STA302/1001: Methods of Data Analysis

Instructor: Fang Yao

Chapter 8: Diagnostics via Residuals

Regression Diagnostics

- also known as model checking
- check if your fitted model is “healthy” or not
- mainly to check if the linear model assumptions are satisfied or not
- up to now, the only tool that you have learnt for model checking is the lack-of-fit test
- we have also looked at some residual plots - but they were not that formal
- now we examine the residuals in a more formal way

Regression Diagnostics: Residuals

- recall: $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$
- then $\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta}$

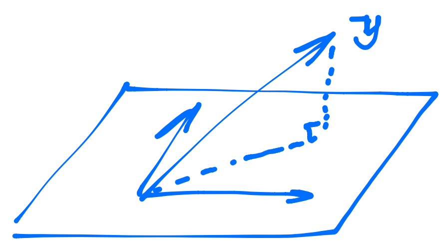
$$= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

- define $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$
- $\mathbf{H}$: hat matrix

  transforms the data $\mathbf{Y}$ into fitted values $\hat{\mathbf{Y}}$
- residuals: $\hat{\mathrm{e}} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{HY} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$
- idompotent projection matrix

$$\mathbf{H}' = \mathbf{H}, \quad \mathbf{HH} = \mathbf{H}, \quad \mathbf{HX} = \mathbf{X}$$

$HY \to \hat{Y}$



Difference between $\hat{\mathrm{e}}$ and e

- assumptions for e (the statistical errors):
- $\mathrm{E}(\mathbf{e}) = \mathbf{0}$, $\mathrm{Cov}(\mathbf{e}) = \sigma^2\mathbf{I}$
- with these assumptions, it is easy to show (later)

$$\mathrm{E}(\hat{\mathbf{e}}) = \mathbf{0} \quad \text{and} \quad \mathrm{Cov}(\hat{\mathbf{e}}) = \sigma^2(\mathbf{I} - \mathbf{H})$$

* different

- note that the variances of $\hat{e}_i$'s are not the same
- Let h_{ii} be the ith diagonal element of $\mathbf{H}$ leverage value
- then $\mathrm{Var}(\hat{e}_i) = \sigma^2(1 - h_{ii})$
- also $\hat{e}_i$'s are correlated—but we usually ignore this
- if intercept is included, $\sum_{i=1}^n \hat{e}_i = 0$ (check SLR case)

SLR: $Y_i = \alpha + \beta X_i + e_i$

fitted regression: $\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i = \bar{Y} - \hat{\beta}\bar{X} + \hat{\beta} X_i = \bar{Y} + \hat{\beta}(X_i - \bar{X})$

$\hat{e}_i = Y_i - \hat{Y}_i = Y_i - \bar{Y} - \hat{\beta}(X_i - \bar{X})$

$\sum_{i=1}^n \hat{e}_i = \sum_{i=1}^n (Y_i - \bar{Y}) + \hat{\beta}\sum(X_i - \bar{X}) = 0$

The Hat Matrix

- verify:

$$\begin{aligned}\mathbf{HH} &= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \cdot \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \\ &= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \\ &= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{H}\end{aligned}$$

$HX = X(X'X)^{-1}X' = X$

$E(\hat{e}) = E[(I-H)Y]$
$= (I-H)X\beta$
$= (X - HX)\beta$
$= 0$

- similarly, $(\mathbf{I}-\mathbf{H})$ is also idempotent

- some direct consequences:

$(I-H)^2 = I + H^2 - 2H$

$$(\mathbf{I}-\mathbf{H})\mathbf{X} = \mathbf{0} \Rightarrow \underline{\mathrm{E}(\hat{\mathrm{e}}) = \mathbf{0}}, \qquad \mathbf{H}(\mathbf{I}-\mathbf{H}) = 0$$

$$\mathrm{Cov}(\hat{\mathrm{e}}, \hat{\mathbf{Y}}) = \mathrm{Cov}((\mathbf{I}-\mathbf{H})\mathbf{Y}, \mathbf{HY}) = \sigma^2\mathbf{H}(\mathbf{I}-\mathbf{H}) = \mathbf{0}$$

$$\mathrm{Cov}(\mathbf{Y}) = \sigma^2\mathbf{I}, \quad \mathrm{Cov}(\hat{\mathbf{Y}}) = \sigma^2\mathbf{HH}' = \sigma^2\mathbf{H}$$

$$\mathrm{Cov}(\hat{\mathrm{e}}) = \sigma^2(\mathbf{I}-\mathbf{H})(\mathbf{I}-\mathbf{H})' = \sigma^2(\mathbf{I}-\mathbf{H})$$

note that $\mathrm{Cov}(\hat{\mathrm{e}}) = \mathrm{Cov}(\mathbf{Y}-\hat{\mathbf{Y}}) = \mathrm{Cov}(\mathbf{Y}) - \mathrm{Cov}(\hat{\mathbf{Y}})$

Diagonal of the Hat Matrix h_{ii}

- Let us look at h_{ii} more carefully:
- $\mathrm{Var}(\hat{e}_i) = \sigma^2(1 - h_{ii})$
- with an intercept, one can show $\frac{1}{n} \leq h_{ii} \leq \frac{1}{r_i}$
  where r_i is # of replicates for $\mathbf{x}_i$
- so, the bigger the h_{ii}, the smaller the $\mathrm{Var}(\hat{e}_i)$
- what does it mean when $\mathrm{Var}(\hat{e}_i) = 0$?
  only the ith observation itself is used to get $\hat{y}_i$
- h_{ii} is sometimes called the leverage of the ith observation
- what does a high-leverage observation mean?

Diagonal of the Hat Matrix h_{ii} - con't

- $\mathbf{H}$ is idompotent, $h_{ii} = h_{ij}^2$, i.e., $h_{ii}(1 - h_{ii}) = \sum_{j \neq i} h_{ij}^2$
- $\hat{y}_i = \sum_{j=1}^n h_{ij} y_j = h_{ii} y_i + \sum_{j \neq i}^n h_{ij} y_j$
- as $h_{ii} \to 1$, $\hat{y}_i \to y_i$, $\hat{y}_i$ is mostly determined by y_i only
  is this what we want?
- with an intercept, use a centered design matrix
  (think about SLR case)

$$h_{ii} = \frac{1}{n} + (\mathbf{x}_i - \bar{\mathbf{x}})'(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{x}_i - \bar{\mathbf{x}})$$

- this is the equation of an ellipsoid centered at $\bar{\mathrm{x}}$
- large values of h_{ii} indicate unusual values for $\mathbf{x}_i$
  (large leverage values $\neq$ outliers)

Large Leverage Values

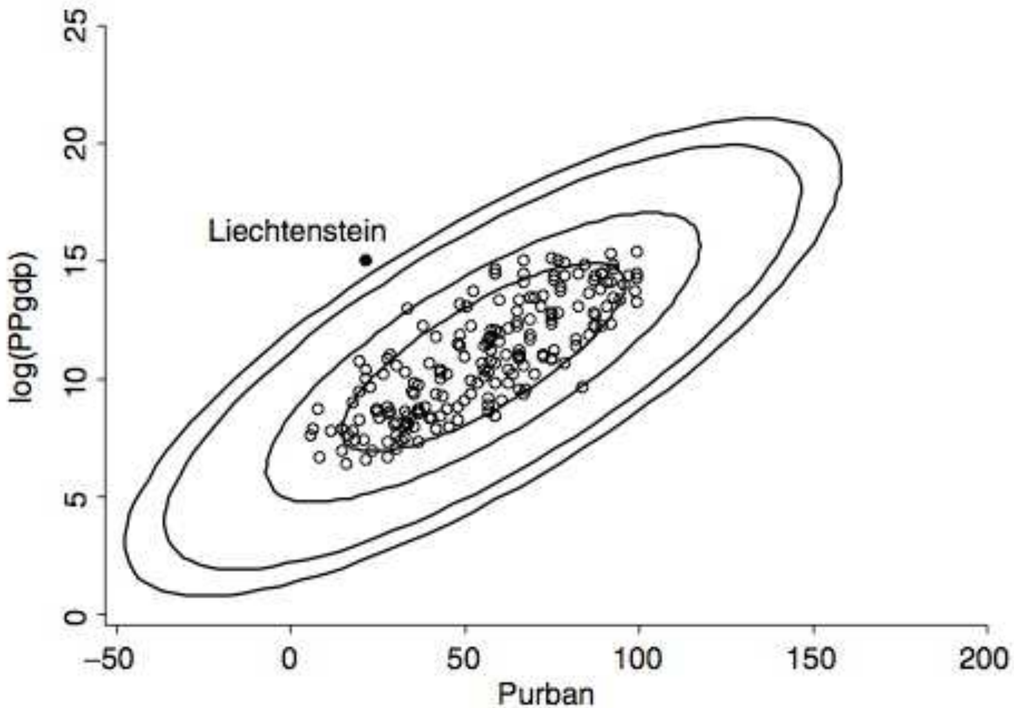


FIG. 8.1 Contours of constant leverage in two dimensions.

When doing WLS

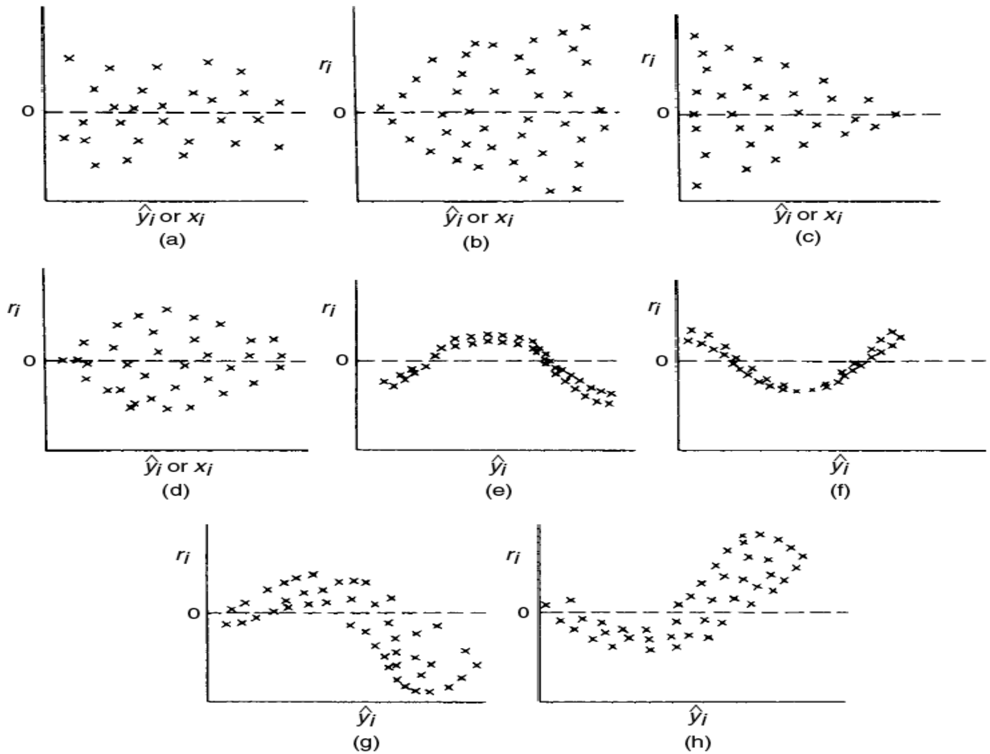
- assumption: $\mathrm{Var}(\mathbf{e}) = \sigma^2 \mathbf{W}^{-1}, \quad \mathbf{W}$: known weights
- then $\mathbf{H} = \mathbf{W}^{1/2}\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}^{1/2}$
- fitted values: $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{H}\mathbf{Y}$
- residuals may be defined in different ways
- definition 1: $\hat{e}_i = y_i - \hat{y}_i$
- definition 2: $\hat{e}_i = \sqrt{w_i}(y_i - \hat{y}_i)$
- we will use definition 2
- in *R*: definition 2 is sometimes known as Pearson residuals, or weighted residuals

When the model is CORRECT...

- let U be any of the terms, or any linear combination of the terms, e.g., fitted value
- then $\mathrm{E}(\hat{e}_i|U_i) = 0$ and $\mathrm{Var}(\hat{e}_i|U_i) = \sigma^2(1 - h_{ii})$
- so a plot of residuals against U should have constant mean zero
- and that the variance function of $\hat{\mathrm{e}}$ is <u>not</u> constant (even if the model is correct)
- the variability will be smaller for large h_{ii}
- so when the model is correct, residual plots should look like null plots

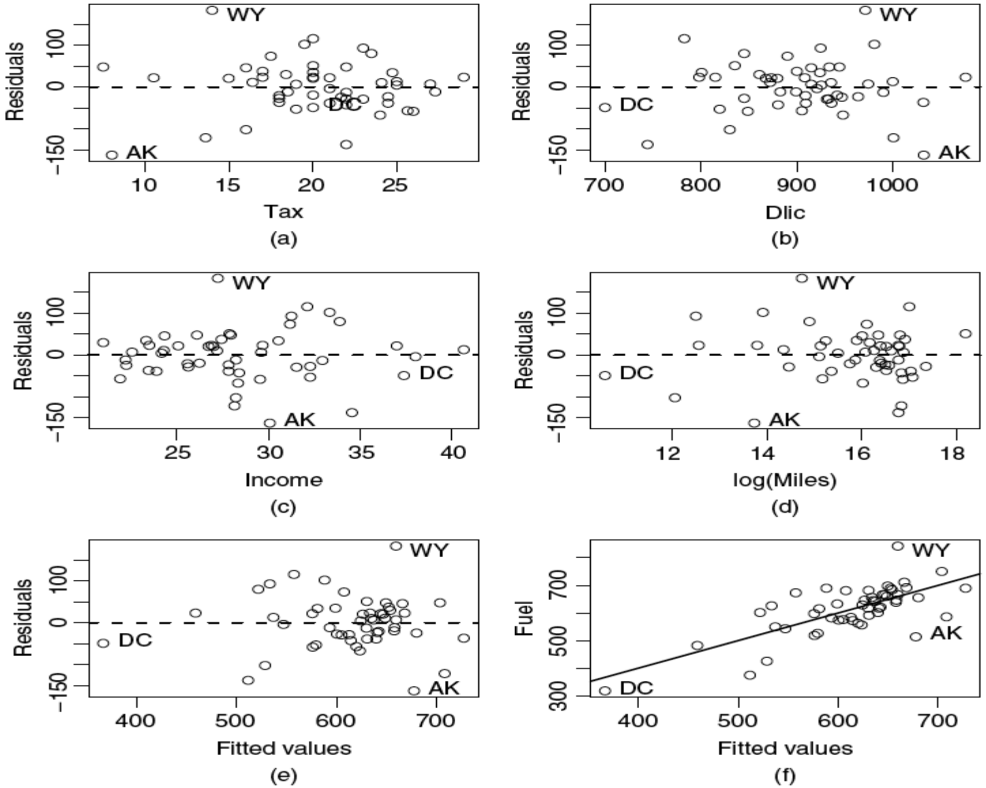
When the model is INCORRECT

- except (a), the rest residuals plots are not null (Fig 8.2)



Fuel Consumption Data

Fig 8.5



Fuel Consumption Data - con't

- three possible problematic data points:
  AK (Alaska), WY (Wyoming), DC (District of Columbia)
- WY: large but sparsely populated with a well-developed road system, people tend to drive longer for daily life
- AK: also large and sparsely populated, but road system is not good, people don't drive that much
- DC: compact urban area with good public transit
- WY and AK: possible outliers (more in next chapter)
  while DC has smaller residuals but unusual values in $\mathbf{x}_i$
- DC indeed has high leverage: $h_{ii} = 0.415$

Testing Curvature in Residual Plot

- sometimes "looking" is not enough
- a simple test for detecting curvature in residual plots
- test $\hat{e}$ versus U, where U can be any terms, combination of terms, or fitted values:
  1. refit the data with the original model + U^2
  2. test the significance of the coefficient of U^2
- if U does not depend on any estimated coefficients (like one of the terms), use t-test
- otherwise (like fitted value), use approximate z-test, called “Tukey’s test for non-additivity”.

Testing for Curvature - con't

TABLE 8.1 Significance Levels for the Lack-of-Fit Tests for the Residual Plots in Figure 8.5

Term	Test Stat.	$\Pr(>\|t\|)$
Tax	−1.08	0.29
Dlic	−1.92	0.06
Income	−0.09	0.93
log(Miles)	−1.35	0.18
Fitted values	−1.45	0.15

- obtained by R function: residualPlots(...)

Nonconstant Variance

- residual plots often show this issue
- many ways to fix this problem, and you will see two
- one option: do WLS
- it's not the simple case with $w_i = n_i$ any more,
  the challenge is how to determine the weights
- another option: variance stabilizing transformation
- our usual model: $\mathrm{Var}(Y|X = \mathbf{x}) = \sigma^2$
- now we have $\mathrm{Var}(Y|X = \mathbf{x}) = \sigma^2 g(\mathrm{E}(Y|X = \mathbf{x}))$
- where $g(\cdot)$ is an increasing (or decreasing) function

Variance Stabilizing Transform

- three common transforms:

$$\sqrt{Y}, \quad \log(Y), \quad \frac{1}{Y}$$

- (actually power transform)
- $\log(Y)$: most common, usually when response is counts or prices
- $\sqrt{Y}$: mild, when $\log$-transform is too much
- Y^{-1}: typically for "time to an event", like "time to heal after surgery"

$E(Y|X) = \mu_x$

e.g. $Var(Y|X) = \mu_x \sigma^2$

$$Var(\sqrt{Y}) = \left(\frac{1}{2\sqrt{\mu_x}}\right)^2 Var(Y) = \frac{1}{4\mu_x} Var = \frac{1}{4\mu_x} g(\mu_x)\sigma^2 = \frac{1}{4\mu_x}\mu_x\sigma^2 = \frac{1}{4}\sigma^2$$

In the linear model: $E(Y|X) = \mu_x = f(\beta_0 + \beta_1 X)$ — f: a monotone differentiable function

$f^{-1}(\mu_x) = \beta_0 + \beta_1 X$

Probability: $E(Y|X) = \mu_x = \dfrac{\exp(\beta_0+\beta_1 X)}{1+\exp(\beta_0+\beta_1 X)}$

$\boxed{\text{This part is an extension}}$

$\log \dfrac{\mu_x}{1-\mu_x} = \beta_0 + \beta_1 X$

Some clarifications:

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

$$\sum_{i=1}^{n} h_{ii} = \overset{\text{trace}}{tr}(H) = tr(\underbrace{X}_{n\times(p+1)}\underbrace{(X'X)^{-1}}_{(p+1)\times(p+1)}\underbrace{X'}_{(p+1)\times n}) = tr(\underbrace{(X'X)^{-1}X'X}_{(p+1)\times(p+1)}) = p+1$$

* $tr(H) = rank(H) = p+1$ here since H is idenpotent.

$H^2 = H \iff Q\Lambda Q'Q\Lambda Q' = Q\Lambda Q'$

can write symmetric matrix $H = Q\Lambda Q'$

$\iff Q\Lambda^2 Q' = Q\Lambda Q'$

$\Rightarrow \Lambda^2 = \Lambda \Rightarrow \lambda^2 = \lambda$

$\Rightarrow \lambda = 0 \text{ or } 1$

$\Rightarrow Tr(H) = 1$