Regression II

Sofia Olhede

November 10, 2020

1. Distributional checks
   - Leverage
   - Weighted Least Squares

Last slide (missing) from last week

- Assume that $Y_1, \ldots, Y_n$ are i.i.d. $N(\mu, \sigma^2)$.

$$\begin{aligned} P(Y_i \leq y) &= P\left(\frac{Y_i - \mu}{\sigma} \leq \frac{y - \mu}{\sigma}\right) \\ &= \Phi\left(\frac{y - \mu}{\sigma}\right) \end{aligned}$$

Then as $n \to \infty$

$$F_n(y_i) \to \Phi\left(\frac{y_i - \mu}{\sigma}\right)$$

(recall from prev chapter that $F_n(y)$ is the Empirical CDF and so

$$\begin{aligned} \Phi^{-1}(F_n(y_i)) &\approx \frac{y_i - \mu}{\sigma} \\ \Phi^{-1}(\text{prop. of obs. } \leq y_i) &\approx \frac{y_i - \mu}{\sigma} \end{aligned}$$

Set–up

- We can generalize this to

$$P(Y \leq y) = F(y)$$

and so as $n \to \infty$

$$F_n(y) \to F(y)$$

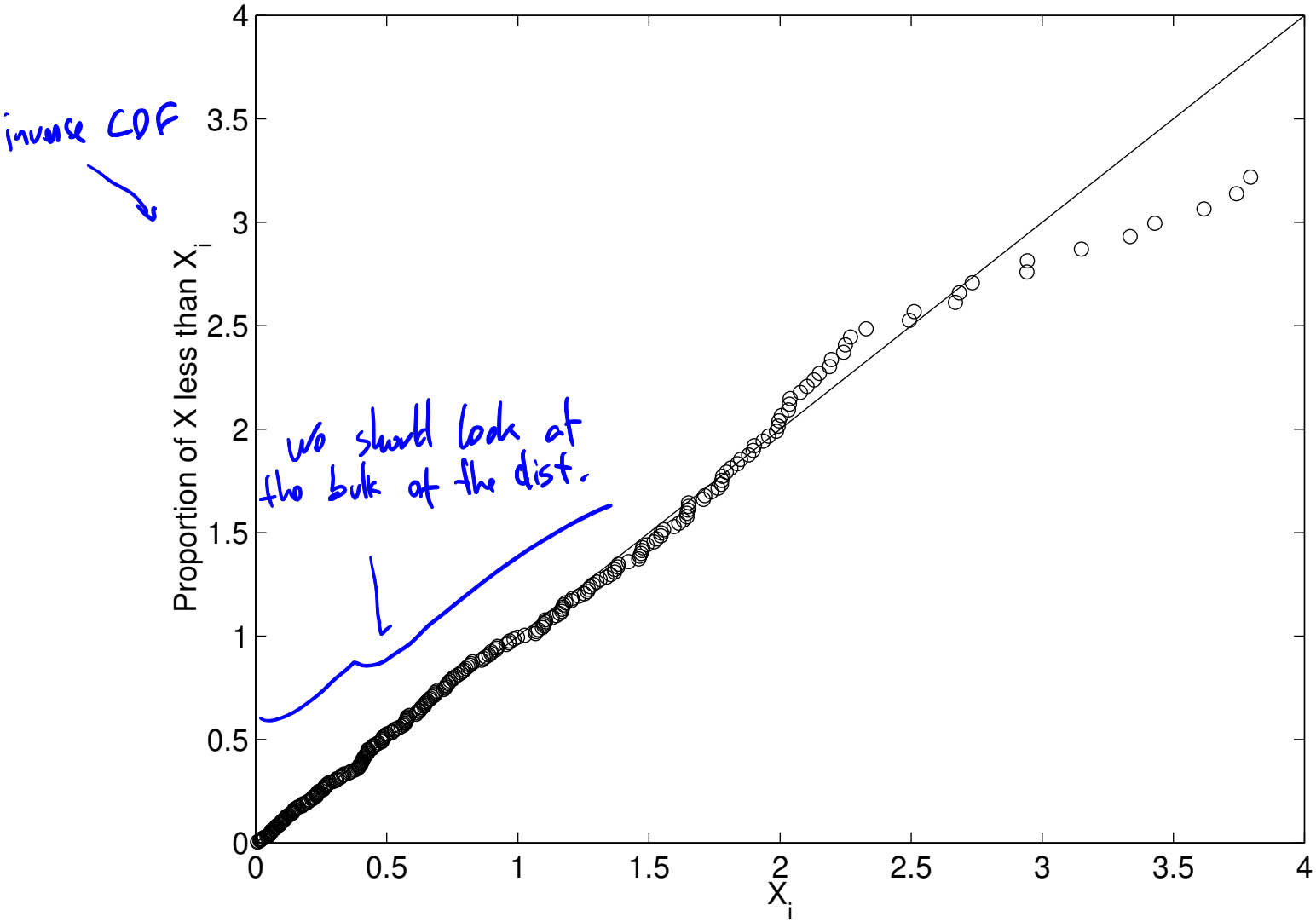
$$\begin{aligned} F^{-1}(F_n(y_i)) &\approx y_i \\ F^{-1}(\text{prop. of obs.} \leq y_i) &\approx y_i. \end{aligned}$$

proportion of observations

"Gormorov-Smirnoff (?) test is useless because it is not powerful enough, because of the asymetri because NULL and the other hypothesis. Also residuals are not a good idea because they're not idd. Q-Q plots is much better, do Q-Q plot" (next slide)

QQ–plot

A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight.



Set–up

- Note that

$$\mathbb{V}\mathrm{ar}\{\mathsf{e}\} = \sigma^2(\mathsf{I}_n - \mathsf{P})$$

hat matrix

- If $p_{ii} \approx 1$ then the variance of the ith residual is very low.
- Totally determined by X, i.e. the design matrix is forcing the ith observation to have high impact.
- The ith observation has **high leverage**.
- $\sum_{i=1}^{n} p_{ii} = p$ so the "average" is p/n and a rule of thumb is to take notice when

$$p_{ii} > \frac{2p}{n}.$$

Weighted Least Squares

- Consider the linear model

$$\mathbb{E}\{Y_i\} = x_i\beta$$

and

$$\mathbb{V}\text{ar}(Y_i) = \frac{\sigma^2}{w_i},$$

where w_i are known weights.

- heteroscedastic variables.
- Using least squares is no longer optimal.
- Cases with small w_i need to be downweighted with respect to the parameter estimation while those with w_i large need to be given more weight.

In statistics, a vector of random variables is heteroscedastic if the variability of the random disturbance is different across elements of the vector. Here, variability could be quantified by the variance or any other measure of statistical dispersion. Heteroscedasticity is the absence of homoscedasticity. A typical example is the set of observations of income in different cities.

The existence of heteroscedasticity is a major concern in regression analysis and the analysis of variance, as it invalidates statistical tests of significance that assume that the modelling errors all have the same variance. While the ordinary least squares estimator is still unbiased in the presence of heteroscedasticity, it is inefficient and generalized least squares should be used instead.[5][6]

in this case: $w_i \to \infty \Rightarrow$ Var is small
$w_i \to 0 \Rightarrow$ Var is huge

Weighted Least Squares

- Find an estimate for β by minimising the weighted sum of squares:

$$\begin{aligned} S(\beta) &= \sum_{i=1}^{n} w_i \left(Y_i - \mathrm{x}_i^T \beta \right)^2 \\ &= \sigma^2 \sum_{i=1}^{n} \frac{(Y_i - E\{Y_i\})^2}{\mathbb{V}\mathrm{ar}\{Y_i\}} \end{aligned}$$

Weighted least squares (WLS), also known as weighted linear regression, is a generalization of ordinary least squares and linear regression in which the errors covariance matrix is allowed to be different from an identity matrix.

Weighted Least Squares

- In vector form we then have:

new matrix D (diagonal)

$$\mathsf{Y} = \mathsf{X}\beta + \mathsf{D}\epsilon$$

$$\mathbb{V}\mathrm{ar}(\epsilon) = \sigma^2 \mathsf{I}_n$$

$$\mathbb{E}\{\epsilon\} = 0$$

and

$$\mathsf{D} = \begin{pmatrix} \frac{1}{\sqrt{w_1}} & 0 & \dots & 0 \\ 0 & \frac{1}{\sqrt{w_2}} & \dots & 0 \\ 0 & \dots & \frac{1}{\sqrt{w_i}} & 0 \\ \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & \frac{1}{\sqrt{w_n}} \end{pmatrix}$$

Weighted Least Squares

- We then multiply through the linear equation by

$$\begin{aligned} \mathsf{D}^{-1}\mathsf{Y} &= \mathsf{D}^{-1}\mathsf{X}\beta + \epsilon \\ \tilde{\mathsf{Y}} &= \tilde{\mathsf{X}}\beta + \epsilon \\ \mathbb{V}\mathrm{ar}\{\epsilon\} &= \sigma^2 \mathsf{I}_n \\ \mathbb{E}\{\epsilon\} &= 0 \end{aligned} \tag{1}$$

This is recognizable as a liner model. The β estimate is given by

$$\begin{aligned} \hat{\beta} &= \left(\tilde{\mathsf{X}}^T\tilde{\mathsf{X}}\right)^{-1}\tilde{\mathsf{X}}^T\tilde{\mathsf{Y}} \\ &= \left(\left(\mathsf{D}^{-1}\mathsf{X}\right)^T\left(\mathsf{D}^{-1}\mathsf{X}\right)\right)^{-1} \\ &\quad \left(\mathsf{D}^{-1}\mathsf{X}\right)^T\mathsf{D}^{-1}\mathsf{Y} \\ &= \left(\mathsf{X}^T\mathsf{V}\mathsf{X}\right)^{-1}\mathsf{X}^T\mathsf{V}\mathsf{Y} \end{aligned} \tag{2}$$

where $\mathsf{V} = \mathsf{D}^{-2}$.

Testing in the Least Squares Set–Up

- Assume that $\boldsymbol{Y} \sim N(X\beta, \sigma^2 \mathsf{I}_n)$.
- Definition: If $\mathsf{Z} \sim N(\boldsymbol{\mu}, \mathsf{I}_n)$ where $\boldsymbol{\mu} \neq 0$ then $\mathsf{Z}^T\mathsf{Z}$ is said to have a non-central χ^2 distribbution on n d. o. f. and non-centrality parameter $\delta > 0$ given by $\delta^2 = \mu^T\mu$.
- $\mu = 0 \Rightarrow \chi^2_n$ as the non-centrality parameter is then zero.
- We normally for the general distribution write $U = \mathsf{Z}^T\mathsf{Z} \sim \chi^2_n(\delta)$.
- The distribution of $\mathsf{Z}^T\mathsf{Z}$ depends on μ only via δ.
- $\mathbb{E}(U) = n + \delta^2$.
- $\mathbb{V}\mathrm{ar}(U) = 2n + 4\delta^2$
- If $U_i \sim \chi^2_{n_i}(\delta_i)$ for $i = 1, \ldots k$ and if the $\{U_i\}$ are all independent then

$$\sum_{i=1}^{k} U_i \sim \chi^2_n(\delta),$$

where $n = \sum_{i=1}^k n_i$ and $\delta^2 = \sum_{i=1}^k \delta_i^2$.

Testing in the Least Squares Set–Up

- Lemma: If $Z \sim N(\boldsymbol{\mu}, I_n)$ and if A is a $n \times n$ symmetric and idempotent matrix of rank r then

In linear algebra, an idempotent matrix is a matrix which, when multiplied by itself, yields itself. i.e. A = A^2

$$Z^T A Z \sim \chi^2_n(\delta),$$

where $\delta^2 = \mu^T A \mu$.

Proof: Let A be a symmetric idempotent matrix ($A^2 = A$). Then A has r eigenvalues that are unity, and $n - r$ eigenvalues that are zero. Because A is symmetric there is an orthogonal $P^T P = I_n$ matrix P st

$$P^T A P = D,$$

where D is diagonal with r ones and $n - r$ zeros down the diagonal. Let $V = P^T Z$. Then

$$V \sim N(P^T \mu, I_n).$$

Testing in the Least Squares Set–Up

- Furthermore it follows that

$$\begin{aligned} Z^T AZ &= (PV)^T A(PV) \\ &= V^T P^T APV \\ &= V^T DV \\ &= V^T D^T DV \\ &= (DV)^T (DV) \\ &= \text{sum of squares of } r \text{ components} \\ &= \chi^2_r(d), \end{aligned}$$

$P^T AP = D$ from prev. slide

D is diagonal and only zeros or ones so $D = D^T D$ eg $\begin{bmatrix} 0 & & \\ & 1 & \\ & & 0 \end{bmatrix}\begin{bmatrix} 0 & & \\ & 1 & \\ & & 0 \end{bmatrix} = \begin{bmatrix} 0 & & \\ & 1 & \\ & & 0 \end{bmatrix}$

for some d. In fact

$$d^2 = \mathbb{E}(DV)^T \mathbb{E}(DV) = (DP^T\mu)^T(DP^T\mu) = \mu^T PDP^T\mu = \mu^T A\mu.$$

Testing in the Least Squares Set–Up

- Lemma: If $Z \sim N(\boldsymbol{\mu}, \mathrm{I}_n)$ and A_1 and A_2 are symmetric idempotent matrices such that $A_1 A_2 = 0$ then $Z^T A_1 Z$ and $Z^T A_2 Z$ are independent.
  Proof: $Z^T A_i Z = (A_i Z)^T (A_i Z)$ for $i = 1, 2$. Consider the two vectors $A_1 Z$ and $A_2 Z$ then

$$\mathbb{C}\mathrm{ov}\{A_1 Z, A_2 Z) = A_1 \, \mathbb{C}\mathrm{ov}(Z) A_2^T \tag{3}$$
$$= A_1 \mathrm{I} A_2 = 0. \tag{4}$$

  This means that every component of $A_1 Z$ is uncorrelated with every component of $A_2 Z$. By normality this means that the components are independent.

- Corollary: If $A_1, \ldots, A_k$ are symmetric and idempotent and if $A_i A_j = 0$ for $i \neq j$ then $\{Z^T A_i Z\}$ are mutually independent.

↳ for all i in $1, \ldots, K$

Testing in the Least Squares Set–Up

Lemma: If $A_1, \ldots A_k$ are symmetric $n \times n$ matrices such that $\sum A_i = \mathsf{I}$ and such that $\mathrm{rank}(A_i) = r_i$ then the following are equivalent:

(a) $\sum_i r_i = n$
(b) $A_i A_j = 0 \; i \neq j$
(c) A_i are idempotent for $i = 1, \ldots, k$.

i.e. if one property holds, the other two will hold too!

Testing in the Least Squares Set–Up

- If

$$Z \sim N(\mu, I_n)$$

and

$$\sum_i A_i = I_n$$

where A_i with ranks r_i are symmetric $n \times n$ matrices such that at least one of

1. $\sum_i r_i = n$
2. $A_i A_j = 0$ for $i \neq j$
3. A_i are idempotent.

holds (and therefore all of them, proof omitted) then

$$Z^T A_i Z$$

are independent distributions of the type:

$$\chi^2_{r_i}(\delta_i)$$

where $\delta_i^2 = \mu^T A_i \mu$.

→ this will be important later. We will find many matrices in this form

↓

eg if we take quadratic form of Z i.e. $Z^T Z$ and we multiply an idempotent matrix, we end up with many χ^2 dist.

Testing in the Least Squares Set–Up

- Proof: A_i are assumed to be idempotent. By the lemma this means

$$Z^T A_i Z \sim \chi^2_{r_i}(\delta_i).$$

Because they are mutually orthogonal by assumption this implies that

$$Z^T A_i Z$$

are independent.

Testing in the Least Squares Set–Up

- Assume we want to test the hypothesis

$$H_0 : \ A\beta = 0$$

versus

$$H_1 : \ A\beta \neq 0$$

where $\text{rank}(A) = s = p - p_0$.

- Under H_0 we get the simpler linear model

$$E\{Y\} = X_0\beta_0$$

where β_0 is $p_0 \times 1$. New hat matrix:

$$P_0 = X_0\left(X_0^T X_0\right)^{-1} X_0^T.$$

Testing in the Least Squares Set–Up

- $\boldsymbol{P}_0$ has trace p_0.
- Consider the likelihood ratio:

$$t = \frac{\text{maximum liklihood under } H_1}{\text{maximum liklihood under } H_0}$$

- From MLE we get a biased estimate of σ and the least squares estimates of $\widehat{\boldsymbol{\beta}}$.
- Plugging in:

$$t = \left(\frac{\widehat{\sigma}^2_{ML,0}}{\widehat{\sigma}^2_{ML}} \right)^{\frac{n}{2}}.$$

NULL model

alternative model

- Consider a monotonic increasing function of t:

$$f(t) = \frac{n-p}{p-p_0}\left(t^{2/n} - 1\right),$$

take to power $\frac{2}{n}$ and subtract 1:

$$t^{\frac{2}{n}} - 1 = \frac{\hat{\sigma}^2_{ML,0}}{\hat{\sigma}^2_{ML}} - 1$$

?

or

$$F = \frac{n-p}{p-p_0}\frac{RSS_0 - RSS}{RSS}.$$

?

Testing in the Least Squares Set–Up

- Use the Fisher-Cochran theorem:

$$I_n = (I_n - P) + (P - P_0) + P_0$$

with ranks

$$n = (n - p) + (p - p_0) + p_0.$$

Let

$$\begin{aligned} A_1 &= (I_n - P) \\ A_2 &= (P - P_0) \\ A_3 &= P_0 \end{aligned}$$

These are are symmetric and idempotent.

Testing in the Least Squares Set–Up

- We may write $P_0 = XB$ for some B of constants. Let

$$Z = \frac{1}{\sigma} Y$$

Note

$$RSS = Y^T A_1 Y = \sigma^2 Z^T A_1 Z,$$

$$RSS_0 - RSS = Y^T A_2 Y = \sigma^2 Z^T A_2 Z$$

and so with NTA by the Fisher-Cochran theorem

$$RSS/\sigma^2 \sim \chi^2_{n-p}$$

and

$$(RSS_0 - RSS)/\sigma^2 \sim \chi^2_{p-p_0}$$

independently (the non-centrality parameters vanish.)

Testing in the Least Squares Set–Up

- We then have

$$\begin{aligned} F &= \frac{\sigma^2}{\sigma^2}\frac{n-p}{p-p_0}\frac{RSS_0 - RSS}{RSS} \\ &\sim \frac{\chi^2_{p-p_0}/(p-p_0)}{\chi^2_{n-p}/(n-p)} \\ &\sim F_{p-p_0,n-p} \end{aligned}$$

- Decompose the **total sum of squares** by

$$\begin{aligned} \mathsf{Y}^T\mathsf{Y} &= \mathsf{Y}^T(\mathsf{I}_n - \mathsf{P})\mathsf{Y} \\ &\quad + \mathsf{Y}^T(\mathsf{P} - \mathsf{P}_0)\mathsf{Y} + \mathsf{Y}^T\mathsf{P}_0\mathsf{Y} \end{aligned}$$

These are the **total sum of squares** (TSS), **residual sum of squares** (RSS), **sum of squares for testing H_0** and the sum of squares due reduction due to β_0. This can be summarized in an ANOVA (ANalysis Of VAriance) table.

Testing in the Least Squares Set–Up

ANOVA table:

P_0 is the new hat matrix, slide 17 (?)

Source	d.o.f.	Sum of Squares	Mean squares	$F_{statistics}$
Red	$p-s$	$\underline{y}^T \mathrm{P}_0 \underline{y}$		
reduction				
H_0	s	$\underline{y}^T(\mathrm{P}-\mathrm{P}_0)\underline{y}$	$M_1 = \frac{\underline{y}^T(\mathrm{P}-\mathrm{P}_0)\underline{y}}{s}$	$\frac{M_1}{M_2}$
Residual	$n-p$	RSS: $\underline{y}^T(\mathrm{I}-\mathrm{P})\underline{y}$	(4) $M_2 = \frac{\underline{y}^T(\mathrm{I}-\mathrm{P})\underline{y}$	
Σ total	n	$\underline{y}^T\underline{y}$		

- Then
- (4) $M_2 = \frac{RSS}{n-p}$ is an *unbiased* estimate of σ^2.
- Reject the null hypothesis at level α if

$$F > f_\alpha$$

where

$$P(F_{s,n-p} > f_\alpha) = \alpha.$$

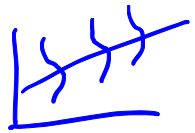
f_α: α quantile for the f-distribution

→ sum of squares (final)
= unexplained/residual sum of sq. $y^T(I-P)y$ on the full model
+ sum of sq for testing $y^T(P-P_0)y$
+ sum of reduction due to P_0 $y^T P_0 y$
(also sum-of-sq)
$= y^T(P_0 + P - P_0 + I - P)y = y^T y$

NOTE: this is one of the most used methods for linear models in the world. Here we assum y is Gassian. Later, we'll learn generalized linear models that allow y not to be Gausssian!

Assumptions in the Least Squares Set–Up

not weighted

- Four basic assumptions inherent in the Gaussian linear regression model:
- Linearity: $\mathbb{E}\{Y\}$ is linear in X.
- Homoskedasticity: $\mathbb{V}\text{ar}\{\epsilon_j\} = \sigma^2$ for all j. (before we mentioned heteroscedasticity)
- Gaussian Distribution: errors are normally distributed. → 
- Uncorrelated Errors: ϵ_i uncorrelated with ϵ_j for $i \neq j$.
- When one of these assumptions fails clearly, then Gaussian linear regression is inappropriate as a model for the data.
- Isolated problems, such as outliers and influential observations also deserve investigation. They may or may not decisively affect model validity.

Assumptions in the Least Squares Set–Up

- Scientific reasoning: impossible to validate model assumptions.
- Cannot prove that the assumptions hold. Can only provide evidence in favour (or against!) them.
- Strategy: Find implications of each assumption that we can check graphically (mostly concerning residuals).
- Construct appropriate plots and assess them (requires experience).
- 'Magical Thinking': Beware of overinterpreting plots!

Outliers

- An outlier is an observation that does not conform to the general pattern of the rest of the data.
- We *standardise* the residuals through:

r_i is the standardized residual = residual divided by estimation of variance

residuals we saw before (Gaussian)

$$r_i = \frac{e_i}{\sqrt{s^2(1-p_{ii})}}$$

$\rightarrow \chi^2$ dist.

variance of residuals (?)

p_{ii} is taken from hat matrix

where

$$s^2 = \frac{RSS}{n-p}.$$

χ^2 dist.

s^2 has $n-p$ degrees of freedom, and note that r_i is *not* student t.

Outliers (more)

- Outliers may be influential: they "stand out" in the "y-dimension".
- However an observation may also be influential because of unusual values in the "x–dimension".
- Such influential observations cannot be so easily detected through plots. But we may wish to automatically detect problems.
- How to find cases having strong effect on fitted model?
- Idea: see effect when case j , i.e., (x_j^T, Y_j) is not kept.
- Let $\boldsymbol{\beta}_{-j}$ be the LSE when model is fitted to data without case j and let $\widehat{\mathrm{Y}}_{-j} = \mathrm{X}\boldsymbol{\beta}_{-j}$ be the fitted value.

Outliers (more)

- Define Cook's distance

$$C_j = \frac{1}{ps^2}\left\{\hat{Y} - \hat{Y}_{-j}\right\}^T \left\{\hat{Y} - \hat{Y}_{-j}\right\}.$$

number of model params · variance

- This measures the scaled distance between the predictions, and we recall that:

$$s^2 = \frac{1}{n-p}\|Y - \hat{Y}\|^2.$$

variance

- It is possible to show that

$$C_j = \frac{r_j^2 p_{jj}}{p(1-p_{jj})},$$

hat matrix element

and thus it can be seen that a large C_j implies either large r_j and/or large p_{jj}.

Outliers (more)

- Cases with $C_j > 8/(n-2p)$ are considered large.
- We therefore plot C_j against j and compare with this cut-off.

very nice / useful slide

Diagnostics

How do we diagnose if a model appears to be fitting well?

- We plot $\boldsymbol{Y}$ against columns of $\boldsymbol{X}$ to check for linearity and outliers.
- We plot the standardized residuals r against the columns of $\boldsymbol{X}$.
- We plot the standardized residuals r against covariates we left out.
- We plot r against $\widehat{Y}$ to check homoscedasticity.
- We make qq plots to check distribution.
- We make the Cook distance plot to check for influential observations.

(last slide missing)