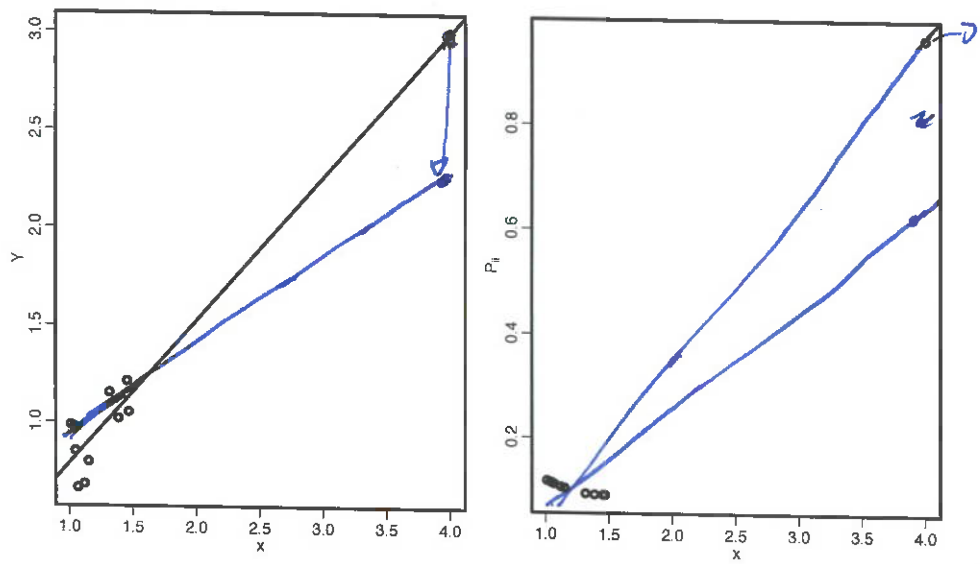
$\sum_{i=1}^n P_{ii} = \text{trace}(P) = \text{rank}(X) =: r$ (see Lemma 12), so the "average" is $r/n$ and a rule of thumb is to take notice when

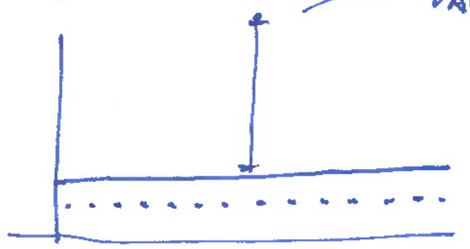
$$P_{ii} > \frac{2r}{n}.$$

**Example 63 (Linear regression)**
$\mathrm{E}\, Y_i = \beta_1 + \beta_2 x_i$




var $e_i \approx 0$
NAMELY
$P_{ii} \approx 1$

$P_{ii} \approx 0$
VAR $e_i$ IS
LARGE



## 11.3 Cook's Distance

To measure how much the $i$th observation changes the estimator $\hat\beta$ one can consider the following measure, called Cook's distance:

$$D_i = \frac{(\hat\beta_{(i)} - \hat\beta)^T X^T X (\hat\beta_{(i)} - \hat\beta)}{p\,\text{RSS}/(n-p)},$$

where $\hat\beta_{(i)}$ is the least squares estimator with the $i$th observation removed. Alternatively,

$$D_i = \frac{(\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_{(i)})^T (\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_{(i)})}{p\,\text{RSS}/(n-p)},$$

where $\hat{\mathbf{Y}}_{(i)} = X\hat\beta_{(i)}$. Rule of thumb: take notice if $D_i$ gets close to 1.

Algebraically equivalent expression:

$$D_i = r_i^2 \frac{P_{ii}}{(1 - P_{ii})r},$$

where $r_i$ is the standardised residual and $r = \text{rank}(X)$. Cook's distance *combines leverage and residual.*

## 11.4 Under/overfitting

**Underfitting**=necessary predictors left out

Let

$$\mathbf{Y} = X\beta + \mathbf{Z}\gamma + \epsilon$$

be the model the observations have come from and let

$$\mathbf{Y} = X\beta + \epsilon$$

be the fitted model.

Suppose we are interested in estimating $\mathbf{c}^T\beta$. $\hat{\beta}$ is biased:

$$\text{E}\,\hat{\beta} = (X^TX)^{-1}X^T\underbrace{(X\beta + \mathbf{Z}\gamma)}_{Y} = \beta + (X^TX)^{-1}X^T\mathbf{Z}\gamma \quad \neq \beta$$

Hence,

$$MSE(\mathbf{c}^T\hat{\beta}) = \text{Var}(\mathbf{c}^T\hat{\beta}) + (E(\mathbf{c}^T\hat{\beta} - \mathbf{c}^T\beta))^2 = \underbrace{\mathbf{c}^T(X^TX)^{-1}\mathbf{c}\sigma^2}_{} + \underbrace{(\mathbf{c}^T(X^TX)^{-1}X^T\mathbf{Z})^2\gamma^2}_{}$$

Let $(\hat{\beta}^F, \hat{\gamma})^T$ be the estimator in the full model. Then

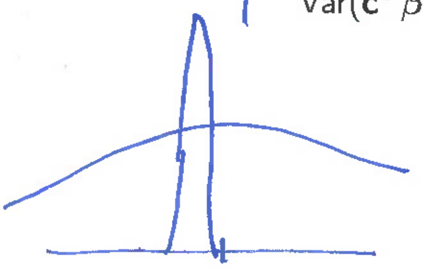
$$\text{cov}\left(\begin{pmatrix}\hat{\beta}^F \\ \hat{\gamma}\end{pmatrix}\right) = \sigma^2 \begin{pmatrix} X^TX & X^T\mathbf{Z} \\ \mathbf{Z}^TX & \mathbf{Z}^T\mathbf{Z}\end{pmatrix}^{-1}$$

Formulas for the inverse of $2 \times 2$ block matrices are known in the literature (see e.g. the Matrix cookbook at http://matrixcookbook.com/). Using these we get

$$\text{cov}(\hat{\beta}^F) = \sigma^2(X^TX)^{-1} + \sigma^2(X^TX)^{-1}X^T\mathbf{Z}(\mathbf{Z}^TQ\mathbf{Z})^{-1}\mathbf{Z}^TX(X^TX)^{-1},$$

where $Q$ is the projection matrix onto $(\text{span}\,X)^\perp$. Hence,

$$\text{Var}(\mathbf{c}^T\hat{\beta}^F) = \sigma^2\mathbf{c}^T(X^TX)^{-1}\mathbf{c} + \frac{\sigma^2}{\mathbf{Z}^TQ\mathbf{Z}}(\mathbf{c}^T(X^TX)^{-1}X^T\mathbf{Z})^2$$





$$\gamma^2 < \frac{\sigma^2}{Z^TQZ} \Rightarrow MSE(c^T\hat{\beta}) < VAR(c^T\hat{\beta}^F)$$

Hence, if $\frac{\sigma^2}{\mathbf{z}^T Q \mathbf{z}} > \gamma^2$ then the estimator from the reduced model has the smaller MSE.

Hence, the mean squared error can be improved by omitting covariates...

Sometimes it pays to use a simpler model!

**Overfitting** = unnecessary predictors included.

This means that some of the components of $\beta$ are 0. Estimator $\hat{\beta}$ is unbiased; however the variance will be larger than in a model where these predictors are left out.

PROBLEM SHEET 10

## 11.5 Weighted Least Squares

So far we have assumed $\text{cov}(\mathbf{Y}) \overset{= \text{Cov}(\epsilon)}{=} \sigma^2 I_n$. Now suppose $\text{cov}(\mathbf{Y}) \overset{= \text{Cov}(\epsilon)}{=} \sigma^2 V$, where $V$ is known, symmetric and positive definite.

**Example 64**
Var $Y_i \propto b_i^2$, $Y_i$'s uncorrelated. Then $V = \begin{pmatrix} b_1^2 & & 0 \\ & \ddots & \\ 0 & & b_n^2 \end{pmatrix}$.

BEST LINEAR UNBIASED ESTIMATOR

How to estimate $\beta$? What is a BLUE in this situation? Main idea: transform the model to a situation in which (SOA) hold true, i.e. in which $\text{cov}(\epsilon) = \sigma^2 I$.

$V$ is symmetric and positive definite. There $\exists$ a nonsingular matrix $T$ such that $T^T V T = I_n$ and $TT^T = V^{-1}$. Indeed, by Lemma 8, $\exists$ an orthogonal matrix $P$ and a diagonal matrix $D$ with the eigenvalues of $V$ on the diagonal s.t.

$$P^T V P = D$$

Let $T = PD^{-1/2}P^T$. Since $P$ is orthogonal, $V = PDP^T$ and thus

$$T^T V T = PD^{-1/2}P^T PDP^T PD^{-1/2}P^T = I.$$

Furthermore, $TT^T = PD^{-1}P^T = (P^T)^{-1}D^{-1}P^{-1} = (PDP^T)^{-1} = V^{-1}$.

$$Z = T^T Y = T^T X\beta + T^T\varepsilon := \tilde{X}\beta + \tilde{\varepsilon}$$

Let $\mathbf{Z} = T^T\mathbf{Y}$. Then

$$= \mathrm{cov}(\tilde{\varepsilon}) = \mathrm{cov}(T^T\varepsilon) = T^T \mathrm{cov}(\varepsilon) T = T^T \sigma^2 V T$$

$$\mathrm{E}(\mathbf{Z}) = \underbrace{T^T X}_{=:\tilde{X}}\beta \quad \text{and} \quad \mathrm{cov}(\mathbf{Z}) = \underbrace{T^T V T}_{=I}\sigma^2 = \sigma^2 I_n$$

Thus the linear model $\mathrm{E}\,\mathbf{Z} = \tilde{X}\beta$ satisfies (SOA). Assuming (FR), we get the following least squares estimator.

$$\begin{aligned}\hat{\beta}_{WLS} &= [\tilde{X}^T\tilde{X}]^{-1}\tilde{X}^T\mathbf{Z}\\ &= [X^T(TT^T)X]^{-1}X^T(TT^T)\mathbf{Y}\\ &= (X^TV^{-1}X)^{-1}X^TV^{-1}\mathbf{Y}.\end{aligned}$$

$$\hat{\beta}_{LS} = (X^TX)^{-1}X^TY$$

$$\hat{\beta}_{WLS} = (X^TV^{-1}X)^{-1}X^TV^{-1}Y$$

$$E[Y] = X\beta$$

Note: $\hat{\beta}$ is an optimal estimator in the sense of the Gauss-Markov theorem.

$$VAR(\hat{\beta}_{WLS}) \leq VAR(\hat{\beta}_{LS})$$

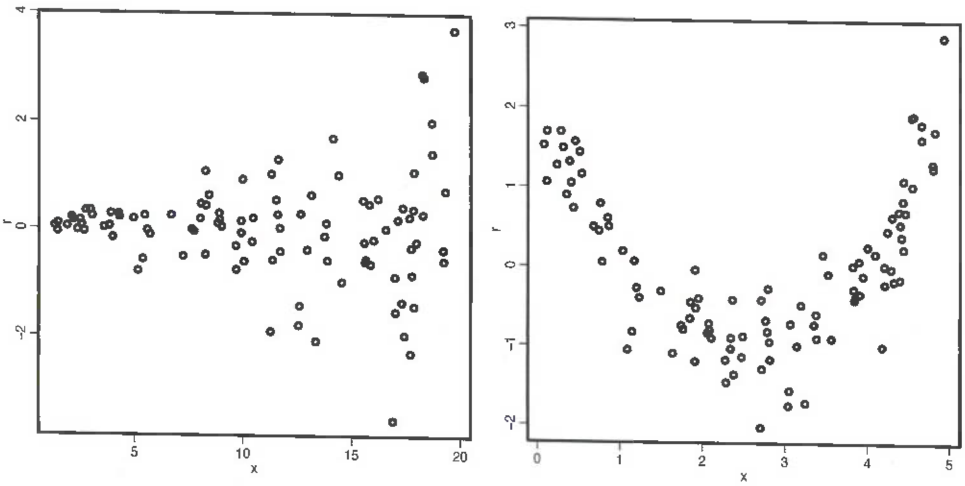
FOR A PROOF OF IT CHECK EXAM 2022 QUESTION 4(a)

## 11.6 Residual Plots

$\hat{\beta}_{WLS}$ IS BLUE

Goal: To detect problems with a model; in other words: to detect a lack of fit of a model:
Approach: Plot standardised residuals against some other variable (e.g. a column of $X$, potentially interesting additional covariates, $\hat{Y}$, ...)

If the model is correct then the resulting plots should just show "noise", with no distinct patterns.




The left plot suggests a non-constant variance (a heteroscedastic error).

The right hand plot indicates that the covariate may have a nonlinear influence.