Response variable: Y. Explanatory variable: $x_1 \cdots x_n$

Simple Linear Regression

Pearson Coefficient: $\rho = \frac{\text{cov}(X,Y)}{\sqrt{\text{Var}(X)\,\text{Var}(Y)}}$

$$S_{xx} = \sum (x_i - \bar{x})^2 = \sum x_i^2 - n\bar{x}^2$$
$$S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - n\bar{x}\bar{y}$$
$$S_{yy} = \sum (y_i - \bar{y})^2 = \sum y_i^2 - n\bar{y}^2$$

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

Least Square Regression:

$$\hat{\beta_1} = \frac{S_{xy}}{S_{xx}} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}}$$
$$\hat{\beta_0} = \bar{y} - \hat{\beta_1}\bar{x}$$

$r_i = y_i - \hat{y_i} = y_i - \hat{\beta_0} - \hat{\beta_1} x_i$
$\sum r_i = 0, \quad \sum r_i x_i = 0 \quad \sum r_i \hat{y_i} = 0$

2 estimators for σ^2: $s^2 = \frac{1}{n-1}\sum (y_i - \bar{y})^2 = \frac{1}{n-1} S_{yy}$
$df = n-1$

OR Mean Square Error:
$$MSE = \frac{1}{n-2}\sum r_i^2 \qquad (\hat{\beta_0} + \hat{\beta_1} \text{ estimator}, \ df = n-2)$$

Maximum-Likelihood: $L(\theta) = f(y_i; \theta)$, $l(\theta) = \log(L(\theta))$
Score function: $s(\theta) = \frac{\partial}{\partial \theta} l(\theta) \rightarrow$ solve this to get $\hat{\theta}$
MLE and LSE has the same $\hat{\beta_0}$ and $\hat{\beta_1}$
Estimator for σ^2: $\frac{1}{n}\sum r_i^2$ (DONT USE $\rightarrow$ biased)
* Note that estimator is a r.v., while estimate is the real value.

Assumptions: $E[\hat{\beta_0}] = \beta_0, \quad E[\hat{\beta_1}] = \beta_1.$

$$\text{Var}[\hat{\beta_0}] = \sigma^2 \frac{\sum x_i^2}{n S_{xx}} \qquad \text{Var}[\hat{\beta_1}] = \frac{\sigma^2}{S_{xx}}$$

$\text{Cov}(y_i, \hat{\beta_1}) = 0 \qquad \text{Cov}(\hat{\beta_0}, \hat{\beta_1}) = -\sigma^2 \frac{\bar{x}}{S_{xx}}$

Assume $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2) \rightarrow \hat{\beta_0} \sim N(\beta_0, \frac{\sigma^2 \sum x_i^2}{n S_{xx}}) = N(\beta_0, \frac{s^2 \sum x_i^2}{n S_{xx}})$

$\hat{\beta_1} \sim N(\beta_1, \frac{\sigma^2}{S_{xx}}) = N(\beta_1, \frac{s^2}{S_{xx}})$

Use $s^2 = \frac{1}{n-2} \sum r_i^2$ to replace σ^2.

Standard Error = $SE(\hat{\beta_1}) = \sqrt{Var[\hat{\beta_1}]} = \sqrt{\frac{s^2}{S_{xx}}} \sim t_{n-2}$

$SE(\hat{\beta_0}) = \sqrt{Var[\hat{\beta_0}]} = \sqrt{\frac{s^2 \sum x_i^2}{n S_{xx}}} \sim t_{n-2}$

Confidence Interval: $\hat{\beta_0} \pm t_{\alpha/2, n-2} \cdot SE(\hat{\beta_0})$

$\hat{\beta_1} \pm t_{\alpha/2, n-2} \cdot SE(\hat{\beta_1})$

Hypothesis Testing: Reject H_0 if $|t| = \left| \frac{\hat{\beta_1} - \beta^*}{SE(\hat{\beta_1})} \right| > t_{\alpha/2, n-2}$ (upper)

p-value: if $T \sim t_{n-2}$, $Pr(|T| > |t|) = 2P(T > |t|)$

Reject H_0 if p-value $< \alpha$

Significance Test: Hypothesis Test when $\beta^* = 0$

* One-sided Test: $H_0: \beta_1 \geq \beta^*$ $H_a: \beta_1 < \beta^*$

Reject H_0 if $t < -t_{\alpha/2, n-2}$ or p-value $= P(T < t)$

Usually don't use this, only if intended.

CANNOT use this after a failed Two-sided Test.

* If H_0 is $\leq$, reject when $t < -t_{\alpha/2}$

Else if H_0 is $\geq$, reject when $t > t_{\alpha/2}$

Estimate of Mean Response: $\mu = E[y | x_p] = \beta_0 + \beta_1 x_p$

$E[\hat{\mu}] = \mu$. $Var[\hat{\mu}] = \sigma^2 [\frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{xx}}]$

$SE[\hat{\mu}] = \sqrt{s^2 [\frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{xx}}]}$ $\frac{\hat{\mu} - \mu}{SE[\hat{\mu}]} \sim t_{n-2}$

Confidence Interval: $\hat{\mu} \pm t_{\alpha/2, n-2} \cdot SE[\hat{\mu}]$

Prediction $\hat{y_p} = \hat{\beta_0} + \hat{\beta_1} x_p$ prediction error $= y_p - \hat{y_p}$

$E[y_p - \hat{y_p}] = 0$, $Var[y_p - \hat{y_p}] = \sigma^2 + \sigma^2 [\frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{xx}}]$

$$\frac{y_p - \hat{y}_p - 0}{SE[y_p - \hat{y}_p]} \sim t_{n-2} \qquad SE[y_p - \hat{y}_p] = \sqrt{s^2 + s^2\left[\frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{xx}}\right]}$$

Confidence Interval = $\hat{y}_p \pm t_{\alpha/2, n-2} \cdot SE[y_p - \hat{y}_p]$

wider than CI for mean response at $x = x_p$

ANOVA

Regression	$SSR = \sum(\hat{y}_i - \bar{y})^2 = \hat{\beta}_1^2 S_{xx}$	$df = 1$	$MSR = \frac{SSR}{df} = SSR$
Error	$SSE = \sum(y_i - \hat{y}_i)^2 = \sum r_i^2$	$df = n-2$	$MSE = \frac{SSE}{df} = \frac{1}{n-2}\sum r_i^2$
Total	$SST = \sum(y_i - \bar{y})^2 = S_{yy}$	$df = n-1$	

$E[MSE] = \sigma^2$. $E[MSR] = \sigma^2 + \beta_1 S_{xx}$

MSR would be larger than MSE when $\beta_1 \neq 0$

Significance Test: $H_0: \beta_1 = 0$

$$F = \frac{MSR}{MSE} \sim F_{1, n-2} \qquad R^2 = \frac{SSR}{SST}$$

Reject H_0 if $F > F_{\alpha; 1, n-2}$ — bigger the better

Multiple linear Regression

$Y = [y_1, \cdots, y_n]^T \in \mathbb{R}^{n \times 1}$, $y_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_p x_{i,p} + \varepsilon_i$

Design Matrix: $\begin{bmatrix} 1 & x_{1,1} & \cdots & x_{1,p} \\ \vdots & & \ddots & \\ 1 & x_{n,1} & \cdots & x_{n,p} \end{bmatrix} \in \mathbb{R}^{n \times (p+1)}$

$\beta = [\beta_0, \cdots, \beta_p]^T \in \mathbb{R}^{(p+1) \times 1}$

$\varepsilon = [\varepsilon_1 \cdots \varepsilon_n]^T \in \mathbb{R}^{n \times 1}$

$$Var[Y] = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \vdots & \sigma_2^2 & \ddots & \vdots \\ \vdots & \vdots & & \sigma_n^2 \end{bmatrix} = \Sigma \in \mathbb{R}^{n \times n}$$

if y_i are independent $\leftrightarrow$ diagonal.

* $E[AY + B] = AE[Y] + B$

$Var[AY + B] = A\,Var(Y)\,A^T$ (if Y is scalar, $Var[AY+B] = A^2 Var[Y]$)

If $Y \sim MVN(\mu, \Sigma)$,

- any partition of $Y = Y_1, Y_2$, $Y_1 \sim MVN(\mu_1, \Sigma_{11})$, $Y_2 \sim MVN(\mu_2, \Sigma_{22})$

  $\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$, $\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$
- $V = AY$, $W = BY$, W & V are independent $\leftrightarrow A\Sigma B^T = 0$

Hypothesis Testing $\beta_j = 0 \rightarrow$ all β are 0, x_j not linearly related to y_j.

$Y = X\beta + \varepsilon \rightarrow$ since $\varepsilon \sim MVN(0, \sigma^2 I)$, $Y \sim (X\beta, \sigma^2 I)$

Least Squares $\hat{\beta} = (X^T X)^{-1} X^T Y$ $\quad E[\hat{\beta}] = \beta$, $Var[\hat{\beta}] = \sigma^2 (X^T X)^{-1}$

$\hat{\beta} \sim MVN(\beta, \sigma^2 (X^T X)^{-1})$, $\hat{\beta}_j \sim N(\beta_j, \sigma (X^T X)^{-1}_{jj})$

Fitted Value $\hat{Y} = X(X^T X)^{-1} X^T Y$, $H = X(X^T X)^{-1} X^T$

symmetric & idempotent

$H^T = H$ $\quad HH = H$

$r = Y - \hat{Y} = (1-H)Y$ $\quad E[r] = 0$, $Var[r] = \sigma^2 (I - H)$

$r \sim MVN(0, \sigma^2 (I-H))$

Same as in linear regression, $\sum r_i = 0$

$(X^T r = 0)$ $\quad \sum r_i x_{i,1} = 0 = \cdots = \sum r_i x_{ip} = 0$

$\sum r_i \hat{y}_i = 0$

Estimator of σ^2 $\frac{SSE}{\sigma^2} = \frac{1}{\sigma^2} r^T r \sim \chi^2_{n-(p+1)}$

$E[\frac{SSE}{\sigma^2}] = n-(p+1) \rightarrow E[SSE] = (n-p-1)\sigma^2$ $\quad *SSE = \sum r_i^2$

$MSE = \frac{SSE}{n-(p+1)}$ is unbiased estimator of σ^2.

Confidence Interval: $\hat{\beta}_j \pm t_{\alpha/2, n-(p+1)} SE[\hat{\beta}_j]$

$SE[\hat{\beta}_j] = \sqrt{s^2 (X^T X)^{-1}_{jj}}$

Hypothesis Test: $H_0 : \beta_j = \beta^*$

Reject H_0 if $|t| = \left| \frac{\beta_j - \beta^*}{SE[\hat{\beta}_j]} \right| > t_{\alpha/2; n-(p+1)}$

We can only fit linear model when $n > p+1$ since $(X^T X)$ should be invertible = $rank(X) = p+1$

Combinations of Coefficients

Estimate mean response at $c = (1, x_1, \cdots, x_p)^T$.

$$\hat{\mu}_c = c^T\hat{\beta} \sim MVN(c^T\beta,\ c^T[\sigma^2(X^TX)^{-1}]c)$$

$$SE[\hat{\mu}_c] = \sqrt{c^T[\hat{\sigma}^2(X^TX)^{-1}]c} \qquad \frac{c^T\hat{\beta} - c^T\beta}{SE[c^T\hat{\beta}]} \sim t_{n-(p+1)}$$

$$\downarrow$$

$$MSE = \frac{SSE}{n-(p+1)}$$

Confidence Interval = $\hat{\mu}_c = c^T\hat{\beta} \pm t_{\alpha/2,\, n-(p+1)}\, SE[c^T\hat{\beta}]$

Prediction

$$y_p = c^T\beta + \varepsilon_p, \quad \hat{y}_p = c^T\hat{\beta}$$

$$E[y_p - \hat{y}_p] = 0, \quad Var[y_p - \hat{y}_p] = \sigma^2 + c^T[\sigma^2(X^TX)^{-1}]c$$

$$SE[y_p - \hat{y}_p] = \sqrt{\hat{\sigma}^2 + c^T[\hat{\sigma}^2(X^TX)^{-1}]c}$$

$$\frac{y_p - \hat{y}_p - 0}{SE[y_p - \hat{y}_p]} \sim t_{n-(p+1)}$$

Prediction Interval: $\hat{y}_p \pm t_{\alpha/2,\, n-(p+1)}\, SE[y_p - \hat{y}_p]$

ANOVA

Regression	$SSR = \sum(\hat{y}_i - \bar{y})^2$	$df = p$	$MSR = SSR/p$
Error	$SSE = \sum(y_i - \hat{y}_i)^2$	$df = n-p-1$	$MSE = SSE/(n-p-1)$
Total	$SST = \sum(y_i - \bar{y})^2$	$df = n-1$	

F-Test for Overall Significance

$H_0: \forall \beta_i = 0$ vs. $H_A: \exists \beta_i \neq 0$

$$F\text{-val} = \frac{SSR/p}{SSE/(n-p-1)} = \frac{MSR}{MSE} \sim F_{\alpha, p, n-p-1}$$

Reject H_0 if $F > F_{\alpha,p,n-p-1}$ or $Pr(>f) < 0.05$

Coefficient of Determination

$$R^2 = \frac{SSR}{SST}$$ – proportion of response variable that can be explained by the model.

Always increases as more variables added

$$R^2_{adj} = 1 - \frac{n-1}{n-p-1}(1-R^2)$$

Higher the better (close to R^2)

Categorical Variable

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 \underset{b}{x_{i,2}} + \beta_3 \underset{c}{x_{i,3}} + \beta_4 \underset{d}{x_{i,4}} + \varepsilon_i$$

β_2 = difference of mean response between type a and b.
β_3 = " " a and c.
β_4 = " " a and d.
$\beta_2 - \beta_3$ = " " b and c

$$\hat{\beta} \sim MVN(\beta, \sigma^2[X^TX]^{-1}), \quad \hat{\beta_j} \sim N(\beta_j, \sigma^2[X^TX]^{-1}_{jj})$$

If test the difference $\beta_2 - \beta_3$, then

$$Var(\hat{\beta_2} - \hat{\beta_3}) = Var(\hat{\beta_2}) + Var(\hat{\beta_3}) - 2Cov(\hat{\beta_2}\hat{\beta_3})$$
$$\sigma^2(X^TX)^{-1}_{22} \quad \sigma^2(X^TX)^{-1}_{33} \quad \sigma^2(X^TX)^{-1}_{23}$$

Interaction Effect

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,1} x_{i,2} + \varepsilon_i$$
(β_1: main effect, β_2: main effect, β_3: interaction effect)

$* \; x_{i,1} + 1 \rightarrow y \mathrel{+}= \beta_1 + \beta_3 x_{i,2}$ (depend on $x_{i,2}$)

Test if the relationship between y and $x_{i,1}$ the same for different $x_{i,2}$ (usually categorical) ?
$H_0: \beta_3 = 0$.

General Linear Hypothesis

$A \in \mathbb{R}^{l \times (p+1)}$ $\quad H_0: A\beta = 0$ vs. $H_A: A\beta \neq 0$

If $\beta_2 = \beta_3 = 0 \rightarrow$ test if categorical variable doesn't matter at all.

$$A = \begin{bmatrix} 0 & 0 & 1 & 0 & \cdots \\ 0 & 0 & 0 & 1 & \cdots \end{bmatrix}$$

If $\beta_2 = \beta_3 \rightarrow$ no difference between categorical variable.

$$A = [0 \; 0 \; 1 \; -1 \; \cdots]$$

If $\beta_1 = 0$, $\beta_2 = \beta_3 \rightarrow$ no effect of x_1, and

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 & \cdots \\ 0 & 0 & 1 & -1 & \cdots \end{bmatrix}$$

$\text{rank}(A) = l$

$L(X) = \{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p\}$, $L_A(X) = \{\beta : A\beta = 0\} \subseteq L(X)$

$Y = \hat{Y}_A + (\hat{Y} - \hat{Y}_A) + r$, $r = Y - \hat{Y}$, $r_A = Y - \hat{Y}_A$

r_A

$\begin{cases} r \perp \hat{Y}_A \\ r \perp \hat{Y} - \hat{Y}_A \\ \hat{Y}_A \perp \hat{Y} - \hat{Y}_A \end{cases}$

$\|Y\|^2 = \|\hat{Y}_A\|^2 + \|\hat{Y} - \hat{Y}_A\|^2 + \|r\|^2$

$\|\hat{Y}_A\|^2$: unadjusted total SS

$\|\hat{Y} - \hat{Y}_A\|^2 = \|r\|^2 - \|r_A\|^2 = SSE - SSE_A$

$\|r\|^2$: SSE

If $A\beta = 0$ is true, then $SSE_A - SSE = \text{Extra SS}$ is small compared to SSE.

$$F\text{-statistic} = \frac{\|\hat{Y} - \hat{Y}_A\|^2 / l}{\|r\|^2 / (n-p-1)} = \frac{(SSE - SSE_A)/l}{SSE/(n-p-1)} \overset{H_0}{\sim} F_{2,l,n-p-1}$$

$SSE/(n-p-1)$: MSE

Reject H_0 if $f > F_{2,l,n-p-1}$

Residual & Random Errors

$r = Y - \hat{Y} = (I - H)X\beta + (I - H)\varepsilon = (I - H)\varepsilon$

If H is small relative to I, $r \cong \varepsilon \rightarrow r \sim MVN(0, \sigma^2 I)$

Standardized residual: $d_i = \frac{r_i}{\hat{\sigma}}$, $E[d_i] = 0$, $Var[d_i] \approx 1$

Studentized residual: $e_i = \frac{r_i}{\sqrt{\hat{\sigma}^2 (1 - H_{ii})}}$, $E[e_i] = 0$, $Var[e_i] = 1$

- $E(e_i) = 0, \text{Var}(e_i) = 1.$

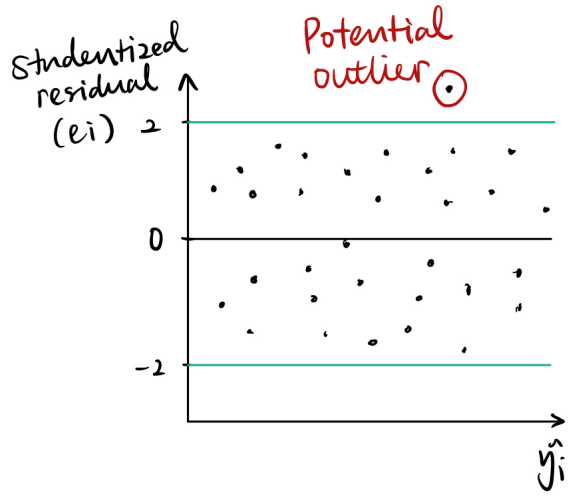
Residual Plots

- Gaphical display of residuals, an effective way to detect departures of model assumptions.
- Various types of plots for different assumptions
- Typically studentized residuals are ploted.

Plot of Residuals vs. Fitted Values

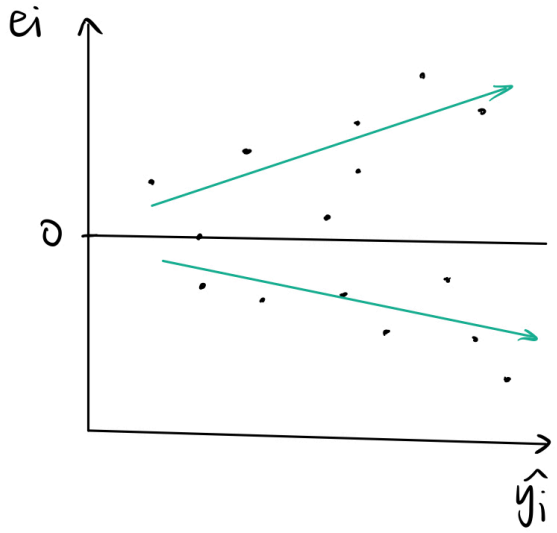
A plot of residual r_i or any of scaled residuals (d_i or e_i) against corresponding fitted value ($\hat{y}_i$)

1. If residuals fluctuates randomly around 0 inside a horizontal band, then **no visible defects**.



Random scatter, majority within ±2 bands (covers approx. 95% points)

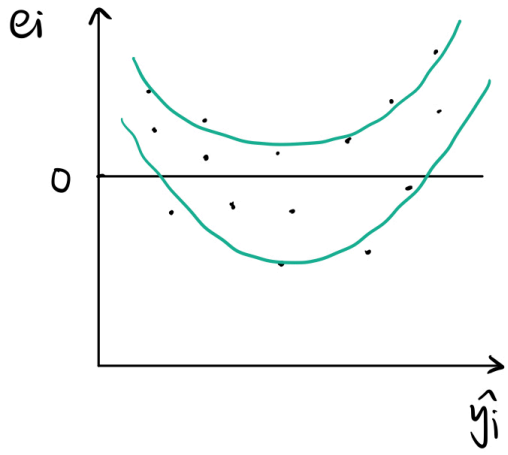
2. If the residuals can be contained in an opening funnel ("fan" shape), it indicates that $\text{Var}(\epsilon_i)$ **is not constant.**



Could use box-cox power transformation.

Variance increases as y increases (Pattern can be flipped)

3. If the residuals are contained inside a curve plot, then it indicates **nonlinearity** (the relationship between y and some x variables is not linear, or some other explanatory variables are needed)

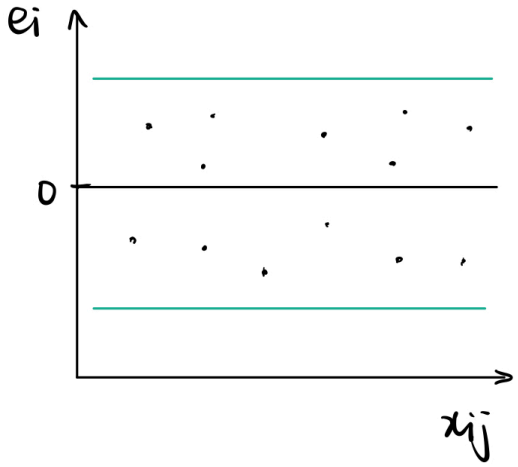


Plots of Residuals vs. Explanatory Variable

A plot of residuals against the values of j-th explanatory variable, x_{ij}'s.

Some interpretation of the plot as the case of plot of residuals vs. $\hat{y}_i$'s.

1. Horizontal band $\rightarrow$ no visible defects



Horizontal bands, no visible defects.

2. Funnel/fan shape $\rightarrow$ Variance is nonconstant $\rightarrow$ could add/transform some explanatory variables.
3. Curvature $\rightarrow$ Nonlinearity (may suggest need x_j^2 in the model)

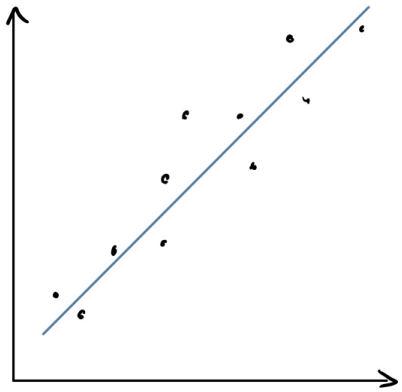
Partial Residual Plots

Most useful in investigating the relationshiop between response y and an explanatory variable x_j.

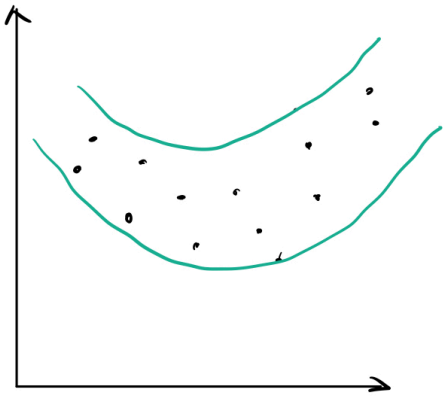
Partial Residual for $x_j, j = 1, \ldots, p$, $r_i^{(j)} = r_i + \hat{\beta}_j x_{ij}$, where r_i is the residual based on all p explanatory variables, adding effect of x_j variable back into the residual.

Plot Partial Residuals $r_i^{(j)}$ *vs.* x_{ij}

1. Linear trend $\rightarrow$ x_j enters model linearly



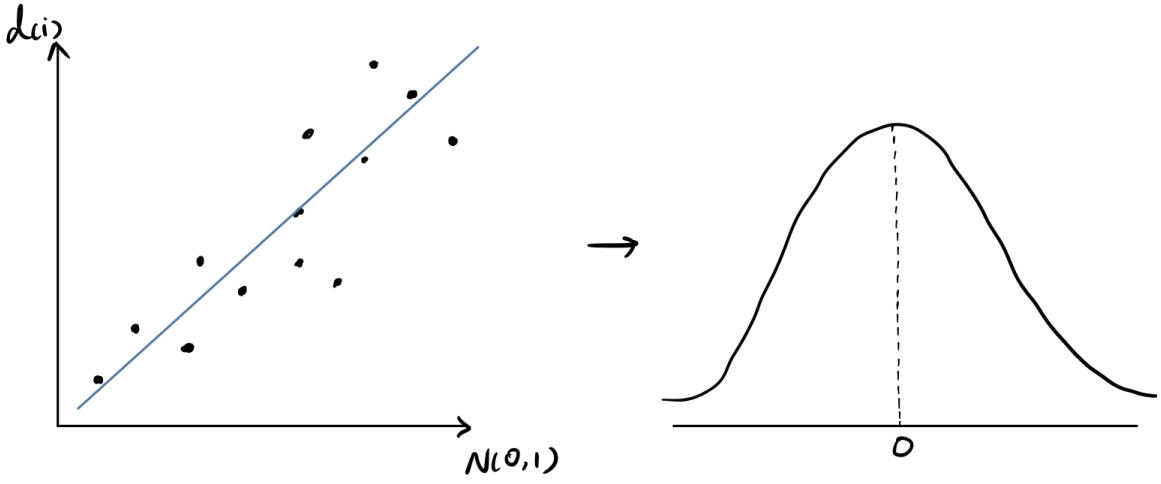
2. Curvature $\rightarrow$ higher order terms in x_j may be helpful



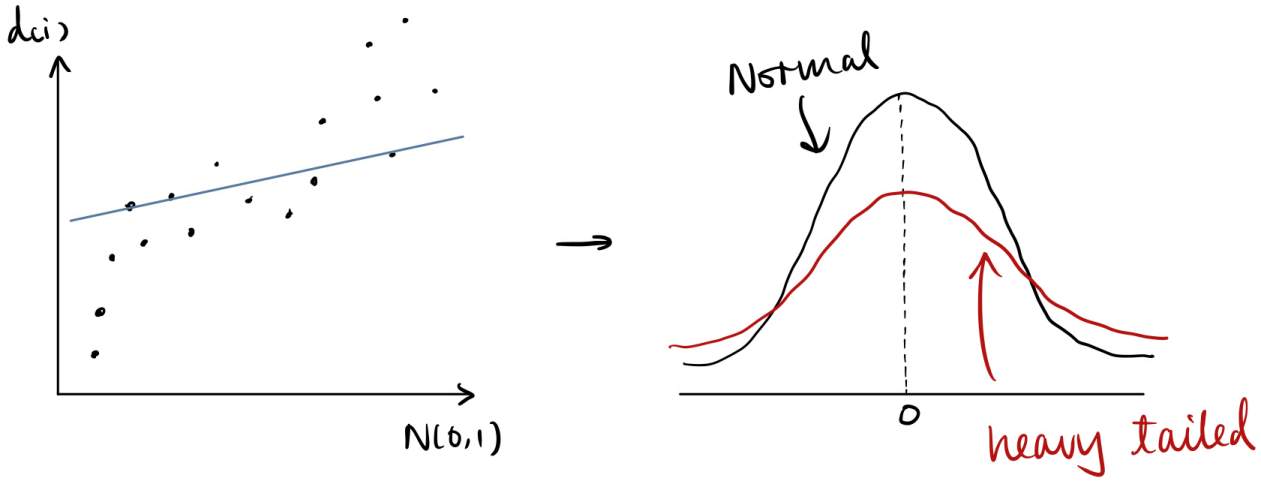
Q-Q Plot for Normal Distribution

This is a graphical technique for detecting substantive **departure from normality**. Plot ordered standardized residuals $d_{(1)} < d_{(2)} < \cdots < d_{(n)}$ against the theoretical quantiles of $N(0,1)$. If normality holds, then the ordered residuals should be align with the quantiles of normal distribution.

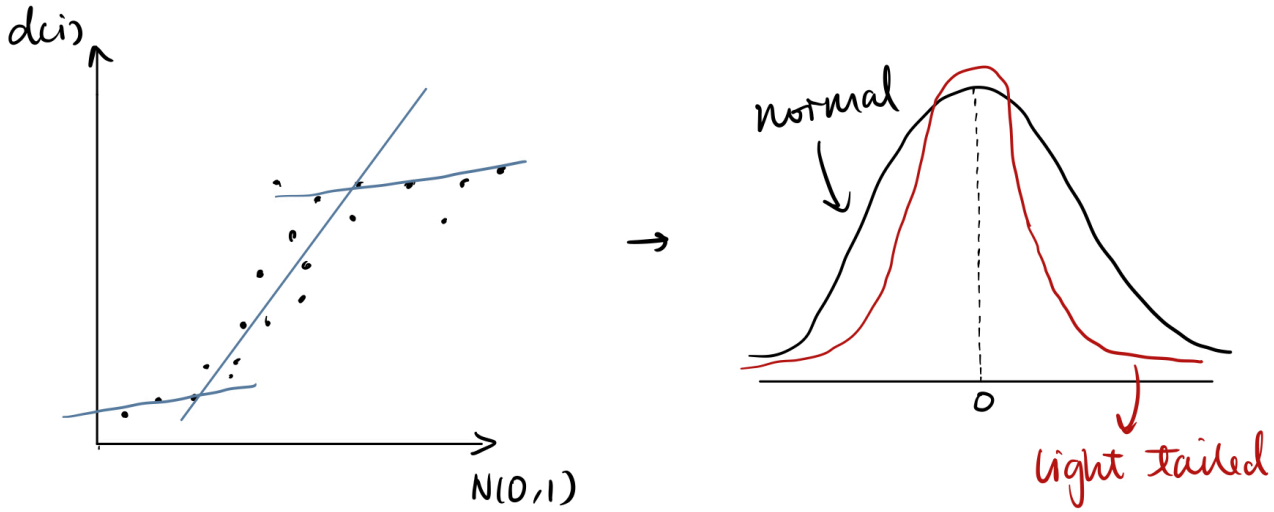
1. Points lies approximately on the straight line $\rightarrow$ underlying distribution is normal.



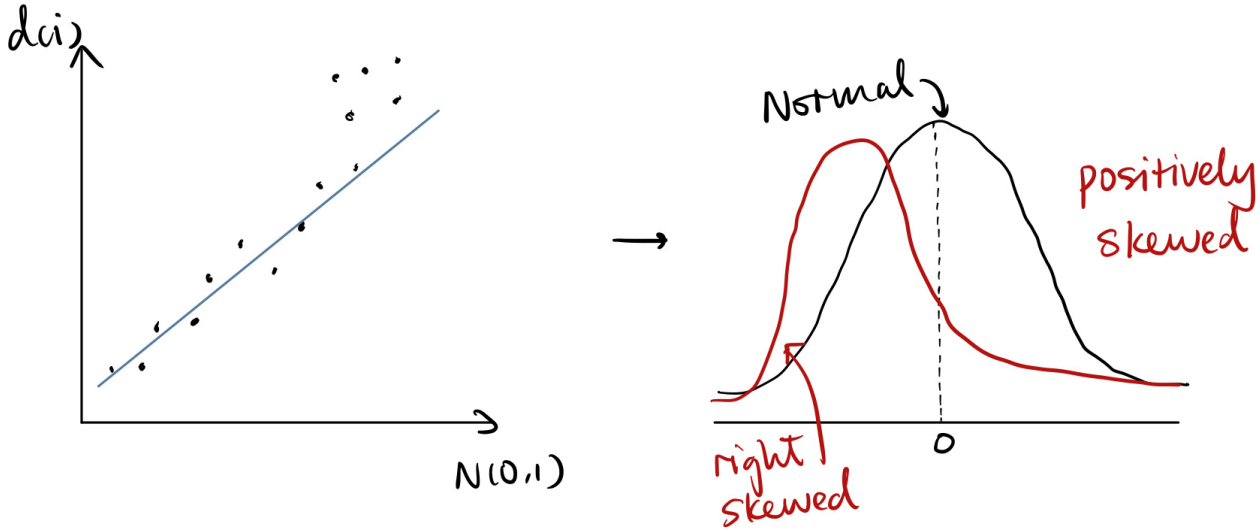
2. Sharp upward and downward curves at both extremes $\rightarrow$ heavy-tailed



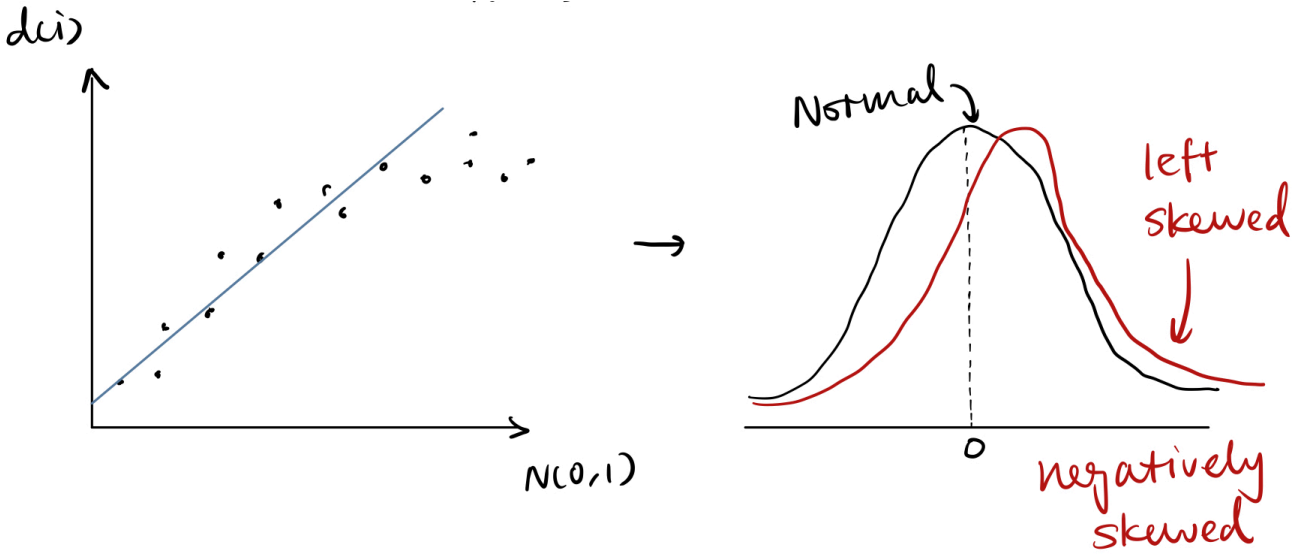
3. Flattening at extremes → light-tailed



4. Sharp change in the trend in an upward direction from the mid → positively skewed (right skewed)



5. Sharp change in the trend in a downward direction from the mid → negatively skewed (left skewed)



Remarks:

- Do not wnat to check normality untill other assumptions been checked and fixed.
- For skewed, transformation of y may help.
- Light tailed, can ignore
- Heavy tailed, problematic.

5.3 Addressing Model Assumption Problems

If residual plots reveal problems with assumptions, we might be able to address via data transformation.

1. Transformation of Response to Stablize Variance

   This can help address non-constant variance identified by the plot r vs. $\hat{y}$ (or r vs. x_j)

   Idea: apply function g and fit regression model on transformed $g(y_i)$

   $$g(y_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i$$

   Rationals: the variance fo response might be a function of mean $\mu_i = E(y_i)$, ie.

   $$\text{Var}(y_i) = \text{Var}(\epsilon_i) = h(\mu_i)\sigma^2 \quad (\text{for some } h(\,) > 0)$$

   in which case we want $\text{Var}(g(y_i)) \approx \sigma^2$

   By first order Taylor Expansion, we have

   $$g(y_i) \approx g(\mu_i) + (y_i - \mu_i)g'(\mu_i)$$

   Then,

   $$\text{Var}(g(y_i)) \approx [g'(\mu_i)]^2 \text{Var}(y_i) \approx [g'(\mu_i)]^2 h(\mu_i)\sigma^2$$

   Thus, we need $[g'(\mu_i)]^2 \propto \frac{1}{h(\mu_i)}$.

   Examples:

   1. $h(\mu_i) = \mu_i \Rightarrow \text{Var}(y_i) = \mu_i\sigma^2$.

      We need $g'(\mu_i) \propto \frac{1}{\sqrt{\mu_i}} = \frac{1}{\sqrt{\mu_i}} \Rightarrow g(\mu_i) = \sqrt{\mu_i}$ will work

      Thus, we can apply $g(y_i) = \sqrt{y_i}$ to abtain approximate constant variance.

   2. $h(\mu_i) = \mu_i^2 \Rightarrow \text{Var}(y_i) = \mu_i^2\sigma^2$.

      We need $g'(\mu_i) \propto \frac{1}{\mu_i} \Rightarrow g(\mu_i) = \ln(\mu_i)$ may work

      Thus, we can apply $g(y_i) = \ln(y_i)$ to abtain approximate constant variance.

   3. Box-cox Power Transformation

      $$g(y_i) = \begin{cases} \frac{{y_i}^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \log(y_i) & \lambda = 0 \end{cases}$$

Data Transformation

Response: address non-constant variance.
let $Var(y_i) = h(\mu_i)\,\sigma^2$, we want $[g'(\mu_i)]^2 \propto \frac{1}{h(\mu_i)}$

Box-Cox Power Transformation

$$g(y_i) = \begin{cases} \frac{y_i^{\lambda}-1}{\lambda} & \lambda \neq 0 \\ \log(y_i) & \lambda = 0 \end{cases}$$

It helps transform non-constant variance of the form
$$Var(y_i) = \mu_i^c\,\sigma^2$$
for some constant c

Special cases: $\lambda = \frac{1}{2}$ → square-root transformation
$\lambda = 0$ → log-transformation
$\lambda = 1$ → identity transformation
$\lambda = -1$ → reciprocal transformation

$g(y_i) = \log(y_i) = y_i = \exp\{\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i\}$
every 1 unit increase in x_j causes $100(e^{\beta_j}-1)$ % change in mean response.
95% CI of mean response at $(1, x_1, \cdots, x_p)^T$ =
$$\exp\{x^T\hat{\beta} \pm t_{\alpha/2,\,n-p-1}\,SE(x^T\hat{\beta})\}$$

Detect Outlier:

Studentized Residual: $e_i = \frac{r_i}{\sqrt{\hat{\sigma}^2(1-h_{ii})}}$

If $|e_i| > 3$, then observation i is an outlier in response

Note that $0 \le h_{ii} \le 1$, if $h_{ii} > 2\bar{h}$, then observation i is an outlier in explanatory variables.
→ If x_i is far from $\bar{x}$, then h_{ii} is large.

Influence of observation i to $\hat{\beta} = \hat{\beta}^{(i)}$.

Cook's Distance: $D_i = \dfrac{(\hat{\beta}-\hat{\beta}^{(i)})^T X^T X (\hat{\beta}-\hat{\beta}^{(i)})}{\hat{\sigma}^2 (p+1)} = e_i^2 \dfrac{h_{ii}}{1-h_{ii}} \dfrac{1}{p+1}$

If $D_i > 0.5$, may be influential.
$D_i > 1$, very likely to be influential.

Model Selection:

Selection Criterion: $R^2_{adj} = 1 - \dfrac{SSE/(n-k-1)}{SST/(n-1)} = 1 - \left(\dfrac{n-1}{n-k-1}\right)(1-R^2)$

(R^2_{adj}: higher the better.)

($1-R^2 = \frac{SSR}{SST}$)

$AIC = 2q - 2\ln(L(\hat{\theta}))$

(q: # of parameters; $L(\hat{\theta})$: likelihood evaluated at $\hat{\theta}$.)

lower the better.

$BIC = q\ln(n) - 2\ln(L(\hat{\theta}))$

lower the better

For models with the same k predictors, R^2_{adj}, AIC, BIC pick the same best model w/ lowest SSE.

Search Strategy:

Forward Search: Start w/ 0 predictors, adding one at a time and pick the best one improving the model.

Backward Elimination: Start w/ p predictors, removing one at a time and pick the best one improving the model.

Forward-backward Stepwise: Start w/ forward selection, then at each step, try forward & backward

Build Predictive Model:

$$MSPE = \frac{1}{v}\sum_{i=1}^{v}(y_i - \hat{y}_i)^2$$

v = # of validation set

y_i: label, $\hat{y}_i$: prediction.

$$RMSE = \sqrt{MSPE}$$

$$\text{Mean Absolute Error} = \frac{1}{v}\sum_{i=1}^{v}|y_i - \hat{y}_i|$$

Split Data into training & validation set:

80/20 split: $\{1, \cdots, n\} \cup \{n+1 \cdots n+v\}$

Cross-Validation: $D = D_1 \cup D_2 \cup \cdots \cup D_K$.

For $k = 1 \cdots K$, D_k is validation set.

Average $MSPE = \frac{1}{K}\sum MSPE_i$