Transformations on X's or Y to meet model assumptions

$$\mathbb{E}Y = X\beta = \begin{pmatrix} x_1^T\beta \\ \vdots \\ x_n^T\beta \end{pmatrix}$$

- $\mathbb{E}e = 0 \quad \mathrm{cov}\, Y = \sigma^2 I$
- Consider $\mathbb{E}Y = \alpha + \beta^T x$

  Transform x by $g(x)$

$$\mathbb{E}Y = \alpha + \beta g(x)$$

Consider simple transformations like

$$g_\lambda(x) = \begin{cases} \ln x, & \lambda = 0 \\ \dfrac{x^\lambda - 1}{\lambda}, & \lambda \neq 0 \end{cases}$$

- $\lim_{\lambda \to 0} \dfrac{x^\lambda - 1}{\lambda} = \ln x.$

In practice, run the regression with

$$\lambda = \{-2, -\tfrac{1}{2}, \tfrac{1}{3}, 0, \tfrac{1}{3}, \tfrac{1}{2}, 1, 2, -1\}$$

Look at SSE when we

- reg y on $g_\lambda(x)$ for each λ. Pick the λ with smallest SSE.

▲ Do same procedure on the response ($y > 0$) & ad-hoc.

▲ Alternatively, use variance stabilizing transformation

So that $V(y)$ is a constant

$$g(Y) = g(\underset{\uparrow}{\mathbb{E}Y}) + (Y - \mu)g'(\mu) + o_p(Y - \mu)$$

- $\mathrm{Var}\, g(Y) \approx \mathrm{Var}\, Y [g'(u)]^2 \equiv \sigma^2 (g'(u))^2$

$$\Rightarrow g'(\mu) = \frac{\sigma^2}{\sqrt{\mathrm{Var}\, Y}}.$$

$$g(\mu) = \int \frac{\sigma}{\sqrt{\mathrm{Var}\, Y}}\, d\mu$$

if Poisson: $\sqrt{Y} = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$

- In ANOVA, k groups may have σ_i^2, $i = 1, 2, \cdots, k$.

$H_0: \sigma_1 = \sigma_2 = \cdots = \sigma_k = \sigma$ vs.

$H: \sigma_i \neq \sigma_j, \quad i \neq j$

Let $\hat{S}_i^2$ be the unbiased estimate of σ_i^2, $i = 1, 2, \cdots, k$. The ratio

Hartley's test $\dfrac{\max_{1 \le i \le k} \hat{S}_i^2}{\min_{1 \le i \le k} \hat{S}_i^2}$

is close to 1 if H_0 holds, otherwise reject.

Levene's test

$Y_{11}, Y_{12}, \cdots, Y_{1n_1}$	$Z_{ij} = \lvert Y_{ij} - \bar{Y}_{i\cdot} \rvert$
$Y_{21}, Y_{22}, \cdots, Y_{2n_2}$	$\bar{Z}_{\cdot\cdot} = \frac{1}{n} \sum_i \sum_j Z_{ij}$
$\vdots \quad \vdots \quad \ddots \quad \vdots$	$\bar{Z}_{i\cdot} = \frac{1}{n_i} \sum_{j=1}^{n_i} Z_{ij}$
$Y_{k1}, Y_{k2}, \cdots, Y_{kn_k}$	

use

$$W = \frac{\sum_i n_i (\bar{Z}_{i\cdot} - \bar{Z}_{\cdot\cdot})^2 / (k-1)}{\sum_i \sum_j (Z_{ij} - \bar{Z}_{i\cdot})^2 / (n-k)}$$

If $\bar{Y}_{i\cdot}$ is replaced by $\tilde{Y}_{i\cdot}$: median in i^{th} group, then it is Brown-Forsythe test

to test H_0 if $W > F_{k-1, n-k}$

Collinearity

Reg Y on $X_1 \cdots X_p$

$$\hat{Y} = X\hat{\beta} = (X_1 \cdots X_p)\begin{pmatrix}\hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p\end{pmatrix}$$

$$Var(\hat{\beta}_k) = \frac{\sigma^2}{\| X_k^\perp \|_2^2} = \sigma^2 \| X_k - P(X_k / V_{(k)}) \|_2^{-2}$$

$V_{(k)} = \mathcal{L}(X_1, \cdots, X_p$ but without $X_k)$

If $X_k \approx$ linear comb. of the other X's, then $\hat{\beta}_k$ will be estimated unreliably.

How to detect collinearity?

$R_j^2 \equiv R^2$ obtained when reg. X_j on $X_{(j)}$.

$$VIF_j \equiv \frac{1}{1-R_j^2}$$ variance inflation factor

If VIF_j is large, say > 10, then X_j is almost linearly related with the rest of X's.

What to do?

① Omit the variable or variables with $VIF > 10$.

② $\mathbb{E}Y = X\beta$

$= XR^{-1}R\beta$, R non-singular

where R is from the QR decomp of X: Householder matrix:

$$X = QR \qquad Hx = \begin{pmatrix}\|x\|_2 \\ 0\end{pmatrix}$$

$$(X_1\ X_2\ \cdots X_p) = \underbrace{(w_1\ w_2 \cdots w_p)}_{Q}\underbrace{\begin{pmatrix}\square & \square & \square \\ 0 & \square & \square \\ 0 & 0 & \square\end{pmatrix}}_{R}$$

$w_i^T w_j = 0$ & $w_i^T w_i = 1$.

$w_1 = \frac{X_1}{\|X_1\|_2}$, $c_{12}w_1 + c_{22}w_2 = X_2$.

$\gamma \equiv R\beta$.

$$\hat{\gamma} = \widehat{R\beta} = (Q^TQ)^{-1}Q^TY = Q^TY = R^{-T}X^TY$$

$$SSReg = Y^T P_{C(Q)} Y = YQQ^TY$$

$$= \sum_{i=1}^{p} (q_i^T Y)^2 \qquad Q = [q_1 \cdots q_p]$$

HW: $\mathbb{E}Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3$

I take obs at $X = 1, 2, \cdots, 5$.

$$X = \begin{pmatrix}1 & 1 & 1 \\ 1 & 2 & 4 & 9 \\ 1 & 3 & 9 & 27 \\ 1 & 4 & 16 & 64 \\ 1 & 5 & 25 & 125\end{pmatrix} \qquad \beta = \begin{pmatrix}\beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3\end{pmatrix}$$

$$\mathbb{E}Y = X\beta = (XR^{-1})(R\beta)$$

$$= (w_1 \cdots w_4)\gamma$$

$$\Rightarrow 0 = \sum_{x=1}^{5} w_i(x) w_j(x)$$

$H_0: \beta_p = 0$ iff $H_0: \gamma_p = 0$.

Heteroscedasticity

(Cook's paper)

$Y_i = X_i^T\beta + \varepsilon_i$, $\mathbb{E}\varepsilon_i = 0$, $Var\,\varepsilon_i = \sigma_i^2$

$\mathbb{E}Y = X\beta$, $r(X) = p$

Assume σ_i^2 depends on covariate

$\sigma_i^2 = f(a\ g\text{-vector covariate } z)$ ↓ Can be one of the x's.

$= \exp\{z_i^T\lambda\}$ &

1st column of X is 1. Then $z = (1, z_1, \cdots, z_k)$.

If $\lambda = \lambda_0 = (1, 0, 0, \cdots, 0)$. Then testing

$H_0: \lambda = \lambda_0$, $\sigma_i^2 = e^{\lambda_0} = \text{const}$

$L(Y_i | \beta, \sigma_i^2) \to L = \prod_{i=1}^{n} L_i \to \left(\frac{\partial \log l}{\partial \beta}\right)\left(\frac{\partial \log l}{\partial \beta}\right)^T$