

b) If $\frac{\hat{\lambda}_{in}}{\sqrt{n}} \xrightarrow{P} \gamma \geq 0$, then

$$\sqrt{n}(\hat{\underline{m}}_R - \underline{m}) \xrightarrow{D} N_{p-1}(-\gamma \nabla \underline{m}, \sigma^2 \underline{V}).$$

proof} By 32), $\hat{\underline{m}}_R =$

$$[\underline{I}_{p-1} - \hat{\lambda}_{in} (\underline{W}^T \underline{W} + \hat{\lambda}_{in} \underline{I}_{p-1})^{-1}] \hat{\underline{m}}_{OLS}, \text{ so}$$

$$\sqrt{n}(\hat{\underline{m}}_R - \underline{m}) = \sqrt{n}(\hat{\underline{m}}_R - \hat{\underline{m}}_{OLS} + \hat{\underline{m}}_{OLS} - \underline{m})$$

$$\sqrt{n}(\hat{\underline{m}}_{OLS} - \underline{m}) + \sqrt{n}(\hat{\underline{m}}_R - \hat{\underline{m}}_{OLS}) =$$

$$\sqrt{n}(\hat{\underline{m}}_{OLS} - \underline{m}) - \underbrace{\sqrt{n} \frac{\hat{\lambda}_{in}}{n}}_{\xrightarrow{P} \gamma} \underbrace{n (\underline{W}^T \underline{W} + \hat{\lambda}_{in} \underline{I}_{p-1})^{-1}}_{n/n=1} \hat{\underline{m}}_{OLS}$$

$$\xrightarrow{D} N_{p-1}(\underline{0}, \sigma^2 \underline{V}) - \gamma \nabla \underline{m} \sim N_{p-1}(-\gamma \nabla \underline{m}, \sigma^2 \underline{V})$$

□

$\sim$ if $\gamma = 0$

34) Lasso and elastic net need

Karush Kuhn Tucker KKT conditions,

which are like normal equations, 97.5

$$35) \quad \underline{y} = \underline{X}\underline{\beta} + \underline{e}, \quad \underline{z} = \underline{W}\underline{\eta} + \underline{e}. \quad \text{Let } \lambda_n \geq 0.$$

The lasso estimator minimizes
RSS = OLS criterion

$$Q_L(\underline{\eta}) = \frac{1}{2} \left(\underline{z} - \underline{W}\underline{\eta} \right)^T \left(\underline{z} - \underline{W}\underline{\eta} \right)$$

$$+ \frac{\lambda_n}{2} \underbrace{\sum_{i=1}^n |\eta_i|}_{\|\underline{\eta}\|_1}$$

for $\underline{\eta} \in \mathbb{R}^{p-1}$

$\|\underline{\eta}\|_1$ one norm or sum norm
or L_1 norm

The lasso estimator $\hat{\underline{\eta}}_L$ does
variable selection: some entries of
 $\hat{\underline{\eta}}_L$ tend to equal 0. From

$$\text{KKT: } -\frac{1}{n} \underline{W}^T (\underline{z} - \underline{W}\hat{\underline{\eta}}_L) + \frac{\lambda_n}{2n} \underline{s}_n = \underline{0}$$

$$\text{or } -\underline{W}^T (\underline{z} - \underline{W}\hat{\underline{\eta}}_L) + \frac{\lambda_n}{2} \underline{s}_n = \underline{0}$$

where $s_{in} \in [-1, 1]$ and $s_{in} = \text{sign}(\hat{\mu}_{iL})$ if

$\hat{\mu}_{iL} \neq 0$. Here $\text{sign}(\mu_i) = 1$ if $\mu_i > 0$
 -1 if $\mu_i < 0$.

$$\text{So } \hat{\underline{\mu}}_L = (\mathbf{W}^T \mathbf{W})^{-1} \mathbf{W}^T \underline{\mathbf{z}} - \frac{\lambda_{in}}{2n} n (\mathbf{W}^T \mathbf{W})^{-1} \underline{\mathbf{s}}_n$$

$$= \hat{\underline{\mu}}_{OLS} - \frac{\lambda_{in}}{2n} n (\mathbf{W}^T \mathbf{W})^{-1} \underline{\mathbf{s}}_n.$$

If $\frac{\lambda_{in}}{\sqrt{n}} \rightarrow 0$, OLS and lasso are asymptotically equivalent even if $\underline{\mathbf{s}}_n$ does not converge in prob to $\underline{\mathbf{s}}$.

36} Lasso CLT! Assume OLS CLT for $\underline{\mathbf{z}} = \mathbf{W} \underline{\boldsymbol{\mu}} + \underline{\mathbf{e}}$ holds.

a) If $\frac{\lambda_{in}}{\sqrt{n}} \xrightarrow{P} 0$, then

$$\sqrt{n} (\hat{\underline{\mu}}_L - \underline{\boldsymbol{\mu}}) \xrightarrow{D} N_{p-1}(\underline{\mathbf{0}}, \sigma^2 \mathbf{V}),$$

b) If $\frac{\hat{\alpha}_n}{\sqrt{n}} \xrightarrow{P} \tau \geq 0$ and $\underline{s}_n \xrightarrow{P} \underline{s}$,

98.5

then $\sqrt{n}(\hat{\eta}_L - \eta) \xrightarrow{D, P-1} N_{P-1}\left(-\frac{\tau}{2} \nabla \underline{s}, \sigma^2 \nabla\right)$.

Proof] If $\frac{\hat{\alpha}_n}{\sqrt{n}} \xrightarrow{P} \tau \geq 0$ and $\underline{s}_n \xrightarrow{P} \underline{s}$, then

$$\sqrt{n}(\hat{\eta}_L - \eta) = \sqrt{n}(\hat{\eta}_L - \hat{\eta}_{OLS} + \hat{\eta}_{OLS} - \eta)$$

$$= \sqrt{n}(\hat{\eta}_{OLS} - \eta) + \sqrt{n}(\hat{\eta}_L - \hat{\eta}_{OLS}) =$$

$$\sqrt{n}(\hat{\eta}_{OLS} - \eta) = \underbrace{\sqrt{n} \frac{\hat{\alpha}_n}{2n}}_{\xrightarrow{P} \tau/2} \cdot \underbrace{n (W^T W)^{-1}}_{\xrightarrow{P} \nabla} \underbrace{\underline{s}_n}_{\xrightarrow{P} \underline{s}}$$

$$\xrightarrow{D, P-1} N_{P-1}\left(0, \sigma^2 \nabla\right) - \frac{\tau}{2} \nabla \underline{s}$$

$$\sim N_{P-1}\left(-\frac{\tau}{2} \nabla \underline{s}, \sigma^2 \nabla\right).$$



37] The elastic net estimator (LS 99)

$\hat{\underline{\mu}}_{EN}$ minimizes $\underbrace{RSS_{\underline{w}}(\underline{\mu})}_{OLS \text{ criterion}} + \lambda_1 \|\underline{\mu}\|_2^2$

$+ \lambda_2 \|\underline{\mu}\|_1$ where $\lambda_1 = (1-\alpha) \lambda$ and

and $\lambda_2 = 2\alpha \lambda$.

KKT conditions:

$$2 \underline{W}^T \underline{W} \hat{\underline{\mu}}_{EN} - 2 \underline{W}^T \underline{z} + 2 \lambda_1 \hat{\underline{\mu}}_{EN} + \lambda_2 \underline{s}_n$$

$= \underline{0}$ or

$$(\underline{W}^T \underline{W} + \lambda_1 \underline{I}_{p-1}) \hat{\underline{\mu}}_{EN} = \underline{W}^T \underline{z} - \frac{\lambda_2}{2} \underline{s}_n$$

$$\text{or } \hat{\underline{\mu}}_{EN} = \hat{\underline{\mu}}_R - n (\underline{W}^T \underline{W} + \lambda_1 \underline{I}_{p-1})^{-1} \frac{\lambda_2}{2n} \underline{s}_n$$

$$= \hat{\underline{\mu}}_{OLS} - \frac{\lambda_1}{n} n (\underline{W}^T \underline{W} + \lambda_1 \underline{I}_{p-1})^{-1}$$

$$- \frac{\lambda_2}{n} n (\underline{W}^T \underline{W} + \lambda_1 \underline{I}_{p-1}) \underline{s}_n$$

$$= \hat{\eta}_{OLS} - n (W^T W + \lambda_1 I_{p-1})^{-1} \left[\frac{\lambda_1}{n} \hat{\eta}_{OLS} + \frac{\lambda_2}{2n} \underline{s}_n \right] \quad 99.5$$

If $\frac{\hat{\lambda}_1}{\sqrt{n}} \xrightarrow{P} \tau$ and $\hat{\alpha} \xrightarrow{P} \psi$, then

$$\frac{\hat{\lambda}_1}{\sqrt{n}} \xrightarrow{P} (1-\psi)\tau \text{ and } \frac{\hat{\lambda}_2}{\sqrt{n}} \xrightarrow{P} 2\psi\tau.$$

For the Th below, get RR CLT if $\psi=0$ and Lasso CLT (using $\frac{\hat{\lambda}_1}{\sqrt{n}} \xrightarrow{P} \tau$) if

$$\psi=1.$$

38) Elastic Net CLT: Assume OLS CLT holds for $\underline{z} = W\underline{y} + \underline{\epsilon}$.

a) If $\frac{\hat{\lambda}_1}{\sqrt{n}} \xrightarrow{P} 0$, then $\sqrt{n}(\hat{\eta}_{EN} - \underline{\eta}) \xrightarrow{D}$

$$N_{p-1}(\underline{0}, \sigma^2 V).$$

b) If $\frac{\hat{\lambda}_1}{\sqrt{n}} \xrightarrow{P} \tau \geq 0$, $\hat{\alpha} \xrightarrow{P} \psi \in [0,1]$ and $\underline{s}_n \xrightarrow{P} \underline{s}$, then

$$\sqrt{n}(\hat{\eta}_{EN} - \underline{\eta}) \xrightarrow{D} N_{p-1}(-V[(1-\psi)\tau\underline{\eta} + \psi\tau\underline{s}], \sigma^2 V).$$

LS 100

$$\text{Proof: } \sqrt{n} (\hat{m}_{EN} - m) = \sqrt{n} (\hat{m}_{EN} - \hat{m}_R + \hat{m}_R - m)$$

$$= \sqrt{n} (\hat{m}_R - m) + \sqrt{n} (\hat{m}_{EN} - \hat{m}_R) \frac{D}{1}$$

$$N_{p+1} \left[-(1-\phi) \gamma \sqrt{V} \underline{m}, \sigma^2 \sqrt{V} \right] - \frac{2\phi\gamma}{2} \sqrt{V} \underline{s}$$

$$\sim N_{p+1} \left(-\sqrt{V} \left[(1-\phi) \gamma \underline{m} + \phi \gamma \underline{s} \right], \sigma^2 \sqrt{V} \right)$$

$\underbrace{\hspace{10em}}_{\underline{Q} \text{ under } \underline{a} \text{ since } \underline{m} \text{ and } \underline{s} \text{ are bounded.}} \quad \square$

39} MLR nonconstant variance

$$y_i = \underline{x}_i^T \underline{\beta} + e_i$$

$$e_i \text{ i.i.d.}, E(e_i) = 0 \quad V(e_i) = \sigma_i^2$$

$$i=1, \dots, n \quad \text{Often } e_i = \sigma_i \cdot \varepsilon_i, \varepsilon_i \text{ i.i.d.} \\ E(\varepsilon_i) = 0, V(\varepsilon_i) = 1.$$

$$\underline{Y} = \underline{X} \underline{\beta} + \underline{e}$$

$$\text{If } \sigma_i^2 = \sigma^2 y_i^2, \text{ can use}$$

weighted least squares (WLS) 100.5

if there are consistent estimators for γ_i^2 , $i=1, \dots, n$. This result rarely holds, so WLS output does not give correct p-values.

exception: $z_j = x_j^T \beta + \varepsilon_j$, $V(\varepsilon_j) = \sigma^2$

and there are m_i z_j at each x_i

$m_i \geq 1$ and at least one $m_i > 1$,

Then replace the z_k with the same

x_i by $y_i = \bar{z}$ (or median z_k).

$V(y_i) = \frac{\sigma^2}{m_i}$. Then $\gamma_i = \frac{1}{m_i}$.

40) OLS gives a consistent estimator for β under 39).

$$\underline{Y} = \underline{X}\underline{\beta} + \underline{e} \quad \text{COV}(\underline{e}) = \text{diag}(\sigma_i^2)$$

$$\underline{\hat{\beta}} = (\underline{X}^T \underline{X})^{-1} \underline{X}^T \underline{Y} \quad \text{has}$$

$$\text{COV}(\underline{\hat{\beta}}) = (\underline{X}^T \underline{X})^{-1} \underline{X}^T \text{diag}(\sigma_i^2) \underline{X} (\underline{X}^T \underline{X})^{-1}$$

It can be shown that replacing

$$\text{diag}(\sigma_i^2) \text{ by } \text{diag}(r_i^2) = \begin{pmatrix} r_1^2 & & \\ & \ddots & \\ & & r_n^2 \end{pmatrix}$$

sometimes gives a consistent estimator of $\text{COV}(\underline{\hat{\beta}})$ (sandwich estimator),

The nonparametric bootstrap appears to work better than sandwich estimators,

Let $\underline{\Sigma_e} = \text{diag}(\sigma_i^2)$. Assume

$$\frac{\underline{X}^T \underline{X}}{n} \xrightarrow{P} W^{-1} \quad \text{and} \quad \frac{\underline{X}^T \underline{\Sigma_e} \underline{X}}{n} \xrightarrow{P} U,$$

Then $\sqrt{n}(\underline{\hat{\beta}}_{OLS} - \underline{\beta}) \xrightarrow{D} N_p(0, V)$ where

$$V = W U W,$$

4) Let $y \perp x \mid \beta_1^T x, \dots, \beta_d^T x$ 10/15

where d is as small as possible:

$$0 \leq d \leq p. \quad y \perp x \mid x = \underbrace{I x}_{p \text{ components}}$$

$d=0$ if $y \perp x$.

The PLS construction was made under the assumption that $d \leq 1$.

For the nonconstant variance MLR model,

$$d=1 \text{ if } \sigma_i^2 = \sigma_i^2(\underline{x}_i^T \underline{\beta}), \text{ but}$$

d could be as high as p if each

σ_i^2 depends on a different linear combination

$$y_i \mid x_i = \underline{\beta}^T \underline{x}_i + e_i, \quad E(e_i) = 0, \quad V(e_i) = \sigma_i^2$$

For the constant variance model

$$y_i \mid \underline{x}_i = \underline{\beta}^T \underline{x}_i + e_i = y_i \mid \underline{x}_i^T \underline{\beta}, \quad V(e_i) = \sigma^2$$

For the nonconstant variance model,

$$y_i \mid \underline{x}_i^T \underline{\beta} = \underline{x}_i^T \underline{\beta} + \varepsilon_i \quad \text{where } V(\varepsilon_i) = \gamma_i^2 = V(y_i \mid \underline{x}_i^T \underline{\beta})$$

but $E(Y_i | \underline{x}_i^T \underline{\beta}) = \underline{x}_i^T \underline{\beta} = m(\underline{x}_i^T \underline{\beta})$

and $V(Y_i | \underline{x}_i^T \underline{\beta}) = \tau_i^2 = \nu(\underline{x}_i^T \underline{\beta})$ depend on $\underline{x}_i$ only through $\underline{x}_i^T \underline{\beta}$. Hence

$Y_i | \underline{x}_i^T \underline{\beta}$ is a linear model that depends on $\underline{x}_i$ only through $\underline{x}_i^T \underline{\beta}$, and

the PLS construction

$$\underline{\hat{\beta}}_{PLS} = \underline{\hat{\beta}}_{OPLS} = \hat{\lambda} \hat{\Sigma}_{xy} \text{ should still}$$

hold. Hence $\hat{\eta} = \hat{\Sigma}_{xy}$ can still

be used to test $H_0 A \underline{\beta} = \underline{0}$ vs $H_A A \underline{\beta} \neq \underline{0}$

42} Expect $\hat{\eta}$ to be useful in high

dimensions. For low dimensions,

the nonparametric bootstrap with OLS

is one of the best methods. Expect

$\underline{\hat{\beta}}_{PLS} = \underline{\hat{\beta}}_{OPLS}$ with the nonparametric

bootstrap and $\hat{\eta}$ without the ^{102.5}
bootstrap may work better,

HW 11 showed $\sqrt{n}(\hat{\eta} - \eta) \xrightarrow{D} N_p(0, \Sigma_w)$

and $\sqrt{n}(\hat{\eta}^* - \hat{\eta}) \xrightarrow{D} N_p(0, \Sigma_w)$

when the nonparametric bootstrap is used.

43) OPLS estimators use

$\hat{\eta}$ and $\hat{\beta}_{\text{OPLS}} = \hat{\lambda} \hat{\eta}$ without

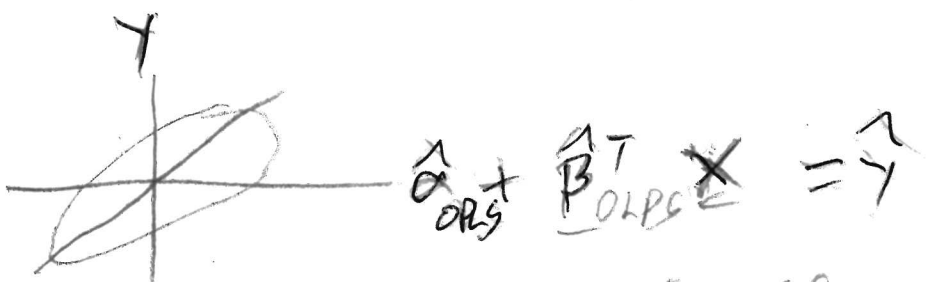
assuming the PLS construction
still have $(x_i^T, y_i)^T$ iid,

$\hat{\beta}_{\text{PPLS}}$

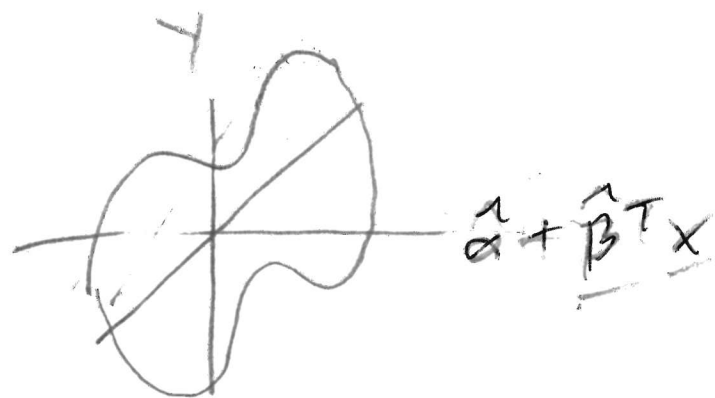
$\hat{\beta}_{\text{PPLS}} = \hat{\beta}_{\text{OPLS}}$

check the model graphically by

plotting $\hat{\eta}^T x$ vs y or $\hat{\beta}_{\text{OPLS}}^T x$ vs y



MLR constant variance



$$\begin{pmatrix} V(Y_i | X_i^T \hat{\beta}) \\ V(Y_i | X_i^T \hat{\beta}) \end{pmatrix}$$

MLR nonconstant variance.

44) Consider binary regression

$Y_i \in \{0, 1\}$, Predictors $x_1, \dots, x_p$

$(x_{i1}, \dots, x_{ip}, Y_i)^T = (\underline{x}_i^T, Y_i)^T$ are the cases,

Binary logistic regression (LR)

$Y_i | \underline{x}_i \sim \text{bin}(n=1, p(\underline{x}_i))$

$$p(\underline{x}_i) = \frac{\exp(\alpha + \underline{x}_i^T \beta)}{1 + \exp(\alpha + \underline{x}_i^T \beta)}$$

There are several other models
(probit, vector support machines).

$$z_i = 2 Y_i - 1$$

45} Let $\pi_j = P(Y=j)$, $j=0,1$

103.5

Let $\mu_j = E(X|Y=j)$, $j=0,1$. Let

$N_i = \# Y_s$ that are equal to i , $i=0,1$.

Then $\hat{\mu}_i = \frac{1}{N_i} \sum_{j: Y_j=i} x_j$ for $i=0,1$

$\hat{\pi}_i = \frac{N_i}{n}$, $\hat{\pi}_1 = 1 - \hat{\pi}_0$, $N_1 + N_2 = n$.

Hence $\hat{\mu}_i = \bar{x}_i$ is the sample mean of the x_k corresponding to $Y_k = i$, $i=0,1$.

Then $\tilde{x}_{xy} = \frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{x} \bar{y}$

$$= \frac{1}{n} \left[\sum_{j: Y_j=1} x_j (1) + \sum_{j: Y_j=0} x_j (0) \right] - \bar{x} \hat{\pi}_1 =$$

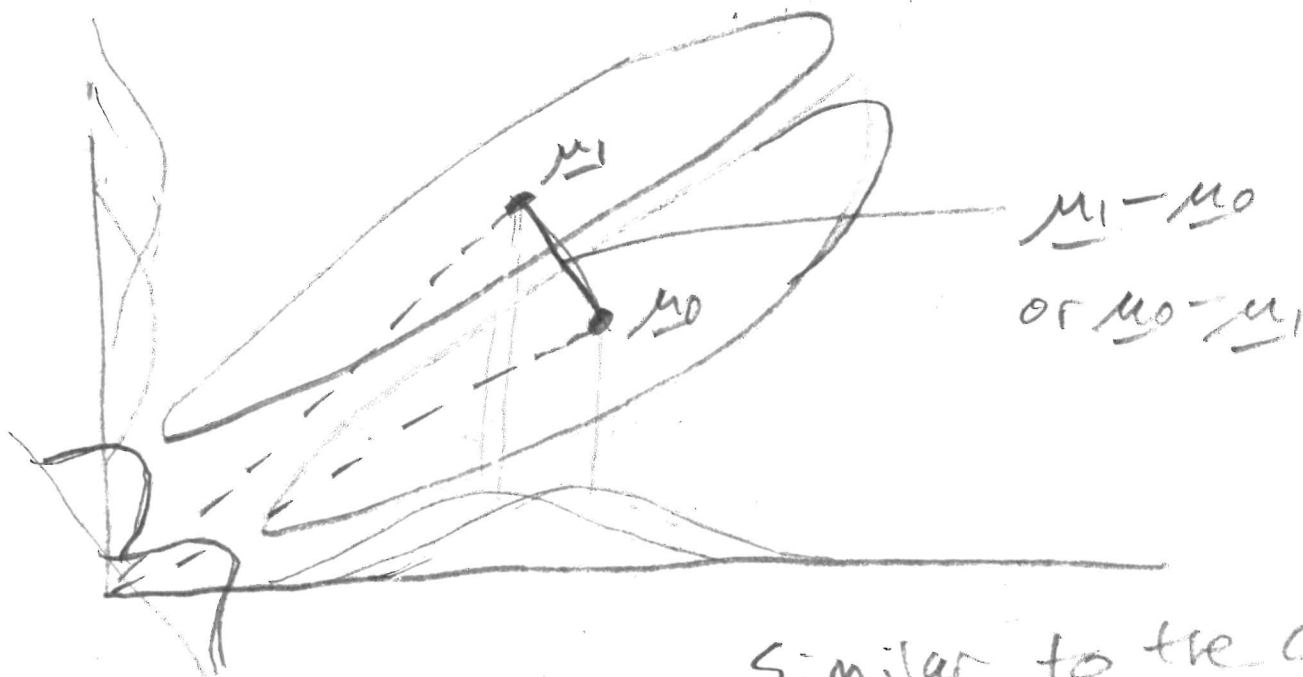
$$\frac{1}{n} (N_1 \hat{\mu}_1) - \frac{1}{n} [N_1 \hat{\mu}_1 + N_0 \hat{\mu}_0] \hat{\pi}_1$$

$$= \hat{\pi}_1 \hat{\mu}_1 - \hat{\pi}_1^2 \hat{\mu}_1 - \hat{\pi}_1 \hat{\pi}_0 \hat{\mu}_0$$

$$= \hat{\pi}_1 (1 - \hat{\pi}_1) \hat{\mu}_1 - \hat{\pi}_1 \hat{\pi}_0 \hat{\mu}_0 = \hat{\pi}_1 \hat{\pi}_0 (\hat{\mu}_1 - \hat{\mu}_0)$$

$$\text{Hence } \underline{\eta} = \text{cov}(X, Y) = \pi_1 \pi_0 (\underline{\mu}_1 - \underline{\mu}_0)$$

and $\underline{\phi} = \underline{\mu}_1 - \underline{\mu}_0$ are quantities of interest for binary regression. LS 104



Similar to the cover
of Dr Cook's Envelopes book

45} If $(\underline{x}_i^T, y_i)^T = \underline{u}_i$ are iid, tests
with $\hat{\underline{\eta}} = \hat{\underline{\beta}}_{xy}$ can be used just
like for MLR. constant variance
not constant variance.

If the $\underline{x}_{ij}$ are random samples (iid)
from pop's $j=0$ and $j=1$, then
tests using $\hat{\underline{\mu}}_1$ and $\hat{\underline{\mu}}_0$ can be used.

46} Let $\underline{z} = \lambda \underline{n} = \gamma \underline{\phi}$. Plug

$\hat{\underline{n}}^T \underline{x}$ or $\hat{\underline{\phi}}^T \underline{x}$ into a binary regression program like LR to get $\hat{\lambda}$ or $\hat{\gamma}$.

$$47} \underbrace{\hat{\underline{n}} (\hat{\underline{n}}^T \hat{\underline{n}})^{-1} \hat{\underline{n}}^T}_{\text{projection matrix on span}(\hat{\underline{n}})} \underline{x} = \frac{\hat{\underline{n}} \hat{\underline{n}}^T \underline{x}}{\hat{\underline{n}}^T \hat{\underline{n}}} \quad \text{projects}$$

$\underline{x}$ onto the span of $\hat{\underline{n}} = \text{span of } \underline{\mu}_1 - \underline{\mu}_0$.

So $\hat{\underline{n}}^T \underline{x}_i \propto \text{coef's } \frac{\hat{\underline{n}}^T \underline{x}_i}{\hat{\underline{n}}^T \hat{\underline{n}}}$ of the projection

Experimental Design OCh7

Assume that there are p groups or populations and $\underline{x}_{ij}$ are iid from the j th pop $i=1, \dots, n_j$.

Then $\sqrt{n_i} (\bar{\underline{x}}_i - \underline{\mu}_i) \xrightarrow{D} N_m(0, \Sigma_i)$ by the MCLT.