应用回归分析

上海财经大学　统计与管理学院

第八章残差分析

章节概括：

- 残差
- 残差检验
- 方差稳定变换
- 异方差函数检验
- 模型检验

残差估计

- 线性模型：

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

$$\mathrm{E}(\mathbf{e}) = \mathbf{0} \qquad \mathrm{Var}(\mathbf{e}) = \sigma^2 \mathbf{I}_n$$

- 最小二乘估计

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

- 回归值

$$\hat{\mathrm{Y}} = \mathrm{X}\hat{\beta} = \mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{Y}$$

- 帽子矩阵

$$\mathrm{H} = \mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'$$

残差估计

- 残差估计

$$\begin{aligned}\hat{\mathbf{e}} &= \mathbf{Y} - \hat{\mathbf{Y}} \\ &= \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}} \\ &= \mathbf{Y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \\ &= (\mathbf{I} - \mathbf{H})\mathbf{Y}\end{aligned}$$

- 对比误差

$$\mathrm{E}(\hat{\mathbf{e}}) = \mathbf{0}$$

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

$$\hat{\mathbf{e}}'\mathbf{1} = \sum \hat{e}_i = 0 \qquad \mathrm{Var}(\hat{e}_i) = \hat{\sigma}^2(1 - h_{ii})$$

帽子矩阵

- 对称

$$\mathbf{HX} = \mathbf{X} \qquad (\mathbf{I} - \mathbf{H})\mathbf{X} = \mathbf{0}$$

$$\mathbf{HH} = \mathbf{H}^2 = \mathbf{H}$$

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

- X 列向量的正交投影

$$h_{ij} = \mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_j = \mathbf{x}_j'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i = h_{ji}$$

帽子矩阵

- $$\sum_{i=1}^{n} h_{ii} = p' \qquad \sum_{i=1}^{n} h_{ij} = \sum_{j=1}^{n} h_{ij} = 1$$

- 杠杆 （leverage）

$$\mathrm{Var}(\hat{e}_i) = \hat{\sigma}^2(1 - h_{ii})$$

$$\hat{y}_i = \sum_{j=1}^{n} h_{ij} y_j = h_{ii} y_i + \sum_{j \neq i}^{n} h_{ij} y_j$$

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

图示

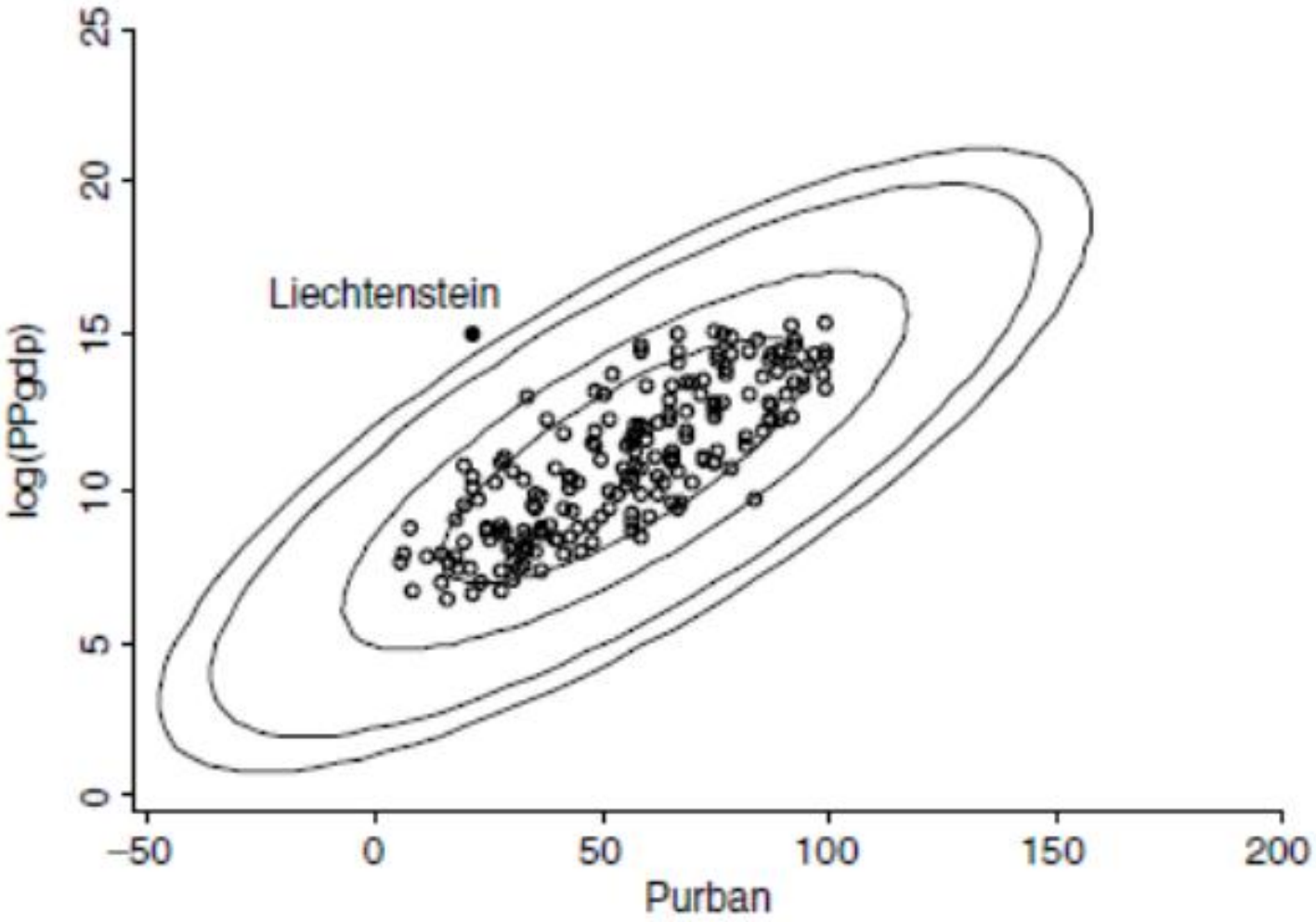


FIG. 8.1 Contours of constant leverage in two dimensions.

异方差

- $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$ $\quad\quad$ $\mathrm{Var}(\mathbf{e}) = \sigma^2\mathbf{W}^{-1}$

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

- 帽子矩阵 $\quad \mathbf{H} = \mathbf{W}^{1/2}\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}^{1/2}$
- 残差

$$y_i - \hat{\boldsymbol{\beta}}'\mathbf{x}_i$$

- 加权残差

$$\hat{e}_i = \sqrt{w_i}(y_i - \hat{\boldsymbol{\beta}}'\mathbf{x}_i)$$

$$RSS(\boldsymbol{\beta}) = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{W}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$$

$$= \sum w_i(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2$$

模型正确

- $\mathrm{E}(\hat{\mathbf{e}}|U) = 0$

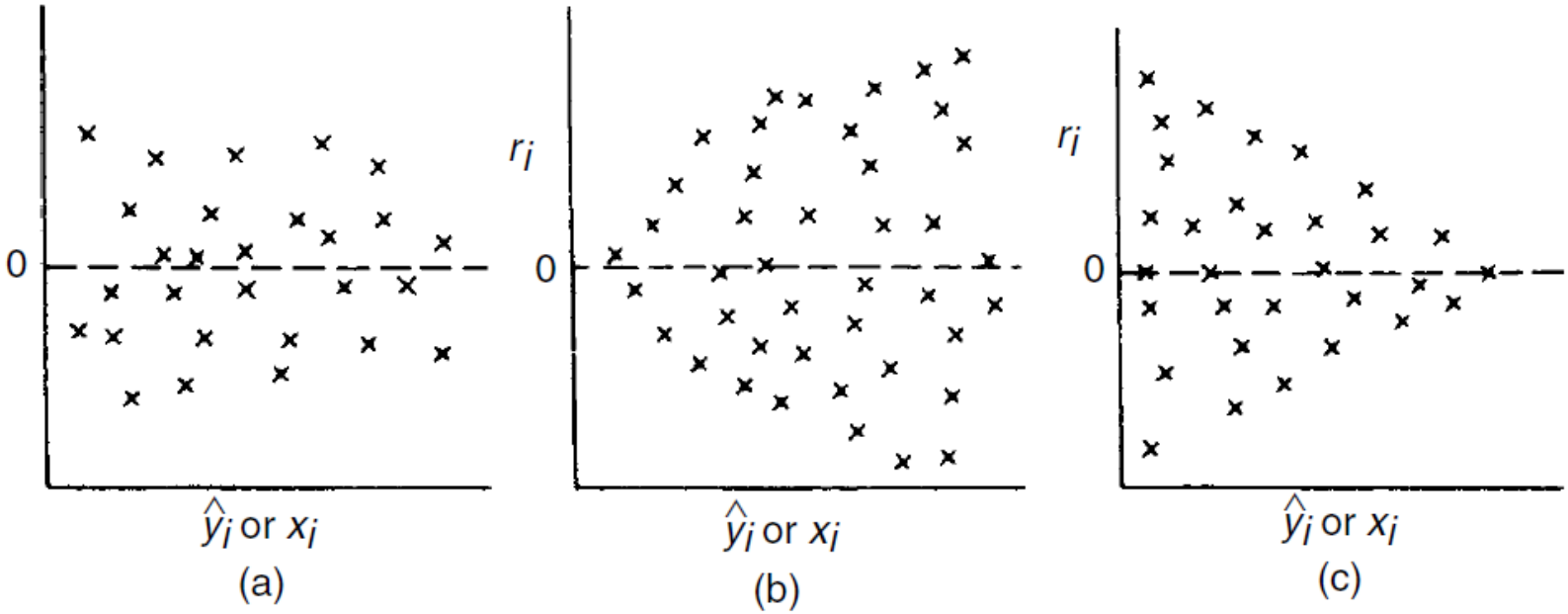
  U是自变量的线性组合

- $\mathrm{Var}(\hat{\mathbf{e}}|U) = \sigma^2(1 - h_{ii})$

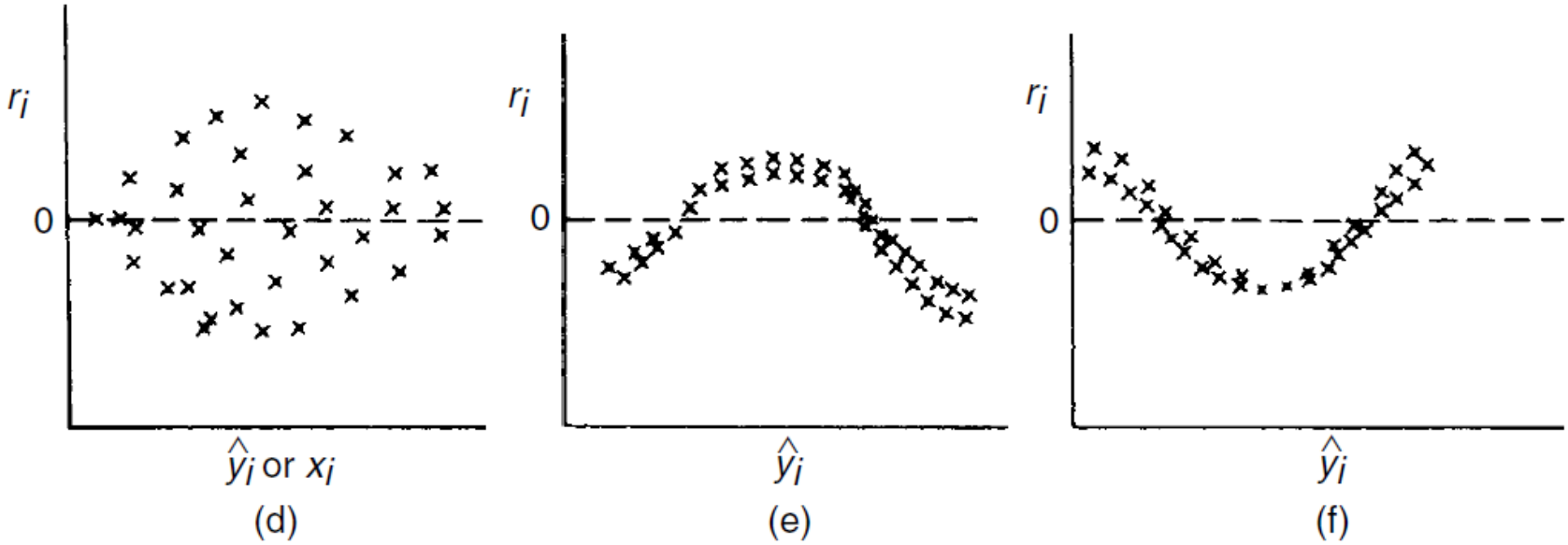
  异方差，高杠杆小方差

- 残差不独立

图例



图例



图例

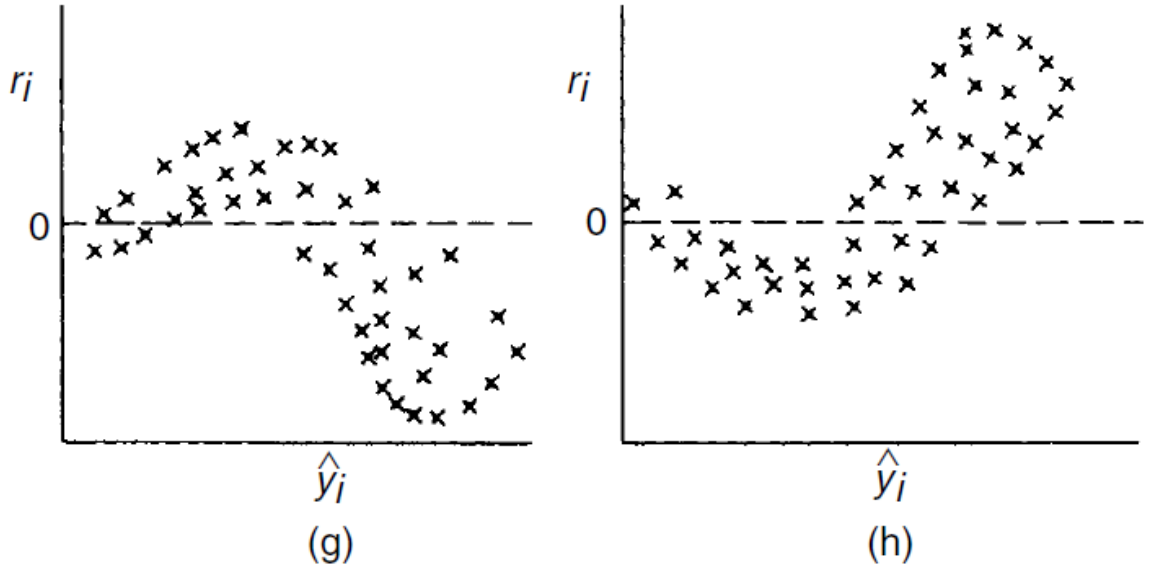


FIG. 8.2 Residual plots: (a) null plot; (b) right-opening megaphone; (c) left-opening megaphone; (d) double outward box; (e)–(f) nonlinearity; (g)–(h) combinations of nonlinearity and nonconstant variance function.

图例

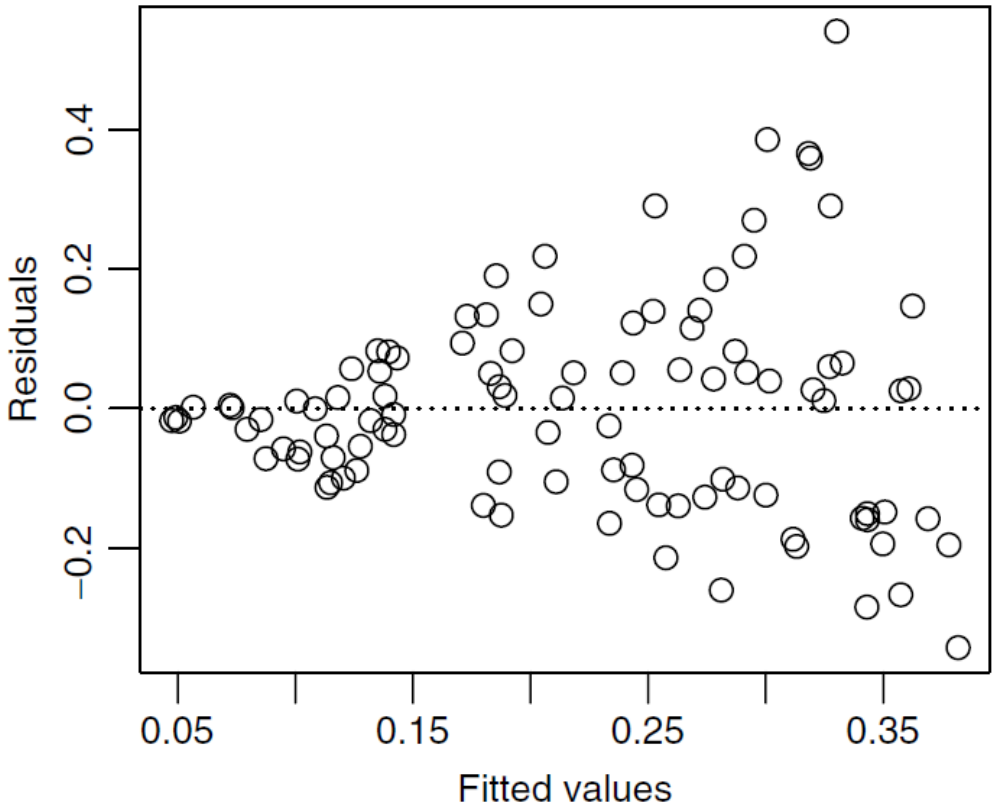


FIG. 8.3 Residual plot for the caution data.

$$\mathrm{E}(Y|X=\mathbf{x}) = \frac{|x_1|}{2+(1.5+x_2)^2}$$

油耗例子

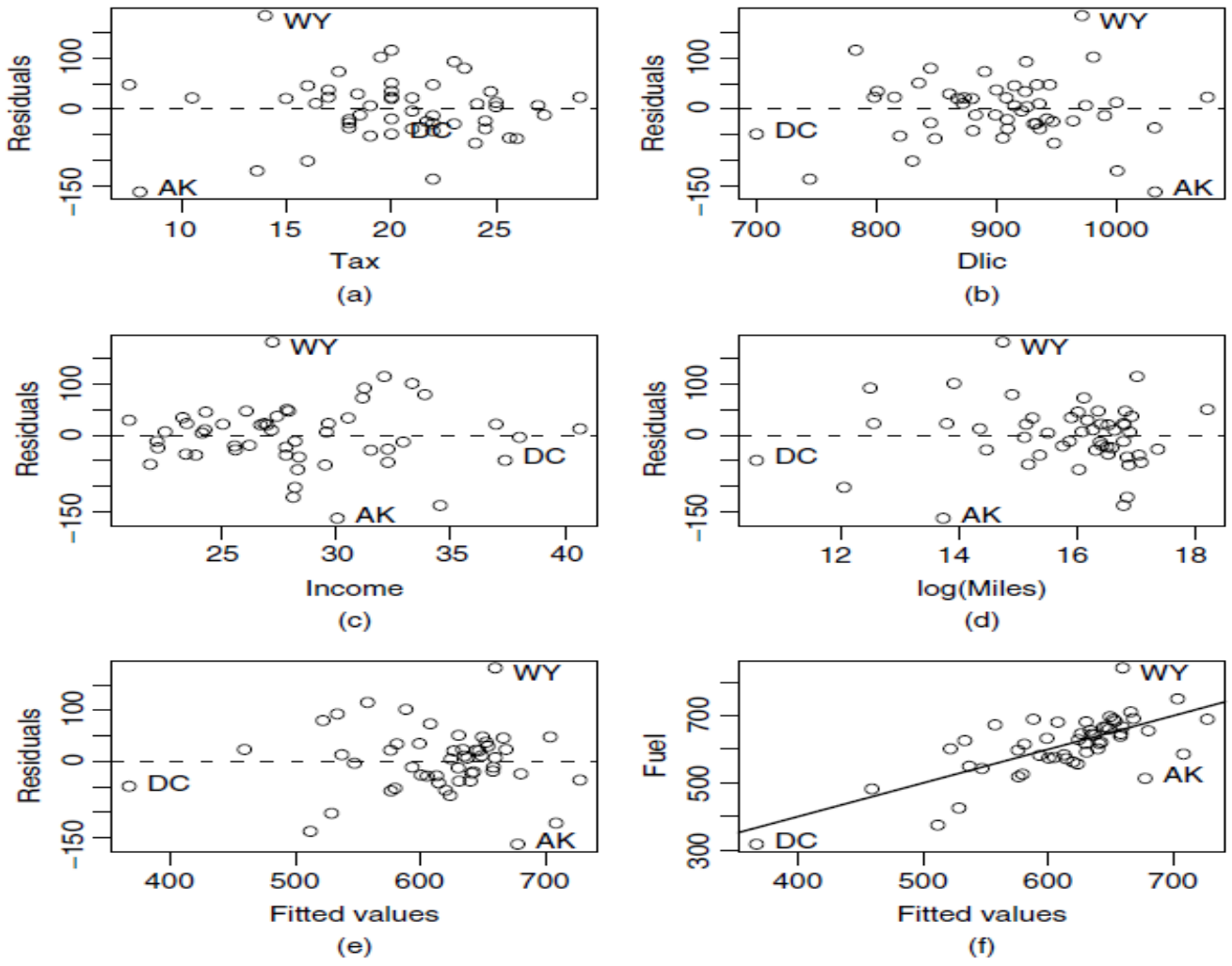


FIG. 8.5 Residual plots for the fuel consumption data.

实际原因

- Wyoming, 面积大，人口稀疏，交通发达
- Alaska, 面积大，人口稀疏， 交通不发达
- DC, 面积小，公共交通发达，车用少

高杠杆

$$h_{9,9} = 0.415$$

- 下章待续
- （e）有非线性的可能

残差检验

- 散点图: $\hat{e}$ 和 U
- 再拟合，加入 U^2

- 若U不依赖于估计的系数，T-test
- 若U依赖于估计的系数，如拟合值，Z-test (Tukey's 检验非可加性)

残差检验

- 油耗例子

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

残差检验

- UN数据

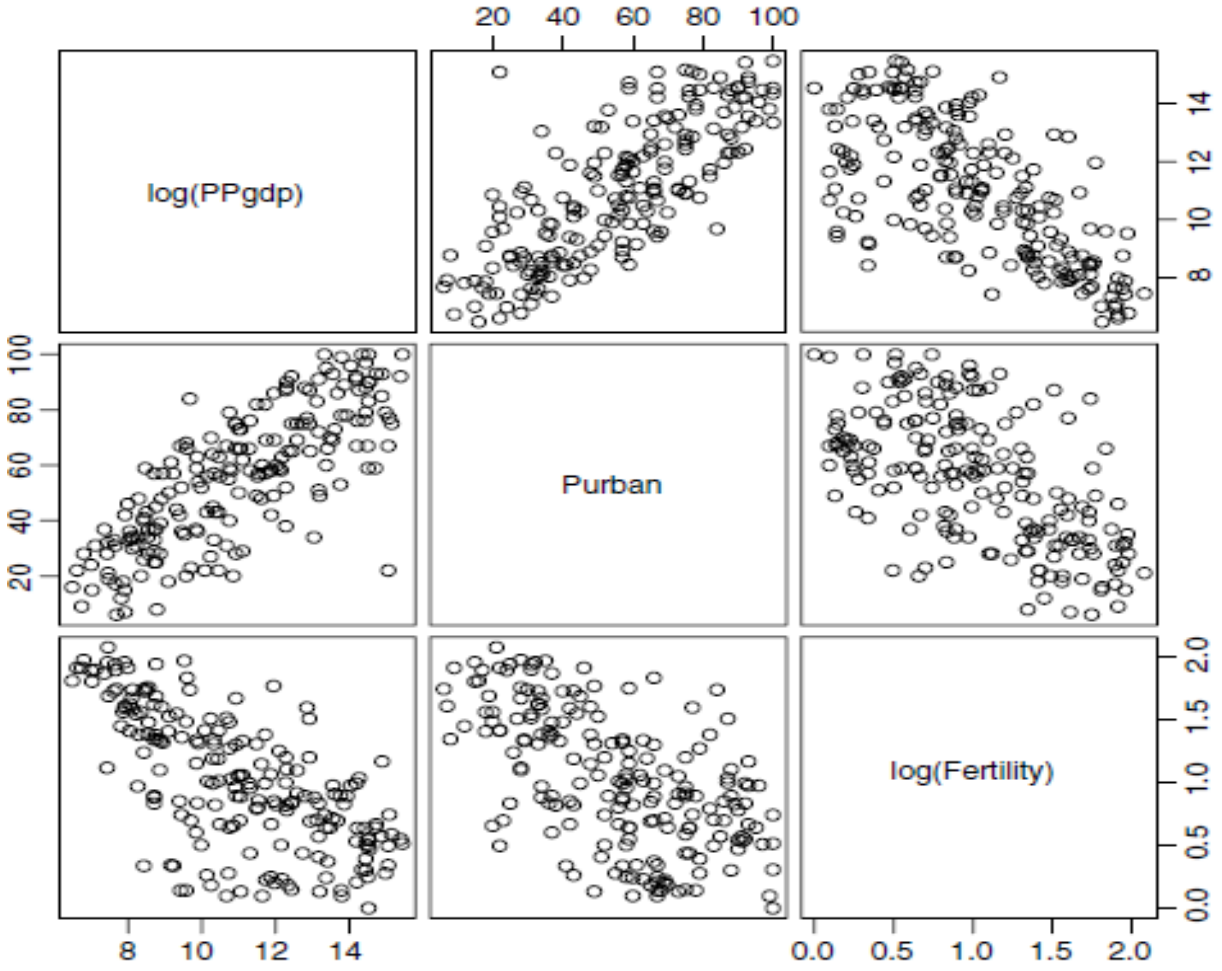


FIG. 8.6 Scatterplot matrix for three variables in the UN data.

残差检验

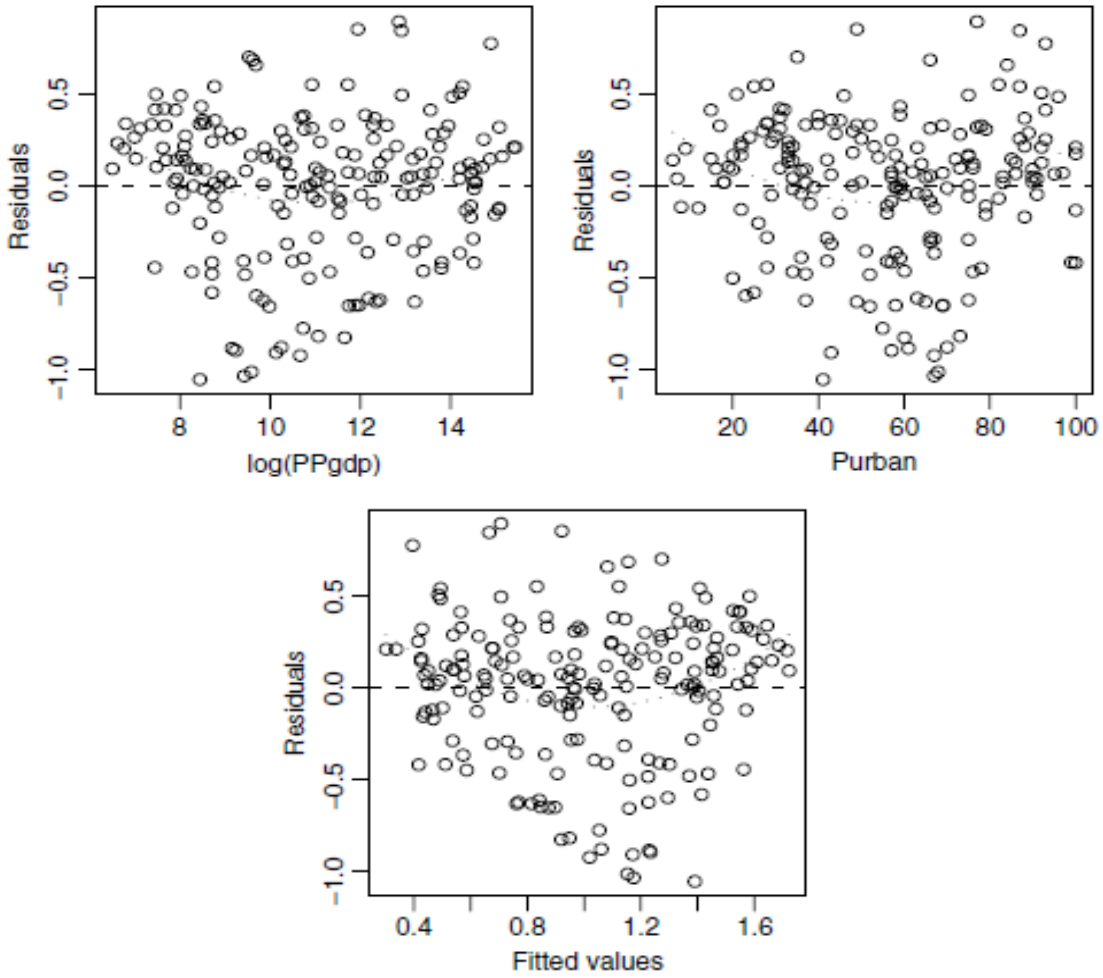


FIG. 8.7 Residual plots for the UN data. The dotted curved lines are quadratic polynomials fit to the residual plot and do not correspond exactly to the lack-of-fit tests that add a quadratic term to the original mean function.

残差检验

TABLE 8.2 Lack-of-Fit Tests for the UN Data

	Test Stat.	Pr(>|t|)
log(*PPgdp*)	3.22	0.00
Purban	3.37	0.00
Tukey test	3.65	0.00

方差稳定变换

- 均值函数/方差函数

$$\mathrm{Var}(Y|X=\mathbf{x}) = \sigma^2 g(\mathrm{E}(Y|X=\mathbf{x}))$$

- Poisson

$$g(\mathrm{E}(Y|X=\mathbf{x})) = \mathrm{E}(Y|X=\mathbf{x})$$

方差稳定变换

TABLE 8.3 Common Variance Stabilizing Transformations

Y_T	Comments
$\sqrt{Y}$	Used when $\mathrm{Var}(Y\|\mathrm{X}) \propto \mathrm{E}(Y\|\mathrm{X})$, as for Poisson distributed data. $Y_T = \sqrt{Y} + \sqrt{Y+1}$ can be used if all the counts are small (Freeman and Tukey, 1950).
$\log(Y)$	Use if $\mathrm{Var}(Y\|X) \propto [\mathrm{E}(Y\|X)]^2$. In this case, the errors behave like a percentage of the response, $\pm 10\%$, rather than an absolute deviation, ± 10 units.
$1/Y$	The inverse transformation stabilizes variance when $\mathrm{Var}(Y\|X) \propto [\mathrm{E}(Y\|X)]^4$. It can be appropriate when responses are mostly close to 0, but occasional large values occur.
$\sin^{-1}(\sqrt{Y})$	The *arcsine square-root* transformation is used if Y is a proportion between 0 and 1, but it can be used more generally if y has a limited range by first transforming Y to the range (0, 1), and then applying the transformation.

异方差判定

- 假定： $\mathrm{Var}(Y|X, Z = \mathbf{z}) = \sigma^2 \exp(\boldsymbol{\lambda}'\mathbf{z})$

- Z已知，可以是Y，X...
- $\boldsymbol{\lambda}$ 未知

- 非负性、单调性
- 同方差即 $\boldsymbol{\lambda} = \mathbf{0}$

检验步骤

- 假设 $\boldsymbol{\lambda} = \mathbf{0}$ 即 $\mathrm{Var}(Y|X, Z = \mathbf{z}) = \sigma^2$
  回归 $\mathrm{E}(Y|X = \mathbf{x}) = \boldsymbol{\beta}'\mathbf{x}$ ，并保存残差 $\hat{e}_i$
- 计算 $$u_i = \hat{e}_i^2/\tilde{\sigma}^2 = n\hat{e}_i^2/[(n - p')\hat{\sigma}^2]$$
  $$\tilde{\sigma}^2 = \sum \hat{e}_j^2/n$$
- 回归 $\mathrm{E}(U|Z = \mathbf{z}) = \lambda_0 + \boldsymbol{\lambda}'\mathbf{z}$
  并计算 $SSreg$
- 当 $\boldsymbol{\lambda} = \mathbf{0}$ 时，
  $S = SSreg/2$ 服从 $\chi^2(q)$

检验步骤

- 若

$$\mathrm{Var}(Y|X, Z = \mathbf{z}) = \frac{\sigma^2}{w}\exp(\lambda'\mathbf{z})$$

- 零假设

$$\mathrm{Var}(Y|X, Z = \mathbf{x}) = \sigma^2/w$$

- 第一步回归用WLS代替OLS，其余相似

Snow Geese 数据

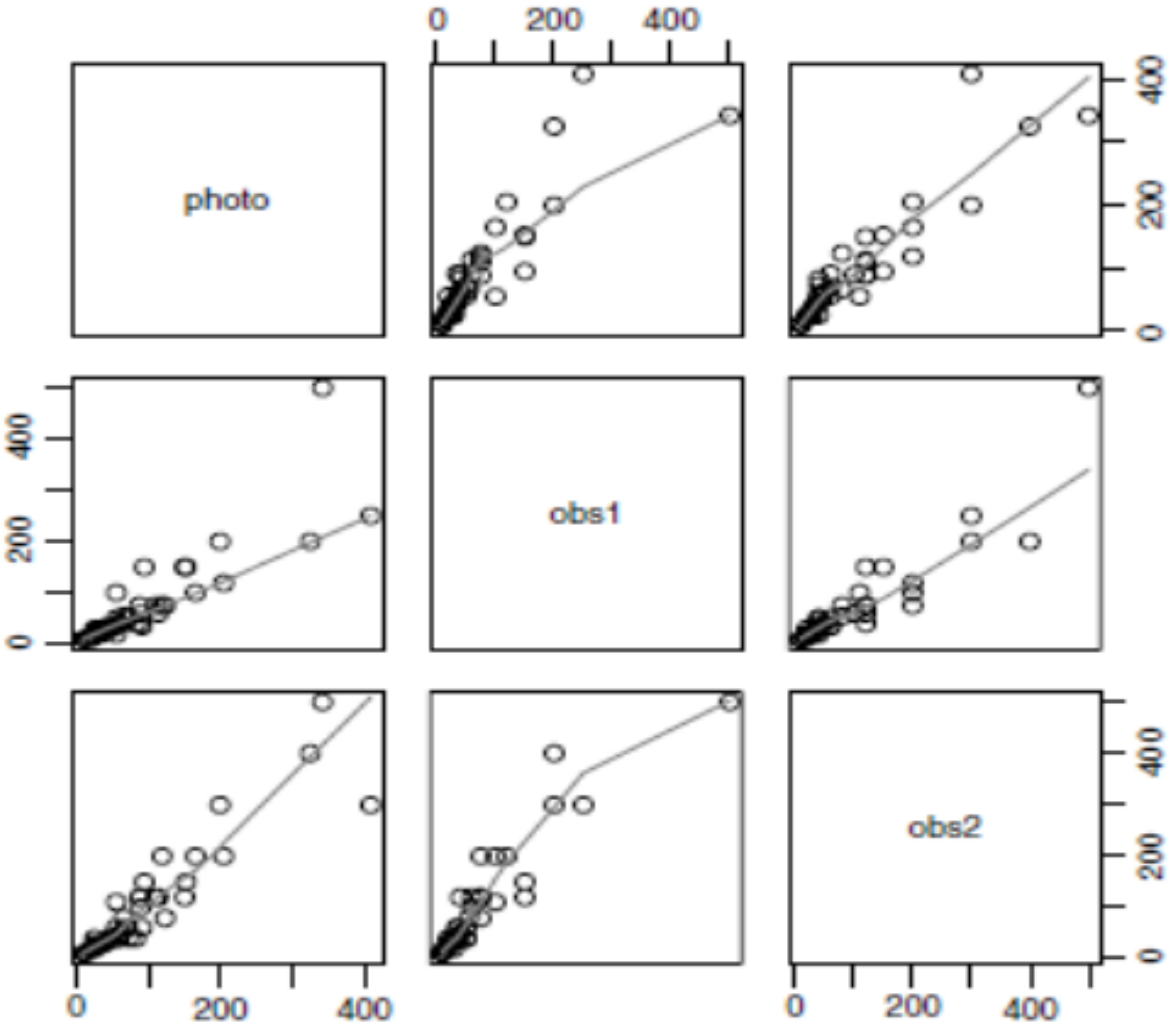


FIG. 8.8 The snow geese data. The line on each plot is a *loess* smooth with smoothing parameter 2/3.

检验结果

- photo vs obs1

$$\hat{E}(photo|obs1) = 26.55 + 0.88obs1$$

- U vs obs1

```
Response: U
            Df  Sum Sq Mean Sq F value    Pr(>F)
obs1         1 162.826 162.826  50.779 8.459e-09
Residuals   43 137.881   3.207
```

- 统计量

$$S = (1/2)SSreg = (1/2)162.83 = 81.41$$

Sniff 数据

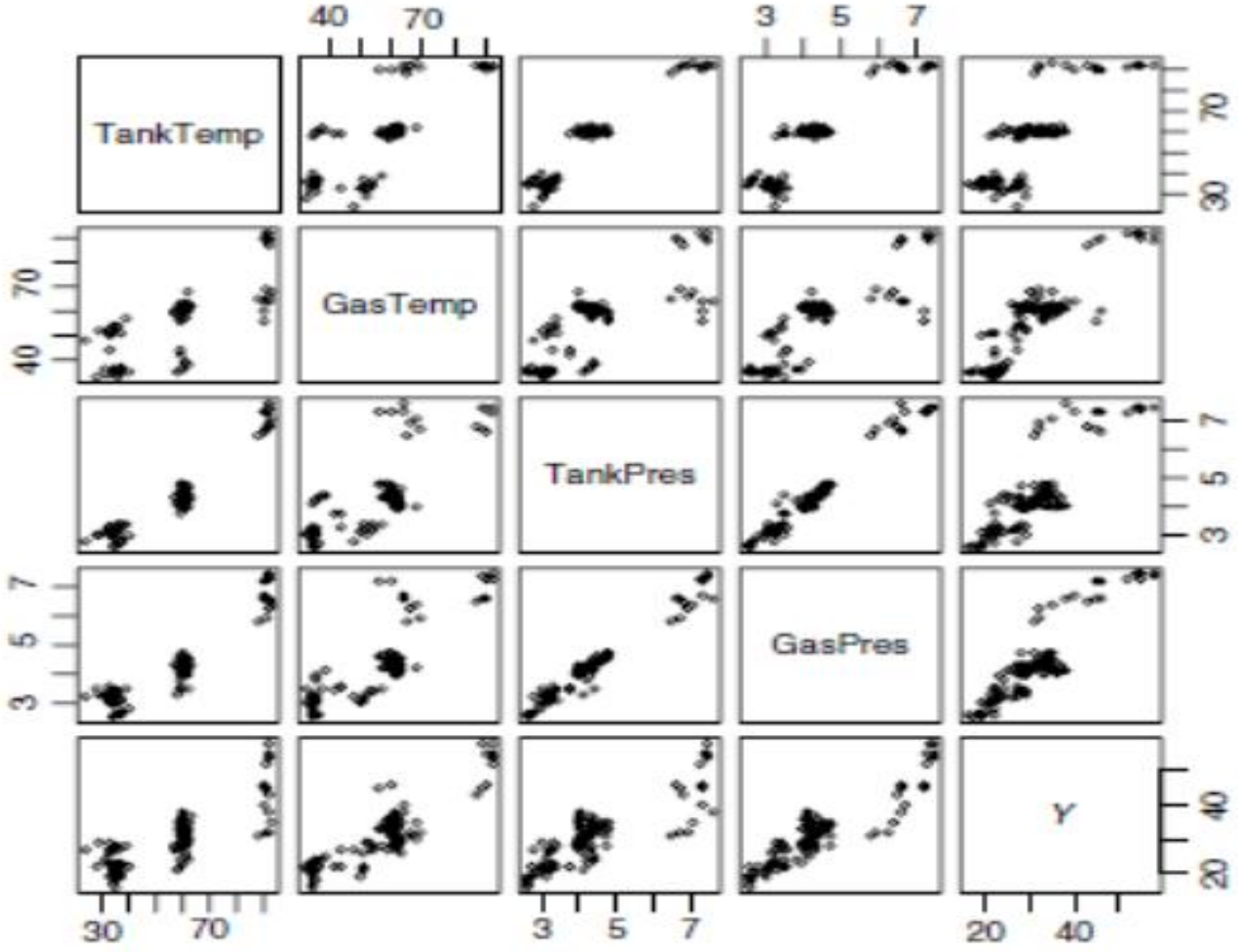


FIG. 8.9 Scatterplot matrix for the sniffer data.

异方差检验

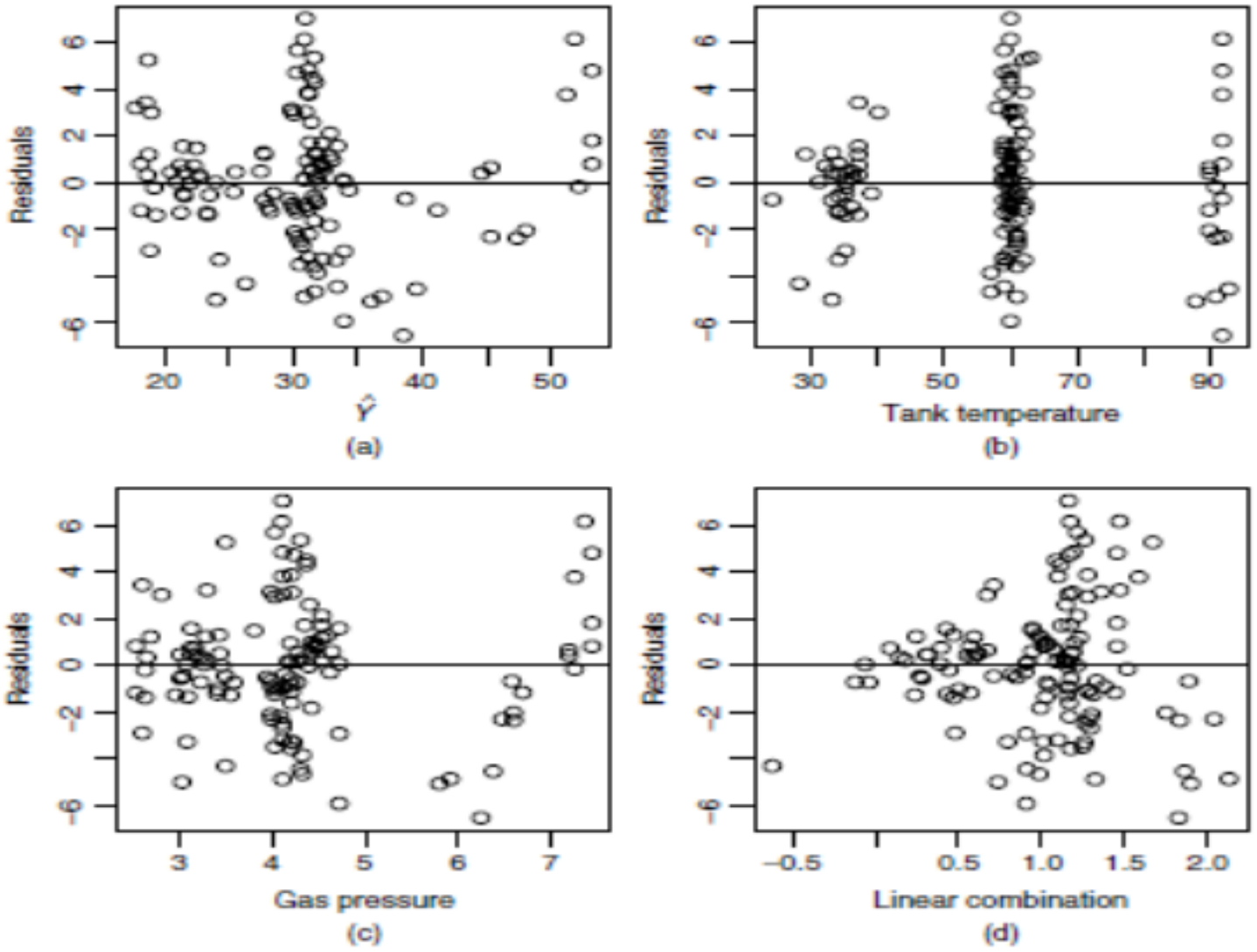


FIG. 8.10 Residuals plots for the sniffer data with variance assumed to be constant.

异方差检验

TABLE 8.4 Score Tests for Sniffer Data

Choice for Z	df	S	p-value
TankTemp	1	5.50	.019
GasPres	1	9.71	.002
TankTemp, GasPres	2	11.78	.003
TankTemp, GasTemp TankPres, GasPres	4	13.76	.008
Fitted values	1	4.80	.028

条件期望

- A.2.4

$$\mathrm{E}(Y) = \mathrm{E}[\mathrm{E}(Y|X=x)]$$

$$\mathrm{Var}(Y) = \mathrm{E}[\mathrm{Var}(Y|X=x)] + \mathrm{Var}(\mathrm{E}(Y|X=x)]$$

$$\begin{aligned}
\mathrm{E}(Y) &= \mathrm{E}[\mathrm{E}(Y|X=x)] \\
&= \mathrm{E}[\beta_0 + \beta_1 x] \\
&= \beta_0 + \beta_1 \mu_x \\
\mathrm{Var}(Y) &= \mathrm{E}[\mathrm{Var}(Y|X=x)] + \mathrm{Var}[\mathrm{E}(Y|X=x)] \\
&= \mathrm{E}[\sigma^2] + \mathrm{Var}[\beta_0 + \beta_1 x] \\
&= \sigma^2 + \beta_1^2 \tau_x^2
\end{aligned}$$

均值函数检验

- 无模型假设，令U是X的函数

  Y vs U 非线性拟和

- 有模型假设，Cook and Weisberg, A.2.4

  $$\mathrm{E}(Y|U=\mathbf{u})=\mathrm{E}[\mathrm{E}(Y|X=\mathbf{x})|U=\mathbf{u}]$$

  用 $\hat{\mathbf{Y}}$ 代替 $\mathrm{E}(Y|X=\mathbf{x})$

  $\hat{\mathbf{Y}}$ vs U非线性 相似

- $\mathrm{E}(Y|\widehat{U}=\mathbf{u})=\mathrm{E}(\hat{\mathbf{Y}}|U=\mathbf{u})$

  对比　Y vs U与 $\hat{\mathbf{Y}}$ vs U 的拟和相似

UN数据

- 模型1：

$$\mathrm{E}(\log(Fertility)|\log(PPgdp), Purban) = \beta_0 + \beta_1 \log(PPgdp) + \beta_2 Purban$$

$$U = \log(PPgdp)$$

- 模型2：

$$\mathrm{E}(\log(Fertility)|\log(PPgdp), Purban) = \beta_0 + \beta_1 \log(PPgdp) + \beta_2 Purban + \beta_{22} Purban^2$$

模型 1

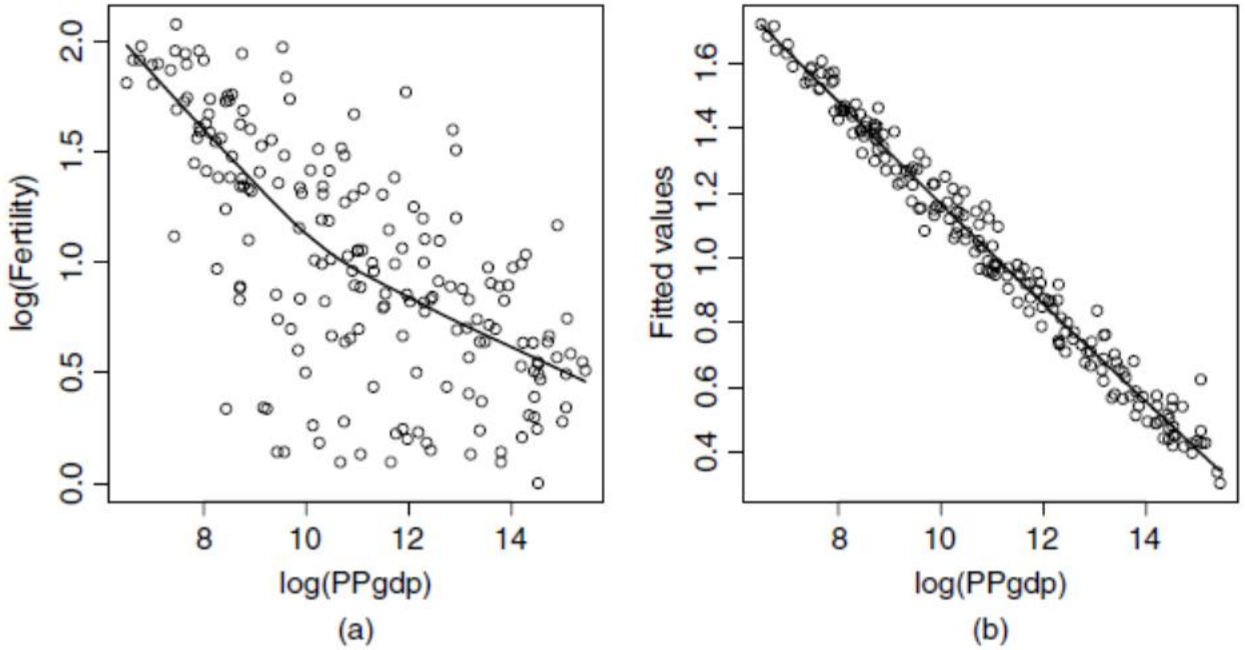


FIG. 8.12 Plots for log(*Fertility*) versus log(*PPgdp*) and $\hat{y}$ versus log(*PPgdp*). In both plots, the curves are *loess* smooths with smoothing parameters equal to 2/3. If the model has the correct mean function, then these two smooths estimate the same quantity.

模型 1

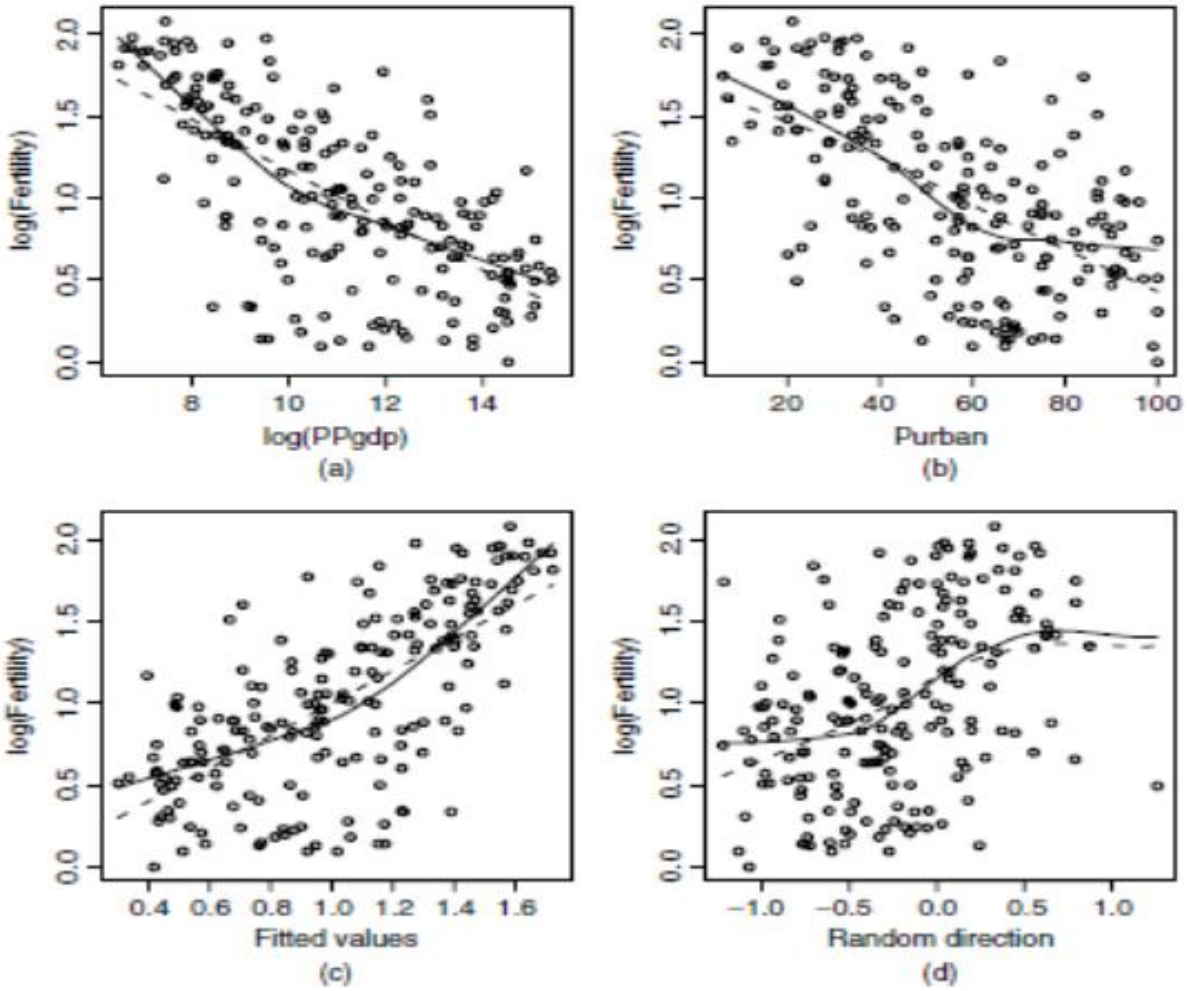


FIG. 8.13 Four marginal model plots, versus the two terms in the mean function, fitted values, and a random linear combination of the terms in the mean function.

模型 2

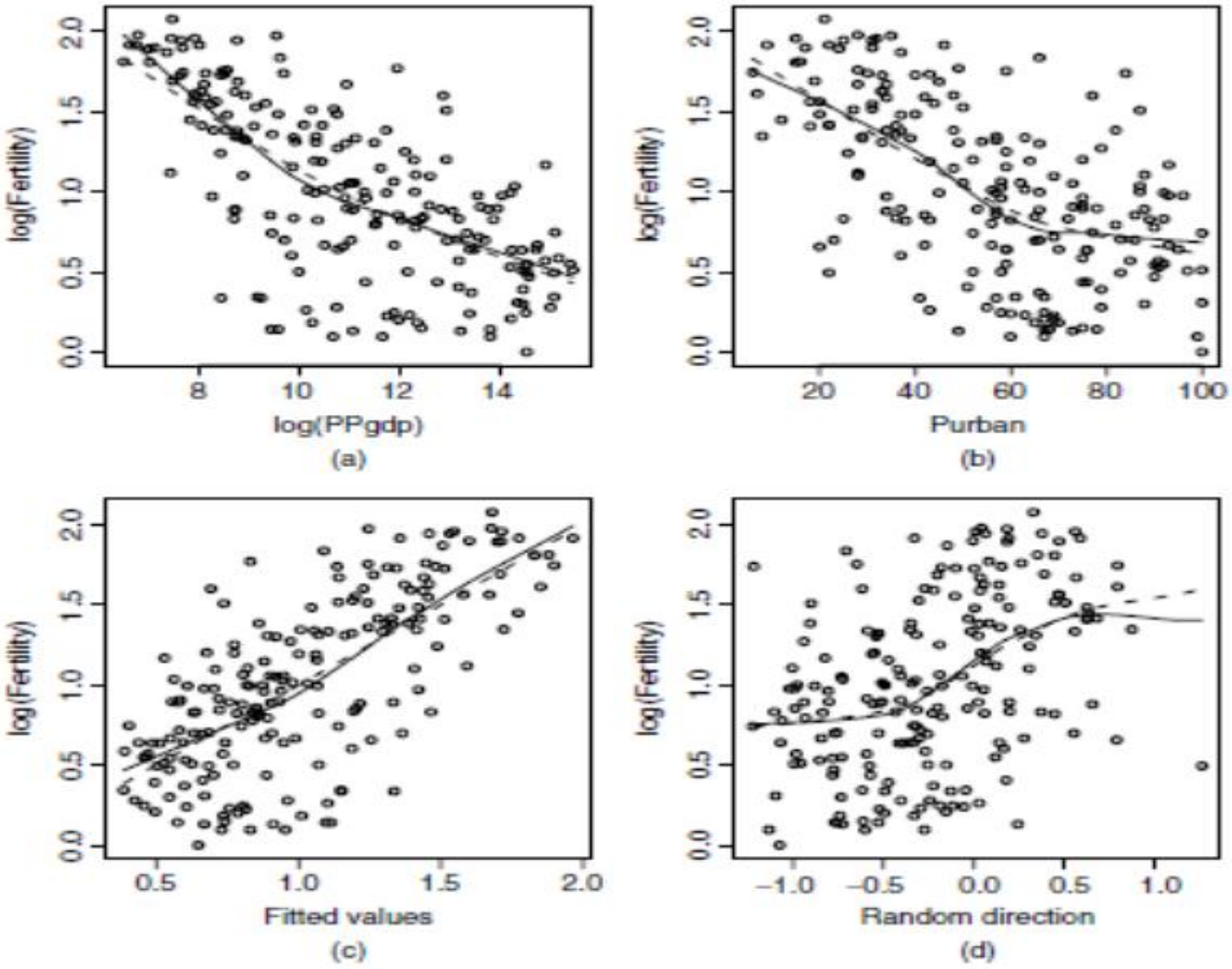


FIG. 8.14 Marginal model plots for the United Nations data, including a quadratic term in *Purban*.

方差函数检验

- 非线性回归，A.5
- 无模型假设，令U是X的函数

$$\mathrm{SD}_{data}(Y|U)$$

- 有模型假设 $\hat{\mathbf{Y}} \approx \mathrm{E}(Y|U=\mathbf{u})$

$$\mathrm{SD}_{model}(Y|U)$$

$$\begin{aligned}\mathrm{Var}(Y|U) &= \mathrm{E}[\mathrm{Var}(Y|X)|U] + \mathrm{Var}[\mathrm{E}(Y|X)|U] \\ &\approx \mathrm{E}[\sigma^2|\mathrm{U}] + \mathrm{Var}[\hat{\mathbf{Y}}|\mathrm{U}] \\ &= \sigma^2 + \mathrm{Var}[\hat{\mathbf{Y}}|\mathrm{U}]\end{aligned}$$

- 对比 $\mathrm{SD}_{data}(Y|U) = \mathrm{SD}_{model}(Y|U)$

UN数据

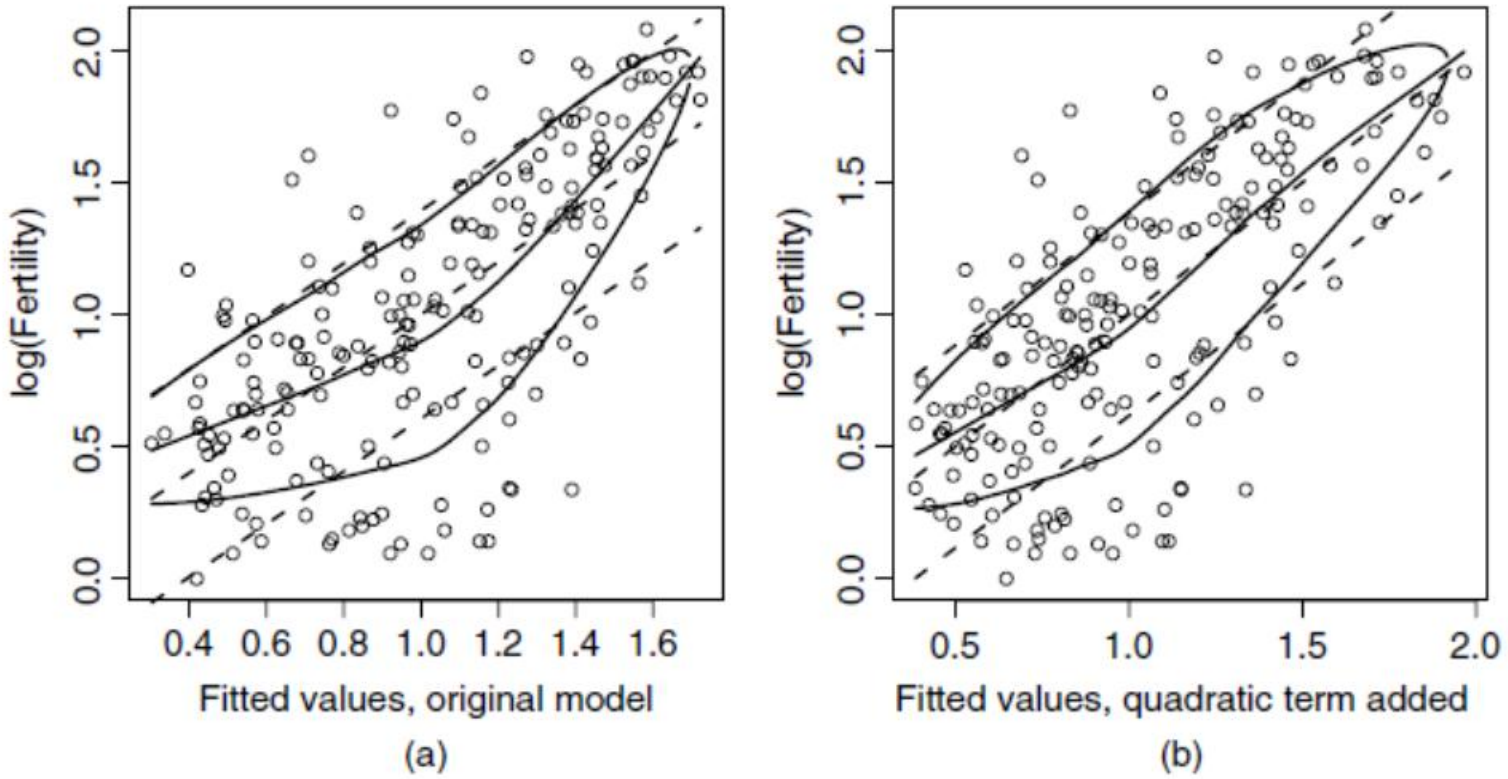


FIG. 8.15 Marginal model plots with standard deviation smooths added. (a) The fit of (8.21). (b) The fit of (8.22).

Thank You !