应用数理统计——第五章

5.1 方差分析

一、方差分析

1.定义

方差分析是研究一个（或多个）分类自变量如何影响一个数值因变量的统计分析方法。

方差分析针对方差相同的多个正态总体，检验它们的均值是否相同。即，同时判断多组数据均值之间差异是否显著。

2.目的

（1）判断某些因素对于我们感兴趣的因变量是否具有“显著”的影响；
（2）如果因素间有交互效应，寻找最佳搭配方案。

3.特点

（1）方差分析与一般的假设检验不同

要比较均值是否相同，可以使用第三章假设检验的方法，但是只能处理两个均值；方差分析处理的是多个均值的情况。

（2）方差分析与回归、相关分析不同

回归与相关处理的是两个数值变量的问题，相应的散点在 x 轴上具有顺序（从小到大）；方差分析的数据在 x 轴上可以任意交换位置。

3.常见方差分析

单因素方差分析、双因素方差分析、多因素方差分析

二、方差分析数学模型

1.变量

（1）响应变量（因变量）：进行随机试验所考察的数值指标；
（2）因素或因子（自变量）：影响因变量的各不同分类原因；
（3）水平：各个因素所构成的组或者类型。

2.模型的构建

$$\begin{cases} y_{11}=\theta_0+\alpha_1+\beta_1+\varepsilon_{11} \\ y_{12}=\theta_0+\alpha_1+\beta_2+\varepsilon_{12} \\ y_{13}=\theta_0+\alpha_1+\beta_3+\varepsilon_{13} \\ y_{21}=\theta_0+\alpha_2+\beta_1+\varepsilon_{21} \\ y_{22}=\theta_0+\alpha_2+\beta_2+\varepsilon_{22} \\ y_{23}=\theta_0+\alpha_2+\beta_3+\varepsilon_{23} \end{cases} \quad \text{矩阵形式} \Rightarrow \quad \begin{pmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \\ y_{23} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \theta_0 \\ \alpha_1 \\ \alpha_2 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} + \begin{pmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{13} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{23} \end{pmatrix}$$

在方差分析中，同一个因素的不同水平看成是模型里的不同变量，而不能看成是同一个自变量在不同试验里的取值。（否则需要 y 对 x 有线性相依关系）。

三、单因素方差分析

1.主要任务

（1）检验假设：$H_0 = \beta_1 = \beta_2 = ... = \beta_r$；

（2）作出未知参数 $\beta_1, \beta_2, ..., \beta_r$ 以及 σ^2 的估计

2.方差分析中未知参数的估计及分布

（1）因素各水平效应的估计采用各个组内平均，

$$\hat{\beta}_i = \bar{y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$$

相应的分布显然是：$\hat{\beta}_i \sim N(\beta_i, \frac{\sigma^2}{n_i}), 1 \le i \le r$

（2）误差方差的估计利用残差平方和，

$$\hat{\sigma}^2 = \frac{RSS}{n-r} = \frac{1}{n-r}\sum_{i=1}^{r}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2$$

（3）$\hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_r, \hat{\sigma}^2$ 之间相互独立

3.方差分析平方和分解公式

（1）总平方和：$TSS = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(y_{ij} - \bar{y})^2$

总平方和是观察到的每个数据与总平均的差异的总和，表示因变量总的变化。它衡量了全部 y_{ij} 的差异，其值越大，说明 y_{ij} 之间的差异越大。

（2）自变量平方和：$CSS = \sum_{i=1}^{r} n_i(\bar{y}_i - \bar{y})^2$

自变量平方和是因为自变量不同的类型而产生的差异，表示自变量在因变量的变化中所占的份额。用Σ（每组平均 总平均）^{2} 来刻化。

（3）残差平方和：$RSS = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2$

残差平方和表示其他原因引起的因变量变化，用Σ（观察值 每组平均）^{2} 来刻化。

（4）关系：$TSS = CSS + RSS$

4.单因素方差分析的检验

（1）检验统计量

$$F\text{比}=\frac{n-r}{r-1}\frac{CSS}{RSS}\sim F(r-1,n-r)$$

（2）拒绝域

$$W=\{F\text{比}\geq F_{\alpha}(r-1,n-r)\}$$

（3）单因素方差分析表

方差来源	平方和	自由度	均方	F-比
分类变量	CSS	r-1	CMS	F 比
残差变量	RSS	n-r	RMS	
总计	TSS	n-1		

其中：$CMS=\frac{CSS}{r-1},RMS=\frac{RSS}{n-r},F\text{-比}=\frac{CMS}{RMS}$

（4）变量关系的强度

$$R^2=\frac{\text{自变量平方和}}{\text{总平方和}}=\frac{CSS}{TSS}$$

5.2 线性回归分析

一、线性模型理论

1.线性模型的定义

y 是可观察的随机变量，$x_1,...,x_m$ 是可观察的分类或数值变量，$\beta_0,...,\beta_k$ 是未知参数，ε 是不可观察随机误差$(\varepsilon \sim N(0,\sigma^2))$。

$$y=\beta_0+\sum_{i=1}^{k} f_i(x_1,...,x_m)\beta_i+\varepsilon$$

称为是线性模型。

2.线性模型的表示

线性模型中“线性”是针对未知参数 β 而言，许多表面上的非线性模型本质也是线性的。

一些统计学家喜欢把线性模型表示成：

$$Ey=\beta_0+x_1\beta_1+...+x_k\beta_k$$

含义是：线性模型就是一个随机变量的数学期望具有未知参数线性结构的统计模型。

3.线性模型的参数估计——最小二乘估计

1.对未知参数 β 的估计：

（1）求解思路：平方和分解

$$\begin{aligned}\|Y-X\beta\|^2&=\|Y-X\hat{\beta}+X\hat{\beta}-X\beta\|^2\\&=\|Y-X\hat{\beta}\|^2+\|X(\hat{\beta}-\beta)\|^2+2(\hat{\beta}-\beta)^T X^T(Y-X\hat{\beta})\end{aligned}$$

使 $\|Y-X\beta\|^2\geq\|Y-X\hat{\beta}\|^2$ 成立的充要条件为 $(\hat{\beta}-\beta)^T X^T(Y-X\hat{\beta})=0$

得到正规方程 $(X^TX)\beta=X^TY$

（2）结论：

最小二乘估计：$\hat{\beta}=(X^TX)^{-1}X^TY$

经验回归函数：$X\hat{\beta}$

经验回归方程：$Y=X\hat{\beta}$

2.对未知参数σ^2的估计：

（1）求解思路：残差平方和

$$Q_e = \| Y - X\hat{\beta} \|^2 = Y^T(I_n - X(X^TX)^{-1}X^T)Y$$

（2）结论：

$$\sigma^2 = \frac{1}{n-k-1}Y^T(I_n - X(X^TX)^{-1}X^T)Y$$

3.随机向量的均值与方差：

（1）期望：$E(Y^TAY) = (EY)^TA(EY) + tr\{A[Var(Y)]\}$

a. β的无偏估计

因为：$E(\hat{\beta}) = E((X^TX)^{-1}X^TY) = E((X^TX)^{-1}X^TX\beta) = \beta$，

所以：$(X^TX)^{-1}X^TY$为β的无偏估计

b. σ^2的无偏估计

因为：

$$\begin{aligned} E(Q_e) &= E(Y^T(I_n - X(X^TX)^{-1}X^T)Y) \\ &= \beta^TX^T(I_n - X(X^TX)^{-1}X^T)X\beta + tr[(I_n - X(X^TX)^{-1}X^T)\sigma^2I_n] \\ &= 0 + \sigma^2 tr(I_n - X(X^TX)^{-1}X^T) = \sigma^2[n - tr(X(X^TX)^{-1}X^T)] \\ &= \sigma^2[n - tr((X^TX)^{-1}X^TX)] = (n-k-1)\sigma^2 \end{aligned}$$

所以：$\frac{1}{n-k-1}Y^T(I_n - X(X^TX)^{-1}X^T)Y$为$\sigma^2$的无偏估计

（2）方差：$Var(BY) = B[Var(Y)]B^T$

4.最小二乘估计量的分布：

对于线性模型$Y = X\beta + \varepsilon, \beta \in R^{k+1}$，$X$是$n \times (k+1)$满秩矩阵，$\varepsilon \sim N(0, \sigma^2I_n)$

（1） β的最小二乘估计服从$k+1$维正态分布，

$$\hat{\beta} = (X^TX)^{-1}X^TY \sim N(\beta, \sigma^2(X^TX)^{-1})$$

（2） σ^2的估计量服从卡方分布，即

$$\frac{n-k-1}{\sigma^2}\hat{\sigma}^2 = \frac{1}{\sigma^2}Y^T(I_n - X(X^TX)^{-1}X^T)Y \sim \chi^2(n-k-1)$$

（3） $\hat{\beta}$与$\hat{\sigma}^2$相互独立

二、一元回归分析

1.统计量的处理方式

（1）表格化处理

x	
y	

（2）平方和处理

$$TSS=\sum_{i=1}^{n}(y_i-\bar{y})^2, RegSS=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2, RSS=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$

$$TSS=RegSS+RSS$$

（3）L 化处理

$$L_{xx}=\sum_{i=1}^{n}(x_i-\bar{x})^2, L_{yy}=\sum_{i=1}^{n}(y_i-\bar{y})^2,$$

$$L_{xy}=\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})=\sum_{i=1}^{n}x_iy_i-n\overline{xy},$$

2.回归与相关分析

（1）回归分析与相关分析：

回归与相关分析是用于讨论数值变量之间关系的统计分析方法。回归分析研究一个或多个自变量的变化如何影响因变量，相关分析研究这两个数值变量的相关程度。

（2）一元线性回归模型

$$y_i=\beta_0+\beta_1x_i+\varepsilon_i, 1\le i\le n$$

其中：$\hat{\beta}_0=\bar{y}-\hat{\beta}_1\bar{x}, \hat{\beta}_1=\frac{L_{xy}}{L_{xx}}, \hat{\sigma}^2=\frac{1}{n-2}(L_{xx}-\hat{\beta}_1L_{yy})$

（3）相关系数

1）表达式：$r^2=\frac{(\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}))^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}=\frac{L_{xy}^2}{L_{xx}L_{yy}}=\frac{RegSS}{TSS}$

2）正相关与负相关：

两个数值变量具有正相关关系，是指因变量将随着自变量的增加而增加；负相关是指因变量将随着自变量的增加而减小。

3）关系强度：相关系数 r 是介于-1 到 1 之间的小数，一般认为：

【-1，-0.75】 【-0.7，-0.3】 【-0.25，0.25】 【0.3，0.7】 【0.75，1】

【负强】 【负适中】 【弱的关系】 【正适中】 【正强】

3.回归系数的假设检验

（1）估计量的分布

1） $\hat{\beta}_0 \sim N(\beta_0, \sigma^2(\frac{1}{n}+\frac{\bar{x}^2}{L_{xx}}))$

2） $\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{L_{xx}})$

3） $\hat{\beta}_0$ 与 $\hat{\beta}_1$ 不独立，协方差为 $Cov(\hat{\beta}_0, \hat{\beta}_1) = -\sigma^2 \frac{\bar{x}}{L_{xx}}$

4） $\hat{\sigma}^2$ 与 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 都独立，并且 $\frac{n-2}{\sigma^2}\hat{\sigma}^2 \sim \chi^2(n-2)$

（2）对 $\hat{\beta}_0$ 的检验

T 检验统计量： $\frac{\hat{\beta}_0}{\hat{\sigma}}\sqrt{\frac{nL_{xx}}{\sum_{i=1}^{n} x_i^2}} \sim t(n-2)$

F 检验统计量： $\frac{\hat{\beta}_0^2}{\hat{\sigma}^2}\frac{nL_{xx}}{\sum_{i=1}^{n} x_i^2} \sim F(1, n-2)$

（3）对 $\hat{\beta}_1$ 的检验

T 检验统计量： $\frac{\hat{\beta}_1}{\hat{\sigma}}\sqrt{L_{xx}} \sim t(n-2)$

F 检验统计量： $\frac{\hat{\beta}_1^2}{\hat{\sigma}^2}L_{xx} = \frac{(n-2)L_{xy}^2}{L_{xx}L_{xy} - L_{xy}^2} = \frac{(n-2)r^2}{1-r^2} = \frac{(n-2)RegSS}{RSS} \sim F(1, n-2)$

4.估计、预测与控制

（1）置信区间

$$\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}}\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \sim t(n-2)$$

$$\Rightarrow [\hat{\beta}_1 - \frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} t_{\alpha/2}(n-2), \hat{\beta}_1 + \frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} t_{\alpha/2}(n-2)]$$

（2）预测区间

$$y_0 - y_0^* \sim N(0, \sigma^2[1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}])$$

$$\Rightarrow \frac{y_0 - y_0^*}{\sigma\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}} \sim N(0,1)(\text{又}\frac{n-2}{\sigma^2}\hat{\sigma}^2 \sim \chi^2(n-2))$$

$$\Rightarrow \hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}} \sim t(n-2)$$

$$\Rightarrow [\hat{\beta}_0+\hat{\beta}_1 x_0 - \hat{\sigma}t_{\alpha/2}(n-2)\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}, \hat{\beta}_0+\hat{\beta}_1 x_0 + \hat{\sigma}t_{\alpha/2}(n-2)\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}]$$

（3）控制区间

取 x_0 使得：

$$A \le y_0^* - \hat{\sigma}t_{\alpha/2}(n-2)\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}, y_0^* + \hat{\sigma}t_{\alpha/2}(n-2)\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}} \le B$$

成立的区间 $[A,B]$。

三、多元回归分析

1.F 统计量的构造

根据定理 5.1.3 可知，有下式成立：

$$\frac{RSS}{\sigma^2} \sim \chi(n-k-1)$$

当 $H_0: \beta_1 = ... = \beta_k = 0$ 也成立时，有：

$$\frac{RegSS}{\sigma^2} \sim \chi(k)$$

所以构造 F 统计量如下：

$$F = \frac{n-k-1}{k}\frac{RegSS}{RSS} \sim F(k, n-k-1)$$

其拒绝域为：

$$W = \{F \ge F_\alpha(k, n-k-1)\}$$

2.回归因子检验

对于零假设 $H_{0i}:\beta_i=0$ 进行检验：

给出矩阵 C 的定义：

$$C=(X^TX)^{-1}$$

由此可构造如下统计量：

（1）T 统计量：

$$T_i=\frac{\hat{\beta}_i}{\sqrt{c_{ii}}\hat{\sigma}}\sim t(n-k-1)$$

（2）F 统计量：

$$F_i=\frac{\hat{\beta}_i^2}{c_{ii}\hat{\sigma}^2}\sim F(1,n-k-1)$$

5.3 题型总结

一、单因素方差分析问题

1.题型描述：利用方差分析，判断某一多水平分类变量对某一数值变量是否影响显著。

2.解题策略：

利用平方和，列出单因素方差分析表进行求解。

【模板】

解：$n=$【数据个数】，$r=$【水平个数】

$TSS=$【总平方和】

$CSS=$【自变量平差和】

$RSS=$【残差平方和】

可列单因素方差分析表如下：

方差来源	平方和	自由度	均方	F-比
分类变量	【CSS】	【$r-1$】	【CMS】	【F比】
残差变量	【RSS】	【$n-r$】	【RMS】	
总计	【TSS】	【$n-1$】		

因为【F比 和 $F_{\alpha}(r-1,n-r)$ 比较】，所以【题干】【有/无】显著影响。

3.题解：

例：试判断在检验水平为 0.05 的条件下，是否可以认为灯丝配料对灯泡寿命有显著影响？

灯丝	使用寿命（小时）							
甲	1600	1610	1650	1680	1700	1720	1800	
乙	1580	1640	1640	1700	1750			
丙	1460	1550	1600	1640	1660	1740	1820	1820
丁	1510	1520	1530	1570	1600	1680		

解：$n=$【26】，$r=$【4】

$TSS=$【227250】

$CSS=$【47399.17】

$RSS=$【179850.8】

可列单因素方差分析表如下：

方差来源	平方和	自由度	均方	F-比
分类变量	【47399.17】	【3】	【15799.7】	【1.9327】
残差变量	【179850.8】	【22】	【8175.04】	
总计	【227250】	【25】		

因为【$F_{0.05}(3,22)=3.05>F$比】，所以【可以认为灯丝配料对灯泡寿命】【有】显著影响。

二、最小二乘法解决一元回归分析问题

1.题型描述：利用最小二乘法，解决一元回归分析的诸多问题。

2.解题策略：

（1）回归方程求解问题：根据数据的处理形式做出不同的解决方法。

【表格化处理】

1） 写出矩阵 X^T 和向量 y^T

2） 求出矩阵 X^TX 和 $(X^TX)^{-1}$

3） 利用公式 $\beta=(X^TX)^{-1}X^Ty$ 求出 β

4） 写出回归方程

【L 化处理】

1） 根据 $\hat{\beta}_1$ 的估计公式 $\hat{\beta}_1=\frac{L_{xy}}{L_{xx}}$ 求出 $\hat{\beta}_1$

2） 根据 $\hat{\beta}_0$ 的估计公式 $\hat{\beta}_0=\bar{y}-\hat{\beta}_1\bar{x}$ 求出 $\hat{\beta}_0$

3） 写出回归方程

（2）剩余方差的计算问题：代入公式求解即可。

【表格化处理】

$$\hat{\sigma}^2=\frac{1}{n-k-1}Y^T(I_n-X(X^TX)^{-1}X^T)Y$$

【L 化处理】

$$\hat{\sigma}^2=\frac{1}{n-k-1}(L_{yy}-\frac{L_{xy}^2}{L_{xx}})$$

（3）假设检验问题：构造 F 统计量进行假设检验。

1）写出矩阵 C

2）构造 F 统计量 $F_i=\frac{\hat{\beta}_i^2}{c_{ii}\hat{\sigma}^2}\sim F(1,n-2)$

3）计算统计量对应的计算值

4）查表比较与计算值的大小

5）得出结论

（4）假设检验的变形问题：利用线性变换，构造 F 统计量求解。

1）写出矩阵 C

2）将原假设转化为 β 的线性变换

3）对矩阵C和向量β做线性变换得到c和β'

4）构造 F 统计量$F=\dfrac{\hat{\beta}'^2}{c\hat{\sigma}^2}\sim F(1,n-2)$

5）查表比较与计算值的大小

6）得出结论

（5）置信区间、预测区间与控制区间问题：根据所求区间，查找对应公式代数即可。

1）置信区间

$$\frac{\hat{\beta}_1-\beta_1}{\hat{\sigma}}\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sim t(n-2)$$

$$\Rightarrow[\hat{\beta}_1-\frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}}t_{\alpha/2}(n-2),\hat{\beta}_1+\frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}}t_{\alpha/2}(n-2)]$$

2）预测区间

$$y_0-y_0^*\sim N(0,\sigma^2[1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}])$$

$$\Rightarrow\frac{y_0-y_0^*}{\sigma\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}}\sim N(0,1)(\text{又}\frac{n-2}{\sigma^2}\hat{\sigma}^2\sim\chi^2(n-2))$$

$$\Rightarrow\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}\sim t(n-2)$$

$$\Rightarrow[\hat{\beta}_0+\hat{\beta}_1x_0-\hat{\sigma}t_{\alpha/2}(n-2)\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}},\hat{\beta}_0+\hat{\beta}_1x_0+\hat{\sigma}t_{\alpha/2}(n-2)\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}]$$

3）控制区间

取x_0使得：

$$A\le y_0^*-\hat{\sigma}t_{\alpha/2}(n-2)\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}},y_0^*+\hat{\sigma}t_{\alpha/2}(n-2)\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}\le B$$

成立的区间$[A,B]$。

3.题解：

例：今有 10 组观察数据由下表给出：

x	0.5	-0.8	0.9	-2.8	6.5	2.3	1.6	5.1	-1.9	-1.5
y	-0.3	-1.2	1.1	-3.5	4.6	1.8	0.5	3.8	-2.8	0.5

应用线性模型

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, i = 1,2,...,10.$$

假定诸 ε_i 独立同服从分布 $N(0,\sigma^2)$

（1） 求 β_0, β_1 的最小二乘估计，并给出回归方程；

（2） 计算剩余方差 δ_e^2；

（3） 在显著性水平 $\alpha = 0.05$ 下检验假设 $H_0 : \beta_1 = 0$；

（4） 在显著性水平 $\alpha = 0.05$ 下检验假设 $H_0 : \beta_0 = \beta_1$；

（5） 求 y 的置信水平为 0.95 的预测区间；

（6） 求 y 的置信水平为 0.95 的置信区间。

解：（1）求 β_0, β_1 的最小二乘估计，并给出回归方程。

1）写出矩阵 X^T 和向量 y^T

$$X^T = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0.5 & -0.8 & 0.9 & -2.8 & 6.5 & 2.3 & 1.6 & 5.1 & -1.9 & -1.5 \end{pmatrix}$$

$$y^T = \begin{pmatrix} -0.3 & -1.2 & 1.1 & -3.5 & 4.6 & 1.8 & 0.5 & 3.8 & -2.8 & 0.5 \end{pmatrix}$$

2）求出矩阵 $X^T X$ 和 $(X^T X)^{-1}$

$$X^T X = \begin{pmatrix} 10 & 9.9 \\ 9.9 & 91.51 \end{pmatrix}, (X^T X)^{-1} = \begin{pmatrix} 0.112 & -0.012 \\ -0.012 & 0.012 \end{pmatrix}$$

3）利用公式 $\beta = (X^T X)^{-1} X^T y$ 求出 β

$\hat{\beta} = (X^T X)^{-1} X^T y = \begin{pmatrix} -0.349 \\ 0.807 \end{pmatrix}$，所以 $\hat{\beta}_0 = -0.349, \hat{\beta}_1 = 0.807$

4）写回归方程

所以回归方程为： $y = -0.349 + 0.807x$

（2）计算剩余方差δ_e^2

$$\delta_e^2=\frac{1}{n-k-1}Y^T(I_n-X(X^TX)^{-1}X^T)Y=\frac{1}{8}\sum_{i=1}^{n}(y_i-(-0.349+0.807x_i))^2=0.8673$$

（3）在显著性水平$\alpha=0.05$下检验假设$H_0:\beta_1=0$；

1）写出矩阵C

$$C=(X^TX)^{-1}=\begin{pmatrix}0.112 & -0.012\\ -0.012 & 0.012\end{pmatrix}$$

2）构造 F 统计量$F_i=\frac{\hat{\beta}_i^2}{c_{ii}\hat{\sigma}^2}\sim F(1,n-2)$

$$F_1=\frac{\hat{\beta}_1^2}{c_{11}\hat{\sigma}^2}\sim F(1,8)$$

3）计算统计量对应的计算值

$$F_1=\frac{0.807^2}{0.012\times 0.8673}=62.57$$

4）查表比较与计算值的大小

查表可得：$F_{0.05}(1,8)=5.32$

5）得出结论

因为 62.57>5.32，所以拒绝H_0，即认为$\beta_1\neq 0$

（4）在显著性水平$\alpha=0.05$下检验假设$H_0:\beta_0=\beta_1$

1）写出矩阵C

$$C=(X^TX)^{-1}=\begin{pmatrix}0.112 & -0.012\\ -0.012 & 0.012\end{pmatrix}$$

2）将原假设转化为β的线性变换

$H_0:\beta_0=\beta_1$等价于$H_0:(1,-1)\begin{pmatrix}\beta_0\\ \beta_1\end{pmatrix}=0$，$H=(1,-1)$

3）对矩阵C和向量β做线性变换得到c和β'

$$c=[H(X^TX)^{-1}H^T]^{-1}=[(1,-1)\begin{pmatrix}0.112 & -0.012\\ -0.012 & 0.012\end{pmatrix}(1,-1)^T]^{-1}=\frac{1}{0.148}$$

$$\beta' = H\hat{\beta} = (1,-1)\begin{pmatrix}-0.349\\0.807\end{pmatrix} = -1.156$$

4）构造 F 统计量 $F = \dfrac{\hat{\beta}'^2}{c\hat{\sigma}^2} \sim F(1,n-2)$

$$F = \frac{\hat{\beta}'^2}{c\hat{\sigma}^2} \sim F(1,8)$$

5）查表比较与计算值的大小

$$F = \frac{(-1.156)^2}{\frac{1}{0.148}\times 0.8673} = 0.228$$

查表可得： $F_{0.05}(1,8) = 5.32$

6）得出结论

因为 0.228<5.32，所以接受 H_0，即认为 $\beta_0 = \beta_1$

（5）求 y 的置信水平为 0.95 的预测区间

$$y_0 - y_0^* \sim N(0,\sigma^2[1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}])$$

$$\Rightarrow \frac{y_0 - y_0^*}{\sigma\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}} \sim N(0,1)(\text{又}\frac{n-2}{\sigma^2}\hat{\sigma}^2 \sim \chi^2(n-2))$$

$$\Rightarrow \hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}} \sim t(n-2)$$

$$\Rightarrow [\hat{\beta}_0+\hat{\beta}_1x_0-\hat{\sigma}t_{\alpha/2}(n-2)\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}, \hat{\beta}_0+\hat{\beta}_1x_0+\hat{\sigma}t_{\alpha/2}(n-2)\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}]$$

所以，预测区间为 $[-0.349+0.807x \pm 2.1476\sqrt{1.1135-0.0273x+0.0138x^2}]$

（6）求 y 的置信水平为 0.95 的置信区间

$$\frac{\hat{\beta}_1-\beta_1}{\hat{\sigma}}\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sim t(n-2)$$

$$\Rightarrow[\hat{\beta}_1-\frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}}t_{\alpha/2}(n-2),\hat{\beta}_1+\frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}}t_{\alpha/2}(n-2)]$$

所以，置信区间为$[0.807\pm\frac{0.9313\times 2.306}{81.709}]$