36} $\sqrt{n}(\hat{\beta}_{MIX} - \beta)$ and $\sqrt{n}(\hat{\beta}_{VS} - \beta)$ are selecting from $\underline{U}_{Hn} = \sqrt{n}(\hat{\beta}_{I_H,0} - \beta)$ and asymptotically from $\underline{U}_H$ (where $\underline{U}_H \sim N_p(\underline{0}, U_{H0})$ if $S \subseteq I_H$). Random selection does not change the dist of $\underline{U}_{Hn}$ and $\underline{U}_H$, but variable selection changes the dist of selected $\underline{U}_{Hn}$ to $\underline{W}_{Hn}$ and the dist of selected $\underline{U}_j$ to $\underline{W}_j$.

37} Variable selection CLT: Assume $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$ and let $\hat{\beta}_{VS} = \hat{\beta}_{I_{H0}}$ with probs π_{Hn} where $\pi_{Hn} \to \pi_H$ as $n \to \infty$. Denote the $\pi_H > 0$ by π_j. Assume $\underline{W}_{jn} = \sqrt{n}(\hat{\beta}^c_{I_j 0} - \beta) \xrightarrow{D} \underline{W}_j$. Then $\underline{W}_n = \sqrt{n}(\hat{\beta}_{VS} - \beta) \xrightarrow{D} \underline{W}$ where the cdf of $\underline{W}$ is $F_{\underline{W}}(t) = \sum_j \pi_j F_{\underline{W}_j}(t)$, Thus $\underline{W}$ is a mixture dist of the $\underline{W}_j$ with prob's π_j.

38) $\hat{\beta}_{S,MIX}$ seems to be a good approx for $\hat{\beta}_{S,VS}$ unless the predictors are highly correlated. Most of the selection bias is due to predictors in E which make selection of S almost random. $\hat{\beta}_{E,MIX}$ and $\hat{\beta}_{E,VS}$ tend to differ, but both use zero padding.

OLS CLT

Let $\underline{\beta} = (\beta_1, \beta_2, 0, 0)' = (\beta_1, \beta_2, \beta_3, \beta_4)'$

ex) Suppose $\sqrt{n}\left(\underline{\hat{\beta}}_S - \underline{\beta}_S\right) \xrightarrow{D} N_2\left(\begin{pmatrix}0\\0\end{pmatrix}, \sigma^2 \underbrace{\begin{pmatrix} v_{11} & v_{12} \\ v_{21} & v_{22}\end{pmatrix}}_{\sigma^2 V_1}\right)$,

$= \sqrt{n}\left(\begin{pmatrix}\hat{\beta}_{1S}\\ \hat{\beta}_{2S}\end{pmatrix} - \begin{pmatrix}\beta_1\\ \beta_2\end{pmatrix}\right)$

Then $\sqrt{n}\left(\underline{\hat{\beta}}_{S,0} - \underline{\beta}\right) \xrightarrow{D} N_4\left(\begin{pmatrix}0\\0\\0\\0\end{pmatrix}, \sigma^2 \underbrace{\begin{pmatrix} v_{11} & v_{12} & 0 & 0 \\ v_{21} & v_{20} & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0\end{pmatrix}}_{\sigma^2 V_{1,0}}\right)$

$= \sqrt{n}\left[\begin{pmatrix}\hat{\beta}_{1S}\\ \hat{\beta}_{2S}\\ 0\\ 0\end{pmatrix} - \begin{pmatrix}\beta_1\\ \beta_2\\ 0\\ 0\end{pmatrix}\right]$

$\underline{\hat{\beta}}_{MIX}$ has a mixture dist of $\underline{\hat{\beta}}_{I_j,0}$ $j = 1, \ldots, 4$ asymptically if $P(S \subseteq I_{min}) \to 1$. $\underline{\beta}_{I_1} = \underbrace{\underline{\beta}_S}_{a_1=2}$, $\underline{\beta}_{I_2} = \underbrace{\begin{pmatrix}\beta_1\\ \beta_2\\ \beta_3\end{pmatrix}}_{a_2=3}$, $\underline{\beta}_{I_3} = \underbrace{\begin{pmatrix}\beta_1\\ \beta_2\\ \beta_4\end{pmatrix}}_{a_3=3}$, $\underline{\beta}_{I_4} = \underbrace{\underline{\beta}}_{a_4=4}$

3g) Geometric Argument. Suppose there is an iid sample $\underbrace{T_{1n}, \ldots, T_{Bn}}_{\text{problem usually } B=1}$ of size B of the statistic. A large sample $100(1-\delta)\%$ prediction region for T_{fn} is $\left\{\underline{w} : D^2_{\underline{w}}\left(\bar{\underline{T}}, \hat{\Sigma}_T\right) \le D^2_{(U_B)}\right\}$

Need $\bar{\underline{T}} \to \underline{\theta}$ faster than $T_i \to \theta$: eg $E(T_i) = \theta$ with MCLT or $\sqrt{n}(T_n - \theta) \xrightarrow{D} N_3(0, \Sigma)$.

Now $D^2_{T_i}\left(\bar{\underline{T}}, \hat{\Sigma}_T\right) = D^2_{\bar{\underline{T}}}\left(T_i, \hat{\Sigma}_T\right)$

($\bar{\underline{T}}$: center of hyperellipsoid)

See Figures on back of HW 10.

$\frac{95}{100}$ T_i are in the prediction region

contains $\approx$ 95% of T_i if n, B are large

So $\bar{T} \in \{\underline{w} : D^2_{\underline{w}}(T_i, \hat{\Sigma}_T) \leq D^2_{(UB)}\}$ iff

$T_i \in \{\underline{w} : D^2_{\underline{w}}(\bar{T}, \hat{\Sigma}_T) \leq D^2_{(UB)}\}$ which occurs

with $\approx 100(1-\delta)\%$ eg 95%. If $\bar{T}$ goes to $\underline{\theta}$

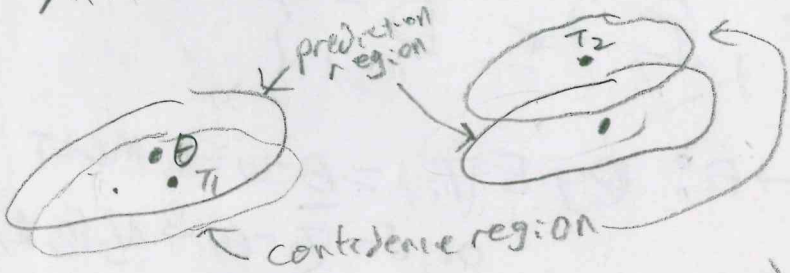
fast enough, then $P[\underline{\theta} \in \{\underline{w} : D^2_{\underline{w}}(T_i, \hat{\Sigma}_T) \leq D^2_{(UB)}\}] \to 1-\delta$.

eg $E\,T_i = \underline{\theta}$ $\quad cov(\bar{T}) = \frac{\Sigma}{B}$ $\quad cov(T_i) = \Sigma$

or $\sqrt{n}(T_i - \underline{\theta}) \xrightarrow{D} N_g(\underline{0}, \Sigma)$ then $\sqrt{n}(\bar{T} - \underline{\theta}) \xrightarrow{D} N_g(\underline{0}, \frac{\Sigma}{B})$

If $\sqrt{n}(T_n - \underline{\theta}) \xrightarrow{D} \underline{U}$ and $\sqrt{n}(T_i^* - T_n) \xrightarrow{D} \underline{U}$

Then the bootstrap sample takes the iid cloud of the $T_1, \ldots, T_B$ centered at $\underline{\theta}$ (or $\bar{T}$) and shifts the cloud to be centered at T_n.



So by the Geometric Argument, the prediction region

Bickel and Ren and Hybrid methods give large sample confidence regions. Need $B \geq 50p$ if $\underline{\theta}$ is $p \times 1$

and need n large, sometimes extremely large. If n is not large enough, undercoverage ($<$ nominal $100(1-\delta)\%$ coverage) can occur. For OLS varsel want $n \geq 20p$, error dist unimodal, not highly skewed.

4.3) For variable selection with the residual and parametric bootstrap (and the nonparametric bootstrap under strong regularity conditions) it can be shown that for large n, the bootstrap sample cloud $T_1^*, \ldots, T_B^*$ tends to be slighty more variable than the iid data cloud $T_1, \ldots, T_B$.

Suppose $S \subseteq I$, then

$E(\hat{\beta}_I) = \beta_I$, $E(\hat{\beta}_{I,0}) = \beta$,

$Cov(\hat{\beta}_I) \to \sigma^2 V_I$, $Cov(\hat{\beta}_{I,0}) \to \sigma^2 V_{I,0}$

$Cov(\hat{\beta}_I^*) \to \sigma^2 V_I$, $Cov(\hat{\beta}_{I,0}^*) \to \sigma^2 V_{I,0}$

$E(\hat{\beta}_I^*) = \hat{\beta}_I$ and $E(\hat{\beta}_{I,0}^*) = \hat{\beta}_{I,0}$.

For the residual bootstrap $Cov(\hat{\beta}^*) \approx \frac{n-p}{n} Cov(\hat{\beta})$.

Let $T_n = A\hat{\beta}_{I_{min},0}$ and $\theta = A\beta$

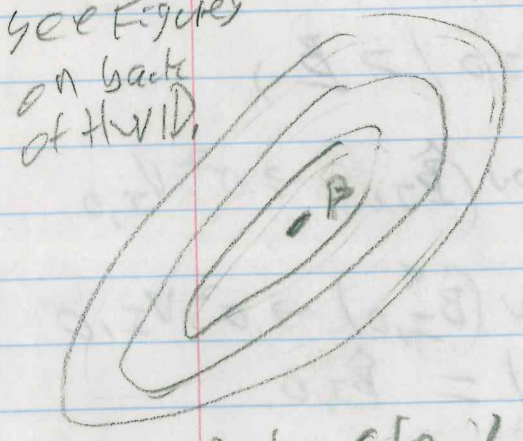
Let the jth component of the bootstrap cloud be $A\hat{\beta}^*_{I_j,0,1}, \ldots, A\hat{\beta}^*_{I_j,B_{jn}}$ (I_j selected B_{jn} times)

while the jth component of the iid cloud is $A\hat{B}_{I_j,1}, \ldots, A\hat{B}_{I_j,n_{j,n}}$ (I_j selected $n_{j,n}$ times),

Asymptotically, the bootstrap and iid sample component clouds have the same variability and shape, but the iid sample component clouds are centered at $E(\hat{B}_{I_j,0}) = \underbrace{B}_{\text{same}}$

while the bootstrap sample component clouds are centered at

$$E(\hat{B}^*_{I_j,0}) = \underbrace{\hat{B}_{I_j,0}}_{\text{different}}$$



→ separate so clouds are centered at $\hat{B}_{I_j,0}$

Makes the overall cloud have greater variability.

Hence by the Geometric Argument expect coverage ≥ nominal.

Go to MREG

Undercoverage can occur if the bootstrap data cloud is less variable than the iid cloud eg $\frac{n-p}{n}$ NOT

Multivariate Linear Regression

Ch8 in course notes

1) Multivariate linear regression (mreg) is a special case of the multivariate linear model where at least one predictor variable takes on many values. MANOVA is another special case where the values of Z are coded often -1, 0, 1.

update

See handout 164) – 190) and references on HW 11

2) The mreg model is $\underset{m\times 1}{\underset{\sim}{y_i}} = \underset{m\times p}{B^T}\,\underset{p\times 1}{\underset{\sim}{x_i}} + \underset{m\times 1}{\underset{\sim}{\varepsilon_i}}$

for $i = 1, \ldots, n$ with $m \geq 2$ response variables $Y_1, \ldots, Y_m$ and p predictor variables

$x_1 \equiv 1, \underbrace{x_2, x_3, \ldots, x_p}_{\text{non trivial predictors}}$. The ith case

$= (x_i^T, y_i^T)^T$ where $x_{i1} = 1$ is often omitted.

data matrix $\begin{pmatrix} x_1^T, & y_1^T \\ \vdots & \vdots \\ x_n^T & y_n^T \end{pmatrix}$ but column of 1s is usually omitted

2) In matrix form,

$$\underset{n\times m}{Z} = \underset{n\times p}{X}\ \underset{p\times m}{B} + \underset{n\times m}{E} \quad \text{where}$$

$$\underset{n\times m}{Z} = (\underset{\sim}{Y}_1, \underset{\sim}{Y}_2, \cdots, \underset{\sim}{Y}_m) = \begin{pmatrix} \underline{y}_1^T \\ \vdots \\ \underline{y}_n^T \end{pmatrix}.$$

$$B = [\underset{\sim}{\beta}_1\ \underset{\sim}{\beta}_2 \cdots \underset{\sim}{\beta}_m], \quad E = [\underline{e}_1 \cdots \underline{e}_m] = \begin{pmatrix} \underset{\sim}{\varepsilon}_1^T \\ \vdots \\ \underset{\sim}{\varepsilon}_n^T \end{pmatrix}.$$

$E(\underset{\sim}{\varepsilon}_h) = \underline{0}$, $cov(\underset{\sim}{\varepsilon}_h) = \Sigma_{\underset{\sim}{\varepsilon}} = (\sigma_{ij})$, $h = 1, \cdots, n$

$E(\underline{e}_i) = \underline{0}$, $cov(\underline{e}_i, \underline{e}_j) = \sigma_{ij} I_n$ for $i, j = 1, \cdots, m$.

3) The $\underline{e}$'s are no longer residuals. Use $\hat{\underline{\varepsilon}}$, $\hat{\Sigma}$

and $\hat{\Sigma}_{ij}$ for residuals.

4) Each response variable Y_j in the m reg model follows a univariate multiple linear regression (MLR) model

$$\underset{n\times 1}{\underset{\sim}{Y_j}} = \underset{n\times p}{\underset{\sim}{X}}\ \underset{p\times 1}{\underset{\sim}{\beta_j}} + \underset{n\times 1}{\underset{\sim}{e_j}} \qquad j = 1, \ldots, m$$

$$Y_{ij} = \underset{\sim}{x_i}^T \underset{\sim}{\beta_j} + e_i = \beta_{0j} + x_{i2}\beta_{2j} + \cdots + x_{ip}\beta_{pj} + e_i$$

↑ error not residual

with $x_{i1} \equiv 1$. X does not depend on j so the same predictors are used for all m response variables.

5) $m = 1 \rightarrow$ MLR model, so for m reg we usually assume $m \geq 2$.

6) Suppose $p = 1$ and $X = \underset{\sim}{1}$. $\underset{n\times m}{Z} = \underset{n\times p}{X}\ \underset{p\times m}{B} + \underset{n\times m}{E}$

Then $\underset{n\times m}{(\underset{\sim}{Y_1} \cdots \underset{\sim}{Y_m})} = \underset{n\times 1}{\underset{\sim}{1}}\ \overbrace{\underset{1\times m}{(\beta_1 \cdots \beta_m)}}^{\underset{\sim}{\mu}^T} + (\underset{\sim}{e_1} \cdots \underset{\sim}{e_m})$

So $\begin{pmatrix} \underset{\sim}{y}_1^T \\ \vdots \\ \underset{\sim}{y}_n^T \end{pmatrix} = \begin{pmatrix} \underset{\sim}{\mu}^T \\ \vdots \\ \underset{\sim}{\mu}^T \end{pmatrix} + \begin{pmatrix} \underset{\sim}{\varepsilon}_1^T \\ \vdots \\ \underset{\sim}{\varepsilon}_n^T \end{pmatrix}$ M

and $\underset{m\times 1}{\underset{\sim}{y}_i} = \underset{m\times 1}{\underset{\sim}{\mu}} + \underset{m\times 1}{\underset{\sim}{\varepsilon}_i}$ $\quad i = 1, \ldots, n$

$\underset{\sim}{B}_i$

m location models $Y_{ij} = \mu_i + e_{ij}$

$i = 1, \ldots, m$, $j = 1, \ldots, n$. Here $\underset{\sim}{y}_1, \ldots, \underset{\sim}{y}_n$ are iid, $E(\underset{\sim}{y}_i) = \underset{\sim}{\mu}$ $\quad Cov(\underset{\sim}{y}_i) = \Sigma_{\underset{\sim}{\varepsilon}} = \Sigma_{\underset{\sim}{y}}$.

This model is the multivariate location and dispersion model and the (multivariate) least squares (LS) estimator is

$$\underset{\sim}{\hat{\mu}} = \overline{\underset{\sim}{y}}_1 = \frac{\sum_{k=1}^{n} \underset{\sim}{y}_k}{n} = \begin{pmatrix} \overline{Y}_1 \\ \overline{Y}_2 \\ \vdots \\ \overline{Y}_m \end{pmatrix}$$ where $\overline{Y}_j = \frac{1}{n}\sum_{i=1}^{n} Y_{ij}$

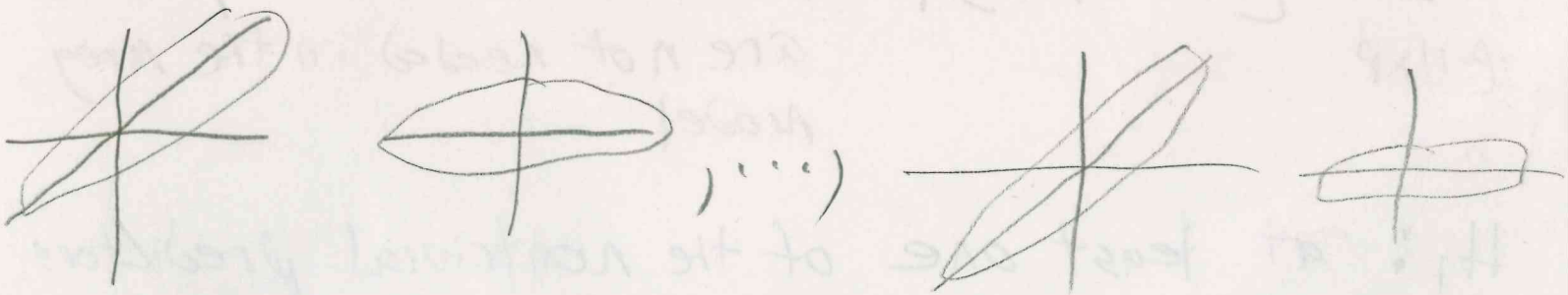
$= \overline{Y}_{0j}$ for $j = 1, \ldots, m$.

Note $\hat{\mu}_j = \overline{Y}_j = \overline{Y}_{0j}$ as in the univariate location model.

7] For each variable Y_k, make a response plot of $\hat{Y}_{ik}$ vs Y_{ik} and a

residual plot of $\hat{Y}_{ik}$ vs $r_{ik} = Y_{ik} - \hat{Y}_{ik}$.

If $n > 10p$ and the mreg model is appropriate, the m pairs of response and residual plots should behave just like those for MLR! the plotted points should scatter about the identity line and $r = 0$ line in roughly evenly populated bands



$k = 1$ $k = m$

8) Consider testing $H_0: LB = 0$ vs $H_1: LB \neq 0$ where $L_{r \times p}$ has full rank $r \leq p$. L was A for MLR. Here $H_1 = HA$.

9) The MANOVA partial F test has a full model using all of the predictors and a reduced model where $r \geq 1$ of the predictors are deleted. The ith row of L has

0's except for a 1 in the position corresponding to the ith variable to be deleted. To delete the last k variables use $L_{k \times p} = [\bar{0} \; I_k]$. The MANOVA F test and F_j test are special cases.

(0) The MANOVA F test uses

$L_{(p-1) \times p} = [\bar{0} \; I_{p-1}]$. H_0: the nontrivial predictors are not needed in the mreg model

H_1: at least one of the nontrivial predictors is needed

This test is the analog of the MLR Anova F_0 test

(1) The F_j test uses $L_j = [\bar{0}, \ldots, 0, \underset{\text{jth position}}{1}, 0, \ldots, 0]$ $(1 \times p)$

H_0 $\underline{b}_j^T = \underline{0}^T \iff$ predictor x_j is not needed in the mreg model given the other predictors are in the model

H_1 $\underline{b}_j^T \neq \underline{0}^T \iff$ predictor x_j is needed

The F_j test is the analog of the MLR t tests for $H_0: \beta_j = 0$.

MLR	MREG
model in matrix form: $\underset{n\times 1}{\underset{\sim}{Y}} = \underset{n\times p}{\mathbb{X}}\,\underset{p\times 1}{\underset{\sim}{\beta}} + \underset{n\times 1}{\underset{\sim}{\varepsilon}}$ (was $\underset{\sim}{e}$) ✓	$\underset{n\times m}{Z} = \underset{n\times p}{\mathbb{X}}\,\underset{p\times m}{B} + \underset{n\times m}{E}$ LM 8.9.5 and 9.11
$E(\underset{\sim}{Y}) = \mathbb{X}\underset{\sim}{\beta}$ ✓	$E(Z) = \mathbb{X}B$
$E(\underset{\sim}{\varepsilon}) = \underset{\sim}{0}$ ✓	$E(E) = 0$, $E(\underset{\sim}{\varepsilon}_j) = \underset{\sim}{0}$
$H = P = \mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'$ ✓	$H = P = \mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'$
$\hat{\underset{\sim}{\beta}} = (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\underset{\sim}{Y}$ ✓	$\hat{B} = (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'Z$; $\hat{\underset{\sim}{\beta}}_j = (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\underset{\sim}{Y}_j$
$\hat{\underset{\sim}{Y}} = P\underset{\sim}{Y}$ ✓	$\hat{Z} = PZ$, $\hat{\underset{\sim}{Y}}_j = P\underset{\sim}{Y}_j$
$\underset{\sim}{r} = \hat{\underset{\sim}{e}} = (I-P)\underset{\sim}{Y}$ ✓	$\hat{E} = (I-P)Z$, $\hat{\underset{\sim}{e}}_j = (I-P)\underset{\sim}{Y}_j$
$E(\hat{\underset{\sim}{\beta}}) = \underset{\sim}{\beta}$ ✓	$E(\hat{B}) = B$, $E(\hat{\underset{\sim}{\beta}}_j) = \underset{\sim}{\beta}_j$
$E(\hat{\underset{\sim}{Y}}) = E(\underset{\sim}{Y}) = \mathbb{X}\underset{\sim}{\beta}$ ✓	$E(\hat{Z}) = E(Z) = \mathbb{X}B$ ($E(\hat{\underset{\sim}{Y}}_j) = \mathbb{X}\underset{\sim}{\beta}_j$)
$\hat{\sigma}^2 = MSE = \frac{\hat{\underset{\sim}{e}}'\underset{\sim}{e}}{n-p} = \frac{\underset{\sim}{e}'\underset{\sim}{e}}{n-p}$ ✓	$\hat{\Sigma}_\varepsilon = \frac{\hat{E}'\hat{E}}{n-p}$, $\hat{\sigma}_j^2 = \frac{\hat{\underset{\sim}{e}}_j'\hat{\underset{\sim}{e}}_j}{n-p}$
response and residual plots	m pairs of response and residual plots
LS CLT ($\frac{\mathbb{X}'\mathbb{X}}{n} \to W^{-1}$) $\sqrt{n}(\hat{\underset{\sim}{\beta}} - \underset{\sim}{\beta}) \xrightarrow{D} N_p(\underset{\sim}{0}, \sigma^2 W)$	MLS CLT ← $\sqrt{n}\,\mathrm{vec}(\hat{B} - B) \xrightarrow{D} N(\underset{pm}{\underset{\sim}{0}}, \Sigma_\varepsilon \otimes W)$; $\sqrt{n}(\hat{\underset{\sim}{\beta}}_j - \underset{\sim}{\beta}_j) \xrightarrow{D} N_p(\underset{\sim}{0}, \sigma_j^2 W)$
$F_{p,n-p} \xrightarrow{D} \frac{1}{p}\chi^2_p$	

MLR	MREG
$H_0: L\underset{\sim}{B} = \underset{\sim}{0}$ $\quad L: r \times p$	$H_0: LB = 0$ $\quad L: r \times p$
$r F_R \xrightarrow{D} \chi^2_r$	$(n-p)\, U(L) \xrightarrow{D} \chi^2_{rm}$
$F_R = \frac{[L\hat{B}]^T \left(L(X^TX)^{-1}L^T\right)^{-1}[L\hat{B}]}{r\,\hat{\sigma}^2_{red}}$	$\frac{(n-p)\,U(L)}{r} =$
$\approx F_{r,n-p}$	$\frac{[\mathrm{vec}(L\hat{B})]^T\left[\hat{\Sigma}_\varepsilon^{-1} \otimes (L(X^TX)^{-1}L^T)^{-1}\right]\mathrm{vec}(L\hat{B})}{r}$
	$\approx F_{rm,\,n-mp}$
$E[y_i] = x_i^T B = B^T x_i$	$E(\underset{\sim}{y}_i) = B^T \underset{\sim}{x}_i$
	$E y_{ij} = \underset{\sim}{x}_i^T \underset{\sim}{B}_j = \underset{\sim}{B}_j^T \underset{\sim}{x}_i$
$V(e_i) = \sigma^2$	$\mathrm{cov}(\underset{\sim}{\varepsilon}_i) = \Sigma_\varepsilon$
$y_i = x_i^T B + e_i$	$\underset{\sim}{y}_i = B^T \underset{\sim}{x}_i + \underset{\sim}{\varepsilon}_i$
$= B^T x_i + e_i$	$\underset{\sim}{y}_i^T = \underset{\sim}{x}_i^T B + \underset{\sim}{\varepsilon}_i^T$