Residuals: $e_i = y_i - \hat{y}_i$

Assessing Assumptions: $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$

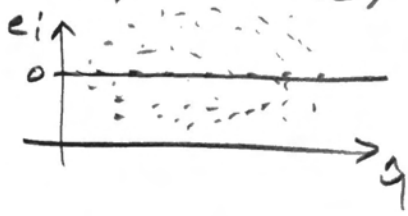
(1) Normality: e_i should look like a sample from a Normal R.V.
↳ boxplots, q-q plots, histograms, etc. (outlier analysis)

(2) Homoskedasticity: Residual plots should not show trends in $X_1, X_2, \ldots$ or $\hat{y}_i$

(3) Linearity:

→ plots of e_i vs. $\hat{y}_i$, e_i vs. X_{i1}, e_i vs. X_{i2}, ---

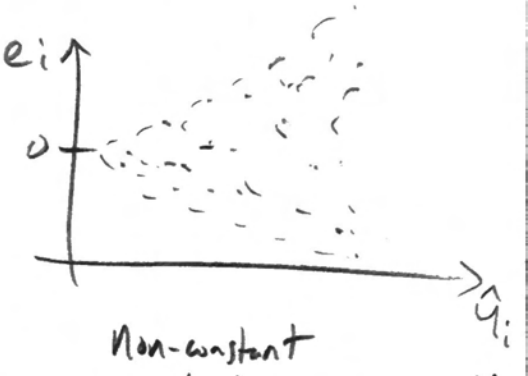
Expect a patternless scatter



Potential Problem Plots:

An important variable may have been omitted

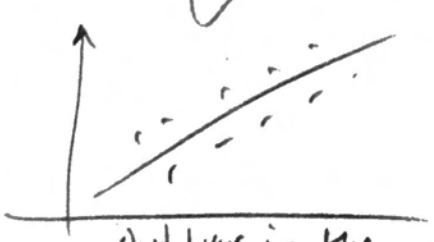
A squared or interaction term has been left out



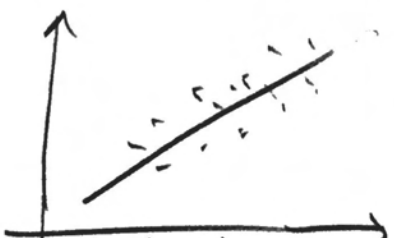
Non-constant variance
⇒ transform y_i?

(4) Independence: typically not assessed using residuals
but instead assessed via the experimental design
or context of data acquisition

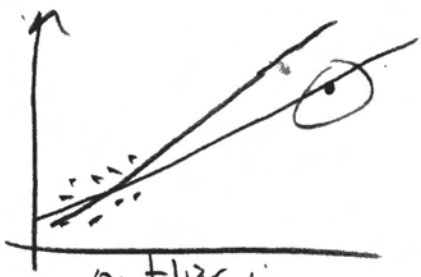
(•) Residuals can be used to assess serial correlation



Outlier in the y-direction



Outlier in the x-direction



Outlier in both directions

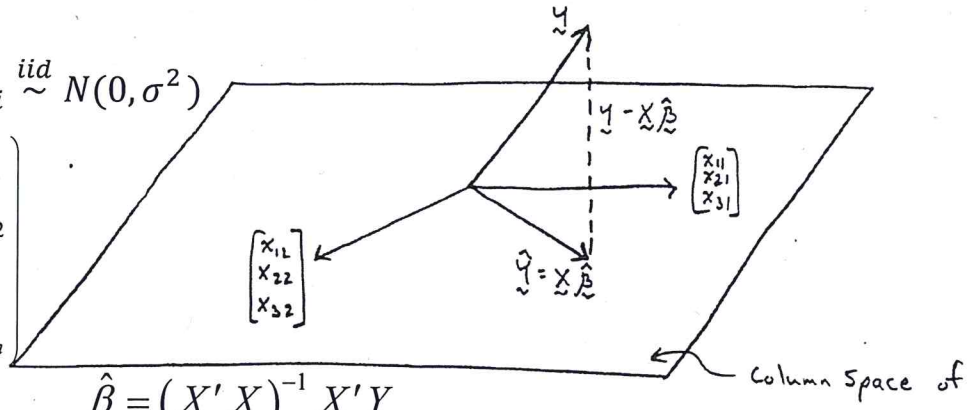
Goal: identify potential outliers and study them

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i \,, \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & X_{11} & X_{21} \\ 1 & X_{12} & X_{22} \\ \vdots & \vdots & \vdots \\ 1 & X_{1n} & X_{2n} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

$$\underset{\sim}{Y} = \underset{\sim}{X} \quad \underset{\sim}{\beta} + \underset{\sim}{\varepsilon}$$

$$\hat{\underset{\sim}{\beta}} = \left(\underset{\sim}{X}' \underset{\sim}{X}\right)^{-1} \underset{\sim}{X}' \underset{\sim}{Y}$$



Leverages (h_{ii}) for SLR

$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ $\quad \underset{\sim}{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$ $\quad \underset{\sim}{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$

$H = \underset{\sim}{X}(\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{X}' = [h_{ij}]$

- $h_{ij} = \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{\sum_{\ell=1}^{n}(x_\ell - \bar{x})^2}$

- $h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}$

h_{ii} measures how far x_i is from the center of the X data

- $\bar{h} = \sum_{i=1}^{n} h_{ii}/n = \frac{k+1}{n}$

Any $h_{ii} > 2\frac{(k+1)}{n}$ should be looked at closely to determine if it is overly influential in fitting the model.

Properties of Residuals

1. $\sum e_i = 0$

2. $\frac{\sum e_i^2}{n-(k+1)} = MSE = s^2_{y|x} = s^2$ (in ch. 14)

3. Same scale as the y_i observations

4. normalized residuals are scale independent

Normalized Residuals

	text	R	Other
$z_i = \frac{e_i - 0}{s}$	Standardized	NA	Not widely used
$r_i = \frac{e_i - 0}{s\sqrt{1-h_{ii}}}$	studentized	rstandard( )	internally studentized
$r_{(-i)} = \frac{e_i - 0}{s_{(-i)}\sqrt{1-h_{ii}}}$	jackknife	rstudent( )	externally studentized

* $s_{(-i)}$ is the MSE computed if we leave out the i^{th} observation

Testing for outliers (in the y-direction)

- $r_{(-i)}$ is better modeled by a T-distribution than r_i

H_0: $r_{(-i)}$ is not from an outlier

Cook's Distance

- Measures the average standardized distance between $\hat{\underset{\sim}{\beta}}$ and $\hat{\underset{\sim}{\beta}}_{(-i)}$
- Measures influence in both the X and Y direction

$$D_i = \left(\frac{r_i^2}{p+1}\right)\left(\frac{h_{ii}}{1-h_{ii}}\right)$$

- If the model is correct $D_i \sim F_{p,\, n-(p+1)}$

Cook's Distance (D_i) for the i^{th} observation is the average standardized distance between $\hat{\beta}$ and $\hat{\beta}_{(-i)}$

```
> dat <- otherdata("gas.dat")
> f.t <- lm(FUEL~TAX)
> anova(f.t)
     Response: FUEL
               Df Sum Sq Mean Sq F value    Pr(>F)
     TAX        1 119690  119690  11.747 0.001294
     Residuals 46 468707   10189

> mse = summary(f.t)$sigma^2 #or anova(f.t)[2,3}

> X    = <- cbind(rep(1,length=48), TAX) #Design matrix
> XtX = t(X) %*% X                       # X'X   matrix

#Compute distance between β̂ and β̂(-40)
> coef1 = lm(FUEL~TAX)$coef
     (Intercept)         TAX
       983.75343   -53.07681

> coef2 = lm(FUEL[-40]~TAX[-40])$coef
     (Intercept)     TAX[-40]
       931.77832   -47.29712

> diff = coef1 - coef2
     (Intercept)         TAX
       51.975107   -5.779688

>(983.75-931.78)^2 + (-53.08 - -47.3)^2
  sum(diff^2)       #2734.8
  t(diff) %*% diff #2734.8



#Compute standardized distance between β̂ and β̂(-40)
> cooks.distance(f.t)[40]
           40
     0.2076523

> ( t(diff) %*% XtX %*% diff ) / (2*mse)
     [1,] 0.2076523


> r.i <- rstandard(f.t) #internally studentized resids
> h.i <- hatvalues(f.t) #leverage or "hat" values
> ( r.i[40]^2 * h.i[40] )/( (1+1) * (1 - h.i[40]) )
           40
     0.2076523

> #Assessing importance
>  pf(.2077,1,46)
     [1] 0.3492797
```

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \begin{pmatrix} 983.75 \\ -53.08 \end{pmatrix}$$

$$\hat{\beta}_{(-40)} = \begin{pmatrix} 931.78 \\ -47.30 \end{pmatrix}$$

$\left\| \hat{\beta} - \hat{\beta}_{(-40)} \right\|$ = the length of this difference

$$D_i = \frac{r_i^2}{k+1} \left(\frac{h_{ii}}{1 - h_{ii}} \right)$$

If the model is correct $D_i \sim F_{k, n-(k+1)}$