§3.3

HD52}

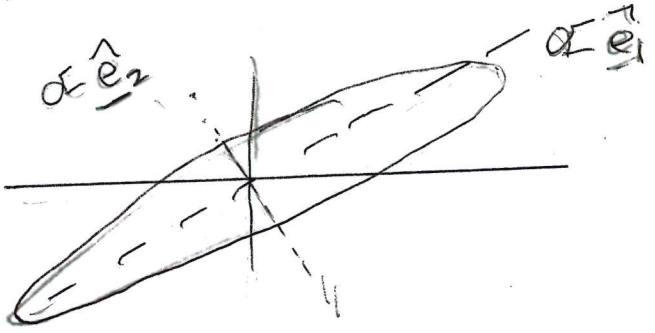
36} Let $D = \Sigma_{\underline{x}}$ or $\underbrace{\rho_{\underline{x}}}_{\text{correlation matrix}}$.

Let $(\hat{\lambda}_1, \hat{\underline{e}}_1), \ldots, (\hat{\lambda}_p, \hat{\underline{e}}_p)$ be the eigenvalue eigenvector pairs of $\hat{D}$ with $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p$. The $\hat{\underline{e}}_i$ are chosen to be orthonormal:

$\hat{\underline{e}}_i^T \hat{\underline{e}}_i = 1$ and $\hat{\underline{e}}_i^T \hat{\underline{e}}_j = 0$, $i \neq j$.

Then $\hat{\underline{m}}_j = \hat{\underline{e}}_j$ and $\underline{v}_i = \begin{pmatrix} \hat{\underline{e}}_1^T \underline{x}_i \\ \vdots \\ \hat{\underline{e}}_K^T \underline{x}_i \end{pmatrix} = \begin{pmatrix} v_{i1} \\ \vdots \\ v_{iK} \end{pmatrix}$.

This procedure is known as a principal components analysis, PCA a big mult analysis Math 585 topic. PCA gives a change of axes.



37} PCR regresses Y on the $\underline{v}_i$, $i = 1, \ldots, n$.

38] Problems: a) no reason why the 1st K $\hat{e}_1^T x_i, \ldots, \hat{e}_K^T x_i$ should be the most highly correlated with Y.

b) $\hat{\beta}_{K\text{PCR}} \neq \hat{\beta}_{OLS}$ unless $\hat{\beta}_{OLS}$ is in the span of $(\hat{e}_1, \ldots, \hat{e}_K)$ (which will occur if $K = p$).

39] Usually fit $\hat{\beta}_{1\text{PCR}}, \ldots, \hat{\beta}_{J\text{PCR}}$ where $J \leq \min(n-1, p)$. Then use a criterion like 10 fold CV to choose $\hat{\beta}_{MS,\text{PCR}} = \hat{\beta}_{K^*\text{PCR}}$. The resulting model is often linear and in low dimensions, $K^* = p$ is common.

40] Partial Least squares (PLS) can be computed in several equivalent ways. One way uses $\hat{w}_j = \hat{\Sigma}_x^{j-1} \hat{\Sigma}_{xy}$ for $j = 1, \ldots, p$.

Then $\hat{\beta}_{KPLS}$ regresses Y on $\hat{n}_1^T x, \ldots, \hat{n}_K^T x$. Hence $\hat{\beta}_{IPLS} = \hat{\beta}_{OPLS}$.

If $\beta_{IPLS} = \beta_{OPLS} = \beta_{OLS}$, it can be shown that $\beta_{KPLS} = \beta_{OLS}$ for $K = 1, \ldots, P$.

Also $\beta_{PPLS} = \beta_{OLS}$. $K = P$

41] A second equivalent way to compute β_{KPLS} is to roughly do a Gram-Schmidt orthogonalization of the $\hat{n}_i$ in 40].

Then $\hat{n}_1 = \dfrac{\hat{\Sigma}_{xy}}{\hat{\Sigma}_{xy}^T \hat{\Sigma}_{xy}}$, $\hat{n}_1^T \hat{n}_1 = 1$,

$\hat{n}_i^T \hat{n}_j = 0$, $i \neq j$.

42] It can be shown that (if $\hat{D} = \hat{\Sigma}_x$) the mth principal component direction $\hat{n}_m = \hat{e}_m$ solves

$$\max_{\|\alpha\| = 1} \operatorname{Var}(X\alpha)$$

$$\hat{n}_\ell^T \hat{\Sigma}_x \alpha = 0, \quad \ell = 1, \ldots, m-1$$

see plot on 36]

and the mth PLS direction $\hat{\underline{n}}_m$

solves $\max_{\|\underline{\alpha}\|=1} [\text{corr}(\underline{Y}, \underline{X}\underline{\alpha})]^2 \text{Var}(\underline{X}\underline{\alpha})$.

$\hat{\underline{n}}_\ell^T \hat{\Sigma}_X \underline{\alpha} = 0, \ell = 1, \ldots, m-1$

Hence expect $[\text{corr}(Y, \hat{\underline{n}}_1^T \underline{x})]^2 \geq [\text{corr}(Y, \hat{\underline{n}}_2^T \underline{x})]^2$

$\geq \ldots \geq [\text{corr}(Y, \hat{\underline{n}}_p^T \underline{x})]^2$.

(supervised learning)

Note that $\underline{X}_1 \hat{\underline{n}}_m = \begin{pmatrix} \underline{x}_1^T \hat{\underline{n}}_m \\ \vdots \\ \underline{x}_n^T \hat{\underline{n}}_m \end{pmatrix} = \underline{c}_m =$

mth component vector (closely related to the projection of $\underline{x}_i$ on $\hat{\underline{n}}_m$ since the $\hat{\underline{n}}_i$ are orthonormal).

43} There are $p+1$ K-component estimators

eg $\hat{\underline{\beta}}^E_{K PLS}$ $K = 1, 2, \ldots, p$ (or J)

and the model selection estimator

$\hat{\underline{\beta}}^E_{MS,PLS} = \underline{\beta}_{K^*,PLS}$ where $K^* \in \{1, \ldots, p\}$ is picked with model selection eg 10 fold CV.

44) $\underset{\sim}{Y} = \mathbb{X}\underset{\sim}{\beta} + \underset{\sim}{e}$

Let the nontrivial predictors $\underset{\sim}{U}_i = (x_{i2}, \ldots, x_{ip})^T$. Let the $n \times (p-1)$ matrix of standardized nontrivial predictors be $\mathbb{W} = (w_{ij})$

where $\sum_{i=1}^{n} w_{ij} = 0$ and $\sum_{i=1}^{n} w_{ij}^2 = n$

jth standardized predictor has $\bar{w}_j = 0$

So $\tilde{\sigma}_j^2 = \frac{1}{n}\sum_{i=1}^{n} w_{ij}^2 = 1$

method of moments sample variance (biased)

See HW 8.

Then the sample correlation matrix of the nontrivial predictors is

$R_{\underset{\sim}{U}} = \frac{\mathbb{W}^T\mathbb{W}}{n}$. Let $\underset{\sim}{Z} = \underset{\sim}{Y} - \bar{\underset{\sim}{Y}}$

where $\bar{\underset{\sim}{Y}} = \bar{Y}\underset{\sim}{1}$.

Many MLR methods fit

$\underset{\sim}{Z} = \underset{\sim}{W}\,\underset{\sim}{n} + \underset{\sim}{e}$ and then find $\hat{\underset{\sim}{\beta}}$ from $\hat{\underset{\sim}{n}}$.

use $\hat{\underset{\sim}{Y}} = \hat{\underset{\sim}{Z}} + \bar{\underset{\sim}{Y}}$.

45] procedure 44) is used so $\hat{\underset{\sim}{\beta}}$ is the same for anyone who uses the same units of measurement,

ex] $Y = ht$ $x_1 =$ constant, $x_2 = ht$ at shoulder, $x_3 = span$.

If we fit $\hat{\underset{\sim}{\beta}}$ in mm and then convert to ft, tend to get different answer it we fit $\hat{\underset{\sim}{\beta}}$ in ft.

point 44] is unitfree so the initial estimator uses the "same units"

ch8 MLR with Heterogeneity

1] MLR1: $\underset{\sim}{Y} = \mathcal{X}\underset{\sim}{\beta} + \underset{\sim}{e}$

e_i are independent $E(e_i) = 0$

$V(e_i) = \sigma_i^2$; $i = 1, \dots, n$.

MLR2: $Y = \alpha + \underset{\sim}{\beta}^T \underset{\sim}{x} + e$

$\underset{\sim}{Y} = \alpha \underset{\sim}{1} + \mathcal{X}_1 \underset{\sim}{\beta} + \underset{\sim}{e}$

2] weighted least squares (WLS) is often recommended, but there are too many parameters $\underset{p+1}{\underset{\sim}{\beta}}$, σ_i^2, $i = 1, \dots, n$.

3] OLS can be shown to be $\sqrt{n}$ consistent and asymptotically normal.

MLR1: $\sqrt{n}(\hat{\beta}_{OLS} - \beta) \xrightarrow{D} N_p(0, V \Omega V)$

where $\frac{X^T X}{n} \xrightarrow{P} V^{-1}$ and $\frac{1}{n}\sum_{i=1}^{n} E[e_i^2 x_i x_i^T] \xrightarrow{P} \Omega$.

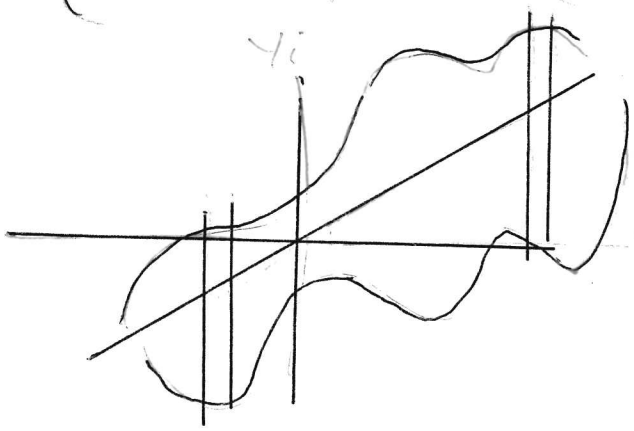
Under iid cases, $\Omega = E[e_i^2 x_i x_i^T]$.

4] $E(Y|w) = m(w)$

$V(Y|w) = v(w)$.

Hence $E(Y|\beta_{OLS}^T x) = \beta^T x$

$V(Y|\beta_{OLS}^T x) = v(\beta_{OLS}^T x)$ } can estimate from plot



$\hat{\beta}_{OLS}^T x_i = x_i^T \hat{\beta}_{OLS} = w_i$

5] Visualize $Y|\underset{\uparrow}{w}$ with a plot of w vs Y
RV

and narrow vertical slices.

6] Suppose $(x_i^T, \varepsilon_i)^T$ are iid

where $E\varepsilon_i = 0$ and $V(\varepsilon_i) = 1$.

Let $Y_i = \underline{x}_i^T \underline{\beta} + \sigma_i \varepsilon_i = \underline{x}_i^T \underline{\beta} + e_i$

where σ_i^2 is a function of $\underline{x}_i$.

Then $(Y_i, \underline{x}_i)$ are iid and earlier MLR OPLS theory still holds.

PhD topic

§4.2 Single index regression model.

1] we can always write (if $(Y, \underline{x})$ have a joint dist)

$$Y_i = E(Y_i | \underline{\eta}^T \underline{x}_i) + e_i$$

$$= m(\underline{\eta}^T \underline{x}_i) + e_i$$

where $e_i = Y_i - m(\underline{\eta}^T \underline{x}_i)$.

$Y_i = m(\underline{\eta}^T \underline{x}_i) + e_i$ is known as a single index model.

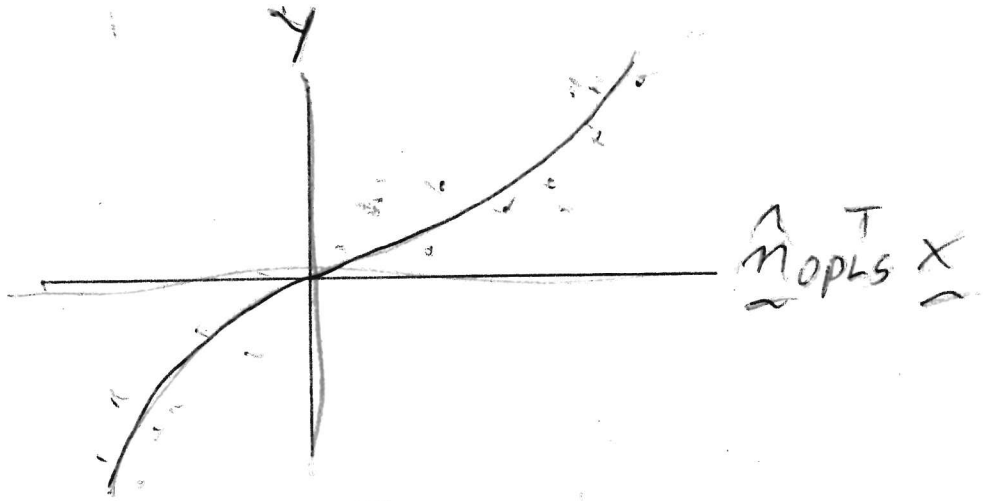
(A multiple index model uses $m(\underline{\eta}_1^T \underline{x}_i, \ldots, \underline{\eta}_K^T \underline{x}_i)$.)

2] If the cases $(Y_i, \underline{x}_i)$ are iid

let $Y_i = m(\underline{\eta}_{OPLS}^T \underline{x}_i) + e_i \quad i = 1, \ldots, n$

$E(e_i) = 0, \; V(e_i) = \sigma_i^2$,

3] Visualize m with a response plot



4) The $\hat{\underline{\Sigma}}_{\underline{x}Y} = \hat{\underline{\eta}}_{OPLS}$ theory still holds $\sqrt{n}(\hat{\underline{\Sigma}}_{\underline{x}Y} - \underline{\Sigma}_{\underline{x}Y}) \xrightarrow{D} N_p(\underline{0}, \underline{\Sigma}_{\underline{w}})$,

PhD topic

§4.4 Poisson Regression

1] $Y_i | \underline{\eta}^T \underline{x} \sim \text{Poisson}[\mu(\alpha + \underline{\eta}^T \underline{x})]$

where $E(Y|\underline{\eta}^T\underline{x}) = \mu(\alpha+\underline{\eta}^T\underline{x}) = m(\alpha+\underline{\eta}^T\underline{x})$

and $V(Y|\underline{\eta}^T\underline{x}) = \mu(\alpha+\underline{\eta}^T\underline{x}) = v(\alpha+\underline{\eta}^T\underline{x})$.

2] A Poisson regression (PR) generalized linear model (GLM) takes $\mu(\alpha+\underline{\beta}_{PR}^T\underline{x}) = \exp(\alpha+\underline{\beta}_{PR}^T\underline{x})$.

3] MMLE ← marginal estimator

Fit $Y \sim \text{Poisson}(e^{\alpha_j+\beta_j x_j})$

or $Y_k \sim \text{Poisson}(e^{\alpha+\beta_j x_{jk}})$ $k=1,\ldots,n$

to get $\hat{\beta}_j$ for $j=1,\ldots,p$.

Then $\underline{\hat{\beta}}_{MMLE} = \begin{pmatrix} \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{pmatrix}$.

4] MMLE variable selection:

standardize the predictors (then do 3] on the W_j) and take the J x_i corresponding to the largest $|\hat{\beta}_i|$.

5) Lasso variable selection for PR can be used in high dimensions.

No theory but if $\lambda = 0$, $\hat{\beta}_L = \hat{\beta}_{PR}$ if $n > p-5$.

Fit the lasso PR model $\hat{\beta}_L$,

then fit the PR (GLM) to the variables with $\hat{\beta}_i \neq 0$.

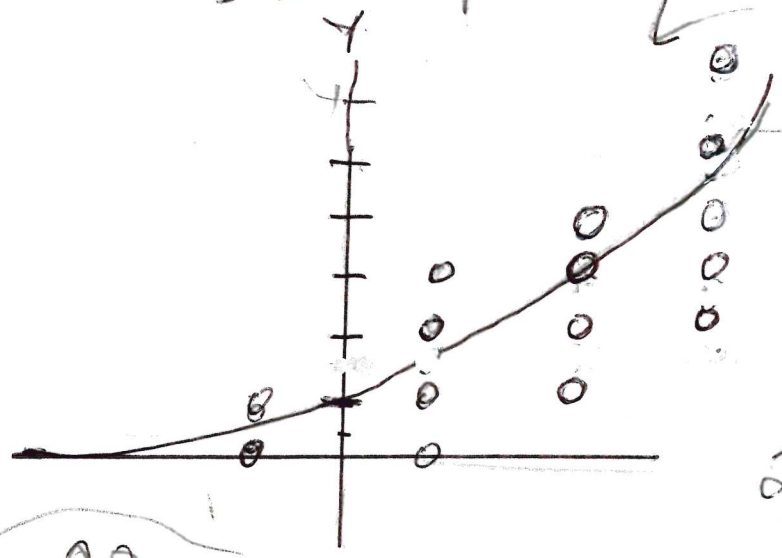
see HW8.

6) Let $z_i = \begin{cases} y_i & \text{if } y_i > 0 \\ 0.5 & \text{if } y_i = 0. \end{cases}$

The minimum chi-square estimator $(\hat{\alpha}_M, \hat{\beta}_M)$ is found from the WLS regression of $\log(z_i)$ on x_i with weights $w_i = z_i$. Equivalently, use the OLS regression (without intercept) of $\sqrt{z_i}\log(z_i)$ on $\sqrt{z_i}(1, x_i^T)^T$.

7) $(\hat{\alpha}_M, \hat{\beta}_M)$ and $(\hat{\alpha}_{PR}, \hat{\beta}_{PR})$ tend to be consistent estimators of (α, β).

if $Y | \underline{\beta}^T \underline{x} \sim$ Poisson $[\exp(\alpha + \underline{\beta}^T \underline{x})]$.



8] PR

$\log(Z) \approx \alpha + \underline{\beta}^T \underline{x} + e$

$\log(Y+1) \approx \alpha + \underline{\beta}^T \underline{x} + e$

} MLR with heterogeneity

If $(\log(Z_i), \underline{x}_i)$ are iid (eg $(Y_i, \underline{x}_i)$ iid)

or $(\log(Y_i+1), \underline{x}_i)$ are iid, then

$$\sqrt{n}\left(\underline{\hat{\beta}}_{\underline{x}(\log(U))} - \underline{\beta}_{\underline{x}(\log(U))} \right) \xrightarrow{D} N_p(\underline{0}, \Sigma_{\underline{w}})$$

where $U = Z$ or $Y+1$.

PhD topic (MMLE, ...)

9] Let $w_i = \underline{\hat{\eta}}^T \underline{x}_i$, $\underline{\hat{\eta}} = \underline{\hat{\beta}}_{MMLE}$ or $\underline{\hat{\beta}}_{\underline{x}(\log(U))}$.

Do PR of Y_i on w_i, get response plot.

10) PIs: Parametric bootstrap

$Y_i^* \overset{\text{bootstrap}}{\sim} \text{Poisson}(\hat{\alpha} + \hat{\underline{\beta}}^T \underline{x}_f)$, $i = 1, \ldots, B$.

$100(1-\delta)\%$ PI $\approx$ $100(1-\delta)\%$ shorth PI

applied to $Y_1^*, \ldots, Y_B^*$ eg $B = 100$ or 1000.

If $Y_i \sim D(\alpha + \underline{\beta}^T \underline{x}, \underline{\gamma})$

and the regression method gives

$\hat{\alpha}$, $\hat{\underline{\beta}}$, and $\hat{\underline{\gamma}}$, then use

$Y_i^* \sim D(\hat{\alpha} + \hat{\underline{\beta}}^T \underline{x}_f, \hat{\underline{\gamma}})$ for $i = 1, \ldots, B$.

(Poisson regression, binomial regression, weibull regression, etc.)

11) weibull regression

$(Y_i, \underline{x}_i)$ follow a weibull proportional hazards (PH) regression model iff $\log(Y_i) = \alpha + \underline{\beta}^T \underline{x}_i + \varepsilon_i$ follows an MLR model = weibull accelerated failure time

(AFT) model.

12] Do lasso for PH, then fit weibull PH model to the x_j with $\hat{\beta}_j \neq 0$ (lasso VS).

13] MMLE fit the p marginal models.

14] usually can't fit OLS to the AFT due to the y_i being censored. Need an estimator of Σ_{xy} when y are censored.

PhD topic

Binary regression and classification of 2 groups:

1] For binary regression, $y_i \in \{0, 1\}$. we may use $z_i = 2y_i - 1 \in \{-1, 1\}$.

$y \mid \underline{x}^T \underline{x} \sim \text{binomial}(m=1, g(\underline{x}^T \underline{x}))$

where $g(\underline{x}^T \underline{x}) = P(y=1 \mid \underline{x}^T \underline{x})$.

2] Multitude of models:

Since Y is binary, if $\begin{pmatrix} Y \\ \underline{w} \end{pmatrix}$ has a joint dist, then $Y|\underline{w} \sim bin(m=1, g(\underline{w}))$ and $Y|\underline{\eta}^T\underline{x} \sim bin[m=1, g(\underline{\eta}^T\underline{x})]$.

Hence if $\begin{pmatrix} Y \\ \underline{x} \end{pmatrix}$ has a joint dist, then

$Y|\underline{\eta}^T\underline{x}$ follows a binary regression model for every $\underset{\sim}{\eta} \in \mathbb{R}^p$.

3] Note that $E(Y|\underline{\eta}^T\underline{x}) = m\, g(\underline{\eta}^T\underline{x}) = g(\underline{\eta}^T\underline{x})$.

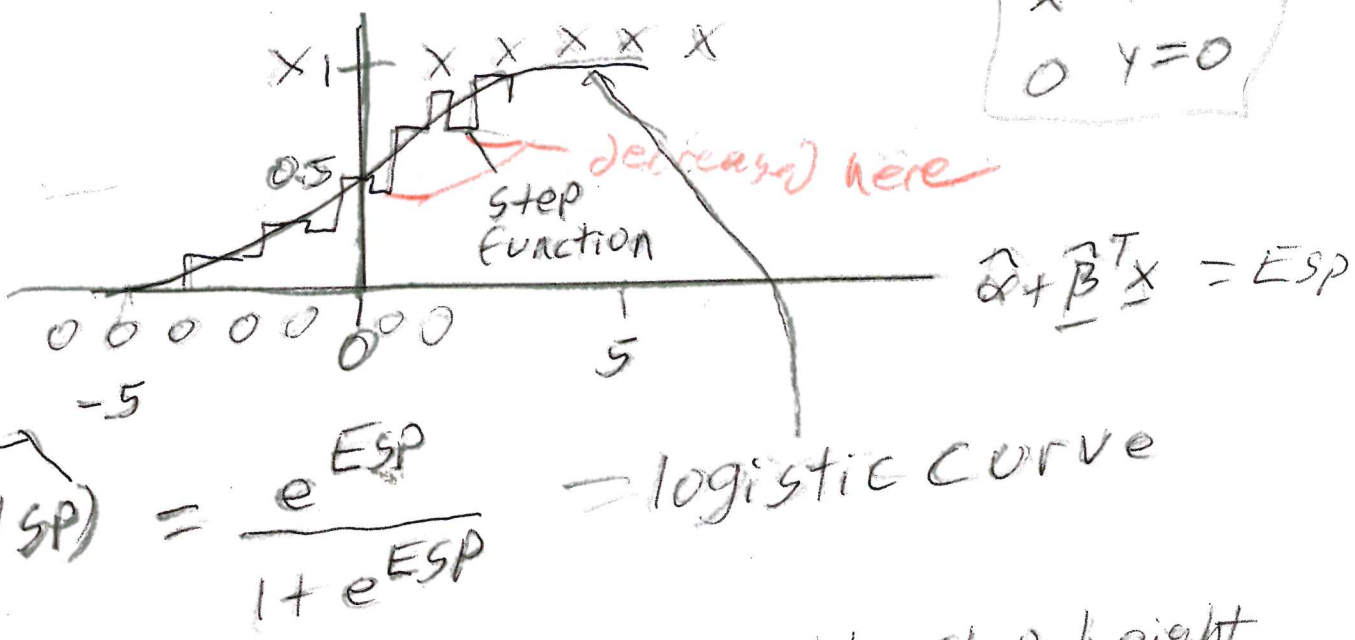
4] One binary regression model is the logistic regression LR model with

$$g(\underline{x}) = g(\alpha + \underline{\beta}^T\underline{x}) = \frac{\exp(\alpha+\underline{\beta}^T\underline{x})}{1+\exp(\alpha+\underline{\beta}^T\underline{x})} \in [0,1].$$

abuse of notation

5] LR response plot

$ESP = \hat{\alpha} + \hat{\beta}^T x$, $\hat{\beta} = \hat{\beta}_{LR}$.



$\widehat{E(Y|ESP)} = \frac{e^{ESP}}{1 + e^{ESP}}$ — logistic curve

Add a step function with step height $= \bar{Y}$ for the cases in the interval. Want the step function to track the logistic curve closely, but the step function often is not an increasing function.

6] Suppose there are 2 groups 0 and 1.

Let $s(\underline{w}) = P(\underline{W} \in \text{group 1}) = P(Y=1|\underline{w})$.

Let $g(sp) = \frac{e^{sp}}{1+e^{sp}}$ and let

$\hat{g}(Esp) = \frac{e^{Esp}}{1+e^{Esp}}$ be the LR

estimator of $g(sp)$. The LR discriminant rule or LR classification rule allocates $\underset{\sim}{w}$ to group

$$\begin{cases} 1 & \text{if } \hat{g}(\underset{\sim}{w}) \geq 0.5 \quad (Esp \geq 0) \\ 0 & \text{if } \hat{g}(\underset{\sim}{w}) < 0.5 \quad (Esp < 0) \end{cases}$$

where $\hat{g}(\underset{\sim}{w}) = \frac{\exp(\hat{\alpha} + \hat{\beta}^T \underset{\sim}{w})}{1+\exp(\hat{\alpha} + \hat{\beta}^T \underset{\sim}{w})}$.

Given $\underset{\sim}{w}$, the training data misclassification rate $\approx \min(\hat{g}, 1-\hat{g})$. $Esp > 2$ or $Esp < -2$ will have low error (misclassification) rate if the LR model is good. $Esp = 0$ $\approx$ 0.5 error rate (coin flip).

7] Lasso variable selection:

Fit lasso LR model $\hat{\beta}_L$.

Fit LR model to the J predictors with nonzero lasso $\hat{\beta}_i$.

8] MMLE:

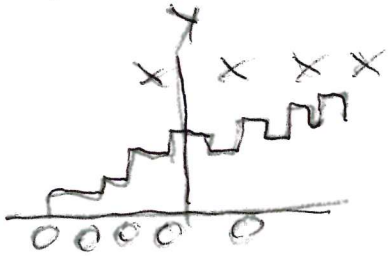
Fit the LR model that regresses Y on x_j to get $\hat{\beta}_j$, Then $\hat{\beta}_{MMLE} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)^T$.

9] MMLE variable selection:

Standardize the predictors (then do 5] on the x_j) and take the J x_j corresponding to the largest $|\hat{\beta}_j|$.

end exam 2 material

10] To visualize 9, use response plot



$\hat{\alpha} + \hat{\beta}^T x = Esp$

11 * Training data:

$\underset{\sim}{x}_{1j}, \ldots, \underset{\sim}{x}_{n_j j}$ group j. $(Y=j)$

use a discriminant or classification method to classify the training data (optimistic compared to test data: estimated error rate is too low).

If m_j of the n_j group j cases are correctly classified, the apparent error rate for group j is $1 - \frac{m_j}{n_j}$. Suppose there are G groups ($G = 2$ for binary). Let $m_A = \sum_{j=1}^{G} m_j$. Then the apparent error rate $\underline{AER} = 1 - \frac{m_A}{n}$

where $n = \sum_{j=1}^{G} n_j$,