Model Adequacy Checking – Part I

Johns Hopkins Engineering

625.461 Statistical Models and Regression

Module 7 – Lecture 7B

Assumptions Underlying Linear Regression Analysis

The major assumptions that we have made thus far in linear regression analysis are:

1) Relationship between y and regressors $\mathbf{x}$ is linear
2) The error term ε has zero mean
3) ε has constant variance σ^2
4) The errors are uncorrelated
5) The errors are normally distributed

Residuals

$$e_i = y_i - \hat{y}_i \qquad i = 1, \ldots, n$$

Plotting residuals is a very effective way to investigate how well the regression model fits the data and to check the assumptions given on slide #2

Properties of Residuals

The residuals have zero mean and approximate average variance estimated by

$$\frac{\sum_{i=1}^{n}(e_i-\bar{e})^2}{n-p}=\frac{\sum_{i=1}^{n}e_i^2}{n-p}=\frac{SS_{\text{Res}}}{n-p}=MS_{\text{Res}}$$

Properties of Residuals

The residuals are **not** independent, as the n residuals have only $(n - p)$ degrees of freedom.

The nonindependence has little effect on their use for model adequacy checking, if n is not small relative to p .

Scaling Residuals

Methods of *scaling residuals* (helpful in finding observations that are outliers or extreme values)

1. Standardized Residuals

$$d_i = \frac{e_i}{\sqrt{MS_{\text{Res}}}}, \quad i = 1, 2, \ldots, n$$

Var(d_i) ≈ 1. So a large residual ($d_i > 3$, say) potentially indicates an outlier.

Scaling Residuals

2. Studentized Residuals

Improve the residual scaling by dividing e_i by the exact standard deviation of the ith residual.

$$\mathbf{e} = (\mathbf{I} - \mathbf{H})\mathbf{y} \qquad \mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

$$\mathbf{e} = (\mathbf{I} - \mathbf{H})(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = \mathbf{X}\boldsymbol{\beta} - \mathbf{H}\mathbf{X}\boldsymbol{\beta} + (\mathbf{I} - \mathbf{H})\boldsymbol{\varepsilon}$$

$$= \mathbf{X}\boldsymbol{\beta} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + (\mathbf{I} - \mathbf{H})\boldsymbol{\varepsilon} = (\mathbf{I} - \mathbf{H})\boldsymbol{\varepsilon}$$

$$\mathrm{Var}(\mathbf{e}) = \mathrm{Var}[(\mathbf{I} - \mathbf{H})\boldsymbol{\varepsilon}] = (\mathbf{I} - \mathbf{H})\,\mathrm{Var}(\boldsymbol{\varepsilon})(\mathbf{I} - \mathbf{H})' = \sigma^2(\mathbf{I} - \mathbf{H})$$

Properties of Residuals

The residuals have different variances and they are correlated, i.e.,

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

$$\mathrm{Cov}(e_i, e_j) = -\sigma^2 h_{ij}$$

Since $0 \leq h_{ii} \leq 1$, using MS_{Res} to estimate the variance of the residuals actually overestimates $\mathrm{Var}(e_i)$.

Properties of Residuals

Since h_{ii} is a measure of the location of the ith point in $\mathbf{x}$ space, the variance of e_i depends on where the point $\mathbf{x}_i$ lies.

Points near the center of the $\mathbf{x}$ space have larger variance than residuals at more remote locations. Violations of model assumptions are more likely at remote points, and these violations may be hard to detect from inspection of e_i (or d_i) because they will be smaller.

Leverage Point

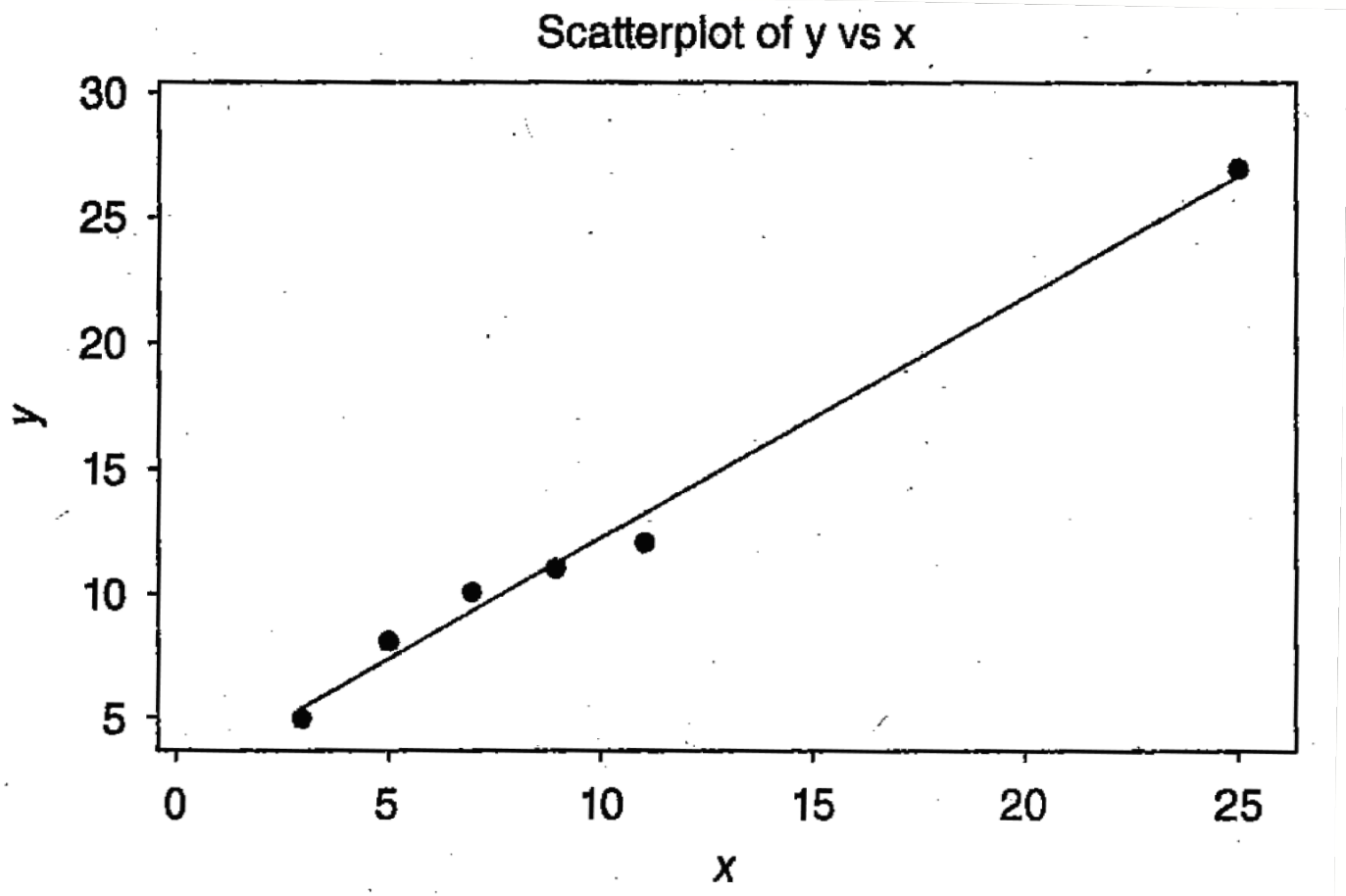

Figure 4.1 Example of a pure leverage point.

Observed value for the response at leverage point is consistent with the prediction based on the other data values

Influential Point

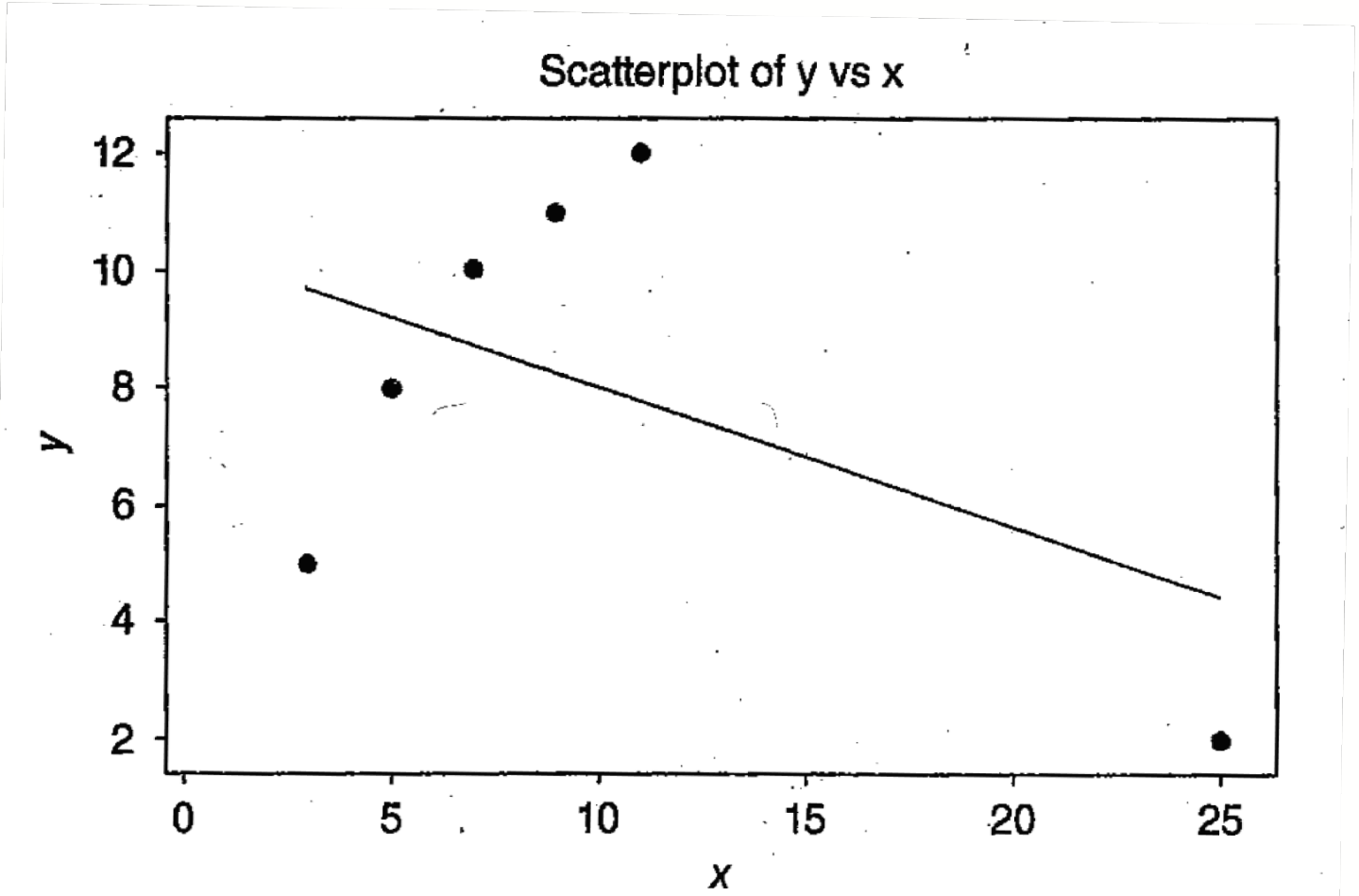

Figure 4.2 Example of an influential point.

Influential point draws the prediction equation to itself

JOHNS HOPKINS

WHITING SCHOOL
of ENGINEERING