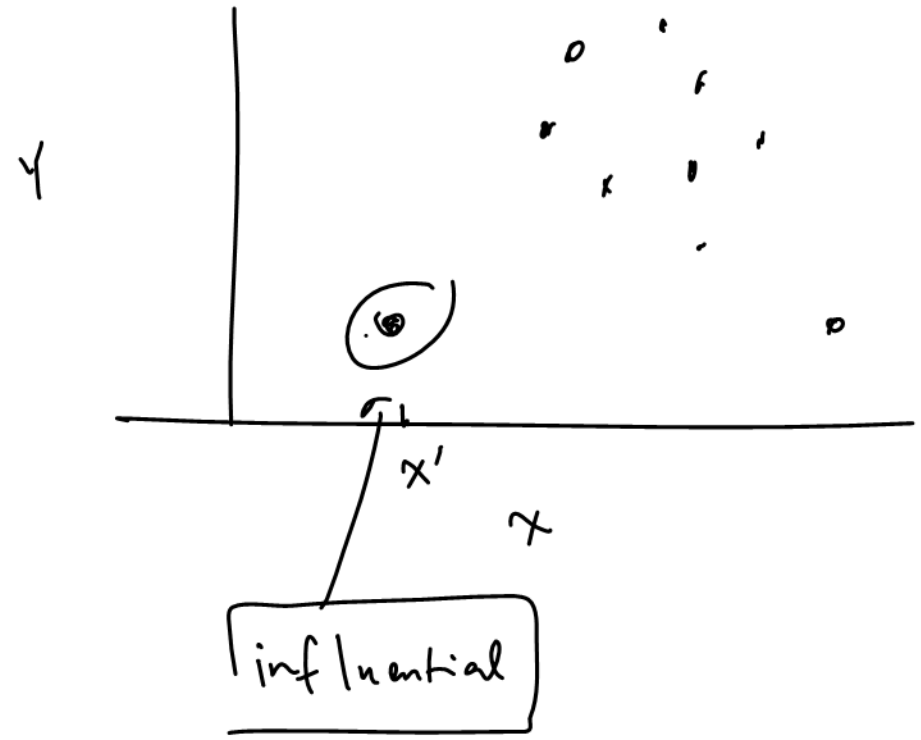
Lecture 09 Oct 11

Outlier & Influential Observations



$\hat{\beta}_1$ (complete) vs $\hat{\beta}_1$ (without (x', Y'))

The estimator $\hat{\beta}_1$:

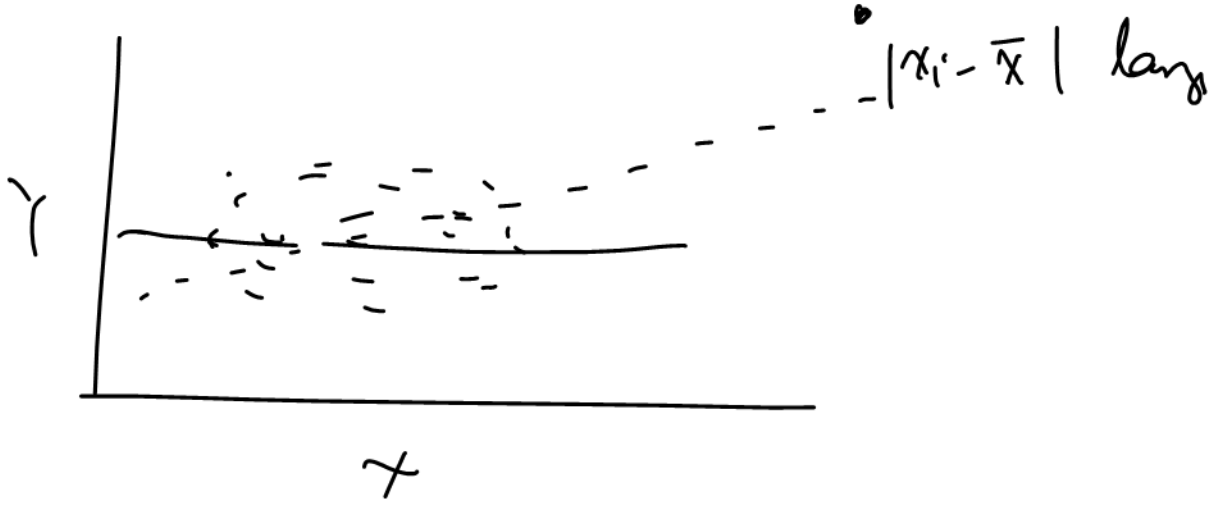
$$\hat{\beta}_1 = \sum w_i Y_i$$

$$= \sum \left[\frac{(x_i - \bar{x})}{\sum (x_j - \bar{x})^2} \right] \cdot Y_i$$

For values $\{x_i\}$ where x_i close to $\bar{x}$

$|x_i - \bar{x}| < \delta \Rightarrow w_i$ small

when $x_i - \bar{x}$ large (in magnitude) then y_i has more weight in estimating β_1.



Let $(x_i, y_i) = z_i$ be a candidate outlier/influential observation.

Let $\mathcal{D} = \{(x_n, y_n),\ n = 1, \ldots, N\}$

$\mathcal{D}_{(i)} = \mathcal{D} - \{(x_i, y_i)\}$

$|\hat{\beta}_1(\mathcal{D}) - \hat{\beta}_1(\mathcal{D}_{(-i)})|$

Multi-collinearity

$$X = \begin{pmatrix} \mathbb{1} & x_1 \\ & x_2 \\ & \vdots \\ & x_n \end{pmatrix}$$

Suppose that $x_i = 10$

$$\begin{pmatrix} \mathbb{1} & 10 \\ & \vdots \\ & 10 \end{pmatrix}$$

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} \equiv \frac{1}{ad-bc}\begin{pmatrix} d & -b \\ -a & \end{pmatrix}$$

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

$$(X'X) = \begin{pmatrix} n & 10n \\ 10n & n(10^2) \end{pmatrix} \qquad (X'X)^{-1} = \frac{1}{\det(X'X)}\begin{pmatrix} & \end{pmatrix}$$

$$\det(X'X) = n^2(10^2) - (10n)^2 = 0.$$

Ridge Regression

If $(X'X)$ is near singular, i.e., $\det(X'X) \simeq 0$

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

$$\hat{\beta}_R = \left[(X'X) + \lambda \cdot I\right]^{-1}X'\underline{Y}, \qquad \lambda > 0$$

Diagnostics

Model $Y_i | x_i \overset{\text{indep}}{\sim} N(\mu(x_i), \sigma^2)$

$$\mu(x_i) = \beta_0 + \beta_1 x_i \leftarrow$$

(1) Is linearity sufficient?
Is the linear structure reasonable?

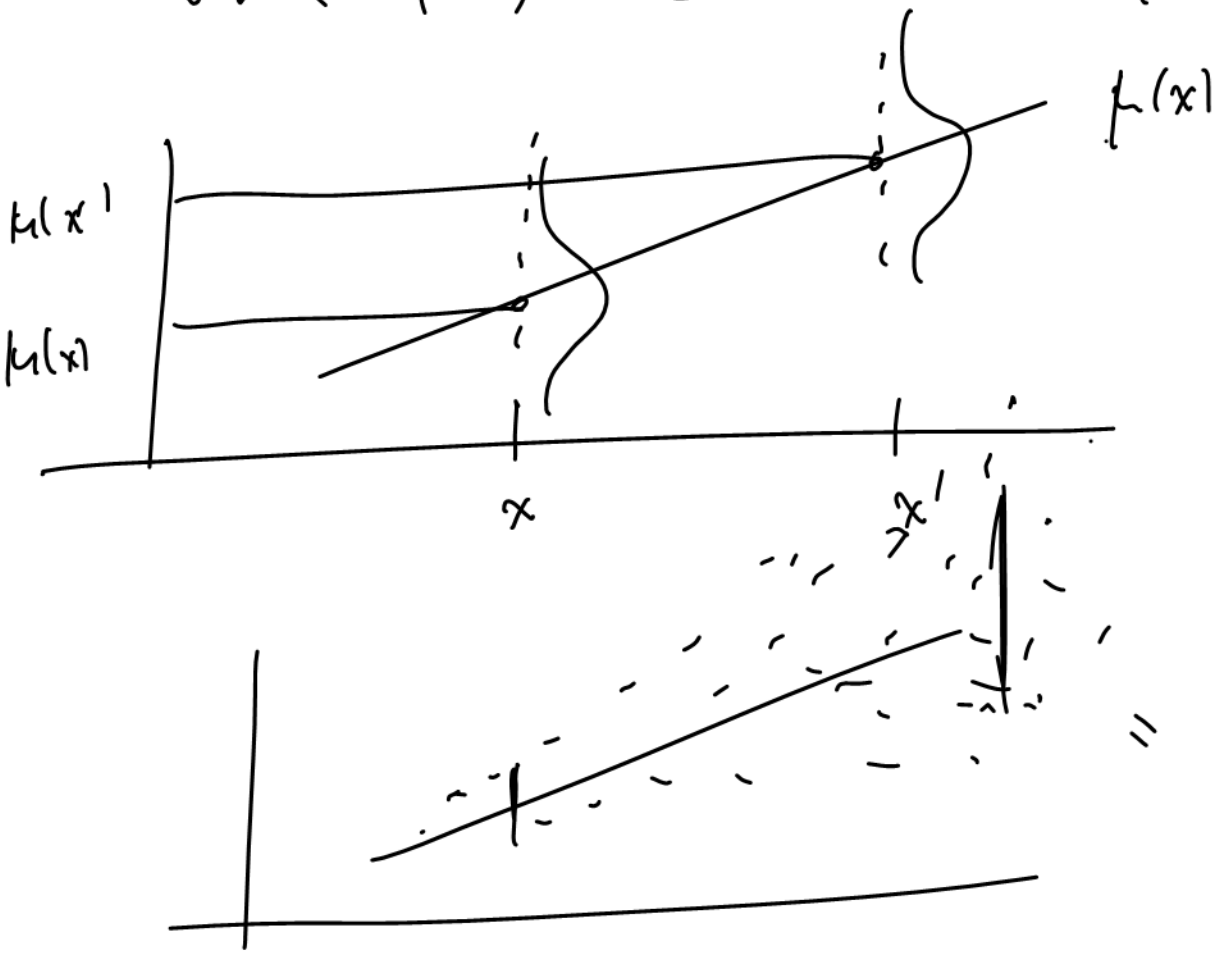
(2) Normality

(3) Equal variance across all x values
Homoscedasticity
Heteros. - -

(4) Independence

According to this model: let x and x' be two values of the indep variable. Then:

$$Var(Y|x) = Var(Y|x')$$

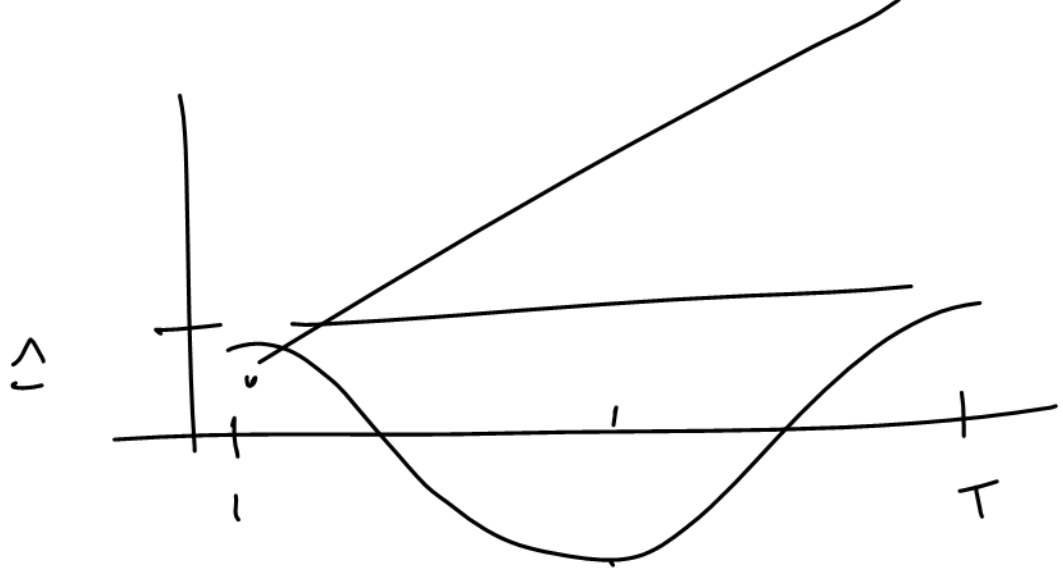


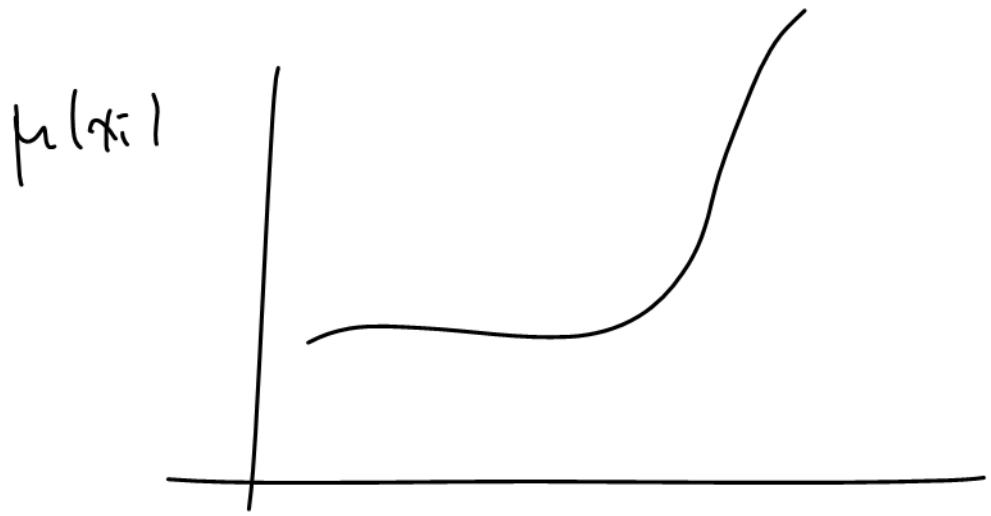
Diagnostics on the mean structure

Suppose that this is the true model:

$$Y_i | x_i \sim N(\mu(x_i), \sigma^2) \qquad x_i = 1, 2, \ldots, T$$

$$\mu(x_i) = \beta_0 + \beta_1 x_i + \alpha_1 \cos\left(2\pi \frac{x_i}{T}\right)$$





Fit this model:

- $Y_i | x_i \sim N(\mu(x_i), \sigma^2)$
- $\mu(x_i) = \beta_0 + \beta_1 x_i$

$\Leftrightarrow$

- $Y_i = (\beta_0 + \beta_1 x_i) + \varepsilon_i$
- $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$

$$\varepsilon_i = Y_i - \mu(x_i)$$ ← not observed

Residuals

$$R_i = Y_i - \hat{\mu}(x_i)$$ } observed from the data

$$= Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$$ }

$$R_i = \overbrace{\mu(x_i)}^{\text{True}} + \varepsilon_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

$$= \overbrace{\beta_0 + \beta_1 x_i + \alpha_1 \cos\left(2\pi \frac{x_i}{T}\right) + \varepsilon_i}^{Y_i}$$

$$- (\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

$$= (\beta_0 - \hat{\beta}_0) + (\beta_1 - \hat{\beta}_1) x_i + \alpha_1 \cos\left(2\pi \frac{x_i}{T}\right)$$

$$+ \varepsilon_i$$

$$\sim \alpha_1 \cos\left(2\pi \frac{x_i}{T}\right) + \varepsilon_i$$

Plot of $\{R_i\}$.

ANALYSIS OF COVARIANCE