1. **Basic assumption in one variable linear regression model**

   (1) $\epsilon_i \sim N(0, \sigma^2)$ and each ϵ_i is i.i.d for all y_i
   (2) Independent observation
   (3) Equal variace (homosceadasticity)

   所谓线性模型，指的是对于参数而言是线性的。即使样本本身特性是非线性的，如二次方、立方甚至指数，模型仍然是线性模型。

2. **Least Square Regression (LSR)**

   To find a straight line go through the points that can minimize the *Sum of Square Error (SSE)*. The expression of the fitted line is simply

   $$y = \beta_0 + \beta_1 \cdot x$$

   Let the points be (x_i, y_i). The the SSE becomes

   $$S(\beta_1, \beta_0) = \sum_{i=1}^{n} (y_i - y)^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 \cdot x_i)$$

   To minimize SSE, just take the partial derivative w.r.t β_1 and β_0 and set both to be zero

   $$\frac{\partial S}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 \cdot x_i) = 0$$

   $$\frac{\partial S}{\partial \beta_1} = -2 \sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 \cdot x_i) = 0$$

   the solution is therefore (hat means it is a estimate, not an estimator for which we usually use $\sim$).

   $$\hat{\beta_0} = \bar{y} - \hat{\beta_1} \bar{x} \quad \& \quad \hat{\beta_1} = \frac{r_{xy} \cdot s_y}{s_x} \ \ or \ \ = \frac{s_{xy}}{s_x^2}$$

   where $r_{xy} = \frac{s_{xy}}{s_x s_y}$. Then the regression line become

   $$y = \beta_0 + \beta_1 x = \bar{y} + s_y r_{xy} \cdot \frac{x - \bar{x}}{s_x}$$

   or a way easier to memorize

   $$\frac{y - \bar{y}}{s_y} = r_{xy} \frac{x - \bar{x}}{s_x}$$

   注：通过以上导数等于零的式子我们可以得到的结论有

   $$\sum_{i=1}^{n} e_i = 0 \ \ \& \ \ \frac{1}{n} \sum_{i=1}^{n} x_i e_i = 0$$

   即 **residual** 之和等于零 and **residual** 以 x 为权重的加权平均值等于零.

3. **Residual Plot**

   Residual refers to the difference between actual value and its prediction which is this case is the point on the regression line with corresponding x value. It is used to check if the linear model is an adequate approximation. It can be expressed as

   $$e_i = y_i - y = y_i - \hat{\beta_0} - \hat{\beta_1} x_i$$

Residuals plot includes (1)Plot of residuals v.s dependent variable (2) Plots of residuals vs. explanatory variable. These are to check the heteroscedasticity and homoscedasticity (Equal variance) (同方差性 & 异方差性). (3)Check normal quantile plot of residuals.

The analysis is based on several assumptions:

(1)Observations are independent and identically distributied
(2) Homoscedasticity of residuals (equal variance)
(3) The response variable is normally distributed
(4) Relationship between response variable and predictor variable(independent variable) is linear

所谓的' 同方差性' 与' 异方差性' 是指——残差项是否与 predictor variable(independent variable) 有关, 即 residual 的方差是否随 x 的变化而有变化. 在 residuals(y) v.s independent variable(x) 图中可以看出

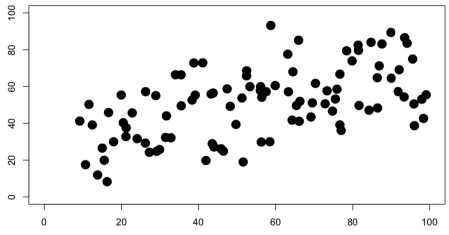


Figure 1: Homo.

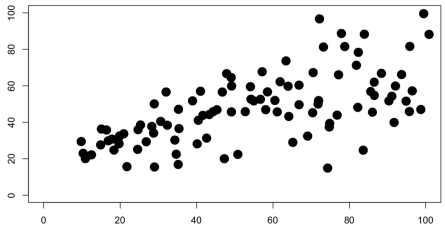


Figure 2: Heter.

Another way to understand is that given different x, the variance of all y values are equal. 同方差意为在给定 x 下的条件方差都是相同的 (做垂线计算线上 residual 的方差都应该相同).

Don't forget to draw the plot v.s independent. Aim is to check the so called homoscedasticity, which is whether there is a pattern of residuals as x changes or not.

What's more, the sum of residuals should equal to, or at least be very close to zero to make a valid regression since the line go through the data clouds, some points are above and some are below the line, which is

$$\sum_{i=1}^{n} e_i \ = 0$$

where e_i are the residuals. The standard deviation of residual is

$$\hat{\sigma} \ = \ \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-2}}$$

4. **Residuals as Random Variable**

For linear regression, we consider the independent variable as non-random and the predicted one as random variable. Without given dataset, now think of residual as a r.v which is the sum of unmeasured effects.

Now assume there is a liner relationship between X and Y, and the model as

$$Y = \beta_0 + \beta_1 X + \epsilon \ \ \& \ \ \epsilon \sim N(0, \sigma^2)$$

where X and ϵ are r.v. Thus we can get

$$Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$$

the regression line is the collection of 'exact value' of given x_i and the deviation from the line is generated by ϵ. It can be shown in the graph as

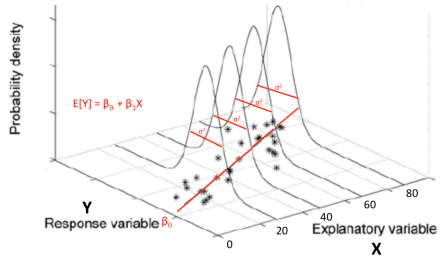


Figure 3: Distribution of residuals

5. **Construction of C.I for estimate of parameter** $\beta_1 \& \beta_0$

To construct a C.I, take the estimate of parameter as a r.v. which is B_1. By some proof we find

$$B_1 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})}{(n-1)s_x^2} Y_i$$

where only Y_i is a r.v.. Then we need to find the SE to build the C.I. The way is to find Var first and then take the square root. Recall that Y_i is a r.v. follows a certain type of normal distribution which is $Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$. So by some algebra we get

$$Var(B_1) = \frac{\sigma^2}{(n-1)s_x^2}$$

where we use sample residuals SE to approximate parameter σ, so we get

$$SE(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{n-1} \cdot s_x^2}$$

then use $t - distribution$ to construct the C.I which is

$$[\beta_1 - t_{n-2,0.975} \frac{\hat{\sigma}}{\sqrt{n-1} \cdot s_x}, \beta_1 + t_{n-2,0.975} \frac{\hat{\sigma}}{\sqrt{n-1} \cdot s_x}]$$

where the 0.975 can be changed for need. Similarly, we can get for B_0 and summarize them together we have 原因是可以证明 $\hat{\beta}_1$ 是 normal distribution 的线性组合。

$$\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{(n-1)s_x^2}) \quad \& \quad \hat{\beta}_0 \sim N(\beta_0, (\frac{1}{n} + \frac{\bar{x}^2}{\sqrt{n-1} \cdot s_x^2}) \cdot \sigma^2)$$

6. **Construction of C.I for estimate of subpopulation mean**

The subpopulation is the set of Y_is under a given value of x. The expression is $\mu_Y(x)$. Now consider it as a r.v

$$\mu_Y(x) = \beta_0 + \beta_1 x$$

then take it as a random variable we have

$$\mu_Y(x) = \hat{\beta}_0 + \hat{\beta}_1 x = \sum_{i=1}^{n} c\frac{1}{n} + \frac{(x-\bar{x})(x_i-\bar{x})}{(n-1)s_x^2}) \cdot Y_i$$

Then for the variance of μ we have

$$Var(\hat{\mu}_Y(x)) = \sum[\frac{1}{n} + \frac{(x-\bar{x})(x_i-\bar{x})}{(n-1)s_x^2})]^2 Var(Y_i) = \sigma^2 \cdot \sum[\frac{1}{n} + \frac{(x-\bar{x})(x_i-\bar{x})}{(n-1)s_x^2}]^2$$

$$= \sigma^2\{\frac{1}{n} + \frac{(x-\bar{x})^2}{[(n-1)s_x^2]^2}\}$$

so finally we have

$$\mu_y(x) \sim N(\beta_0 + \beta_1 x, \sigma^2\{\frac{1}{n} + \frac{(x-\bar{x})^2}{(n-1)s_x^2}\})$$

so the 95% C.I is constructed same as previous.

7. Prediction and Prediction Interval

Suppose now we have already generated form the sample that $Y(x) = \beta_0 + \beta_1 x$ and set up the model. Now we want to predict the Y^* value with given x^*. If in the sample we do not have the Y value at x^*, then this prediction is called the ***out of sample prediction*** and the corresponding Y^* predicted, which is our best estimation, is called the ***point estimation***.

No matter how perfect the model is, the prediction ($\hat{Y} = \hat{B}_0 + \hat{B}_1 x$) always have difference from the real value ($Y(x) = \beta_0 + \beta_1 x + \epsilon$) . The error between prediction and true value is

$$\hat{Y}(x^*) - Y(x^*) = (\hat{B}_0 - \beta_0) + (\hat{B}_1 - \beta_1)x^* - \epsilon(x^*)$$

and the variance of this error

$$Var(\Delta) = Var[(\hat{B}_0 - \beta_0) + (\hat{B}_1 - \beta_1)x^*] + Var(\epsilon)$$

the cov. term is 0 because of independence thus

$$= \sigma^2 \cdot [\frac{1}{n} + \frac{(x^*-\bar{x})^2}{(n-1)s_x^2}] + \sigma^2$$

Notice the term $Var[(\hat{B}_0 - \beta_0) + (\hat{B}_1 - \beta_1)x^*] = Var(\hat{\mu}_Y(x^*))$ the var. of subpopulation mean. Thus consequently the estimated SE of the prediction error is

$$SE = \hat{\sigma} \times \sqrt{1 + \frac{1}{n} + \frac{(x^*-\bar{x})^2}{(n-1)s_x^2}}$$

note this does not decrease to 0 as n approaching ∞. So the prediction interval for a $Y(x^*)$ with a x^* out of samples is

$$\hat{Y}(x^*) \pm t_{n-2,0.975} \times SE$$

注 *1:* 置信区间 *v.s* 预测区间

(1) 置信区间是对某 parameter 的估计而预测区间是对给定 x 的对应单个随机变量值 y 的估计。

(2) 预测区间往往比置信区间要宽——置信区间估计的 parameter 往往是客观存在的常数而预测区间估计的却是随机变量的值，因此除了估计方法本身的误差外，预测区间还需要加上其想要预测的随机变量的标准差。

注 *2:* 统计推断途径

$$\theta \longrightarrow \hat{\theta} \longrightarrow \hat{\Theta} \longrightarrow \begin{cases} \mathbb{E}(\hat{\Theta}) \\ Var(\hat{\Theta}) \end{cases} \longrightarrow SE(\hat{\theta}) \ \ appx\ .\ Var(\hat{\Theta}) \longrightarrow C.I.$$

Target para.→ 一个样本值 → 将样本值理解为 r.v→ 求期望判定无偏估计 & 计算样本方差以近似 r.v 总体方差 → 置信区间

注 *3:* 通过 *OLS* 得到的线性回归方程需要进行检验看各项系数是否显著。我们可以通过 t 检验、F 检验和相关系数 r 的显著性检验判断且三者在一元线性回归中完全等价且结论一致。*306* 只介绍 t 检验*make up other two method after lesson*

8. **The t distribution in interval estimation**

For all those t test above, we have the null hypothesis

$$H_0 \ : \ \beta_1 = 0$$

it can be proved that if the population variance σ is unknown, then the substitution by the sample variance s will make the standardized β_1 follow t_{n-2}. (Proof can be found here https://www.zhihu.com/question/34670804.) The null hypothesis indicates that with the mean at 0, we have

$$\frac{B_1}{SE(B_1)} \sim t_{n-2}$$

and then we can do C.I estimation by t critical value.

注: 95% 置信区间指的是有 95% 个区间包含真正的总体参数 β_1 而非 $\hat{\beta}_1$.

9. **Multivariable least square**

Multivariable linear regression is the extension of single variable regression. We use matrix to represent the relevant informations because of the dramatically increasing data. The fitted plane of regression is

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where all terms in bold face are vectors. The expression of this is

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \dots \\ \dots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \dots & \dots & \dots & & \dots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \dots \\ \dots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \epsilon_0 \\ \epsilon_1 \\ \epsilon_2 \\ \dots \\ \dots \\ \epsilon_n \end{pmatrix}$$

for all the subscript e.g x_{ij}, i means the ith observed dataset and j means the jth value.

Similar to the single regression assumptions, we still presume that $\epsilon_i \sim N(0, \sigma^2)$ for all i. 值得注意的是，在一元线性回归中 ϵ_i 的意思是给定 x_i 的 y 值作为随机变量的分散程度，而多元线性回归中则是给定 x_1, x_2 的 y 值在经过该点的离散程度。Since now we have vector, we can write

$$\boldsymbol{\epsilon} \sim N(0, \sigma^2 \boldsymbol{I}_n)$$

then the r.v y have

$$\boldsymbol{y} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n)$$

in a matrix form it is

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \dots \\ \dots \\ y_n \end{pmatrix} \sim \begin{pmatrix} \mu_1 & \sigma^2 & 0 & 0 \\ \mu_2 & 0 & \sigma^2 & 0 \\ \dots & .. & .. & .. \\ \mu_n & 0 & 0 & \sigma^2 \end{pmatrix}$$

where the right part is the covariance matrix of vector $\boldsymbol{y}$.

(a) **Multilinear OLS**

Exactly the same as single regression, repeal to the differentiation method, we want to minimize the squared error by taking the derivative and set to zero. Finally we should get

$$\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}^T\boldsymbol{y} \quad \equiv \quad \hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

where $\boldsymbol{X}$ and $\boldsymbol{y}$ are all known. 这里的 $\boldsymbol{X}$ 被称为资料矩阵，在取样过程中可以被调控以服务某特定目的。另外，回归系数 β_i 也被成为偏回归系数 (partial regression coefficient) 因为其特指在其他的自变量不变的情况下每增加以单位的 x_i 而使得 y 增加的数量。得到估计的 β，则

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + ... + \hat{\beta}_p x_p$$

被称为经验回归方程。将其以矩阵形式表示为

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$$

介于我们通过 OLS 得到了估计的回归参数值，也可以写成

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

注意 $\boldsymbol{y}$ 是观测到的样本值，因此 $\hat{\boldsymbol{y}}$ 是 $\boldsymbol{y}$ 的回归值。从代数角度看，可以看做 y 通过矩阵的投影。The hat matrix maps the observed value of values of $\boldsymbol{y}$ onto the vector of fitted values $\hat{\boldsymbol{y}}$. 还有些关于线代的内容复习后补充 p62

(b) **Hat matrix**

The above matrix can be considered as a projection matrix on observed $\boldsymbol{y}$ which is

$$H = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

it is also called the *Hat matrix*. 投影后的面、结构就是我们拟合出的模型。The entry H_{ii}, the diagonal is a good measure of how much influence the ith observation has on the fitted model. So the estimate of $\boldsymbol{\epsilon}$ is

$$\hat{\boldsymbol{\epsilon}} \sim Nor(0, \sigma^2 diag[\boldsymbol{I}_n - H])$$

proof needed in screenshot folder The sum of diagonal elements(trace) equals the num of column of infor. matrix. The diagonal entries are called the *leverage*. 名称其实很形象，某本书上讲 'A point has high leverage if omitting it causes a big change in fit'，即该 observation 对拟合的模型影响很大。课件上讲 $\hat{y}_i = p_{ii} + \sum_{j \neq i} p_{ij} y_j$，因此 p_{ii} 可以看做观测值 y_i 对 fitted model 的影响，更详细内容参见文件夹下 ExplainHatMatrix。

注：Hat matrix 有三个性质:

- *Influence* It is easy to check that
$$\frac{\partial \hat{\boldsymbol{Y}}_i}{\partial Y_j} = H_{ij}$$
其意为"H_{ij} is the rate at which the ith fitted value changes as we vary the j^{th} observation", 这也是我们将其理解为 influence 的原因。
- *Symmetry* This is $\boldsymbol{H}^T = H$
- *Idempotency* This is $\boldsymbol{H}^2 = \boldsymbol{H}$.

(c) **Cook's distance**

(d) **Multiple correlation coefficient** 多元决定系数

与在一元回归中相同，可以定义决定系数,
$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$
称 $R = \sqrt{\frac{SSR}{SST}}$ 为样本复相关系数。Also, $R \in [0,1]$. The closer R^2 to 1, the better the regression, the closer x to 0, the worse the regression (see the appendix about the correlation coefficient). 注意, R^2 可以清楚的反应回归拟合的结果，但不能作为严格的显著性检验。Another version is the adjusted corr. coeff. which is
$$adjR^2 = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)} = 1 - \frac{\hat{\sigma}^2}{s_y^2}$$
调整目的是不让其随着自变量的增多而变大，从而可以比较在不同多的 variable 情况下哪一种拟合地更好。注意 $R^2 \geq 0$ 但是 $adjR^2$ 在回归不理想时可以负数。appendix make up

(e) **Interval estimate for parameters $\boldsymbol{\beta}$**
proof or sth will be made up 5later. Let $\hat{\boldsymbol{B}} = \boldsymbol{\beta}$ be an random vector w.r.t $\boldsymbol{Y}$. The variance of the estimator is
$$Var(\boldsymbol{B}) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1} \equiv \sum_{\hat{\boldsymbol{\beta}}}$$
which is the covariance matrix of vector beta. It can be seen that the $Var(\beta_i)$ the entry at position ii in the matrix. So we can find the SE of the β by
$$se(\hat{\beta}_i) = \hat{\sigma}\sqrt{[(\boldsymbol{X}^T\boldsymbol{X})^{-1}]_{ii}}$$
and the appropriate SE is multiplied by $t_{n-k,0.975}$ to get the margin of error to add/subtract form the point estimate. The diagonal of the matrix consisting of the variance of each entry of $\boldsymbol{\beta}$.

(f) **Interval estimation for subpopulation mean**

Consider the subpopulation mean $\mu_Y(x^*)$ with given $\boldsymbol{x}^* = (1, x_1^*, x_2^*, ..., x_p^*)^T$, as a random variable, the variance is
$$Var[\mu_Y(x^*)] = Var[\boldsymbol{x}^{*T}\hat{\boldsymbol{B}}] = \boldsymbol{x}^{*T}\sum_{\hat{\beta}}\boldsymbol{x}^* = \sigma^2\boldsymbol{x}^{*T}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}^*$$
then the SE is
$$se[\mu_Y(x^*)] = \hat{\sigma}\sqrt{\boldsymbol{x}^{*T}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}^*}$$
finally the 95% C.I becomes
$$\hat{\mu}_Y(x^*) \pm t_{n-k,0.975} \cdot se[\mu_Y(x^*)]$$

(g) **Prediction interval for out-of-sample point estimation**

Same as the single variable regression, we are estimating the value of a random variable. Then the prediction error is

$$\begin{aligned} E(error) &= Predicted - True = \hat{Y}(\boldsymbol{x}^*) - Y(\boldsymbol{x}^*) \\ &= \hat{Y}(\boldsymbol{x}^*) - [\beta_0 + \beta_1 x_1^* + ... + \beta_p x_p^* + \epsilon(\boldsymbol{x}^*)] \end{aligned}$$

so the var of error is

$$\begin{aligned} Var(E) &= Var[\boldsymbol{x}^{*T}\hat{\boldsymbol{B}}] + Var[\boldsymbol{\epsilon}(\boldsymbol{x}^*)] \\ &= \sigma^2 \boldsymbol{x}^{*T}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}^* + \sigma^2 \end{aligned}$$

consequently the SE becomes

$$se(E) = \hat{\sigma}\sqrt{1 + \boldsymbol{x}^{*T}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}^*}$$

then the 95% prediction interval becomes

$$\hat{Y}(\boldsymbol{x}^*) \pm t_{n-k,0.975} \cdot se[E]$$

理解 SE 中的 1 来自何处，其为被预测的 r.v 自身的 variance.

(h) **The linear combination of Y**

Let $\boldsymbol{Y} \sim N_n(\boldsymbol{\mu}, \sum)$. If we have a linear combination of $\boldsymbol{Y}$, say $\boldsymbol{CY}$, where $\boldsymbol{c}$ is a $(q \times n)$ matrix, then

$$\boldsymbol{CY} \sim N_q(\boldsymbol{C\mu}, \boldsymbol{C}\sum\boldsymbol{C'})$$

10. **Categorical variable**

Taking into consideration of categorical variable which is introduced into the model by using of dummy variable. For simplicity, consider the model with one continuous variable and one categorical variable. The from of linear regression with categorical variable is

$$\begin{aligned} \mu_Y(x_i, z_i, x_i z_i) &= \beta_0 + \beta_1 x_i + \beta_2 z_i + \beta_3 x_i z_i \\ &= \begin{cases} \beta_0 + \beta_1 x_i & category\,1(baseline) \\ \beta_0 + \beta_1 x_1 + \beta_2 + \beta_3 x_i & category\,2 \end{cases} \end{aligned}$$

其中 z_i 通过 dummy 值 1 或 0 来选择 category。cross term 的意思是，当选择完分类型变量后，与 baseline 相比两直线斜率不同，因此通过添加该项来拟合直线。此处用 μ 而不是 Y_i 是为了体现出拟合的直线且可以省略掉 $random\,term$。未完待续

11. **Multicolinearity** 多重共线性

In regular multilinear regression, we assumed that the infor. matrix $\boldsymbol{X}$ must be of $p+1$ dimension/full rank (p is # of x), or equivalently, using the formal definition in linear algebra, if there exists at least one c_i not equals 0, s.t

$$c_0 + c_1 x_{i1} + c_2 x_{i2} + ... + c_n x_{in} = / \approx 0$$

then the predictor variables are said to be multi-colinearity, 也称复共线性。以上情况也等价于资料矩阵存在线性相关的 x。往往完全线性无关情况很少，因此如果近似为 0. 即 $\approx$ 我们也认为是线性相关，且这并非是一种违背线性回归模型基本假设的情况，除非是完全共线性。该现象可以用语言表述为

完全共线性带来的影响是，对于参数 $\boldsymbol{\beta}$ 的估计值中的 $\boldsymbol{X}^T\boldsymbol{X}$ 不可逆，也因此该估计不合理，即' 在完全共线性下的参数估计量不存在'（上述等于零的情况）。多重共线性产生的原因往往是 *poorly designed study* 或是 *Similar problem to having no control group*。

如果是近似多重共线性，OLS 估计的 $\boldsymbol{\beta}$ 仍然存在，但是由于 $\boldsymbol{X}^T\boldsymbol{X}$ 的 determinate ≈ 0, 所以使得 $\hat{\boldsymbol{\beta}}$ 的对角线元素（即各个 β_i）很大，因此估计精度很低。

(a) 多重共线性的诊断

应用回归分析课本中共有三种方法，306 只介绍了一种 VIF 判断法。VIF 原理是，将其中的一个解释变量用剩余的解释变量解释（回归）并衡量其被解释的程度。VIF 全称是 *Varience Inflation Factor* 方差扩大因子，定义为

$$VIF_j = \frac{1}{1 - R^2_{x_j \cdot \boldsymbol{X}_{-j}}}$$

如果 $VIF_j >> 1$，那么说明对于解释变量 x_j 存在其他的解释变量与其有多重共线性。可以解释 $SE(\hat{\beta}_j)$ 为什么非常大。应用回归分析课本上有更精确的定义

12. **Variable selection Algorithms**

306 只介绍了一种变量的选择准则 C_p 统计量。其定义为

$$C_p = \frac{SSE_p}{\hat{\sigma}^2} - n + 2p$$

其提出的原理是：即使全模型正确，但选模型仍然有可能有较小的预测误差。公式中的 $\hat{\sigma}^2$ 是全模型 的 σ^2 的无偏估计。即

$$MS(Res : x_1, x_2, ..., x_m) = \hat{\sigma}^2(x_1, ..., x_m) = \frac{SSE_m}{n - m - 1}$$

因此，能使得该统计量最小的自变量子集对应的回归方程就是最优回归方程。注意分母上的减一是因为有 β_o 的原因，即此处我们用 m 表示 x 的个数所以未包含 β_0。

另外，可以通过计算不同变量组的 SSE 结合 $NULL$ model 的 SSE 来计算决定系数 R^2 即，

$$R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = 1 - \frac{SSE}{\boldsymbol{y}^T\boldsymbol{y} - n\bar{y}^2} = 1 - \frac{SSE}{SSE(NULL)}$$

NULL model 的 SSE 即最原本的 '方差'，没有包含任何的变量，是常数。计算时先算出，一劳永逸。注意，有时候 C_p statistics 的结果与 R^2 不一定一致，因此需要结合做决策。C_p 统计量实际上有些过时，目前常用的统计量是赤池统计量，考完试补充

(a) **Forward selection**

Start with one variable, add one variable at a time.

(b) **Backward Elimination**

Start with the full model/all potential variable, remove one variable at a time.

13. **Cross-Validation** 交叉检验

交叉检验本质是将收集到的数据集 (dataset, say size is n) 分成训练集 (training set, say size is n_1) 和测试集 (testing set, say size is n_2)。前者用于拟合模型，后者用于检验模型是否合适。定义 out-of-sample root mean square error 为

$$RMSE_{holdout}(x_1, ..., x_m) = \sqrt{\frac{\sum_{i \in holdout}(y_i - \hat{y}_{i|train})^2}{n_2}}$$

通常情况下训练集要大于测试集。进一步我们还可以将整个 dataset 分成等 size 的子集。每一次以其中一个作为 testing set，其他的作为 training set。Let the dataset be divided into G equal size subset/fold, then the G-fold CVRMSE is

$$CVRMSE_{Gfold} = G^{-1} \cdot \sum_{i=1}^{G} RMSE_{holdout_i}(x_1, ..., x_p)$$

下面介绍两种交叉检验方法。

(a) **Leave-one-out**

The notation of OLS for parameter is $\hat{\beta}_{-i}$, which means delete the ith observation and using the remaining to fit the model. The CVRMSE is

$$CVRMSE_{leaveoneout}(x_1,...,x_p) = \sqrt{n^{-1}\sum_{i=1}^{n}(y_i - \hat{y}_{i|-i})^2}$$

即要计算 n 个 regression，拟合 n 个 model，让每一个 observation 都被 leave out 一次。一个简化版的公式是

$$y_i - \hat{y}_{i|-i} = \frac{y_i - \hat{y}_i}{1 - P_{ii}}, \quad P_{ii} = \pm x_i^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i$$

where $x_i^T = (1, x_{i1}, ..., x_{ip})$. 貌似式子里的 $\hat{y}_i$ 就是包含了所有变量的模型得到的估计值 check

14. **Transform and Nonlinearity**

有时候我们可以通过某种变换使得模型具备同方差性，例如取 log 或是开方。表示为

$$Y_i = g(x_{i1}, ..., x_{ip}) + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2(x_i))$$

where x_i is the vector of explanatory variable，即方差是关于变量的函数。常用的变换是：如果 Y 对于 x 呈现 $1/x$ 的形状，则将 x 取 log；如果发现 Y 的值非常分散，可以考虑将 Y 值开根号

15. **Logistic regression** 逻辑回归

用于处理 binary 的回归。首先介绍 signo 函数，定义为

$$g(z) = \frac{1}{1+e^{-z}}, \quad g : \mathbb{R} \to (0,1)$$

其特点为将 R 上的值映射到 0 和 1 之间，可以在此处作为概率函数。定义 $Y_i \sim Bernoli(\pi_i)$，其概率分布可以描述为

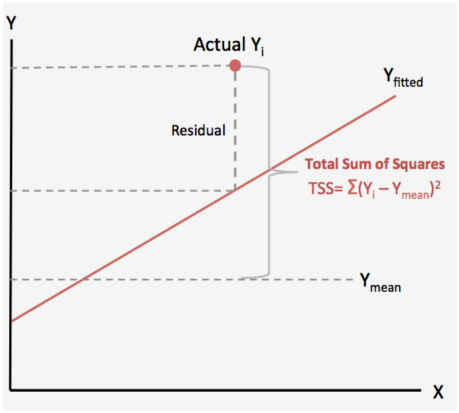
$$P(Y_i = y_i) = \pi_i^{y_i}(1-\pi_i)^{1-\pi_i}$$

由于 Y_i 只能取 1 或 0，所以上式合理地描述了两种情况。由于是 Bernoli 分布，其方差为 $\pi_i(1-\pi)$，且应注意随之而来的二项分布。令 $z = \boldsymbol{X}^T\boldsymbol{\beta}$，beta 是线性模型中的全部参数。然后可以通过二项分布连乘得到似然函数并求得 beta 的 MLE。由于是 bineray，所以无法应用 OLS 进行估计 还有很多需要搞

在逻辑回归中常用的统计量是 AIC 赤池统计量，越小越好。

Appendixes

1. **The decomposition of SST:**



The SS Total is simply the variation of the set of data, so it is

$$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2$$

$$= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + 2\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

then for the middle cross term we have

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = \sum_{i=1}^{n} e_i(\hat{\beta}_0 + \hat{\beta}_1 x_i - \bar{y})$$

$$= (\hat{\beta}_0 - \bar{y})\sum_{i=1}^{n} e_i + \hat{\beta}_1 \sum_{i=1}^{n} e_i x_i = 0$$

this result is from the derivative of SSE.

2. **Basic Formulas**

$$r_{xy} = \frac{\sum z_x z_y}{n-1} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y} = \frac{s_{sy}}{s_x s_y}$$

$$s_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

The $R^2 = r^2$, they are the same thing. 相关系数是剔除量纲影响后的标准化协方差。当我们说协方差时，分清说的是样本协方差还是总体协方差。计算总体协方差时，我们将 X，Y 看做是随机变量并且用概率分布的角度去理解；样本协方差可以看做是对总体协方差的估计：通常除以 (n-1) 是无偏估计。

3. **Specify between error and residual**

Error is difference between true value and observed value; residual is difference between observed and fitted value (What we observed can not always be true).