Pure Mathematics 理论数学, 2019, 9(8), 871-879
Published Online October 2019 in Hans. http://www.hanspub.org/journal/pm
https://doi.org/10.12677/pm.2019.98114

Hans汉斯

Using Statistical Software to Analyze the Relationship between the Rehabilitation Rate and the Percentage of Low-Income Students

Huanyu Yang

Department of Mathematics, ZhanJiang Preschool Education College, Zhanjiang Guangdong
Email: yhyjy2014@163.com

Received: Sep. 16th, 2019; accepted: Oct. 4th, 2019; published: Oct. 11th, 2019

Abstract

On July 23, 2006, the Houston Chronicle published an article entitled "Reading: First-Grade Standard Too Tough for Many". The article claimed in part that "Some people think that more students (across Texas) do not have to rehabilitate the first grade, and experts believe that students in poor areas should be rebuilt in the first grade." The article presents data for each of 61 Texas counties on Y = Percentage of students repeating first grade x = Percentage of low-income students for both 2004~2005 and 1994~1995. The data can be found on the book web site in the file HoustonChronicle.csv. Analysis of covariance is used to decide whether: 1) an increase in the percentage of low income students is associated with an increase in the percentage of students repeating first grade; 2) there has been an increase in the percentage of students repeating first grade between 1994~1995 and 2004~2005; 3) an association between the percentage of students repeating first grade and the percentage of low-income students differs between 1994~1995 and 2004~2005.

Keywords

Analysis of Covariance, R Program

利用统计软件分析重修率与低收入家庭学生的百分比之间的关系

杨环瑜

湛江幼儿师范专科学校数学系，广东 湛江
Email: yhyjy2014@163.com

文章引用: 杨环瑜. 利用统计软件分析重修率与低收入家庭学生的百分比之间的关系[J]. 理论数学, 2019, 9(8): 871-879. DOI: 10.12677/pm.2019.98114

收稿日期：2019年9月16日；录用日期：2019年10月4日；发布日期：2019年10月11日

摘　要

2006年7月23日，休斯敦Chronicle发表了题为*Reading: First-Grade Standard Too Tough for Many*的文章，称“部分人认为更多学生(横跨得克萨斯州)都不必重修一年级，而专家认为应该让贫困地区的学生重修一年级。”该文章介绍的数据为2004~2005年和1994~1995年每个在得克萨斯州61个县的学生：Y = 一年级学生重修率，X = 低收入家庭学生百分比。这些数据来自于书中网站上的文件HoustonChronicle.csv。用R语言采用协方差分析来决定：1) 一年级学生重修率与低收入家庭学生百分比之间的关系；2) 1994~1995年和2004~2005年一年级学生重修率的差异；3) 1994~1995年和2004~2005年一年级学生重修率与低收入家庭学生百分比之间关系的差异。

关键词

协方差分析，R语言

1. 引言

重修率与低收入家庭学生的百分比之间的关系是我们现在教学必须研究的课题，本文采用 R 语言程序强大的绘图功能描述它们之间的差异，采用协方差分析方法，结合理论与数学模型，在图形与命令结果中直观地看出它们的差异，为以后教学的研究提供可行性借鉴。

本次论文在方法的使用与模型的构造上都使用协方差分析方法，是在线性回归的基础上做主要的方差分析[1]，最终结果用偏 F 检验，同时在 R 程序命令中以图形显示。

2. 单因子方差分析

2.1. 数学模型

设试验只有一个因子(又称为因素) A 有 r 个水平 $A_1, A_2, \cdots, A_r$，现在水平 A_i 下进行 n_i 次独立预测，得到观测数据为 $X_{ij}, j=1,2,\cdots,n_i, i=1,2,\cdots,r$，则单因素模型[2]可表示为

$$X_{ij} = \mu + \alpha_i + \varepsilon_{ij}, i=1,2,\cdots,r, j=1,2,\cdots,n_i$$

$$\varepsilon_{ij} \sim N\left(0,\sigma^2\right)，且\ \varepsilon_{ij}\ 独立，$$

其中 μ 为总平均，α_i 是第 i 个水平的效应，ε_{ij} 是随机误差。若 $n_1 = n_2 = \cdots = n_r$，称模型是平衡的，否则称为非平衡的。

我们的目的是要比较因素 A 的 r 个水平的京郊是否有显著差异，这可归结为检验假设 $H_0: \alpha_1 = \alpha_2 = \cdots = \alpha_r \leftrightarrow H_1: \alpha_1, \alpha_2, \cdots, \alpha_r$ 不全相等。

如果 H_0 被拒绝，则说明因素 A 的各水平的效应之间有显著的差异，否则，差异不明显。

按照方差分析的思想，将总离差平方和分解为二部分，即

$$SS_T = SS_E + SS_A$$

其中

$$SS_T = \sum_{i=1}^{r}\sum_{j=1}^{n_i}\left(X_{ij} - \bar{X}\right)^2, \bar{X} = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{n_i}X_{ij},$$

$$SS_E = \sum_{i=1}^{r}\sum_{j=1}^{n_i}\left(X_{ij} - \overline{X_i}\right)^2, \overline{X_i} = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{n_i}X_{ij},$$

$$SS_A = \sum_{i=1}^{r}\sum_{j=1}^{n_i}\left(X_i - \bar{X}\right)^2$$

这里称 SS_T 为总离差平方和(或称总变差)，它是所有数据 X_{ij} 与总平均值 $\bar{X}$ 之差的平方和，描绘所有观察数据的离散程度；SS_E 为误差平方和(或组内平方和)，是对固定的 i，观测值 $X_{i1}, X_{i2}, \cdots, X_{in}$ 之间的差异大小的度量。SS_A 为因素 A 的效应平方(和或组间平方和)，表示因子 A 各水平下的样本均值和总平均值之差的平方和。

可以证明，当 H_0 成了时

$$\frac{SS_E}{\sigma^2} \sim \chi^2(n-r), \frac{SS_A}{\sigma^2} \sim \chi^2(n-1),$$

且 SS_A 与 SS_E 独立。于是

$$F = \frac{SS_A/(r-1)}{SS_E/(n-1)} \sim F(r-1, n-1)$$

若 $F > F_\alpha(r-1, n-r)$，则拒绝原假设，认为因素 A 的 r 个水平有显著差异，反之“接受”原假设。这也可以通过检验的 p 值来决定是接受还是拒绝原假设 H_0。

2.2. 对低收入家庭学生百分比建立数学模型

我们对 X(低收入家庭学生的百分比)划分为 1994 年份和 2004 年份两部分，设它们分别为 A_1, A_2，均值分别为 $\overline{X_1}, \overline{X_2}$。由 R 程序输出结果为：

summary(aov.mis)

	Df	Sum Sq	Mean Sq	F value	Pr (>F)
A	1	301	300.9	0.828	0.365
Residuals	120	43606	363.4		

说明：上述结果中，Df 表示自由度；sum Sq 表示平方和；Mean Sq 表示均方和；F value 表示 F 检验统计量的值，即 F 比；Pr (>F)表示检验的 p 值；A 就是因素 A；Residuals 为残差。

可以看出 P = 0.365 > 0.05，说明不能拒绝原假设，即认为两个年份低收入家庭学生百分比没有较大的显著差异。

通过函数 plot( )绘图可直观描述两个年份低收入家庭学生百分比之间的差异，R 程序中运行：plot(miscellany$X~miscellany$A)。

得到图 1，从图形上也可以看出，两个年份低收入家庭学生百分比没有较大的显著差异。

2.3. 对一年级学生重修率建立数学模型

我们对 Y(一年级学生重修率)划分为 1994 年份和 2004 年份两部分，设它们分别为 B_1, B_2，均值分别为 $\overline{Y_1}, \overline{Y_2}$。由 R 程序输出结果为：

summary(aov.mis)

	Df	Sum Sq	Mean Sq	F value	Pr (>F)
B	1	37.8	37.79	2.309	0.131
Residuals	120	1964.4	16.37		

可以看出 $F = 2.309 < F_{0.05}(2-1,122-2)$，或者 $p = 0.131 > 0.05$，说明没有理由拒绝原假设，即认为两个年份一年级学生重修率没有显著差异。

通过函数 plot( )绘图可直观描述两个年份一年级学生重修率之间的异同，R 程序中运行：plot(miscellany$Y~miscellany$B)

得到图 2，从图形上也可以看出，两个年份一年级学生重修率没有显著差异：

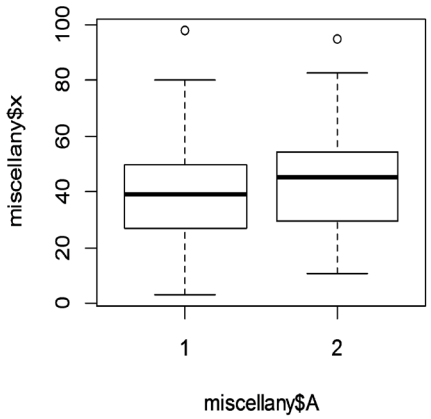

Figure 1. Percentage box chart for low-income families in two years
图 1. 两个年份低收入家庭学生百分比箱型图

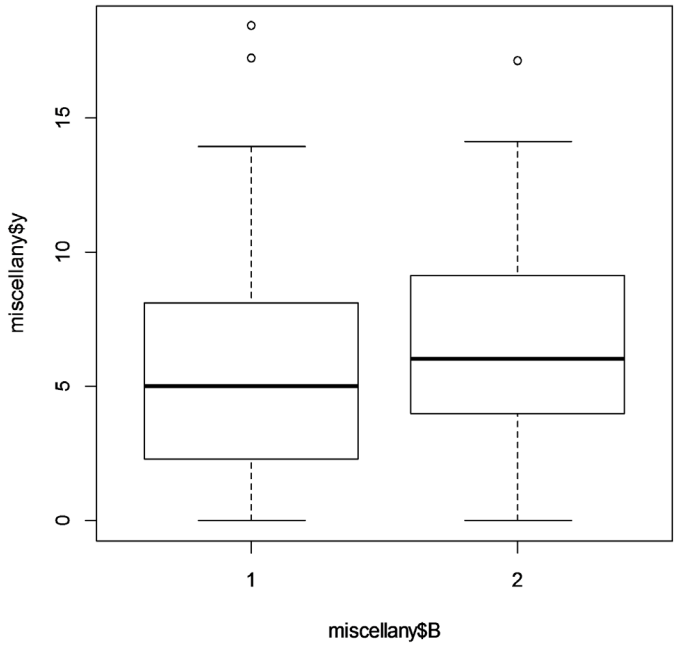

Figure 2. Two-year first-year student rework rate box chart
图 2. 两个年份一年级学生重修率箱型图

3. 协方差分析

协方差的原理

考虑在此我们想模拟一个响应变量的情况下，即在变量 Y 基础上作连续预测，X 和一个虚拟变量 d。假设 X 上的效果 Y 是线性的。这种情况的通常所说的最简单的版本协方差分析[2]，因为预测包括两个定量变量(即，X)和定性变量(即，d)所示。

重合回归线：最简单的模型在给定的情况是在该虚设变量 Y 上没有影响，也就是：

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

和回归线是完全相同的哑的两个值变量。

平行回归线：另一种模式来考虑这个情况是在该虚拟变量只产生 Y 中的哑变量的变化，即：

$$Y = \beta_0 + \beta_1 X + \beta_2 d + \varepsilon = \begin{cases} Y = \beta_0 + \beta_1 X + \varepsilon; \ d = 0 \\ Y = \beta_0 + \beta_1 X + \beta_2 + \varepsilon; \ d = 1 \end{cases}$$

在这种情况下，回归系数 β_2 测量在 Y，主要是由于哑变量改变虚拟变量。

回归线与截距相等，但不同的斜率：第三个模型考虑这种情况是在其中的虚拟变量 X 只改变 Y 大小的效果，也就是说：

$$Y = \beta_0 + \beta_1 X + \beta_3 dX + \varepsilon = \begin{cases} Y = \beta_0 + \beta_1 X + \varepsilon; \ d = 0 \\ Y = \beta_0 + (\beta_1 + \beta_3) X + \varepsilon; \ d = 1 \end{cases}$$

不相关的回归线：最普遍的模式是哑变量的变化能改变适当的虚拟变量 Y，也改变在 X 变量的基础上 Y 的大小。在这种情况下，合适的模型是：

$$Y = \beta_0 + \beta_1 X + \beta_2 d + \beta_3 dX + \varepsilon = \begin{cases} Y = \beta_0 + \beta_1 X + \varepsilon; \ d = 0 \\ Y = \beta_0 + \beta_2 + (\beta_1 + \beta_3) X + \varepsilon; \ d = 1 \end{cases}$$

在不相关的回归线模型，回归系数 β_2 能对 Y 提供一点变化主要是由于哑变量 d，然而回归系数 β_3 改变在变量 X 的基础上 Y 的大小也是由于哑变量 d。

4. 协方差分析方法的实例分析

4.1. 一年级学生重修率与低收入家庭学生百分比之间的关系

由于协方差分析是建立在线性回归分析和基本的方差分析的基础上，于是我们先对 Y (一年级学生重修率)和 X (低收入家庭学生的百分比)做线性处理，分析 X、Y 之间的联系。使用函数：lm(formula = y ~ 1 + x)

在 R 程序中运行结果如下：

Call:

lm(formula = y ~ 1 + x)

Residuals:

Min	1Q	Median	3Q	Max
−8.9845	−2.5072	−0.4184	1.8505	11.1067

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.91419	0.83836	3.476	0.000709 ***

x 0.07550 0.01823 4.141 6.47e−05 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.821 on 120 degrees of freedom

Multiple R-squared: 0.125, Adjusted R-squared: 0.1177

F-statistic: 17.14 on 1 and 120 DF, p-value: 6.472e-05

结论：从上述输出结果 p-值可以看出回归方程通过回归参数的检验与回归方程的检验，由此得到回归方程 $Y = 2.91419 + 0.0755X$，还可以对误差项独立同正态分布的假设进行检验[3]，运行结果见图 3。

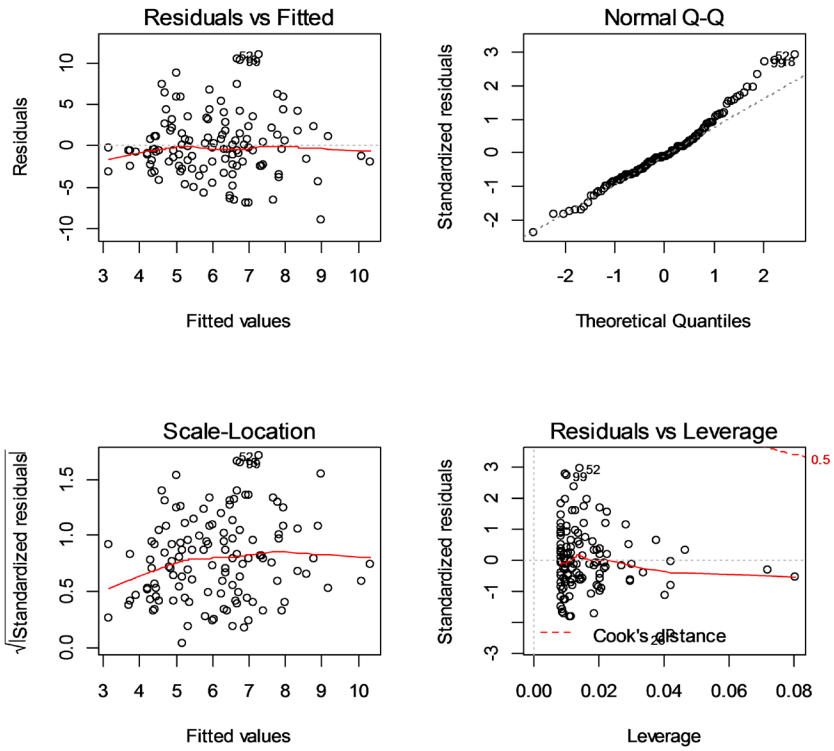

Figure 3. Test result correlation diagram
图 3. 检验结果相关图

1) Residual vs Ftted 为拟合值对残差的图形，可以看出，数据点都基本均匀地分布在直线 y = 0 的两侧，无明显趋势；

2) Normal QQ-plot 图中数据点分布趋于一条直线，说明残差是服从正态分布的；

3) Scale-Location 图显示了标准化残差(standardized residuals)的平方根的分布情况。最高点为残差最大值点；

4) Cook 距离(Cook’s distance)图显示了对回归的影响点。

由上面的分析我们得出，一年级学生重修率与低收入家庭学生百分比之间的关系可以用线性方程 $Y = 2.91419 + 0.0755X$ 表示。

4.2. 1994~1995 年和 2004~2005 年学生一年级重修率的差异

由于 2.1.2 节用单因子方差分析一年级学生重修率，结果显示两年份对这两变量没有较大的差异。于

是本小题主要是使用平行回归线模型，d 为构造的哑变量：

$$Y=\beta_0+\beta_1X+\beta_2d+\varepsilon=\begin{cases}Y=\beta_0+\beta_1X+\varepsilon;\ d=0\\ Y=\beta_0+\beta_1X+\beta_2+\varepsilon;\ d=1\end{cases}$$

在 R 程序中对两年份一年级学生重修率百分比分析结果如图 4：

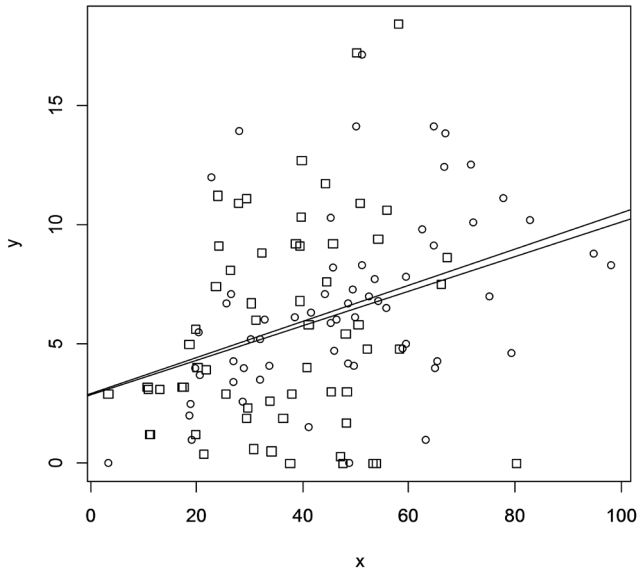

Figure 4. Two-year first-year student rework rate percentage scatter plot and linear regression line
图 4. 两年份一年级学生重修率百分比散点图及线性回归线

图形中三角形的散点代表 1994 年，加号的散点代表 2004 年，$d=0$ 为斜率较小的虚拟回归线，即 1994 年的散点图形虚拟回归：$Y=\beta_0+\beta_1X+\varepsilon$；$d=1$ 为斜率较大的虚拟回归线，即 2004 年的散点图形虚拟回归：$Y=\beta_0+d+\beta_1X+\varepsilon$。

在 R 程序中输出的回归结果：

Call:

lm(formula = y ~ 1+ x + d)

Residuals:

Min	1Q	Median	3Q	Max
−8.6768	−2.5451	−0.4769	1.6624	11.3469

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.84900	0.84995	3.352	0.001076 **
x	0.07248	0.01917	3.782	0.000245 ***
d	0.38311	0.72716	0.527	0.599274

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.832 on 119 degrees of freedom

Multiple R-squared: 0.127, Adjusted R-squared: 0.1124

F-statistic: 8.659 on 2 and 119 DF, p-value: 0.0003083

当 $d=0$，即 1994 年时，$y=0.07x+2.849$；当 $d=1$，即 2004 年时，$y=0.07x+3.232$。两模型之间的差异为 0.38311，即哑变量系数的估计值。因此，我们得出在 1994 年和 2004 年一年级学生重修率百分比的差值约为 38%。

4.3. 1994~1995 年和 2004~2005 年一年级学生重修率与低收入家庭学生百分比之间关系的差异

本小题主要采用不相关回归：

$$Y=\beta_0+\beta_2 d+\beta_1 X+\beta_3 dX+\varepsilon=\begin{cases}Y=\beta_0+\beta_1 X+\varepsilon;\ d=0\\ Y=\beta_0+\beta_2+\left(\beta_1+\beta_3\right)X+\varepsilon;\ d=1\end{cases}$$

在 R 程序中显示图形，图 5 中三角形的散点代表 1994 年，加号的散点代表 2004 年，$d=0$ 为斜率较小的虚拟回归线，即 1994 年的散点图形虚拟回归：$Y=\beta_0+\beta_1 X+\varepsilon$；$d=1$ 为斜率较大的虚拟回归线，即 2004 年的散点图形虚拟回归：$Y=\beta_0+\beta_2+\left(\beta_1+\beta_3\right)X+\varepsilon$。

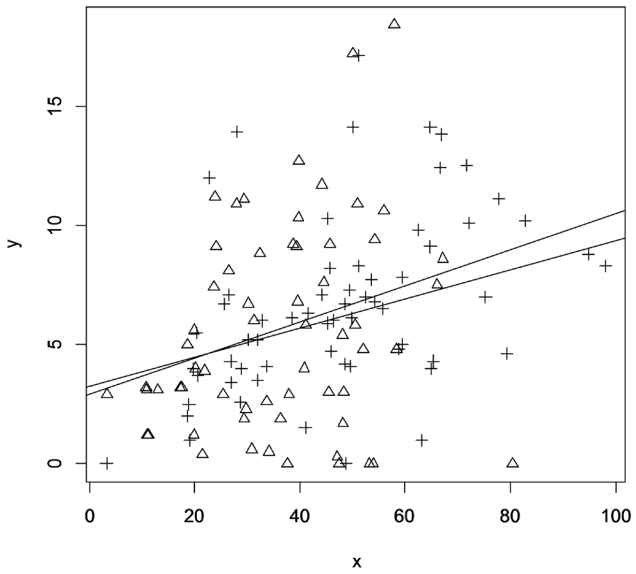

Figure 5. Repetition rate of first-year students in two years and percentage scatter plot and linear regression line of low-income students

图 5. 两年份一年级学生重修率与低收入家庭学生百分比散点图及线性回归线

对模拟回归线进行分析，R 程序运行结果为：

Call:

lm(formula =y ~ 1 + x + d + d * x)

Residuals:

Min	1Q	Median	3Q	Max
−8.1606	−2.6121	−0.5576	1.7495	11.6014

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.27194	1.22347	2.674	0.00855**

x 0.06080 0.03093 1.966 0.05167 .

d −0.38956 1.76109 −0.221 0.82532

x:d 0.01903 0.03949 0.482 0.63066

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.845 on 118 degrees of freedom

Multiple R-squared: 0.1288, Adjusted R-squared: 0.1066

F-statistic: 5.813 on 3 and 118 DF, p-value: 0.0009689

当 $d=0$ 即 1994 年时， $y=0.06x+3.27$ ；当 $d=1$ 即 2004 时， $y=0.07x+2.88$ 。

从上述的回归方程显示，我们初步认为在 1994 年和 2004 年这两年份时，一年级学生重修率和低收入家庭学生百分之间没有较大差异：分析 β_1 ，一个模型的 β_1 估计值是 6%，另一个模型 β_1 估计值是 7%，截距分别为 3.27，2.88。为了确保结果的可靠性，我们再一步对模型作偏 F 检验，在 R 程序中的结果为：

Analysis of Variance Table

Model 1: y ~ 1 + x

Model 2: y ~ 1 + x + d + d * x

Res. Df RSS Df Sum of Sq F Pr(>F)

1 120 1751.9

2 118 1744.4 2 7.512 0.2541 0.7761

$$F=\frac{\left(RSS_{(d=0)}-RSS_{(d=1)}\right)\Big/\left(Df_{(d=0)}-Df_{(d=1)}\right)}{RSS_{(d=1)}\Big/Df_{(d=1)}}$$

$$=\frac{(1751.9-1744.4)/(120-118)}{1744.4/118}\approx 0.2537$$

正如预期的结果那样，这两模型没有较大的差异，即年份没有对一年级学生重修率和低收入家庭学生百分比之间的关系产生较大影响。因此，我们宁愿选择重合线模型，而不是不相关的回归线模型。

参考文献

[1] 汤银才. R 语言与统计分析[M]. 北京: 高等教育出版社, 2008.

[2] 薛毅. 统计建模与 R 软件[M]. 北京: 清华大学出版社, 2007.

[3] Sheather, S. (2009) A Modern Approach to Regression with R. Springer, New York.