^*\|^2.$$

Since X full rank, then $\tilde{R}$ is surjective, so can make the first term 0.
Ie, $X = QR$!
If X **not** full-rank, write $X = QRS^T$. Q takes the p columns and finds ONB. Then S^T takes the resultant rows and finds a r dimensional ONB. Yields:

$$\|y - X\beta\|^2 = \|Q^Ty - RS^T\beta\|^2$$

Let $Q^Ty = \begin{pmatrix} y_1^* \in \mathbb{R}^r \\ y_2^* \in \mathbb{R}^{n-r} \end{pmatrix}$

$$\iff \|y - X\beta\|^2 = \|y_1^* - \tilde{R}\beta_1^*\|^2 + \|y_2^*\|^2$$

Again make first term 0 since rank $\tilde{R} = r$.

Fact 12 (Why QR?)
QR decomposition is useful for the above, since if we can compute it efficiently, it's easy to solve an upper triangular system and we don't have to instantiate X^TX.

6.1 Householder Transforms

Definition 13 (Householder Transforms (HHT))
Any matrix $Q = I - uu^T$ with $\|u\|_2 = 1$ is a HHT.

Fact 14 (Properties of HHT)
Some properties:

1. Symmetric
2. Orthogonal
3. u is eigen vector with evalue -1
4. All elements of $\{u\}^{\perp}$ are e-vectors with evalue 1 (ie, invariant subspace)

Fact 15 (Existence of HHT s.t $a \mapsto b$)
For any pair of vectors a, b of same length, $\exists$ HHT that transforms $a \longrightarrow b$. Namely,

$$u = \frac{b - a}{\|b - a\|}$$

Our goal here is to transform X via orthogonal matrix to get upper triangular R.

Fact 16 (QR Decomposition via HHT)
Recipe:

1. If necessary, permute the columns of X st first $r = \operatorname{rank} X$ are linearly independent (Permutation matrices are orthogonal)
2. Let Q_1 be HHT that takes $x_1 \mapsto \|x_1\| e_1$
3. Then $Q_1 X$ has first column all 0's except first entry.
4. Repeat for the submatrix that is not yet upper diagonal.
5. $QX = Q_p \dots Q_1 X = R$

6.2 Given's Rotation

6.3 Gram-Schmidt

Fact 17 (Orthogonal Predictors in OLS)
If predictors are orthogonal:

$$\hat{\beta}_j = \frac{\langle y, x_j \rangle}{\|x_j\|^2},$$

since $X^T X$ is diagonal.
Note this is analogous to if we just have a single predictor x and do regression through the origin. Orthogonal predictors lets us just do regression separately for each predictor.
Further, if Q are orthonormal predictors, then $\hat{\beta} = Q^T y$.

The idea here is to convert predictors into orthogonal predictors, solve easy OLS, then convert back.
GS: $q_i = x_i - \sum_{k=1}^{i-1} \frac{\langle x_i, q_k \rangle}{\|q_k\|^2} q_k$, then $e_i = q_i / \|q_i\|$. In matrix form:

$$X = \tilde{Q}_{n\times p}\tilde{R}_{p\times p},$$

could complete the basis to get a full QR decomp.

Fact 18 (Gram Schmidt/OLS connection)
If we first calculate coefficients $\hat{\beta}^*$ of y on $\{q_j\}_{j\leqslant p}$, then

$$\hat{\beta}_{OLS} = \tilde{R}\hat{\beta}^* = \tilde{R}Q^T y.$$

The idea is we can solve the easy problem in the orthogonalized coordinates, then **convert back using the upper diagonal matrix from the Gram-Schmidt process**.

Nice trick is since $\tilde{R}$ has 1 on diagonal, then for the *last* OLS coefficient,

$$\hat{\beta}_p = \hat{\beta}^*_p = \frac{y^T q_p}{\|q_p\|^2} \quad \text{where } q_p = x_p - \sum_{i<p} \frac{\langle x_p, q_k\rangle}{\|q_k\|^2} = x_p - \text{Proj}_{\text{span}\, x_{(-p)}}(x_p).$$

In general, the coefficient for each x_j is the coefficient in a simple regression of y on x_j, but then adjusted for x_{-j}: we can always just reorder and get the same solution.

6.4 Modified Gram Schmidt

Better numerical stability than regular GS.

6.5 SVD

$$X = U_{n\times n}D_{n\times p}V^T_{p\times p} = U_{n\times r}D_{r\times r}V^T_{r\times p}$$

Example 19 (Given $\{x_i\}_{i\leqslant n}$, find best fitting hyperplane)

$$\min_{\alpha_0,\gamma_i,V\,:\,V^TV=I} \sum_{i=1}^n \|x_i - (\alpha_0 + V\gamma_i)\|^2$$

We can just solve by the 305c style PCA low rank approx

$$\|X_{n\times p} - \Gamma_{n\times p}V^T_{p\times p}\|_F^2$$

where $\Gamma = \left(\gamma_1, \dots, \gamma_n\right)^T$

Example 20 (Errors in Variables Regression with SVD)
Model:

$$y_i = z_i^T\beta + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma_\epsilon^2)$$

but instead of z_i, we actually observe

$$x_{ij} = z_{ij} + e_{ij}, \quad e_{ij} \sim N(0, \sigma_E^2)$$

$$\check{} = \sum_{i=1}^n \frac{y_i - z_i^T\beta}{\sigma_\epsilon^2} + \sum_{j=1}^p\sum_{i=1}^n \frac{(x_{ij} - z_{ij})^2}{\sigma_E^2.}$$

If $\sigma_\epsilon^2 = \sigma_E^2$, then def $X^* = [y\ :\ X]$, $B = [\beta\ :\ I]$, then minimize $\|X^* - ZB\|_F^2$, ie just a low rank approximation.

6.6 Updating/Downdating LS Computations

Fact 21 (Woodbury Inversion)

$$(A + uv^T)^{-1} = A^{-1} - A^{-1}u(I + v^T A^{-1}u)^{-1}v^T A^{-1}$$

Suppose we've computed LS fit and want to **update** the fit using a new point. New $(X^T X)^{-1}$:

$$(X^T X + x_{n+1}x_{n+1}^T)^{-1}$$

Update $\hat{\beta}$ using Woodbury.

Can use the same trick for downdating (ie LOO fit); $(X^T X - x_i x_i^T)^{-1}$. Idea that $X^T y = X_{(-i)}^T y_{(-i)} + x_i y_i$.

7 Model Selection (ESL Ch7)

8 Regularization (Ridge, Lasso)

8.1 Ridge

Ridge objective:

$$\hat{\beta}_{ridge} = \arg\min_{\beta} \frac{1}{2}\|y - X\beta\|^2 + \lambda\|\beta\|_2^2 = (X^T X + \lambda I_p)^{-1} X^T y$$

Degrees of freedom:

$$df = \sum_{j=1}^{p} \frac{\sigma_j^2}{\sigma_j^2 + \lambda} \quad \text{where } \sigma_j \text{ are singular values of } X.$$

8.2 Lasso

Lasso objective:

$$\hat{\beta}_{lasso} = \arg\min_{\beta} \frac{1}{2}\|y - X\beta\|^2 + \lambda\|\beta\|_1$$