*Weighte residual analysis

Week3-02424

! linearity assumption of the input ! in GLM besides ~~gua~~ gaussanity

- $N(\beta X, \sigma^2 I)$

parameter values ↑ ↑ Design matrix

↑ independen. assumption

- col 1 = col2 + col3 ⇒ linear dependence, i.e. not structural identifyable

col 1 = col4 + col5 + col 6

(Slide 7) (Same as singular X)

- fix:

	B_1	B_2	B_3
A_1	μ_0	$\mu_0+\beta_2$	$\mu_0+\beta_3$
A_2	$\mu_0+\alpha_2$	$\mu_0+\alpha_2+\beta_2$	$\mu_0+\alpha_2+\beta_3$

- Distance: $\delta_\Sigma(y_1, y_2) = y_1^T \Sigma^{-1} y_2 \rightarrow \|y\|_\Sigma = \sqrt{\delta_\Sigma(y,y)}$

- Deviance: $D(y;\mu) = \delta_\Sigma(y-\mu, y-\mu)$

- factor: discrete

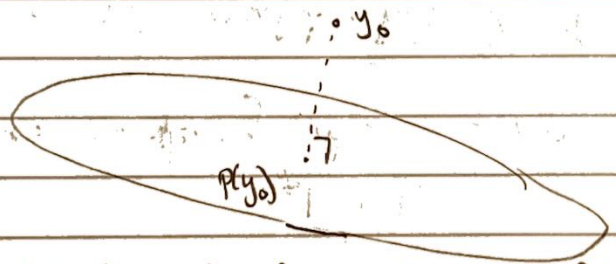
covariat: continous

Week 3 - 02424

- Correlation in data which is not taken in consideration will lead to a too confident model.
  - Very correlated data do not contain n information but ρn information. Hence too much information is assumed
- orthogonal: $\delta_\Sigma(p_0(y); y - p_0(y)) = 0$ (slide 27)

- CI: $\hat{\beta}_i \pm t^{t(n-p)}_{1-\alpha/2} SE_i$ ← $\dim(\beta)$ fordi vi estimere σ^2 ; thm fra slide 28. Vi tager højde for usikkerheden i estimatet, for σ^2 ved at bruge en t-fordeling i stedet for en N.

- Residuals are orthogonal to the model (slide 35)



Vi kan kun variere i de resterende dimensioner n-p fordi de p dimensioner kan beskrives ved modellen.

- exercise slide 40

$$\left.\begin{aligned} y_1 &= \beta_A + \varepsilon_1 \\ y_2 &= \beta_A + \underbrace{(\beta_B - \beta_A)}_{\gamma} + \varepsilon_2 \\ y_3 &= 2\beta_A + \gamma + \varepsilon_3 \end{aligned}\right\} \Rightarrow X = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 2 & 1 \end{bmatrix} \Rightarrow \text{Same H as other design from slide 39}$$

under antagelse at $y_i \sim N(0,1)$ $\forall i$

- $Y^TY = Y^T(I-H)Y + Y^THY$

$X = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$ $H = \begin{bmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{bmatrix}$ under $X \sim H$ (Slide 44)

$$\underbrace{\Sigma(y_i-\bar{y})^2}_{\chi^2_{n-1}} = \underbrace{(\Sigma y_i^2)}_{\chi^2_n} - \underbrace{n\bar{y}^2}_{\chi^2_1}$$

- R^2_{adj} takes complexity into account because

R^2 for the full model has $R^2 = 1$ but is VERY complex

- Slide 50:

$$F(y) = \frac{D(p_1(y); p_0(y))/(m_1-m_0)}{D(y; p_1(y))/(n-m_1)}$$

- Where $D(p_1(y); p_0(y))$ can be understood as $Y^T(H_0-H_1)Y$ and distributed as $\chi^2_{m_1-m_0}$
- And $D(y; p_0(y))$ can be understood as $Y^T(I-H_1)Y$ and distributed as $\chi^2_{n-m_1}$
- We then normalize with the degrees of freedom to obtain the F-distribution

i.e. we have the squared distances distributed by χ^2-distribution and the more degrees of freedoms the larger the variation. In the F-test we then test if it is worth to go down in complexity. See slide 49.