

第二章、经典单方程计量经济学模型：多元线性回归模型

一、内容提要

本章将一元回归模型拓展到了多元回归模型，其基本的建模思想与建模方法与一元的情形相同。主要内容仍然包括模型的基本假定、模型的估计、模型的检验以及模型在预测方面的应用等方面。只不过为了多元建模的需要，在基本假设方面以及检验方面有所扩充。

本章仍重点介绍了多元线性回归模型的基本假设、估计方法以及检验程序。与一元回归分析相比，多元回归分析的基本假设中引入了多个解释变量间不存在（完全）多重共线性这一假设；在检验部分，一方面引入了修正的可决系数，另一方面引入了对多个解释变量是否对被解释变量有显著线性影响关系的联合性 F 检验，并讨论了 F 检验与拟合优度检验的内在联系。

本章的另一个重点是将线性回归模型拓展到非线性回归模型，主要学习非线性模型如何转化为线性回归模型的常见类型与方法。这里需要注意各回归参数的具体经济含义。

本章第三个学习重点是关于模型的约束性检验问题，包括参数的线性约束与非线性约束检验。参数的线性约束检验包括对参数线性约束的检验、对模型增加或减少解释变量的检验以及参数的稳定性检验三方面的内容，其中参数稳定性检验又包括邹氏参数稳定性检验与邹氏预测检验两种类型的检验。检验都是以 F 检验为主要检验工具，以受约束模型与无约束模型是否有显著差异为检验基点。参数的非线性约束检验主要包括最大似然比检验、沃尔德检验与拉格朗日乘数检验。它们仍以估计无约束模型与受约束模型为基础，但以最大似然原理进行估计，且都适用于大样本情形，都以约束条件个数为自由度的 χ^2 分布为检验统计量的分布特征。非线性约束检验中的拉格朗日乘数检验在后面的章节中多次使用。

二、典型例题分析

例 1. 某地区通过一个样本容量为 722 的调查数据得到劳动力受教育的一个回归方程为

$$Y = 10.36 - 0.094X_1 + 0.131X_2 + 0.210X_3$$

$$R^2 = 0.214$$

式中，Y 为劳动力受教育年数， X_1 为该劳动力家庭中兄弟姐妹的个数， X_2 与 X_3 分别为母

亲与父亲受到教育的年数。问

(1) X_1 是否具有预期的影响？为什么？若 X_2 与 X_3 保持不变，为了使预测的受教育水平减少一年，需要 X_1 增加多少？

(2) 请对 X_2 的系数给予适当的解释。

(3) 如果两个劳动力都没有兄弟姐妹，但其中一个的父母受教育的年数为 12 年，另一个的父母受教育的年数为 16 年，则两人受教育的年数预期相差多少？

解答：

(1) 预期 X_1 对劳动者受教育的年数有影响。因此在收入及支出预算约束一定的条件下，子女越多的家庭，每个孩子接受教育的时间会越短。

根据多元回归模型偏回归系数的含义， X_1 前的参数估计值 -0.094 表明，在其他条件不变的情况下，每增加 1 个兄弟姐妹，受教育年数会减少 0.094 年，因此，要减少 1 年受教育的时间，兄弟姐妹需增加 $1/0.094=10.6$ 个。

(2) X_2 的系数表示当兄弟姐妹数与父亲受教育的年数保持不变时，母亲每增加 1 年受教育的机会，其子女作为劳动者就会预期增加 0.131 年的受教育机会。

(3) 首先计算两人受教育的年数分别为

$$10.36 + 0.131 \times 12 + 0.210 \times 12 = 14.452$$

$$10.36 + 0.131 \times 16 + 0.210 \times 16 = 15.816$$

因此，两人的受教育年限的差别为 $15.816 - 14.452 = 1.364$

例 2. 以企业研发支出 (R&D) 占销售额的比重为被解释变量 (Y)，以企业销售额 (X_1)

与利润占销售额的比重 (X_2) 为解释变量，一个有 32 容量的样本企业的估计结果如下：

$$Y = 0.472 + 0.32 \log(X_1) + 0.05X_2$$

$$(1.37) \quad (0.22) \quad (0.046)$$

$$R^2 = 0.099$$

其中括号中为系数估计值的标准差。

(1) 解释 $\log(X_1)$ 的系数。如果 X_1 增加 10%，估计 Y 会变化多少个百分点？这在经济上是一个很大的影响吗？

(2) 针对 R&D 强度随销售额的增加而提高这一备择假设，检验它不随 X_1 而变化的假设。分别在 5% 和 10% 的显著性水平上进行这个检验。

(3) 利润占销售额的比重 X_2 对 R&D 强度 Y 是否在统计上有显著的影响？

解答：

(1) $\log(X_1)$ 的系数表明在其他条件不变时， $\log(X_1)$ 变化 1 个单位， Y 变化的单位数，

即 $\Delta Y = 0.32 \Delta \log(X_1) \approx 0.32(\Delta X_1 / X_1) = 0.32 \times 100\%$ ，换言之，当企业销售 X_1 增长 100%时，企业研发支出占销售额的比重 Y 会增加 0.32 个百分点。由此，如果 X_1 增加 10%， Y 会增加 0.032 个百分点。这在经济上不是一个较大的影响。

(2) 针对备择假设 $H_1: \beta_1 > 0$ ，检验原假设 $H_0: \beta_1 = 0$ 。易知计算的 t 统计量的值为 $t = 0.32/0.22 = 1.468$ 。在 5% 的显著性水平下，自由度为 $32-3=29$ 的 t 分布的临界值为 1.699（单侧），计算的 t 值小于该临界值，所以不拒绝原假设。意味着 R&D 强度不随销售额的增加而变化。在 10% 的显著性水平下， t 分布的临界值为 1.311，计算的 t 值小于该值，拒绝原假设，意味着 R&D 强度随销售额的增加而增加。

(3) 对 X_2 ，参数估计值的 t 统计值为 $0.05/0.46 = 1.087$ ，它比在 10% 的显著性水平下的临界值还小，因此可以认为它对 Y 在统计上没有显著的影响。

例 3. 表 3-1 为有关经批准的私人住房单位及其决定因素的 4 个模型的估计量和相关统计值（括号内为 p -值）（如果某项为空，则意味着模型中没有此变量）。数据为美国 40 个城市的数据。模型如下：

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \mu$$

式中， Y 为实际颁发的建筑许可证数量， X_1 每平方英里的人口密度， X_2 自有房屋的均值（单位：百美元）， X_3 平均家庭的收入（单位：千美元）， X_4 为 1980~1992 年的人口增长百分比， X_5 失业率， X_6 人均交纳的地方税， X_7 人均缴纳的州税。

表 3-1

变量	模型 A	模型 B	模型 C	模型 D
C	813 (0.74)	-392 (0.81)	-1279 (0.34)	-973 (0.44)
X_1	0.075 (0.43)	0.062 (0.32)	0.042 (0.47)	
X_2	-0.855 (0.13)	-0.873 (0.11)	-0.994 (0.06)	-0.778 (0.07)
X_3	110.41 (0.14)	133.03 (0.04)	125.71 (0.05)	116.60 (0.06)
X_4	26.77 (0.11)	29.19 (0.06)	29.41 (0.001)	24.86 (0.08)
X_5	-76.55 (0.48)			
X_6	-0.061 (0.95)			
X_7	-1.006 (0.40)	-1.004 (0.37)		
RSS	4.763E+7	4.843E+7	4.962E+7	5.038E+7
R^2	0.349	0.338	0.322	0.312
$\hat{\sigma}^2$	1.488E+6	1.424E+6	1.418E+6	1.399E+6
AIC	1.776E+6	1.634E+6	1.593E+6	1.538E+6

(1) 检验模型 A 中的每一个回归系数在 10% 水平下是否为零（括号中的值为 p -值）。根据检验结果，你认为应该把变量保留在模型中还是去掉？

- (2) 在模型 A 中, 在 10% 水平下检验联合假设 $H_0: \beta_i = 0 (i=1, 5, 6, 7)$ 。说明被择假设, 计算检验统计值, 说明其在零假设条件下的分布, 拒绝或接受零假设的标准。说明你的结论。
- (3) 哪个模型是“最优的”? 解释你的选择标准。
- (4) 说明最优模型中有哪些系数的符号是“错误的”。说明你的预期符号并解释原因。确认其是否为正确符号。

解答:

(1) 直接给出了 p-值, 所以没有必要计算 t-统计值以及查 t 分布表。根据题意, 如果 p-值 < 0.10 , 则我们拒绝参数为零的原假设。

由于表中所有参数的 p-值都超过了 10%, 所以没有系数是显著不为零的。但由此去掉所有解释变量, 则会得到非常奇怪的结果。其实正如我们所知道的, 多元回归中去掉变量时一定要谨慎, 要有所选择。本例中, X_2 、 X_3 、 X_4 的 p-值仅比 0.1 稍大一点, 在略掉 X_5 、 X_6 、 X_7 的模型 C 中, 这些变量的系数都是显著的。

(2) 针对联合假设 $H_0: \beta_i = 0 (i=1, 5, 6, 7)$, 其对应的备择假设为 $H_1: \beta_i (i=1, 5, 6, 7)$ 中至少有一个不为零。检验假设 H_0 , 实际上就是参数的约束性检验, 非约束模型为模型 A, 约束模型为模型 D, 检验统计值为

$$F = \frac{(RSS_R - RSS_U)/(k_U - k_R)}{RSS_U/(n - k_U - 1)} = \frac{(5.038e + 7 - 4.763e + 7)/(7 - 3)}{(4.763e + 7)/(40 - 8)} = 0.462$$

显然, 在 H_0 假设下, 上述统计量满足 F 分布, 在 10% 的显著性水平下, 自由度为 (4, 32) 的 F 分布的临界值位于 2.09 和 2.14 之间。显然, 计算的 F 值小于临界值, 我们不能拒绝 H_0 , 所以 $\beta_i (i=1, 5, 6, 7)$ 是联合不显著的。

(3) 模型 D 中的 3 个解释变量全部通过显著性检验。尽管 R^2 值相对较小、残差平方和相对较大, 但相对来说其 AIC 值最低, 所以我们选择该模型为最优的模型。

(4) 随着收入的增加, 我们预期住房需要会随之增加。所以可以预期 $\beta_3 > 0$, 事实上其估计值确是大于零的。同样地, 随着人口的增加, 住房需求也会随之增加, 所以我们预期 $\beta_4 > 0$, 事实上其估计值也是如此。随着房屋价格的上升, 我们预期对住房的需求人数减少, 即我们预期 β_5 估计值的符号为负, 回归结果与直觉相符。出乎预料的是, 地方税与州税为不显著的。由于税收的增加将使可支配收入降低, 所以我们预期住房的需求将下降。虽然模型 A 是这种情况, 但它们的影响却非常微弱。

例 4、在经典线性模型基本假定下, 对含有三个自变量的多元回归模型:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \mu$$

你想检验的虚拟假设是 $H_0: \beta_1 - 2\beta_2 = 1$ 。

(1) 用 $\hat{\beta}_1, \hat{\beta}_2$ 的方差及其协方差求出 $\text{Var}(\hat{\beta}_1 - 2\hat{\beta}_2)$ 。

(2) 写出检验 $H_0: \beta_1 - 2\beta_2 = 1$ 的 t 统计量。

(3) 如果定义 $\beta_1 - 2\beta_2 = \theta$ ，写出一个涉及 β_0 、 θ 、 β_2 和 β_3 的回归方程，以便能直接得到 θ 估计值 $\hat{\theta}$ 及其标准误。

解答：

(1) 由数理统计学知识易知

$$\text{Var}(\hat{\beta}_1 - 2\hat{\beta}_2) = \text{Var}(\hat{\beta}_1) - 4\text{Cov}(\hat{\beta}_1, \hat{\beta}_2) + 4\text{Var}(\hat{\beta}_2)$$

(2) 由数理统计学知识易知

$$t = \frac{\hat{\beta}_1 - 2\hat{\beta}_2 - 1}{\text{se}(\hat{\beta}_1 - 2\hat{\beta}_2)}, \text{ 其中 } \text{se}(\hat{\beta}_1 - 2\hat{\beta}_2) \text{ 为 } (\hat{\beta}_1 - 2\hat{\beta}_2) \text{ 的标准差。}$$

(3) 由 $\beta_1 - 2\beta_2 = \theta$ 知 $\beta_1 = \theta + 2\beta_2$ ，代入原模型得

$$\begin{aligned} Y &= \beta_0 + (\theta + 2\beta_2)X_1 + \beta_2 X_2 + \beta_3 X_3 + \mu \\ &= \beta_0 + \theta X_1 + \beta_2(2X_1 + X_2) + \beta_3 X_3 + \mu \end{aligned}$$

这就是所需的模型，其中 θ 估计值 $\hat{\theta}$ 及其标准误都能通过对该模型进行估计得到。

例 5、对于涉及到三个变量 Y 、 X_1 、 X_2 的数据做以下回归：

$$\text{I} \quad Y_i = \alpha_0 + \alpha_1 X_{1i} + \mu_{1i}$$

$$\text{II} \quad Y_i = \beta_0 + \beta_1 X_{2i} + \mu_{2i}$$

$$\text{III} \quad Y_i = \gamma_0 + \gamma_1 X_{1i} + \gamma_2 X_{2i} + \mu_{3i}$$

问在什么条件下才能有 $\hat{\alpha}_1 = \hat{\gamma}_1$ 及 $\hat{\beta}_1 = \hat{\gamma}_2$ ，即多元回归与各自的一元回归所得的参数估计值相同。

解答：由回归模型 I 与 II 分别知：

$$\hat{\alpha}_1 = \frac{\sum x_{1i} y_i}{\sum x_{1i}^2}, \quad \hat{\beta}_1 = \frac{\sum x_{2i} y_i}{\sum x_{2i}^2}$$

对模型 III，令其样本回归模型的离差形式为

$$y_i = \gamma_1 x_{1i} + \gamma_2 x_{2i} + e_{3i}$$

求 $\sum e_i^2 = (y_i - \gamma_1 x_{1i} - \gamma_2 x_{2i})^2$

的最小值，可得如下正规方程组：

$$\sum y_i x_{1i} = \gamma_1 \sum x_{1i}^2 + \gamma_2 \sum x_{1i} x_{2i}$$

$$\sum y_i x_{2i} = \gamma_1 \sum x_{1i} x_{2i} + \gamma_2 \sum x_{2i}^2$$

解此方程组得

$$\hat{\gamma}_1 = \frac{(\sum y_i x_{1i})(\sum x_{2i}^2) - (\sum y_i x_{2i})(\sum x_{1i} x_{2i})}{\sum x_{1i}^2 \sum x_{2i}^2 - (\sum x_{1i} x_{2i})^2}$$

$$\hat{\gamma}_2 = \frac{(\sum y_i x_{2i})(\sum x_{1i}^2) - (\sum y_i x_{1i})(\sum x_{1i} x_{2i})}{\sum x_{1i}^2 \sum x_{2i}^2 - (\sum x_{1i} x_{2i})^2}$$

可见，当 $\sum x_{1i} x_{2i} = 0$ 时，即 X_1 与 X_2 完全线性无关时（正交），有 $\hat{\alpha}_1 = \hat{\gamma}_1$ 及 $\hat{\beta}_1 = \hat{\gamma}_2$ 。由此得多元回归的一个重要的结论：当各解释变量没有线性相关性时，多元回归中各解释变量的参数等于分别进行一元回归时解释变量的参数。

三、教材练习题及其参考解答

1、多元线性回归模型的基本假设是什么？试说明在证明最小二乘估计量的无偏性和有效性的过程中，哪些基本假设起了作用？

答：多元线性回归模型的基本假定仍然是针对随机误差项与针对解释变量两大类的假设。针对随机误差项的假设有：零均值、同方差、无序列相关且服从正态分布；针对解释变量的假设有：解释变量应具有非随机性，如果是随机的，则不能与随机误差项相关；各解释变量之间不存在（完全）线性相关关系。

在证明最小二乘估计量的无偏性中，利用了解释变量非随机或与随机误差项不相关的假定；在有效性的证明中，利用了随机误差项同方差且无序列相关的假定。

2、在多元线性回归分析中， t 检验与 F 检验有何不同？在一元线性回归分析中二者是否有等价的作用？

答：在多元线性回归分析中， t 检验常被用作检验回归方程中各个参数的显著性，而 F

检验则被用作检验整个回归关系的显著性；各解释变量联合起来对被解释变量有显著的线性关系，并不意味着每一个解释变量分别对被解释变量有显著的线性关系。在一元线性回归分析中，二者具有等价作用，因为二者都是对共同的假设——解释变量的参数等于零——进行检验。

3、为什么说对模型参数施加约束条件后，其回归的残差平方和一定不比未施加约束的残差平方和小？在什么样的条件下，受约束回归与无约束回归的结果相同？

答：对模型参数施加约束条件后，就限制了参数的取值范围，寻找到的参数估计值也是在此条件下使残差平方和达到最小，它不可能比未施加约束条件时找到的参数估计值使得残差平方达到的最小值还要小。但当约束条件为真时，受约束回归与无约束回归的结果就相同了。

4、在一项调查大学生一学期平均成绩（ Y ）与每周在学习（ X_1 ）、睡觉（ X_2 ）、娱乐（ X_3 ）与其他（ X_4 ）等各种活动所用时间的关系的研究中，建立如下回归模型：

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \mu$$

如果这些活动所用时间的总和为一周的总小时数 168。问：保持其他变量不变，而改变其中一个变量的说法是否有意义？该模型是否有违背基本假设的情况？如何修改此模型以使其更加合理。

解答：由于 $X_1 + X_2 + X_3 + X_4 = 168$ ，当其中一个变量变化时，至少有一个其他变量也得变化，因此，保持其他变量不变，而改变其中一个变量的说法是无意义的。

显然，由于四类活动的总和为一周的总小时数 168，表明四个 X 间存在完全的线性关系，因此违背了解释变量间不存在（完全）多重共线性的假设。

可以去掉其中的一个变量，如去掉代表“其他”活动的变量 X_4 ，则新构成的三变量模型更加合理。如这时 β_1 就测度了当其他两变量不变时，每周增加 1 小时的学习时间所带来的学习成绩的平均变化。这时，即使睡觉和娱乐的时间保持不变，也可以通过减少其他活动的时间来增加学习的时间。而这时三个变量间也不存在明显的共线性问题。

5、考虑下列两个模型：

$$\text{I} \quad Y_i = \alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_{2i} + u_i$$

$$\text{II} \quad Y_i - X_{1i} = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + v_i$$

(1) 证明: $\hat{\beta}_1 = \hat{\alpha}_1 - 1$, $\hat{\beta}_0 = \hat{\alpha}_0$, $\hat{\beta}_2 = \hat{\alpha}_2$ 。

(2) 证明: 两个模型的最小二乘残差相等, 即对任何 i , 有 $\hat{u}_i = \hat{v}_i$ 。

(3) 在什么条件下, 模型 II 的 R^2 小于模型 I 的 R^2 ?

解答: (1) 对模型 II 变形如下:

$$Y_i = \beta_0 + (\beta_1 + 1)X_{1i} + \beta_2 X_{2i} + v_i$$

因此, 在与模型 I 有相同的样本下进行 OLS 估计, 有

$$\hat{\alpha}_1 = \hat{\beta}_1 + 1, \quad \hat{\beta}_0 = \hat{\alpha}_0, \quad \hat{\beta}_2 = \hat{\alpha}_2$$

或
$$\hat{\beta}_1 = \hat{\alpha}_1 - 1, \quad \hat{\beta}_0 = \hat{\alpha}_0, \quad \hat{\beta}_2 = \hat{\alpha}_2$$

(2) 在 (1) 成立的条件下,

$$\begin{aligned}\hat{u}_i &= Y_i - \hat{\alpha}_0 - \hat{\alpha}_1 X_{1i} - \hat{\alpha}_2 X_{2i} \\ &= Y_i - \hat{\beta}_0 - (\hat{\beta}_1 + 1)X_{1i} - \hat{\beta}_2 X_{2i} \\ &= Y_i - X_{1i} - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i} = \hat{v}_i\end{aligned}$$

$$(3) \text{ 对模型 I, } R^2 = 1 - \frac{\sum \hat{u}_i^2}{\sum (Y_i - \bar{Y})^2}$$

$$\text{对模型 II, } R^2 = 1 - \frac{\sum \hat{v}_i^2}{\sum [(Y_i - X_{1i}) - (\bar{Y} - \bar{X}_2)]^2}$$

由 (2) 知 $\sum \hat{u}_i = \sum \hat{v}_i$, 故, 只有当 $\sum [(Y_i - X_{1i}) - (\bar{Y} - \bar{X}_2)]^2 < \sum (Y_i - \bar{Y})^2$ 时, 即模型 II 的总变差 (解释变量的离差平方和) 小于模型 I 的总变差 (解释变量的离差平方和) 时, 才会有模型 II 的 R^2 小于模型 I 的 R^2 。

6、考虑下列三个试验步骤

(1) 对 $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$ 进行回归

(2) 对 $X_{1i} = \alpha_0 + \alpha_1 X_{2i} + v_i$ 进行回归, 计算残差 $\hat{v}_i$

(3) 对 $Y_i = \gamma_0 + \gamma_1 \hat{v}_i + \gamma_2 X_{2i} + w_i$ 进行回归

试证明 $\hat{\beta}_1 = \hat{\gamma}_1$, 并直观地解释该结果。

证明: 由 (2) 计算残差 $\hat{v}_i$: $\hat{v}_i = X_{1i} - \hat{\alpha}_0 - \hat{\alpha}_1 X_{2i}$, 代入到 (3) 的回归中得:

$$Y_i = \gamma_0 + \gamma_1(X_{1i} - \hat{\alpha}_0 - \hat{\alpha}_1 X_{2i}) + \gamma_2 X_{2i} + w_i$$

或
$$Y_i = (\gamma_0 - \hat{\alpha}_0 \gamma_1) + \gamma_1 X_{1i} + (\gamma_2 - \hat{\alpha}_1 \gamma_1) X_{2i} + w_i$$

可见，模型形式与步骤（1）中的完全相同，因此必有 $\hat{\beta}_1 = \hat{\gamma}_1$ 。直观地看， v_i 测度的是 X_2 以外的因素对 X_1 的影响。因此对（3）中的模型来说， $\hat{v}_i$ 对 Y 的影响只能归结到 X_1 对 Y 的影响上来，与 X_2 无关。所以，（1）中模型的 β_1 与（3）中模型的 γ_1 都是测度排除了 X_2 后的 X_1 对 Y 的影响，二者的回归结果应是相等的。

7、考虑以下过原点回归

$$Y_i = \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + e_i$$

（1）求参数的 OLS 估计量；

（2）对该模型，是否仍有结论 $\sum e_i = 0$ ， $\sum e_i X_{1i} = 0$ ， $\sum e_i X_{2i} = 0$ 。

解：（1）根据最小二乘原理，需求适当的 $\hat{\beta}_1$ 、 $\hat{\beta}_2$ ，使得残差平方和最小：

$$\text{Min } \sum e_i^2 = \sum (Y_i - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i})^2$$

由微积分的知识，对上式分别关于 $\hat{\beta}_1$ 、 $\hat{\beta}_2$ 求偏导，并令导数值为零得如下正规方程组：

$$\sum (Y_i - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i}) X_{1i} = 0$$

$$\sum (Y_i - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i}) X_{2i} = 0$$

或

$$\hat{\beta}_1 \sum X_{1i}^2 + \hat{\beta}_2 \sum X_{1i} X_{2i} = \sum X_{1i} Y_i$$

$$\hat{\beta}_1 \sum X_{1i} X_{2i} + \hat{\beta}_2 \sum X_{2i}^2 = \sum X_{2i} Y_i$$

解得

$$\hat{\beta}_1 = \frac{(\sum Y_i X_{1i})(\sum X_{2i}^2) - (\sum Y_i X_{2i})(\sum X_{1i} X_{2i})}{\sum X_{1i}^2 \sum X_{2i}^2 - (\sum X_{1i} X_{2i})^2}$$

$$\hat{\beta}_2 = \frac{(\sum Y_i X_{2i})(\sum X_{1i}^2) - (\sum Y_i X_{1i})(\sum X_{1i} X_{2i})}{\sum X_{1i}^2 \sum X_{2i}^2 - (\sum X_{1i} X_{2i})^2}$$

（2）由（1）中的正规方程组知，对该模型，仍有

$$\sum e_i X_{1i} = 0$$

$$\sum e_i X_{2i} = 0$$

但不存在 $\sum e_i = 0$ 。即过原点的残差和不一定为零。

8、对多元线性回归模型 $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\mu}$ ，试证明随机误差项 $\boldsymbol{\mu}$ 的方差的无偏估计量为

$$\hat{\sigma}^2 = \frac{\mathbf{e}'\mathbf{e}}{n-k-1}。其中 \mathbf{e} 为相应样本回归模型的残差向量。$$

证：

由于被解释变量的估计值与观测值之间的残差

$$\begin{aligned}\mathbf{e} &= \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}} \\ &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\mu} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\mu}) \\ &= \boldsymbol{\mu} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\mu} \\ &= (\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\boldsymbol{\mu} \\ &= \mathbf{M}\boldsymbol{\mu}\end{aligned}$$

残差的平方和为：

$$\mathbf{e}'\mathbf{e} = \boldsymbol{\mu}'\mathbf{M}'\mathbf{M}\boldsymbol{\mu}$$

因为 $\mathbf{M} = (\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')$ 为对称等幂矩阵，即

$$\begin{aligned}\mathbf{M} &= \mathbf{M}' \\ \mathbf{M}^2 &= \mathbf{M}'\mathbf{M} = \mathbf{M}\end{aligned}$$

所以有

$$\begin{aligned}\mathbf{e}'\mathbf{e} &= \boldsymbol{\mu}'\mathbf{M}\boldsymbol{\mu} \\ E(\mathbf{e}'\mathbf{e}) &= E(\boldsymbol{\mu}'(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\boldsymbol{\mu}) \\ &= \sigma^2 \text{tr}(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}') \\ &= \sigma^2 (\text{tr}\mathbf{I} - \text{tr}(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')) \\ &= \sigma^2 (n - (k+1))\end{aligned}$$

其中符号“tr”表示矩阵的迹，其定义为矩阵主对角线元素的和。于是

$$\sigma^2 = \frac{E(\mathbf{e}'\mathbf{e})}{n-k-1}$$

以上过程既导出了随机误差项方差的估计量为

$$\hat{\sigma}^2 = \frac{\mathbf{e}'\mathbf{e}}{n-k-1}$$

也证明了该估计量是无偏估计量。

9、对多元线性回归模型 $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\mu}$ ，试证明普通最小二乘估计量 $\hat{\boldsymbol{\beta}}$ 具有最小方差性。

证：

设 β^* 是其他方法得到的关于 β 的线性无偏估计量：

$$\beta^* = C^* Y$$

其中， $C^* = C + D = (X'X)^{-1}X' + D$ ， D 为一固定矩阵，于是

$$\beta^* = C^* Y = C^* X\beta + C^* \mu$$

$$E(\beta^*) = C^* X\beta$$

β^* 的无偏性要求 $C^* X = I$ 。由于

$$C^* X = (X'X)^{-1}X'X + DX$$

于是， $C^* X = I$ 当且仅当 $DX = 0$ 。

β^* 的方差-协方差矩阵为：

$$\begin{aligned} \text{Cov}(\beta^*) &= E((\beta^* - \beta)(\beta^* - \beta)') \\ &= E[(C^* Y - \beta)(C^* Y - \beta)'] \\ &= E[(C^* \mu)(C^* \mu)'] \\ &= E[(X'X)^{-1}X' + D]\mu\mu'[X(X'X)^{-1} + D'] \\ &= \sigma^2[(X'X)^{-1}X'X(X'X)^{-1} + (X'X)^{-1}X'D' + DX(X'X)^{-1} + DD'] \\ &= \sigma^2(X'X)^{-1} + \sigma^2 DD' \\ &= \text{Cov}(\hat{\beta}) + \sigma^2 DD' \end{aligned}$$

DD' 为主对角线元素非负的对称矩阵，由此得 β^* 的方差大于或等于最小二乘估计量 $\hat{\beta}$ 的方差。

10、对下列模型：

$$\text{I} \quad Y_i = \alpha + \beta X_i + 2Z_i + u_i$$

$$\text{II} \quad Y_i = \alpha + \beta X_i - \beta Z_i + u_i$$

求出 β 的最小二乘估计值；并将结果与下面的三变量回归方程的最小二乘估计值作比较：

$$\text{III} \quad Y_i = \alpha + \beta X_i + \gamma Z_i + u_i$$

你认为哪一个估计值更好？

解答：将模型 I 改写成 $(Y_i - 2Z_i) = \alpha + \beta X_i + u_i$ ，则 β 的估计值为：

$$\hat{\beta} = \frac{\sum (x_i)(y_i - 2z_i)}{\sum (x_i)^2}$$

将模型 II 改写成 $Y_i = \alpha + \beta(X_i - Z_i) + u_i$ ，则 β 的估计值为：

$$\hat{\beta} = \frac{\sum (x_i - z_i)y_i}{\sum (x_i - z_i)^2}$$

对模型 III， β 的估计值为

$$\hat{\beta} = \frac{(\sum y_i x_i)(\sum z_i^2) - (\sum y_i z_i)(\sum x_i z_i)}{\sum x_i^2 \sum z_i^2 - (\sum x_i z_i)^2}$$

显然，模型 I 与模型 II 分别是模型 III 的参数在如下约束下的变形式：

$$\gamma = 2, \quad \gamma = -\beta$$

因此，如果限制条件正确，则三个回归结果相同。当然，从参数估计的表达式上看，模型 I 与模型 II 的回归算法更简洁。但如果限制条件不正确，则模型 I 与模型 II 的回归参数是有偏的。

11、下表给出三变量模型的回归结果：

方差来源	平方和 (SS)	自由度 (d.f.)	平方和的均值(MSS)
来自回归	65965	—	—
来自残差	—	—	—
总离差(TSS)	66042	14	

求：（1）样本容量 n ；残差平方和 RSS ；回归平方和 ESS 及残差平方和 RSS 的自由度。

（2）求拟合优度 R^2 及调整的拟合优度 $\bar{R}^2$

（3）检验假设： X_2 和 X_3 对 Y 无影响。应采用什么假设检验？为什么？

（4）根据以上信息，你能否确定 X_2 和 X_3 各自对 Y 的影响吗？

解答：

（1）样本容量为： $n = d.f. + 1 = 15$

$$RSS = TSS - ESS = 66042 - 65965 = 77$$

$$ESS \text{ 的自由度为: } d.f. = 14 - 2 = 12$$

$$RSS \text{ 的自由度为: } d.f. = n - 3 = 12$$

$$(2) R^2 = \frac{ESS}{TSS} = \frac{65965}{66042} = 0.9988$$

$$\bar{R}^2 = 1 - (1 - R^2) \frac{n-1}{n-k-1} = 1 - 0.0012 \times \frac{14}{12} = 0.9986$$

(4) 应该采用联合假设检验，即 F 检验，理由是只有这样做才能判断 X_2 、 X_3 一起是否对 Y 有影响。

(5) 不能。因为仅通过上述信息，可初步判断 X_2 、 X_3 联合起来对 Y 有线性影响，两者的变化解释了 Y 变化的 99.8%。但由于无法知道回 X_2 、 X_3 前参数的具体估计值，因此还无法判断它们各自对 Y 的影响有多大。

12、在一项对某社区家庭对某种消费品的消费需要调查中，得到表 3-3 的资料。请用手工与软件两种方式对该社区家庭对该商品的消费需求支出作二元线性回归分析。其中手工方式要求以矩阵表达式进行运算。

- (1) 估计回归方程的参数及随机误差项的方差 $\hat{\sigma}^2$ ，计算 R^2 及 $\bar{R}^2$ ；
- (2) 对方程进行 F 检验，对参数进行 t 检验，并构造参数 95% 的置信区间。
- (3) 如果商品单价变为 35 元，则某一月收入为 20000 元的家庭对其的消费支出估计是多少？构造该估计值的 95% 的置信区间。

表 3-3

序号	对某商品的 消费支出 Y (元)	商品单价 X_1 (元)	家庭月收 入 X_2 (元)	序号	对某商品的 消费支出 Y (元)	商品单价 X_1 (元)	家庭月收 入 X_2 (元)
1	591.9	23.56	7620	6	644.4	34.14	12920
2	654.5	24.44	9120	7	680	35.3	14340
3	623.6	32.07	10670	8	724	38.7	15960
4	647	32.46	11160	9	757.1	39.63	18000
5	674	31.15	11900	10	706.8	46.68	19300

解：(1) 以矩阵形式表达，二元样本回归方程为

$$Y = X\hat{\beta} + e$$

参数的估计值为

$$\hat{\beta} = (X'X)^{-1}(X'Y)$$

由于

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} 5.32536028 & -0.36302110 & 0.00053817 \\ -0.3630211 & 0.03381604 & -0.00005958 \\ 0.00053817 & -0.0005958 & 0.00000011 \end{pmatrix}, \quad \mathbf{X}'\mathbf{Y} = \begin{pmatrix} 6703.3 \\ 228956.63 \\ 89275178 \end{pmatrix}$$

于是

$$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Y}) = \begin{pmatrix} 626.509 \\ -9.79057 \\ 0.02862 \end{pmatrix}$$

根据随机误差项方差的估计式 $\hat{\sigma}^2 = \sum e_i^2 / (n - k - 1)$ 得到

$$\hat{\sigma}^2 = \mathbf{e}'\mathbf{e} / (n - k - 1)$$

而

$$\begin{aligned} \mathbf{e}'\mathbf{e} &= (\mathbf{Y} - \hat{\mathbf{Y}})'(\mathbf{Y} - \hat{\mathbf{Y}}) = (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \\ &= \mathbf{Y}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y} + \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} \\ &= \mathbf{Y}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y} + \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \\ &= \mathbf{Y}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}\hat{\boldsymbol{\beta}} \\ &= 4515072 - 4512955 = 2116.85 \end{aligned}$$

$$\text{故} \quad \hat{\sigma}^2 = \mathbf{e}'\mathbf{e} / (n - k - 1) = \frac{2116.85}{10 - 2 - 1} = 302.41$$

$$\text{又由于} \quad TSS = \sum (Y_i - \bar{Y})^2 = \sum (Y_i^2 - 2\bar{Y}Y_i + \bar{Y}^2) = \sum Y_i^2 - n\bar{Y}^2$$

$$\begin{aligned} &= \mathbf{Y}'\mathbf{Y} - n\bar{Y}^2 \\ &= 4515072 - 10 \times 449342.3 = 21648.74 \end{aligned}$$

$$\text{故} \quad R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\mathbf{e}'\mathbf{e}}{\mathbf{Y}'\mathbf{Y} - n\bar{Y}^2} = 1 - \frac{2116.85}{21648.74} = 0.9022$$

$$\bar{R}^2 = 1 - (1 - R^2) \frac{n-1}{n-k-1} = 0.8743$$

(2) 方程的总体线性性检验由下面的 F 检验进行

$$\begin{aligned} F &= \frac{ESS/k}{TSS/(n-k-1)} = \frac{(TSS - RSS)/k}{TSS/(n-k-1)} \\ &= \frac{(21648.74 - 2116.85)/2}{21648.74/(10-2-1)} = 32.29 \end{aligned}$$

在 5% 的显著性水平下, 自由度为 (2, 7) 的 F 分布的临界值为 $F_{0.05}(2, 7) = 4.74$, 可见

32.29 > 4.74, 表明方程的总体线性性显著成立。

由于

$$S_{\hat{\beta}_0} = \sqrt{\hat{\sigma}^2 c_{00}} = \sqrt{302.41 \times 5.32536} = \sqrt{1610.42} = 40.13$$

$$S_{\hat{\beta}_1} = \sqrt{\hat{\sigma}^2 c_{11}} = \sqrt{302.41 \times 0.033816} = \sqrt{10.2262} = 3.1978$$

$$S_{\hat{\beta}_2} = \sqrt{\hat{\sigma}^2 c_{22}} = \sqrt{302.41 \times 0.00000011} = \sqrt{0.00003408} = 0.005838$$

故常数项与 X_1 、 X_2 前参数的估计值的 t 检验值分别为

$$t = \frac{\hat{\beta}_0}{S_{\hat{\beta}_0}} = \frac{626.509}{40.13} = 15.612$$

$$t = \frac{\hat{\beta}_1}{S_{\hat{\beta}_1}} = \frac{-9.79057}{3.1978} = -3.062$$

$$t = \frac{\hat{\beta}_2}{S_{\hat{\beta}_2}} = \frac{0.028618}{0.005838} = 4.902$$

在 5% 的显著性水平下, 自由度为 8 的 t 分布的临界值为 $t_{0.025}(8) = 2.306$, 可见常数项及 X_1

与 X_2 的总体参数值均显著地异于零。

常数项、 X_1 与 X_2 参数的 95% 的置信区间分别为

$$\hat{\beta}_0 \pm t_{0.025} \times S_{\hat{\beta}_0} = 626.509 \pm 2.306 \times 40.13 \quad \text{或} \quad (533.97, \quad 719.05)$$

$$\hat{\beta}_1 \pm t_{0.025} \times S_{\hat{\beta}_1} = -9.791 \pm 2.306 \times 3.1978 \quad \text{或} \quad (-17.16, \quad -2.42)$$

$$\hat{\beta}_2 \pm t_{0.025} \times S_{\hat{\beta}_2} = 0.0286 \pm 2.306 \times 0.0058 \quad \text{或} \quad (0.0152, \quad 0.0421)$$

在 Eviews 软件下, 回归结果如下:

Dependent Variable: Y				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	626.5093	40.13010	15.61195	0.0000
X1	-9.790570	3.197843	-3.061617	0.0183
X2	0.028618	0.005838	4.902030	0.0017
R-squared	0.902218	Mean dependent var		670.3300
Adjusted R-squared	0.874281	S.D. dependent var		49.04504
S.E. of regression	17.38985	Akaike info criterion		8.792975
Sum squared resid	2116.847	Schwarz criterion		8.883751
Log likelihood	-40.96488	F-statistic		32.29408
Durbin-Watson stat	1.650804	Prob(F-statistic)		0.000292

(3) 将 $X_1 = 35$, $X_2 = 20000$ 代入回归方程, 可得:

$$Y = 626.51 - 9.7906 \times 35 + 0.0286 \times 20000 = 856.20(\text{元});$$

由于

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} 5.32536028 & -0.36302110 & 0.00053817 \\ -0.36302110 & 0.03381604 & -0.00005958 \\ 0.00053817 & -0.00005958 & 0.00000011 \end{pmatrix}$$

因此，取 $\mathbf{X}_0 = (1 \quad 35 \quad 20000)$ ，Y 均值的预测的标准差为

$$S_{\hat{Y}_0} = \sqrt{\hat{\sigma}^2 \mathbf{X}_0 (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}_0'} = \sqrt{302.41 \times 4.539} = \sqrt{1372.62} = 37.05$$

在 5% 的显著性水平下，自由度为 $10-2-1=7$ 的 t 分布的临界值为 $t_{0.025}(7) = 2.365$ ，于是

Y 均值的 95% 的预测区间为：

$$856.20 \pm 2.365 \times 37.05 \quad \text{或} \quad (768.58, 943.82)$$

同样容易得到 Y 个值的预测的标准差为

$$S_{\hat{Y}_0} = \sqrt{\hat{\sigma}^2 (1 + \mathbf{X}_0 (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}_0')} = \sqrt{302.41 \times 1.2661} = \sqrt{1675.03} = 40.93$$

于是，Y 个值的 95% 的预测区间为：

$$856.20 \pm 2.365 \times 40.93 \quad \text{或} \quad (759.41, 952.99)$$

13、表 3-4 列出了中国 2000 年按行业分的全部制造业国有企业及规模以上制造业非国有企业的工业总产值 Y，资产合计 K 及职工人数 L。设定模型：

$$Y = AK^{\alpha} L^{\beta} e^{\mu}$$

要求：（1）利用上述资料，进行回归分析。

（2）中国 2000 年的制造业总体呈现规模报酬不变状态吗？

表 3-4

序号	工业总产值 Y (亿元)	资产合计 K (亿元)	职工人数 L (万人)	序号	工业总产值 Y (亿元)	资产合计 K (亿元)	职工人数 L (万人)
1	3722.70	3078.22	113	17	812.70	1118.81	43
2	1442.52	1684.43	67	18	1899.70	2052.16	61
3	1752.37	2742.77	84	19	3692.85	6113.11	240
4	1451.29	1973.82	27	20	4732.90	9228.25	222
5	5149.30	5917.01	327	21	2180.23	2866.65	80
6	2291.16	1758.77	120	22	2539.76	2545.63	96
7	1345.17	939.10	58	23	3046.95	4787.90	222
8	656.77	694.94	31	24	2192.63	3255.29	163
9	370.18	363.48	16	25	5364.83	8129.68	244
10	1590.36	2511.99	66	26	4834.68	5260.20	145
11	616.71	973.73	58	27	7549.58	7518.79	138
12	617.94	516.01	28	28	867.91	984.52	46
13	4429.19	3785.91	61	29	4611.39	18626.94	218

14	5749.02	8688.03	254	30	170.30	610.91	19
15	1781.37	2798.90	83	31	325.53	1523.19	45
16	1243.07	1808.44	33				

解：（1）在 Eviews 软件下，选中 Quick\Estimate Question，在出现的对话框中输入“log(Y

C log(K) log(L)”，得如下回归结果：

Dependent Variable: Y				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	1.153994	0.727611	1.586004	0.1240
LOG(K)	0.609236	0.176378	3.454149	0.0018
LOG(L)	0.360796	0.201591	1.789741	0.0843
R-squared	0.809925	Mean dependent var		7.493997
Adjusted R-squared	0.796348	S.D. dependent var		0.942960
S.E. of regression	0.425538	Akaike info criterion		1.220839
Sum squared resid	5.070303	Schwarz criterion		1.359612
Log likelihood	-15.92300	F-statistic		59.65501
Durbin-Watson stat	0.793209	Prob(F-statistic)		0.000000

于是，样本回归方程为

$$\ln \hat{Y} = 1.154 + 0.609 \log K + 0.361 \ln L$$

$$(1.59) \quad (3.45) \quad (1.79)$$

$$R^2 = 0.8099 \quad \bar{R} = 0.7963 \quad F=59.66$$

给定显著性水平 5%，自由度为（2，28）的 F 分布的临界值为 $F_{0.05}(2,28) = 3.34$ ，因此总体上看， $\ln K$ 、 $\ln L$ 联合起来对 $\ln Y$ 有着显著的线性性影响。在 5% 的显著性水平下，自由度为 28 的 t 分布的临界值为 $t_{0.025}(28) = 2.048$ ，因此， $\ln K$ 的参数通过了该显著性水平下的 t 检验，但 $\ln L$ 未通过检验。如果设定显著性水平为 10%，t 分布的临界值为 $t_{0.05}(28) = 1.701$ ，这时 $\ln L$ 的参数通过了显著性水平检验。

$\bar{R} = 0.7963$ 表明，工业总产值对数值的 79.6% 的变化可以由资产合计的对数与职工的对数值的变化来解释，但仍有 20.4% 的变化是由其他因素的变化影响的。

（2）从上述回归结果看， $\hat{\alpha} + \hat{\beta} = 0.97 \approx 1$ ，即资产与劳动的产出弹性之和近似为 1，表明中国制造业在 2000 年基本呈现规模报酬不变的状态。下面进行参数的约束检验，检验的零假设为 $H_0: \alpha + \beta = 1$ 。

如果原假设为真，则可估计如下模型：

$$\ln(Y/L) = C + \alpha \ln(K/L) + \mu$$

通过上述资料，估计结果如下：

Dependent Variable: Y				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	1.026048	0.596769	1.719339	0.0962
LOG(K/L)	0.608141	0.173590	3.503324	0.0015
R-squared	0.297366	Mean dependent var	3.100040	
Adjusted R-squared	0.273138	S.D. dependent var	0.491331	
S.E. of regression	0.418891	Akaike info criterion	1.159927	
Sum squared resid	5.088613	Schwarz criterion	1.252443	
Log likelihood	-15.97888	F-statistic	12.27328	
Durbin-Watson stat	0.846460	Prob(F-statistic)	0.001511	

容易看出，该估计方程通过了 F 检验与参数的 t 检验。

在原假设为真的条件下，

$$F = \frac{(RSS_R - RSS_U)/1}{RSS_U/(31-2-1)} = \frac{(5.0886 - 5.0703)}{5.0703/28} = 0.1011$$

在 5% 的显著性水平下，自由度为(1, 28)的 F 分布的临界值为 4.20。0.1011 < 4.20，不拒绝原假设，表明 2000 年中国制造业呈现规模报酬不变的状态。

在 Eviews 软件中，当估计完第一个模型后，选中 View\Coefficient Test\Wald Coefficient Restrictions，在出现的对话框中输入 “C(2)+C(3)=1”，单击 OK 后则出现

Null Hypothesis: C(2)+C(3)=1			
F-statistic	0.101118	Probability	0.752851
Chi-square	0.101118	Probability	0.750492

因此，得出的结论仍是不拒绝原假设，原假设为真。

四、补充练习题

3-1. 解释下列概念：

- | | |
|-------------|--------------------------|
| 1) 多元线性回归模型 | 2) 偏回归系数 |
| 3) 正规方程组 | 4) 调整的多元可决系数 $\bar{R}^2$ |
| 5) 多重共线性 | 6) 联合假设检验 |
| 7) 受约束回归 | 8) 无约束回归 |

3-2. 观察下列方程并判断其变量是否呈线性？系数是否呈线性？或都是？或都不是？

- 1) $Y_i = \beta_0 + \beta_1 X_i^3 + \varepsilon_i$
- 2) $Y_i = \beta_0 + \beta_1 \log X_i + \varepsilon_i$
- 3) $\log Y_i = \beta_0 + \beta_1 \log X_i + \varepsilon_i$

$$4) \quad Y_i = \beta_0 + \beta_1(\beta_2 X_i) + \varepsilon_i$$

$$5) \quad Y_i = \frac{\beta_0}{\beta_1 X_i} + \varepsilon_i$$

$$6) \quad Y_i = 1 + \beta_0(1 - X_i^{\beta_1}) + \varepsilon_i$$

$$7) \quad Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}/10 + \varepsilon_i$$

3-3. 多元线性回归模型与一元线性回归模型有哪些区别？

3-4. 为什么说最小二乘估计量是最优的线性无偏估计量？多元线性回归最小二乘估计的正规方程组，能解出唯一的参数估计的条件是什么？

3-5. 请说明区间估计的含义。

3-6. 什么是正规方程组？分别用非矩阵形式和矩阵形式写出模型：

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \mu_i, \quad i = 1, 2, \dots, n$$

的正规方程组及其推导过程。

3-7. 为什么从计量经济学模型得到的预测值不是一个确定的值？预测值的置信区间和置信度的含义是什么？在相同的置信度下如何才能缩小置信区间？

3-8. 设有模型： $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \mu$ ，试在下列条件下：

$$(1) \quad \beta_1 + \beta_2 = 1$$

$$(2) \quad \beta_1 = \beta_2$$

分别求出 β_1 和 β_2 的最小二乘估计量。

3-9. 假设要求你建立一个计量经济模型来说明在学校跑道上慢跑半小时或半小时以上的人数，以便决定是否修建第二条跑道以满足所有的锻炼者。你通过整个学年收集数据，得到两个可能的解释性方程：

$$\text{I} \quad \hat{Y} = 125.0 - 15.0X_1 - 1.0X_2 + 1.5X_3 \quad \bar{R}^2 = 0.75$$

$$\text{II} \quad \hat{Y} = 123.0 - 14.0X_1 + 5.5X_2 - 3.7X_4 \quad \bar{R}^2 = 0.73$$

其中： Y 为某天慢跑者的人数， X_1 为该天的降雨量（毫米）， X_2 为该天日照时间（小时数）， X_3 为该天的最高温度（华氏温度）， X_4 为第二天需交学期论文的班级数。请回答下列问题：

(1) 这两个方程你认为哪个更合理些，为什么？

(2) 为什么用相同的数据去估计相同变量的系数却得到不同的符号?

3-10. 有人以校园内食堂每天卖出的盒饭数量作为被解释变量, 盒饭价格、气温、附近餐厅的盒饭价格、学校当日的学生数量作为解释变量, 进行回归分析; 假设你看到如下的回归结果 (括号中是标准差), 但并不知道各解释变量是哪一项。试判定每项结果对应着哪一个变量, 说明理由。

$$\hat{Y}_i = 10.6 + 28.4X_{1i} + 12.7X_{2i} + 0.61X_{3i} - 5.9X_{4i}$$

(2.6) (6.3) (0.61) (5.9)

$$\bar{R}^2 = 0.63 \quad n = 35$$

3-11. 下面给出依据 15 个观察值计算得到的数据:

$$\bar{Y} = 367.693, \quad \bar{X}_1 = 402.760, \quad \bar{X}_2 = 8.0, \quad \sum y_i^2 = 66042.269$$

$$\sum x_{1i}^2 = 84855.096, \quad \sum x_{2i}^2 = 280.0, \quad \sum y_i x_{1i} = 74778.346$$

$$\sum y_i x_{2i} = 4250.9, \quad \sum x_{1i} x_{2i} = 4796.0$$

其中小写字母代表了各值与其样本均值的离差。要求:

- (1) 估计 β_0 、 β_1 、 β_2 三个多元回归系数, 求出 R^2 与 $\bar{R}^2$?
- (2) 求 $\hat{\beta}_1$ 、 $\hat{\beta}_2$ 的标准差, 并估计 β_1 、 β_2 在 95% 置信度下的置信区间;
- (3) 在显著性水平 $\alpha = 5\%$ 下, 检验估计的每个回归系数的统计显著性;
- (4) 在 $\alpha = 5\%$ 下检验假设: 所有的参数都为零。

3-12. 一个关于个人收入与物价水平及失业率的关系的回归方程如下 (括号内为估计标准差):

$$\hat{W}_i = 8.562 + 0.364P_i + 0.004P_{i-1} - 2.560U_i$$

(0.080) (0.072) (0.658)

$$n = 19 \quad R^2 = 0.873$$

其中: W 为 t 年的每位雇员的工资和薪水, P 为 t 年的物价水平, U 为 t 年的失业率。

要求: (1) 对个人收入估计的斜率系数进行假设检验;

(2) 讨论 P_{t-1} 在理论上的正确性, 对本模型的正确性进行讨论; P_{t-1} 是否应从方程中删除? 为什么?

3-13. 参考补充练习题 2-18 给出的数据, 要求:

- (1) 建立一个多元回归模型，解释 MBA 毕业生的平均初职工资，并且求出回归结果；
- (2) 如果模型中包括了 GPA 和 GMAT 分数这两个解释变量，先验地，你可能会遇到什么问题，为什么？
- (3) 如果学费这一变量的系数为正、并且在统计上是显著的，是否表示进入最昂贵的商业学校是值得的。学费这个变量可用什么来代替？

3-14. 根据 100 对 (X_1, Y) 的观察值计算出：

$$\sum x_1^2 = 12 \quad \sum x_1 y = -9 \quad \sum y^2 = 30$$

要求：

- (1) 求出一元模型 $Y = \beta_0 + \beta_1 X_1 + u$ 中的 β_1 的 OLS 估计量及其相应的标准差的估计量；

(2) 后来发现 Y 还受 X_2 的影响，于是将一元模型改为二元模型

$$Y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + v$$

收集 X_2 的相应观察值并计算出：

$$\sum x_2^2 = 6 \quad \sum x_2 y = 8 \quad \sum x_1 x_2 = 2$$

求二元模型中的 α_1, α_2 的 OLS 估计量及其相应的标准差估计量；

- (3) 一元模型中的 $\hat{\beta}_1$ 与二元模型中的 $\hat{\alpha}_1$ 是否相等？为什么？

3-15. 考虑以下预测的回归方程：

$$\hat{Y}_t = -120 + 0.10F_t + 5.33R_t \quad \bar{R}^2 = 0.50$$

其中： Y_t 为第 t 年的玉米产量（吨/亩）， F_t 为第 t 年的施肥强度（公斤/亩）， R_t 为第 t 年的降雨量（毫米）。要求回答下列问题：

- (1) 从 F 和 R 对 Y 的影响方面，说出本方程中系数 0.10 和 5.33 的含义；
- (2) 常数项 -120 是否意味着玉米的负产量可能存在？
- (3) 假定 β_F 的真实值为 0.40，则估计值是否有偏？为什么？
- (4) 假定该方程并不满足所有的古典模型假设，即并不是最佳线性无偏估计值，则是否意味着 β_R 的真实值绝对不等于 5.33？为什么？

3-16. 已知数据如表 3-5。要求：

(1) 先根据表中数据估计以下回归模型的方程（只估计参数不用估计标准差）：

$$Y_i = \alpha_0 + \alpha_1 X_{1i} + \mu_{1i}$$

$$Y_i = \lambda_0 + \lambda_2 X_{2i} + \mu_{2i}$$

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \mu_i$$

(2) 回答下列问题： $\alpha_1 = \beta_1$ 吗？为什么？ $\lambda_2 = \beta_2$ 吗？为什么？

表 3-5

Y	X_1	X_2
1	1	10
3	2	9
8	3	5
15	4	1
28	5	-6

3-17、表 3-6 列出了某地区鸡肉家庭人均年消费量（Y）与家庭月平均收入（X）、鸡肉价格（ P_1 ）、猪肉价格（ P_2 ）与牛肉价格（ P_3 ）的相关数据。要求：

(1) 求出该地区关于家庭鸡肉消费需求的如下模型：

$$\ln Y = \beta_0 + \beta_1 \ln X + \beta_2 \ln P_1 + \beta_2 \ln P_2 + \beta_3 \ln P_3 + \mu$$

(2) 请分析，鸡肉的家庭消费需求是否受猪肉及牛肉价格的影响？

表 3-6

	Y (公斤)	X (元)	P_1 (元/公斤)	P_2 (元/公斤)	P_3 (元/公斤)
1980	2.78	397	4.22	5.07	7.83
1981	2.99	413	3.81	5.2	7.92
1982	2.98	439	4.03	5.4	7.92
1983	3.08	459	3.95	5.53	7.92
1984	3.12	492	3.73	5.47	7.74
1985	3.33	528	3.81	6.37	8.02
1986	3.56	560	3.93	6.98	8.04
1987	3.64	624	3.78	6.59	8.39
1988	3.67	666	3.84	6.45	8.55
1989	3.84	717	4.01	7	9.37
1990	4.04	768	3.86	7.32	10.61
1991	4.03	843	3.98	6.78	10.48
1992	4.18	911	3.97	7.91	11.4
1993	4.04	931	5.21	9.54	12.41

1994	4.07	1021	4.89	9.42	12.76
1995	4.01	1165	5.83	12.35	14.29
1996	4.27	1349	5.79	12.99	14.36
1997	4.41	1449	5.67	11.76	13.92
1998	4.67	1575	6.37	13.09	16.55
1999	5.06	1759	6.16	12.98	20.33
2000	5.01	1994	5.89	12.8	21.96
2001	5.17	2258	6.64	14.1	22.16
2002	5.29	2478	7.04	16.82	23.26

五、补充练习参考答案

3-1. 答：

(1) 在现实经济活动中往往存在一个变量受到其他多个变量的影响的现象，表现为在线性回归模型中有多个解释变量，这样的模型被称为多元线性回归模型，多元指多个解释变量。

(2) 在多元回归模型中，每一个解释变量前的参数即为偏回归系数，它测度了当其他解释变量保持不变时，该变量增加 1 个单位对被解释变量带来的平均影响程度。

(3) 正规方程组指采用 OLS 法估计线性回归模型时，对残差平方和关于各参数求偏导，并令偏导数为零后得到的一组方程，其矩阵形式为 $\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{Y}$ 。

(4) 调整的多元可决系数 $\bar{R}^2$ ，又称多元判定系数，是一个用于描述伴随模型中解释变量的增加、多个解释变量对被解释变量的联合影响程度的量。它与 R^2 有如下关系：

$$\bar{R}^2 = 1 - (1 - R^2) \frac{n-1}{n-k}。$$

(5) 多重共线性是多元回归中特有的一个概念，指多个解释变量间存在线性相关的情形。如果存在完全的线性相关性，则模型的参数就无法求出，OLS 回归无法进行。

(6) 联合假设检验是相对于单个假设检验来说的，指假设检验中的假设有多个，不止一个。如多元回归中的方程的显著性检验就是一个联合假设检验，而每个参数的 t 检验就是单个假设检验。

(7) 在实际经济活动中，常常需要根据经济理论对模型中变量的参数施加一定的约束条件，对模型参数施加约束条件后进行回归，称为受约束回归。

(8) 无约束回归是与受约束回归相对的一个概念，无需对模型中变量的参数施加约束条

件进行的回归称为无约束回归。

3-2. 答:

1)、2)、3) 变量非线性、系数线性;

4) 变量线性、系数非线性;

5)、6)、7) 变量、系数均为非线性。

3-3. 答:

多元线性回归模型与一元线性回归模型的区别表现在如下几方面: 一是解释变量的个数不同; 二是模型的经典假设不同, 多元线性回归模型比一元线性回归模型多了“解释变量之间不存在线性相关关系”的假定; 三是多元线性回归模型的参数估计式的表达更为复杂。

3-4. 在多元线性回归模型中, 在模型满足经典假定的条件下, 参数的最小二乘估计量具有线性性、无偏性以及最小方差性, 所以被称为最优线性无偏估计量 (BLUE)。

对于多元线性回归最小二乘估计的正规方程组, 能解出唯一的参数估计量的条件是 $(X'X)^{-1}$ 存在, 或者说各解释变量间不完全线性相关。

3-5. 答:

区间估计是指寻找一个以未知参数的点估计值为中心的区间, 以使真实的未知参数以一定的概率落入该区间内。区间估计是用来考察参数的点估计值与真实参数间的“接近”程度的。显然, 如果找到的区间越窄, 同时真实参数落入区间的概率越大, 则说明点估计值越“接近”真实的未知参数。

3-6. 答:

在 OLS 估计中, 由求残差平方和最小化而导出的一组含有待估参数的方程组称为正规方程组。

正规方程组的非矩阵形式如下:

$$\begin{cases} \Sigma(\hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \cdots + \hat{\beta}_k X_{ki}) = \Sigma Y_i \\ \Sigma(\hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \cdots + \hat{\beta}_k X_{ki}) X_{1i} = \Sigma Y_i X_{1i} \\ \Sigma(\hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \cdots + \hat{\beta}_k X_{ki}) X_{2i} = \Sigma Y_i X_{2i} \\ \vdots \\ \Sigma(\hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \cdots + \hat{\beta}_k X_{ki}) X_{ki} = \Sigma Y_i X_{ki} \end{cases}$$

正规方程组的矩阵形式如下:

$$(X'X)\hat{\beta} = X'Y$$

推导过程参见教材§3.2 节。

3-7. 答:

原因有二个方面。(1)模型中的参数估计量不确定, 它们随着抽样的不同而不同;(2)其他随机因素的影响, 即使找到了参数的真实值, 由于其他随机因素的影响, 也会使通过估计的模型得到的预测值具有不确定性。

正是由于预测值的不确定性, 得到的仅仅是预测值的一个估计值。真实的预测值仅以某一个置信度处于以该估计值为中心的一个区间中。预测值的置信区间指: 在给定 $(1-\alpha)$ 的置信度下, 被解释变量预测值 Y_0 的置信区间为:

$$\hat{Y}_0 - t_{\alpha/2} \times \hat{\sigma} \sqrt{1 + X_0(X'X)^{-1}X_0'} < Y_0 < \hat{Y}_0 + t_{\alpha/2} \times \hat{\sigma} \sqrt{1 + X_0(X'X)^{-1}X_0'}$$

预测值的置信度又称预测值的置信水平, 指预测值出现在上述区间的概率, 是表明预测值的可靠程度的量。

在相同的置信度下, 通过增加样本容量、提高模型的拟合优度和提高样本观测值的分散度可以达到缩小置信区间的目的。

3-8. 解:

(1) 由条件 $\beta_1 + \beta_2 = 1$, 容易将原模型变换为如下一元回归:

$$Y - X_2 = \beta_0 + \beta_1(X_1 - X_2) + \mu$$

因此

$$\hat{\beta}_1 = \frac{\sum (x_1 - x_2)(y - x_2)}{\sum (x_1 - x_2)^2}, \quad \hat{\beta}_2 = 1 - \frac{\sum (x_1 - x_2)(y - x_2)}{\sum (x_1 - x_2)^2}$$

式中, 小写字母表示对其均值的离差。

(2) 由条件 $\beta_1 = \beta_2$, 容易将原模型变换为如下一元回归:

$$Y = \beta_0 + \beta_1(X_1 + X_2) + \mu$$

因此

$$\hat{\beta}_1 = \frac{\sum (x_1 + x_2)y}{\sum (x_1 + x_2)^2}, \quad \hat{\beta}_2 = 1 - \frac{\sum (x_1 + x_2)y}{\sum (x_1 + x_2)^2}$$

3-9. 答:

(1) 方程 II 更合理些。原因是: 方程 II 中的参数估计值的符号与现实更接近些, 如与日

照的小时数同向变化, 天长则慢跑的人会多些; 与第二天需交学期论文的班级数成反向变化, 这一点在学校的跑道模型中是一个合理的解释变量。方程 I 相对来说不太合理, 因为日照小时数前的符号与预期的正号不相符, 而且所选的变量“日照小时数”与“该天的最高温度”有较强的相关性。

(2) 方程 I 和方程 II 中由于选择了不同的解释变量, 如方程 I 选择的是“该天的最高温度”而方程 II 选择的是“第二天需交学期论文的班级数”, 由此造成 X_2 与这两个变量之间的关系不同, 所以用相同的数据估计相同的变量得到不同的符号。其中变量“日照小时数”与“该天的最高温度”的较强相关性很大程度上导致了 X_2 的符号为负。

3-10. 答:

答案并不唯一, 猜测为: X_1 为学生数量, X_2 为附近餐厅的盒饭价格, X_3 为气温, X_4 为校园内食堂的盒饭价格; 理由是被解释变量应与学生数量成正比, 并且应该影响显著; 与本食堂盒饭价格成反比, 这与需求理论相吻合; 与附近餐厅的盒饭价格成正比, 因为彼此是替代品; 与气温的变化关系不是十分显著, 因为大多数学生不会因为气温升高不去食堂吃饭。

3-11. 解:

(1) 易知

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum y_i x_{1i} \sum x_{2i}^2 - \sum y_i x_{2i} \sum x_{1i} x_{1i}}{\sum x_{1i}^2 \sum x_{2i}^2 - \sum x_{1i} x_{2i} \sum x_{1i} x_{2i}} \\&= \frac{74778.346 \times 280 - 4250.9 \times 4796.0}{84855.096 \times 280 - 4796.0^2} \\&= \frac{550620}{757810} \\&= 0.7266 \\ \hat{\beta}_1 &= \frac{\sum y_i x_{2i} \sum x_{1i}^2 - \sum y_i x_{1i} \sum x_{1i} x_{2i}}{\sum x_{1i}^2 \sum x_{2i}^2 - \sum x_{1i} x_{2i} \sum x_{1i} x_{2i}} \\&= \frac{4250.9 \times 84855.096 - 74778.346 \times 4796.0}{84855.096 \times 280 - 4796.0^2} \\&= \frac{2073580}{757810} \\&= 2.7363 \\ \hat{\beta}_0 &= \bar{Y} - \hat{\beta}_1 \bar{X}_1 - \hat{\beta}_2 \bar{X}_2 \\&= 367.693 - 0.7266 \times 402.760 - 2.7363 \times 8.0 \\&= 53.1572\end{aligned}$$

$$\begin{aligned}\text{由于 } \sum e_i^2 &= \sum e_i e_i = \sum e_i (y_i - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i}) = \sum e_i y_i \\ &= \sum y_i^2 - \hat{\beta}_1 \sum y_i x_{1i} - \hat{\beta}_2 \sum y_i x_{2i}\end{aligned}$$

故

$$\begin{aligned}R^2 &= 1 - \frac{RSS}{TSS} = 1 - \frac{\sum y_i^2 - \hat{\beta}_1 \sum y_i x_{1i} - \hat{\beta}_2 \sum y_i x_{2i}}{\sum y_i^2} \\ &= \frac{\hat{\beta}_1 \sum y_i x_{1i} + \hat{\beta}_2 \sum y_i x_{2i}}{\sum y_i^2} = 0.9988\end{aligned}$$

$$\bar{R}^2 = 1 - (1 - R^2) \frac{n-1}{n-k} = 0.9986$$

(2) 如果记样本回归模型的离差形式为

$$y_i = \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + e_i$$

则容易知

$$\text{Var}(\hat{\beta}_i) = \sigma^2 (x'x)_{ii}^{-1} \quad i = 1, 2$$

$$\text{其中: } \mathbf{x}'\mathbf{x} = \begin{pmatrix} \sum x_{1i}^2 & \sum x_{1i}x_{2i} \\ \sum x_{1i}x_{2i} & \sum x_{2i}^2 \end{pmatrix}$$

由线性代数的知识易知:

$$(\mathbf{x}'\mathbf{x})^{-1} = \begin{pmatrix} \frac{\sum x_{2i}^2}{\sum x_{1i}^2 \sum x_{2i}^2 - (\sum x_{1i}x_{2i})^2} & -\frac{\sum x_{1i}x_{2i}}{\sum x_{1i}^2 \sum x_{2i}^2 - (\sum x_{1i}x_{2i})^2} \\ -\frac{\sum x_{1i}x_{2i}}{\sum x_{1i}^2 \sum x_{2i}^2 - (\sum x_{1i}x_{2i})^2} & \frac{\sum x_{1i}^2}{\sum x_{1i}^2 \sum x_{2i}^2 - (\sum x_{1i}x_{2i})^2} \end{pmatrix}$$

由于

$$\begin{aligned}\sigma^2 &= \frac{\sum e_i^2}{n-3} = \frac{\sum y_i^2 - \hat{\beta}_1 \sum y_i x_{1i} - \hat{\beta}_2 \sum y_i x_{2i}}{15-3} \\ &= \frac{66042.269 - 0.7266 \times 74778.346 - 2.7363 \times 4250.9}{12} \\ &= 6.3821\end{aligned}$$

于是

$$S_{\hat{\beta}_1} = \sqrt{\sigma^2 \frac{\sum x_{2i}^2}{\sum x_{1i}^2 \sum x_{2i}^2 - (\sum x_{1i} x_{2i})^2}}$$

$$= \sqrt{6.3821 \times \frac{280}{84855.096 \times 280 - 4796^2}} = 0.0486$$

$$S_{\hat{\beta}_2} = \sqrt{\sigma^2 \frac{\sum x_{1i}^2}{\sum x_{1i}^2 \sum x_{2i}^2 - (\sum x_{1i} x_{2i})^2}}$$

$$= \sqrt{6.3821 \times \frac{84855.096}{84855.096 \times 280 - 4796^2}} = 0.8454$$

样本容量为 $n=12$ ，查 5%显著性水平下自由度为 $15-2-1=12$ 的分布表的临界值为

$t_{0.025}(12) = 2.179$ ，因此 β_1 、 β_2 在 95%置信度下的置信区间分别为

$$0.7266 \pm 2.179 \times 0.0486, \quad \text{或} \quad 0.6207 \leq \beta_1 \leq 0.8325$$

$$2.7362 \pm 2.179 \times 0.8454, \quad \text{或} \quad 0.8942 \leq \beta_2 \leq 4.5784$$

(3) 针对每个参数都为零的假设，易有下面的 t 检验值

$$t_{\beta_1} = \frac{0.7266 - 0}{0.0486} = 14.9509$$

$$t_{\beta_2} = \frac{2.7363 - 0}{0.8454} = 3.2367$$

显然，两估计参数计算的 t 值大于临界值 2.179，拒绝它们各自为零的原假设。

(4) 需检验联合假设：两参数的值同时为零。易计算 F 统计量：

$$F = \frac{ESS}{RSS} = \frac{TSS - RSS}{RSS} = \frac{\sum y_i^2 - \sum e_i^2}{\sum e_i^2}$$

$$= \frac{32981.509}{6.6042} = 4994.0203$$

在 5%的显著性水平下，自由度为 (2, 12) 的 F 分布的临界值为 $F_{0.05}(2,12) = 3.89$ ，

计算的 F 值大于该临界值，所以拒绝原假设。

3-12. 解：

(1)对给定在 5%的显著水平下，可以进行 t 检验。计算的 t 值如下：

$$P_t \text{ 参数的 t 值: } 0.364/0.080=4.55$$

$$P_{t-1} \text{ 参数的 t 值: } 0.004/0.072=0.056$$

U_t 参数的 t 值: $-2.560/0.658=-3.89$

在 5% 的显著性水平下, 自由度为 $19-3-1=15$ 的 t 分布的临界值为 $t_{0.025}(15) = 2.131$, 比较得 P_t 、 U_t 的参数显著地不为零, 而不能拒绝 P_{t-1} 为零的假设。

(2) 回归式表明影响工资水平的主要是当期的物价水平, 前期的物价水平对它的影响并不大, 而失业率与工资水平呈反向变动也符合经济理论。故可将 P_{t-1} 从模型中删除。

3-13. 解:

(1) 利用补充练习题 2-18 的数据建立多元回归模型, Eviews 的回归结果如下:

Dependent Variable: ASP				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-310301.5	58715.98	-5.284788	0.0000
GPA	25676.01	22106.48	1.161470	0.2560
GMAT	442.8452	115.9222	3.820195	0.0007
X	1.083847	0.475704	2.278409	0.0312
R-squared	0.768182	Mean dependent var	68260.00	
Adjusted R-squared	0.741434	S.D. dependent var	18187.78	
S.E. of regression	9248.370	Akaike info criterion	21.22585	
Sum squared resid	2.22E+09	Schwarz criterion	21.41267	
Log likelihood	-314.3877	F-statistic	28.71905	
Durbin-Watson stat	1.453354	Prob(F-statistic)	0.000000	

因此, 有如下回归方程:

$$\hat{ASP} = -310301.5 + 25676.0GPA + 442.8GMAT + 1.1X$$

(2) 从回归方程看, GPA 和 GMAT 的系数为正, 因为 GPA 的分数和 GMAT 分数越高, 通常可以认为学生的水平越高, 从而可以获得更高的年薪;

(3) 从回归方程看, X 系数为正且显著, 表明进入最昂贵的商业学校是值得的; X 这个变量可以用商业学校的排名代替。

3-14. 解:

$$(1) \hat{\beta}_1 = \frac{\sum x_1 y}{\sum x_1^2} = \frac{-9}{12} = -0.75$$

$$\hat{\sigma}_u^2 = \frac{\sum y^2 - \hat{\beta}_1 \sum x_1 y}{n-2} = \frac{30 - 0.75 \times (-9)}{100-2} = 0.375$$

$$Var(\hat{\beta}_1) = \frac{\hat{\sigma}_u^2}{\sum x_1^2} = \frac{0.375}{12} = 0.03125$$

$$S_{\hat{\beta}_1} = \sqrt{0.03125} = 0.1768$$

(2)记样本的二元线性模型离差形式为

$$y = \hat{\alpha}_1 x_1 + \hat{\alpha}_2 x_2 + \hat{v}$$

则有

$$\hat{\alpha}_1 = \frac{\begin{vmatrix} \sum x_1 y & \sum x_1 x_2 \\ \sum x_2 y & \sum x_2^2 \end{vmatrix}}{\begin{vmatrix} \sum x_1^2 & \sum x_1 x_2 \\ \sum x_2 x_1 & \sum x_2^2 \end{vmatrix}} = \frac{\begin{vmatrix} -9 & 2 \\ 8 & 6 \end{vmatrix}}{\begin{vmatrix} 12 & 2 \\ 2 & 6 \end{vmatrix}} = \frac{-70}{68} = -1.0294$$

$$\hat{\alpha}_2 = \frac{\begin{vmatrix} \sum x_1^2 & \sum x_1 y \\ \sum x_2 x_1 & \sum x_2 y \end{vmatrix}}{\begin{vmatrix} \sum x_1^2 & \sum x_1 x_2 \\ \sum x_2 x_1 & \sum x_2^2 \end{vmatrix}} = \frac{\begin{vmatrix} 12 & -9 \\ 2 & 8 \end{vmatrix}}{\begin{vmatrix} 12 & 2 \\ 2 & 6 \end{vmatrix}} = \frac{114}{68} = 1.6765$$

$$\hat{\sigma}_v^2 = \frac{\sum y^2 - \hat{\alpha}_1 \sum x_1 y - \hat{\alpha}_2 \sum x_2 y}{n-3} = \frac{30 - [-1.0294 \times (-9)] - 1.6765 \times 8}{100-3} = 0.2665$$

由补充题 3-11 知:

$$\begin{aligned} S_{\hat{\alpha}_1} &= \sqrt{\hat{\sigma}_v^2 \frac{\sum x_2^2}{\sum x_1^2 \sum x_2^2 - (\sum x_1 x_2)^2}} \\ &= \sqrt{0.2665 \times \frac{6}{12 \times 6 - 2^2}} = \sqrt{0.0235} = 0.1533 \end{aligned}$$

$$\begin{aligned} S_{\hat{\alpha}_2} &= \sqrt{\hat{\sigma}_v^2 \frac{\sum x_1^2}{\sum x_1^2 \sum x_2^2 - (\sum x_1 x_2)^2}} \\ &= \sqrt{0.2665 \times \frac{12}{12 \times 6 - 2^2}} = \sqrt{0.0470} = 0.2169 \end{aligned}$$

(3)由上述计算结果知,一元模型中的 $\hat{\beta}_1$ 与二元模型中的 $\hat{\alpha}_1$ 不相等。主要原因在于 X_1 与 X_2 有相关性。如果它们是线性无关的,即 $\sum x_1 x_2 = 0$,则容易验证二者相等。

3-15. 解:

(1)在降雨量不变时,每亩增加一公斤肥料将使第 t 年的玉米产量增加 0.1 吨/亩;在每亩施肥量不变的情况下,每增加一毫米的降雨量将使第 t 年的玉米产量增加 5.33 吨/亩;

(2)在种地的一年中不施肥、也不下雨的现象同时发生的可能性极小,所以玉米的负产量不可能存在;事实上,这里截距项为负无实际意义。

(3) 如果 β_F 的真实值为 0.40，则表明该估计值与真值有偏误，但一般不说 0.1 是有偏的估计。理由是 0.1 是参数的一个估计值，而所谓估计量的有偏性是针对估计的期望来说的，即如果取遍所有可能的样本，这些参数估计值的平均值与 0.40 有偏误的话，就说估计是有偏的。

(4) 不一定。即便该方程并不满足所有的古典模型假设、不是最佳线性无偏估计值，也有可能得出的估计系数等于 5.33。

3-16. 解：

(1) 对于 $Y_i = \alpha_0 + \alpha_1 X_{1i} + u_{1i}$ 的估计结果如下：

Dependent Variable: Y				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-8.800000	3.942926	-2.231845	0.1118
X1	6.600000	1.188837	5.551644	0.0115
R-squared	0.911297	Mean dependent var	11.00000	
Adjusted R-squared	0.881729	S.D. dependent var	10.93161	
S.E. of regression	3.759433	Akaike info criterion	5.775587	
Sum squared resid	42.40000	Schwarz criterion	5.619363	
Log likelihood	-12.43897	F-statistic	30.82075	
Durbin-Watson stat	1.529245	Prob(F-statistic)	0.011526	

即有 $Y = -8.8 + 6.6X_1$

对于 $Y_i = \lambda_0 + \lambda_2 X_{2i} + u_{2i}$ 的估计结果如下：

Dependent Variable: Y				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	17.34075	0.481444	36.01820	0.0000
X2	-1.668618	0.069060	-24.16180	0.0002
R-squared	0.994887	Mean dependent var	11.00000	
Adjusted R-squared	0.993183	S.D. dependent var	10.93161	
S.E. of regression	0.902551	Akaike info criterion	2.921991	
Sum squared resid	2.443794	Schwarz criterion	2.765766	
Log likelihood	-5.304977	F-statistic	583.7925	
Durbin-Watson stat	1.948272	Prob(F-statistic)	0.000155	

即有 $Y = 17.34 - 1.66X_2$

对于 $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$ 的估计结果如下：

Dependent Variable: Y				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	21.92222	4.355258	5.033507	0.0373
X1	-1.177778	1.113026	-1.058176	0.4009
X2	-1.944444	0.269316	-7.219949	0.0186
R-squared	0.996722	Mean dependent var	11.00000	
Adjusted R-squared	0.993445	S.D. dependent var	10.93161	
S.E. of regression	0.885061	Akaike info criterion	2.877389	

Sum squared resid	1.566667	Schwarz criterion	2.643052
Log likelihood	-4.193473	F-statistic	304.1064
Durbin-Watson stat	2.912057	Prob(F-statistic)	0.003278

即有 $Y = 21.92 - 1.178X_1 - 1.944X_2$

(2)从上述回归结果可知, $\hat{\alpha}_1 = 6.6$, $\hat{\beta}_1 = -1.18$, 显然 $\hat{\alpha}_1 \neq \hat{\beta}_1$;

同样地, $\hat{\lambda}_2 = -1.67$, $\hat{\beta}_2 = -1.94$, 显然 $\hat{\lambda}_2 \neq \hat{\beta}_2$;

二元回归与分别对 X_1 与 X_2 所做的一元回归, 其相应的参数估计不相等, 主要原因在于 X_1 与 X_2 有很强的相关性, 事实上两者的相关系数为 $r=-0.9679$ 。

3-17 解

(1) Eviews 软件的回归结果如下:

Dependent Variable: LOG(Y)				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.731520	0.296947	-2.463467	0.0241
LOG(X)	0.345257	0.082565	4.181649	0.0006
LOG(P1)	-0.502122	0.109891	-4.569294	0.0002
LOG(P2)	0.146868	0.099006	1.483420	0.1553
LOG(P3)	0.087185	0.099852	0.873137	0.3941
R-squared	0.982474	Mean dependent var	1.361301	
Adjusted R-squared	0.978579	S.D. dependent var	0.187659	
S.E. of regression	0.027465	Akaike info criterion	-4.162123	
Sum squared resid	0.013578	Schwarz criterion	-3.915276	
Log likelihood	52.86441	F-statistic	252.2633	
Durbin-Watson stat	1.824820	Prob(F-statistic)	0.000000	

即有如下回归方程:

$$\ln \hat{Y} = -0.7315 + 0.3463X - 0.5021P_1 + 0.1469P_2 + 0.0872P_3$$

$$(-0.297) \quad (4.182) \quad (-4.569) \quad (1.483) \quad (0.873)$$

$$\bar{R}^2=0.9786 \quad F=252.26 \quad RSS=0.0135$$

容易验证, 家庭收入水平与鸡肉的价格对鸡肉的消费需求有显著的影响, 而猪肉价格及牛肉价格对鸡肉的消费影响不显著, 尤其是牛肉价格的影响很小。但方程总体的线性关系是显著的。

(2) 那么是否猪肉价格与牛肉价格真的对鸡肉的消费需求没有影响呢? 可检验如下原

假设: $H_0: \beta_3 = \beta_4 = 0$

对 Y 关于 X 、 P_1 做回归得

Dependent Variable: LOG(Y)				
Variable	Coefficient	Std. Error	t-Statistic	Prob.

C	-1.125797	0.088420	-12.73237	0.0000
LOG(X)	0.451547	0.024554	18.38966	0.0000
LOG(P1)	-0.372735	0.063104	-5.906668	0.0000
R-squared	0.980287	Mean dependent var	1.361301	
Adjusted R-squared	0.978316	S.D. dependent var	0.187659	
S.E. of regression	0.027634	Akaike info criterion	-4.218445	
Sum squared resid	0.015273	Schwarz criterion	-4.070337	
Log likelihood	51.51212	F-statistic	497.2843	
Durbin-Watson stat	1.877706	Prob(F-statistic)	0.000000	

即有如下回归方程：

$$\ln \hat{Y} = -1.1258 + 0.4515X - 0.3727P_1$$

$$(-12.73) \quad (18.38) \quad (-5.90)$$

$$\bar{R}^2 = 0.9783 \quad F = 497.28 \quad RSS = 0.0153$$

为了检验原假设，求如下的 F 统计量

$$F = \frac{(RSS_R - RSS_U)/2}{RSS_U/(23-4-1)} = \frac{(0.0153 - 0.0135)/2}{0.0135/18} = 1.2$$

在 5% 的显著性水平下，自由度为 (2, 18) 的 F 分布的临界值为 $F_{0.05}(2,18) = 3.55$ ，因此，没有理由拒绝原假设，即该地区猪肉与牛肉价格确实对家庭的鸡肉消费需求不产生显著影响。