Catagorical Data: Nominal, Ordinal
Numerical Data: Discrete, Continuous
Other: Survival, Longitudinal, Time Series

样本偏度 Sample Skewness: $n\sum_{i=1}^{n}(X_i-X^-)^3/(S^3(n-1)(n-2))$
总体偏度 Population Skewness: $E(X-\mu)^3/\sigma^3$
Sample Excess Kurtosis:
$$\frac{n(n+1)}{(n-1)(n-2)(n-3)}\sum_{i=1}^{n}\frac{(X_i-\bar{X})^4}{S^4}-\frac{3(n-1)^2}{(n-2)(n-3)}$$
Population Excess Kurtosis: $E(X-\mu)^4/\sigma^4 - 3$
Excess Kurtosis > 0 → 此分布较正态分布更容易产生离群值

无序类别变量的独立性检验
Pearson's χ^2： $\chi^2=\sum_{i=1}^{r}\sum_{j=1}^{c}(O_{ij}-E_{ij})^2/E_{ij}$.
Cramer's V: $V=\surd(\chi^2/(n*\min(r-1,c-1)))$

有序类别变量的独立性检验
Kendall's τ=(#Concordant Pairs-#Discordant Pairs)/#Total Pairs
Concordant: $i<j, X_i<X_j \leftrightarrow Y_i<Y_j; X_i>X_j \leftrightarrow Y_i>Y_j$
Spearman's ρ： R=Rank
$r_S=\sum_{i=1}^{n}(R(X_i)-R(X)^-)(R(Y_i)-R(Y)^-)/\surd\sum(R(X_i)-R(X)^-)^2\sum(R(Y_i)-R(Y)^-)^2$

样本相关系数：
$$r=\frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2\sum_{i=1}^{n}(Y_i-\bar{Y})^2}}$$
总体相关系数： $\rho=Cov(X, Y) / \sigma_X\sigma_Y$.

缺失值：
MAR(系统性): $Pr(M=1|X, Y)=Pr(M=1,X)$
MCAR(完全随机): $Pr(M=1|X, Y)=Pr(M=1)$
MNAR(值相关，不可忽视)
MCAR 处理: Listwise Deletion, Pairwise Deletion
其他方式：单值插补，依分布插补(Multiple Imputation)

I 型错误：弃真；II 型错误：取伪；固定样本时，无法同时降低

	H_0 is True	H_1 is True
Accept H_0	Right	Type II Error
Reject H_0	Type I Error	Right

p-value = Pr(H_0 为真时观测到观测数据)
power = 1 - β

III 假设检验

III 假设检验：当 p 值<显著水平或检验统计量落入拒绝域时，可以宣称零假设 H_0 在显著水平α下拒绝，反之则无法拒绝 H_0
Z-test: H_0: $p = p_0$ v.s. H_1: $p</\neq/>p_0$; $p^{hat}=\sum X_i/n$
$Z = (p^{hat} - p_0)/\surd(p_0(1 - p_0)/n)\sim_{H0}N(0, 1)$(渐近)
p-value = $Pr(Z<z_{obs}|H_0)$ or $Pr(|Z|>|z_{obs}||H_0)$ or $Pr(Z>z_{obs}|H_0)$
e.g. $Pr(Z>z_{obs})=1-\Phi(z_{obs})$. Φ(.)是 N(0,1)的累积分布函数
$z_{0.05}=1.645, z_{0.025}=1.96, z_{0.01} = 2.33, z_{0.005} = 2.58$
CI: $p^{hat}\pm z_{\alpha/2}\surd(p^{hat}(1-p^{hat})/n)$; 令 Effect Size $p_d=p_1-p_0\neq_{H1}0$.
则 $Z=p_d$^$/\surd(p_0(1-p_0)/n)\sim_a N(p_d/\surd(p_0(1-p_0)/n), p_1(1-p_1)/(p_0(1-p_0)))$
$\beta=1-\Phi((|p_d|/\surd(p_0(1-p_0)/n)-z_{\alpha/2})/\surd(p_1(1-p_1)/(p_0(1-p_0)))$.
样本大小 $n\approx(z_{\alpha/2}\surd(p_0(1-p_0))+z_\beta\surd(p_1(1-p_1)))^2/(p_d)^2$.
使用另一个 Z 时，$Z= (p^{hat} - p_0)/\surd(p$^$(1 - p$^$)/n)\sim_{H0}N(0,1)$(渐近)
则 $Z\sim_{H1}N(p_d/\surd(p_1(1-p_1)/n), 1)$
$\beta=1-\Phi(|p_d|/\surd(p_1(1-p_1)/n)-z_{\alpha/2})$. $n\approx(z_{\alpha/2}+z_\beta)^2p_1(1-p_1)/p_d^2$.

2-sample Z-test: H_0: $p_1 - p_2 = 0$ v.s. H_1: $p_1-p_2\neq0$
$SE=\surd(p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2)$ | 假设 $p_1=p_2=p$
$Z=(p_1^{hat}-p_2^{hat})/\surd(p(1-p)(1/n_1+1/n_2))\sim_{H0}N(0,1)$(渐近)
p_1 和 p_2 很小时用 Relative Risk p_1/p_2 或 Odds Ratio $p_1(1-p_2)/(p_2(1-p_1))$
$p^{hat}=(n_1p_1^{hat}+n_2p_2^{hat})/(n_1+n_2)$, pooled proportion
$Z=(p_1^{hat}-p_2^{hat})/\surd(p^{hat}(1-p^{hat})/(1/n_1+1/n_2))$
CI: $p_1^{hat}-p_2^{hat}\pm z_{\alpha/2}\surd(p_1^{hat}(1-p_1^{hat})/n_1+p_2^{hat}(1-p_2^{hat})/n_2)$ Wald Interval

CI: 检查 $n_1p_1^{hat}, n_1(1-p_1^{hat}), n_2p_2^{hat}, n_2(1-p_2^{hat})\geq10$
假设检验: 检查 $n_1p^{hat}, n_1(1-p^{hat}), n_2p^{hat}, n_2(1-p^{hat})\geq10$

拟合优度检验 Goodness of fit
检验 H_0: $p_1=p_{01}, ..., p_m=p_{0m}$. ($\sum p_{0i}=1$, p_{0i} 是设置好的)
$\chi^2=\sum_{i=1}^{m}(O_i-E_i)^2/E_i\sim\rightarrow\chi^2(m-1)$, O_i:第 i 类的观测频数, $E_i=np_{0i}$
2 类特例：$\chi^2=n(X^--p_{01})^2/(p_{01}(1-p_{01}))=Z^2$.

独立性、同质性检验(2 维类别型变量，r 组×c 类)
H_0: $p_{1j}=p_{2j}$ j = 1, 2, ..., c
前提：样本独立采样；样本容量足够大（各类期望均≥10）
$\chi^2=\sum_{i=1}^{r}\sum_{j=1}^{c}(O_{ij}-E_{ij})^2/E_{ij}$. $E_{ij}=np_{i.}p_{.j}\sim\chi^2_{(r-1)(c-1)}$|Assume Independence
χ^2 检验的备择假设总是双边的。

McNemar's Test for Paired Samples
Paired: 观测集中的每一样本均与另一观测集中对应唯一样本有联系
SE^$(p_1$^$-p_2$^$)=(\surd(b+c-(b-c)^2/n))/n=_{H0}(\surd(b+c))/n$
$\chi^2=(b-c)^2/(b+c)\sim_{H0}\chi^2(1)$. (Asymptotically, b 或 c 小时，渐近不好)
CI: $(b-c)/n\pm z_{\alpha/2}(\surd(b+c-(b-c)^2/n))/n$
p-value=$2\sum_{k=b}^{b+c}C_{b+c}^{k}p^k(1-p)^{m-k}$ (b~Bin(b+c, p), p=0.5)

连续变量检验

单样本 t 检验
H_0: $\mu=\mu_0$ vs. H_1: $\mu\neq\mu_0$ or $\mu>\mu_0$ or $\mu<\mu_0$.
$S=\surd(\sum_{i=1}^{n}(X_i-X^-)^2/(n-1))$, $T=(X^--\mu_0)/\surd(S^2/n)$
CI: $X^-\pm t(\alpha/2, n-1)S/\surd n$

双样本 t 均值检验
H_0: $\mu_1=\mu_2$ vs. H_1: $\mu_1\neq\mu_2$ or $\mu_1>\mu_2$ or $\mu_1<\mu_2$.
Pooled $S_p=\surd(((n_1-1)S_1^2+(n_2-1)S_2^2)/(n_1+n_2-2))$
$T=(X_1^--X_2^-)/\surd(S_p^2(1/n_1+1/n_2))$
CI for $\mu_1-\mu_2$: $X_1^--X_2^-\pm t(\alpha/2, n_1+n_2-2)S_p\surd(1/n_1+1/n_2)$.

F 等方差检验
H_0: $\sigma_1^2=\sigma_2^2$ H_1: $\sigma_1^2\neq\sigma_2^2$. | $(n_1-1)S_1^2/\sigma_1^2\sim\chi^2(n_1-1)$.
$F=S_1^2\sigma_2^2/S_2^2\sigma_1^2=_{H0}S_1^2/S_2^2\sim F(n_1-1, n_2-1)$.

F 检验被拒绝时，t 检验改为 Satterthwaite/Welch's t-test
自由度$\nu=(S_1^2/n_1^2+S_2^2/n_2^2)^2/((S_1^2/n_1)^2/(n_1-1)+(S_2^2/n_2)^2/(n_2-1))$
$T_S=(X_1^--X_2^-)/\surd(S_1^2/n_1+S_2^2/n_2)\sim|_{H0}t(\nu)$ (自由度取整数部分)

均值/中位数的非参数检验
Sign test: H_0: $m=m_0$ vs. H_1: $m\neq/</>m_1$. 可转化为 z-test.
即等价于检验 X_i-m_0 的符号是否服从 p=0.5 的二项分布。
Wilcoxon Signed Rank Test:
①为 X_i-m_0 的绝对值从 1 到 n 编号，差为 0 则丢弃
②对于绝对值相同的，取排位的平均
③定义 S^+为正差值排位的和，S^-为负差值排位的和
④定义 $W=S^+$或 S^+-S^-或 $\min(S^+, S^-)$.
H_1: $m\neq m_0$: 当 $w\leq d$ 或 $w\geq n(n+1)/2 - d$ 时拒绝 H_0.(查表得 d)
H_1: $m>m_0$: 当 $w\geq n(n+1)/2 - d$ 时拒绝 H_0.
H_1: $m<m_0$: 当 $w\leq d$ 时拒绝 H_0.

Family-wise Error Rate FWER (Type I Error/α Inflation)
Pr(At Least One Significant Result)=1-Pr(No Significant Results).
Bonferroni 调整: $p_i^{\sim}=\min(mp_i, 1), i=1, ..., m$ 或$\alpha^{\sim}=\alpha/m$。
Holm/Step-down 调整 :
$p_{(1)}<\alpha/m$? 拒绝 $H_{(10)}$ 并继续 : 停止. / $p_{(1)}^{\sim}=\min(1, mp_{(1)})$
....$p_{(i)}<\alpha/(m-i+1)$? 拒绝 $H_{(i0)}$ 并继续 : 停止.
/ $p_{(i)}^{\sim}=\min(1, \max((m-i+1)p_{(i)}, p_{(i-1)}^{\sim}))$

IV 多元线性回归 Multiple Linear Regression

LSE: $\mathbf{y}$^$=X\boldsymbol{\beta}$^$=X(X'X)^{-1}X'\mathbf{y}=H\mathbf{y}$. ε_i^$=y_i-y_i$^.
min RSS=$\sum_{i=1}^{n}\varepsilon_i^{2\wedge}$=SSE($\boldsymbol{\beta}$^)=$||\mathbf{y}-X\boldsymbol{\beta}||^2$. $E(\boldsymbol{\beta}$^$)=\boldsymbol{\beta}$, $Var(\boldsymbol{\beta})=\sigma^2(X'X)^{-1}$.
$\sigma^{2\wedge}=\mathbf{y}'(I-H)\mathbf{y}/(n-p-1)=RSS/(n-p-1)$, $E(\sigma^{2\wedge})=\sigma^2$.
MLE
$\sigma^{2\wedge}=RSS/n$, $\boldsymbol{\beta}$^$\sim N(\boldsymbol{\beta}, \sigma^2(X'X)^{-1})$. $RSS/\sigma^2\sim\chi^2(n-p-1)$.

Theorem 4.1 $A\mathbf{y}\sim N(A\boldsymbol{\mu}, A\Sigma A')$, $A\mathbf{y}\perp B\mathbf{y}\leftrightarrow A\Sigma B'=0$
$SSE(\boldsymbol{\beta})/\sigma^2=\sum_{i=1}^{n}\varepsilon_i^2/\sigma^2\sim\chi^2(n)$.
$SSE(\boldsymbol{\beta})=RSS+(\boldsymbol{\beta}^{\wedge}-\boldsymbol{\beta})'X'X(\boldsymbol{\beta}^{\wedge}-\boldsymbol{\beta})$
$(X'X)^{1/2}(\boldsymbol{\beta}^{\wedge}-\boldsymbol{\beta})/\sigma\sim N(0, I_{p+1})\rightarrow(\boldsymbol{\beta}^{\wedge}-\boldsymbol{\beta})'X'X(\boldsymbol{\beta}^{\wedge}-\boldsymbol{\beta})/\sigma^2\sim\chi^2(p+1)$.

回归系数检验

单回归系数检验：H_0: $\beta_i=0$ vs. H_1: $\beta_i\neq0$. 记$(X'X)^{-1}$的第 i 对角元为 c_{ii}
$\beta^{\wedge}_i\sim N(\beta_i, c_{ii}\sigma^2)$. $T=(\beta_i^{\wedge}-0)/\surd(c_{ii}\sigma^2)\sim_{H0} t(n-p-1)$
多回归系数检验：H_0: $\beta_{k+1}=\beta_{k+2}=...=\beta_p=0$, H_1: 存在$\beta_i\neq0, k+1\leq i\leq p$
Full: $Y=\beta_0+\sum_{i=1}^{p}\beta_iX_i+\varepsilon$，Reduced: $Y=\beta_0+\sum_{i=1}^{k}\beta_iX_i+\varepsilon$
$RSS_R\geq RSS_F$. $F=(RSS_R-RSS_F)(n-p-1)/(RSS_F(p-k))\sim F(p-k,n-p-1)$.

有效性检验 Reduced: $Y=\beta_0+\varepsilon$

Src	DF	SS	MS	F	Pr>F
Model	p	SSM	SSM/p	MSM/MSE	
Error	n-p-1	SSE	SSE/(n-p-1)		
Total	n-1	SST			

R^2=SSM/SST=1-SSE/SST.

Gauss-Markov 定理
若 $E(\varepsilon_i)=0, Var(\varepsilon_i)=\sigma^2<\infty, \forall i, Cov(\varepsilon_i, \varepsilon_j)=0, \forall i\neq j$.
则 LSE 估计的$\boldsymbol{\beta}$是线性无偏估计量中方差最小的，称为 BLUE.

V 模型选择与诊断 Model Selection and Diagnosis

R^2 在添加自变量至模型中时不减
$R_{adj}^2=1-(1-R^2)(n-1)/(n-k-1)$. 越大越好
Mallow's C_p: $C_p=SSE_k/MSE_p-(n-2k-2)$. $E(C_p)=k+1$.选择最接近期望的
AIC=$-2\ln(L$^$)+2(k+1)=_{线性模型} n\ln(SSE_k/n)+2(k+1)$ 越小越好
BIC=$-2\ln(L$^$)+(\ln n)(k + 1)$

前向选择 Forward Selection
定义标准（F, R_{adj}^2, C_p, AIC, BIC）→从仅有截距的模型开始→依照标准选择 1 个变量加入模型，直到没有变量能够被加入

后向消除 Backward Elimination
从全量模型开始，依次消除一个变量直到没有变量可以被删去

逐步选择 Stepwise Selection
同时考虑待加入的变量和已加入的变量，加入最大贡献的待加入变量并删去最小贡献的已加入变量。

收缩方法 Shrinkage Method
岭回归：$SSE(\boldsymbol{\beta}; \lambda)=\sum_{i=1}^{n}(y_i-\beta_0-\sum_{k=1}^{p}\beta_kx_{ik})^2+\lambda\sum_{k=1}^{p}\beta_j^2$
$\boldsymbol{\beta}_{ridge}^{\wedge}=(X'X+\lambda I_p)^{-1}X'\mathbf{y}=\mathrm{argmin}_{\boldsymbol{\beta}}(\sum_{i=1}^{n}(y_i-\beta_0-\sum_{k=1}^{p}\beta_kx_{ik})^2, \sum_{k=1}^{p}\beta_k^2\leq t$.
LASSO: $\boldsymbol{\beta}_{LASSO}^{\wedge}=\mathrm{argmin}_{\boldsymbol{\beta}}(\sum_{i=1}^{n}(y_i-\beta_0-\sum_{k=1}^{p}\beta_kx_{ik})^2$, s.t. $\sum_{k=1}^{p}|\beta_k|\leq t$.

Least Angle Regression
$\boldsymbol{\beta}$设为 0....

交叉验证 Cross Validation
Leave-one-out CV:
K-fold CV:

模型诊断

线性回归的假设：线性、误差同方差、独立误差、正态误差
关于同方差性 Homoscedasticity
在拟合值-残差图中，观察误差均值是否近似为 0，且散布一致

Fitted vs. Residual Plot: 检验线性和同方差性
Normal Quantile-Quantile Plot：检验正态性
Histogram of Residuals: 检验正态性

Breusch-Pegan Test: 检验同方差性
对于残差的平方，拟合一个线性模型，检验所有系数是否等于 0
改进：$LM=nR^2$ 检验是否服从$\chi^2(k)$. (仅从 $X_1, ..., X_p$ 中选 k 个拟合)
White Test: 检验同方差性
相比于上一个检验，拟合的模型中添加了高阶项 $x_{ij}^{2\wedge}$和交叉项 x_{ij}^x_{ik}^。

Shapiro-Wilk Test: 检验正态性
$0\leq W=(\sum_{i=1}^{n}a_ir_{(i)})^2/\sum_{i=1}^{n}(r_i-r^-)^2\leq1$,过小拒绝正态性,一般 n≤2000.
Kolmogorov-Shirnov Test: 检验正态性
基于经验分布函数（EDF）$F_n(x)=(\sum_{i=1}^{n}I(r_i\leq x))/n$.
$D=\sup_x|F_n(x)-F(x)|$(F(x)为样本均值方差正态分布的累计分布函数)
n≥2000 时使用，满足正态性时 D 渐近趋向于 Kolmogorov 分布。

检验独立性：根据数据来源判断数据是否是独立采集的
Durbin-Watson Test: 检验数据是否有时序性(一阶自相关性)
$DW=\sum_{t=2}^{n}(\varepsilon_t^{\wedge}-\varepsilon_{t-1}^{\wedge})^2/\sum_{t=1}^{n}\varepsilon_t^{\wedge2}$. (模型：$\varepsilon_t=\rho\varepsilon_{t-1}+u_t$, H_0: ρ=0)
DW=2→无自相关; 0≤DW<1→强正自相关; 3<DW≤4→强负自相关

冲突恢复
线性：做变换、加入高阶项
同方差性：
①***稳定方差变换*** Variance Stablizing Transformation
假设 $Var(Y)=h(E(Y))=h(\mu)$，变换 $g(Y)\propto\int dY/\surd h(Y)$
Poisson: $Var(Y)=\mu\rightarrow g(Y)=\surd Y$
Exponential: $Var(Y)=\mu^2\rightarrow g(Y)=\ln Y$.
②***带权最小二乘回归*** Weighted LS Estimation
$\boldsymbol{\beta}^{\wedge WLS}=(X'WX)^{-1}X'W\mathbf{y}$. $W=diag(1/\sigma_i^2)$.
σ_i^2 未知时，先进行常规拟合得到拟合值 y_i^残差 r_i 然后对 $y_i^和|r_i|$进行拟合得到第 2 个模型，第二个模型的拟合值记为 v_i^. 令 $w_i=1/v_i^{\wedge2}$
正态性：
Box-Cox 变换：$g_\lambda(y) = \lambda ? (y^\lambda - 1) / \lambda : \ln y$. λ由最大似然决定。
不寻常样本的识别：
Leverage 杠杆点：在解释变量上取到极端值的样本 $h_{ii}>2(p+1)/n$
Hat/Projection Matrix: $H = X(X'X)^{-1}X'$.
$Z=[[\mathbf{x_1}'-\mathbf{x}^-']... [\mathbf{x_n}'-\mathbf{x}^-']$, $h_{ii}=1/n + (\mathbf{x}_i-\mathbf{x}^-)'(Z'Z)^{-1}(\mathbf{x}_i-\mathbf{x}^-)$
$\sum h_{ii}=tr(H)=p+1$

Outlier 离群点：在拟合的模型上得到巨大残差的样本 $|r_i^{stu}| > 3$
Residual: $\mathbf{r}=\boldsymbol{\varepsilon}^{\wedge}=\mathbf{y}-\mathbf{y}^{\wedge}=(I-H)\mathbf{y}$
Studentized Residual: $r_i^{stu}=(y_i-y_i^{\wedge})/(\sigma^{\wedge}\surd(1-h_{ii}))$
$\sigma^{\wedge2}=RSS/(n-p-1)$; $\sigma_{(i)}^{\wedge2}=RSS_{(i)}/(n-p-2)$

Influence 影响点：删去后会显著改变拟合系数的样本 $D_i>F(0.5, p+1, n-p-1)$
Cook's Distance: $D_i=(r_i^{stu})^2h_{ii}/((1-h_{ii})(p+1))$
（或 $D_i>4/n$）

Multicollinearity Detection 共线性检测 VIF > 10
共线性：存在不全为 0 的系数 c_i，使 $c_0+c_1X_1+...+c_pX_p=0$.
共线性存在时，X 可能不满秩；X'X 近乎奇异；$\boldsymbol{\beta}$^不稳定
Tolerance: $TOL_j=1-R_j^2$. R_j^2 是删去第 j 个样本后拟合模型的 R^2.
Variance Inflation Factor: $VIF_j=1/(TOL_j)=1/(1-R_j^2)$
可以考虑删去异常样本；或使用岭回归、LASSO 回归等

VI 方差分析 Analysis of Variance

One-way ANOVA: 对有 1 个类别自变量的回归分析
H_0: $\mu_1=...=\mu_k$. vs. H_1: μ_i 不全相等.
模型：$Y_{ij}=\mu+\alpha_i+\varepsilon_{ij}\sim N(\mu_i, \sigma^2)$. $\alpha_i=\mu_i-\mu$衡量第 i 组的主要影响（Main Effect）
$\mu=\sum_{i=1}^{k}n_i\mu_i/n$ 为全局均值
于是假设变为 H_0: $\alpha_i=0, \forall i$, vs. H_1: $\exists i, \alpha_i\neq0$
组 i 的样本均值和样本标准差：$Y_i^-=\sum_{j=1}^{ni}Y_{ij}/n_i$, $S_i=\surd(\sum_{j=1}^{ni}(Y_{ij}-Y_i^-)^2/(n-1))$
组间方差：$SSB=\sum_{i=1}^{k}\sum_{j=1}^{ni}(Y_i^--Y^-)^2=\sum_{i=1}^{k}n_i(Y_i^--Y^-)^2\sim|_{H0}\chi^2(k-1)$
组内方差：$SSW=\sum_{i=1}^{k}\sum_{j=1}^{ni}(Y_{ij}-Y_i^-)^2=\sum_{i=1}^{k}(n_i-1)S_i^2\sim|_{H0}\chi^2(n-k)$
总方差：$SST=\sum_{i=1}^{k}\sum_{j=1}^{ni}(Y_{ij}-Y^-)^2\sim|_{H0}\chi^2(n-1)$
$E(SSB)=(k-1)\sigma^2+\sum_{i=1}^{k}n_i(\mu_i-\mu)^2=|_{H0}(k-1)\sigma^2$; $E(SSW)=(n-k)\sigma^2$.
$F=SSB(n-k)/(SSW(k-1))\sim|_{H0}F(k-1, n-k)$.

Src	DF	SS	MS	F	Pr>F
Between	k-1	SSB	SSB/(k-1)	MSB/MSW	
Within	n-k	SSW	SSW/(n-k)		
Total	n-1	SST			

Reject H_0 if F is large.

同方差性检验：H_0: $\sigma_1^2=...=\sigma_k^2=\sigma^2$. vs. H_1: σ_i^2 不全相等.
Bartlett 检验：对正态性的偏离很敏感
Pooled Sample Variance: $S^2=SSW/(n-k)=MSW=\sum_{i=1}^{k}(n_i-1)S_i^2/(n-k)$
$B=((n-k)\ln S^2-\sum_{i=1}^{k}(n_i-1)\ln(S_i^2))/(1+(-1/(n-k)+\sum_{i=1}^{k}1/(n_i-1))/(3k-3))$
$B\sim|_{H0}\chi^2(k-1)$, reject H_0 if $B>\chi^2(\alpha, k-1)$.
Levene 检验、Brown-Forsythe 检验：
定义基于原始因变量 Y_{ij} 的散布（Dispersion）变量 Z_{ij}.
Levene: $Z_{ij}=(Y_{ij}-Y_i^-)^2$ 或$|Y_{ij}-Y_i^-|$
Brown-Forsythe: $Z_{ij}=|Y_{ij}-m_i|$, m_i 是第 i 组的样本中位数
对 Z_{ij} 进行方差分析，计算 SSB、SSW、F 统计量和 p 值.
方差分析的结果即是齐方差检验的结果。
正态性检验：
Kruskal-Wallis 检验：检验各组中位数是否相等
对所有 Y_{ij} 从 1 至 n 进行排序，相同值的排名改为其排名的均值
记 R_{ij} 为 Y_{ij} 的排名，$R_i^-=\sum_{j=1}^{ni}R_{ij}/n_i$, $R^-=\sum_{i=1}^{k}\sum_{j=1}^{ni}R_{ij}/n=(n+1)/2$
$KW=(n-1)\sum_{i=1}^{k}n_i(R_i^--R_i)^2/(\sum_{i=1}^{k}\sum_{j=1}^{ni}(R_{ij}-R^-)^2)\sim|_{H0}\chi^2(k-1)$.
$KW>\chi^2(k-1)$时拒绝相等假设。

Multiple Comparisons
...

Two-way ANOVA: 对有 2 个类别自变量的回归分析
类别变量 A 有 a 个水平，B 有 b 个水平，定义了 ab 个类别
交互模型 Interaction Model：$Y_{ijk}=\mu_{ij}+\varepsilon_{ijk}=\mu+\alpha_i+\beta_j+\gamma_{ij}+\varepsilon_{ijk}$. (总是从此开始)
加和模型 Additive Model: $Y_{ijk}=\mu+\alpha_i+\beta_j+\varepsilon_{ijk}$.
B 模型 Factor B Only Model: $Y_{ijk}=\mu+\beta_j+\varepsilon_{ijk}$.
A 模型 Factor A Only Model: $Y_{ijk}=\mu+\alpha_i+\varepsilon_{ijk}$.
零模型 Null Model: $Y_{ijk}=\mu+\varepsilon_{ijk}$.
满足：$\sum_{i=1}^{a}n_{i.}\alpha_i=\sum_{j=1}^{b}n_{.j}\beta_i=\sum_{i=1}^{a}n_{ij}\gamma_{ij}=\sum_{j=1}^{b}n_{ij}\gamma_{ij}=0$
$\mu=\sum_{i=1}^{a}\sum_{j=1}^{b}n_{ij}\mu_{ij}/n$, $\mu_{i.}=\sum_{j=1}^{b}n_{ij}\mu_{ij}/n_{i.}$, $\mu_{.j}=\sum_{i=1}^{a}n_{ij}\mu_{ij}/n_{.j}$.
$\alpha_i=\mu_{i.}-\mu$, $\beta_j=\mu_{.j}-\mu$, $\gamma_{ij}=\mu_{ij}-(\mu+\alpha_i+\beta_j)=\mu_{ij}-\mu_{i.}-\mu_{.j}+\mu$.
$SSE=\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{nij}(Y_{ijk}-\mu_{ij})^2=\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{nij}(Y_{ijk}-\mu-\alpha_i-\beta_j-\gamma_{ij})^2$.
①检验交叉项为 0：H_0^{AB}: $\gamma_{ij}=0, \forall i, j$, vs. H_1^{AB}: $\exists i, j$, s.t. $\gamma_{ij}\neq0$.
②H_0^{AB} 接受时，检验：H_0^A: $\alpha_i=0, \forall i$, vs. H_1^A: $\exists i$, s.t. $\alpha_i\neq0$.
③H_0^{AB} 接受时，检验：H_0^B: $\beta_j=0, \forall j$, vs. H_1^B: $\exists j$, s.t. $\beta_j\neq0$.
组间方差：$SSM=\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{nij}(Y_{ij}^--Y^-)^2=\sum_{i=1}^{a}\sum_{j=1}^{b}n_{ij}(Y_{ij}^--Y^-)^2\sim|_{H0}\chi^2(ab-1)$
组内方差：$SSW=\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{nij}(Y_{ijk}-Y_{ij}^-)^2\sim|_{H0}\chi^2(n-ab)$
总方差：$SST=\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{nij}(Y_{ijk}-Y^-)^2=SSM+SSE\sim|_{H0}\chi^2(n-1)$
$F=SSB(n-ab)/(SSW(ab-1))\sim|_{H0}F(k-1, n-k)$. 不好，仍需计算 A 和 B 的。
A 组间方差：$SSA=\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{nij}(Y_{i.}^--Y^-)^2=\sum_{i=1}^{a}n_{i.}(Y_{i.}^--Y^-)^2\sim|_{H0}\chi^2(a-1)$
B 组内方差：$SSB=\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{nij}(Y_{.j}^--Y^-)^2=\sum_{j=1}^{b}n_{.j}(Y_{.j}^--Y^-)^2\sim|_{H0}\chi^2(b-1)$
交互方差：$SSAB=\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{nij}(Y_{ij}^--Y_{i.}^--Y_{.j}+Y^-)^2=\sum_{i=1}^{a}\sum_{j=1}^{b}n_{ij}(Y_{ij}^--Y_{i.}^--Y_{.j}+Y^-)^2$
$E(SSA)=(a-1)\sigma^2+\sum_{i=1}^{a}n_{i.}\alpha_i^2$; $E(SSB)=(b-1)\sigma^2+\sum_{j=1}^{b}n_{j.}\beta_j^2$

$E(SSAB)=(a-1)(b-1)\sigma^2+\sum_{i=1}^{a}\sum_{j=1}^{b}n_{ij}\gamma_{ij}^2$.
$F^{AB}=SSAB(n-ab)/(SSE(a-1)(b-1))$;
$F^{A}=SSA(n-ab)/(SSE(a-1))$; $F^{B}=SSB(n-ab)/(SSE(b-1))$
$SST=SSA+SSB+SSAB+SSE$.

Type I & Type III SS in SAS 处理样本量不平衡的问题。

Type I: A, B, AB 按顺序加入模型；SSM=SSA+SSB+SSAB, 顺序相关
SSA=SSE(null)-SSE(A),
SSB=SSE(A)-SSE(A, B),
SSAB=SSE(A, B)-SSE(A, B, AB)
Type III: 考虑全模型和删减模型的差异 SSM≠SSA+SSB+SSAB
SSA=SSE(B, AB)-SSE(A, B, AB),
SSB=SSE(A, AB)-SSE(A, B, AB),
SSAB=SSE(A, B)-SSE(A, B, AB)
样本量相等时，TypeI SS 与 TypeIII SS 相等。

协方差分析 Analysis of Covariance ANCOVA

X 是连续变量

Model I: $Y_{ij}=\mu+\alpha_i+\beta X_{ij}+\beta_i X_{ij}+\varepsilon_{ij}$, 第 i 组第 j 个样本，共 k 组，每组 n_i 个样本

Model II: $Y_{ij}=\mu+\alpha_i+\beta X_{ij}+\varepsilon_{ij}$, Model III: $Y_{ij}=\mu+\alpha_i+\varepsilon_{ij}$, Model IV: $Y_{ij}=\mu+\beta X_{ij}+\varepsilon_{ij}$,

H_0^{AX}: $\beta_i=0, \forall i$, vs. H_1^{AX}: $\exists i, \beta_i\neq0$. H_0^{AX} 被拒绝时，选择模型 I

H_0^{A}: $\alpha_i=0, \forall i$, vs. H_1^{AX}: $\exists i, \alpha_i\neq0$; H_0^{X}: $\beta=0$, vs. H_1^{X}: $\beta\neq0$.

均被拒绝时，选模型 II; 仅 H_0^{A} 被拒绝，模型 III; 仅 H_0^{X} 被拒绝，模型 IV.

所有假设均被接受，选用零模型 $Y=\mu+\varepsilon_{ij}$.

VII 广义线性模型

指数分布族

$f(\mathbf{y};\boldsymbol{\theta})=h(\mathbf{y})\exp[(\boldsymbol{\eta}(\boldsymbol{\theta})\cdot\mathbf{T}(\mathbf{y})-A(\boldsymbol{\theta}))/\varphi]$，φ称为 Dispersion par.

自然形式：$\boldsymbol{\eta}(\boldsymbol{\theta})=\boldsymbol{\theta}$. $f(\mathbf{y};\boldsymbol{\theta})=h(\mathbf{y})\exp(\sum_{i=1}^{k}\theta_iT_i(\mathbf{y})-A(\boldsymbol{\theta}))$

总是能够重参数化将一个指数族化为自然形式

T(Y)是自然参数**θ**的充分统计量，**T(y)=y** 时称为自然指数族

二项分布、泊松分布、指数分布、正态分布均是指数族

Bin(n,p): $f(y;p)=C_n^y\exp(y\ln(p/(1-p))+n\ln(1-p))$
Poisson(λ): $f(y;\lambda)=1/y! * \exp(y\ln\lambda-\lambda)$
$N(\mu,\sigma^2)$: $f(y;\mu)=(2\pi\sigma^2)^{-1/2}\exp(-y^2/2\sigma^2+y\mu/\sigma^2-\mu^2/2\sigma^2)$
Multi(n, **θ**):

$M_{\mathbf{T}}(\mathbf{t})=E(e^{\mathbf{t}\cdot\mathbf{T}(\mathbf{Y})})=\exp(A(\mathbf{t}+\boldsymbol{\theta})-A(\boldsymbol{\theta}))$, $E(\mathbf{T}(\mathbf{Y}))=\partial A(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$.
$Cov(\mathbf{T}(\mathbf{Y}))=\varphi\cdot\partial^2A(\boldsymbol{\theta})/\partial\boldsymbol{\theta}^2$.

广义线性模型·组成部分：

①因变量 $Y_1, ..., Y_n$ 独立，且均服从指数分布族中的同一个分布
②有一系列自变量 $X_1, ..., X_p$ 和参数**β**.
③有一个单调的连接函数 g 使得$\mu_i=E(Y_i|\mathbf{X_i}=\mathbf{x_i})$.

$$g(E(Y_i|\mathbf{X}_i=\mathbf{x}_i))=g(\mu_i)=\beta_0+\sum_{k=1}^{p}\beta_kx_{ik}=\mathbf{x}_i'\boldsymbol{\beta}=\eta_i.$$

辨析：在线性模型的转换中，有 $E(g(Y_i)|\mathbf{X}_i=\mathbf{x}_i)=\mathbf{x}_i'\boldsymbol{\beta}$

考虑具有参数θ_i的自然指数族 $f(y_i;\theta_i)=h(y_i)\exp(\theta_iy_i-A(\theta_i))$
$E(T(Y_i))=E(Y_i)=A'(\theta_i)$, $g(E(Y_i))=g(A'(\theta_i))=\eta_i\to g(\cdot)=(A')^{-1}(\cdot)$.
g(·)被称为典型连接函数（Canonical Link Function）。
$\mu_i=A'(\theta_i)$, $\eta_i=g(\mu_i)$, $\eta_i=\mathbf{x}_i'\boldsymbol{\beta}$.
似然函数 $L(\beta)=\Pi_{i=1}^{n}h(y_i)\exp(\theta_iy_i-A(\theta_i))$
对数似然函数 $l(\boldsymbol{\beta})=\sum_{i=1}^{n}\ln(h(y_i))+\theta_iy_i-A(\theta_i)$.

$$\frac{\partial l}{\partial\boldsymbol{\beta}}=\sum_{i=1}^{n}\frac{\partial l_i}{\partial\theta_i}\frac{\partial\theta_i}{\partial\mu_i}\frac{\partial\mu_i}{\partial\eta_i}\frac{\partial\eta_i}{\partial\boldsymbol{\beta}}=\sum_{i=1}^{n}\frac{\boldsymbol{x}_i(y_i-\mu_i)}{\mathrm{Var}(Y_i)g'(\mu_i)}$$
$$\overset{\text{Canonical Link}}{=}\sum_{i=1}^{n}\boldsymbol{x}_i(y_i-\mu_i)=\sum_{i=1}^{n}\boldsymbol{x}_i(y_i-g^{-1}(\boldsymbol{x}_i^T\boldsymbol{\beta}))$$

得分函数 $\mathbf{U}(\boldsymbol{\beta})=\partial l/\partial\boldsymbol{\beta}=\mathbf{0}$. 使用迭代法从$\boldsymbol{\beta}_{(0)}$出发求解**β**.

信息矩阵 Information Matrix: $J(\boldsymbol{\beta}_{(0)})=-\partial^2l/\partial\boldsymbol{\beta}^2|_{\boldsymbol{\beta}=\boldsymbol{\beta}(0)}$

$\mathbf{U}(\boldsymbol{\beta})\approx\mathbf{U}(\boldsymbol{\beta}_{(0)})-J(\boldsymbol{\beta}_{(0)})(\boldsymbol{\beta}-\boldsymbol{\beta}_{(0)})$. $\to\boldsymbol{\beta}_{(m+1)}=\boldsymbol{\beta}_{(m)}+[J(\boldsymbol{\beta}_{(m)})]^{-1}\mathbf{U}(\boldsymbol{\beta}_{(m)})\to_{m\to\infty}\boldsymbol{\beta}^{\wedge}$.
推导结果：$\boldsymbol{\beta}^{\wedge}=(X'W_{(m)}X)^{-1}X'W_{(m)}\mathbf{z}_{(m)}$,其中 $z_i^{(m)}=\eta_i^{(m)}+(y_i-\mu_i^{(m)})g'(\mu_i^{(m)})$.
$w_i^{(m)}=1/(Var(Y_i|\beta_{(m)})[g'(\mu_i^{(m)})]^2)$. $W_{(m)}=diag(w_i^{(m)})$.
$Var(\mathbf{U})=E(\mathbf{UU}')=[\sum_{i=1}^{n}x_{ij}x_{jk}/(Var(Y_i)[g'(\mu_i)]^2)]_{jk}=X'WX=V$

MLE:

在常见的条件下，近似地，$\mathbf{U}\sim N(0, V)$
$\mathbf{U}(\boldsymbol{\beta}^{\wedge})=0$, $\mathbf{U}(\boldsymbol{\beta})\approx\mathbf{U}(\boldsymbol{\beta}^{\wedge})-J(\boldsymbol{\beta}^{\wedge})(\boldsymbol{\beta}-\boldsymbol{\beta}^{\wedge})\to(\boldsymbol{\beta}-\boldsymbol{\beta}^{\wedge})\approx[J(\boldsymbol{\beta}^{\wedge})]^{-1}\mathbf{U}(\boldsymbol{\beta})$.
$J(\boldsymbol{\beta}^{\wedge})=-\partial^2l/\partial\boldsymbol{\beta}^2|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{\wedge}}$称为观测信息矩阵。
$E(\mathbf{U})=\mathbf{0}\to E(\boldsymbol{\beta}^{\wedge}-\boldsymbol{\beta})\approx\mathbf{0}$, 渐近地 $E(\boldsymbol{\beta}^{\wedge})=\boldsymbol{\beta}$. $E(J(\boldsymbol{\beta}))=V=X'WX$
渐近地$\boldsymbol{\beta}^{\wedge}\sim N(\boldsymbol{\beta}, V^{-1})=N(\boldsymbol{\beta}, (X'WX)^{-1})$
CI: $\beta_j\in[\beta_j^{\wedge}\pm z_{\alpha/2}\sqrt{(V^{-1\wedge})_{jj}}]$

假设检验 Wald Test

H_0: $\beta_j=0$ vs. H_1: $\beta_j\neq0$. $Z_j=\beta_j^{\wedge}/\sqrt{V^{-1\wedge}_{jj}}\sim_{asymp, H0}N(0,1)$

似然比检验 Likelihood-Ratio Test

H_0: $\beta_{k+1}=...=\beta_p=0$ vs. H_1: $\beta_j\neq0$, $k+1\leq j\leq p$.
新的限制加入模型时，L(**β**)不增。
$\Lambda=-2\ln(L(\boldsymbol{\beta}^{\wedge}_{Reduced})/L(\boldsymbol{\beta}^{\wedge}_{Full}))=2(l(\boldsymbol{\beta}^{\wedge}_{Full})-l(\boldsymbol{\beta}^{\wedge}_{Reduced}))\sim_{a, H0}\chi^2(p-k)$.
拒绝域：$\{\Lambda\geq\chi^2(\alpha, p-k)\}$。注意拒绝域是单边的。

指数分布族二三事

一般的自然指数分布族 $f(y;\theta,\varphi)=h(y;\varphi)\exp((y\theta-A(\theta))/\varphi)$.
$E(Y)=A'(\theta)$; $Var(Y)=\varphi A''(\theta)=\varphi\partial\mu/\partial\theta=\varphi V(\mu)$. V(μ)是方差函数

分布	θ	φ	A(θ)	μ	V(μ)
N	μ	σ^2	$\theta^2/2$	μ	1
Bin	ln p/(1-p)	1	$n\ln(1+e^\theta)$	np	μ(1-μ/n)
Poisson	ln λ	1	e^θ	λ	μ
$\Gamma_{\mu,v}$	-1/μ	1/v	-ln(-θ)	μ	μ^2

$\Gamma(y;\mu,v)$: $f(y;\mu,v)=(yv/\mu)^ve^{-yv/\mu}/(y\Gamma(v))$.

拟合优度 Goodness-of-Fit

饱和模型 Saturated Model 令$\mu_i^S=y_i$
Scaled Deviance $D^*(M)=2(l(\boldsymbol{\beta}^{\wedge}_S)-l(\boldsymbol{\beta}^{\wedge}_M))$
Deviance $D(M)=\varphi D^*(M)$.
......

Logistic Regression for Binary Variables

Logit Link: $logit(p_i)=\mathbf{x}_i'\boldsymbol{\beta}=\beta_0+\beta_1x_{i1}...+\beta_kx_{ik}$. $logit(p)=\ln(p/(1-p))$.
Recall Y~Bernoulli(p), $f(y;p)=p^y(1-p)^{1-y}=\exp(y\ln(p/(1-p))+\ln(1-p))$.
$h(y)=1$, $\theta=\ln(p/(1-p))$, $\varphi=1$, $A(\theta)=-\ln(1-p)=\ln(1+e^\theta)$
$A'(\theta)=e^\theta/(1+e^\theta)\to g(p)=(A')^{-1}(p)=\ln(p/(1-p))$.

于是 Logit Link:[0, 1]→R 是典型连接函数.

其他连接函数：

ProbitLink $g(p)=\Phi^{-1}(p)$
Complementary log-log Link: $g(p)=\ln(-\ln(1-p))$.

odds=p/(1-p)=Pr(Y=1)/Pr(Y=0).
odds ratio OR=$p_1(1-p_2)/(p_2(1-p_1))$.
若 x_j 指示类别变量处于第 l 水平，则 $\exp(\beta_j)$是 Y=1 时 l 水平的总体和非 l 水平总体的 odds ratio.

Iteratively Reweighted Least Squares Algorithm IRLS

用于得到$\boldsymbol{\beta}^{\wedge}$的 MLE.
$\boldsymbol{\beta}^{\wedge}\sim_{asymp}N(\boldsymbol{\beta}, (X'WX)^{-1})$, $W=diag(p_i^{\wedge}(1-p_i^{\wedge}))$.
$l(\boldsymbol{\beta})=\sum_{i=1}^{n}y_i\ln(p_i)+(1-y_i)\ln(1-p_i)$.
$l(\boldsymbol{\beta}^{\wedge}_S)=\sum_{i=1}^{n}y_i\ln(y_i)+(1-y_i)\ln(1-y_i)=0$
$D=-2\sum_{i=1}^{n}y_i\ln(p_i^{\wedge})+(1-y_i)\ln(1-p_i^{\wedge})$
for binomial data,$D=-2\sum_{i=1}^{n}y_i\ln(y_i/(n_ip_i^{\wedge}))+(1-y_i)\ln((n_i-y_i)/(n_i(1-p_i^{\wedge})))$

Or $\chi^2=\sum_{i=1}^{n}(y_i-n_ip_i^{\wedge})^2/(n_ip_i^{\wedge}(1-p_i^{\wedge}))$
D 和χ^2均渐近服从$\chi^2(n-p-1)$. 要求 $n_i\geq10$, $n_ip_i^{\wedge}\geq5$.

n_i不够大时，使用 Hosmer-Lemeshow(HL)检验拟合优度

预测能力 Predictive Power

C=#Concordant Pairs(类别为 1 样本的预测值比类别为 0 的高)
D=#Discordant Pairs
T=#Ties
Somer's D=(C-D)/(C+D+T)
Gamma=(C-D)/(C+D)
Tau-a=(C-D)/N
Concordance Index c=(C+0.5T)/(C+D+T).

Confusion Matrix

	Y^=1	Y^=0
Y=1	TP	FN
Y=0	FP	TN

Correct Prediction Rate/Accuracy=(TP+TN)/n
True Positive Rate/Sensitivity/Recall=TP/(TP+FN)
True Negative Rate/Specificity=TN/(TN+FP)
Positive Predicted Value/Precision=TP/(TP+FP)
Negative Predicted Value=TN/(TN+FN)
调整阈值不能使 Sensitivity 和 Specificity 同时增加。
ROC 曲线：1-Specificity - Sensitivity.
AUC: ROC 的曲线下面积，和 Concordance Index 一样。其意义是随机选取的正样本的预测值高于随机选取的负样本的预测值的概率

计数值的泊松回归 Poisson Regression for Counts

因变量为离散的计数值。$Pr(Y=y)=\mu^ye^{-\mu}/y!$, $y\in N$.
二项分布的极限分布是泊松分布。
$E(Y)=Var(Y)=\mu$. $h(y)=1/y!$, $\theta=\ln\mu$, $A(\theta)=\mu=e^\theta$. $\varphi=1$. $(A')^{-1}(\mu)=\ln\mu$
Poisson Regression Model / Log-linear Model
$\ln(\mu_i)=\mathbf{x}_i'\boldsymbol{\beta}+(\ln T_i)$. $\ln T_i$ 称为 offset.

过度散布的调整 Adjusting for Overdispersion

$\boldsymbol{\beta}^{\wedge}$不受过度散布影响，但其方差会。φ可以解释过度散布 $Var(Y_i)=\varphi\mu_i$.
φ可以被 Pearson χ^2/DF 或 Deviance/DF 估计.
或者可以用负二项分布解释过度散布。1/r 称为负二项散布参数。
记 $E(Y)=pr/(1-p)=\mu$，则 $Var(Y)=pr/(1-p)^2=\mu+\mu^2/r>\mu$.

零膨胀模型 Zero-Inflated Models

对于某些计数数据，其零的个数会显著超过假设的分布。这时，应该对过量 0 的产生单独建模。
Zero-Inflated Poisson Model
$Y_i=0$ with $Pr=\pi_i$; $Y_i\sim Poisson(\mu_i)$ with $Pr=1-\pi_i$.
$Pr(Y_i=0)=\pi_i+(1-\pi_i)e^{-\mu i}$; $Pr(Y_i=k)=(1-\pi_i)e^{-\mu i}\mu_i^k/k!$.
$logit(\pi_i)=\mathbf{z}_i'\boldsymbol{\gamma}$; $\ln\mu_i=\mathbf{x}_i'\boldsymbol{\beta}$.
过量的 0 也是一种过度散布的形式。如果其仍有过度散布存在，可以使用 Zero-Inflated Negative Binomial(ZINB)模型。

多水平因变量的逻辑回归 Logistic Regression for Multilevel Response

Proportional Odds Model for Ordinal Response

以有 3 个水平的有序类别变量为例
p_{ij1}, p_{ij2}, p_{ij3}代表其他类别变量为 i，j 时，有序类别变量为 1-3 的概率。
记θ_{ijk}为累积概率，如$\theta_{ij2}=p_{ij1}+p_{ij2}$.
累积 logit: $logit(\theta_{ij1})=\ln(p_{ij1}/(p_{ij2}+p_{ij3}))$; $logit(\theta_{ij2})=\ln((p_{ij1}+p_{ij2})/p_{ij3})$...
拟合模型：$logit(\theta_{ijk})=\alpha_k+\mathbf{x}_{ij}'\boldsymbol{\beta}_k$.
Proportional Odds Assumption: $logit(\theta_{ijk})=\alpha_k+\mathbf{x}_{ij}'\boldsymbol{\beta}$.
对比子总体 $logit(\theta_{ijk})-logit(\theta_{ij'k})=(\mathbf{x}_{ij}-\mathbf{x}_{ij'})'\boldsymbol{\beta}$.
自由度为 p(r-2). p 为其他自变量的个数、r 为 logit 的个数

作业题

Show that the weighted least squares estimate defined by Eq. (5.17) in the lecture notes is the best linear unbiased estimate (BLUE) of β.

Solution: Now we want to show that is the best linear unbiased estimate of β, where $W=[Var(y)]^{-1}$. The variance of $\beta^{\wedge WLS}$ is

$Var(\beta^{\wedge WLS})=((X'WX)^{-1}X'W)Var(\mathbf{y})((X'WX)^{-1}X'W)'=(X'WX)^{-1}$

Assume there is another linear estimation of

β: $\beta^{\sim}=A\mathbf{y}$, $A=(X'WX)^{-1}X'W+B$, then it must satisfies

$E(\beta^{\sim})=E(((X'WX)^{-1}X'W+B)(X\beta+\varepsilon))=(I+BX)\beta=\beta$.

i.e. BX = 0, or it would not be unbiased estimator of **β**. Yet

$Var(\beta^{\sim})=(X'WX)^{-1}+BW^{-1}B'=Var(\beta^{\wedge WLS})+BW^{-1}B'>Var(\beta^{\wedge WLS})$

then it is clear that $\beta^{\wedge WLS}$ is the best linear unbiased estimate of **β**.

$Y_1,...,Y_n$ are independent and $Y_i\sim Poisson(\mu_i)$. Let M be a model of interest and $\hat{y}i$ be the estimated value of Yi under model M (y_i is the observed value of Y_i).

(1) Show that the deviance of model M is
$D=2[\sum yi\log(yi/\hat{y}i)-\sum(yi-\hat{y}i)]$.

Solution:The log likelihood function of parameters is

$$l(\boldsymbol{\mu};\boldsymbol{y})=\sum_{i=1}^{n}y_i\ln\mu_i-\sum_{i=1}^{n}\mu_i-\sum_{i=1}^{n}\ln y_i!$$

It is obvious that the MLE. of μ is μi=yi for all i's, hence

$$l(\hat{\boldsymbol{\mu}}_S;\boldsymbol{y})=\sum_{i=1}^{n}y_i\ln y_i-\sum_{i=1}^{n}y_i-\sum_{i=1}^{n}\ln y_i!$$

Yet for the model of interest M, the MLE. is derived by

$$l(\hat{\boldsymbol{\mu}}_M;\boldsymbol{y})=\sum_{i=1}^{n}y_i\ln\hat{\mu}_i-\sum_{i=1}^{n}\hat{\mu}_i-\sum_{i=1}^{n}\ln y_i!$$
$$=\sum_{i=1}^{n}y_i\ln\hat{y}_i-\sum_{i=1}^{n}\hat{y}_i-\sum_{i=1}^{n}\ln y_i!$$

Thus, the deviance is given by

$$D(M)=2(l(\hat{\boldsymbol{\mu}}_S;\boldsymbol{y})-l(\hat{\boldsymbol{\mu}}_M;\boldsymbol{y}))=2(\sum_{i=1}^{n}y_i\ln\frac{y_i}{\hat{y}_i}-\sum_{i=1}^{n}(y_i-\hat{y}_i))$$

(2) Let $\log(\mu_i)=\beta_0+\beta_1x_{i1}+\cdots+\beta_px_{ip}$.
Show that the score statistic for β_0 is $U_0=\sum(y_i-\mu_i)$.

Solution: We can know that

$$\mu_i=\exp(\beta_0+\sum_{k=1}^{p}\beta_kx_{ik})$$

Then the score statistic for β_0 is

$$U_0=\frac{\partial l(\boldsymbol{\beta};\boldsymbol{y})}{\partial\beta_0}$$
$$=\sum_{i=1}^{n}y_i\frac{\partial\ln\mu_i}{\partial\beta_0}-\sum_{i=1}^{n}\frac{\partial\mu_i}{\partial\beta_0}$$
$$=\sum_{i=1}^{n}(y_i-\mu_i)$$

(3) Show that the deviance of model M simplifies to
$D=2\sum y_i\log(y_i/\hat{y}_i)$.

Solution: When applying the model M, we need to satisfy

$$U_0=\sum_{i=1}^{n}(y_i-\mu_i)=0$$

That is to say, the estimation must satisfies

$$\sum_{i=1}^{n}\hat{y}_i=\sum_{i=1}^{n}\hat{\mu}_i=\sum_{i=1}^{n}y_i$$

Then the deviance is simplified to

$$D(M)=2(\sum_{i=1}^{n}y_i\ln\frac{y_i}{\hat{y}_i}-\sum_{i=1}^{n}(y_i-\hat{y}_i))$$
$$=2\sum_{i=1}^{n}y_i\ln\frac{y_i}{\hat{y}_i}$$

Let $Xi\sim^{\text{i.i.d.}}\mathcal{N}(\mu_1,\sigma_1^2)$ and $Y_i\sim^{\text{i.i.d}}.\mathcal{N}(\mu_2,\sigma_2^2)$ are two independent samples. The corresponding sample size, sample mean, and sample standard deviation are given below: $n_1=17$,$\bar{X}=13.5$, $S_1=5.5$ $n_2=19$,$\bar{Y}=10.5$, $S_2=4.5$
(1) Test for equal variance: H_0: $\sigma_1^2=\sigma_2^2$ vs.H_1: $\sigma_1^2\neq\sigma_2^2$ at $\alpha=0.05$.
Solution: We want to test H_0: $\sigma_1=\sigma_2$ against H1: σ1≠σ2.
Under H_0, we have S_1^2/S_2^2 as test statistic:

$$\frac{\sigma_2^2S_1^2}{\sigma_1^2S_2^2}=\frac{S_1^2}{S_2^2}\sim F(n_1-1,n_2-1)=F(16,18)$$

Because $S_1^2/S_2^2=1.4938>1$, the critical region is given by

$$\overline{[F(0.975,16,18),F(0.025,16,18)]}=(-\infty,0.368)\cup(2.640,+\infty)$$

where $F(\alpha, n_1, n_2)$ is the upper α-th quantile of $F(n_1, n_2)$.
p-value=2Pr(F(16, 18)>1.4938|H_0)=0.41>0.05, hence we cannot reject H_0 under the significance level of 95%.

(2) Assuming $\sigma_1^2=\sigma_2^2$, construct a 95% confidence interval for $\mu_1-\mu_2$.
Solution: With $\sigma_1^2=\sigma_2^2$, denoted as σ^2, we have

$$\bar{X}\sim N(\mu_1,\frac{\sigma^2}{n_1}),\bar{Y}\sim N(\mu_2,\frac{\sigma^2}{n_2}),$$
$$\bar{X}-\bar{Y}\sim N(\mu_1-\mu_2,(\frac{1}{n_1}+\frac{1}{n_2})\sigma^2),$$
$$Z=\frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sigma\sqrt{1/n_1+1/n_2}}\sim N(0,1)$$

,

and

$$S_p^2=\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2},$$
$$W=\frac{(n_1+n_2-2)S_p^2}{\sigma^2}\sim\chi^2(n_1+n_2-2)$$

therefore,

$$T=\frac{Z}{\sqrt{W/(n_1+n_2-2)}}=\frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{S_p\sqrt{1/n_1+1/n_2}}\sim t(n_1+n_2-2)$$
$$\because S_p=\sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}}=4.9956,$$
$$\therefore T=\frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{S_p\sqrt{1/n_1+1/n_2}}=\frac{3-(\mu_1-\mu_2)}{1.6678}$$

Hence a 95% CI of μ1-μ2 is given by

$$[\bar{X}-\bar{Y}\pm t(\alpha/2,34)S_p\sqrt{1/n_1+1/n_2}]=[-0.3894,6.3894]$$

(3) Suppose that the underlying true mean difference between the two population is $\mu_1-\mu_2=2.5$ and it is given that $\sigma_1=\sigma_2=\sigma=5$, compute the Type II error rate if the two-sample z-test is applied to test H_0: $\mu_1-\mu_2=0$ vs. H_1: $\mu_1-\mu_2>0$ with samples of sizes $n_1=17$ and $n_2=19$. Use $\alpha=0.05$.
Solution: Now we want to compute the Type II error rate of the test H_0: $\mu_1-\mu_2=0$ against H_1: $\mu_1-\mu_2>0$.

$$\because Z=\frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sigma\sqrt{1/n_1+1/n_2}}\sim N(0,1)$$

the critical region is given by

$$[\Phi(1-\alpha)\sqrt{1/n_1+1/n_2}\sigma,+\infty)=[2.744,+\infty)$$

Therefore, by using the fact of

$$Z=\frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sigma\sqrt{1/n_1+1/n_2}}\sim N(0,1)$$

,

we have

$$\beta=\Pr(\bar{X}-\bar{Y}<\Phi(1-\alpha)\sqrt{1/n_1+1/n_2}\sigma|H_1)$$
$$=\Pr(N(0,1)<z_\alpha-\frac{\mu_1-\mu_2}{\sqrt{1/n_1+1/n_2}\sigma})$$
$$=\Phi(z_\alpha-\frac{\mu_1-\mu_2}{\sqrt{1/n_1+1/n_2}\sigma})$$
$$=0.5584$$

where z_α is the upper α-th quantile of standard normal distribution, and Φ(·) is the c.d.f. of standard normal distribution.

(4) Suppose that the underlying true mean difference between the two population is $\mu_1-\mu_2=2.5$ and it is given that$\sigma_1=\sigma_2=\sigma=5$, compute the minimum total sample size $n=n_1+n_2$ such that the power is at least 0.8 when applying the two-sample ztest to test H_0: $\mu_1-\mu_2=0$ vs. H_1: $\mu_1-\mu_2>0$. Use $\alpha=0.05$.
Solution: Now we want p = 1-β ⩾ 0.8, that is to say,

$$\Phi(z_\alpha - \frac{\mu_1-\mu_2}{\sqrt{1/n_1+1/n_2\sigma}}) \le 0.2$$

$$\therefore z_\alpha - \frac{\mu_1-\mu_2}{\sqrt{1/n_1+1/n_2\sigma}} \le z_{0.8}$$

$$\therefore 1/n_1+1/n_2 = \frac{n_1+n_2}{n_1n_2} \le (\frac{\mu_1-\mu_2}{(z_\alpha - z_{0.8})\sigma})^2 = 0.040436$$

$$\because n_1+n_2 \le 0.040436 n_1n_2 \le 0.040436\frac{(n_1+n_2)^2}{4}$$

$$\therefore n_1+n_2 \ge \frac{4}{0.040436} > 98$$

When $n_1 = 49$, $n_2 = 50$, $(n_1 + n_2) / (n_1n_2) = 0.040408 < 0.040436$, hence the minimum of $n_1 + n_2$ is 99.

SAS 代码

SAS 基础

```
DATA
INPUT  字符变量  $  长度  .  数字变量;
 字符缺失值''  数字缺失值.  小于所有数字
SET  从存在的数据集中加载;
ARRAY song (5) wj -- ttr;
DO i = 1 TO 5;
   IF song(i) = 9 THEN song(i) = .;
END;

PROC IMPORT DATAFILE="file.csv" OUT=mydata
REPLACE;

PROC SORT DATA = mydata;
BY = mycol;

DATA MergeData;
   Merge Data1 Data2;
   By mycol;

PROC MEANS DATA=mydata MAXDEC=3; /*保留 3 位小数*/
   VAR mycol;
   CLASS mycls1 mycls2;  声明类别变量
   TYPES() mycls1 * mycls2;  声明需要总结的分类  ()代表不分类
   TITLE "my title";

PROC MEANS DATA=mydata MIN MAX MEDIAN;
   BY = section;  分类
   VAR score;  用于统计的变量
   CLASS mycls1 mycls2;
   TITLE1 "title line 1";
   TITLE2 "title line 2";

PROC FREQ DATA=mydata;
   TABLES mycol1 * mycol2;  产生交叉表
   TABLES mycol1 * mycol2 / MISSING NOPERCENT NOCOL NOROW;

PROC FORMAT;
   VALUE myfmt 1="E" 2="W" 3="S" 4="N";  左边替换为右边

PROC TABULATE DATA=mydata;
   CLASS cls1 cls2 cls3;
   VAR myvar;
   TABLE cls1 * cls2, cls3 * myvar * FORMAT=dollar12.;  纵轴，横轴
   TABLE cls1 * cls2, (cls3 ALL) * myvar * (SUM MEAN N); ALL:不分类
   FORMAT cls1 myfmt. ...;

PROC SGPLOT DATA=mydata;
   SCATTER x=myx y=myy / GROUP=mygrp;
/
   HISTOGRAM myvar;  条形图（连续变量）
   DENSITY myvar / TYPE=KERNEL;  密度图（可选核密度图）
/
   VBAR mycls / GROUP=mygrp GROUPDISPLAY=CLUSTER;  频数条形
图
/
   VBOX myvar / CATEGORY=mycls;
```

假设检验 Hypothesis Testing

```
PROC FREQ DATA=mydata;
   TABLES pred / Binomial (Level=1 P=0.5 Wald Wilson Exact)
Alpha=0.05;
   EXACT Binomial;  检验变量 pred 为 Level 的概率是否显著不同于 P.


PROC FREQ DATA=mydata;
   TABLES cls1 * cls2 / RISKDIFF (EQUAL VAR=NULL CL=Wald
NORISKS);
   WEIGHT N;  指定频数列
   检验 cls1 不同是否在 cls2 中无影响。(p1=p2)

PROC FREQ DATA=mydata;
   TABLES cls1 * cls2 / CHISQ;
   EXACT BARNARD;
   WEIGHT N;  指定频数列

PROC FREQ DATA=mydata;
   TABLES cls1 * cls2 / AGREE;  提供 McNemar 检验
   EXACT MCNEM;  提供精确的 McNemar 检验
   WEIGHT N;  指定频数列

PROC MEANS DATA=mydata T PRT MEAN STD ALPHA=0.05 CLM;
   VAR X;  检验均值为 0，提供 t 统计量、双边 p 值、双边置信区间

PROC UNIVARIATE DATA=mydata MU0=200 CIBASIC ALPHA=0.05;
   VAR Y;
MU0 定义零假设的μ0. CIBASIC 提供均值方差等的置信区间（基于正
态假设）

PROC TTEST DATA=mydata H0=200 SIDES=2 ALPHA=0.05;
   VAR Y; (SIDES = 2 / L / U)  提供条形图、箱线图、均值区间、QQ 图

PROC TTEST DATA=mydata;  提供双样本 t 检验
   CLASS cls;
   VAR Y;

PROC MULTTEST PDATA(pvalue) = Pvals BON HOLM;
   提供 p 值的调整方法，如 Bonferroni/Holm 等。pvalue 是数据集
```

多元线性回归、模型选择与诊断

```
PROC GLM DATA=mydata;
   MODEL Y=X1 | X2;  等价于 Y=X1 X2 X1*X2

PROC GLMSELECT DATA = mydata
   PLOTS=(CriterionPanel CoefficientPanel);
   CLASS cls;
   MODEL Y = X1 X2 X3 X4 X5 X6 cls
      / SELECTION=STEPWISE(SELECT=SL)
STATS=(ADJRSQ CP AIC SBC SL);
   SL: F 统计量对应的显著水平

PROC GLMSELECT DATA = mydata PLOTS=CandidatesPlot SEED=123;
   CLASS cls;
   MODEL Y = X1 X2 X3 X4 X5 X6 cls /
      SELECTION=STEPWISE(SELECT=CV DROP=Competitive)
      CVMETHOD=Random(5);
CV:  将 Predicted Residual Sum of Squares 用于交叉检验
DROP=BeforeAdd 先考虑丢弃再加入/Competitive 同时考虑丢弃加入
DROP 仅对于 STEPWISE 有效;SELECTION=FORWARD/BACKWARD/...

PROC GLMSELECT DATA = mydata PLOTS=(CriterionPanel ASEPlot);
   PARTITION FRACTION(Validate=0.2 Test=0.2); 60%用于训练
   CLASS cls;
   MODEL Y = X1 X2 X3 X4 X5 X6 cls
      / SELECTION=STEPWISE(SELECT=AIC CHOOSE=Validate);

PROC MODEL DATA=SimModel1;
   PARMS b0 b1;  声明模型参数
   y = b0 + b1 * x;    用 White 和 Breusch-Pegan 检验异方差性
   FIT y / WHITE PAGAN=(1 x);

PROC GLM DATA=SimModel1;MODEL y = x;
OUTPUT OUT = SimModel1_fitted (Keep=y x r fv) Residual=r
Predicted=fv;
PROC UNIVARIATE DATA=SimModel1_fitted NORMAL;
VAR r;提供 Shapiro-Wilk(n<2000)、Kolmogorov-Smirnov 正态性检验
PROC MODEL DATA=SimModel1_fitted;
   PARMS b0;    r = b0;
FIT r/NORMAL;Shapiro-Wilk(n<2000)Kolmogorov-Smirnov(n ⩾2000)
/* Obtain ordinary least squares estimate */
PROC GLM DATA=SimModel2 PLOTS=DIAGNOSTICS(UNPACK);
   MODEL y = x; OUTPUT OUT = SimModel2_fitted (Keep=y x r fv)
      Residual=r Predicted=fv;
DATA SimModel2_fitted; SET SimModel2_fitted; rAbs = ABS(r);

/* Auxilliary regression of the absolute residuals */
PROC GLM DATA=SimModel2_fitted;
   MODEL rAbs = x;
   OUTPUT OUT = Temp (Keep=y x fv2) Predicted=fv2;

DATA TEMP; SET TEMP; w = 1 / (fv2 ** 2);

/* Obtain weighted least squares estimate */
PROC GLM DATA=Temp;
   MODEL y = x;    WEIGHT w;

PROC GLM DATA=Crime PLOTS=DIAGNOSTICS(UNPACK);
   Model crime = pctMetro poverty single;
   OUTPUT OUT = CrimeRes1(KEEP = state crime pctmetro poverty
      single rstu lev cd) Rstudent=rstu h=lev CookD=cd;
PROC PRINT DATA=CrimeRes1;  找到异常样本
   WHERE ABS(rstu) > 2 OR lev > 8/51 OR cd > 4 / 51;
PROC GLM DATA=Crime PLOTS=DIAGNOSTICS(UNPACK);
   MODEL crime = pctMetro poverty single;
   WHERE state ne "DC";  手动删去异常点
```

方差分析

```
/* One-way ANOVA with raw SAS programing */
DATA _NULL_;
   n1 = 5; n2 = 7; n3 = 7; k = 3; n = n1 + n2 + n3;
   ybar1 = MEAN(60.8, 74.0, 69.8, 71.6, 67.5);
   ybar2 = MEAN(102.6, 102.1, 98.7, 106.8, 89.5, 96.5, 99.7);
   ybar3 = MEAN(87.9, 84.2, 90.3, 77.6, 86.9, 75.2, 82.7);
   ybar = (n1 * ybar1 + n2 * ybar2 + n3 * ybar3) / n;
   s1 = STD(60.8, 74.0, 69.8, 71.6, 67.5);
   s2 = STD(102.6, 102.1, 98.7, 106.8, 89.5, 96.5, 99.7);
   s3 = STD(87.9, 84.2, 90.3, 77.6, 86.9, 75.2, 82.7);
  SSB=n1*(ybar1-ybar)**2+n2*(ybar2-ybar)**2+ n3*(ybar3-ybar)**2;
   SSW = (n1 - 1) * s1 * s1 + (n2 - 1) * s2 * s2 + (n3 - 1) * s3 * s3;
   F_stat = (SSB / (k - 1)) / (SSW / (n - k));
   p_value = 1 - CDF("F", F_stat, k - 1, n - k);

/* Perform ANOVA */
PROC GLM DATA = Feeds;
   CLASS feed;
   MODEL weight = feed;
   MEANS feed / HOVTEST = LEVENE (TYPE = ABS);
MEANS 为每个类别变量的各类计算均值标准差
   HOVTEST 还为各组做同方差检验
HOVTEST=BARTLETT/LEVENE(ABS SQUARE)/BF (Brown-Forsythe)

* Take the log transformation on survtimes;
DATA CancerSurvival2;
   SET CancerSurvival;
   logsurv = LOG(survtime);  同方差性被拒绝，可考虑对数变换
* Perform the ANOVA on the transformed survtime;
PROC GLM DATA = CancerSurvival2;
   CLASS organ;
   MODEL logsurv = organ;  提供多元比较检验↓
   MEANS organ / HOVTEST = BF TUKEY DUNNETT("breast");
OUTPUT OUT=CancerSurvOut PREDICTED = pred RESIDUAL = resid;

* Perform the Kruskal-Wallis Test;
* The WILCOXON option requests an analysis of Wilcoxon scores;
* The ANOVA option requests a standard analysis of variance on the raw data;
* The DSCF option requests a particular multiple comparison procedure;
PROC NPAR1WAY DATA = CancerSurvival2 WILCOXON ANOVA DSCF;
   CLASS organ;
   VAR logsurv;

/* Fit interaction model */
PROC GLM DATA = PygmalionEffect;
   CLASS Company Treat;
   MODEL Score = Company Treat Company * Treat;
/* Fit interaction model with different order */
PROC GLM DATA = PygmalionEffect;
   CLASS Company Treat;
   MODEL Score = Treat Company Company*Treat;
/* Fit additive model */
PROC GLM DATA = PygmalionEffect;
   CLASS Company Treat;
   MODEL Score = Company Treat;
/* Fit factor treatment only model */
PROC GLM DATA = PygmalionEffect;
   CLASS Treat;
   MODEL Score = Treat;
/* Perform two-sample t-test */
PROC TTEST DATA = PygmalionEffect;
   CLASS Treat;
   VAR Score;

DATA DrugTest;INPUT Drug $ PreTreatment PostTreatment @@;
DATALINES;
A 11   6     A   8   0     A   5   2     A 14   8     A 19 11
A   6   4    A 10 13     A   6   1     A 11   8     A   3   0
D   6   0    D   6   2     D   7   3     D   8   1     D 18 18
D   8   4    D 19 14     D   8   9     D   5   1     D 15   9
F 16 13     F 13 10     F 11 18     F   9   5     F 21 23
F 16 12     F 12   5     F 12 16     F   7   1     F 12 20;

/* Fit Model I */
PROC GLM DATA = DrugTest;
   CLASS Drug;
   MODEL PostTreatment = Drug PreTreatment Drug * PreTreatment;
/* Fit Model II */
PROC GLM DATA = DrugTest;
   CLASS Drug;
   MODEL PostTreatment = Drug PreTreatment;
/* Fit Model II */
PROC GLM DATA = DrugTest;
    CLASS Drug;
    MODEL PostTreatment = PreTreatment Drug;

/* Boxplot of PreTreatment by Drug group */
PROC SGPLOT DATA = DrugTest;
         VBOX PreTreatment / GROUP = Drug;

/* Fit Model IV, i.e., simple linear regression */
PROC GLM DATA = DrugTest;
    MODEL PostTreatment = PreTreatment;

* Compare MEANS and LSMEANS;
* The SOLUTION option produces a solution to the normal equations
(parameter estimates);
PROC GLM DATA = DrugTest;
    CLASS Drug;
    MODEL PostTreatment = Drug PreTreatment / SOLUTION;
    LSMEANS Drug;
    MEANS Drug;

* Add options to LSMEANS;
* The PDIFF option requests that p-values for differences of the LS-means be
produced;
* The STDERR option produces the standard error of the LS-means and the
probability level for the hypothesis
H0: LS-mean = 0;
* The ADJUST option requests a multiple comparison adjustment for the
p-values and confidence limits for the differences of LS-means;
PROC GLM DATA = DrugTest;
    CLASS Drug;
    MODEL PostTreatment = Drug PreTreatment;
    LSMEANS Drug / PDIFF STDERR ADJUST=TUKEY;
```

广义线性模型

```
* Fit logistic regression with main effects;
* The FREQ statement identifies a variable that contains the frequency of
occurrence of each observation;
* The PARAM option specifies the parameterization method for the
classification variable or variables;
PROC LOGISTIC DATA = Byss;
   FREQ count;
   CLASS envir years(REF = first) smoke(REF = first) / PARAM = REF;
   MODEL complaint(Event=last) = envir years smoke;

/* 'complaint(EVENT = last)' is equivalent with the DESCENDING option */
PROC LOGISTIC DATA = Byss DESCENDING;
   FREQ count;    REF=first 意味着以第一个遇到的类别作为基准
   CLASS envir years(REF = first) smoke(REF = first) / PARAM = REF;
   MODEL complaint = envir years smoke;

/* Fit without PARAM = REF, i.e., use the default PARAM = EFFECT */
PROC LOGISTIC DATA = Byss;
   FREQ count;
   CLASS envir years(REF = first) smoke(REF = first);
   MODEL complaint(EVENT = last) = envir years smoke;

/* Obtain plots for odds ratio and probabilities */
PROC LOGISTIC DATA = Byss
   PLOTS(ONLY) = (EFFECT(CLBAND YRANGE = (0,0.3)
    X = envir*years*smoke)ODDSRATIO);
   FREQ count;
   CLASS envir years(REF = first) smoke(REF = first) / PARAM = REF;
   MODEL complaint(EVENT = last) = envir years smoke;

/* Fit logistic regression with all pairwise interactions */
PROC LOGISTIC DATA = Byss;
   FREQ count;
   CLASS envir years(REF = first) smoke(REF = first) / PARAM = REF;
   MODEL complaint(EVENT = last) = envir|years|smoke@2;
```

```
* Exclude the nonsignificant smoke*year interaction;
* The INTERACTION option displays a curve of predicted values versus an explainatory variable grouped by the levels of a CLASS effect;
* The LINK option displays the fit on the scale of the link function, i.e., the linear predictor;
PROC LOGISTIC DATA = Byss;
    FREQ count;
    CLASS envir years(REF = first) smoke(REF = first) / PARAM = ref;
    MODEL complaint(EVENT = last) = envir years smoke envir*years envir*smoke;
    EFFECTPLOT INTERACTION (X = envir) / AT(years = all smoke = all) LINK NOOBS;
    ODDSRATIO envir;

/* An alternative data format */
DATA Byss2;
INPUT envir $ years $ smoke $ numComp numTotal;
DATALINES;
dusty     <10      yes      30     233
dusty     <10      no        7     126
dusty     >=10     yes      57     218
dusty     >=10     no       11      92
not       <10      yes      14    1354
not       <10      no       12    1016
not       >=10     yes      24    1384
not       >=10     no       10     996;

/* Fit logistic regression with main effects */
PROC LOGISTIC DATA = Byss2;
    CLASS envir years(REF = first) smoke(REF = first) / PARAM = REF;
    MODEL numComp/numTotal = envir years smoke;

* Access goodness-of-fit;
* The AGGREGATE option specifies the subpopulations on which the Pearson chi-square test
statistic and the likelihood ratio chi-square test statistic (deviance) are calculated;
* The SCALE= option enables you to supply the value of the dispersion parameter or to specify the method for estimating the dispersion parameter;
PROC LOGISTIC DATA = Byss DESCENDING;
    FREQ count;
    CLASS envir years(REF = first) smoke(REF = first) / PARAM = ref;
    MODEL complaint = envir years smoke envir*years envir*smoke / AGGREGATE SCALE=NONE;

PROC LOGISTIC DATA = Byss DESCENDING;
              FREQ count;
              CLASS envir years(REF = first) smoke(REF = first) / PARAM = ref;
              MODEL complaint = envir years smoke / AGGREGATE SCALE=NONE;

* Request the Hosmer-Lemeshow test with the LACKFIT option;
PROC LOGISTIC DATA = Coronary DESCENDING;
              CLASS sex (REF = first) / PARAM = REF;
              MODEL cad = sex ecg age / AGGREGATE SCALE=NONE LACKFIT;

/* Output the classification table and ROC curve */
PROC LOGISTIC DATA = Coronary DESCENDING PLOTS(ONLY) = ROC;
              CLASS sex (REF = first) / PARAM = REF;
              MODEL cad = sex ecg age / CTABLE;

/* Specify a threshold for the classification table and add threshold values to the ROC curve */
PROC LOGISTIC DATA = Coronary DESCENDING PLOTS(ONLY) = ROC(ID = CUTPOINT);
              CLASS sex (REF = first) / PARAM = REF;
              MODEL cad = sex ecg age / CTABLE PPROB = 0.5;

/* Compare the ROC curves for different models */
PROC LOGISTIC DATA = Coronary DESCENDING;
              CLASS sex (REF = first) / PARAM = REF;
              MODEL cad = sex ecg age;
              ROC 'omit sex' age ecg;
              ROC 'omit sex age' ecg;

/* Fit the Poisson regression model with offset */
* TYPE1 tests the significance of variables when variables are added sequentially to the model;
* TYPE3 tests the significance of variables when other variables are already in the model;
PROC GENMOD DATA = Melanoma;
    CLASS region (REF = 'north') age (REF = '<35') / PARAM = ref;
    MODEL cases = age region
        / DIST = POISSON LINK = LOG OFFSET = ltotal TYPE1 TYPE3;

/* equivalent code */
PROC GENMOD DATA = Melanoma;
    CLASS region (REF = 'north') age (REF = '<35') / PARAM = ref;
    MODEL cases / total = age region
        / DIST = POISSON LINK = LOG TYPE1 TYPE3;

/* Not correct */
PROC GENMOD DATA = Melanoma;
    CLASS region (REF = 'north') age (REF = '<35') / PARAM = ref;
    MODEL cases = age region
        / DIST = POISSON LINK = LOG OFFSET = LOG(total) TYPE1 TYPE3;

/* Obtain two example incidence density ratios */
PROC GENMOD DATA = Melanoma;
    CLASS region (REF = 'north') age (REF = '<35') / PARAM = ref;
    MODEL cases = age region / DIST = POISSON LINK = LOG OFFSET = ltotal;
    ESTIMATE "45-54 vs. <35" age 0 1 0 0 0 / EXP;
    ESTIMATE "South vs. North" region 1 / EXP;
```

	Infected	Not Infected	Sum
Placebo	12	3	15
Vaccine	7	8	15
Sum	19	11	30

Fisher Exact Test:

在维持行和和列和不变的情况下，遍历所有可能的表
计算可能出现的表中，比观测到的表相当或更极端的表出现的概率。

P-value=$(C_{15}^{12}C_{15}^{7}+C_{15}^{13}C_{15}^{6}+C_{15}^{14}C_{15}^{5}+C_{15}^{15}C_{15}^{4})/C_{30}^{19}=$ 641/10005

P-value=$(C_{19}^{12}C_{11}^{3}+C_{19}^{13}C_{11}^{2}+C_{19}^{14}C_{11}^{1}+C_{19}^{15}C_{11}^{0})/C_{30}^{15}=$ 641/10005

——§終わり§——