STA302/1001 - Methods of Data Analysis I

(Week 05 lecture notes)

Wei (Becky) Lin

Oct. 10, 2016

UNIVERSITY OF TORONTO

Midterm

- No cheat sheet. A formula page is provided. Bring: a calculator, student card.
- Cover the first 4 lectures (questions: proof and analysis).
- TA office hour: 12-2pm, Monday, Oct 17, SS2108.(For L0101+L2001).
- TA office hour: 4-6pm,Wednesday, Oct 19,UC161. (For L5101).
- My office hour: 3-4pm, Monday, Oct 17, SS6011 or SS6007.

Day section L0101+L2001

- Time: 10:10-11:40am, Tuesday, Oct 18. (No class on Oct 20)
- Location:
  - Last name: A-C, please write your midterm at BL205.
  - Last name: D-Hu, please write your midterm at CB114.
  - Last name: Huang -Z, please write your midterm at MS2158.

Evening Section L5101

- Time: 6:00-8:00pm, Thursday, Oct 20. (No class from 5-6pm)
- Location: EX100

Review for Midterm: Variance and Covariance

- Let a, b, a_i, b_i, c be some constant.
- Variance
  - Definition: $V(X) = E(X^2) - [E(X)]^2 = E([X - E(X)]^2)$
  - $V(c) = 0, V(cX) = c^2 V(X), V(X + c) = V(X)$
  - $V(X \pm Y) = V(X) + V(Y) \pm 2Cov(X, Y)$
- Covariance
  - Definition:
    $Cov(X, Y) = E(XY) - E(X)E(Y) = E[(X - \mu_x)(Y - \mu_y)]$
  - Symmetry: $Cov(X, Y) = Cov(Y, X)$
  - Bilinearity:
    - $Cov(cX, Y) = Cov(X, cY) = cCov(X, Y)$
    - $Cov(X_1 + X_2, Y) = Cov(X_1, Y) + Cov(X_2, Y)$
    - $Cov(a + X, b + Y) = Cov(X, Y)$
  - Product formula
    - $Cov(\sum_1^n X_i, \sum_1^m Y_j) = \sum_i \sum_j Cov(X_i, Y_j)$
    - $Cov(\sum_1^n a_i X_i, \sum_1^m b_j Y_j) = \sum_i \sum_j a_i b_j Cov(X_i, Y_j)$

Variance and Covariance:

Show $Cov(Y_i, Y_j)I(i \neq j) = 0$ and $V(b_1) = \sigma^2/S_{xx}$

From $Cov(a + X, b + Y) = Cov(X, Y)$, we have

$$i \neq j : Cov(Y_i, Y_j) = Cov(\beta_0 + \beta_1 X_i + \epsilon_i, \beta_0 + \beta_1 X_j + \epsilon_j) = Cov(\epsilon_i, \epsilon_j) = 0$$

$$V(b_1) = V(\sum_i K_i Y_i) \tag{1}$$

$i=j$: n terms

$$= \sum_i K_i^2 V(Y_i) + 2 \sum_i \sum_j I(i \neq j) Cov(Y_i, Y_j) \tag{2}$$

$$= \sigma^2 \sum_i K_i^2 + 0 \tag{3}$$

$$= \frac{\sigma^2}{S_{xx}} \tag{4}$$

Equation (2) uses the third formula for variance in previous slide. Second term in (3) is 0 since Y_j and Y_i are uncorrelated for $i \neq j$.

Variance and Covariance:

Show $Cov(b_1, \bar{Y}) = 0$

$k_i = \frac{X_i - \bar{X}}{S_{xx}}$, $S_{xx} = \sum_{i=1}^{n} (X_i - \bar{X})^2$

$\sum k_i = \frac{1}{S_{xx}} \sum (X_i - \bar{X})$

$= \frac{1}{S_{xx}} (\sum X_i - n\bar{X})$ (5)

$= 0 \qquad = 0$

$$= \sum_{i}^{n} \sum_{j}^{n} K_i \frac{1}{n} Cov(Y_i, Y_j) \tag{6}$$

$$= \sum_{i=j=1}^{1} K_i \frac{1}{n} Cov(Y_i, Y_j) + \sum \sum_{i \neq j} K_i \frac{1}{n} Cov(Y_i, Y_j) \tag{7}$$

$$= \sum_{i=1^n} K_i \frac{1}{n} Cov(Y_i, Y_i) + 0 \tag{8}$$

$$= \sigma^2 \frac{1}{n} \sum_{i=1^n} K_i \tag{9}$$

$= 0$

$$= 0 \tag{10}$$

Last Week

- Correlation analysis.
- The equivalence between the conditional distribution and the SLR.
- Two measurements of coefficient correlation: $r_{Pearson}$ and $r_{Spearman}$
- Coefficient of determination: R^2
  - how to estimate it
  - how to interpret it
  - what limitations of using it?
- Regression with dummy variable.
  - the estimate of the intercept is the mean of Y under reference level.
  - the slope estimate is the mean difference of Y between other level and the reference level.
- Diagnostic of predictor variable
  - identify problematic data points: influence and high leverage point
  - sequence plot to examine the independence

$\hat{\beta}_0 = E(Y|x=\text{ref level})$

$\hat{\beta}_1 = b_1 = E(Y|x\neq\text{ref level}) - \hat{\beta}_0$

Review: a leverage point

$$\hat{Y}_i = \sum_{j=1}^{n} [\frac{1}{n} + \frac{(X_j - \bar{X})(X_i - \bar{X})}{S_{XX}}] Y_j = \sum_{j=1}^{n} h_{ij} Y_j$$

- A leverage point is a point whose x-value is distant from the mean. Define the leverage of the i^{th} data point as

$$h_{ii} = \frac{1}{n} + \frac{(X_i - \bar{X})^2}{S_{XX}} \in (0, 1)$$

$(X_i - \bar{X})^2$ larger $\Rightarrow h_{ii}$ Move toward 1

- Properties of h_{ij}
  - $\sum_j h_{ij} = 1$;
  - $\sum_j h_{ij}^2 = h_{ii}$
  - $\bar{h}_{ii} = 2/n$;
  - $h_{ij} = h_{ji}$

sum row = 1

$$\begin{pmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \vdots \\ \hat{Y}_n \end{pmatrix} = \begin{pmatrix} h_{11} & h_{12} & h_{13} & \cdots & h_{1n} \\ h_{21} & h_{22} & h_{23} & \cdots & h_{2n} \\ \vdots & & & & \vdots \\ h_{n1} & h_{n2} & h_{n3} & \cdots & h_{nn} \end{pmatrix} \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}$$

sum(row element^2) $= h_{ii}$

- Identifying leverage points if

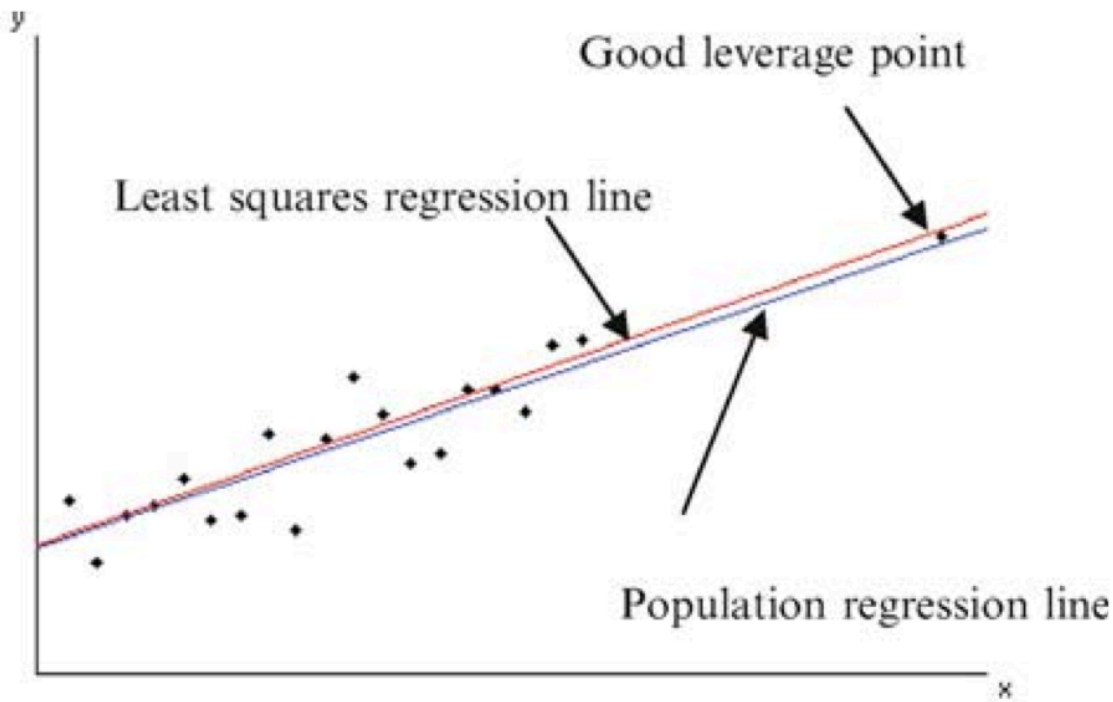
$$h_{ii} > 2\bar{h}_{ii} = \frac{4}{n}$$

A good leverage point

- A web-based applet (by Robert McCulloch) to illustrate leverage points:

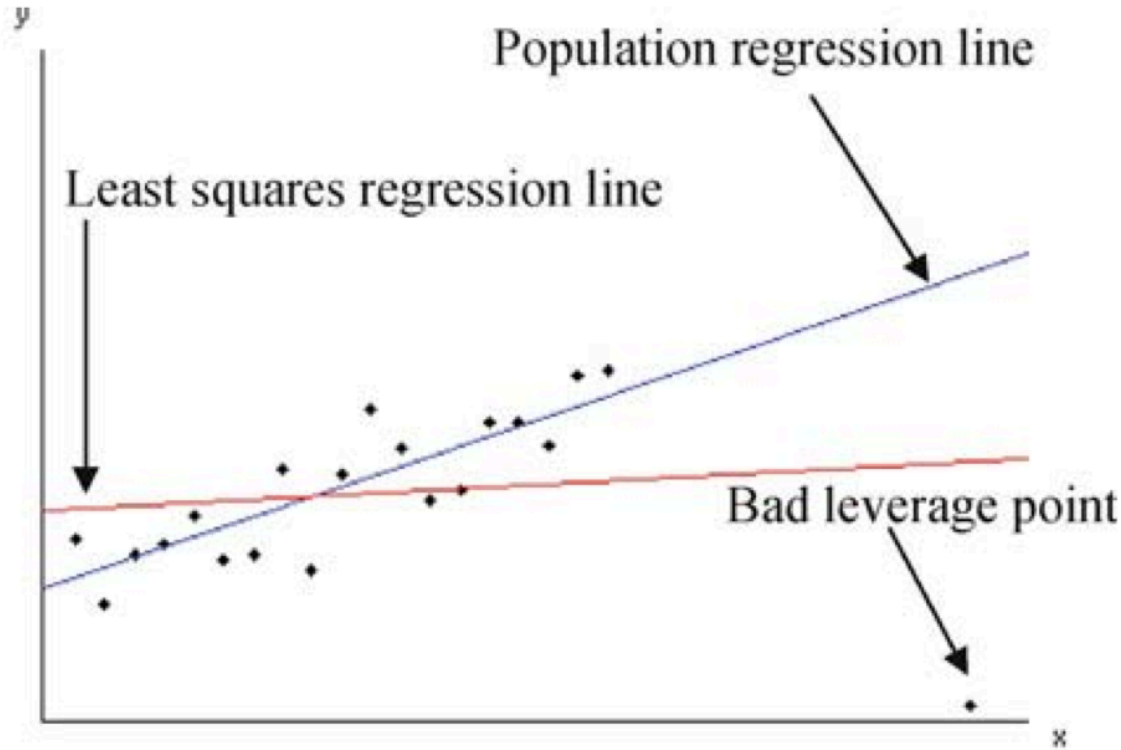
  `http://www.rob-mcculloch.org/teachingApplets/Leverage/index.html`

- A good leverage point if its y-value closely follows the upward or downward trend pattern set by other points. In other words, a good leverage point is a leverage point which is not an outlier.

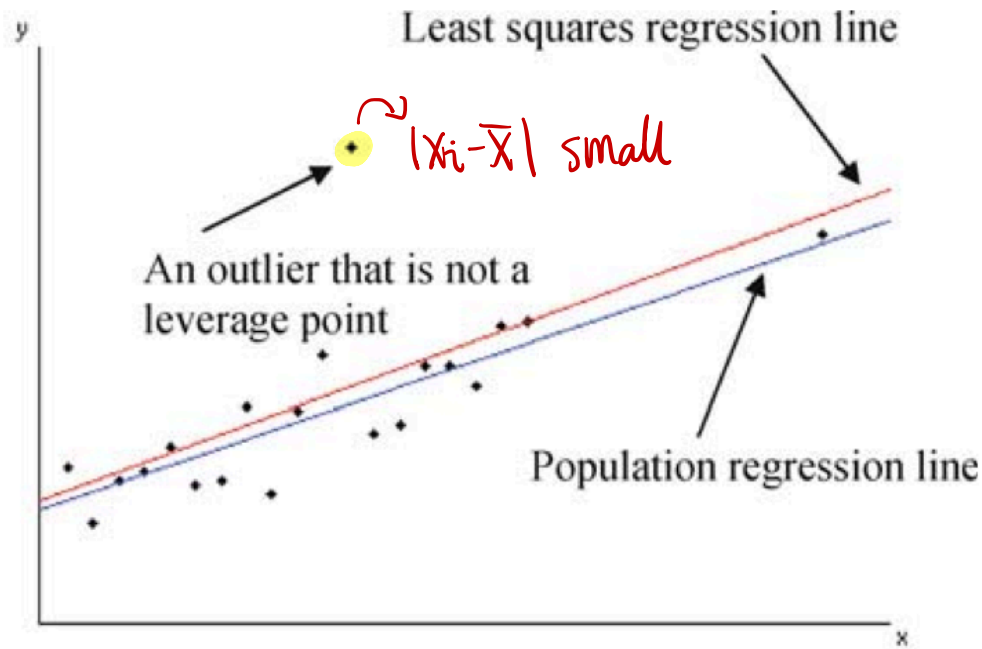


A bad leverage point

- A bad leverage point if its y-value does not follow the pattern set by the other data points. In other words, a bad leverage point is leverage point which is also an outlier.



An outlier that is not a leverage point.



Notice how in the least squares regression has changed relatively little in response to changing the Y-value of centrally located x.

Week 04- Learning objectives & Outcomes

- Unusual and influential data points
- Diagnostics for residual
- Influence Metrics: DFFITS, DFBETAS,COOK's distance
- Case study

Unusual and Influence data points

- Outliers: we classify points as outliers if their standardized residuals have absolute value greater than 2 (4) for a small or moderate sample (for a large sample).
  - An observation whose y-value is unusual given its x-value.
  - An outlier may indicate a sample peculiarity or may indicate a data entry error or other problem.
- Leverage: an observation with extreme x-value
  - Leverage is a measure of how far the predictor variable deviates from its mean.
  - These leverage points can have an effect on the estimate of regression coefficients.
- Influence: influence can be thought of as the product of leverage and outliers
  - Removing the observation substantially changes the estimate of coefficients.
  - But on what sense we can classify a data point as influential point? Answer: DFBETAS measure

Diagnostics for Residuals

Why residual analysis?

- Observed error

$$e_i = Y_i - \hat{Y}_i$$

→ ∴ observable estimate of ϵ_i

- True error

$$\epsilon_i = Y_i - E(Y_i)$$

→ never observe it

- If the linear model is appropriate for the data at hand, the observed residuals, e_i, should then reflect the properties assumed for the ϵ_i.

Properties of residuals e_i

- $\sum e_i = 0$, i.e. $\bar{e} = 0$
- $\sum e_i X_i = 0$, $\sum e_i \hat{Y}_i = 0$,the residual is orthogonal to the subspace spanned by the regressors.
- $s^2 = \frac{1}{n-2} \sum e_i^2 = SSE/(n-2) = MSE$, E(MSE)$=\sigma^2$
- Residuals $e_i (i = 1, \ldots, n)$, are NOT independent (since $\text{Cov}(e_i, e_j) \neq 0$ for $i \neq j$), but for large n we can ignore their correlation.

Show $\mathrm{Var}(\hat{Y}_i) = \sigma^2 h_{ii}$

$$\hat{Y}_i = \sum_{j=1}^{n} h_{ij} Y_j$$

$$\Rightarrow \mathrm{Var}(\hat{Y}_i) = V\left(\sum_{j=1}^{n} h_{ij} Y_j\right)$$

$$= \sum_{j=1}^{n} h_{ij}^2 V(Y_j) + 2 \sum\sum_{m \neq j} \mathrm{cov}(h_{im} Y_m, h_{ij} Y_j)$$

$$= \sigma^2 \sum_{j=1}^{n} h_{ij}^2 + 2 \sum\sum_{m \neq j} h_{im} h_{ij} \underbrace{\mathrm{cov}(Y_m, Y_j)}_{=0,\ m \neq j}$$

$$= \sigma^2 \underbrace{\sum_{j=1}^{n} h_{ij}^2}$$

$$= \sigma^2 h_{ii}$$

Show residuals do NOT have the same variance.

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

$e_i = Y_i - \hat{Y}_i = Y_i - \sum_{j=1}^{n} h_{ij} Y_j$

$\mathrm{Var}(e_i) = V(Y_i - \sum_{j=1}^{n} h_{ij} Y_j)$

$= V(Y_i) + V(\sum_{j=1}^{n} h_{ij} Y_j) - 2\,\mathrm{cov}(Y_i, h_{i1}Y_1 + \cdots h_{ii}Y_i + \cdots Y_{in}Y_n)$

$\sigma^2 h_{ii}$, rest is 0

$= \sigma^2 + \sum_{j=1}^{n} h_{ij}^2 V(Y_j) + 2 \sum\sum_{m \neq j} h_{im} h_{ij}\, \mathrm{cov}(Y_m, Y_j) - 2\sigma^2 h_{ii}$

$\mathrm{cov}(Y_m, Y_j) = 0$

$= \sigma^2 + \sigma^2 \sum_{j=1}^{n} h_{ij}^2 + 0 - 2\sigma^2 h_{ii}$

$\sum_{j=1}^{n} h_{ij}^2 = h_{ii}$

$= \sigma^2 + \sigma^2 h_{ii} - 2\sigma^2 h_{ii}$

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

Q.E.D.

Show $\text{Cov}(e_i, e_j) = -\sigma^2 h_{ij}$, $i \neq j$

$$\text{cov}(e_i, e_j) = \text{cov}(Y_i - \hat{Y}_i, e_j)$$

$$= \text{cov}(Y_i, e_j) - \boxed{\text{cov}(\hat{Y}_i, e_j)} \quad = 0$$

Exercise this Week

$$= \text{cov}(Y_i, Y_j - \hat{Y}_j)$$

$$= \text{cov}(Y_i, Y_j) - \text{cov}(Y_i, \hat{Y}_j), \quad i \neq j$$

$$= 0 - \text{cov}(Y_i, \textstyle\sum_{m=1}^{n} h_{jm} Y_m)$$

$$= 0 - \text{cov}(Y_i, h_{j1}Y_1 + \cdots + h_{ji}Y_i + \cdots + h_{jn}Y_n)$$

$\sigma^2 h_{ji}$, rest is 0

$$= -\sigma^2 h_{ji}$$

$$= \sigma^2 h_{ij}$$

Q.E.D.

Variance and covariance matrix for a vector of residuals

In previous slides, we have

$$\mathrm{Var}(e_i) = \sigma^2(1-h_{ii}), \quad \mathrm{cov}(e_i, e_j) = -\sigma^2 h_{ij}, \ i \neq j$$

$$\Rightarrow \mathrm{Var}\begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix} = \begin{pmatrix} v(e_1) & c(e_1,e_2) & \cdots & c(e_1,e_n) \\ c(e_2,e_1) & v(e_2) & \cdots & c(e_2,e_n) \\ \vdots & & & \\ c(e_n,e_1) & c(e_n,e_2) & \cdots & v(e_n) \end{pmatrix}$$

$$= \sigma^2 \begin{pmatrix} 1-h_{11} & -h_{12} & \cdots & -h_{1n} \\ -h_{21} & 1-h_{22} & \cdots & -h_{2n} \\ \vdots & \vdots & & \vdots \\ -h_{n1} & -h_{n2} & \cdots & 1-h_{nn} \end{pmatrix}$$

Identity Matrix

Hat matrix

$$= \sigma^2 \left[\begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix} - \begin{pmatrix} h_{11} & h_{12} & \cdots & h_{1n} \\ h_{21} & h_{22} & \cdots & h_{2n} \\ \vdots & & & \\ h_{n1} & h_{n2} & \cdots & h_{nn} \end{pmatrix} \right] = \sigma^2(I-H)$$

Show $\text{Corr}(e_i, e_j) = \dfrac{-h_{ij}}{\sqrt{(1-h_{ii})(1-h_{jj})}}$, $i \neq j$

$$\text{Corr}(e_i, e_j) = \frac{\text{cov}(e_i, e_j)}{\sqrt{\text{var}(e_i)\,\text{var}(e_j)}}$$

$$= \frac{-\sigma^2 h_{ij}}{\sqrt{\sigma^2(1-h_{ii})\,\sigma^2(1-h_{jj})}}$$

$$= -\frac{h_{ij}}{\sqrt{(1-h_{ii})(1-h_{jj})}}$$

Q.E.D.

Standardized and studentized residuals

- Semi-studentized residuals
  - $Var(\epsilon_i) = \sigma^2$ and MSE is an unbiased estimate of σ^2.
  - Semi-studentized residuals are "standardized" residuals

$$e_i^* = \frac{e_i - \bar{e}}{\sqrt{MSE}} = \frac{e_i}{\sqrt{MSE}}$$

  - Variance of e_i^* should be ≈ 1.

- Studentized residuals
  - Variance of e_i is: $V(e_i) = \sigma^2(1 - h_{ii})$
  - Studentized residuals are properly standardized residuals

$$r_i = \frac{e_i}{\sqrt{MSE(1 - h_{ii})}}$$

  - This is actually an *internally* studentized residual.
  - If the i^{th} case is deleted (change estimate of σ^2, say estimate is $\text{MSE}_{(i)}$), r_i uses $\text{MSE}_{(i)}$ is the *externally* studentized residual.

Diagnostics for Residuals

- Assume the normal error model.
- Use residuals to study the following 6 types of departures from the SLR model
  - The regression function is not linear.
  - The error terms does not have constant variance.
  - The model fits all but one or a few outlier observations
  - The error terms are not normally distributed.
  - One or several important predictor variables have been omitted for the model.

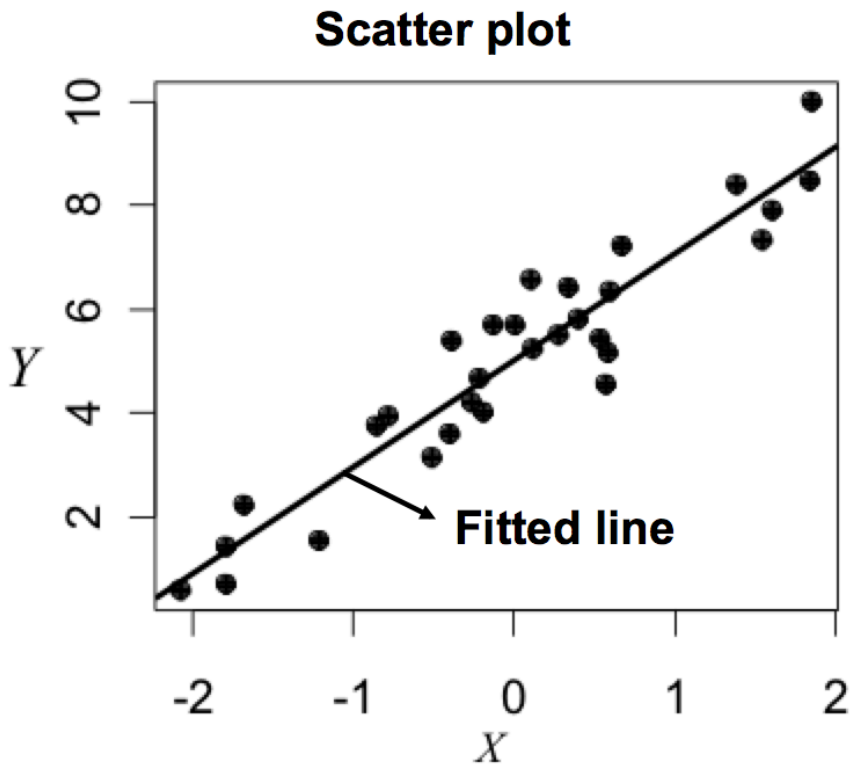
1.) Checking for Linearity

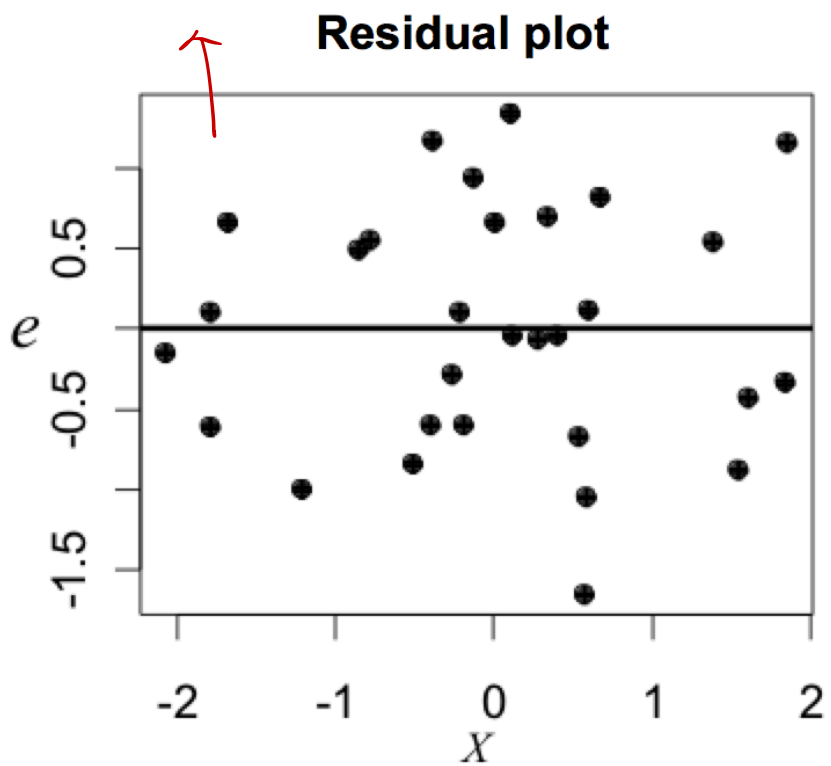
- To check for linearity
  - Plot residuals vs predictor variable (e_i vs X_i)
  - Plot residuals vs fitted values (e_i vs $\hat{Y}_i$)
- A scatter plot (Y vs X) can also be used. But residual plots are preferred because
  - Can spot nonlinearity more easily.
  - Can check other assumptions, e.g. constant variance and etc.
  - Applied to multiple linear regression
    - (e_i vs X_i) gives equivalent information (e_i vs $\hat{Y}_i$) in SLR but not in MLR (multiple linear regression).

Scatter vs Residual plot

- Linear relationship

randomly spread, no clear pattern

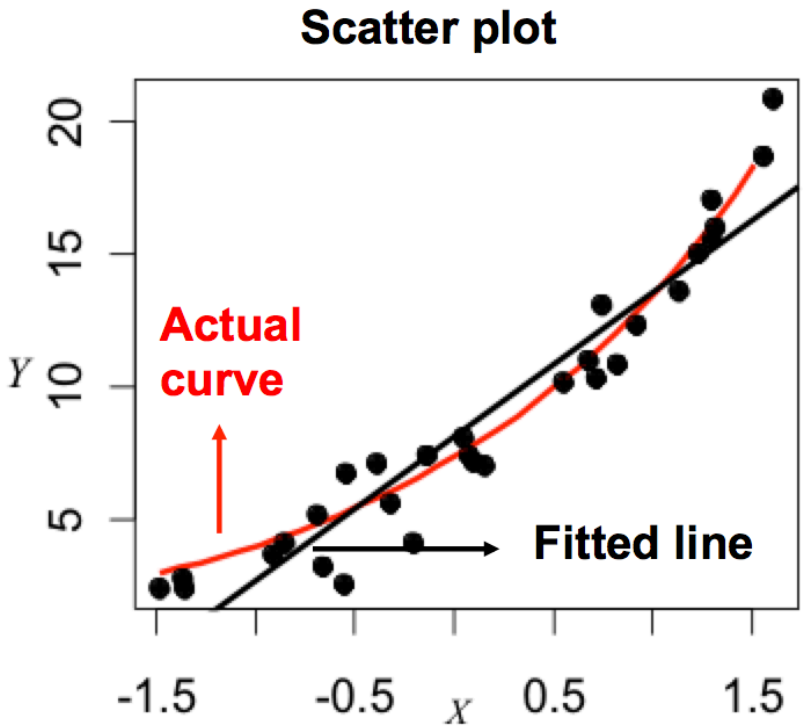


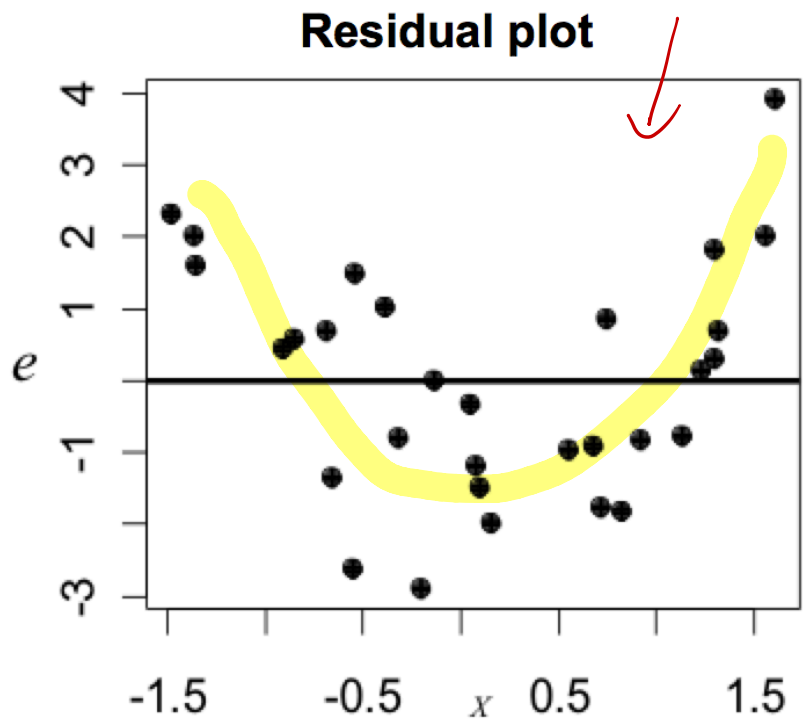


Scatter vs Residual plot

- Nonlinear relationship

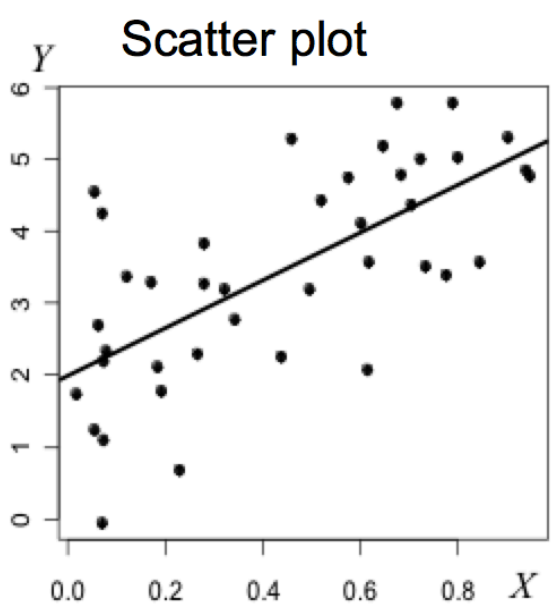
a clear pattern

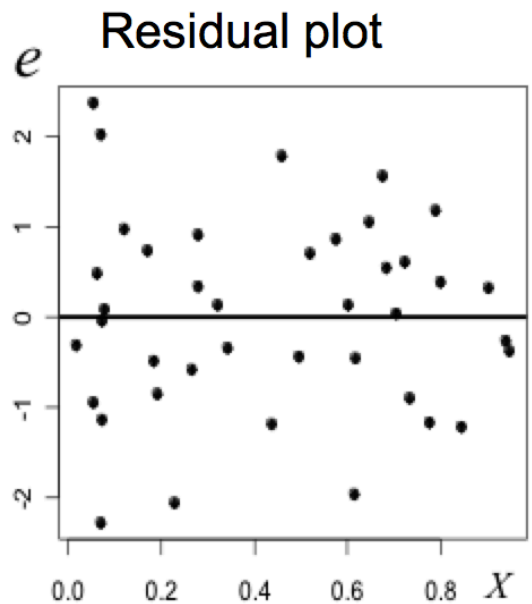


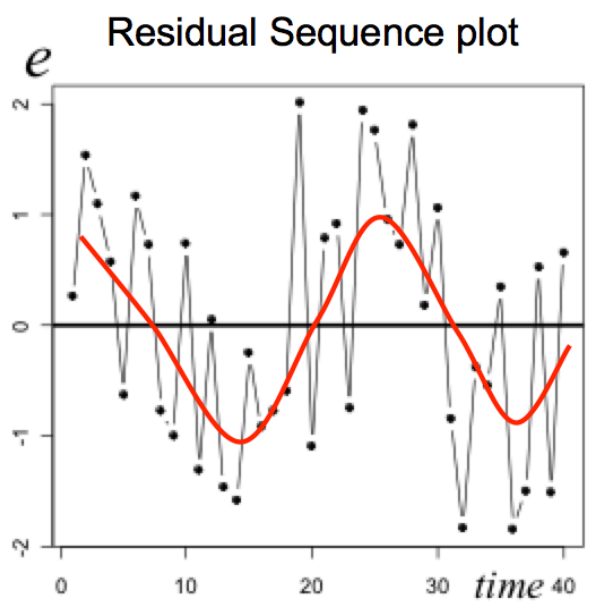


2.) Checking for Error Independence

- Residual sequence plot (e_i vs time or distance)
  - Check for temporal or spatial dependence

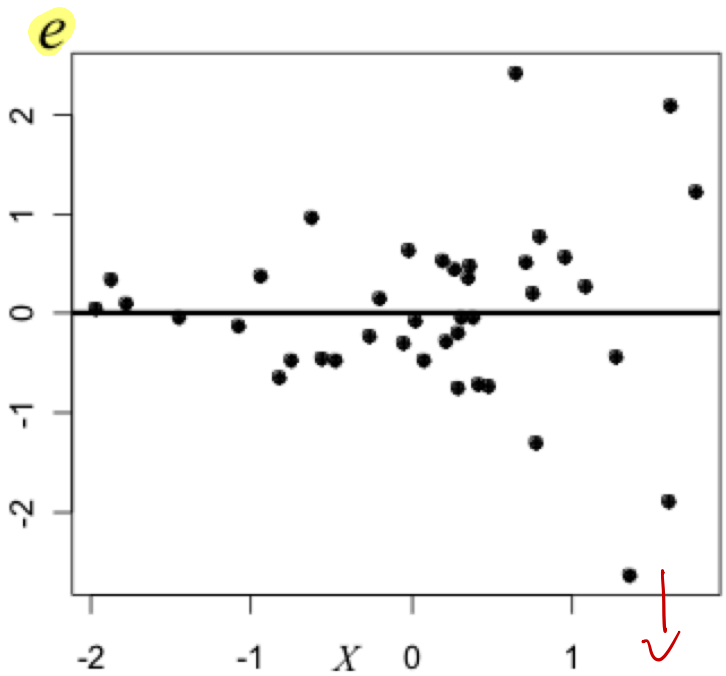




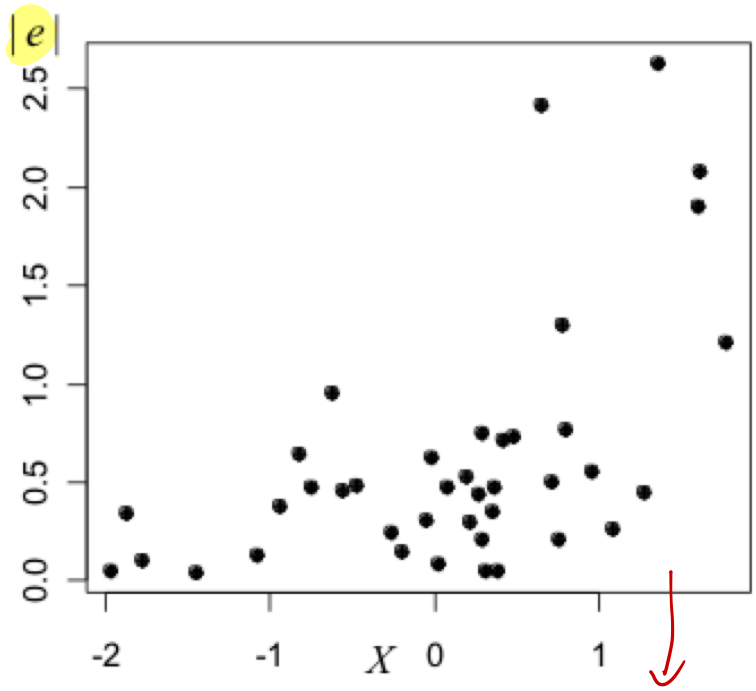


3.) Checking for Constant Variance

- Residual or absolute residual plot
  - Should be spread out evenly across X.



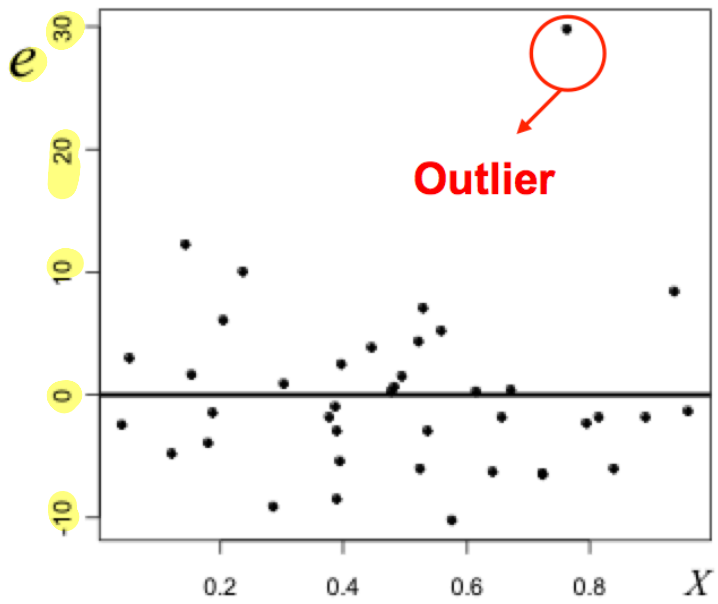
nonconstant σ^2

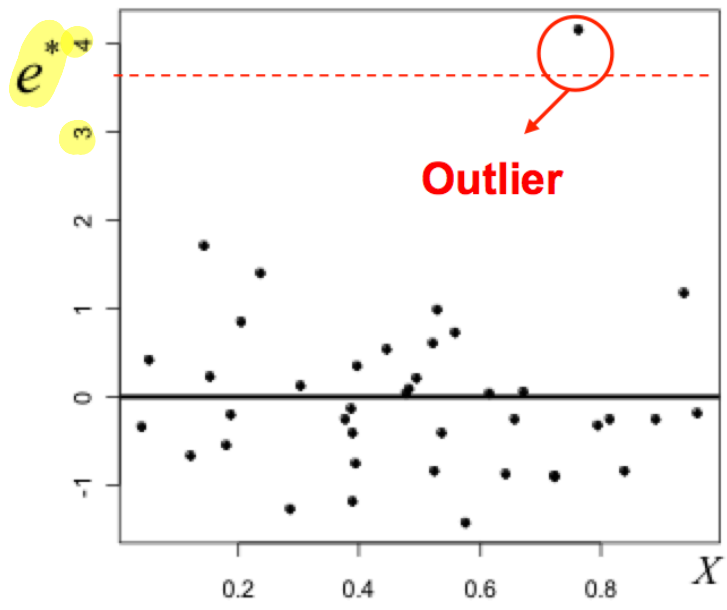


|e| ↑ as x ↑

4.) Checking for Outliers

- Residual or semi-studentized residual plot.
  - Rule of thumb: observation i is an outlier if $|e_i^*| > k$ where k=2 for small or moderate samples and k= 4 for large samples.

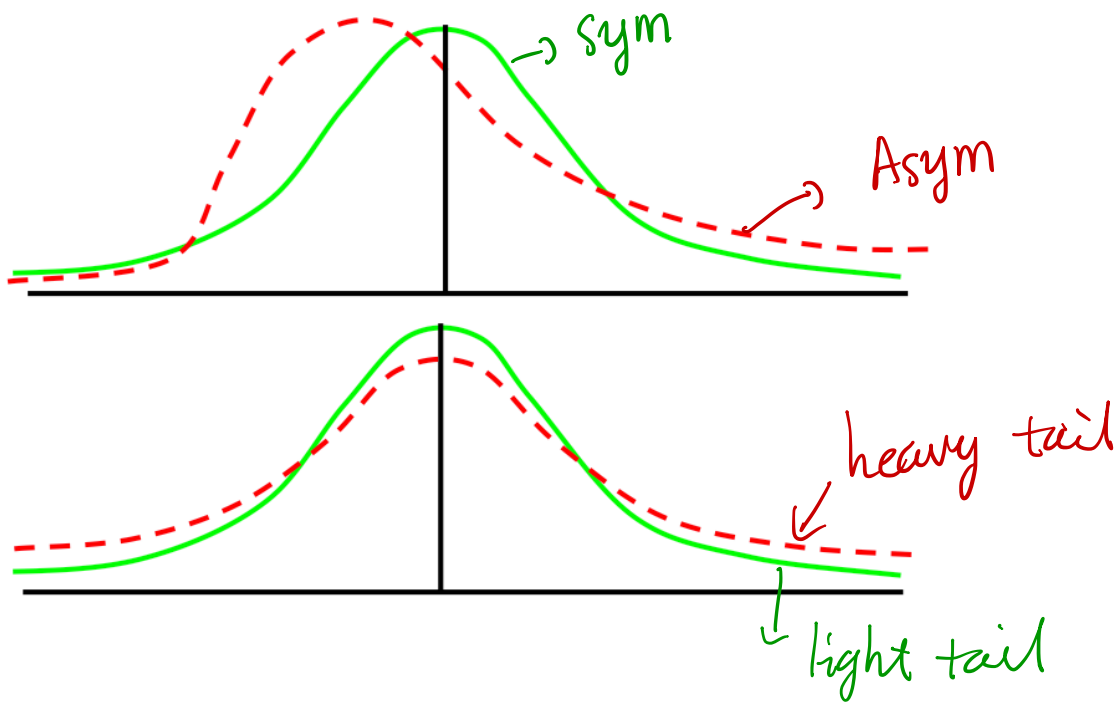




✓ easy to identify outliers

5.) Checking for the normality of error

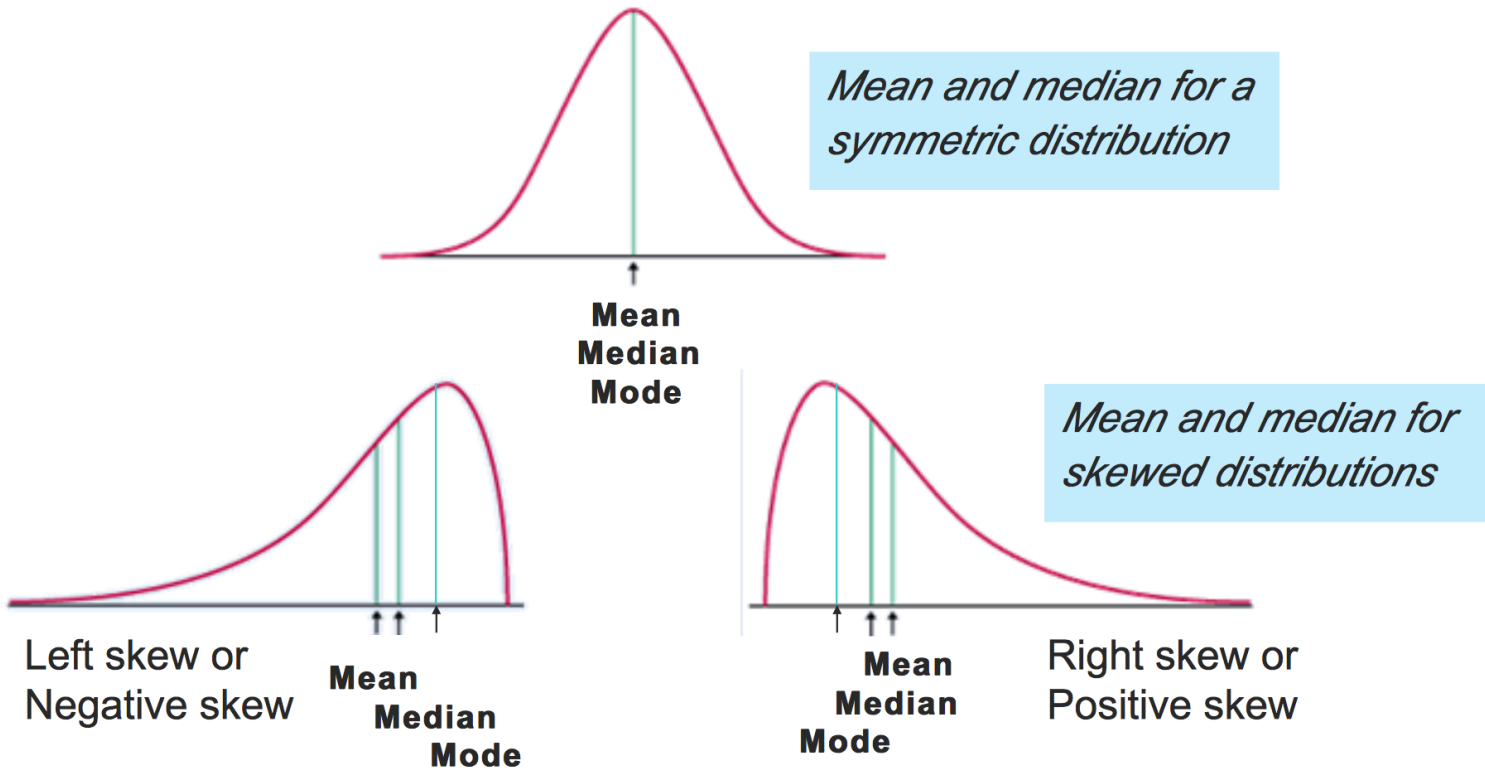
- Look for 2 possible deviations from Normality
  - Asymmetry
  - Heavey tails
- Shapiro-Wilk test of normality
  - in R, use **shapiro-wilk()** where H_0 : is normal



The normality is critical for inference but less important for fitting regression line.

Right-skewed and left-skewed

- For a right-skewed distribution, mean$\geq$median. The right tail (positive) side is longer than on the left tail.
- For a left-skewed distribution, mean$\leq$median. The left tail (negative) side is shorter than on the right tail.



What is normal quantile-quantile (QQ) plot?

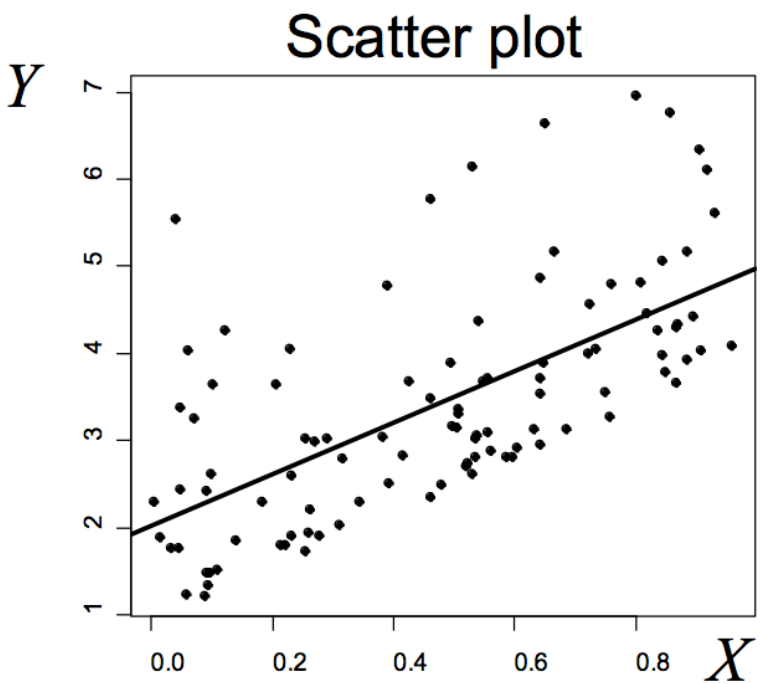
- A 12-minute tutorial on youtube ← very Good. Watch it.

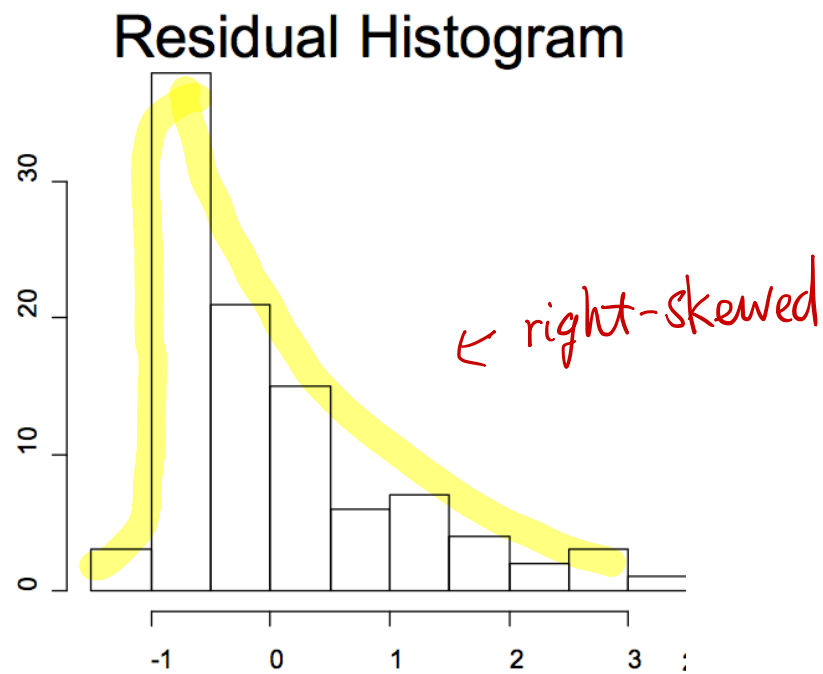
  `https://www.youtube.com/watch?v=X9_ISJ0YpGw`

- Assessing normality is important since most SLR inference procedures assume true error is normally distributed.
- A normal Q-Q plot is a graphical tool to assess if a set of observations is normally distributed or not. If it does, then a normal QQ-plot of the observations will result in an approximately straight line.
- How: by comparing the quantiles of a dataset and a set of theoretical quantiles from a normal probability distribution.

Checking for the normality of error

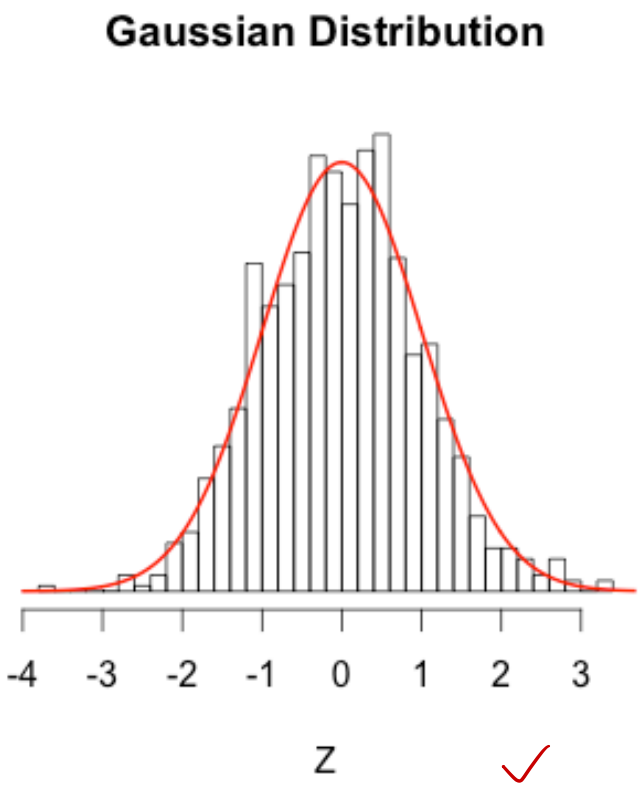
- To check symmetry: any plot describing the data distribution (box-plot, histogram, dot-plot...)

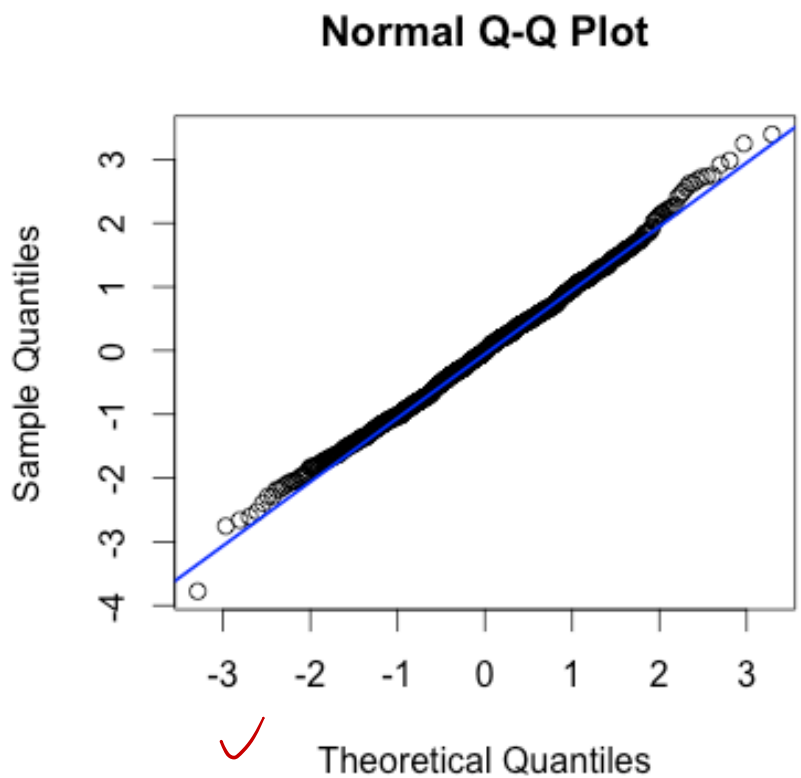




Checking for the normality of error

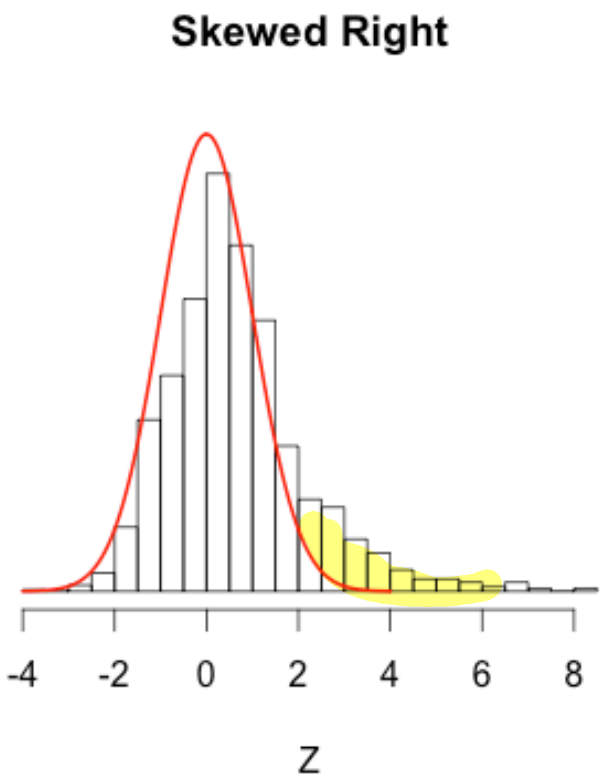
- Normal: Normal QQ-plot

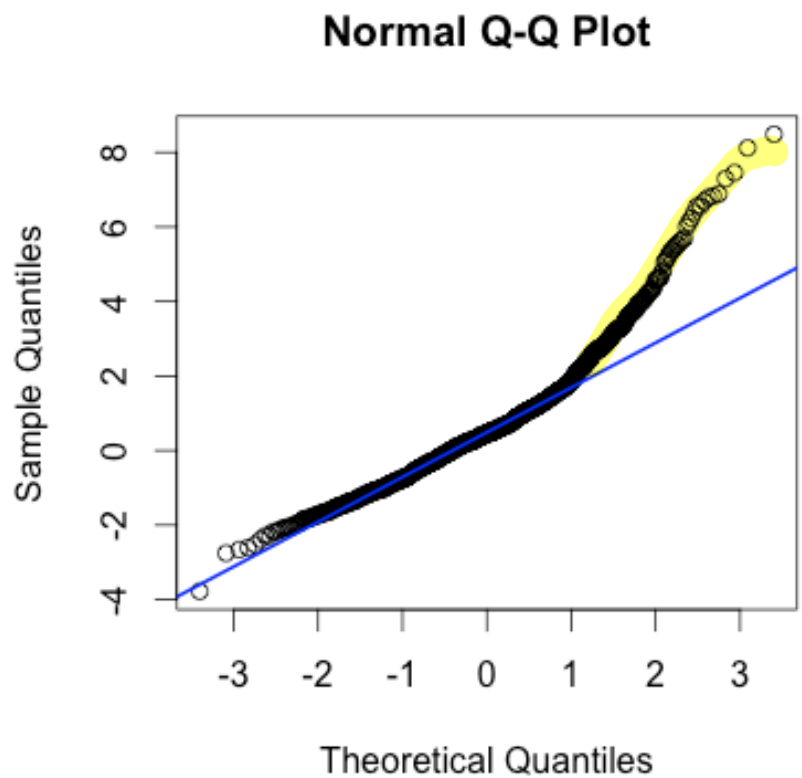




Checking for the normality of error

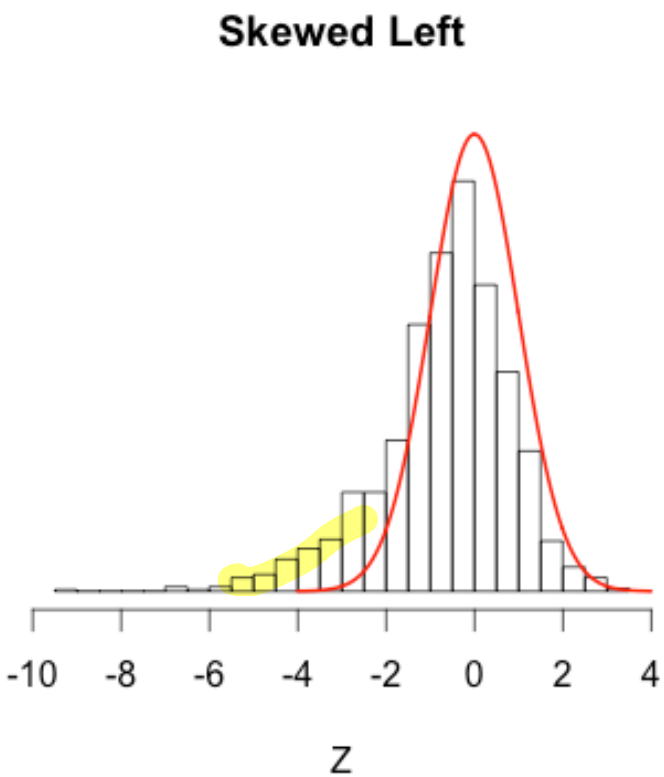
- Right-skewed distribution: Normal QQ-plot

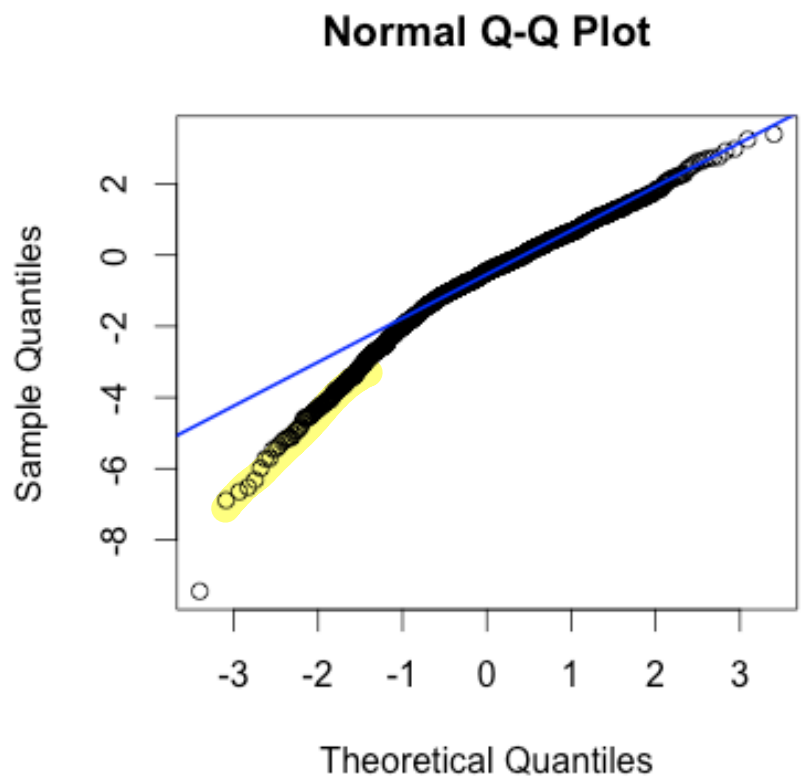




Checking for the normality of error

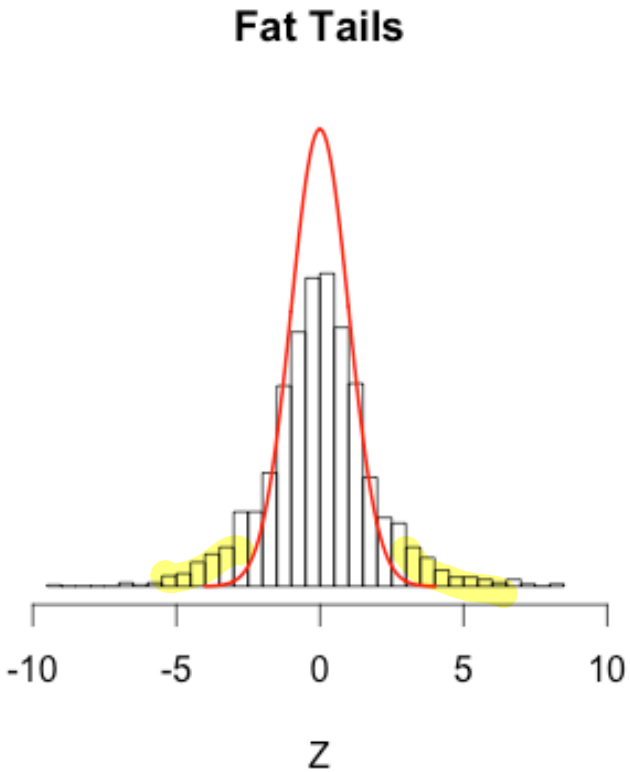
- Left-skewed distribution: Normal QQ-plot

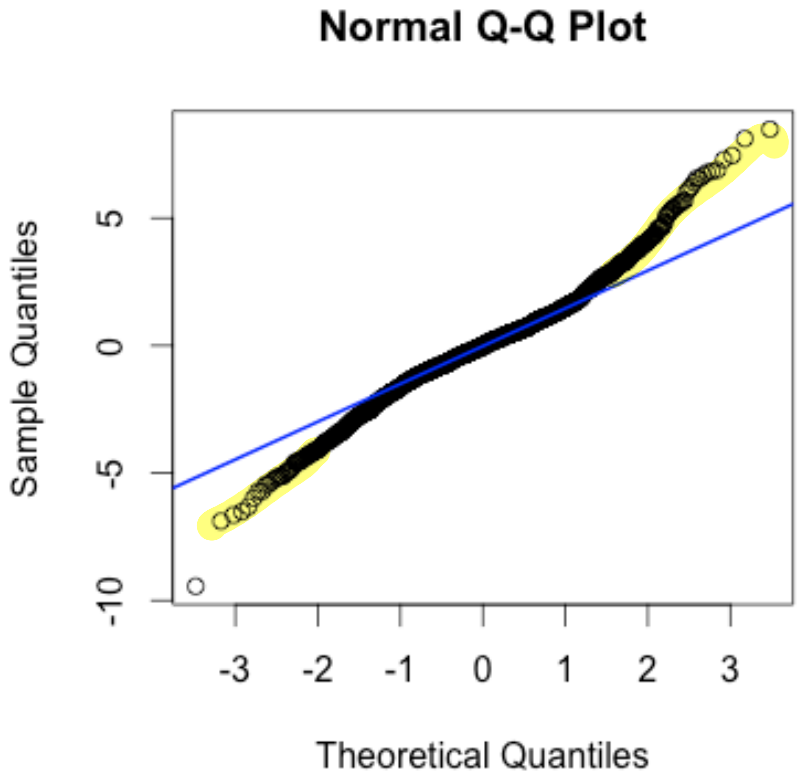




Checking for the normality of error

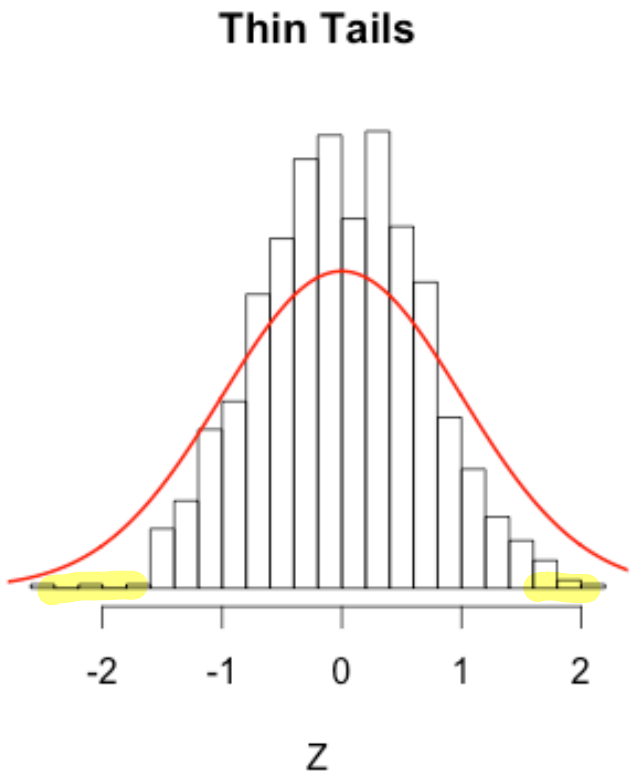
- Fat-tailed (heavy tails) distribution: Normal QQ-plot

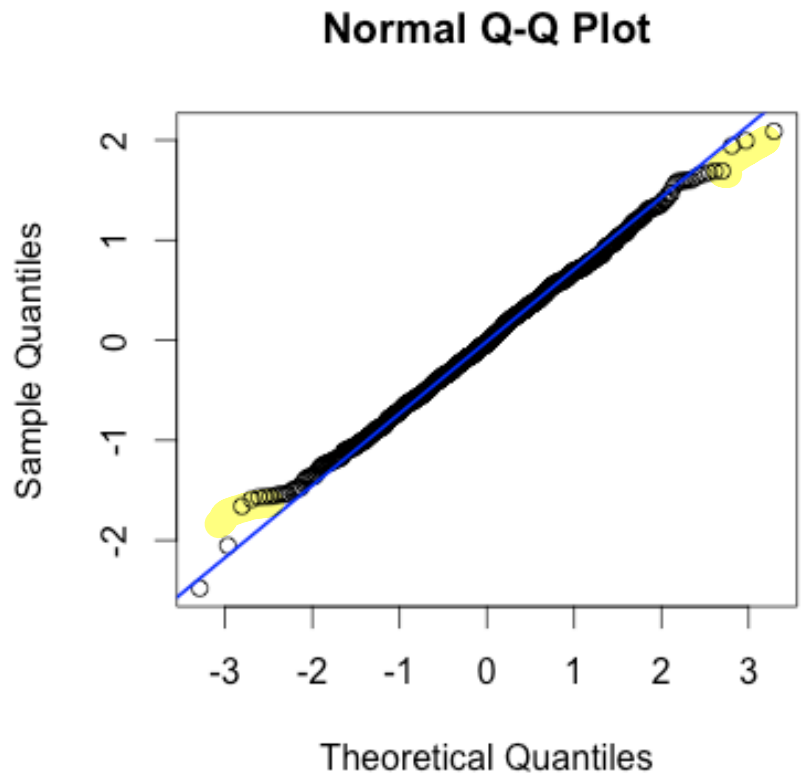




Checking for the normality of error

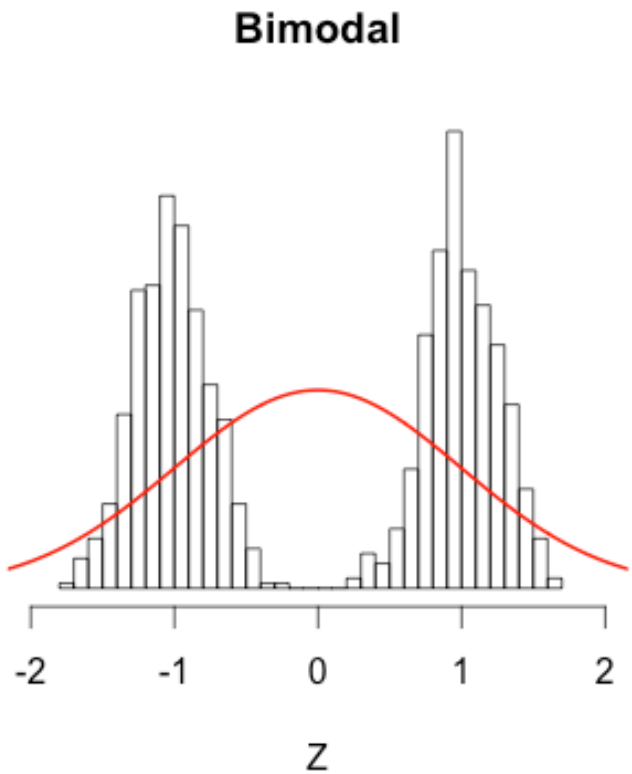
- Thin-tailed distribution: Normal QQ-plot

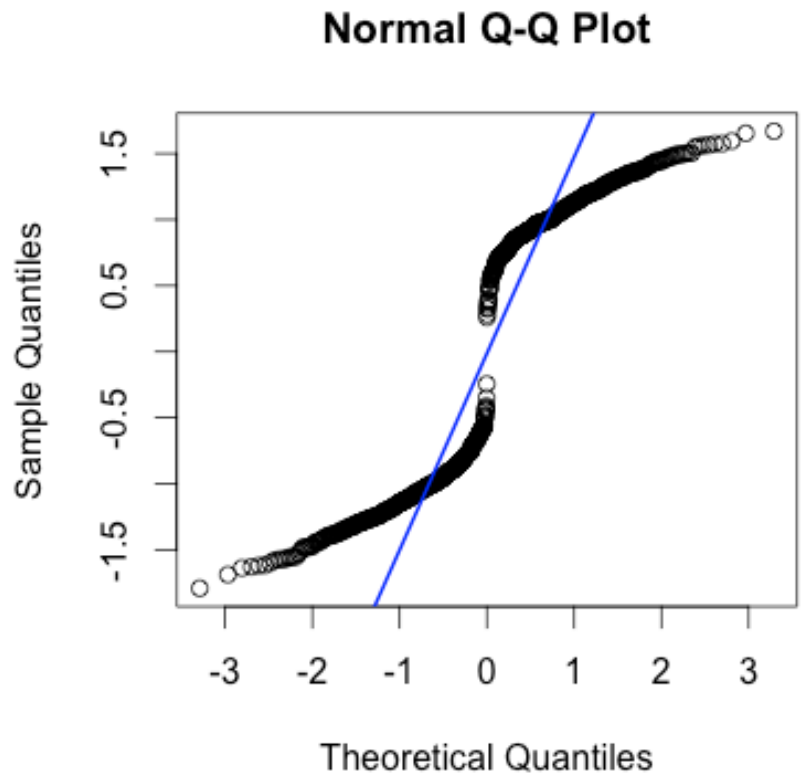




Checking for the normality of error

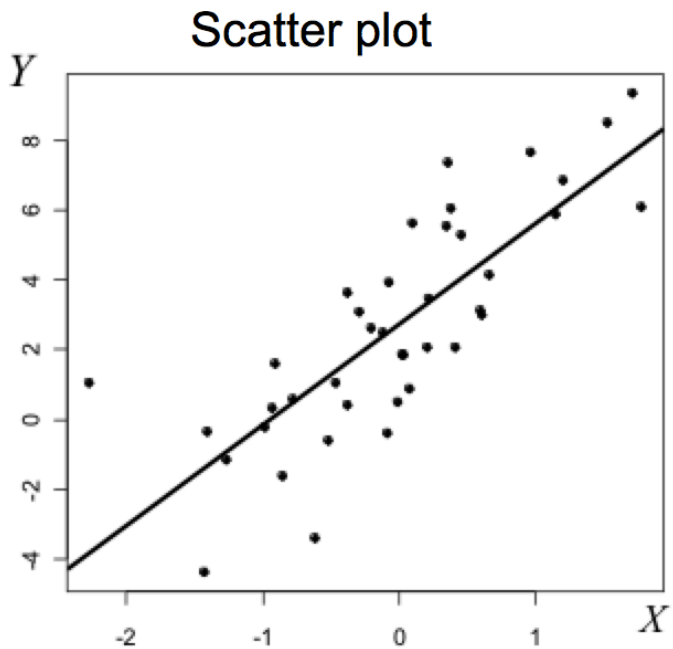
- Bimodal distribution: Normal QQ-plot

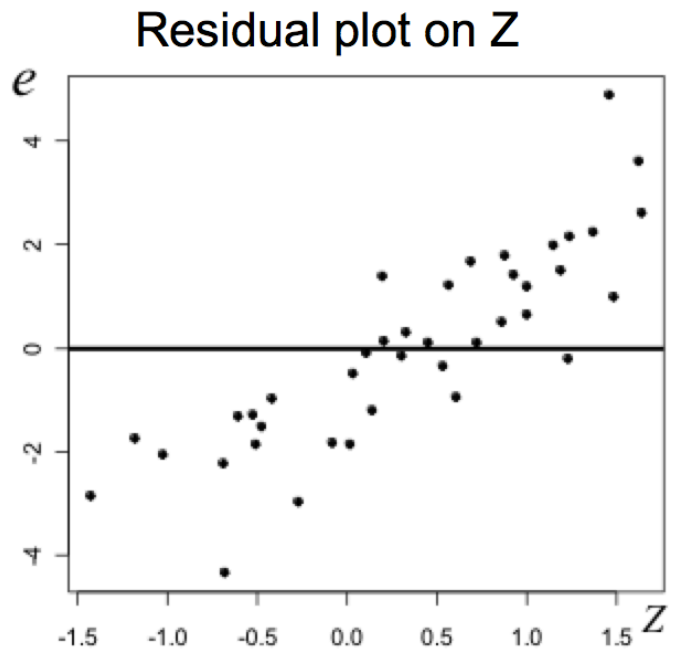




6.) Checking for important predictor variables

- Residual plot vs possible predictor variables
  - Fit $Y = b_0 + b_1 X + e$
  - Z is another potential predictor varible, plot e vs Z





Remedical Measures

- Even if one or more of the diagnostics show problems, the fitted model may be improved
  - 1. **Regression function is not linear**: in some cases, a *variable transformation* can make the data "more linear". If not, a different (e.g. nonlinear) model might be better.
  - 2. **Error terms are not independent**: If dependence can be modeled (e.g. serially correlated ϵ's), it can be incorporated in regression.
  - 3. **Error terms do not have constant σ^2**: if you can model variance as a function of X, an weighted regression can be used. Sometimes, variable transformation also stabilize variance.

Remedical Measures

- be continued...
  - 4. **Presence of outliers**: Sometimes outliers can be omitted from the data such as entry error. Alternatively, different fitting methods can be used to reduce their impact (robust regression).
  - 5. **Error terms are not normally distributed**: Often, variable transformations can help. If not, try modelling the error term with a different distribution.
  - 6. **One or more important predictor variables are omitted from model**: include them in the model (multiple regression).

Influence Metrics:
DFFITS, DFBETAS, Cook's Distance

DFFITS

$$DFFIT_i = \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\sqrt{MSE_{(i)} h_{ii}}}$$

Predict it from reg. with n cases

Predict it from reg. with n-1 cases (drop i^{th} case)

where

- $\hat{Y}_i$: i^{th} fitted value from a regression model based on sample of size n.
- $\hat{Y}_{i(i)}$: i^{th} fitted value from a regression model based on sample of size n-1 (dropping i^{th} data point).
- $MSE_{(i)}$: MSE from the fitted model without i^{th} data point.

What does DFFITS do?

- Defined as the change in the predicted value of a point when that point is removed.
- Measure of the influence of a point on a fitted value.

DFBETAS

$$DFBETA_{ij} = \frac{b_j - b_{j(i)}}{SE_{\beta_{j(i)}}}$$

- Defined as the standardized change in the regression coefficient j when point i is removed.
- Measure of the influence of a point on a regression coefficient.
- Usually don't bother with the intercept, so for SLR:

$$DFBETA_{i1} = \frac{b_1 - b_{1(i)}}{SE_{\beta_{1(i)}}} = \frac{b_1 - b_{1(i)}}{\sqrt{MSE_{(i)}/S_{XX}}}$$

Cook's Distance

$D_i > 1$: influential pts

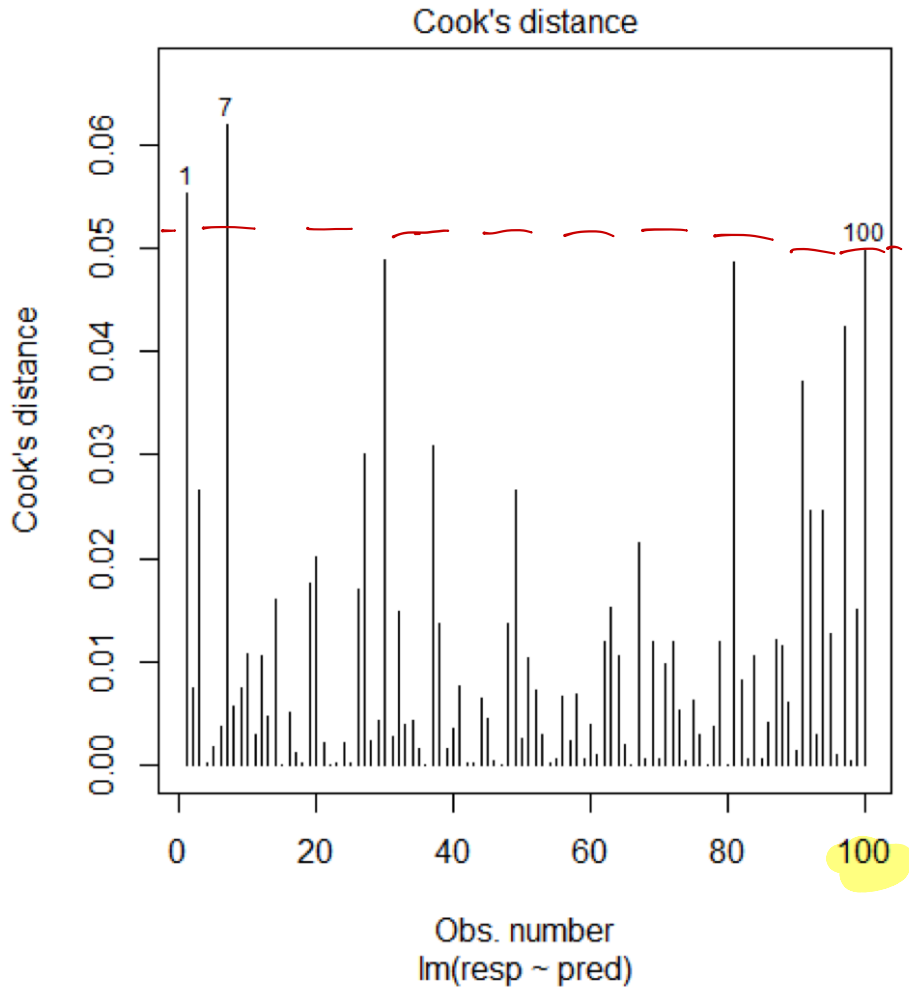
$D_i > \frac{4}{n}$ or $\frac{4}{n-k-1}$

$$D_i = \frac{\sum_j (\hat{Y}_{j(i)} - \hat{Y}_j)^2}{2 \cdot MSE} = \frac{r_i^2}{2} \frac{h_{ii}}{1 - h_{ii}}$$

where r_i is the studentized residual.

- Cook's distance refers to how far, on average, predicted y-values will move if the observation in question is dropped from the data set.
- Some texts tell you that points for which Cook's distance is higher than 1 are to be considered as influential. Other texts give you a threshold of $4/n$ or $4/(n-k-1)$ where n is the sample size and k is the number of predictor variables.

Cook's Distance



$n = 100$

$D_i > \frac{4}{n-k-1}$

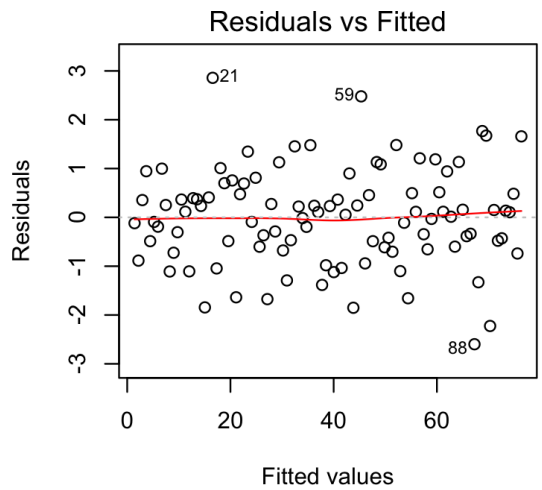
Points 1, 7, and 100 are influential points.

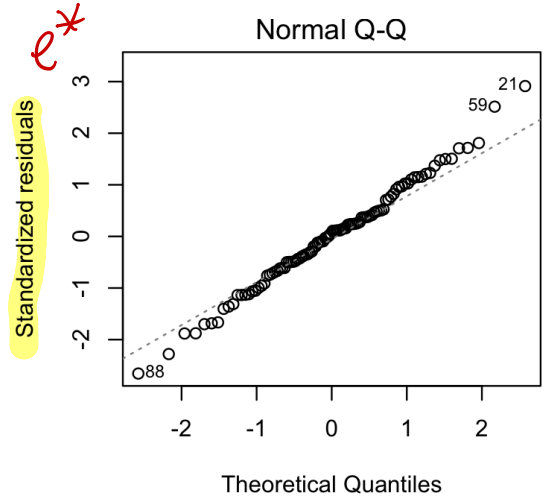
Summary of Influence Metrics

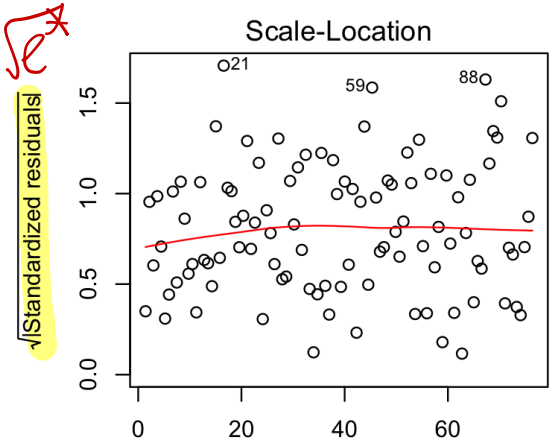
- DFFITS
  - Measure of the influence of a point on a fitted value.
- DFBETAS
  - Measure of the influence of a point on a regression coefficient.
- Cook's Distance
  - Combined measures of the leverage and outliers magnitude of a point.
- More discussion
  - Cook's distance refers to how far, on average, predicted y-values will move if the observation in question is dropped from the data set. DFBETA refers to how much a parameter estimate changes if the observation in question is dropped from the data set.
  - Cook's distance is presumably more important to you if you are doing predictive modelling, whereas DFBETA is more important in explanatory modelling.

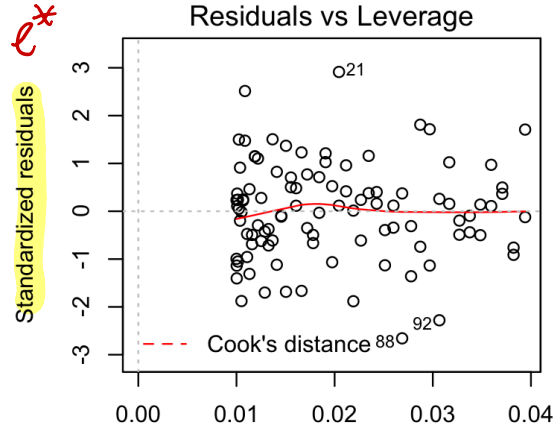
Simulation study

```
set.seed(1000); x = 1:100; y=1+0.75*x+rnorm(100,0,1) # generate x and y
par(mfrow=c(2,2)) # split the panel as 2 by 2
plot(lm(y~x))     # model diagostic
```









All suggest that assumptions for SLR hold.

Simulation study

```
set.seed(1000); x = 1:100; y=1+0.75*x+rnorm(100,0,1) # generate x and y
par(mfrow=c(2,2)) # split the panel as 2 by 2
influence.measures(lm(y~x))$infmat[1:7,]
```

```
##         dfb.1_       dfb.x       dffit    cov.r       cook.d       hat
## 1 -0.02468749  0.02132701 -0.02468841 1.062272 0.0003078536 0.03940594
## 2 -0.18151974  0.15600696 -0.18154782 1.043384 0.0165085793 0.03822982
## 3  0.07106854 -0.06075510  0.07109404 1.057165 0.0025497843 0.03707771
## 4  0.18737375 -0.15929894  0.18749711 1.038503 0.0175878587 0.03594959
## 5 -0.09465330  0.08001053 -0.09475381 1.052175 0.0045238242 0.03484548
## 6 -0.01779918  0.01495622 -0.01782729 1.056196 0.0001605292 0.03376538
## 7 -0.03571379  0.02982398 -0.03579311 1.054417 0.0006469247 0.03270927
```

Case study: residual plot

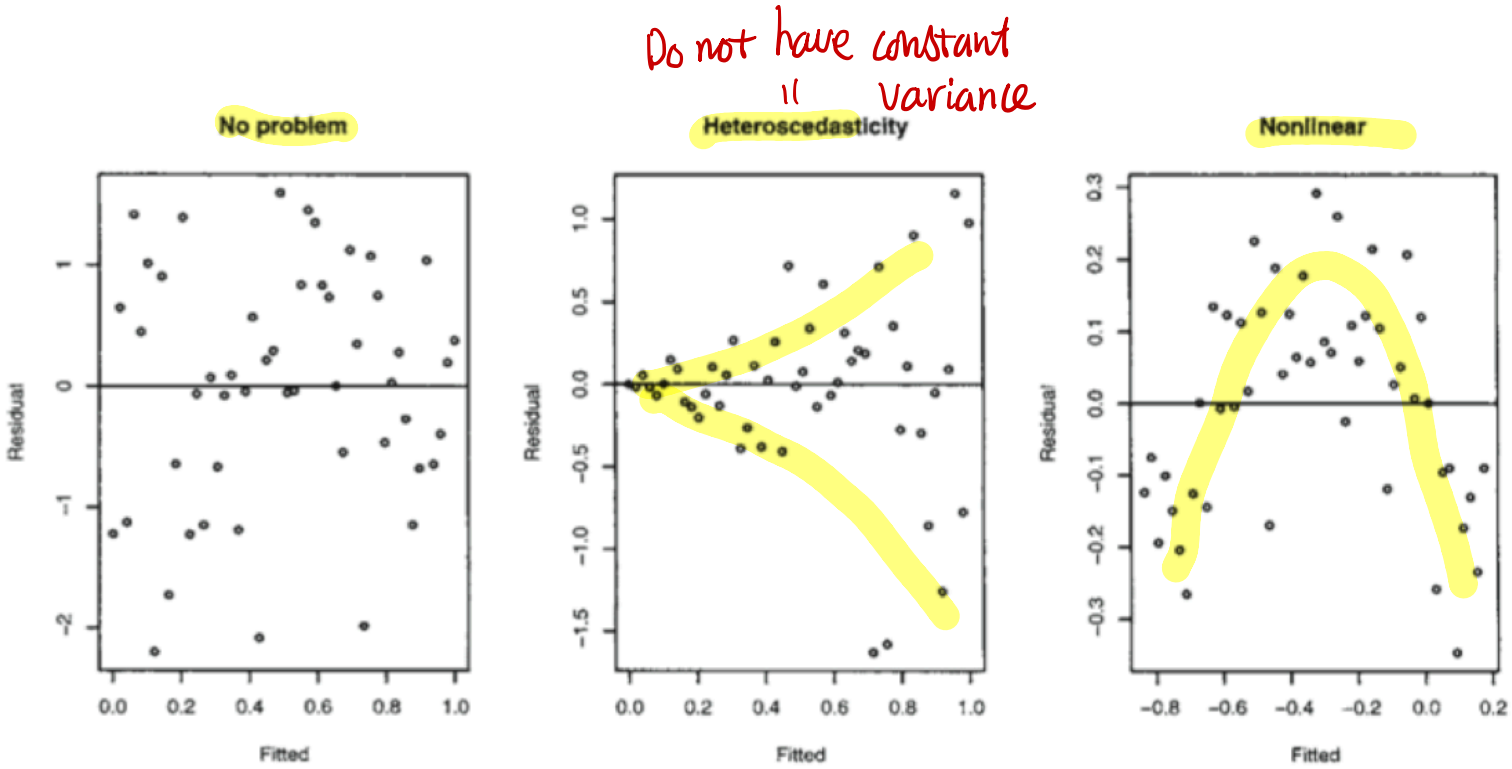
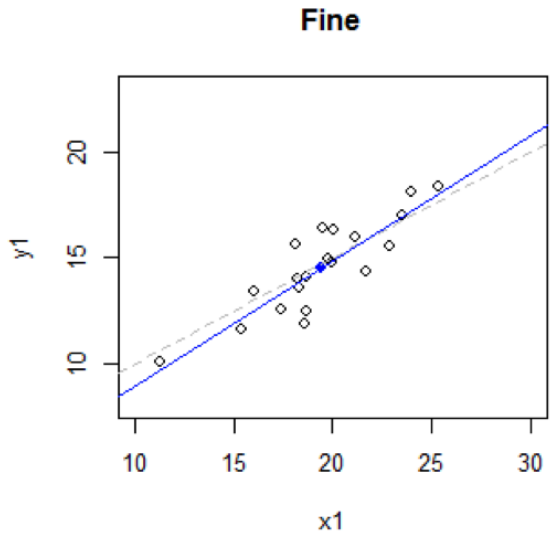
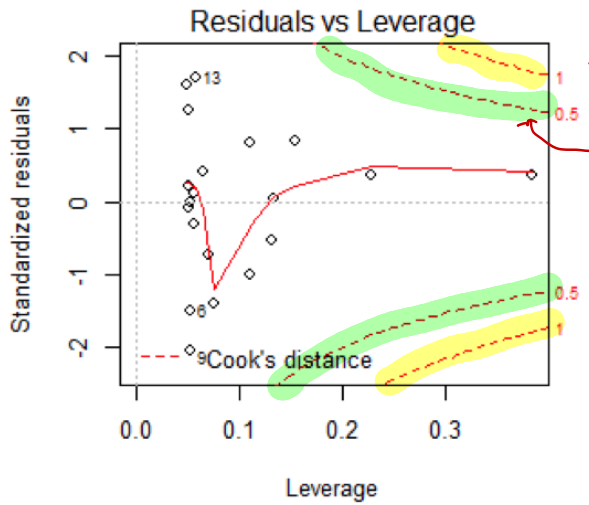


Figure 4.1 *Residuals vs. fitted plots—the first suggests no change to the current model while the second shows nonconstant variance and the third indicates some nonlinearity, which should prompt some change in the structural form of the model.*

From Faraway's *Linear Models with R* (2005)

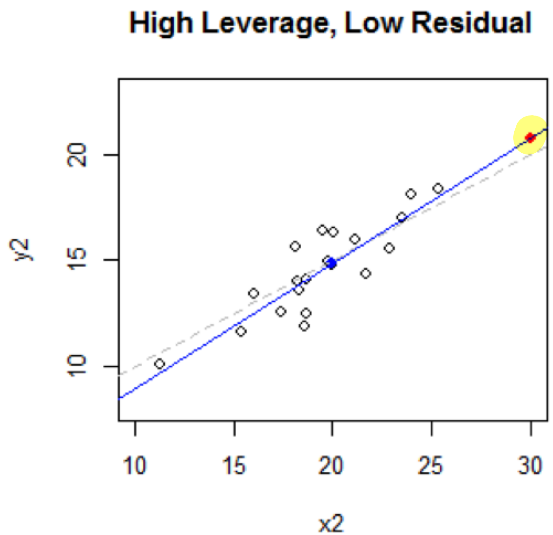
Case study: Residual vs Leverage

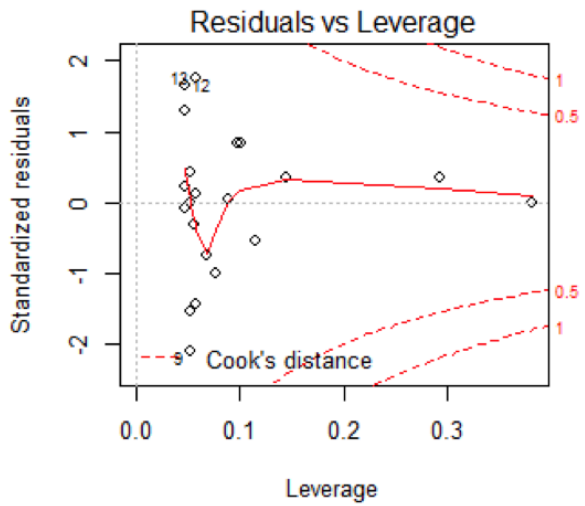




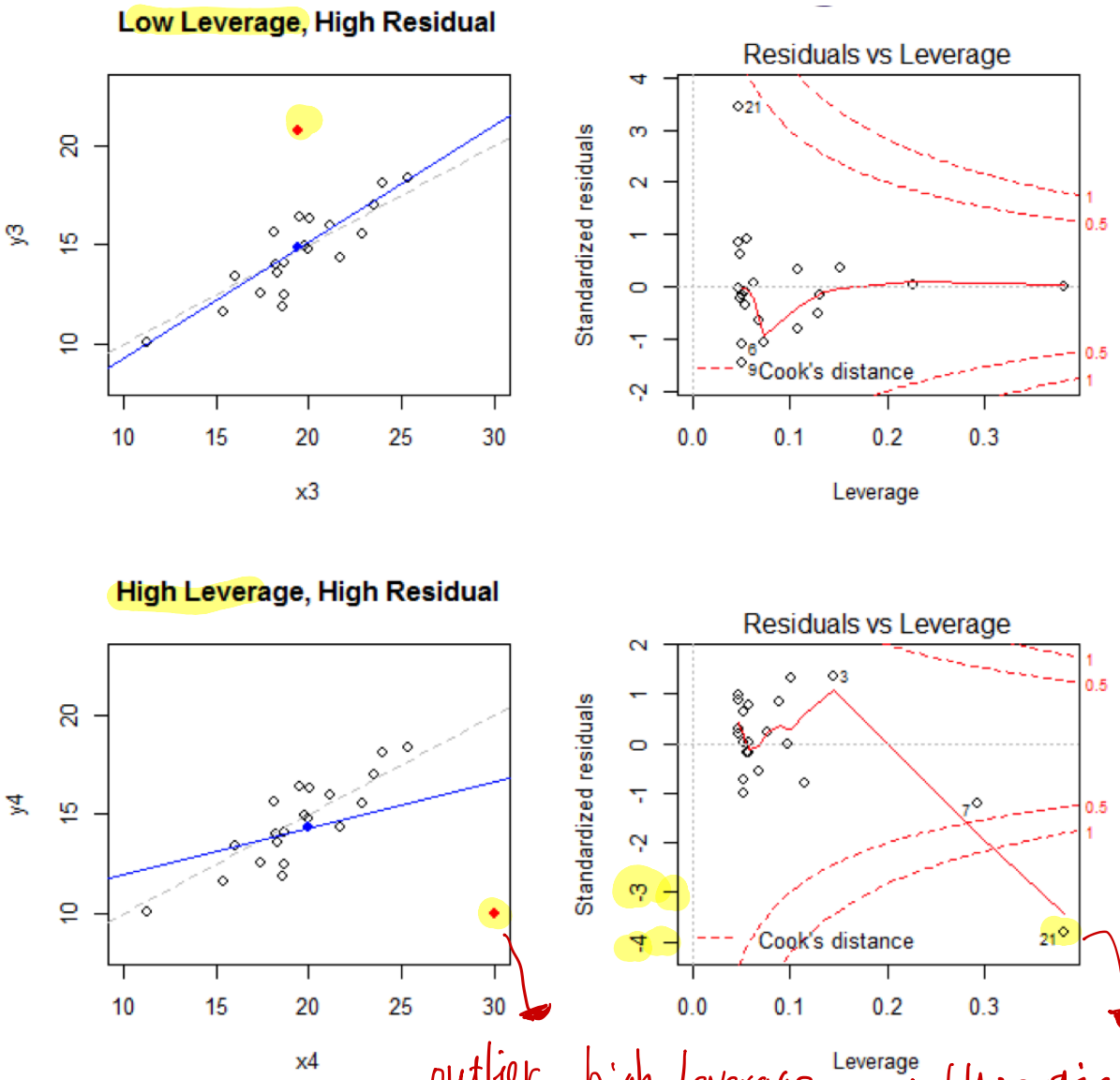
= cook Dist = 1

= cook Dist = 0.5

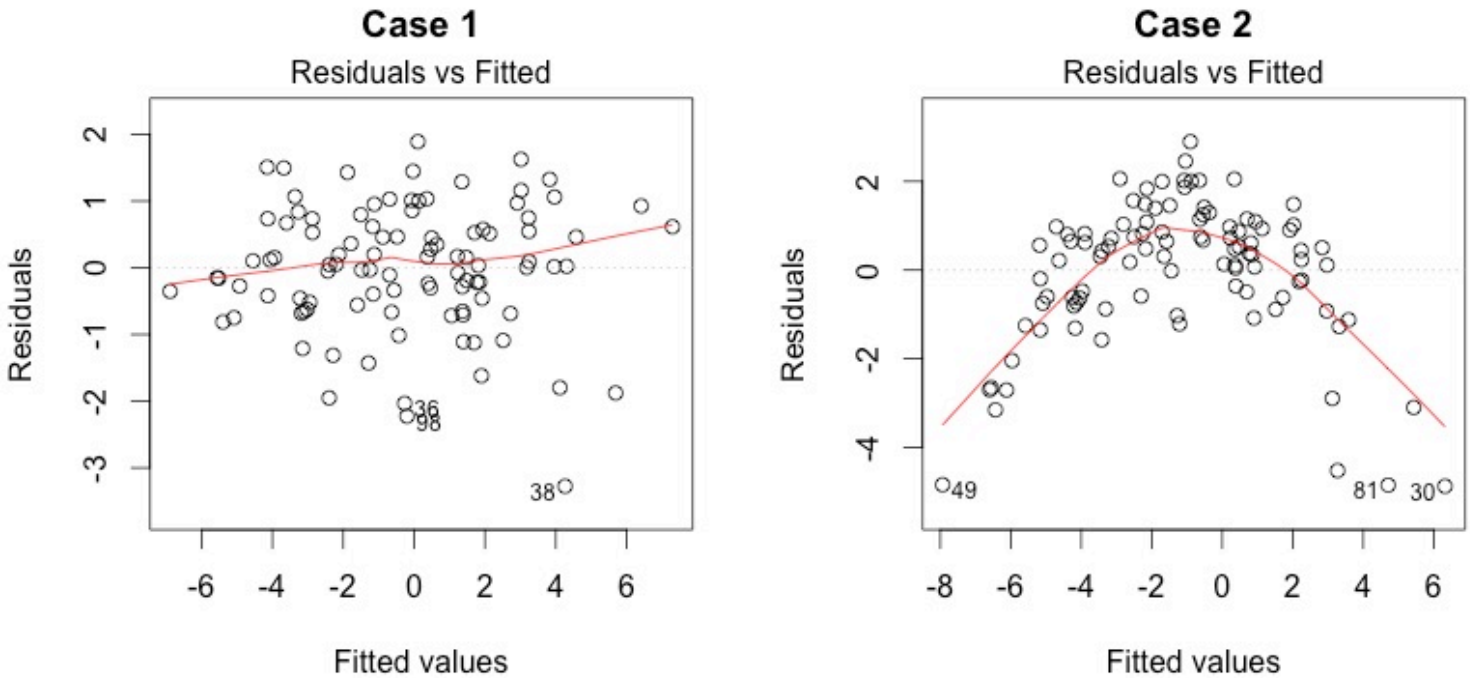




Case study: Residual vs Leverage

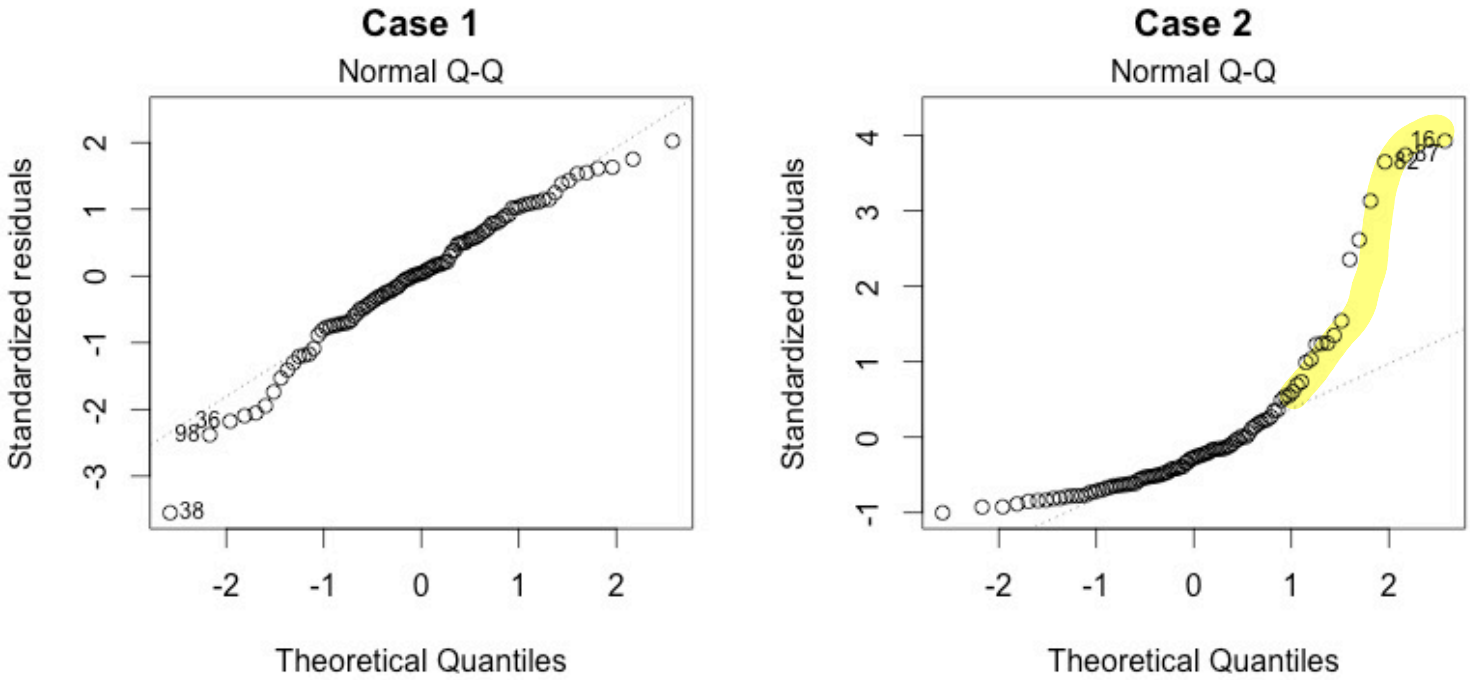


A: linearity checking



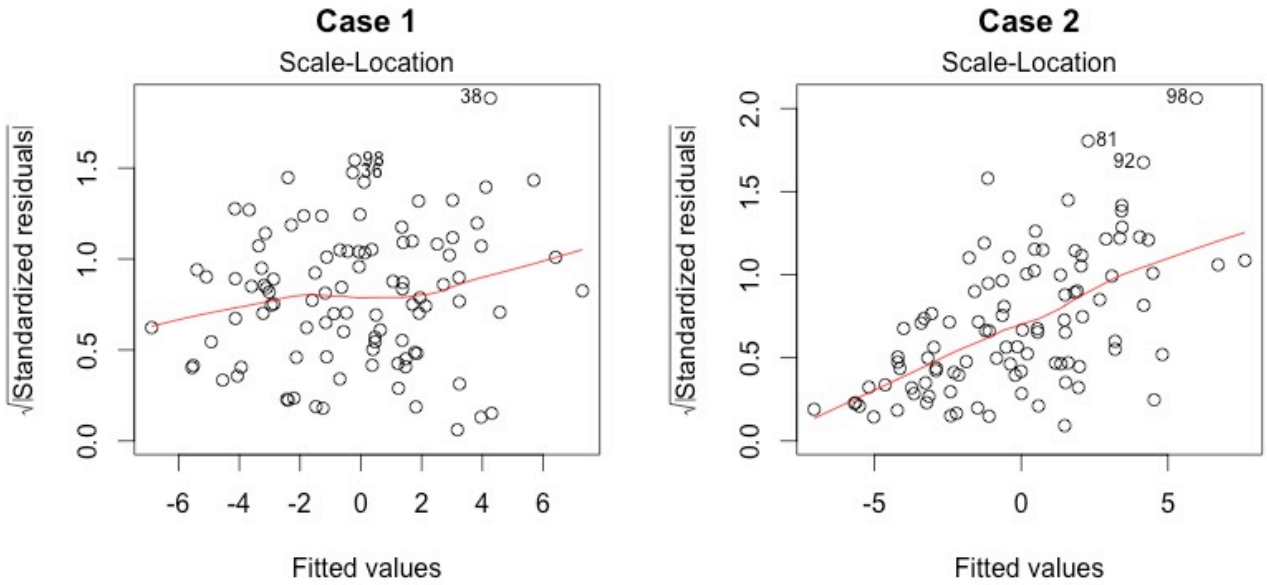
- Residuals vs Fitted
  - No distinctive pattern in Case 1.
  - A clear pattern (a parabola) in case 2. Nonlinearity exists.

B: Normal Q-Q plot



- Normal Q-Q plot
  - Most data points lie on straight line and this normality assumption seems ok.
  - Observe heavy right tail in case 2. Data is right-skewed distributed.

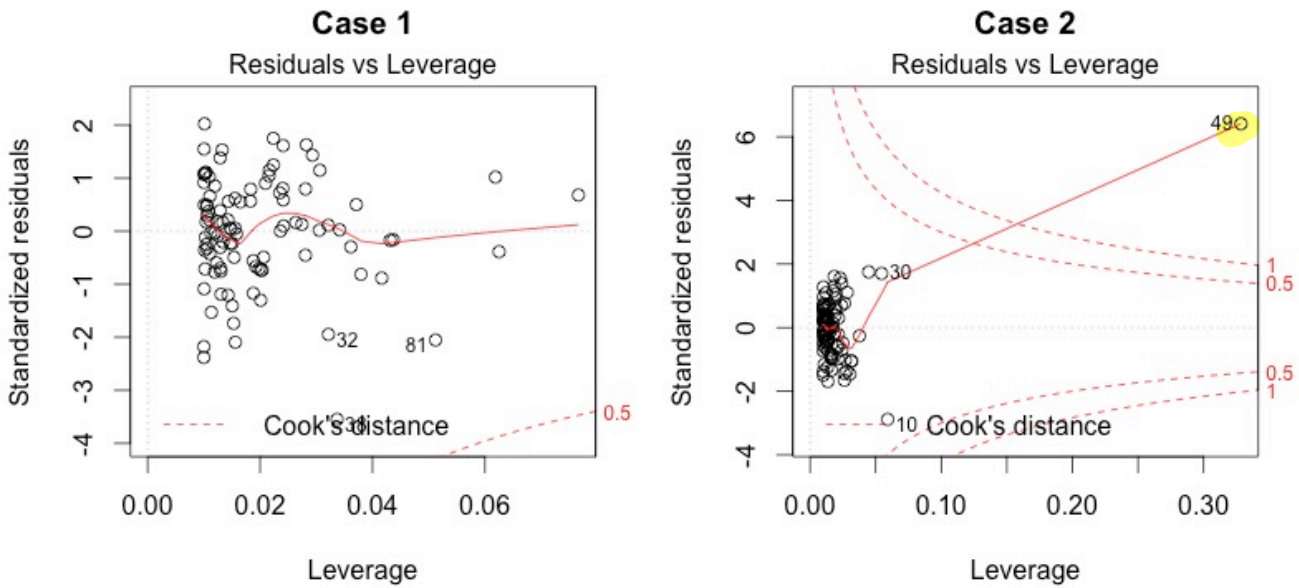
C: Scale-Location



- Scale-Location plot, is also called Spread-location plot
  - This plot shows if residuals are spread equally along the ranges of predictors.
  - This is how you can check the assumption of equal variance (homoscedasticity).
  - In Case 1, the residuals appear randomly spread.
  - In Case 2, the residuals begin to spread wider along the x-axis as it passes around 5.

D: Residuals vs Leverage

- This plot helps us to find influential cases (i.e., subjects) if any.
- Look for cases outside of a dashed line, Cook's distance (upper right corner or lower right corner). When cases are outside of the Cook's distance (have high Cook's distance scores), the cases are influential to the regression results. The regression results will be altered if we exclude those cases.



- Case 1 is the typical look when there is no influential case, or cases.
- In case 2, the plot identified the influential observation as #49

Practice problems and upcoming topics

- Practice problems after today's lecture: Chapter 3: 3.1, 3.2, 3.4 (b, d, e – no Table lookup, f, h), 3.5 (b-f), 3.9, 3.18, 3.19, 3.20.
- Show $Cov(e_i, \hat{Y}_i) = 0$ and $Cov(e_i, \hat{Y}_j) = 0, i \neq j$
- Upcoming topics
  - Variable transformation
  - Ch4: other topics in SLR