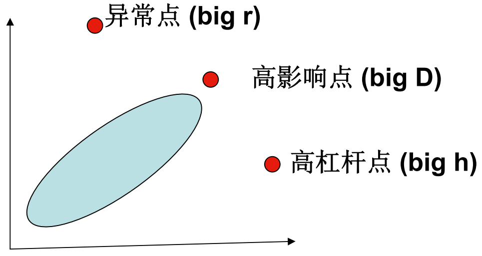
2014.5.6

第二十二讲. 置信域和同时置信区间

回顾影响分析



定义：$DFFITS_i=\dfrac{\hat{Y}_i-\hat{Y}_{i\,(-i)}}{\sqrt{h_{ii}\hat{\sigma}^2_{(-i)}}}$，其中$\hat{Y}_{i\,(-i)}=\tilde{x}_i\hat{\beta}_{(-i)}$ = 删除第i行后对Y_i的预测

证明：$DFFITS_i=r_i^*\sqrt{\dfrac{h_{ii}}{1-h_{ii}}}$，其中$r_i^*=\dfrac{e_i}{\hat{\sigma}_{(-i)}\sqrt{1-h_{ii}}}$

定义：Cook 距离 $D_i=\dfrac{\|\hat{Y}-\hat{Y}_{(-i)}\|^2}{p\hat{\sigma}^2}=\dfrac{(\hat{\beta}-\hat{\beta}_{(-i)})'X'X(\hat{\beta}-\hat{\beta}_{(-i)})}{p\hat{\sigma}^2}$

证明：$D_i=\dfrac{1}{p}\times\dfrac{h_{ii}}{1-h_{ii}}r_i^2$，其中$r_i=\dfrac{e_i}{\sqrt{1-h_{ii}}\hat{\sigma}}$为标准化残差。

证明：$D_i=\dfrac{1}{p}\times\dfrac{h_{ii}}{1-h_{ii}}r_i^2$，其中$r_i=\dfrac{e_i}{\sqrt{1-h_{ii}}\hat{\sigma}}$。

证：删除样本点i，拟合模型$Y_{(-i)}=X_{(-i)}\beta+\varepsilon_{(-i)}$，得$\hat{\beta}_{(-i)}=(X_{(-i)}'X_{(-i)})^{-1}X_{(-i)}'Y_{(-i)}$

记$X=\begin{pmatrix}\tilde{x}_1'\\ \cdots\\ \tilde{x}_n'\end{pmatrix}$，则$X'X=\sum_{i=1}^{n}\tilde{x}_i x_i'$，$X'Y=\sum_{i=1}^{n}\tilde{x}_i y_i$

$$\Rightarrow X'X=X_{(-i)}'X_{(-i)}+\tilde{x}_i\tilde{x}_i',\quad X'Y=\sum_{j=1}^{n}\tilde{x}_j y_j=X_{(-i)}'Y_{(-i)}+\tilde{x}_i y_i$$

利用事实：设$A_{p\times p}$对称，x，y为$p\times 1$向量，则$(A+xy')^{-1}=A^{-1}-\dfrac{A^{-1}xy'A^{-1}}{1+x'A^{-1}y}$，

$$\Rightarrow \hat{\beta}_{(-i)}=(X_{(-i)}'X_{(-i)})^{-1}X_{(-i)}'Y_{(-i)}=(X'X-\tilde{x}_i\tilde{x}_i')^{-1}(X'Y-\tilde{x}_i y_i)$$

$$=\left((X'X)^{-1}+\frac{(X'X)^{-1}\tilde{x}_i\tilde{x}_i'(X'X)^{-1}}{1-\tilde{x}_i'(X'X)^{-1}\tilde{x}_i}\right)(X'Y-\tilde{x}_i y_i)$$

注意到$h_{ii}=\tilde{x}_i'(X'X)^{-1}\tilde{x}_i$

$$\hat{\beta}_{(-i)}=\left((X'X)^{-1}+\frac{(X'X)^{-1}\tilde{x}_i\tilde{x}_i'(X'X)^{-1}}{1-\tilde{x}_i'(X'X)^{-1}\tilde{x}_i}\right)(X'Y-\tilde{x}_i y_i)$$

$$=\hat{\beta}-(X'X)^{-1}\tilde{x}_i y_i+\frac{1}{1-h_{ii}}(X'X)^{-1}\tilde{x}_i\tilde{x}_i'(X'X)^{-1}X'Y-\frac{1}{1-h_{ii}}(X'X)^{-1}\tilde{x}_i\tilde{x}_i'(X'X)^{-1}\tilde{x}_i y_i$$

$$=\hat{\beta}-(X'X)^{-1}\tilde{x}_i y_i+\frac{1}{1-h_{ii}}(X'X)^{-1}\tilde{x}_i\tilde{x}_i'\hat{\beta}-\frac{h_{ii}}{1-h_{ii}}(X'X)^{-1}\tilde{x}_i y_i$$

$$=\hat{\beta}-\frac{1}{1-h_{ii}}\left\{(X'X)^{-1}\tilde{x}_i\left(y_i-\tilde{x}_i'\hat{\beta}\right)\right\}=\hat{\beta}-\frac{1}{1-h_{ii}}\left\{(X'X)^{-1}\tilde{x}_i e_i\right\}$$

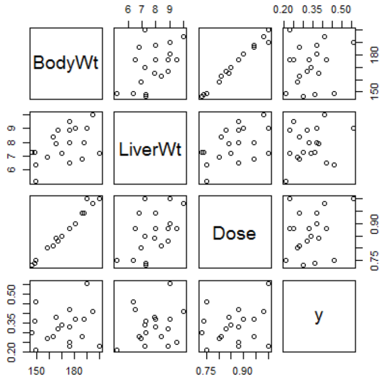
$$\Rightarrow \hat{\beta}-\hat{\beta}_{(-i)}=\frac{1}{1-h_{ii}}\left\{(X'X)^{-1}\tilde{x}_i e_i\right\}$$

所以 $D_i=\dfrac{(\hat{\beta}-\hat{\beta}_{(-i)})'X'X(\hat{\beta}-\hat{\beta}_{(-i)})}{p\hat{\sigma}^2}=\dfrac{1}{p}\times\dfrac{h_{ii}}{1-h_{ii}}r_i^2$

注：由$\hat{Y}_i-\hat{Y}_{i\,(-i)}=\left(X\hat{\beta}-X\hat{\beta}_{(-i)}\right)_i=\dfrac{1}{1-h_{ii}}\left\{\tilde{x}_i'(X'X)^{-1}\tilde{x}_i e_i\right\}=\dfrac{h_{ii}e_i}{1-h_{ii}}\Rightarrow DFFITS_i=r_i^*\sqrt{\dfrac{h_{ii}}{1-h_{ii}}}$

例2. (Dennis Cook数据)。 一项试验希望研究动物肝脏对某种药物的吸收能力。为此随机选取19只小白鼠，口服该药物，剂量大小(Dose)由体重决定 (大约为 40mg/kg 乘以体重)。一段时间后，测出肝中与所含药物重量，除以服用的剂量，得到肝吸收百分比 y。我们研究 y 与体重BodyWt, 肝重LiverWt，剂量Dose的关系。

理论上，这种决定药物剂量的方法决定了所测得到的 y 应该与三个变量无关。



lm(y ~ ., data = rat)

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	0.266	0.195	1.4	0.19	
BodyWt	-0.021	0.008	-2.7	0.02	*
LiverWt	0.014	0.017	0.8	0.42	
Dose	4.178	1.523	2.7	0.02	*

Multiple R-squared: 0.36,
F-statistic: 2.9 on 3 and 15 DF, p-value: 0.072

BowyWt和Dose都显著。

因为BodyWt, Dose几乎完全成正比，模型中去掉BodyWt 或Dose, 变量不再显著：

lm(y ~ . - Dose, data = rat)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.17836	0.22778	0.8	0.4
BodyWt	0.00035	0.00151	0.2	0.8
LiverWt	0.01233	0.02041	0.6	0.6

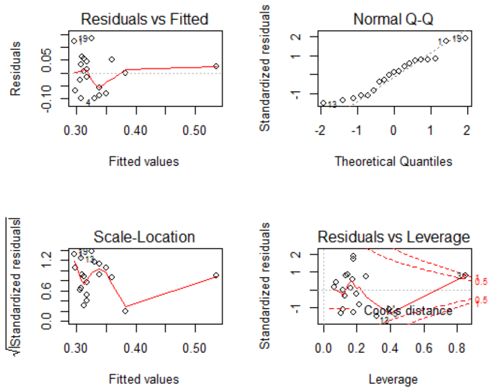
lm(y ~ . - BodyWt, data = rat)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.1174	0.2191	0.5	0.6
LiverWt	0.0087	0.0201	0.4	0.7
Dose	0.1736	0.2863	0.6	0.6

如何解释这种现象？

从第一个图可以发现第3个小白鼠的拟合值过大，其杠杆值 h =0.85；

从第4个图发现其D =0.93, 是高影响点。



Influence measures of
lm(formula = y ~ ., data = rat) :

	dfb.1_	dfb.BdyW	dfb.LvrW	dfb.Dose	dffit	cov.r	cook.d	hat
1	-0.038	0.315	-0.704	-0.244	0.892	0.631	0.169	0.178
2	0.143	-0.098	-0.482	0.126	-0.609	1.016	0.089	0.179
3	-0.231	-1.668	0.305	1.747	1.905	7.401	0.930	0.851
4	0.125	-0.127	-0.304	0.140	-0.494	0.860	0.057	0.108
5	0.522	-0.396	0.550	0.275	-0.909	1.524	0.203	0.392
6	0.002	0.014	0.029	-0.017	0.043	1.567	0.000	0.161
7	-0.184	0.150	-0.083	-0.118	0.310	1.289	0.025	0.137
8	-0.297	0.059	0.246	-0.040	0.426	1.520	0.047	0.254
9	-0.010	0.018	0.000	-0.017	0.043	1.402	0.000	0.067
10	-0.006	0.010	-0.003	-0.009	-0.014	1.496	0.000	0.120
11	-0.291	0.194	0.101	-0.173	-0.410	1.066	0.041	0.120
12	0.217	-0.025	0.052	-0.009	0.269	1.444	0.019	0.172
13	-0.772	0.144	0.766	-0.120	-1.099	0.972	0.273	0.316
14	-0.035	-0.046	-0.077	0.060	-0.142	1.461	0.005	0.131
15	0.019	0.041	-0.055	-0.038	0.119	1.359	0.004	0.076
16	0.123	-0.006	0.329	-0.049	-0.447	1.375	0.051	0.217
17	-0.104	0.015	-0.028	0.002	-0.126	1.607	0.004	0.195
18	-0.154	0.190	0.162	-0.189	0.352	1.270	0.032	0.149
19	0.856	-0.250	-0.295	0.171	0.995	0.517	0.200	0.178

删除第三行数据重新分析：

	BodyWt	LiverWt	Dose	y
1	176	6.5	0.88	0.42
2	176	9.5	0.88	0.25
3	190	9	1	0.56
4	176	8.9	0.88	0.23
5	200	7.2	1	0.23
6	167	8.9	0.83	0.32
7	188	8	0.94	0.37
8	195	10	0.98	0.41
9	176	8	0.88	0.33
10	165	7.9	0.84	0.38
11	158	6.9	0.8	0.27
12	148	7.3	0.74	0.36
13	149	5.2	0.75	0.21
14	163	8.4	0.81	0.28
15	170	7.2	0.85	0.34
16	186	6.8	0.94	0.28
17	146	7.3	0.73	0.3
18	181	9	0.9	0.37
19	149	6.4	0.75	0.46

lm(formula = y ~ ., data = rat[-3,])

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.311	0.205	1.52	0.15
BodyWt	-0.008	0.019	-0.42	0.68
LiverWt	0.009	0.019	0.48	0.64
Dose	1.485	3.713	0.40	0.70

Multiple R-squared: 0.0211,
F-statistic: 0.1 on 3 and 14 DF, p-value: 0.958

去除BodyWt, Dose之一：

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.3268	0.1957	1.67	0.12
BodyWt	-0.0003	0.0013	-0.25	0.81
LiverWt	0.0065	0.0170	0.38	0.71

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.3227	0.198	1.63	0.12
LiverWt	0.0061	0.017	0.36	0.72
Dose	-0.0553	0.253	-0.22	0.83

所有变量都不显著。

lm(y ~ ., data = rat)

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	0.266	0.195	1.4	0.19	
BodyWt	-0.021	0.008	-2.7	0.02	*
LiverWt	0.014	0.017	0.8	0.42	
Dose	4.178	1.523	2.7	0.02	*

lm(formula = y ~ ., data = rat[-3,])

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.311	0.205	1.52	0.15
BodyWt	-0.008	0.019	-0.42	0.68
LiverWt	0.009	0.019	0.48	0.64
Dose	1.485	3.713	0.40	0.70

lm(y ~ . - Dose, data = rat)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.17836	0.22778	0.8	0.4
BodyWt	0.00035	0.00151	0.2	0.8
LiverWt	0.01233	0.02041	0.6	0.6

lm(y ~ . - Dose, data = rat[-3,])

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.3268	0.1957	1.67	0.12
BodyWt	-0.0003	0.0013	-0.25	0.81
LiverWt	0.0065	0.0170	0.38	0.71

lm(y ~ . - BodyWt, data = rat)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.1174	0.2191	0.5	0.6
LiverWt	0.0087	0.0201	0.4	0.7
Dose	0.1736	0.2863	0.6	0.6

lm(y ~ . - BodyWt, data = rat[-3,])

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.3227	0.198	1.63	0.12
LiverWt	0.0061	0.017	0.36	0.72
Dose	-0.0553	0.253	-0.22	0.83

Influence measures of
lm(formula = y ~ . - Dose, data = rat) :

	dfb.1_	dfb.BdyW	dfb.LvrW	dffit	cov.r	cook.d	hat	inf
1	0.00102	0.28867	-0.43179	0.5333	1.103	0.092246	0.1647	
2	0.12469	0.16552	-0.49159	-0.6007	1.054	0.114968	0.1717	
3	-0.78638	0.53400	0.34751	1.1370	0.375	0.296102	0.1350	*

Influence measures of
lm(formula = y ~ . - BodyWt, data = rat) :

	dfb.1_	dfb.LvrW	dfb.Dose	dffit	cov.r	cook.d	hat	inf
1	0.03135	-0.38849	0.24283	0.4831	1.128	0.076530	0.1558	
2	0.11665	-0.48512	0.17765	-0.5862	1.081	0.110240	0.1747	
3	-1.11935	0.15633	0.95488	1.3922	0.430	0.453009	0.1986	*

这可能表明，某些高影响点，只有在有严重复共线性时才是高影响的。最初模型(所有数据、所有变量)分析的问题在于自变量之间的高度共线性性（复共线性性）导致出现了高影响点。

置信域

模型： $Y = X_{n\times p}\beta + \varepsilon$

A为$q\times p$行满秩矩阵，则正态假 设下$A\beta$的$(1-\alpha)100\%$置信域：

$$\left\{ b = A\beta\text{: } (b - A\hat\beta)'\left[A(X'X)^{-1}A'\right]^{-1}(b - A\hat\beta)/q\hat\sigma^2 \le F_{q,n-p}(1-\alpha) \right\}$$

由 $\hat\beta \sim N(\beta, \sigma^2(X'X)^{-1}) \Rightarrow A\hat\beta \sim N(A\beta, \sigma^2 A(X'X)^{-1}A')$

$$\Rightarrow \frac{(A\hat\beta - A\beta)'(A(X'X)^{-1}A')^{-1}(A\hat\beta - A\beta)}{\sigma^2} \sim \chi^2_q$$

$$\Rightarrow \frac{(A\hat\beta - A\beta)'(A(X'X)^{-1}A')^{-1}(A\hat\beta - A\beta)/q\sigma^2}{RSS/(n-p)\sigma^2} \sim F_{q,n-p}$$

分子只与$\hat{Y} = P_X Y$有关，分母只与$e = (I - P_X)Y$有关，故独立。

$$= \frac{(A\hat\beta - A\beta)'(A(X'X)^{-1}A')^{-1}(A\hat\beta - A\beta)}{q\hat\sigma^2}\text{是枢轴变量}$$

特别地, $A = \mathbf{a}'$是行向量时, $\mathbf{a}'\beta$的$(1-\alpha)100\%$置信区间:

$$\left\{ \mathbf{a}'\beta\text{: } |\mathbf{a}'\beta - \mathbf{a}'\hat\beta| \le t_{n-p}(\alpha/2)\hat\sigma\sqrt{\mathbf{a}'(X'X)^{-1}\mathbf{a}} \right\}$$

特别地, $A = (0,\ldots,1,0,\ldots0)'$, $A\beta = \beta_k$的$(1-\alpha)100\%$置信区间:

$$\left\{ \beta_k\text{: } |\beta_k - \hat\beta_k| \le t_{n-p}(\alpha/2)\hat\sigma\sqrt{(X'X)^{-1}_{kk}} = t_{n-p}(\alpha/2)\hat\sigma/\|\mathbf{x}_k^\perp\| \right\}$$

其中$\mathbf{x}_k$为X的第 k 列， $\mathbf{x}_k^\perp = \mathbf{x}_k - P_{X_{(-k)}}\mathbf{x}_k$

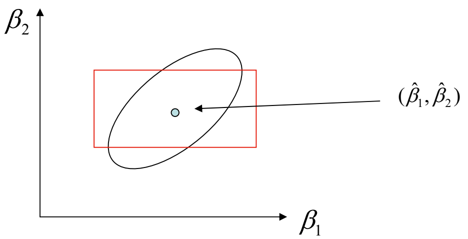
$A = (\mathbf{a}_1,\ldots,\mathbf{a}_q)'$, $A\beta$的置信域即q个参数$\mathbf{a}_1'\beta,\ldots,\mathbf{a}_q'\beta$的同时置信估计（为一椭 球）。

同时置信区间

为了简便实用，对于多个参数,能否构造长方形置信域

$$\left\{(\beta_1,\ldots,\beta_q):\ \mathrm{L_j} \le \beta_j - \hat{\beta}_j \le \mathrm{U_j},\ j=1,\ldots,q\right\}$$

称为同时置信区间simultaneous confidence interval



最为常用的是Bonferroni同时置信区间，其它还有Sheffe同时置信区间，对于方差分析模型，有Tukey同时置信区间.

Bonferroni 同时置信区间：

$\mathbf{a}_1'\beta,\ldots,\mathbf{a}_q'\beta$的Bonferroni $(1-\alpha)100\%$同时置信区间为

$$\bigcap_{j=1}^{q}\left\{\mathbf{a}_j'\beta:\ |\mathbf{a}_j'\beta - \mathbf{a}_j'\hat{\beta}| \le t_{n-p}\left(\frac{\alpha}{2q}\right)\hat{\sigma}\sqrt{\mathbf{a}_j'(X'X)^{-1}\mathbf{a}_j}\right\}$$

注意每一个 $a_j'\beta$的置信区间 I_j的置信水平是 $(1-\alpha/q)$而不是$1-\alpha$

证明: $\mathrm{P}(I_j) = 1-\dfrac{\alpha}{q}$,所以$P\left(\bigcap_{j=1}^{q} I_j\right) = 1 - P\left(\bigcup_{j=1}^{q} I_j^c\right) \ge 1 - \sum_{j=1}^{q} P(I_j^c) = 1-\alpha$

即Bonferroni同时置信区间的置信度为$1-\alpha$

Bonferroni 不等式：$P\left(\bigcup_{j=1}^{q}\mathrm{E}_j\right) \le \sum_{j=1}^{q} P(\mathrm{E}_j)$

特别地, $\beta_{i_1},\ldots,\beta_{i_q}$的$(1-\alpha)100\%$Bonferroni同时置信区间为

$$\bigcup_{j=i_1,\ldots i_q}\left\{\beta_j:\ |\beta_j - \hat{\beta}_j| \le t_{n-p}(\alpha/2q)\hat{\sigma}/\|\mathrm{x}_j^{\perp}\|\right\}$$

注1: 当$I_j, j=1,..,q$独立时($\mathrm{a}_1,\ldots,\mathrm{a}_q$ 正交时)

$$P\left(\bigcap_{j=1}^{q} I_j\right) = \bigcap_{j=1}^{q} P(I_j) = (1-\alpha/q)^q \approx 1-\alpha$$

即Bonferroni方法基本能精确地控制置信水平.

注2：而不独立时，Bonferroni过于严格（置信水平可能会远大于1-α），因而是一种过于保守的方法。

Scheffe同时置信区间:

Scheffe同时置信区间：

$$\mathrm{P}\left(|\mathbf{a}'\beta - \mathbf{a}'\hat{\beta}| \le \hat{\sigma}\sqrt{\mathbf{a}'(X'X)^{-1}\mathbf{a}}\sqrt{pF_{p,n-p}(1-\alpha)},\ \forall \mathbf{a}\in R^p\right) \ge 1-\alpha$$

引理：$\sup_{\mathrm{u}}\left\{(\mathrm{u'v})^2/\mathrm{u'}A\mathrm{u}\right\} \le \mathrm{v'}A^{-1}\mathrm{v}, \mathrm{u},\mathrm{v}\in \mathrm{R}^n, A>0$

证：由Cauchy不等式 $(\mathrm{u'v})^2 \le \mathrm{u'}A\mathrm{u}\times \mathrm{v'}A^{-1}\mathrm{v}$.

证：$\mathrm{P}\left(|\mathbf{a}'\beta - \mathbf{a}'\hat{\beta}| \le \hat{\sigma}\sqrt{\mathbf{a}'(X'X)^{-1}\mathbf{a}}\sqrt{pF_{p,n-p}(1-\alpha)},, \forall \mathbf{a}\in R^p\right)$

$$= \mathrm{P}\left(\sup_{\mathbf{a}\in R^p}\frac{|\mathbf{a}'\beta - \mathbf{a}'\hat{\beta}|}{\sqrt{\mathbf{a}'(X'X)^{-1}\mathbf{a}}} \le \hat{\sigma}\sqrt{pF_{p,n-p}(1-\alpha)}\right)$$

$$= \mathrm{P}\left((\beta - \hat{\beta})\mathrm{X'X}(\beta - \hat{\beta})/p\hat{\sigma}^2 \le F_{p,n-p}(1-\alpha)\right) = 1-\alpha$$

注：Scheffe同时置信区间含无穷多个区间，过长而较少使用

多重检验（multiple testing）

与建立$\beta_1,...,\beta_q$的同时置信区间类似的一个问题是同时检验多个假设

$$H_1:\beta_1=0,\ ...,\ H_q:\beta_q=0$$

更一般地，实际问题中可能考虑同时检验多个假设 $H_1, ..., H_q$
如果每个检验的水平都是α，而且各个检验是独立的，那么同时检验q个假设的I型错误率为$1-(1-\alpha)^q\approx q\alpha$。

更一般地，实际问题中可能考虑同时检验多个假设 $H_1, ..., H_q$
如果每个检验的水平都是α，而且各个检验是独立的，那么同时即使各个检验不独立，同时检验q个假设的I型错误率也会远大于α（由Bonfernnoni不等式，至多为$q\alpha$）。

为此需要校正（下调）单个检验的水平.

Bonfernnoni校正方法将单个检验的水平调整为α/q，由Bonferroni不等式，q个检验的I型错误率不会超过α。

换个角度看，Bonferroni方法将q个检验的最小p值$\min p_j$调整为$q\times\min p_j$

类似于Bonferroni同时置信区间，将单个检验的水平下调为α/q对于独立情形是精确的；对于不独立情形可能太谨慎/保守了。所以Bonferroni方法主要用于检验之间独立，或者不独立当相关结构未知的情形。


```
> lm(log(Species)~., data=lakes)->a
Coefficients:
              Estimate Std. Error t  value   Pr(>|t|)
(Intercept)   3.638e+00  5.313e-01  6.847   2.87e-07 ***
MaxDepth     -3.299e-04  2.815e-03 -0.117   0.9076
MeanDepth     6.357e-04  3.993e-03  0.159   0.8747
Cond         -2.334e-05  2.203e-04 -0.106   0.9164
Elev         -1.976e-04  8.751e-05 -2.258   0.0326 *
Lat          -2.345e-02  9.115e-03 -2.573   0.0161 *
Long         -4.422e-05  3.000e-03 -0.015   0.9884
Dist         -8.352e-02  4.569e-02 -1.828   0.0790 .
NLakes        6.471e-05  1.577e-04  0.410   0.6850
Photo         1.653e-04  2.298e-04  0.719   0.4784
Area          1.337e-07  6.051e-08  2.209   0.0362 *
---
F-statistic: 5.226 on 10 and 26 DF,  p-value: 0.0003291
```

比如上述结果中如果同时检验10个自变量是否存在某个或某些是显著的，最小p值为0.0161，Bonferroni校正后的p值：0.161不显著。(与回归方程的显著性F检验相比过于保守）。