37] The squared Euclidean distance of $\underset{\sim}{x}$ from T is $(\underset{\sim}{x}-T)^T(\underset{\sim}{x}-T) = D^2_{\underset{\sim}{x}}(T, I_p)$

38] Outlier detection if $p > n$, but more than $\frac{n}{2}$ cases in the bulk of the data.

a) use Euclidean distances from the coordinatewise median $D_i(MED(W), I_p)$.

b) Let $MED_j$ be the coordinatewise median computed from cases $\underset{\sim}{x}_i$ with $D_i^2 \leq D_i^2(MED_{j-1}, I_p)$ where $MED_0 = MED(W)$. often use $j=0$ or $j=9$. Let $D_i = D_i(MED_j, I_p)$.

Let $w_i = \begin{cases} 1 & \text{if } D_i \leq MED(D_1, \ldots, D_n) + 5MAD(D_1, \ldots, D_n) \\ 0 & \text{else.} \end{cases}$

The covmb2 set B consists of the $m \geq \frac{n}{2}$ cases with weight $w_i = 1$. The covmb2 estimator $(T, C)$ is the

sample mean and sample covariance matrix of the m cases:

$$T = \frac{\sum_{i=1}^{n} w_i \underset{\sim}{x}_i}{\sum_{i=1}^{n} w_i}, \quad C = \frac{\sum_{i=1}^{n} w_i (\underset{\sim}{x}_i - T)(\underset{\sim}{x}_i - T)^T}{\sum_{i=1}^{n} w_i - 1}.$$

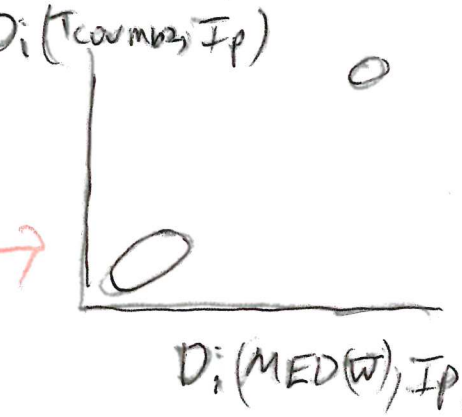
lots of w's

39] Let $\underset{\sim}{w}_i = (Y_i, \underset{\sim}{x}_i^T)^T$ and let the continuous predictors from $\underset{\sim}{x}_i$ be $\underset{\sim}{v}_i$ (the predictors that take on many values, so not gender). Apply the regression (or classification) method to the m cases $\underset{\sim}{w}_i$ corresponding to the covmb2 set $B = \{i_1, \ldots, i_m\}$ applied to $\underset{\sim}{v}_1, \ldots, \underset{\sim}{v}_n$.

MLR, GLMs, GAMs, LDA, QDA, kNN etc.

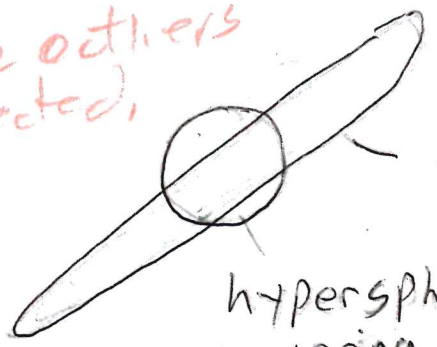
40] The function DDplots plots $D_i(MED(\underset{\sim}{w}), I_p)$ vs $D_i(T_{covmb2}, I_p)$.

The plotted points tend to cluster about the identity line with outliers in the upper right corner of the plot with a gap between the bulk of the data and the outliers.




see HW3

41] To detect outliers in one group and the bulk of the data in another group, the distance of the outliers from the bulk of the data increases roughly with $\sqrt{p}$. otherwise the outliers can't be detected,




outliers in the hypersphere can sometimes be detected with j steps, but not always.

hypersphere covering $\approx$ $\frac{1}{2}$ of the data.

42) For $\begin{pmatrix} Y \\ \underline{x} \end{pmatrix} \sim N_{p+1}(\underline{\mu}, \Sigma)$,

$$1.4826\, MAD(Z_j) \xrightarrow{P} \sqrt{V(Z_j)} = SD(Z_j)$$

where $Z_j = Y$ or $Z_j = X_j$ ($j$th variable).

Since $cov(X,Y) = \frac{V(X+Y) - V(X-Y)}{4}$, (HW1) $\widehat{Cov}(X_i, Y)$.

$$\approx \frac{(1.4826)^2}{4}\left([MAD(X_i+Y)]^2 - [MAD(X_i-Y)]^2\right)$$

$$\approx cov(X_i, Y).$$

So $\hat{\underset{\sim}{\Sigma}}_M = \begin{pmatrix} \widehat{cov}(X_1, Y) \\ \vdots \\ \widehat{cov}(X_p, Y) \end{pmatrix}$ is a robust estimator of $\underset{\sim}{\Sigma}_{XY}$.

(see HW2 C)

43) If the predictors $(\underset{\sim}{X}_1, \ldots, \underset{\sim}{X}_p)$ are continuous, get the covmbz set $B$ applied on $(\underset{\sim}{X}_1, \ldots, \underset{\sim}{X}_p, \underset{\sim}{Y})$ and let

← assuming all continuous

$\hat{\underset{\sim}{\Sigma}}_{MB}$ be the sample covariance vector $\hat{\underset{\sim}{\Sigma}}_{XY}(MB)$ applied to the cases in $B$.

(unlike 39), apply covmbz to $\underset{\sim}{X}$ and $Y$)

HD 24

44} ✓know compute pop OLS $\Sigma_x^{-1}\Sigma_{xy}$ (see HW4 Q4)

and pop OPLS $\lambda$ $\Sigma_{xy}$ given

$\Sigma_x$ and either $\Sigma_{xy}$ or $\beta = \beta_{OLS}$.

use $\Sigma_{xy} = \Sigma_x \beta$.

OPLS paper near bottom of course webpage

ex} $p=4$, $x \sim N_p\left[0, \underbrace{diag(1,2,3,4)}_{\Sigma_x}\right]$ (not needed)

$y_i | x_i = y_i = x_{i1} + x_{i2} + e_i$, $i = 1, \ldots, n$, ind cases

$e_i \sim N(0,1)$ (not needed) $\perp x_i$.

So $\alpha = 0$, $\beta = (1,1,0,0)^T$. $(\alpha, \beta) = (\alpha_{OLS}, \beta_{OLS})$.

$$\Sigma_{xy} = \Sigma_x \beta_{OLS} = \begin{pmatrix} 1 & & & \\ & 2 & & \\ & & 3 & \\ & & & 4 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 0 \\ 0 \end{pmatrix}$$

$\lambda_{OPLS} =$

$$\lambda = \frac{\Sigma_{xy}^T \Sigma_{xy}}{\Sigma_{xy}^T \underbrace{\Sigma_x \Sigma_{xy}}} = \frac{5}{9}, \quad 5 = 1^2 + 2^2 + 0^2 + 0^2$$

$$(1,2,0,0) \begin{pmatrix} 1 \\ 4 \\ 0 \\ 0 \end{pmatrix} = 1(1) + 2(4) = 9$$

so $\underline{\beta}_{OPLS} = \lambda \underline{\Sigma}_{xy} = \left(\frac{5}{9}, \frac{10}{9}, 0, 0\right)^T$

$\neq \underline{\beta}_{OLS} = (1, 1, 0, 0)^T$.

know for E1, HW4, Q4

45] Given $Y | \underbrace{\underline{n}^T \underline{x}}_{w} \sim N(\mu_w, \sigma_w^2)$ as in 23], be able to compute

$$\sigma_w^2 = \sigma_{\underline{n}}^2 = \sigma_Y^2 - \frac{(\underline{\Sigma}_{xy}^T \underline{n})^2}{\underline{n}^T \underline{\Sigma}_x \underline{n}}$$

if $\underline{n} = \underline{\Sigma}_{xy}$ or $\underline{n} = \underline{\beta}_{OLS}$.

ex] In last ex, if $\underline{n} = \underline{\Sigma}_{xy}$, then

$$\sigma_w^2 = \sigma_Y^2 - \frac{(\underline{\Sigma}_{xy}^T \underline{\Sigma}_{xy})^2}{\underline{\Sigma}_{xy}^T \underline{\Sigma}_x \underline{\Sigma}_{xy}} =$$

$$\sigma_Y^2 - \frac{5^2}{9} = \sigma_Y^2 - \frac{25}{9}$$

If $\underset{\sim}{n} = \underset{\sim}{n}_{OLS}$,

then $\sigma_w^2 = V(Y|\underset{\sim}{x}) = \sigma_Y^2 - \underset{\sim}{\Sigma}_{xy}^T \underset{\sim}{\Sigma}_x^{-1} \underset{\sim}{\Sigma}_{xy} =$

$\text{diag}(1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}) \begin{pmatrix} 1 \\ 2 \\ 0 \\ 0 \end{pmatrix}$

$$\left(\sigma_Y^2 - \frac{\left(\underset{\sim}{\Sigma}_{xy}^T \underset{\sim}{\Sigma}_x^{-1} \underset{\sim}{\Sigma}_{xy}\right)^2}{\underset{\sim}{\Sigma}_{xy}^T \underbrace{\underset{\sim}{\Sigma}_x^{-1} \underset{\sim}{\Sigma}_x}_{I} \underset{\sim}{\Sigma}_x^{-1} \underset{\sim}{\Sigma}_{xy}}\right) = \sigma_Y^2 - (1, 2, 0, 0) \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \end{pmatrix}$$

$$= \sigma_Y^2 - 3 \qquad = \sigma_Y^2 - \frac{27}{9}.$$

§3.10

46] The marginal maximum likelihood estimator (MMLE) is simple: regress $\underbrace{Y \text{ on } x_i \text{ to get } \hat{\beta}_i}_{\text{marginal regression estimator}}$,

then $\underset{\sim}{\hat{\beta}}_{MMLE} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)^T$.

47] For MLR, the MMLE is sometimes

called the marginal least squares estimator, and $\hat{\beta}_i$ comes from the simple linear regression (SLR) of $Y_i$ on $X_i$.

48] Pop SLR is OLS pop reg of $Y$ on $X_i$: $\beta_i = \Sigma_{x_i}^{-1} \Sigma_{x_i y}$

$$= \frac{\text{cov}(X_i, Y)}{V(X_i)}.$$

$$\hat{\beta}_i = \frac{\widehat{\text{cov}}(X_i, Y)}{\hat{V}(X_i)},$$

Hence $\underline{\beta}_{MMLE} = [\text{diag}(\Sigma_{\underline{x}})]^{-1} \Sigma_{\underline{x} y}$

diag matrix with diagonal of $\Sigma_{\underline{x}}$

and $\hat{\underline{\beta}}_{MMLE} = [\text{diag}(\hat{\Sigma}_{\underline{x}})]^{-1} \hat{\Sigma}_{\underline{x} y}$.

49] If the $\underline{w}_i$ are vectors of standardized predictors,

such that the sample variances $=1$, then $\hat{\underline{B}}_{MMLE} = \hat{\underline{B}}_{MMLE}(\underline{w}, y)$

$= \hat{\Sigma}_{\underline{w}y} = I^{-1}\hat{\Sigma}_{\underline{w}y} = \hat{\underline{\eta}}_{OPLS}(\underline{w}, y)$.

50] Let $V = \text{diag}(\Sigma_{\underline{x}}) = \text{diag}(\sigma_1^2, \ldots, \sigma_p^2)$.

Let $D = \text{diag}\left(\frac{1}{\sigma_1}, \ldots, \frac{1}{\sigma_p}\right)$.

Let $\text{diag}(\hat{\Sigma}_{\underline{x}}) = \text{diag}(s_1^2, \ldots, s_p^2)$

and $\hat{D} = \text{diag}\left(\frac{1}{s_1}, \ldots, \frac{1}{s_p}\right)$.

Let $\underline{w}_i = \hat{D}\,\underline{x}_i$ and $\underline{v}_i = D\,\underline{x}_i$.

Then $\Sigma_{\underline{v}}$ is the pop correlation matrix of the $\underline{x}_i$. For iid cases,

$\underline{B}_{MMLE} = V^{-1}\Sigma_{\underline{x}y} = V^{-1}\Sigma_{\underline{x}}\Sigma_{\underline{x}}^{-1}\Sigma_{\underline{x}y} =$

$V^{-1}\Sigma_{\underline{x}}\underline{B}_{OLS}$. Then $\underline{B}_{MMLE} = \underline{B}_{OLS}$

if $\underline{\beta}_{OLS} = \underline{0}$, if $(\underline{V}^{-1} - \underline{\Sigma}_x^{-1})\underline{\Sigma}_{xy} = \underline{0}$,

or if $\underline{\beta}_{OLS}$ is an eigenvector of $\underline{V}^{-1}\underline{\Sigma}_x$ with eigenvalue 1.

5.) MMLE theory: For MLR with iid cases, let $\underline{w}_i$ be the standardized predictors and assume $\hat{\underline{\Sigma}}_{\underline{w}y} \xrightarrow{P} \underline{\Sigma}_{\underline{w}y}$ and $\hat{\underline{\Sigma}}_{\underline{w}} \xrightarrow{P} \underline{\Sigma}_{\underline{w}}$ where the $\underline{\Sigma}_{\underline{w}}$ are nonsingular for large enough $n$ and $\underline{\Sigma}_{\underline{v}}$ is nonsingular.

a) $\hat{\beta}_{MMLE}(\underline{w}, Y) = \hat{\underline{\Sigma}}_{\underline{w}y} = \hat{\underline{n}}_{OPLS}(\underline{w}, Y)$

$\xrightarrow{P} \underline{\Sigma}_{\underline{v}y} = n_{OPLS}(\underline{v}, Y) =$

$\underline{\Sigma}_{\underline{v}} \underline{\Sigma}_{\underline{v}}^{-1} \underline{\Sigma}_{\underline{v}y} = \underline{\Sigma}_{\underline{v}} \beta_{OLS}(\underline{v}, Y)$.

b) Let $\underline{\beta}_{OLS} = \beta_{OLS}(\underline{v}, Y)$. Then

$\underline{\beta}_{MMLE} = \beta_{MMLE}(\underline{v}, Y) = \underline{\Sigma}_{\underline{v}} \underline{\beta}_{OLS} = \underline{\beta}_{OLS}$

if $\underset{\sim}{\beta}_{OLS} = \underset{\sim}{0}$ or if $\underset{\sim}{\beta}_{OLS}$ is an eigenvector of $\hat{\Sigma}_{\underset{\sim}{v}}$ with eigenvalue 1.

## Ch 2 Variable selection and Data splitting

1] One common HD technique is to stack marginal estimators: $\bar{\underset{\sim}{x}}$, MED(W), $\hat{\Sigma}_{\underset{\sim}{x}Y}$, $\hat{\underset{\sim}{\beta}}_{OPLS}$, $\hat{\underset{\sim}{\beta}}_{MMLE}$. Another common technique is to use variable selection to reduce the number of predictors.

2] For now, consider general regression models: MLR, PR, BR, PH, AFT, GLM. In low dimensions, $\underset{\sim}{\beta} = \underset{\sim}{\beta}(\underset{\sim}{x}, Y)$ is for the full model (that uses all predictors). regress Y on $\underset{\sim}{x}$

$$\underset{\sim}{x}^T \underset{\sim}{\beta} = \underset{\sim}{x}^T \underset{\sim}{\beta}_S + \underset{\sim}{x}_E^T \underset{\sim}{\beta}_E = \underset{\sim}{x}_S^T \underset{\sim}{\beta}_S$$

If $S \subseteq I$, $\underset{\sim}{x}^T \underset{\sim}{\beta} = \underset{\sim}{x}_S^T \underset{\sim}{\beta}_S = \underset{\sim}{x}_I^T \underset{\sim}{\beta}_I + \underset{\sim}{x}_O^T \underset{\sim}{\beta}_O = \underset{\sim}{x}_I^T \underset{\sim}{\beta}_I$

($\underset{\sim}{x}_I^T \underset{\sim}{\beta}_I$: in the model; $\underset{\sim}{x}_O^T \underset{\sim}{\beta}_O$: out of the model)

3] Variable selection is a search for a subset of predictors that can be deleted

{ without important loss of information if $n \geq 10p$
{ so that model $I$ is good for prediction if $n < 5p$ (HD)

$I$ = remaining predictors in the model

4] Often there are $J$ candidate submodels $I_1, \ldots, I_J$. Let $I_{min}$ be the model that minimizes criterion $C(I)$

$I = I_1, \ldots, I_J$. $AIC(I)$, $BIC(I)$, $EBIC(I)$ are common. $C_p(I)$ for MLR.

5] Forward selection; models have $sp = \alpha + \beta^T x$ (take $\alpha = 0$ if there is no constant). Predictors $x_1, \ldots, x_p$.

$I_1$ uses $x_1^*$ where $x_1^*$ minimizes criterion $C(x_i)$, $i = 1, \ldots, p$: $p$ models fit

$I_2$ uses $(x_1^*, x_2^*)$ where $x_2^*$ minimizes

criterion $C(x_1^*, x_j)$ : $p-1$ models fit

$I_J$ uses $x_1^*, \ldots, x_J^*$

where $x_J^*$ minimizes $C(x_1^*, x_2^*, \ldots, x_{J-1}^*, x_j)$ : $p-J+1$ models fit

often $J = \min(n-r, p)$

where $r \in \{0, 1, 2, 3, 4, 5\}$. (could use $\min(\alpha n, p)$ e.g. $\alpha = 0.1$)

$I_{min}$ has the smallest value of $C(x_1^*, \ldots, x_j^*)$ $j = 1, \ldots, J$.

6] Forward selection can be too slow if $n$ and $p$ are large:

fit $p + p-1 + p-2 + \ldots + p-J+1$ models

step 1 2 3 J

$$= \sum_{i=1}^{J} (p - i + 1) = J(p+1) - \sum_{i=1}^{J} i$$

$$= J(p+1) - \frac{J(J+1)}{2} \approx nP - \frac{n^2}{2} \approx n\left(p - \frac{n}{2}\right) \text{ if } p >> n.$$

more details later

7] Lasso variable selection uses a grid of $J$ $\lambda$ values to get $J$ submodels $I_1, \ldots, I_J$.
often $J = 100$, but larger $J$ may be useful in high dimensions.

8] MMLE variable selection:
Standardize the predictors.
Take the $J$ variables $x_1^*, \ldots, x_J^*$ that have the largest $|\hat{\beta}_i|$.
This method can be fast in high dimensions.
these variables are not standardized

9] using 8) with $r \approx \min(n, p)$ and then forward selection or lasso variable selection on the $x_{i_1}^*, \ldots, x_{i_r}^*$ selected by 8) to get $x_1^*, \ldots, x_a^*$ can reduce the multicollinearity that can result from 8).
e.g. lots of highly correlated predictors.

PhD topic

10} Let $\hat{\beta}_I$ be a $q \times 1$. Form the $p \times 1$ vector $\hat{\beta}_{I,0}$ by adding 0's corresponding to the omitted variables.

e.g. from OLS or PH or GLM

ex} $p = 4$, $\hat{\beta}_{I_{min}} = (\hat{\beta}_1, \hat{\beta}_3)^T$. Then

$$\hat{\beta}_{VS} = \hat{\beta}_{I_{min},0} = (\hat{\beta}_1, 0, \hat{\beta}_3, 0)^T$$

11} As a statistic, $\hat{\beta}_{VS} = \hat{\beta}_{I_K,0}$

with probs $\pi_{K,n} = P(I_{min} = I_K)$, $K = 1, \ldots, J$.

ex} $Y$ = height, $x_1$ = height at right shoulder, $x_2$ = height at left shoulder, $x_3, \ldots, x_p$. For MMLE, if $\hat{\beta}_1 = \frac{6}{5}$, then $\hat{\beta}_2 \approx \frac{6}{5}$. For OPLS, if $\hat{\beta}_1 = \hat{\lambda}\frac{6}{5}$ then $\hat{\beta}_2 = \hat{\lambda}\frac{6}{5}$. For forward selection and lasso VS, probably only one of $x_1$ and $x_2$

standardized data

is included in $I_{min}$.

MMLE variable selection will likely include $x_1$ and $x_2$ which have correlation $\approx 1$.

$p = 4$

ex)

| model $I_j$ | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $\hat{\beta}_{I_j,0}$ |
|---|---|---|---|---|---|
| $I_1$ | | | * | | $(0, 0, \hat{\beta}_3, 0)^T$ |
| $I_2$ | | | * | * | $(0, 0, \hat{\beta}_3, \hat{\beta}_4)^T$ |
| $I_3$ | * | | * | * | $(\hat{\beta}_1, 0, \hat{\beta}_3, \hat{\beta}_4)^T$ |
| $I_4$ | * | * | * | * | $(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4)^T$ |

full model

| model | $I_1$ | $I_2$ | $I_3$ | $I_4$ |
|---|---|---|---|---|
| $C(I)$ | 33.21 | 17.43 | 3.41 | 5.00 |

$I_{min} = I_3 = \{1, 3, 4\}$

$\underset{\sim}{\hat{\beta}}_{I_{min}} = (\hat{\beta}_1, \hat{\beta}_3, \hat{\beta}_4)^T$

$\underset{\sim}{\hat{\beta}}_{I_{min},0} = (\hat{\beta}_1, 0, \hat{\beta}_3, \hat{\beta}_4)^T$

· see HW4

($\hat{\beta}_i$ depends on model $I_j$)