Model Diagnostics

Ch 5, 6, 8, 9 out of order

Model $\underline{y} = X\beta + \varepsilon$ is wrong

How ~~harmful is it~~ can we tell

How harmful is it.

Look at Residuals

$$e = \underline{y} - \hat{\underline{y}}$$

Detect

- outliers
- Curvilinear trends
- Non-constant variance
- Departure from normality

Look at Plots

- $\underline{y}$ vs $\hat{\underline{y}}$
- $\hat{\underline{y}}$ vs e
- IVs vs residuals
- Time sequence vs residuals

Plot $\hat{y}$ vs y

What should we see if model is correct?

Squared sample correlation between y & $\hat{y}$ equals $R^2 = \frac{SSR}{SST}$

Assume $\sum e_i = 0$

$$SST = SSE + 0 + SSR$$

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 \quad \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \nearrow + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

$$\geq \sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y})$$

Because $\frac{\sum_{i=1}^{n} y_i}{n} = \frac{\sum_{i=1}^{n} \hat{y}_i}{n}$, $\bar{\hat{y}}_i = \bar{y}$

$$r = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2 \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}}$$

$$= \sum_{i=1}^{n}(y_i - \hat{y}_i)\hat{y}_i = 0$$

$$- \sum_{i=1}^{n} y_i \hat{y}_i$$

$$\bar{y}\sum_{i=1}^{n}(y_i - \bar{y})$$

$$= \sqrt{SST \cdot SSR}$$

$$= \frac{\sum_{i=1}^{n} y_i \hat{y}_i - \bar{y} \sum_{i=1}^{n} \hat{y}_i}{\sqrt{SST \cdot SSR}} = \frac{y'\hat{y} - n\bar{y}^2}{\sqrt{SST \cdot SSR}}$$

of course $\hat{y} \neq y$

Note $y'\hat{y} = y'Xb = \hat{y}'\hat{y}$ because

$$\hat{y}'\hat{y} = (Xb)'Xb = b'X'X(X'X)^{-1}X'y$$

$$= b'X'Xb = b'\underbrace{X'X(X'X)^{-1}}_{I}X'y$$

$$= b'X'y = (Xb)'y = \hat{y}'y$$

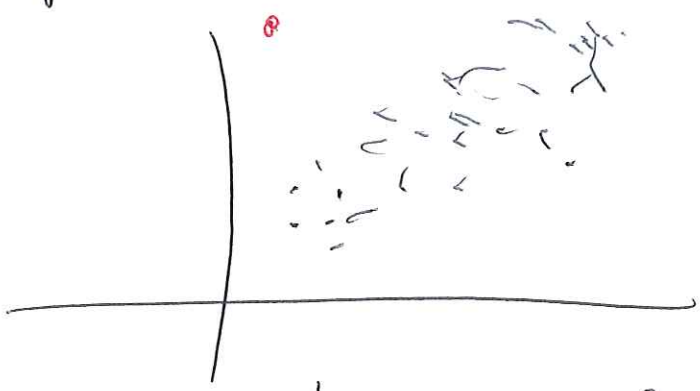
$$= \hat{y}'\hat{y}$$

$$= \frac{\sum_{i=1}^{n} \hat{y}_i^2 - n\bar{y}^2}{\sqrt{SST \cdot SSR}} = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sqrt{SST \cdot SSR}}$$

$$= \frac{SSR}{\sqrt{SST \cdot SSR}} = \sqrt{\frac{SSR}{SST}}$$

$$= \sqrt{R^2}$$

Scatterplot will have pos slope



Reflect strength of relationship

Could see outliers
curvilinear trends
Non-constat variance

Linear regression means linear in the Betas, so

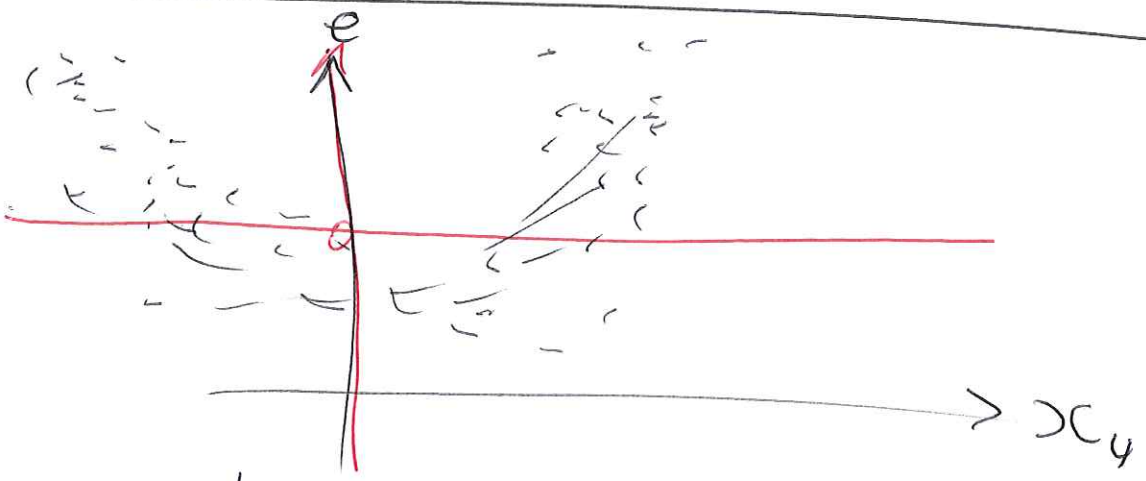
$$y_i = \beta_0 + \beta_1 \sqrt{x_1} + \beta_2 x_2^2 + \varepsilon_i$$

is Linear

$$y_i = \beta_0 e^{\cos(\beta x_i)} + \varepsilon_i$$

NONLINEAR

Plots of e_i vs x variables in the model



Q. What should you see?

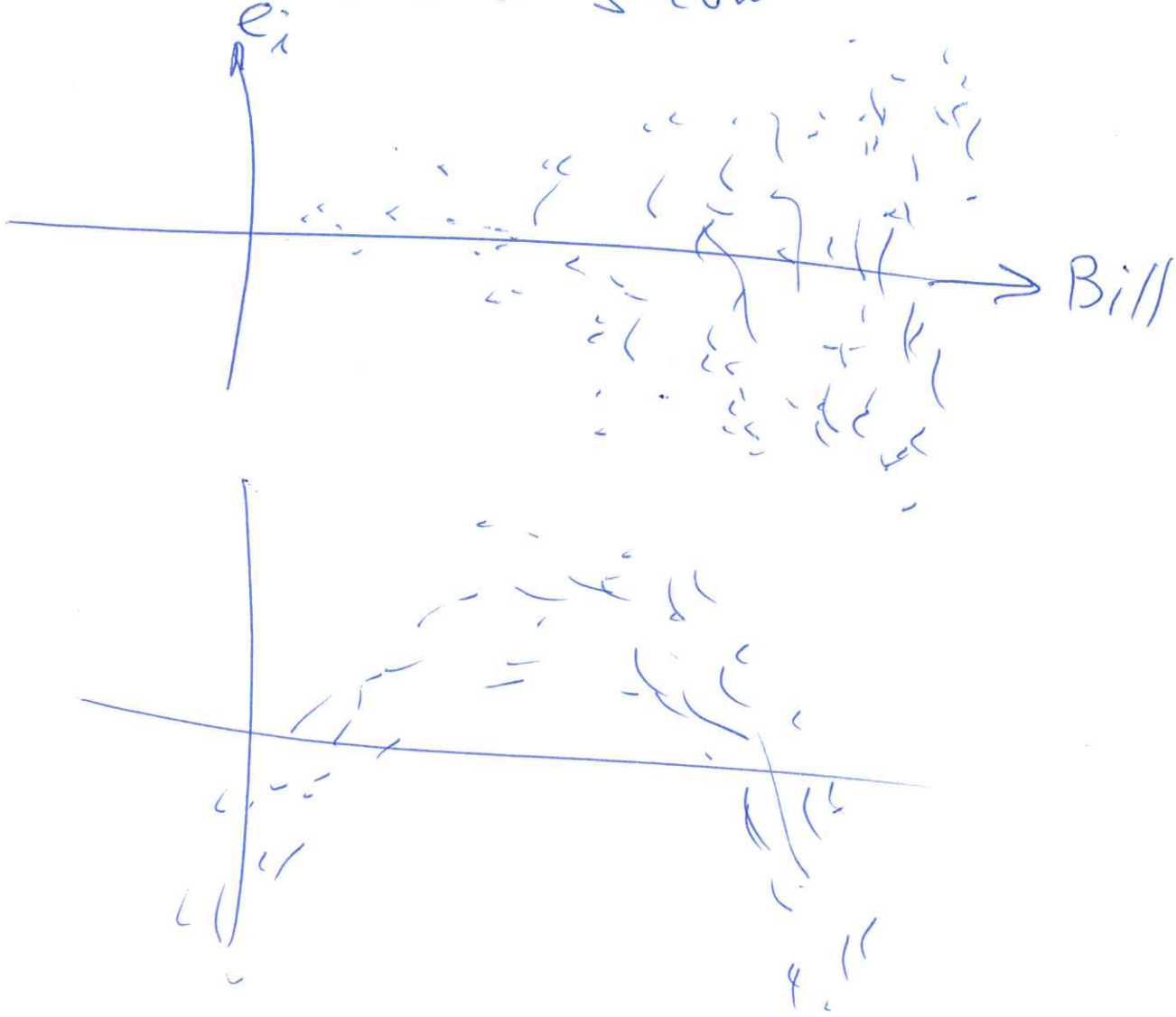
$$r_{x_j,e} = \frac{\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)(e_i-0)}{\sqrt{\quad}}$$

$$top = \sum_{i=1}^{n} x_{ij}e_i - \bar{x}_j \sum_{i=1}^{n} e_i$$

$$\begin{bmatrix} 1 & 1 & 1 \\ x_1 & & \\ x_{1j} & x_{2j} & x_{nj} \end{bmatrix} \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix} = 0$$

X' $\quad e$

Plots could show



Can plot e_i vs variables not in equation.