CI (mean, for all): $\hat{Y} \pm t_{n-2} \times S_{Y|X}\sqrt{\frac{1}{n}+\frac{(X_0-\bar{X})^2}{(n-1)S_X^2}}$ // PI (single): $\hat{Y} \pm t_{n-2} S_{Y|X}\sqrt{1+\frac{1}{n}+\frac{(X_0-\bar{X})^2}{(n-1)S_X^2}}$

X值离样本平均更远, interval更宽.

X值相同时, PI比CI更宽.

ANOVA TABLE

Source	DF	Sum of Squares	Mean Square	F statistic
Regression	k	SSR	$MSR=\frac{SSR}{k}$	$F=\frac{MSR}{MSE}$
Error	n-k-1	SSE	$MSE=\frac{SSE}{n-k-1}$	
Total	n-1	SSY		

Lec7 多元 (multiple) $\hat{Y}=\beta_0+\beta_1X_1+\beta_2X_2$ (有数值) number of predictions

$S_{Y|X}=\sqrt{MSE}$ model $Y=\beta_0+\beta_1X_1+\cdots+\varepsilon$.

解释 coefficient: 1. 0 & 0 --would have.

1. For every additional A, the predicted of总 will ↗或↘ by ⟹, holding B constant.

Residual (真实 - 预测($\hat{Y}$))

Reduced Model: $Y=\beta_0+\varepsilon$

? 选择最合适的模型

R^2 对比full model有帮助.

Residual std D. (S) 不好.

$H_0: \beta_1=\beta_2=\beta_3=0$ vs H_a: at least $\beta_i \neq 0$.

解释: tend to deviate from predicted values by 0 in each direction. CV. $F_{r,E,0.05}=$.

Lec8

$SS(X_1,X_2,X_3)=SS(X_1)+SS(X_2|X_1)+SS(X_3|X_1,X_2)$.

$SS(X_2|X_1)\neq SS(X_2)$

$SS(X_1,X_2)=SS(X_1)+SS(X_2|X_1)=SS(X_2)+SS(X_1|X_2)$

$SS(X_3,X_4|X_1,X_2)=SS(X_1,X_2,X_3,X_4)-SS(X_1,X_2)$

对于SSR(A,B), AB以何为不重要

对于SSR(A|B), A B 重要

SSR(A)没有变动为等个总的situation.

$SS(X_1|X_2)=SS(X_1,X_2)-SS(X_2)$

Model Reduced 那个 model 哪些是不是0.

加大p值不会降低SSR why

新加入元素 compete with former.

most variability has been explained by former.

Partial R^2 $R^2_{Y,X_1^*,\cdots X_q^*|X_1,\cdots X_p}=\frac{SSR(X_1^*\cdots X_q^*|X_1\cdots X_p)}{SSE(X_1,\cdots X_p)}$

1个q可以解释为 of variability that $X_1\cdots X_p$ 不能解释

有了β_1,β_2,β_3 6种setting, 能说明什么.

enter order不改变 SSR total

q & n-p-q-1, intercept

tested reduced

1° $\beta_1=\beta_2=\beta_3=0$

2° $\beta_1=0$.

3° $\beta_2=\beta_3=0$, X_1在model

4° $\beta_3=0$, X_1,X_2在 model

5° $\beta_2=0$, X_1在model.

Overall ($X_1,\cdots X_k$ is better than empty). $\begin{cases} Y=\beta_0+\beta_1X_1+\cdots\beta_kX_k+\varepsilon \\ Y=\beta_0+\varepsilon \end{cases}$ $F=\frac{SSR/k}{SSE/(n-k-1)}$ k & n-k-1 (df)

都为0

Partial (X_i是否对已有$X_1\cdots X_p$的model有为0.

$\begin{cases} Y=\cdots\beta_pX_p+\beta_iX_i+\varepsilon \\ Y=\beta_0+\beta_1X_1+\cdots\beta_pX_p+\varepsilon \end{cases}$ $F=\frac{SS(X_i|X_1\cdots X_p)}{MSE(X_1,\cdots,X_p,X_i)}$ 1 & n-p-1-1 df

有一个不为0?

Multi Pa. (X_i,X_j 是否对已有..)

$F=\frac{SS(X_i,X_j|X_1,X_2;X_p)/ij\cdots}{MSE(X_i,\cdots X_p,X_i,X_j\cdots)}$ $\begin{cases} Y=\cdots\beta_pX_p+\beta_iX_i+\beta_jX_j+\varepsilon \\ Y=\sim \end{cases}$

ij ... 对比个

n-p-ij-1 df

Lec3 X-predictor Y-response

Sum of Squares: $SSX = \sum_{i=1}^{n}(X_i - \bar{X})^2$ $S_X = \sqrt{\frac{SSX}{n-1}}$ $S_{XY} = \frac{\sum(X_i-\bar{X})(Y_i-\bar{Y})}{n-1} = \frac{SSXY}{n-1}$

↓ covariance

correlation coefficient: $r = \frac{SSXY}{\sqrt{SSX \cdot SSY}} = \frac{S_{XY}}{S_X S_Y}$ (-1, 1) 0.4-0.7: moderate ; 0.7-1: strong

Simple $S_{Y|X} = \sqrt{\frac{SSE}{n-1-(k-1)}}$ (k=1)

Simple Linear Regression Line

$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$

Intercept (-) 用却取值 slope: $\frac{S_Y}{S_X} r = \frac{\sum(X_i-\bar{X})(Y_i-\bar{Y})}{\sum(X_i-\bar{X})^2}$

Intercept: 当 0 …… is predicted to do … Intercept.

Slope: For every additional …, the predicted … of …… ↗/↘ by slope.

Residual (误差): 合 = 实际值 - 预测值 (>0 则 above line) RSE (残差标准差) $= \sqrt{\frac{\sum(Y-\hat{Y})^2}{n-2}}$ → SSE

$S_{Y|X}$.

Observed values tend to deviate from predicted values by … $S_{Y|X}$

Lec4:

R-Squared: $\frac{SSY-SSE}{SSY} = r^2 = 1 - \frac{SSE}{SSY} = \frac{\text{Regression}}{\text{Total}}$ $SSY = SSR + SSE$ (Regression, Error)

coefficient of determination: % of the variability in the …. is explained by the ….

Certain? X. Linear model: $Y = \beta_0 + \beta_1 X + E$ (不能有参数次).

$SSR = \sum(\hat{Y}_i - \bar{Y})^2$

$SSE = \sum(Y_i - \hat{Y}_i)^2$

$SSY = \sum(Y_i - \bar{Y})^2$

Independence: 一个变量的 x 是否影响另一个变量的 y.

Linearity: Sample means 是否 approximately a straight line.

Normality: Histogram of residuals 是否 approximately normal. { 是否 skewed.

Homoscedasticity: Residuals 是否 have same spread for each predicted value.

Impact of unusual obs:

Lec5:

significant predictor: $H_0: \beta_1 = 0$ vs. $H_a: \beta_1 \neq 0$. Slope

Test static: $t = \frac{\hat{\beta}_1 - \beta_1}{SE}$ > CV: $t_{df, .025} = \pm$ …, p小, 拒绝原假设

whether > xx. correlation = $\sqrt{R\text{-squ}}$

Test: $t = \frac{\beta_1 - xx}{SE}$ < CV: $t_{df, 0.95}$ X reject

xx is a plausible factor.

是否有 intercept: $H_0: \beta_0 = 0$ vs $H_a: \beta_0 \neq 0$ $t = \frac{\hat{\beta}_0 - \beta_0}{S_{Y|X}\sqrt{\frac{1}{n} + \frac{\bar{X}^2}{(n-1)S_X^2}}}$ (Intercept)

// base xx, β_0 ? xx, 同上是

95% CI: $\hat{\beta}_1 \pm t_{n-2} \times SE$ ($SE = \frac{S_{Y|X}}{S_X\sqrt{n-1}}$)

95% confiden that … for … the predicted y …

测试结果是否符合 CI?

Reject H_0 Not Contained

Fail to … Contained

Lec6:

Test 0 correlation $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$

p小, 拒绝, 是着差异

outlier 对 regression 的影响:

make the model fit worse

increase SSE => increase residual standard deviation => derrease R square.

nonzero ($H_0: \rho = 0$ vs $\neq 0$)

Fisher's Z.

$Z = \frac{\frac{1}{2}\ln(\frac{1+r}{1-r}) - \frac{1}{2}\ln(\frac{1+\rho_0}{1-\rho_0})}{1/\sqrt{n-3}}$

对 xx r

// $\frac{\frac{1}{2}\ln(\frac{1+r_1}{1-r_1}) - \frac{1}{2}\ln(\frac{1+r_2}{1-r_2})}{\sqrt{\frac{1}{n_1-3} + \frac{1}{n_2-3}}}$

P小，拒绝，差异显著。Test值 > CV // $H_0: \beta_1 = ?$ vs $H_A: \beta_1 \neq$

L10

3个 dummy variable

$\begin{pmatrix} 0 & 1 \\ 1 & 0 \\ 0 & 0 \end{pmatrix}$ → 没在表中出现的

Equation: $\hat{Y}$ = 基础(前项) + 对应 DV 改变的 slope 和 interrept

Model: $Y = \beta_0 + E + \beta_1 X + \beta_2 Z$

图：slope不一样，interaction就显著。DV 翻转，1,0不变

0时，predition 为 x，every additional ?, increases / decrease by XX.

confounding variable：解释不同时，注意 extraneous variable

check：记算 [ ]，若 β_{1C} 不在 [ ]，则为 confounder

best set → 扰乱关系 X 和 Y

control variable：防止 confusion

决定是否 confound 比 [ ]

1° CV qt(0.975, 148)

2° 算 [ ]：$\beta_1 \pm CV \cdot SE$

3° 看完整模型中该元素 β_1(E) 是否在 [ ]。coefficient

4° 不在 [ ] 内 => 是 confound.

什么样的 model 最好？([ ] 最窄)

是否 confound 和 一般都和 model 的 [ ] 比。

L11 Interaction.

Model 加入 $\beta_3 XZ$

判断 relationship

$Y = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ + E$

Source	DF	SS	Mean Square	F-stat
Regression	k	SSR	$MSR = \frac{SSR}{k}$	$F = \frac{MSR}{MSE}$
Error	n-k-1	SSE	$MSE = \frac{SSE}{n-k-1}$	
Total	n-1	SSY		

$r^2 = \frac{SSR}{SSY}$ Residual Standard Dev $= S_{Y|X} = \sqrt{MSE} = \frac{S_X}{S_Y}\hat{\beta}_1$

① coincidence (是否 same slope & intercept) $H_0: \beta_2 = \beta_3 = 0$ (Z, XZ)

P小，拒绝，X coincident, significant

② parallelism (是否 same slope but different intercept)

$H_0: \beta_3 = 0$ (XZ) P小，拒绝，not parallel

① $F = \frac{(\beta_2, \beta_3 \text{对应的 SSR}) / 2}{\text{Full model Res / DF}}$

② $F = \frac{(\text{Full中3个SSR} - \text{Reduced中2个SSR})}{\text{Full中Res / DF}}$

not reject C, $\beta_0 \beta_1 E$

reject C, not P: $\beta_0 \beta_1 \beta_2 E$

X and Z on Y indep

$\beta_0 \beta_1 \beta_2 \beta_3 E$

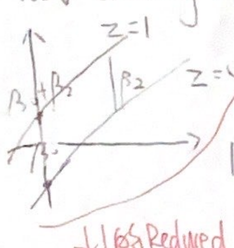
reject P: X & Y differ on Z

Legal	1	1
Illegal	0	0

// Z=1 时，β_2 调整 intercept，β_3 调整 slope

L12 协方差 (must be parallel // no significant inter-)

test dummy variables 的参数是否为 0.



必须平行才能 ANCOVA

步骤：

1.1 先测 prereq 即 $H_0: \beta_3 = 0$. → 是否 parallel

方法在上 (实际换 Z_i)

2.1 Full: $Y = \beta_0 + \beta_1 X + \beta_2 Z + E$ vs. Reduced: $\hat{Y} = \beta_0 + \beta_1 X + E$.

$F = \frac{\text{Full's SSR} - \text{Reduced SSR}}{\text{Full中Res / df}}$ test Z 的参数

pairs 是否 determined 的显著不同

DV 之间有对比呈它比

Adjusted Means.

Full: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 Z_1 + \beta_4 Z_2 + E$

Reduced: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + E$.

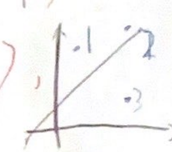
$H_0: \beta_3 = \beta_4 = 0$

H_1: at least one $\beta_i \neq 0$

$F = \frac{(\text{Seqss(Full - Reduced)'s Total Regression}) / df}{\text{Full's Res / df}}$ 根据 p 去选择 DV

求值：根据 output 列 $\hat{Y}$，代入 mean 求值.

L13 回归诊断 (三者必须都满足)



3种 point

1° outlier：有很大且极端的 res.

2° High leverage observation. X 距离其他很远，Y 的极端值，Res 很高，影响 R line ($> \frac{3(k+1)}{n}$)

3° Influential point：影响 slope / R line 的方向，Y 和 Res 都极端，cook's distance $> \frac{4}{n}$

hi Leverage

(0,1) cut-off: $\frac{2(k+1)}{n}$ // k 为有几个 predictor 元素，n 为 observation 数量

高 leverage 是指极端值

Res 类型

1° standardized Residual: $z_i = \frac{\hat{E}_i}{S_{Y|X}} = \frac{Y - \hat{Y}}{RSE}$

$= \frac{Y - Y_{cpred}}{RSE}$

2° studentized res: $r_i = \frac{Y - \hat{Y}}{S_{Y|X}\sqrt{1 - h_i}}$ → L

3° Jackknife $r_{(-i)} = \frac{Y - \hat{Y}}{S_{Y|X(-i)}\sqrt{1 - h_i}}$ // 作 without 的 model，改 $S_{Y|X}$

outlier CV

为 qt(.025, df)

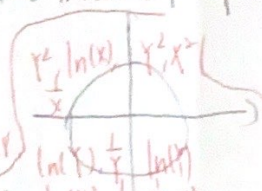
obs - pre数 - 1

Cook's distance $= \frac{4}{n_{obs}}$ // cutoff $d_i = \frac{1}{k+1} \cdot \frac{h_i}{1 - h_i} r_i^2$ num of pred

cook决定 influential, out r决定 outlier, leverage决定 high L

六角满足那3个条件，才能safe to remove influential

{ normality straight-line
linearity 是否 obvious curve in resplot
Homo ⟨ 这样不好，像 random scatter



L14 转 { tran Y, if homo & normal X
tran X, if linear X, but other ✓ → linear 经验

{ ln(Y): Y↑, variance of res↑时，应变量>0.
$\sqrt{Y}$: 同ln(Y): ——; $\frac{1}{Y}$: 同 ——.
$\ln(\frac{Y}{1-Y})$ response is proportion or rate, (0,1)~(-∞,∞)

Log tran化为 ↗先代进X进tran过的，再把Y还原
通过 reg eq 求 predicted response.

L15 多项: $Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_k x^k + \varepsilon$.

修正linearity ⇒ 用 lack of fit test ①

{ H_0: model合适 → X lack
H_a: model不合适 → lack of fit

$$F = \frac{\text{lack of fit SSLF}/(d-k-1)}{\text{pure SSPE}/(n-d)}$$

{ 1. Remove a subset. → data-based.
2. Create a linear combination of preds.
3. Center predictors. → structural

有多 correlation 高的 pair，但要与 response 有强 cor

去除无关元素 前解释，2个多 pair 高 cor

P值小的留.

为什么一个元素高cor去P高p? (因为有高corr with other元素).

另一个元素 贡献大的SSR，留下的约定不多.

留他元素高解释的东西不多.

引入多项. test for addition of β_i.

$$F = \frac{\text{extra SS for addition}}{\text{MSE Full model}}, df(1, n-k-1)$$

$$F = \frac{SS(\text{extra})}{\text{pure } df}, \text{可} \sim f(0.95, 1, n-k-1)$$

p小，拒绝，add显著提升模型.

当X lack of fit，可为 best model.

如果lack，可cubic，考虑要不要

如果straight-line和quadratic都有lack of fit

看addition of cubic term是否significant，再test

L16 collinearity: predictors彼此高度相关 ⇒ 不可

$VIF = \frac{1}{1-R^2}$ → r-square.; remove.

cluster of collinearity, R^2 越小代表越好

$$= \frac{\sum(\hat{Y}-\bar{Y})^2/d-k+1}{\sum(Y-\hat{Y})^2/n-d} = \frac{\text{lack of fit}/df}{\text{pure error}/df}, (V = q(0.95, df))$$

Center X. — only works when higher order terms are involved. Every term in this model is linear and centering — X to ↓ collinearity

L Combination: $X_1 \cdots$; $\beta \cdots$. They have different units of measurement and should not be combined using — bc the units would not make sense.

①: p小，差异显著(sgn) if test > cv，拒绝原假设，not pla...

LEC3 $r = \frac{SSXY}{\sqrt{SSX \cdot SSY}} = \frac{SXY}{SX\,SY} = \frac{SX}{SY} \cdot \hat{\beta}_1$ [no, weak, moderate, strong]

$SSXY = \sum (X_i - \bar{X})(Y_i - \bar{Y})$ $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$ $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$

Model 无参数 ↗ intercept Equation 有参数

$\hat{\varepsilon} = Y - \hat{Y}$ Error 实际 vs 预测

LEC4

Model: $Y = \beta_0 + \beta_1 X + \varepsilon$

linearity: ~~是否 straight line~~, obvious curve in plot.

homo: 不对，需要 random scatter，

normality: 是否 straight line

F-stat is $(t\text{-stat})^2$

相关性测试就测斜率是否为0.

LS CI: coef ± CV · SE coef

$t = \frac{\text{coef} - \text{var 一般值}}{\text{SE coef}}$ (df: n-2).

df = obs - 表里有几列

LEC6. Z(见仅面, correlation)

L7. 多元.

coefficient: 系数. 每增加

结果是否符合 CI.
{Reject H_0 Not 包含 Fail… contained.

residual: 预估值 - 实际值

L8 Muti part..

$F = \frac{(SSR(Full) - SSR(Reduced)}{Res(Full)/df}$.

L10. DV.

3个. 先0后1.

$\begin{pmatrix} 0 & 1 \\ 1 & 0 \\ 0 & 0 \end{pmatrix}$ → 表里没出现的0

$\hat{Y} = E + \beta_0 + \beta_1 X_1 + \beta_2 Z + \beta_3 X_1 Z + \beta_4 X_2 + \beta_5 X_2 Z \ldots$

student res 永远大于 res

图: slope 不一样, interaction 就显著 = 被包含.

CI 最宽的为 model 最好. β系数在最初模型以内, 就 confound.

是否 confound 未 = 最初 model 的为 CI 化. (都做了)

L11. interaction. Model 加入 $\beta_3 X Z$, 判断关系.

$Y = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ + E$, $R = \beta_0 + \beta_1 X + \varepsilon$

① coincidence (是否 slope & intercept 都相同)

$H_0: \beta_{3XZ} = \beta_{2Z} = 0$. p小, 拒绝, 显著, X coincident.

② parallel (是否 same slope but diff intercept).

$H_0: \beta_{3XZ} = 0$ …, not parallel.

ANOVA Table

Src	DF	Sum of Squares	Mean Square	F-stat
Regression	k	SSR	MSR = SSR/k	F = MSR/MSE
Error	n-k-1	SSE	MSE = SSE/(n-k-1)	
Total	n-1	SSY		

$SSR = \sum (\hat{Y}_i - \bar{Y})^2$, $SSE = \sum (Y_i - \hat{Y})^2$ $S_{Y|X} = \sqrt{\frac{SSE}{n-k-1}} = \sqrt{MSE}$ n-2?

slope $\hat{\beta}_1 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$ $R^2 = \frac{SSR}{SSY}$ residual standard deviation RSD.

$R^2 = \frac{\sum (\hat{Y}_i - \bar{Y})^2}{\sum (Y_i - \bar{Y})^2}$ R^2% … is explained by

$S_{\hat{\beta}_1} = \frac{S_{Y|X}}{\text{Stand Deviation}\sqrt{n-1}}$

CI (for all) $\hat{Y} \pm t_{n-2\text{会给}} \times S_{Y|X} \sqrt{\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{(n-1)S_X^2}}$.

PI (single). $\hat{Y} \pm \ldots \sqrt{1 + \frac{1}{n} + \ldots}$

outlier: 使模型更差, 提高 SSE, RSE, ↓R^2.

值离 mean 更远, interval 更宽

值相同, PI 比 CI 宽

$S_{Y|X}$ 每个值 deviate 多少.

L12 ANCOVA

新生: 新手不能测.

用 $H_0: \beta_3 = 0$. $X_i Z_i$ ~ . P值大, test < CV

$F = \frac{\text{Full送} - \text{R送}}{\text{Full Res/df}}$ 让XZ不重要.

adjust mean. Full 没有 XZ

{Full = $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 Z_1 + \beta_4 Z_2$ ~

Re = $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + E$

{$H_0: \beta_3 = \beta_4 (Z) = 0$, vs at (east

$F = \frac{(\text{Full送} - \text{R送})/df}{\text{Full' res/df}}$

求值: 根据 output 列个个代入 mean. 之前告知 Z 比

$F = \frac{(\beta_2, \beta_3 \text{对应的} SSR)/2}{\text{Full model Res/DF}}$.

$F = \frac{(\text{Full中3个SSR} - \text{R中2个SSR})}{\text{Full Res/df}}$ Full送 - R送

Slope: $H_0: \beta_1 = 值$ vs $H_1: \neq$ (或 $=$ $>$). $t = \frac{\hat{\beta}_1 - 值}{\frac{S_{Y|X}}{S_X\sqrt{n-1}}}$ CI: $\beta_1 \pm t_{n-2}\left(\frac{S_{Y|X}}{S_X\sqrt{n-1}}\right)$ //test < CV, 值 plausible

Intercept: $H_0: \beta_0 = 值$, $H_0 \neq$. $t = \frac{\beta_0 - 值}{S_{Y|X}\sqrt{\frac{1}{n} + \frac{\bar{X}^2}{(n-1)S_X^2}}}$ CI: $\beta_0 \pm t_{n-2}\left(S_{Y|X}\sqrt{\frac{1}{n} + \frac{\bar{X}^2}{(n-1)S_X^2}}\right)$.

Zero correlation. $H_0: \rho = 0$ vs $\neq 0$. $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$

non-zero: $\rho = 值$ vs $\neq 值$. $Z = \frac{\frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) - \frac{1}{2}\ln\left(\frac{1+值}{1-值}\right)}{1/\sqrt{n-3}}$.

Two correlation. $\rho_1 = \rho_2$ vs $\neq$. $Z = \frac{\frac{1}{2}\ln\left(\frac{1+r_1}{1-r_1}\right) - \frac{1}{2}\ln\left(\frac{1+r_2}{1-r_2}\right)}{\sqrt{\frac{1}{n_1-3} + \frac{1}{n_2-3}}}$.

Test for 特定 add. $F = \frac{Reg(X_3|Reg(X_3)/2}{2对应Error/DF}$

$SS(X_1,X_2,X_3) = SS(X_1) + SS(X_2|X_1) + SS(X_3|X_1,X_2)$. //单SS, 看第一行是那个的.

$SS(X_3|X_1,X_2) = SS(X_1,X_2,X_3) - SS(X_1,X_2)$

$SS(X_3,X_4|X_1,X_2) = SS(X_1,X_2,X_3,X_4) - SS(X_1,X_2)$ // 对 SS(A,B), 顺序不重要. SS(A|B), 顺序重要.

$SS(X_2|X_1) \neq SS(X_2)$

$SS(X_1,X_2) = SS(X_1) + SS(X_2|X_1) = SS(X_2) + SS(X_1|X_2)$

X_2 在模型里, X_1 加入增多少? $SS(X_1|X_2) = SS(X_1,X_2) - SS(X_2)$ (已经在模型里的)

Reg X_1, X_2, X_3 (后加, 测它的addition) Full, reduced model. $F = \frac{Reg(X_3)/1}{对应Error\ SS/DF}$

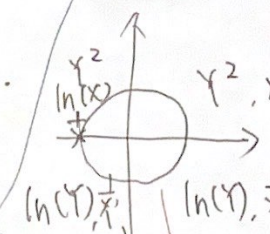
X_3 的贡献, 已经在模型里. Test for addation of X_3

Overall Reg (X_1, X_2 是否比空本模型好) Full: $Y = \beta_0 + \beta_1 X_1 + \cdots \beta_k X_k + \varepsilon$; R: $Y = \beta_0 + \varepsilon$; $H_0: \beta_1 = \cdots = \beta_k = 0$ vs at least one $\neq 0$. $F = \frac{SSR/k}{SSE/(n-k-1)}$ df 同理.

Partial F (测试单个 β^*) $F = \frac{SS(X^*|X_1 \cdots X_p)}{MSE(X_1 \cdots X_p, X^*)}$ 1, $n-p-1-1$.

Multi.

U4 转变



变曲 vs 特定回归形状.

{ tran Y: if homo & normal & linear 可以转倒数.
tran X: if linear X, but other ✓.

Log tran 比较好用 $(0,1) \sim (-\infty, \infty)$

将值代进 tran 的式子, 再 $e^{\ln(Y)}$ 求预值.

L3 回归诊断 ! 必须3条件都得满足. standard res $Z_i = \frac{Y-\hat{Y}}{S_{Y|X}}$

stu- res $r_i = \frac{Y-\hat{Y}}{S_{Y|X}\sqrt{1-h_i}}$

Jack Res $r_{(-i)} = \frac{Y-\hat{Y}}{S_{Y|X(-i)}\sqrt{1-h_i}}$

look $d_i = \left(\frac{1}{k+1}\right)\left(\frac{h_i}{1-h_i}\right)r_i^2$ ↓ 决定是哪种 influ, hL, outlier!

1° outlier: 很大且在远的 res

2° High leverage obs: 距离其他很远, $> \frac{2(k+1)}{n}$ //k 为 pred 变量

3° Influential point: 移动 slope. (X 和 Y 都极端)

大于就归类. cut off

Lib. collinearity $0 \sim \infty, \sim 1$

$VIF = \frac{1}{1-r^2}$ VIF 越低越好.

r-square 此时低好. 去掉表中变化值大的.

combine (相应单位 unit of measurement 相同)

center 指数作用, 但方差式. 可以 center.

② L20 模型选择

model 最多多少 variable? $n > 5k$ → predictor 数

variable 数

即对于一个 pred 至少 5 个 obs.

Max Model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 Z_1 + \beta_5 Z_2 + \varepsilon$ (全变量)

$R^2 = 1 - \frac{SSE(n-k-1)}{SSY(n-1)}$ Res 出现几个 pred

MSE 在 SSE 右边

多个新 pred, R^2 ↑ (no到 SS(R) ↑ SSR

模型: reduced 减少 predictor 数

$C_p = \frac{SSE(Reduced)}{MSE(Full)} - [n - 2(p+1)]$

$\frac{SSE}{n-k-1}$ $\frac{SSE}{n-k-1}$ 时所处 diff $= C_p - p - 1$, 越小越好

R^2 越大越好

$AIC = n\ln(SSE) - n\ln(n) + 2p$

$BIC = n\ln(SSE) - n\ln(n) + p\ln(n)$

obs 数

L18

A. B 值越小越好 并不一定同模型 有几个参数, 记 sin for

用 SSE 衡量模型

减去 $n\ln(n)$ 惩罚样本

对大模型有惩罚

Forward: 添加 pred 直到 p 是 pred 减 (opt. $p = pred + 1$

AIC(B) 会增加.

Backward: 删除 pred 让其减小.

Stepwise: 两者结合 在每一步做决定

F: 看单一模型, 找比里面最小的, 再看 2 元素有 X_1 的, 找自上降低 AIC, 再从 2 已确定的向选第三个. 找到最小 AIC.

B: 从缺一的模型里选 B 比最小的再缺二… 约减!

L21 时间序列

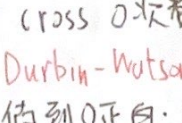
描述 linear trend: No, volatile(波动力).

sky rocket (暴升) plummet (暴跌)

Positive auto correlation: 给定符号自动倾向跟随相同符号

Negative autocorrelation: …跟随相反, 自动降

高点低点很长: positive, independence condition 不成立



Consecutive positive residuals tend to be followed by positive resuidals.

The same tends to happen once the residuals flip to being negative. (time series plot of residuals cross 0 次数少).

Durbin-Watson Test 自相关是否存在 $\begin{cases} H_0: \rho = 0 \\ H_a: \rho \neq 0 \end{cases}$

值到 0 正自; 到 4 负自

$\hat{Y} = E + \beta_1 L_1 + \beta_2 L_2 + \beta_3 L_3 \cdots$ (前几期自动)

predicted value: 代入最近的几个值

Auto regressive model

1.0, 0.9, 0.8, 0.7 …

Lagging: 将数据移动一定时间段以自适应

说明今天、明天 highly correlated 但时间过去, 预测性降低.

Cp 筛选: 从每个 size 中挑 Cp 最小的. 算 diff, 对比.

LE(1 交叉验证 { 一次删一个, 剩余继续拟合, 预测删除的. 适合小样本 (<1000)

Leave-one-out CV:

$SSR = PRESS$ $RMSE = \sqrt{\frac{PRESS}{n}}$ RMSE 越小越好, 说明模型精准确.

k fold CV k=n 时与 LOOCV 相同

$RMSE = \sqrt{\frac{RMSE_1^2 + RMSE_2^2 + \cdots RMSE_k^2}{k}}$

LE(2 逻辑回归

logistic reg model

$\text{logit}(p) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$ $p = \frac{e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k}}{1 + e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k}}$

g model

$e^?$ … 代入对应数值

$odds = \frac{p}{1-p}$ 95% CI = Coef ± 1.96 SE Coef conf 人参数

obs

	Fail	surress	Total
Fail	TN	FN	N_{pred}
Success	FP	TP	P_{pred}
Total	N	P	n

Accuracy = $\frac{TP+TN}{n}$

Sensitivity = $\frac{TP}{TP+FN}$

specificity = $\frac{TN}{TN+FP}$

1° Accuray 为准确预测模型

2° sensitivity 上 % 表现 misclassified as rushes pass

3° sp: 上 % 表现 rush as pass.

{ 都很好

$OR_{A\ vs\ B} = \frac{odds(A)}{odds(B)}$ $OR = e^{\beta_i}$

CI for slope: $\beta_i \pm Z_{1-\alpha/2} S_{\beta_i}$ (不含 0, 显著)

CI for odds ratio: $\exp(\beta_i \pm Z_{1-\alpha/2} S_{\beta_i})$ 不含 1 显著

根据 D-W Test 判断是否自相关: DW 接近 2 表示无相关.

p 值大, 无显著差异, 接受 H_0, 无相关.

0 2 4

L22 missing data

MAR: 关注因缺失而

MCAR: 缺失数据与特征完全无关. 不同的变量.

MAR: 值缺失的概率与所观测变量值有关.

如何确定? (比较变量缺失/不缺)

MCAR: 根据已知值的分布估算总体 mean/median.

自回归 model 特点: DW 2, p 值大, 无法拒绝 H_0 满足 不足

? diagonal 有相同相关性

predictors 由 lagging 创建, 相同的值出现于不同的 row, 以不同步 correlation.

§ 下 2 11. 中.

LS解决 Lack of fit. 从 obs, 个pred. $\hat{\beta}$

① Add of X to $Y = \beta_0 + \varepsilon$ 1- , X-2

② Lack of Fit $Y = \beta_0 + \beta_1 X + \varepsilon$ Y-2 , X-Y ,

③ Add X^2 $Y = \beta_0 + \beta_1 X + \varepsilon$ 1 X-3

④ LOF- $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$ Y-3 X-Y .

⑤ Add X^3 $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$ 1 X-4

⑥ LOF- $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \varepsilon$ Y-4 X-Y.

Y为Add, 上面体现型为Product. add of X'to. 用②vs③ to //F > (V, P值小, 不合适, 说明有必要

3为LOF, 4变量为 appropriate. 不需要变换了

LOF. F = $\frac{\text{Lack of Fit df}}{\text{PURE Error df}}$

Add F= $\frac{\text{新的的SSE}}{\text{Error df}}$

Default ( ) more likely to ~ compared to ~ needing the less by. () < 1, less likely to pass

1 more likely

equal

find down second down

1

对于adv currently

一个是为的

b. 到底 - 为LOF T-test <<V, 说明多

d - 比 - 1 > 1.

跟变量无关.

对称symmetric: 用mean来impute.

看missing与什么有关:

表中, (No) (a 差异很大, 将那两个格子给对立

例. 其它等多有的; 没的 mean 填补.

把有的mean给有的的人, 把没 … 给 …

类别数据.

描述一下data, 分类预测数据. categorical

大约 %的数据是什么. diff很小不足以得出你的结论

如何impute

MCAR: 用imputation替代丢失值 with plausible alternative.

↓ to

no difference

MAR: 删除, obs; pred, 替代

imputation 完整dataset, 更精确, 但有风险.

Us 口完全不同 loner levels.

如果像) 缺失对应类别的数据.