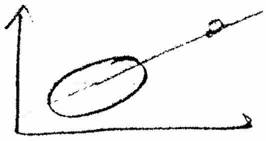
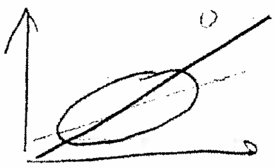
Regularized regression

Outliers

$\Rightarrow$ overestimate of σ^2 $\Rightarrow$ a problem when it comes to inference
$\Rightarrow$ drop & mention



$\Rightarrow$ problems w. interpolation, this one observation will give the appearance of a better fit than the data as a whole supports $\Rightarrow$ drop & discuss



$\Rightarrow$ This observation influences the fit. Note, you would not spot this in a residual plot since it pulled reg. line towards it. $\Rightarrow$ drop it & discuss it

Leverage

– "potential to do harm"

$\hat{y} = Hy$ $\quad \hat{y}_i = \sum_j h_{ij} y_j$ $\quad$ h_{ii} weight assigned to y_i in estimate $\hat{y}_i$.

$\frac{\partial \hat{y}_i}{\partial y_i} = h_{ii}$ – "rate of influence"

Plot it

h_{ii} in univariate case $h_{ii} = \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right)$

large if x_i far from $\bar{x}$, i.e. extreme in x.

(2)

Multivariate case

$$H = X(X'X)^{-1}X'$$

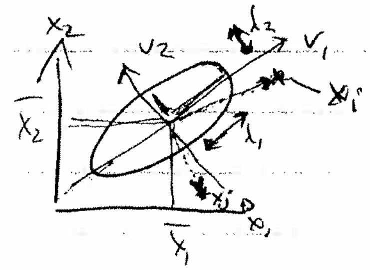
$$h_{ij} = \underset{1\times p}{\underset{\sim}{x}_i}\ \underset{p\times p}{(X'X)^{-1}}\ \underset{p\times 1}{\underset{\sim}{x}_j'}$$

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1\,p-1} \\ 1 & x_{21} & x_{22} & \cdots & x_{2\,p-1} \\ \vdots & & & & \\ 1 & x_{i1} & x_{i2} & \cdots & x_{i\,p-1} \\ \vdots & & & & \\ 1 & x_{n1} & x_{n2} & \cdots & x_{n\,p-1} \end{pmatrix} \quad \boxed{1\ x_{i1}\ x_{i2}\ \cdots\ x_{i\,p-1}} = \underset{\sim}{x}_i$$

Now $X'X$ is a symmetric matrix

– SVD, Decomposition $\Rightarrow$
- $\underset{n\times p}{X} = \underset{n\times p}{U}\,\underset{p\times p}{D}\,\underset{p\times p}{V'}$, $U'U = I_p$, $V'V = I_p$
- $X'X = VDU'UDV' = VD^2V'$ — nonzero eigenvalues of $X'X$ ("SS matrix")
- $(X'X)^{-1} = \underset{p\times p}{V}\,\underset{p\times p}{D^{-2}}\,\underset{p\times p}{V'}$

– $H = X(X'X)^{-1}X' = UU'$



$$- h_{ij} = \underset{\sim}{x}_i (X'X)^{-1} \underset{\sim}{x}_j'$$

$$= \underset{\sim}{x}_i\, V D^{-2} V'\, \underset{\sim}{x}_j'$$

$$= \underset{\sim}{x}_i \left(\sum_{k=1}^{p} d_k^{-2}\, \underset{\sim}{v}_k \underset{\sim}{v}_k' \right) \underset{\sim}{x}_j'$$

$$= \|x_i\|\,\|x_j\| \sum d_k^{-2} \cos\theta_{ik} \cos\theta_{kj}$$

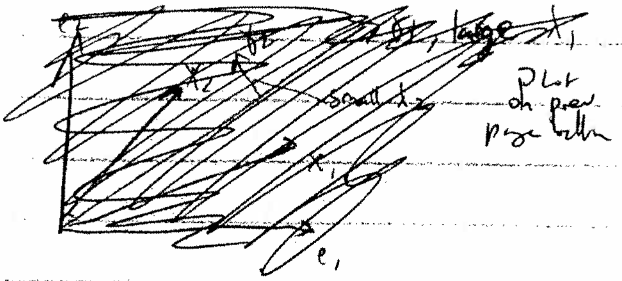
angle between $\underset{\sim}{x}_i$ and k-th eigenvector

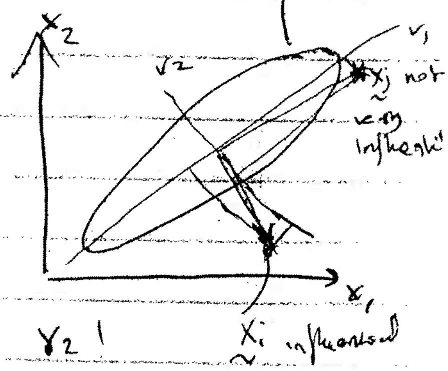
$$- h_{ii} = \sum_k \left(\frac{\|x_i\|^2}{d_k^2} \right) \cos^2\theta_{ik}$$

large h_{ii}: $\underset{\sim}{x}_i$ parallell to an eigenvector with small eigenvalue

⇒ **Meaning:** If $\tilde{x}_i$ extreme in a direction of small variance → can influence the fit.

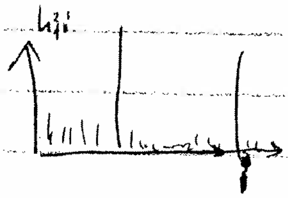
X_1 & X_2 correlated predictors





high leverage ⇒ $\tilde{x}_i$ extreme in direction v_2!

$\boxed{\text{Plot, plot, plot!}}$



Studentized residuals

$e = (I-H)y \qquad V(e_i) = \sigma^2(1-h_{ii})$

→ $e_{(i)}$ = residual @ obs i, when obs i **not** used to fit the line!

Turns out → easy formula to compute this

- $\boxed{e_{(i)} = \dfrac{e_i}{1-h_{ii}}}$!

- $\hat{\beta} - \hat{\beta}_{(i)} = \dfrac{(X'X)^{-1}x_i e_i}{1-h_{ii}}$

OK - so we have leverage to worry about
- and deviating observations/outliers

We have h_{ii} for one, and $e_{(i)}$ for the other

$\Rightarrow$ Cook's distance $C_i = \frac{1}{p} \frac{h_{ii}}{1-h_{ii}} e_{(i)}$

* Which is large if h_{ii} and/or $e_{(i)}$ large

$\boxed{\text{Plot it!}}$

More on multivariate regression

- $Y = X\beta + \varepsilon$, X $n \times p$ matrix
- partition into $X = (X_1, X_2)$

subset of predictors — another subset

- Let's first regress Y on X_1 $\Rightarrow$ hat matrix $H_1 = X_1(X_1'X_1)^{-1}X_1'$
- Now get residuals from first fit: $e_1 = (I - H_1)y$
and residuals from regressing X_2 on X_1: $W_2 = (I - H_1)X_2$
- Regress e_1 on W_2 $\Rightarrow$ hat matrix $H_2 = W_2(W_2'W_2)^{-1}W_2'$

The hatmatrix $H = X(X'X)^{-1}X' = H_1 + H_2$

and $H_2 = X_2(X_2'X_2)^{-1}X_2$ only if

X_2 and X_1 are orthogonal (uncorrelated)

What does this mean?

multi-collinearity

If X_1 & X_2 are correlated
- can't interpret sign & magnitude of coefficients very easily
- the coefficients are affected by both X-components.

If X_1 & X_2 are closely correlated
$\Rightarrow$ numerical instability problems.

What should we do?

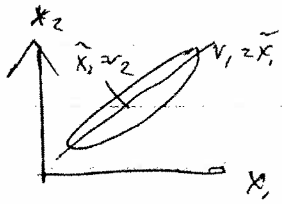
a) If we know a lot about the data reduce the number of predictors to a set of least redundant information

b) Try to regularize or improve the fit.

(b) How to improve the fit?

One option is principal component regression (PC)

- $X = UDV'$
- $X'X = VD^2V'$, $V'V = I = VV'$



- $y = X\beta + \varepsilon = XVV'\beta + \varepsilon = \tilde{X}\tilde{\beta} + \varepsilon$

$\tilde{X}$ is X projected onto orthogonal components

so now $\tilde{X}'\tilde{X} = V'X'XV = D^2$ — a diagonal matrix

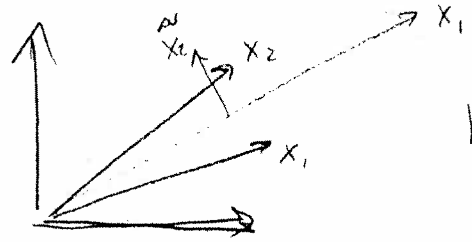
- $$\begin{cases} \hat{\tilde{\beta}} = (\tilde{X}'\tilde{X})^{-1}\tilde{X}'y = D^{-2}\tilde{X}'y \\ V(\hat{\tilde{\beta}}) = \sigma^2 D^{-2} \end{cases}$$

- Now D^2 is the diagonal matrix $\begin{pmatrix} d_1^2 & & \\ & d_2^2 & \\ & & \ddots \\ & & & d_p^2 \end{pmatrix}$

where $d_1^2 \geq d_2^2 \geq \dots \geq d_p^2$
and $d_1^2 = Var(\tilde{X}_1)$, $d_2^2 = Var(\tilde{X}_2)$, ...

- $\hat{\tilde{\beta}}_1$ has smallest variance, and is the slope associated with the maximum variance variable $\tilde{X}_1$

- To regularize the fit → use only the $\tilde{X}_j$ with large d_j^2



Know: reliable slope estimates if var of X large ⇒ so use a linear combination $\tilde{X}_1$ with largest possible variance as a new predictor.

(+) A stable fit, good estimates of slopes of high variance components, component are orthogonal

(−) The linear combinations $\tilde{X}_1, \tilde{X}_2$ ... can be difficult to interpret

Comp (a) <u>Subset regression</u>
- select a smaller subset of predictors that are non-redundant

(+) stable fit, interpretable coefficients

(−) subjective choice of subset to use