

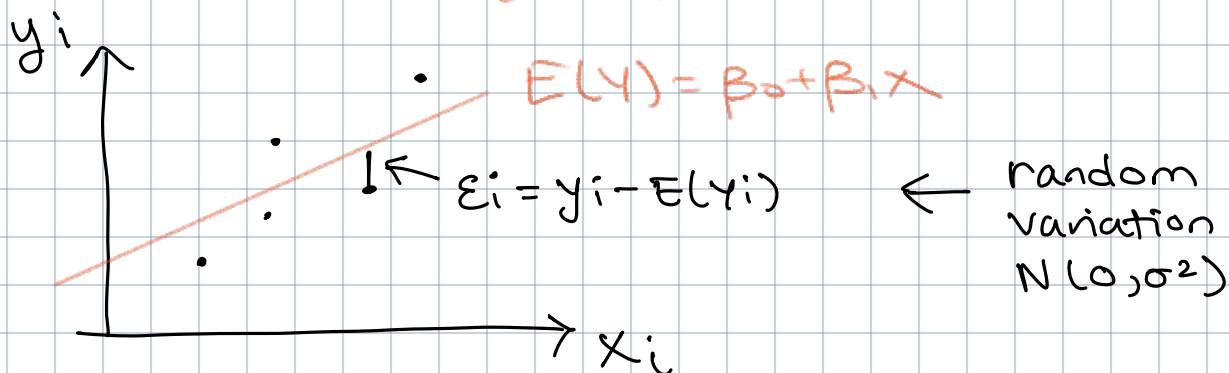
Chap. 9 Model check

Linear model

$$\vec{Y} = X\vec{\beta} + \vec{\varepsilon} \quad \vec{\varepsilon} \sim \text{MVN}(\vec{0}, \sigma^2 I)$$

- $E(Y_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki}$
- $\text{Var}(\varepsilon_i) = \sigma^2$
- $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0$ (independent)
- $\varepsilon_i \sim N(0, \sigma^2)$ (identical dist.)

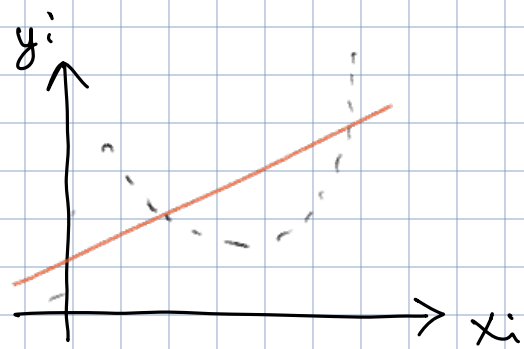
Simple : $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$
consider known



Not

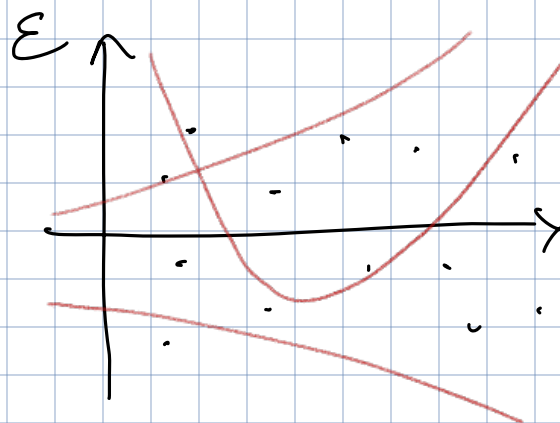


$\text{Var}(Y_i) = \text{Var}(\varepsilon_i)$
depends on x



$E(Y_i)$ not linear
→ need transform
of x .

NB1: Cannot generalize this plot to more than one $x \rightarrow$ need something else for multiple linear models.



Not

$$E(Y) = \beta_0 + \beta_1 x$$

$\uparrow$ can have $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots$

NB2: We do not know β -s and we cannot observe the ϵ -s.

Check 1

$\hat{\epsilon}$	vs	$\hat{Y} = E(Y)$
$\uparrow$		$\uparrow$
residuals		fitted values

Example: $Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i$

(Slides)

1) Random $\checkmark$

2)

3)

Why $\hat{\varepsilon}$ vs $\hat{y}$ "works"?

$$\hat{y} = E(\vec{y}) = X\hat{\beta} = H\vec{y}$$

$$\hat{\varepsilon} = \vec{y} - \hat{y} = (I - H)\vec{y}$$

from 3.62

$$\begin{aligned} \text{cov}(\hat{\varepsilon}, \hat{y}) &= (I - H) \text{cov}(\vec{y}) H^T = (I - H) \sigma^2 I H \\ &= \sigma^2 (H - H \cdot H) = 0. \end{aligned}$$

But all our plots show no linear trend

→ need independence to match plot I

→ $\text{cov} = 0 \Rightarrow$ independence in normal distn.

Distn $\hat{y}, \hat{\varepsilon}$:

$$E(\hat{y}) = X\vec{\beta} \quad \text{cov}(\hat{y}) = \sigma^2 H H = \sigma^2 H$$

$$E(\hat{\varepsilon}) = \vec{0} \quad \text{cov}(\hat{\varepsilon}) = \sigma^2 (I - H)$$

$$\begin{array}{ccc} H & n \times n & , \quad I - H \quad n \times n \\ \text{rank } k-1 & & \text{rank } n-k-1 \end{array} \quad \begin{array}{c} \cdot \cdot \\ \sim \end{array}$$

Recall that book has a strict def of MVN → requires covariance matrix with non-zero determinant. We are not so strict...

$$\hat{y} \sim \text{MVN}(X\vec{\beta}, \sigma^2 H) \quad \hat{\varepsilon} \sim \text{MVN}(\vec{0}, \sigma^2 (I - H))$$

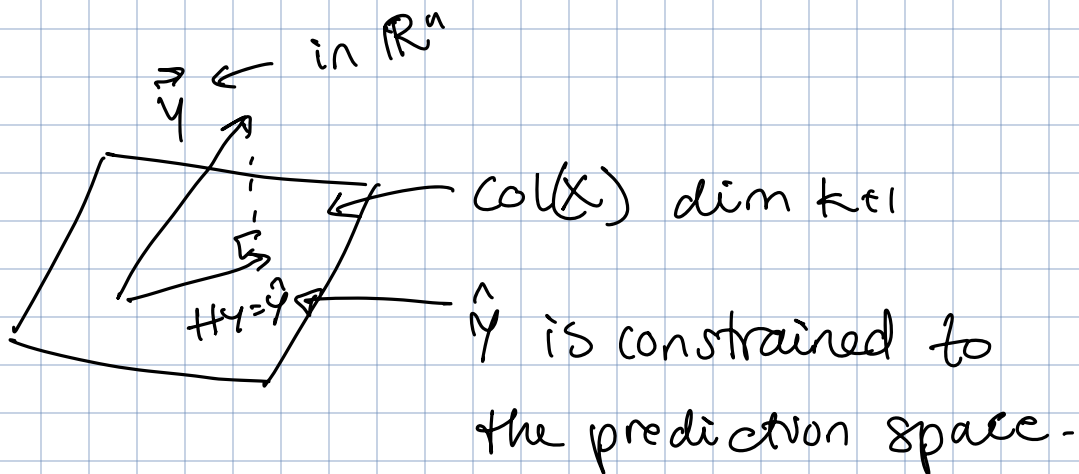
→ $\hat{y}, \hat{\varepsilon}$ independent.

If time

* Note: $\hat{y}$ and $\hat{e}$ are normal
but 'degenerate'.

Ex: $\vec{y}$ takes values in $\mathbb{R}^n$
 $\rightarrow$ n-dimensional space

X has $\text{Dim}(\text{col}(X)) = k+1$



Check 2 : QQ plot

QQ = quantile-quantile

check $\vec{\varepsilon} \sim \text{MVN}(\vec{0}, \sigma^2 \mathbf{I})$

We have $\hat{\varepsilon} \sim \text{MVN}(\vec{0}, \sigma^2 (\mathbf{I} - \mathbf{H}))$

$$\rightarrow \text{Var}(\hat{\varepsilon}_i) = \sigma^2 (1 - h_{ii})$$

Studentized :

↑ i th diag. element of $\mathbf{H}$.

$$r_i = \frac{\hat{\varepsilon}_i}{\sqrt{s^2 (1 - h_{ii})}}$$

r_i 's approx $N(0, 1)$

[slides]

Check 3 : Outliers and influential points

(slides)

Discuss : which is worse?

A) outlier in y only $\rightarrow$
changes $\hat{\beta}$ slightly, $\hat{\varepsilon}_i$ large

B) outlier in x only $\rightarrow$ high leverage
minimal change $\hat{\beta}$, $\hat{\varepsilon}_i$ small

c) Now we are in trouble,
outlier in both x and y
→ $\hat{\epsilon}_i$ large
→ leverage large

Def: h_{ii} is the leverage of obs. i

$$\hat{y}_i = h_{ii} y_i + \sum_{j \neq i} h_{ij} y_j$$

$$\frac{1}{n} \leq h_{ii} \leq 1$$

$$\sum_{i=1}^n h_{ii} = k+1$$

$$-0.5 < h_{ij} < 0.5$$

if h_{ii} 'large' → $\hat{y}_i$ forced towards y_i .

Also $\text{Var}(\hat{y}_i) = \sigma^2 h_{ii}$ large.

Plot $\hat{\epsilon}$ vs $\hat{y}$, QQ-plots

Cooks distance

$$D_i = \frac{r_i}{k+1} \left(\frac{h_{ii}}{1-h_{ii}} \right)$$

r_i and h_{ii}
large
→ D_i large.

Check $D_i > 1$

Check 4 Multicollinearity

If time,
otherwise
in model-
selection

→ highly correlated X-s give
unstable estimates

Def: Variance inflation factor for β_j

$$VIF_j = \frac{1}{1 - R_j^2}$$

R_j^2 : R^2 from
regressing x_j on
all other covariates

model variance σ^2 est. $\rightarrow \hat{\sigma}^2$

$$\text{Var}(\hat{\beta}_j) = \frac{\hat{\sigma}^2}{(n-1) S_{X_j}^2} \cdot VIF_j$$

↑ sample variance x_j

inflates variance
of coeff-est.

check $VIF_j > 10$