Transformations

purpose: to achieve linear relation but more important: normality of errors!

1. Linearizable models:
   a. \( y = \alpha x^\beta \rightarrow \log y = \log \alpha + \beta \log x \) (linear)
   b. \( y = \alpha e^{\beta x} \rightarrow \log y = \log \alpha + \beta x \) (linear)
   c. \( y = \alpha + \beta \log x \rightarrow \text{linear} \)
   d. \( y = \frac{x}{\alpha x - \beta} \rightarrow \frac{1}{y} = \alpha - \beta \frac{1}{x} \) (linear)
   e. \( y = \frac{e^{\alpha + \beta x}}{1 + e^{\alpha + \beta x}} \rightarrow \log \left( \frac{y}{1-y} \right) = \alpha + \beta x \) (linear)

2. Graphs:
   a. \( \beta > 1, \beta = 1, 0 < \beta < 1 \)
   b. \( \beta > 0, \beta < 0 \)
   c. \( \beta > 0, \beta < 0 \)

3. Note that however that if you fit
   \( \log y = \log \alpha + \beta x + \varepsilon \)
   then
   \( y = e^{\log \alpha + \beta x} \)
   and \( e^\varepsilon \) is log-normal, not normal, etc.
**Box-Cox Transform:**

Let $y_i^{(\lambda)} = \begin{cases} \frac{y_i^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \log y_i & \text{if } \lambda = 0 \end{cases}$, note

$$y_i^\lambda - 1 = \frac{y_i^\lambda - y_i^0}{\lambda} = e^{(0+\lambda)(\log y_i^\lambda - \log y_i^0)} - 0$$

So

$$\lim_{\lambda \to 0} \frac{y_i^\lambda - 1}{\lambda} = \frac{d}{dt} (e^{t \log y_i}) \bigg|_{t=0}$$

So with Box-Cox Transform, by Jacobian Method

$$\frac{1}{\sqrt{2 \pi \sigma^2}} \exp \left\{ -\frac{1}{2} (y_i^{(\lambda)} - \hat{\beta})'(y_i^{(\lambda)} - \hat{\beta}) \right\}$$

To find $\lambda$ that optimizes the log likelihood

$$\text{Likelihood has a maximum likelihood of } (w. v. t \hat{\beta}, \sigma^2)$$

$$\frac{-n}{2} \hat{\sigma}^2 = \frac{-n}{2} \left( \frac{1}{n} \left[ y^{(\lambda)} \right]'(I - x(x'x)^{-1}x')y^{(\lambda)} \right)$$

Then optimize $L(\lambda) = -\frac{1}{2} n \log(\text{RSS}(\lambda)) + (\lambda - 1) n \sum \log y_i$.
Overfitting / Underfitting:

**Underfitting:** suppose we fit $(x'x)^{-1} x' y$ but $\mathbb{E}(y) = x\beta + \epsilon$.

Then $\mathbb{E}(\hat{\beta}) = (x'x)^{-1} x' \mathbb{E}(y) = (x'x)^{-1} x' (x\beta + \epsilon) = \\
= \beta + \frac{(x'x)^{-1} x' \epsilon}{\text{bias}}$

If $\sigma^2$ known, $\text{Var}(\hat{\beta}) = \sigma^2 (x'x)^{-1} \rightarrow \text{ok}$

If $\sigma^2$ unknown, however, $\mathbb{E}(\hat{\sigma}^2) = \mathbb{E}\left(\frac{y' (I - x (x'x)^{-1} x') y}{n-p}\right) = \\
= \sigma^2 + \frac{1}{n-p} y' z' (I - P_x) z y$

$\mathbb{E}(\hat{y}) = P_x \mathbb{E}(y) = x\beta + P z y$

some information of $z$ still used

$\mathbb{E}(e) = \mathbb{E}(y) - \mathbb{E}(\hat{y}) = (I - P_x) z y$

structure shown in residual plot

$\text{Var}(e) = (I - P) \sigma^2$

unchanged

**Summary:**

underfitting is biased, 
biased for $\sigma^2$, has apparent structure in residuals.

Overfitting: let $X = (x_1; x_2)$ but $\mathbb{E}(y) = x_1 \beta_1$. Then

$\mathbb{E}(\hat{\beta}) = (x'x)^{-1} x' x_2 \beta_1 = (x'x)^{-1} x' (x_1; x_2) \begin{pmatrix} \beta_1 \\ 0 \end{pmatrix} = \begin{pmatrix} \beta_1 \\ 0 \end{pmatrix}$

also $\mathbb{E}(\hat{y}) = \mathbb{E}(x\hat{\beta}) = x \mathbb{E}(\hat{\beta}) = x_1 \beta_1$

however, note $(x'x)^{-1} = (x'x_1)^{-1} + \text{other page}$
\[
(X'X)^{-1} = \left( (x_i'x_i) + (x_i'x_i) x_i'x_j (x_j'x_j)^{-1} \right) x_j'x_j (x_i'x_i)^{-1}
\]

\[
= \left( x_i'x_i \right)^{-1} + \frac{LML'}{p.d.}
\]

Therefore, \(\text{Var}(\hat{\beta}_i) = \sigma^2 \left( \left( x_i'x_i \right)^{-1} + \frac{LML'}{p.d.} \right)\)

\[> \sigma^2 \left( x_i'x_i \right)^{-1} = \text{Var}(\hat{\beta}_i\text{ conducted})\]

However, \(\text{E}(\hat{\sigma}^2) = \text{E}\left( y'(I-P_{x_1x_2})y \right) / n-p = \frac{(n-p)\sigma^2 + \beta'X'(I-P_{x_1x_2})X_2\beta_1}{n-p} = \sigma^2\)

so \(\hat{\sigma}^2\) is unbiased

**Summary**

- overfitting has \(\hat{\beta}, \hat{\gamma}, \hat{\sigma}\) unbiased
- but \(\text{Var}(\beta)\) is inflated

**Variance Inflating Factor** (254)

Let \(y_i = \gamma_0 + \gamma_1 (x_{i1} - \bar{x}_1) + \ldots + \gamma_{p-1} (x_{ip-1} - \bar{x}_{p-1}) + \epsilon_i\)

Then \(\text{Var}(\hat{\gamma}_j) = \frac{\sigma^2}{1 - R_j^2}\) where \(R_j^2\) is \(R^2\) for \(j\)

\(X_j^* = c_0 + c_1 x_{j1} + \ldots + c_{j-1} x_{j-1} + c_j x_{j1} + \ldots + c_{p-1} x_{jp-1}\)

(also multicollinearity)

Rule of thumb: \(\text{VIF} > 5\) problem

Also spectral decomposition: \(X'X = (UDV')'UDV' = VD^2V'\)

Inverse: \(VD^{-2}V'\), also A.H.G A - T - B p.d. for \(H\) small
Heteroskedasticity: \[ \Sigma = \begin{pmatrix} \sigma_i^2 \\ \sigma_n^2 \end{pmatrix} \]

Then optimal estimator is \((X' \Sigma^{-1} X)^{-1} X' \Sigma^{-1} \)

However \(\Sigma\) is unknown.

Note if we use \((X'X)^{-1} X'y = \hat{\beta}\), then \(\mathbb{E}(\hat{\beta}) = \beta\) but \(\text{Var}(\hat{\beta}) = (X'X)^{-1} \Sigma X'X\hat{\beta}\)
(\(\hat{\beta}\) violates BLUE assumptions in this case)

Naively: check whether \(\hat{\sigma}^2(x)\) or \(\hat{\sigma}^2(y)\), use these as weights.

Frequently in GLM: Poisson: \(\text{Var}(Y) = \mu\)

Binomial: \(\text{Var}(Y) = \mu(1-\mu)\)

Time dependency: not in midterm!

Multicollinearity: see VIF.

Confidence Intervals

Bonferroni: \(\hat{\beta}_j \pm \frac{q_{1-\alpha/2}}{n-p} \sqrt{\text{det}(D_{jj})}\), where \(D_{jj} = (X'X)_{jj}\)

(Schaffe: since \((\hat{\alpha} \hat{\beta} - \hat{\alpha} \hat{\beta})' (A(X'X)'A)\)^{-1}(A\hat{\alpha} \hat{\beta} - A\hat{\alpha} \hat{\beta}) \sim F_{\text{rank}(A) \leq 2, n-p}\)

Then \(\hat{\beta} \pm \frac{q_{1-\alpha/2}}{n-p} \sqrt{\text{det}(D_{jj})}\)

Jointly for \(A=I, h = g\): \(\hat{\beta}_j \pm (F_{p, n-p, \alpha/2})^{1/2} \sqrt{(X'X)_{jj}}\)

Particularly for \(A=I, h = g\): \(\hat{\beta}_j \pm (F_{p, n-p, \alpha/2})^{1/2} \sqrt{(X'X)_{jj}}\)
Outliers:

Leverage points: check if $h_{ii} > \frac{2p}{n}$ implies leverage.

Outliers test: (10.53): $\frac{(n-p-1)e_i^2}{(1-h_i)\text{RSS} - e_i^2} \sim F_{1, n-p-1}$

Equivalent: $\frac{(n-p-1)r_i^2}{n-p-r_i^2}$

Note: $e_i = y_i - \hat{y}_i$

$r_i = \frac{e_i}{S(1-h_i)^{1/2}}$ (studentized, internally)

$t_i = \frac{e_i}{S(i)(1-h_i)^{1/2}}$ (externally studentized or jackknifed residual)

Cook's Distance: $D_i = \frac{e_i^2 h_{ii}}{pS^2(1-h_{ii})^2} = \frac{r_i^2}{p} \frac{h_{ii}}{1-h_{ii}}$

Seber & Lee's: $D_i > F_{p, n-p}^{0.10}$

Cook & Weisberg: $D_i > 1$

Professor Zhang: $D_i > \frac{4}{n-p}$
Variable Selection Algorithms

- parsimony principle
- overfitting

Use $\hat{\sigma}^2$ full model for $\sigma^2$ in general

- adjusted $R^2$: $\overline{R}^2 = 1 - \frac{(1-R^2) \cdot n}{n-p}$

- based on fit
- can be used for stepwise reg.
- meaningful only for normal data

- AIC:

$$AIC = -2 \log f(y, \hat{\Theta}(y)) + 2r$$

where $r = \dim(\Theta)$ → number of parameters

→ regular regression case: $AIC = n \log 2\pi + n \log \left(\frac{RSS}{n}\right) + n + 2(p+1)$

- based on fit (maximum likelihood)
- stepwise
- more general (doesn't require normality)

- BIC:

$$BIC = -2 \log f(y, \hat{\Theta}(y)) + r \log(n)$$

$r = \dim(\Theta)$

→ regression: $BIC = n \log 2\pi + n \log \left(\frac{RSS}{n}\right) + n + (p+1) \log n$

- favors parsimonious models more than AIC
- doesn't require normality
- based on fit
Mallow's $C_p$ 

\[
\hat{C}_p = \frac{\text{RSS}}{\hat{\sigma}^2} + 2p - n \Rightarrow \frac{\text{E}(\text{RSS})}{\hat{\sigma}^2} + \frac{\text{E}(\text{ME})}{\hat{\sigma}^2} = \frac{\text{E}(\text{RSS})}{\hat{\sigma}^2} + 2p - n
\]

\[
\text{E}(\hat{C}_p) \approx \frac{\text{E}(\text{RSS})}{\hat{\sigma}^2} + 2p - n = \frac{1}{n}(I-\hat{R})\mu + n - p + 2p - n
\]

\[
\approx p.
\]

4. Based on prediction

when first: (p 386)

prediction error: $P_E = E_{\hat{y}_0}(||\hat{y}_0 - x_{0}\hat{\beta}||^2)$

we want to predict new $m$ observations $\hat{y}_0$ with $x_0$

Cross-validation:

if there is prediction training data available, $\frac{1}{m} \sum_{i=1}^{m} (y_{0i} - x_{0i}'\hat{\beta})^2$ based on $(y_2, x_2), \ldots, (y_n, x_n)$.

often unavailable

leave one out: $CV(1) = \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i'\hat{\beta}(i))^2$

leave out many: $CV(d) = (n-d)^{-1} \sum_{D} (y_D - x_D\hat{\beta})(I-H_D)^{-2}(y-D\hat{\beta})$

- Based on prediction
- Ridiculously expensive in computational effort

\[
H_D = X_D(X'X)^{-1}X_D' \rightarrow \text{not idempotent?}
\]
Ridge Regression:

Let

\[ Q(\beta) = (y-x\beta)'(y-x\beta) + \lambda \|\beta\|^2 \]

\[ \hat{\beta} = y'y - 2y'x\beta + \beta'x'x\beta + \lambda \beta \]

\[ = y'y - 2\beta'x'y + \beta'(x'x + \lambda I)\beta \]

\[ \frac{\partial Q}{\partial \beta} = -2x'y + 2(x'x + \lambda I)\beta = 0 \]

\[ \Rightarrow \hat{\beta}_R = (x'x + \lambda I)^{-1}x'y \]

This is a Bayesian estimator where \(\lambda\) is a prior for \(\Sigma\).

Also usually \(x^*\) is scaled (scaled \(X\)).

\(CV\) estimate of \(\lambda\) is minimize of

\[ \frac{1}{n} \sum_{i=1}^{n} \frac{(y_i - x_i'\hat{\beta}(\lambda))^2}{\left(1 - x_i'A(\lambda)\right)^2} \]

where \(A = X(x'x + \lambda I)^{-1}x'\)

\(CV\) : \[ \frac{1}{n} \sum (y_i - x_i'\hat{\beta}(\lambda))^2 \]

\[ \left(1 - x_i'A(\lambda)\right)^2 \]

- shrinks variance of \(\hat{\beta}\)
- shrinks \(\hat{\beta}\) when other by \(\hat{\beta}(\lambda) = \frac{1}{1+\lambda} \hat{\beta}\) uniformly

LASSO:

\[ Q(\beta) = (y-x\beta)'(y-x\beta) + \lambda \|\beta\|_1 \]
Principal Components:

Sources of variation in $X'X$:

$$X'X = VDV'UDV' = VD^2V' \rightarrow \text{spectral decomposition}$$

$D^2 \rightarrow \text{eigenvalues}$

pick those which explain say 95% of variability

then $V_i'X$ is the linear combination of $X$ that best explains $(X'X)_i$. Let $Z_i = V_i'X$, then

- drawback: ignores $Y$
- nice: PC are orthogonal! particularly $\hat{\theta}_m = \frac{<z_m, y>}{<z_m, z_m>}$
- Partial Least Squares:

Scale $X$

let $\hat{y}^{(0)} = \bar{y} 1 \_\_ , \ x_j^{(0)} = x_j , \ j = 1, \ldots, p$

- for $m = 1, 2, \ldots, p$
  
  $\hat{z}_m = \frac{1}{p} \sum_{j=1}^{p} \hat{\theta}_{mj} x_j^{(m-1)} \ \text{when} \ \hat{\theta}_{mj} = \frac{<x_j^{(m-1)}, y>}{<z_m, z_m>}$
  
  $\hat{\theta}_m = \frac{<z_m, y>}{<z_m, z_m>}$
  
  $\hat{y}^{(m)} = \hat{y}^{(m-1)} + \hat{\theta}_m z_m$
  
  Gram-Schmidt step: $x_j^{(m)} = x_j^{(m-1)} - \frac{<z_m, x_j^{(m-1)}>}{<z_m, z_m>} z_m$
  
  $\hat{y}^{(m)}$, can recover $\hat{\beta}_{\text{PLS}}$ since $\hat{y}^{(m)} = X\hat{\beta}_{\text{PLS}}$
  
  similar to PCR but uses fit on $Y$ as weights.
Midterm:

Mallow’s $C_p$: \[ C_p(p) = \frac{RSS(p)}{S^2} + 2p - n \]

Akaike’s: \[ AIC(p) = n \log(2\pi) + n \log \left( \frac{RSS(p)}{n} \right) + n + 2(p+1) \]

Bayesian: \[ BIC(p) = n \log(2\pi) + n \log \left( \frac{RSS(p)}{n} \right) + n + \log(n)(p+1) \]

1. $S^2$ from lm 1234 \( \Rightarrow S^2 = (0.05095)^2 = 0.0026 \)

also \( TSS = \frac{RSS}{1-R^2} = \frac{0.0026 \times 14}{1-0.849} = 0.1894 \)

\[ R^2 = 1 - \frac{RSS}{TSS} \]

\[ RSS(p) = (1-R^2(p)) \times S^2 \]

also for any choice of model: \[ R^2(p) = 0.1894 \]

\[ \bar{R}^2 = 0.7996, \quad \bar{R}^2_{12,3} = 0.8086, \quad \bar{R}^2_{1234} = 0.7542 \]

\[ F = \frac{(RSS_{13} - RSS_{123})/8}{MSE_{1234}} = 1.625 \rightarrow p \text{ value } 0.23 \]
b) \[ \text{since VIF}(\beta_j) = \frac{1}{1-R_j^2} \]

\[ \text{note VIF}(\beta_2) = \frac{1}{1-0.7306} = 3.712 < 5 \]

all others smaller than this.

c) \[ \frac{2p}{n} = \frac{2 \times 5}{16} = 0.65 \]

\[ \text{hat}(x)_3 = 0.811 > 0.65 \text{ ~ issue} \]

\[ \text{internal std residual} = \frac{e_i}{0.05095 \sqrt{1-h_i}} \]

\[ r_3 = -2.102 \]

bad but could be worse.

Cook's distance: \[ D_3 = 1.8039 > \frac{4}{n-p} = \frac{4}{11} = 0.36 \]

\[ \rightarrow \text{abnormal.} \]

T/F:

- 1 F
- 2 T
- 3 T
- 4 T
- 5 F

\[ \begin{array}{cccc}
6 & F & \left[ \begin{array}{c} 11 \\ 12 \\ 13 \\ 14 \\ 15 \\ \end{array} \right] & F \\
7 & F & F (ans. key T) \\
8 & T & F \\
9 & T & F \\
10 & F & F \\
\end{array} \]