Sample pearson: $\gamma=\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})/(\sum_{i=1}^{n}(X_i-\overline{X})^2\cdot\sum_{i=1}^{n}(y_i-\overline{y})^2)^{\frac{1}{2}}$ Two Ordinal Variables: (not int in assoc, but direction)

estimation: $\rho=E[(X-\mu_X)(Y-\mu_Y)]/\sigma_X\sigma_Y$

Cramér's V: $V=\sqrt{\frac{\chi^2/n}{\min(r-1,c-1)}}$ r: row $\in(0,1)$, high→ stronger tendency; c: col for cat of X ~ cat of Y (particular)

1 Basic Hypothesis Testing

Measures of variability: standard deviation& range:$x_{(n)}-x_{(1)}$ & interquartile range(IQR):Q_3-Q_1, first, second, third quartiles.
Measures of shape: positive/rights kewed, negative/left skewed
skewness=$\frac{E[(X-\mu)^3]}{\sigma^3}$ (sample=$\frac{n}{(n-1)(n-2)}\sum_{i=1}^{n}\frac{(X_i-\overline{X})^3}{S^3}$)
kurtosis=$\frac{E[(X-\mu)^4]}{\sigma^4}-3$, measures the heaviness of tails, compared to a normal distribution.
sample=$\frac{n(n+1)}{(n-1)(n-2)(n-3)}\sum_{i=1}^{n}\frac{(X_i-\overline{X})^4}{S^4}-\frac{3(n-1)^2}{(n-2)(n-3)}$.

3 Basic Hypothesis Testing

3.1 Basic concepts

Type I error(α): reject H_0 when H_0 is actually the truth. Type II error()β): fail to reject H_0 when H_1 is the truth.
α=Pr($\boldsymbol{x}\in RR|H_0$ is true), β=Pr($\boldsymbol{x}\notin RR|H_1$ is true).
The p-value is the prob of obtaining test result at least as extreme as the result actually observed during the test, assuming H_0 is truth.
Due to the randomness of the observed data, p-value is r.v., which $\sim U[0,1]$
Statistic Power: the prob of rejecting H_0 when H_1 is true./ power=1-β
Sampling dist: dist of the point estimate based on samples of a fixed size from a population.
Interpretation of CI: having numerous sample datasets and the 95% CI is computed for each sample dataset, then the fraction of cmputed CI that encompass the true parameter would tend toward 95%.

3.2 Hypothesis Testing for Categorical Variables

One-sample z test: $\hat{p}=\frac{\sum X_i}{n}$, SE of estimate $SE(\hat{p})=\sqrt{p(1-p)/n}$
test statistic: $Z=\frac{\hat{p}-p_0}{\sqrt{p_0(1-p_0)/n}}\sim_{asy}N(0,1)$.
Wald interval: $\hat{p}\pm z_{\alpha/2}\times\sqrt{\hat{p}(1-\hat{p})/n}$
Need for sample size:(CI)$n\hat{p}\geq 10$&$n(1-\hat{p})\geq 10$ (test on p)$np_0\geq 10$& $n(1-p_0)\geq 10$
There are other types of intervals, e.g., the Wilson(or score)interval, the Clopper-Pearson(or exact)interval, etc
Two-sample z test: $H_0:p_1-p_2=0$, we have $\hat{p_1}=\frac{\sum X_{1i}}{n_1}$, $\hat{p_2}=\frac{\sum X_{2i}}{n_2}$
$SE(\hat{p_1}-\hat{p_2})=\sqrt{\frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2}}$, under H_0,
$SE(\hat{p_1}-\hat{p_2})=\sqrt{p(1-p)(\frac{1}{n_1}+\frac{1}{n_2})}$
test statistic: $Z=\frac{(\hat{p}_1-\hat{p}_2)}{\sqrt{p(1-p)(\frac{1}{n_1}+\frac{1}{n_2})}}\sim_{asy}N(0,1)$ where p is estimated by pooled porportion $\hat{p}=\frac{n_1\hat{p}_1+n_2\hat{p}_2}{n_1+n_2}$
Need for sample size:(CI)$n_i\hat{p}_i\geq 10$&$n_i(1-\hat{p}_i)\geq 10$ (test on p)$n_i\hat{p}\geq 10$& $n_i(1-\hat{p})\geq 10$ for i=1,2
Test for contingency table: Pearson's chi-square test can be used to assess: Goodness of fit&Homogeneity&Independence
For Got:$H_0:p_1=p_{01},...,p_m=p_{0m}$. Pearson's chi-square test statistic is $\chi^2=\sum_{i=1}^{m}\frac{(O_i-E_i)^2}{E_i}\sim_{asy}\chi^2_{m-1}$, O_i is the obs freq of the ith category, $E_i=np_{0i}$ is the expected freq under H_0.
For homog&indep:groups=r, category=c, $\chi^2=\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$ $\sim_{asy}\chi^2_{(r-1)(c-1)}$ under hom or indep ass, where E_{ij} is the expected freq of cell (i,j) assuming indep: $E_{ij}=np_{i\cdot}p_{\cdot j}$, when r=c=2, χ^2 equiv two sample z test for binary variables.
Ass for applay this test:(1)obs X_{1i} and X_{2i} are indep samples (2)sample size is enough(cell counts greater or equal to 10)
When sample is small, apply exact tests to compute p-values: 1. Fisher's exact test: with the same margins(same row and col sums) 2.Barnard's exact test: only the row margins are fixed, more powerful than the Fisher's
McNemar's Test for paired samples: $H_0:p_1=p_2$

3.3 Hypothesis testing for Continuous Variable

One-sample t test:(for normal population) $H_0:\mu=\mu_0$, test statistic $T=\frac{\sqrt{n}(\overline{X}-\mu_0)}{S}\sim t_{n-1}$(exact)
The larger the d.f., the more closely the dist approxmiates N(0,1)
By CLT, T asy follows N(0,1), under H_0, the t-test provides an exact test.
Two-sample t test:(for 2 indep normal populations) $H_0:\mu_1=\mu_2$, if assuming that $\sigma_1=\sigma_2=\sigma$, then pooled sample standard deviation
$S_p=\sqrt{\frac{1}{n_1+n_2-2}(\sum_{i=1}^{n_1}(X_{1i}-\overline{X}_1)^2+\sum_{i=1}^{n_2}(X_{2i}-\overline{X}_2)^2)}=\sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}}$, test statistic:$T=\frac{\overline{X}_1-\overline{X}_2}{\sqrt{S_p^2(\frac{1}{n_1}+\frac{1}{n_2})}}\sim t_{n_1+n_2-2}$(exact)
F test, $H_0:\sigma_1^2=\sigma_2^2$, test statistic: $F=\frac{S_1^2}{S_2^2}\sim F(n_1-1,n_2-1)$
when the variance are not equal, $SE(\overline{X}_1-\overline{X}_2)$ is better estimated by $\sqrt{S_1^2/n_1+S_2^2/n_2}$, thus the test statistic is $T_s=\frac{\overline{X}_1-\overline{X}_2}{\sqrt{S_1^2/n_1+S_2^2/n_2}}\sim_{asy}t_v$,
where $v=\frac{(S_1^2/n_1+S_2^2/n_2)^2}{\frac{(S_1^2/n_1)^2}{n_1-1}+\frac{(S_2^2/n_2)^2}{n_2-1}}$ is called Satterthwaite/Welch's t-test.
Nonparametric test for Means/Median: **Sign test:** $H_0:m=m_0$, let N^+ be the number of positive signs obtained upon calculating X_i-m_0 for $i=1,..,n$,under H_0, $N^+\sim Bin(n,p)$ with $p=0.5$, take one-sample z-test.(robust)
Wilcoxon signed-rank test: compute $\{X_i-m_0\}_{i=1}^{n}$→order $\{|X_i-m_0|\}_{i=1}^{n}$ and assign ranks→sums of ranks(positive)= S^+
When sample size n is large(>20), by CLT, under H_0, we have $W=S^+\sim_{asy}N(\frac{n(n+1)}{4},\frac{n(n+1)(2n+1)}{24})$

3.4 Multiple Comparsions

Type I error/alpha inflation. To control the family-wise error rate(FWER), i.e., the prob of incorrectly rejecting at least one H_0.
Bonmferroni:$\tilde{p}_i$=min$\{m\times p_i,1\}$ or $\tilde{\alpha}_i=\alpha/m$ is conservative when m is large or the tests are highly positively correlated.
Holm adjustment: step 1:if $p_{(1)}\leq\alpha/m$, reject $H_{(1)0}$ and continue, else stop$\cdots$ step m:if $p_{(m)}\leq\alpha$, rejcet $H_{(m)0}$. with larger threshold (more powerful):$\tilde{\alpha}_i=\alpha/(m-i+1)$ & adjusted p-value: $\tilde{p}_{(i)}=\{1,max\{(m-i+1)p_{(i)},\tilde{p}_{(i-1)}\}\}$

4 Linear Regression: Model Fitting

4.1 The Multiple Linear Regression

Regression analysis: describe the mean of the distribution of one variable (response) as a function of other variables (explanatory):$E(Y|X)=f(X)$.
Regression Model:$\boldsymbol{y}=\boldsymbol{X\beta}+\boldsymbol{\epsilon}$: $\boldsymbol{X}$, design matrix, $\boldsymbol{\beta}$:vector of parameter.
Least Square Method: assumptions: 1.All explanatory variables X_i are fixed 2.random errors are uncorrelated with $E(\epsilon)=0$&$Var(\epsilon)=\sigma^2$.
Two results: $\frac{\partial\boldsymbol{a}^T\boldsymbol{x}}{\partial\boldsymbol{x}}=\frac{\partial\boldsymbol{x}^T\boldsymbol{a}}{\partial\boldsymbol{x}}=\boldsymbol{a}$ and $\frac{\partial(x^TAx)}{\partial\boldsymbol{x}}=(A+A^T)\boldsymbol{x}$
$SSE(\boldsymbol{\beta})=||\boldsymbol{y}-\boldsymbol{X\beta}||^2$→$\hat{\beta}=(\boldsymbol{X^TX})^{-1}\boldsymbol{X^Ty}$→fitted value(orthogonal projection):$\hat{\boldsymbol{y}}=\boldsymbol{X}(\boldsymbol{X^TX})^{-1}\boldsymbol{X^Ty}=\boldsymbol{Hy}$,$\boldsymbol{H}$:hat matrix(projection matrix)
$E(\hat{\boldsymbol{\beta}})=\boldsymbol{\beta}$,$Var(\hat{\boldsymbol{\beta}})=\sigma^2(\boldsymbol{XX^T})^{-1}$, $\hat{\sigma}^2=\frac{RSS}{n-p-1}$ is a unbiased estimator.
Maximum Likelihood Estimation: Assumptions: 1. All explanatory variables X_i are fixed 2.random errors are i.i.d. $N(0,\sigma^2)$
Under ass, have important results: 1. $\hat{\boldsymbol{\beta}}\sim N(\boldsymbol{\beta},\sigma^2(\boldsymbol{X^TX})^{-1})$ 2. $\frac{n\hat{\sigma}^2}{\sigma^2}=\frac{RSS}{\sigma^2}\sim\chi^2_{n-p-1}$ 3. $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ are independent. need has two cons
If $\boldsymbol{A},\boldsymbol{B}$(scalars matrices) and $\boldsymbol{y}\sim N(\boldsymbol{\mu},\boldsymbol{\Sigma})$, then 1. $\boldsymbol{Ay}\sim N(\boldsymbol{A\mu},\boldsymbol{A\Sigma A^T})$ 2.$\boldsymbol{Ay}$ and $\boldsymbol{By}$ are indep iff $\boldsymbol{A\Sigma B}^T=0$

4.2 Testing the Regression Coefficients

Single:$H_0:\beta_i=0$. Denote (i+1)-th diagonal element of $(\boldsymbol{X^TX})^{-1}$ as c_{ii}, as $\hat{\boldsymbol{\beta}}\sim N(\boldsymbol{\beta},\sigma^2(\boldsymbol{X^TX})^{-1})$, have $\hat{\beta}_i\sim N(\beta_i,c_{ii}\sigma^2)$, test statistic: $T=\frac{\hat{\beta}_i-0}{c_{ii}\hat{\sigma}^2}\sim t_{n-p-1}$ under H_0
Several:$H_0:\beta_{k+1}=\cdots=\beta_p=0$. Def two models, full model: $Y=\beta_0+\beta_1X_1+\cdots+\beta_kX_k+\cdots+\beta_pX_p+\epsilon$ and reduced/restricted model($k<p$): $Y=\beta_0+\beta_1X_1+\cdots+\beta_kX_k+\epsilon$→$RSS_R\geq RSS_F$, test statistic: $F=\frac{(RSS_R-RSS_F)/(p-k)}{RSS_F/(n-p-1)}\sim F(p-k,n-p-1)$ under H_0
Overall significance,$H_0:\beta_1=\cdots=\beta_p=0$ → ANOVA table **SST**:total sum of squares, **SSM**:explained sum of sqaures of the model, **SSE**: residual sum of squares
R Squared/Coefficient of determination: $R^2=SSM/SST$, represents the proportion of variance in the response variable that is explained by the explanatory variables, the remaining can be attributed to unknown variables or inherent variability.
Interactive effects: two exp vars are said to interact if the effect that one of them has on the mean response depends on the value of the other.
Gauss-Markov Theorem: in linear regression model, if $\epsilon_1,\cdots,\epsilon_n$ satisfie:$E(\epsilon_i)=0$ and $Var(\epsilon_i)=\sigma^2<\infty$; $Cov(\epsilon_i,\epsilon_j)=0,\forall i\neq j$, then $\hat{\beta}_{LSE}$has the lowest sampling variance within the class of linear unbiased estimators, termed the BLUE.

5 Linear Regression: Model Selection and Diagnosis

5.1 Model Selection

With less variables: overfitting, simplicity/Interpretation
Model-fitting criterion: $R^2_{adj}=1-\frac{MSE_k}{MST}=1-(1-R^2)\frac{n-1}{n-k-1}$, largest R^2_{adj} is equiv to choose model smallest MSE
Mallows's C_p: $C_p=\frac{SSE_k}{SSE_p/(n-p-1)}-(n-2k-2)$, under full model, k=p, $C_p=k+1=p+1$, can be proven that $E(C_p)=k+1$, choose the model with C_p closest to k+1 and k is small.
$AIC=-2log(\hat{L})+2(k+1)$, $BIC=-2log(\hat{L})+logn(k+1)$, when $n\geq 8$, BIC imposes heavier penalty on k than AIC,
$l(\beta,\sigma^2)=-\frac{n}{2}log\sigma^2-\frac{(y-X^T\beta)^T(y-X^T\beta)}{2\sigma^2}$+c, and $\hat{\sigma}^2=\frac{(y-X^T\beta)^T(y-X^T\beta)}{n}$
AIC=$nlog(\frac{SSE_k}{n})+2(k+1)+c$, BIC=$nlog(\frac{SSE_k}{n})+(k+1)logn+c$
Sequential Selection: Begin with the current model, sequentially add and/or drop one explanatory variable at a time based on whether the resulting model is superior.
forward selection/ backward elimination/ stepwise selection
Shrinkage method:
Ridge, minimize $SSE(\boldsymbol{\beta},\lambda)=\sum_{i=1}^{n}(y_i-\beta_0-\cdots-\beta_px_{ip})^2+\lambda\sum_{j=1}^{p}\beta_j^2$ is equiv $SSE(\boldsymbol{\beta})=\sum_{i=1}^{n}(y_i-\beta_0-\cdots-\beta_px_{ip})^2$ subject $\lambda\sum_{j=1}^{p}\beta_j^2\leq t$
$\hat{\boldsymbol{\beta}}^{ridge}=(\boldsymbol{X^TX}+\lambda\boldsymbol{I_p})^{-1}\boldsymbol{X^Ty}$, addition of $\lambda\boldsymbol{I}_p$ makes nonsingular.
Lasso: Because of the nature of the constraint, making sufficiently small (sufficiently large) will cause some of the coefficients to be exactly zero.
Cross validation: to test the model's ability to predict new data to obtain an insight on how the model will generalize to an unknown dataset.
Leave-one-out CV & k-fold CV: Shuffle the data randomly and split the data into groups of approximately equal size.

5.2 Model Diagnosis

Linearity& Homoscedasticity&Independence&Normality, lin¿hom¿nor
Residual plots:$r_i=y_i-\hat{y}_i$, first check the linear and hom by Fitted(X) versus Residual Plot(Y), i.e., scatterplots of the residuals against the fitted.
Noraml Qunantile-Q plot: check the normality ass, we expect that the points in the Q-Q plot will closely lie on a straight line, or histogram.
Hypothesis: **Hom**: Breusch-Pagan test and White test
BP: auxiliary regression moel: $r_i^2=\gamma_0+\gamma_1z_{i1}+\cdots+\gamma_kz_{ik}+e_i$, $H_0:\gamma_1=\cdots=\gamma_k=0$, using F-statistic. White: All explan vars, all square vars,all intera terms are included. another form:$r_i^2=\gamma_0+\gamma_1\hat{y}_i+\gamma_2\hat{y}_i^2+e_i$
Normality: Shapiro-Wilk test and Kolmogorov-Smirnov test
SW: test statistic: $W=\frac{(\sum_{i=1}^{n}a_ir_{(i)})^2}{\sum_{i=1}^{n}(r_i-\overline{r})^2}$, $0\leq W\leq 1$ and small values of W lead to rejection of normality.
Dist of W under norm has no closed form, only applied when $n\leq$2000.
KS: based on empirical (edf), $F_n(x)=\frac{1}{n}\sum_{i=1}^{n}I_{(r_i\leq x)}$, test statistic: $D=sup_x|F_n(x)-F(X)|$, F is normal cdf, $D\sim_{asy}$Kolmogorov dist under normality, K-S test requires a relatively large($n>2k$) to take proper cdf
Independence: Durbin-Watson Test, we can judge whether it is reasonable to assume independence based on the nature of how the data were collected. for time series data, test statistic: $DW=\frac{\sum_{t=2}^{n}(\hat{\epsilon}_t-\hat{\epsilon}_{t-1})^2}{\sum_{t=1}^{n}\hat{\epsilon}_t^2}$, detect the first order auto-corr(ass $\epsilon_t=\rho\epsilon_{t-1}+u_t$), H_0: $\rho=0$. $1\leq DW\leq 4$, and DW=2 indicates no auto-corr, DW<1 means strong postive auto-corr, DW>3 means strong negative auto-corr.

5.3 Unusual Observation

6 Analysis of Variance

Source	DF	Sum of Squares	Mean Squares	F Value	Pr > F
Model	p	SSM	$MSM=\frac{SSM}{p}$	$F=\frac{MSM}{MSE}$	
Error	$n-p-1$	SSE	$MSE=\frac{SSE}{n-p-1}$		
Total	$n-1$	SST			

6.1.1 Definition and F-test

ANOVA is used to analyze the differences among group means in a sample

The model is $Y_{ij}\sim N(\mu_i,\sigma^2), i=1,2,\cdots,k$.

means model: $Y_{ij}=\mu_i+\epsilon_{ij}$ and **effect model**: $Y_{ij}=\mu+\alpha_i+\epsilon_{ij}$

where $\epsilon_{ij}\sim_{iid}N(0,\sigma^2)$ is random error, $\alpha_i=\mu_i-\mu$: main effect of group i.

Hypothesis: $H_0:\mu_1=\mu_2=\cdots=\mu_k$ vs. H_1 : not all μ_i are equal.

Assumptions: Normality & Homoscedasticity & Independence

SSB=$\sum_{i=1}^k\sum_{j=1}^{n_i}(\overline{Y}_i-\overline{Y})^2=\sum_{i=1}^k n_i(\overline{Y}_i-\overline{Y})^2$, variation between groups

SSW=$\sum_{i=1}^k\sum_{j=1}^{n_i}(\overline{Y}_{ij}-\overline{Y}_i)^2=\sum_{i=1}^k(n_i-1)S_i^2$, variation within groups

Under assumptions of independence and equal variance, we have $E(SSB)=(k-1)\sigma^2+\sum_{i=1}^k n_i(\mu_i-\mu)^2$ and $E(SSW)=(n-k)\sigma^2$

The test statisitc: $F=\frac{SSB/(k-1)}{SSW/(n-k)}$

SSB$\perp$SSW under H_0, and SSB+SSW=SST=$\sum_{i=1}^k\sum_{j=1}^{n_i}(Y_{ij}-\overline{Y})^2$

With normality & under H_0: $\frac{SST}{\sigma^2}\sim\chi^2_{n-1}$, $\frac{SSB}{\sigma^2}\sim\chi^2_{k-1}$, $\frac{SSW}{\sigma^2}\sim\chi^2_{n-k}$

Therefore, under H_0 : $F=\frac{SSB/(k-1)}{SSW/(n-k)}\sim F(k-1,n-k)$ (one-side test)

Source	df	SS	MS	F Value
Between	$k-1$	SSB	MSB	$F=\frac{\text{MSB}}{\text{MSW}}$
Within	$n-k$	SSW	MSW	
Total	$n-1$	SST		

6.1.2 Testing Equality of Group Variance(homoscedasticity)

Hypothesis: $H_0:\sigma_1^2=\sigma_2^2=\cdots=\sigma_k^2$ vs. H_1 : Not all σ_i^2 are equal.

Bartlett's: For $a_1,a_2,\cdots,a_k>0$, the weighted arithmetic mean and geometric mean are $\overline{a}^A=\sum_{i=1}^k w_ia_i$, $\overline{a}^G=\Pi_{i=1}^k a_i^{w_i}$.

For $\sum_{i=1}^k w_i=1$, $\overline{a}^A>\overline{a}^G$ and attain equality iff $a_1,a_2,\cdots,a_k$ are all equal. Let $w_i=\frac{n_i-1}{n-k}$, then $\overline{S}_A^2=\sum_{i=1}^k w_iS_i^2$, $\overline{S}_G^2=\Pi_{i=1}^k(S_i^2)^{w_i}$.

Test statistic: $B=\frac{(n-k)(log\overline{S}_A^2-log\overline{S}_G^2)}{1+\frac{1}{3(k-1)}[(\sum_{i=1}^k\frac{1}{n_i-1})-\frac{1}{n-k}]}\sim_{approx}\chi^2_{k-1}$ under H_0

(approx need normality&large sample)

Also $\frac{(n-k)InS^2-\sum_{i=1}^k(n_i-1)S_i^2}{1+\frac{1}{3(k-1)}[(\sum_{i=1}^k\frac{1}{n_i-1})-\frac{1}{n-k}]}$ where $S^2=\sum_{i=1}^k(n_i-1)S_i^2/(n-k)$

Levene's and Brown-Forsythe test: Transform the original values of Y_{ij} to dispersion variable Z_{ij}, perform ANOVA on Z_{ij}.

Levene's: $Z_{ij}=(Y_{ij}-\overline{Y}_i)^2$ or $|Y_{ij}-\overline{Y}_i|$ (these two tests are robust)

BF use $Z_{ij}=|Y_{ij}-m_i|$, m_i is sample median of group i.

6.1.3 Kruskal-Wallis Test (Nonparameter test)

When normality is violated, nonparametric alternative to one-way ANOVA, relpacing y_{ij} by rank$\to$test the equality of population group median.

Rank all Y_{ij}'s from all groups together, denoted R_{ij}, $\overline{R}_i=\sum_{j=1}^{n_i}R_{ij}/n$ and $\overline{R}=(n+1)/2$

Test statistic: $KW=\frac{(n-1)\sum_{i=1}^k n_i(\overline{R}_i-\overline{R})^2}{\sum_{i=1}^k\sum_{j=1}^{n_i}(R_{ij}-\overline{R})^2}\sim_{approx.}\chi^2_{k-1}$ under H_0.

H_0 can be rejected if $KW>\chi^2_{\alpha,k-1}$.

Multiple Comparsions: if the CI of $\mu_i-\mu_j$ contains 0, μ_i and μ_j are not significantly different. where Tukry-Kramer for comparsions bet all pairs of means and Dunnett for comparsions bet a control and all other means.

6.2 Two-way ANOVA

Models: means: $Y_{ijk}=\mu_{ij}+\epsilon_{ijk}$ & effects: $Y_{ijk}=\mu+\alpha_i+\beta_i+\gamma_{ij}+\epsilon_{ijk}$

α_i,β_i are the main effect of factor, γ_{ij} is the interaction effect bet A and B. Interaction effect means the effect of one factor depends on the level of the other factor.

For inreraction model: $\alpha_i=\mu_{i.}-\mu$, $\beta_j=\mu_{.j}-\mu$

then $\gamma_{ij}=\mu_{ij}-(\mu+\alpha_i+\beta_j)=\mu_{ij}-\mu_{i.}-\mu_{.j}+\mu$

The SSE to be minimized is $SSE=\sum_{i=1}^a\sum_{b=1}^b\sum_{k=1}^{n_{ij}}(Y_{ijk}-\mu_{ij})^2$

LSE estimator: $\hat{\mu_{ij}}=\overline{Y}_{ij}$, $\hat{\mu}=\overline{Y}$, $\hat{\alpha_i}=\overline{Y}_{i\cdot}-\overline{Y}$, $\hat{\beta_i}=\overline{Y}_{\cdot j}-\overline{Y}$, and $\hat{\gamma}_{ij}=\overline{Y}_{ij}-\overline{Y}_{i\cdot}-\overline{Y}_{\cdot j}+\overline{Y}$

Test interaction effect: $H_0^{AB}:\gamma_{ij}=0$ for $i=1,\cdots,a, j=1,\cdots,b$

If H_0^{AB} is not reject, then test the main effect of each factor

$SSM=\sum_{i=1}^a\sum_{j=1}^b n_{ij}(\overline{Y}_{ij}-\overline{Y})^2$, $SSW=\sum_{i=1}^a\sum_{j=1}^b\sum_{k=1}^{n_{ij}}(Y_{ijk}-\overline{Y}_{ij})^2$

$SST=\sum_{i=1}^a\sum_{j=1}^b\sum_{k=1}^{n_{ij}}(Y_{ijk}-\overline{Y})^2$ and SST=SSM+SSE

F-test statistic: $F=\frac{SSM/(ab-1)}{SSE/(n-ab)}$ to H_0: interaction model vs. H_1: null model

furtherly decompose the variation between groups to difference sources.

Consider $\overline{Y}_{ij}-\overline{Y}=(\overline{Y}_{i\cdot}-\overline{Y})+(\overline{Y}_{\cdot j}-\overline{Y})+(\overline{Y}_{ij}-\overline{Y}_{i\cdot}-\overline{Y}_{\cdot j}+\overline{Y})$

Define $SSA=\sum_{i=1}^a n_{i\cdot}(\overline{Y}_{i\cdot}-\overline{Y})^2$ (variation between groups due to factor A), define $SSB=\sum_{j=1}^b n_{\cdot j}(\overline{Y}_{\cdot j}-\overline{Y})^2$, define

$SSAB=\sum_{i=1}^a\sum_{j=1}^b n_{ij}(\overline{Y}_{ij}-\overline{Y}_{i\cdot}-\overline{Y}_{\cdot j}+\overline{Y})^2$ (variation between groups due to the interaction of factor A and B)

Under ass of indep&homoscedasticity, $E(SSA)=(a-1)\sigma^2+\sum_{i=1}^a n_{i\cdot}\alpha_i^2$,

$E(SSB)=(b-1)\sigma^2+\sum_{i=1}^a n_{\cdot j}\beta_j^2$,

$E(SSAB)=(a-1)(b-1)\sigma^2+\sum_{i=1}^a\sum_{j=1}^b n_{ij}\gamma_{ij}^2$

F-test: H_0^{AB}: All $\gamma_{ij}=0$, H_0^A: All $\alpha_i=0$, H_0^B: All $\beta_j=0$

Source	df	SS	MS	F Value
A	$a-1$	SSA	MSA	$F^A=\frac{\text{MSA}}{\text{MSE}}$
B	$b-1$	SSB	MSB	$F^B=\frac{\text{MSB}}{\text{MSE}}$
A*B	$(a-1)(b-1)$	SSAB	MSAB	$F^{AB}=\frac{\text{MSAB}}{\text{MSE}}$
Error	$n-ab$	SSE	MSE	
Total	$n-1$	SST		

However, SSM=SSA+SSB+SSAB is only true when all n_{ij} are equal.

6.2.3 Type I and Type III SS

We can consider the SS for a given source to be the extra variability explained when the respective term is added to the model, i.e., reduction in when the term is added.

SSA=SSE(null)-SSE(A), SSB=SSE(A,B)-SSE(A,B,AB), SSAB=SSE(A,B)-SSE(A,B,AB)

The order in which terms are entered into the model matters.

Difference: In Type I, effects are added sequentially. In Type III, assumed that all the effects are already in the model other than the effect of interest.

SSA=SSE(B,AB)-SSE(A,B,AB), SSAB=SSE(A,B)-SSE(A,B,AB)

In a balanced design, Type I and Type III SS are the same, because each effect provides unique information and doesn't take away from what another effect explains.

The order in which terms are entered into the model does not change the Type III SS, not satisfy SSM=SSA+SSB+SSAB

7 Generalized Linear Models

7.1 Exponential Family and Generalized Linear Models

A pmf/pdf belongs to exponential family of distributions if it is of the form: $f(\boldsymbol{y};\boldsymbol{\theta})=h(\boldsymbol{y})exp\{\boldsymbol{\eta}(\boldsymbol{\theta})\cdot\boldsymbol{T}(\boldsymbol{y})-A(\boldsymbol{\theta})\}$

If $\boldsymbol{\eta}(\boldsymbol{\theta})=\boldsymbol{\theta}$, called canonical form(where θ called canonical parameter)

T($\boldsymbol{y}$) is the sufficient statistic of the natural parameter $\boldsymbol{\theta}$

If $T(\boldsymbol{y})=\boldsymbol{y}$, then the family is called a natural exponential family.

Binomial, poisson, exponential, normal belongs to exponential family.

Property of exp_fam: for canonical form: $f(\boldsymbol{y};\boldsymbol{\theta})=h(\boldsymbol{y})exp\{\boldsymbol{\theta}\cdot\boldsymbol{T}(\boldsymbol{y})-A(\boldsymbol{\theta})\}$, (1) mgf of $T(\boldsymbol{Y})$ is $M_T(\boldsymbol{t})=exp\{A(\boldsymbol{t}+\boldsymbol{\theta})-A(\boldsymbol{\theta})\}$,(2)$E(T(\boldsymbol{Y}))=A'(\theta)$, $Var[T(\boldsymbol{Y})]=A''(\theta)$

if consider the ϕ,we have $E(T(\boldsymbol{Y}))=A'(\theta)$, $Var[T(\boldsymbol{Y})]=\phi A''(\theta)$

link function g s.t. $g(\mu_i)=\beta_0+\beta_1x_{i1}+\cdots+\beta_px_{ip}=\boldsymbol{x}_i^T\boldsymbol{\beta}=\eta_i$ linear predictor.

canconical link: $g(\cdot)=(A')^{-1}(\cdot)$ s.t. $g(E(Y_i))=(A')^{-1}(A'(\theta_i))=\theta=\eta_i$

parameter estimation: $l(\boldsymbol{\beta})=\sum_{i=1}^n[\theta_iy_i-A(\theta_i)]$

if use canonical link: $\frac{\partial l}{\partial\boldsymbol{\beta}}=\sum_{i=1}^n\boldsymbol{x_i}(y_i-\mu_i)=\sum_{i=1}^n\boldsymbol{x_i}(y_i-g^{-1}(\boldsymbol{x_i^T\beta}))$

general link: $\frac{\partial l}{\partial\boldsymbol{\beta}}=\sum_{i=1}^n\frac{\boldsymbol{x}_i(y_i-\mu_i)}{Var(Y_i)g'(\mu_i)}$ & score function $\boldsymbol{U}(\boldsymbol{\beta})=\frac{\partial l}{\partial\boldsymbol{\beta}}$

iterative alg: $\boldsymbol{\beta}^{(m+1)}=\boldsymbol{\beta}^{(m)}+[\boldsymbol{J}(\boldsymbol{\beta^{(m)}})]^{-1}\boldsymbol{U}(\boldsymbol{\beta^{(m)}})$ where $\boldsymbol{J}(\boldsymbol{B})$ is usually replaced by $\boldsymbol{I}(\boldsymbol{\beta})=E[\boldsymbol{J}(\boldsymbol{\beta})]$ Fisher inofrmation matrix

Another Iteratively Reweighted Least Squares(IRLS):

$\boldsymbol{\beta}^{(m+1)}=(\boldsymbol{X^TW^{(m)}X})^{-1}\boldsymbol{X^TW^{(m)}z^{(m)}}$ with $\boldsymbol{W}^{(m)}=diag\{w_i^{(m)}\}$

where $z_i^{(m)}=\eta_i^{(m)}+(y_i-\mu_i^{(m)})g'(\mu_i^{(m)})$(working response) and $w_i^{(m)}=\frac{1}{Var(Y_i|\beta^{(m)})[g'(\mu_i^{(m)})]^2}$(working weight matrix)

CI: score statistic: $\boldsymbol{U}=\frac{\partial l}{\partial\boldsymbol{\beta}}=\sum_{i=1}^n\frac{\boldsymbol{x}_i(y_i-\mu_i)}{Var(Y_i)g'(\mu_i)}$ we have $E(\boldsymbol{U})=0$, variance matrix of $\boldsymbol{U}$ is $\boldsymbol{V}=E(\boldsymbol{UU^T})$with (j,k) element:

$v_{jk}=\sum_{i=1}^n\frac{x_{ij}x_{ik}}{Var(Y_i)[g'(\mu_i)]^2}$ rewrite as: $\boldsymbol{V}=\boldsymbol{X^TWX}$, where $\boldsymbol{W}$ is an n by n matrix with elements $w_{ii}=\frac{1}{Var(Y_i)[g'(\mu_i)^2]}$.

$\boldsymbol{U}\sim N(\boldsymbol{0},\boldsymbol{V})$ (asymptotically), $\hat{\boldsymbol{\beta}}\sim N(\boldsymbol{\beta},\boldsymbol{V^{-1}})$(asymptotically)

Goodness of fit Deviance: $D(M)=2(l(\hat{\boldsymbol{\beta}}_S)-l(\hat{\boldsymbol{\beta}}_M))$ where $g^{-1}(\hat{\beta}_i^S)=\hat{\mu}_i^S=y_i$, note that $D^*=\frac{D}{\phi}\sim\chi^2_{n-p-1}$ (asy), deviance residual: $r_{Di}=sign(y_i-\hat{\mu}_i)\sqrt{d_i}$, d_i is the contribution of the ith obs to the deviance

Generalized Pearson's Chi-Square statistic; $\chi^2=\sum_{i=1}^n\frac{(y_i-\hat{\mu}_i)^2}{V(\hat{\mu}_i)}$,

$r_{Pi}=\frac{(y_i-\hat{\mu}_i)}{\sqrt{V(\hat{\mu}_i)}}$, scaled version $\chi^{2*}=\frac{\chi^2}{\phi}\sim_{asy}\chi^2_{n-p-1}$

Logistic regression: why logit link: 1.canonical link 2.to ∞ 3.good interpretation. odds=$\frac{p}{1-p}$, An event with odd>1 is more likely to happen than not happen. OR=$\frac{odds_1}{odds_2}$, OR>1 indicates that the event is more likely to happen in the first population.

When X_j is a continuous explanatory variable, with all other x_l's fixed, if x_j increases by 1 unit, the odds of Y=1 changes by a multiplicative factor of exp(β_j).

A worker under...has 14.42 times the odds of...compared to

Outliers: Leverage, Outlier, Influence:

Leverage: $\underset{\sim}{H}=\underset{\sim}{X}(\underset{\sim}{X}^T\underset{\sim}{X})^{-1}\underset{\sim}{X}^T$. h_{ii}: leverages.

centered version: $\underset{\sim}{Z}=\begin{pmatrix}(x_1-\bar{x})^T\\ \vdots\\ (x_n-\bar{x})^T\end{pmatrix}$

→ can show $h_{ii}=\frac{1}{n}+(x_i-\bar{x})^T(Z^TZ)^{-1}(x_i-\bar{x})$

$h_{ii}\uparrow$. more far away from $\bar{x}$.

Since $\sum h_{ii}=p+1$. → $h_{ii}>2\bar{h}=\frac{2(p+1)}{n}$ (high)

Outlier: $\underset{\sim}{r}=\underset{\sim}{y}-\hat{\underset{\sim}{y}}=(\underset{\sim}{I}-\underset{\sim}{H})\underset{\sim}{y}$. under linear

→ $Var(r_i)=(1-h_{ii})\hat{\sigma}^2$. → studentized res:

$r_i^{stu}=\frac{y_i-\hat{y}_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$

$\begin{cases}\hat{\sigma}^2=RSS/n-p-1=MSE\\ \hat{\sigma}_{(i)}^2=RSS_{(i)}/n-p-2=MSE_{(i)}\end{cases}$ externally stu res $\sim t_{n-p-2}$ (well)

$\begin{cases}\text{if } |r_i^{stu}|<2.\ \text{not outlier}\\ 2\le|r_i^{stu}|\le3.\ \text{possible}\\ |r_i^{stu}|>3.\ \text{is outlier}\end{cases}$

influence: Cook's distance.

$D_i=\frac{(\hat{\beta}-\hat{\beta}_{(i)})^T\underset{\sim}{X}^T\underset{\sim}{X}(\hat{\beta}-\hat{\beta}_{(i)})}{(p+1)\hat{\sigma}^2}=\frac{(y_i-\hat{y}_i)^2}{(p+1)MSE}\cdot\frac{h_{ii}}{(1-h_{ii})^2}=\frac{1}{p+1}(r_i^{stu})^2\frac{h_{ii}}{1-h_{ii}}$

large stud res / higher lever $\xrightarrow{may}$ high D_i

$\begin{cases}\text{if } D_i<F(0.2,p+1,n-p-1)\ \text{not influential}\\ \text{if } F(0.2,p+1,n-p-1)\le D_i<F(0.5,p+1,n-p-1).\ \text{possible}\\ \text{if } D_i>F(0.5,p+1,n-p-1)\ \text{is influe}\end{cases}$

Simple: $D_i>4/n$ (empirical)

Multicollinearity:

tolerance(TOL). variance inflation factor(VIF):

$TOL_j=1-R_j^2$, & $VIF=\frac{1}{TOL_j}=\frac{1}{1-R_j^2}$

TOL_j less than 0.1 (or ⇔ $VIF_j\ge10$) may be consider small enough to suspect serious multicollinearity.

Spearman's rank correlation coefficient & Kendall's Tau

- Spearman's rank correlation coefficient is:

$$r_S=\frac{\sum_{i=1}^n[R(X_i)-\overline{R(X)}][R(Y_i)-\overline{R(Y)}]}{\sqrt{\sum_{i=1}^n[R(X_i)-\overline{R(X)}]^2\sum_{i=1}^n[R(Y_i)-\overline{R(Y)}]^2}},$$

- $R(\cdot)$ denotes the rank operation;
- the Pearson correlation coefficient applied to the rank variables;
- assesses monotonic relationships;
- ranges from -1 to 1;
- suitable for continuous variables and is hardly affected by outliers.
- $(X_i,Y_i),(X_j,Y_j)$ is called a concordant pair if: $(X_i>X_j\ \&\ Y_i>Y_j)$ or $(X_i<X_j\ \&\ Y_i<Y_j)$
- $(X_i,Y_i),(X_j,Y_j)$ is called a discordant pair if: $(X_i>X_j\ \&\ Y_i<Y_j)$ or $(X_i<X_j\ \&\ Y_i>Y_j)$
- Kendall's Tau is:

$$\tau=\frac{(\#\text{ of concordant pairs})-(\#\text{ of discordant pairs})}{\text{total number of pairs}}$$

- ranges from -1 to 1;

```
* generate two two-way frequency tables with different options;
PROC FREQ DATA = Grade;
    TABLES Gender * Section;
    TABLES Gender * Section / MISSING NOPERCENT NOCOL NOROW;
    TITLE 'Student Gender by Class Section';
RUN;
```

```
PROC GLM DATA = sim2 PLOTS = DIAGNOSTICS(UNPACK);
    MODEL loss = effort hours;
    OUTPUT OUT = sim2_fitted (Keep = loss effort hours) Residual = r Predicted = fv;
RUN;
```

```
PROC GLM DATA = Work;
    CLASS X1 X2;
    MODEL Y = X1 X2 X1*X2;
    OUTPUT OUT = DataRegOut RESIDUAL = resid;
RUN;
```

```
PROC MEANS DATA=Grade MAXDEC=3;
    VAR Score;
    CLASS Status Year;
    TYPES () Status*Year;
    TITLE 'Final Exam Grades fo
RUN;
```

逻辑回归 Y~Binomial(1,p) $g(p)=\log(\frac{p}{1-p})$ $f(y;p)=p^y(1-p)^{1-y}$

Odds $=\frac{p}{1-p}=\frac{Pr(Y=1)}{Pr(Y=0)}$ Odds Ratio $\underline{OR}=\frac{odds_1}{odds_2}=\frac{p_1(1-p_2)}{p_2(1-p_1)}$ OR>1,更可能发生在人群1 在p_1,p_2小的时比p_1-p_2准

X_j 连续 其它 x_i's fixed, x_j 增加1, odds of Y=1 变化 $\exp(\beta_j)$

X_j dummy 代表 1 level(吸烟), $\exp(\beta_j)$ 是吸烟和没吸烟的人肺癌的odds ratio

```
PROC LOGISTIC DATA = Byss; DESCENDING
    FREQ count;
    CLASS envir years(REF = first) smoke(REF = first) / PARAM = REF;
    MODEL complaint(EVENT = last) = envir|years|smoke@2;
```

加入 ODDSRATO envir 可看优比表 (-1和1) 默认为Effect 用于算P 列所有类别变量 为0 为0(NO,YES) 为1 无交互: envir years smoke

三种输入数据方法: Count NumComp/NumTotal 都输入(条条)

拟合优度L $n_i\ge10$ $n_i\hat{p}_i\ge5$ χ^2_{n-k-1} $D=2\sum_{i=1}^n\left[y_i\log\left(\frac{y_i}{n_i\hat{p}_i}\right)+(n_i-y_i)\log\left(\frac{n_i-y_i}{n_i(1-\hat{p}_i)}\right)\right]$ $\chi^2=\sum_{i=1}^n\frac{(y_i-n_i\hat{p}_i)^2}{n_i\hat{p}_i(1-\hat{p}_i)}$ 两个差得大, n_i不足

对于Raw binary data ①分成G组 ② $\sum_{i=1}^G(O_i-E_i)^2/E_i\sim\chi^2_{G-2}$

在MODEL的最后加入: /AGGREGATE SCALE=NONE LACKFIT (引入HL Test)

预测准度L Somer's D (Gini) $=\frac{C-D}{C+D+T}$, Gamma $=\frac{C-D}{C+D}$, Tau − a $=\frac{C-D}{N}$, $c=\frac{C+0.5T}{C+D+T}$ concordence

Ob\Pre	1	0
1	TP	FN
0	FP	TN

Sensitivity $TPR=\frac{TP}{TP+FN}$ 不能同时大 Specific $TNR=\frac{TN}{TN+FP}$ Precis $PPV=\frac{TP}{TP+FP}$ AUC $NPV=\frac{TN}{TN+FN}$

比较ROC curve和AUC值 画ROC

```
PROC LOGISTIC DATA = Coronary DESCENDING PLOTS(ONLY) = ROC(ID = CUTPOINT);
```

在Model后加 CTABLE PPROB=0.5对比

obs \ Pre	1	0
1	FP	FN
0		TN

泊松回归 时空固定,事件独立,rate为常数 $Pr(Y=y)=\frac{\mu^ye^{-\mu}}{y!}$ $E(Y)=Var(Y)=\mu$

$g(\mu_i)=\log(\mu_i)$ 用IRLS估β, 用Wald test和Likelihood ratio test看变量显著, 用deviance和Pearson's chi-square检验拟合优度

Unit time/space: $\log(\mu_i/T_i)=x_i^T\beta\longrightarrow\log(\mu_i)=x_i^T\beta+\log(T_i)$ offset

```
PROC GENMOD DATA = Melanoma;
    CLASS region (REF = 'north') age (REF = '<35') / PARAM = ref;
    MODEL cases = age region / DIST = POISSON LINK = LOG OFFSET = ltotal;
```

BINOMIAL LOGIT

OverdispersionP 均值方差应相等,可能因不独立,rate不为常数造成 ①β估计值无影响 ②φ ③用Chi和Deviance估计 ④Rich model和异常点,

用Negative Binomial来代替Poisson $f(y;r,p)=Pr(Y=y)=\binom{y+r-1}{y}(1-p)^rp^y$

$Var(Y)=\mu+\frac{\mu^2}{r}=(1+\frac{\mu}{r})\mu>\mu$ $\frac{1}{r}$是dispersion parameter DIST=NEGBIN

- The AUC is the same with the concordance index.
- AUC can be interpreted as the probability that a randomly chosen positive sample has a higher predicted value than a randomly chosen negative one.

obs \ Pre	1	0
1	TP	FN
0	FP	TH