## • Violations

$Y = X\beta + \varepsilon,\quad r(X) = p,\quad \varepsilon \sim N_n(0, \sigma^2 I)$

• $\hat\beta = (X^TX)^{-1}X^TY.\quad Cov\,\hat\beta = \sigma^2(X^TX)^{-1}$

$\mathbb{E}\,Y = X\beta$ working model

$\mathbb{E}\,Y = X\beta + Z\gamma$ true model

$\Rightarrow \mathbb{E}\,\hat\beta = (X^TX)^{-1}X^T(X\beta + Z\gamma)$

$= \beta + (X^TX)^{-1}X^TZ\gamma$

unbiased if $X^TZ = 0$

$S^2 = \frac{1}{n-p} Y^TQY$

$\mathbb{E}\,S^2 = \frac{1}{n-p}(\mathbb{E}Y^T Q\, \mathbb{E}Y + Tr\,Q(\sigma^2 I))$

$= \frac{1}{n-p}[\gamma^TZ^TQZ\gamma + \sigma^2(n-p)]$

$= \sigma^2 + \underbrace{\frac{1}{n-p}\gamma^TZ^TQZ\gamma}_{} > 0$

$S^2$ overestimates $\sigma^2$ unless $QZ = 0$

Fitted Model?

$\hat Y = X\hat\beta$

$\mathbb{E}\hat Y = X\beta + X(X^TX)^{-1}X^TZ\gamma$

$= X\beta + P_X Z\gamma$

• Residuals $e = Y - \hat Y$

$(L = (X^TX)^{-1}X^TZ)$

$\mathbb{E}e = \mathbb{E}Y - \mathbb{E}X\hat\beta$

$= X\beta + Z\gamma - X(\beta + L\gamma)$

$= QZ\gamma \neq 0$

So residual plots do not band around the $y=0$ line. What about $Var\ e$?

$Var\ e = Var(QY)$

$= \sigma^2 Q$ no change

Verify as far as prediction at $x_0$

$\hat y = x_0^T\hat\beta$ under fitted model

True: $\hat y = x_0^T\hat\beta + z_0^T\hat\gamma$

$Var\ \hat y(x_0, z_0) \geq Var\ \hat y(x_0)$

Similar calculations for overfitting.

working model: $\mathbb{E}\,Y = X_1\beta_1 + X_2\beta_2$

true model: $\mathbb{E}\,Y = X_1\beta_1$

$\hat\beta = (X^TX)^{-1}X^TY$

$\mathbb{E}\,\hat\beta = (X^TX)^{-1}X^T(X_1\ X_2)\begin{pmatrix}\beta_1\\ 0\end{pmatrix}$

$\mathbb{E}\begin{pmatrix}\hat\beta_1\\ \hat\beta_2\end{pmatrix} = \begin{pmatrix}\beta_1\\ 0\end{pmatrix} \Rightarrow$

$\mathbb{E}\hat\beta_1 = \beta_1$ is still unbiased.

• Covariance is misspecified:

Assume $Cov\,\varepsilon = \sigma^2 I$

If $Cov\,\varepsilon = V\sigma^2$, then

$\hat\beta_w = (X^TV^{-1}X)^{-1}X^TV^{-1}Y.$

If $\mathbb{E}\,Y = X\beta$ holds

$Cov\,\hat\beta_w = \sigma^2(X^TV^{-1}X)^{-1}.$

Is $\hat\sigma^2 = \frac{1}{n-p}Y^TQY$ biased for $\sigma^2$ when $Cov\,\varepsilon = V\sigma^2$?

$\boxed{\text{claim}}$ If $\mu_1 \leq \mu_2 \leq \dots \leq \mu_n$ are eigvals of $V$.

$\frac{1}{n-p}\sum_{i=1}^{n-p}\mu_i \leq \mathbb{E}\frac{\hat\sigma^2}{\sigma^2} \leq \frac{1}{n-p}\sum_{i=n-p+1}^{n}\mu_i$

**Lemma** $H^T = H = H^2$ & $r(H) = r$. if $A^T = A$, then
$$\sum_{i=1}^{r} \lambda_i \le Tr(HA) \le \sum_{i=n-r+1}^{n} \lambda_i$$
where $\lambda_1 \le \lambda_2 \le \cdots \le \lambda_n$ are eig-vals of $A$.

Pf: $H^T = H = H^2 \Rightarrow \exists P$ s.t. $P^T H P = \begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix} = D_r$

$$Tr\, HA = Tr\, P D_r P^T A$$
$$= Tr\, (D_r^2 P^T A P)$$
$$= Tr\left[(PD_r)^T (APD_r)\right]$$
$$= Tr \begin{pmatrix} P_1^T \\ P_2^T \\ \vdots \\ P_r^T \\ 0 \end{pmatrix} A (P_1 \cdots P_r\, 0) = \sum_{i=1}^{r} P_i^T A P_i$$

- $PD_r = (P_1 \cdots P_n)(e_1 \cdots e_r\ 0)$

$$\Rightarrow Tr\, HA = \sum_{i=1}^{r} P_i^T A P_i, \quad \|P_i\|_2 = 1$$

Verify:
$$\sum_{i=1}^{r} \lambda_i \le Tr\, HA \le \sum_{i=n-r+1}^{n} \lambda_i$$

# LACK OF FITNESS

$H_0: EY = X\beta$ vs $H_a: EY \ne X\beta$

Assumptions:

$n_1$ obs on $y$ at $x_1^T = (x_{11}, x_{12}, \ldots, x_{1k})$

$n_2$ obs on $y$ at $x_2^T = (x_{21}, x_{22}, \ldots, x_{2k})$

$\vdots$

$n_g$ obs on $y$ at $x_g^T = (x_{g1}, \ldots, x_{gk})$

$n = \sum_{i=1}^{g} n_i$ (total # of obs)

Then:
$$Y = \begin{pmatrix} y_{11} \\ y_{21} \\ \vdots \\ y_{n_1 1} \\ y_{12} \\ \vdots \\ y_{n_2 2} \\ \vdots \\ y_{1g} \\ \vdots \\ y_{n_g g} \end{pmatrix}, \quad X = \begin{pmatrix} x_1^T \\ \vdots \\ x_1^T \\ x_2^T \\ \vdots \\ x_2^T \\ \vdots \\ x_g^T \\ \vdots \\ x_g^T \end{pmatrix} \begin{matrix} \left.\vphantom{\begin{matrix} a \\ a \\ a \end{matrix}}\right\} n_1 \\ \left.\vphantom{\begin{matrix} a \\ a \\ a \end{matrix}}\right\} n_2 \\ \vdots \\ \left.\vphantom{\begin{matrix} a \\ a \\ a \end{matrix}}\right\} n_g \end{matrix}$$

Reg $y$ on $x_1, x_2, \ldots, x_k$ &
$$SSE = Y^T Q Y.$$

Let
$$\mathcal{U} = \begin{pmatrix} I - \frac{11^T}{n_1} & & & 0 \\ & I - \frac{11^T}{n_2} & & \\ & & \ddots & \\ 0 & & & I - \frac{11^T}{n_g} \end{pmatrix}$$
(i.e. $\sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2$)

$$SSE = Y^T (Q - \mathcal{U} + \mathcal{U}) Y$$
$$= Y^T (Q - \mathcal{U}) Y + Y^T \mathcal{U} Y$$
$$= SSLoF + SSPE.$$

- $(Q - \mathcal{U})(Q - \mathcal{U}) = Q - \mathcal{U}Q + \mathcal{U} - Q\mathcal{U}$

why $X^T \mathcal{U} = 0$ ? (verify it)

$\Rightarrow Y^T (Q - \mathcal{U}) Y \perp\!\!\!\perp \underbrace{Y^T \mathcal{U} Y}_{Y^T \mathcal{U} Y}$

So $SSLOF = Y^T (Q - \mathcal{U}) Y \perp\!\!\!\perp SSPE$

& $\frac{SSLOF}{\sigma^2} \sim \chi^2_{g-k}$

$\frac{SSPE}{\sigma^2} \sim \chi^2_{n-g}$

& so test statistic for
$H_0: EY = X\beta$

is $F = \frac{SSLOF/(g-k)}{SSPE/(n-g)} \sim F_{g-k,\, n-g}$

$\underbrace{df\, SSE}_{n-k} = (g-k) + (n-g).$