AMS 572 Data Analysis I
Simple Linear Regression

Pei-Fen Kuan

Applied Math and Stats, Stony Brook University

Analysis of Variance

- Recall under $H_0 : \beta = 0$

$$t = \frac{\hat{\beta}}{\sqrt{s_{y.x}^2 / \sum_i (X_i - \bar{X})^2}} \sim t_{N-2}$$

- In general, if $T \sim t_\nu$, then $T^2 \sim F_{1,\nu}$. Thus

$$t^2 \sim F_{1,N-2}$$

Analysis of Variance

- Note

$\hat{\beta} = \frac{\sum(X_i Y_i - N\bar{X}\bar{Y})}{\sum(X_i - \bar{X})^2}$

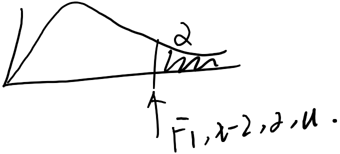
$$SSR = \sum(\hat{Y}_i - \overline{Y})^2 = \sum \hat{\beta}^2 (X_i - \overline{X})^2$$
$$= \frac{(\sum(X_i - \bar{X})(Y_i - \bar{Y}))^2}{\sum(X_i - \bar{X})^2}$$

- Thus

$MSE = S^2_{Y|X}$

$$t^2 = \frac{SSR}{MSE} = \frac{SSR}{SSE/(N-2)} \sim F_{1, N-2}$$

Analysis of Variance



- For $H_0 : \beta = 0$ vs $H_A : \beta \neq 0$, can use F with

$$C_\alpha = \{F : F > F_{1-\text{[redacted]};1,N-2}\}_{\alpha,u}.$$

- For two sided alternative F and t tests equivalent
- For one sided alternative, use t

Analysis of Variance

MSR = SSR/1

- ANOVA table:

Source	df	SS	MS	F
Regression	1	SSR	SSR=MSR	MSR/MSE
Residual	$N-2$	SSE	SSE/$(N-2)$	
Total	$N-1$	SST		

Diagnostics

- Assumptions for linear regression
  1. Linearity: $Y_i = \alpha + \beta X_i + \epsilon_i$
  2. X's are fixed constants
  3. ϵ_i iid $\sim N(0, \sigma^2)$

constant

Diagnostics

- Assumptions: Linear model and homogeneity of variance
- *Residual plot*: Scatterplot of

$$(\hat{Y}_i, r_i) = (\hat{Y}_i, Y_i - \hat{Y}_i)$$

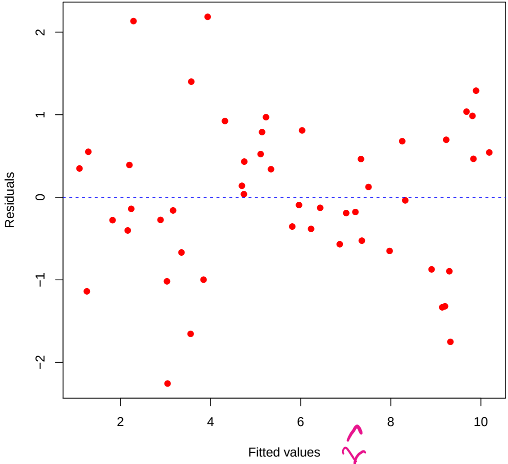
- If we see lack of homogeneity of variance or linearity, consider transformations

Diagnostics

- The following three slides are prototypical residual plots indicating
  1. linear regression model is appropriate
  2. assumption of linearity questionable
  3. assumption of constant variance questionable

Residual plots

$\hat{r}_i$

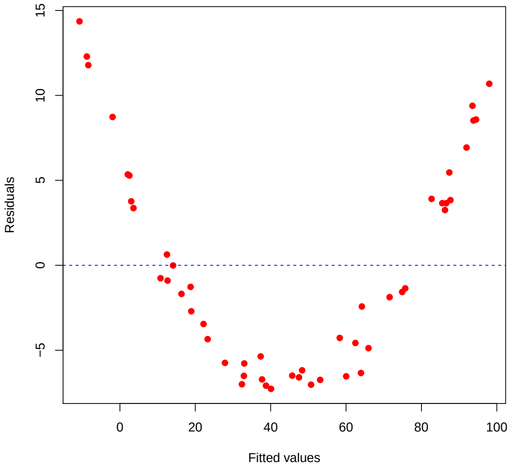


$\hat{Y}$

Residual plots

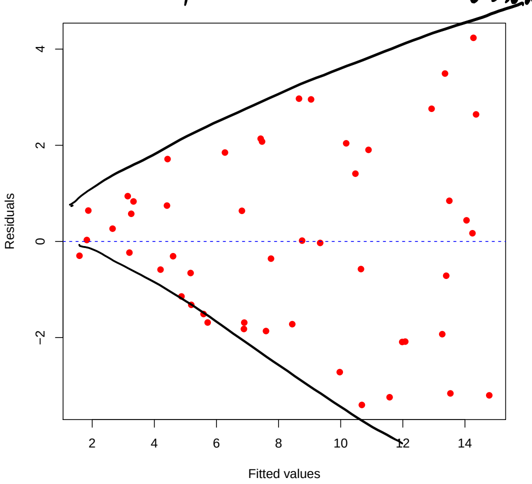
$Y = 2 + \beta_1 X + \beta_2 X^2$

sth not right.



Residual plots

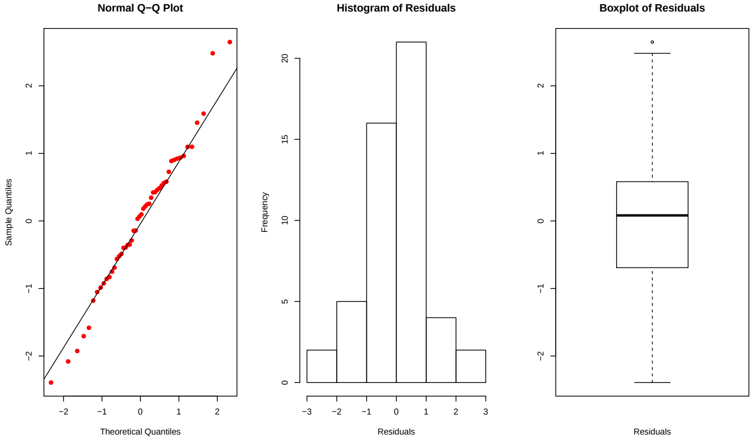
funnal pattern: constant variance.
assumption questionable



Normality Diagnostics

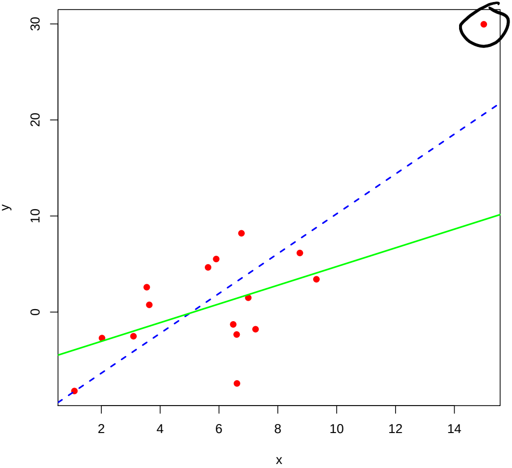
- Assumption: ϵ_i's are normally distributed
- This assumption is not as important if N is large (CLT)
- Inference robust to small departures from normality
- Violations of other assumptions can suggest non-normality
- qq-plot, histogram, boxplot of residuals

Residual plots



Regression: Diagnostics

- Beware influential observations; always check scatterplot



after delete

Remedial Measures

- Transformations, e.g., $\log(Y) = \alpha + \beta X$
- Multiple regression, e.g., $Y = \alpha + \beta_1 X + \beta_2 X^2$
- Nonparametric procedures, e.g., Kendall's tau
- More sophisticated models allowing for
  - dependencies/clusters (e.g., GEE)
  - heterogeneity of variance (e.g., weight least squares)

non constant / hetero scedascity.