Chapter 2 Problem 7

Problme

考虑储蓄函数
$sav = \beta_0 + \beta_1 inc + u, u = \sqrt{inc} \cdot e$

式中，e 为一个随机变量，且有 $\mathrm{E}(e) = 0$ 和 $\mathrm{Var}(e) = \sigma_e^2$ ，假设 e 独立于inc。
（i）证明：若 $\mathrm{E}(u \mid \mathrm{inc}) = 0$ ，则满足零条件均值这个关键假设（假设 SLR．4)。[提示：若 e 独立于 inc，则 $\mathrm{E}(e \mid \mathrm{inc}\ ) = \mathrm{E}(e)$ 。]
（ii）证明：若 $\mathrm{Var}(u \mid inc) = \sigma_r^2 inc$ ，则不满足同方差假设 SLR 5。特别地，sav 的方差随着 inc 而增加。 [提示：若 e 和 inc 独立，则 $\mathrm{Var}(e \mid inc) = \mathrm{Var}(e)$ 。]
（iii）讨论支持储蓄方差随着家庭收人而递增的证据。

Solution (i)

为了证明满足零条件均值假设，我们需要证明 $\mathrm{E}(u \mid \mathrm{inc}) = 0$。

根据题意，$u = \sqrt{inc} \cdot e$。因此，
$\mathrm{E}(u \mid \mathrm{inc}) = \mathrm{E}(\sqrt{inc} \cdot e \mid \mathrm{inc})$

由于 $\sqrt{inc}$ 是 inc 的函数，在给定 inc 的条件下，$\sqrt{inc}$ 是一个常数。因此，可以将其从条件期望中提出来：
$\mathrm{E}(u \mid \mathrm{inc}) = \sqrt{inc} \cdot \mathrm{E}(e \mid \mathrm{inc})$

题目中假设 e 独立于 inc，这意味着 e 的分布与 inc 的取值无关。因此，e 在给定 inc 下的条件期望等于其无条件期望：
$\mathrm{E}(e \mid \mathrm{inc}) = \mathrm{E}(e)$

又因为题目中给定 $\mathrm{E}(e) = 0$，所以：
$\mathrm{E}(e \mid \mathrm{inc}) = 0$

将此结果代入，我们得到：
$\mathrm{E}(u \mid \mathrm{inc}) = \sqrt{inc} \cdot 0 = 0$

因此，零条件均值假设 SLR.4 成立。

Solution (ii)

为了证明不满足同方差假设，我们需要考察 $\mathrm{Var}(u \mid inc)$ 是否为一个常数。

根据条件方差的性质，对于一个仅依赖于 inc 的函数 $g(inc)$，我们有：
$\mathrm{Var}(g(inc) \cdot e \mid inc) = [g(inc)]^2 \cdot \mathrm{Var}(e \mid inc)$

在本题中，$u = \sqrt{inc} \cdot e$，所以 $g(inc) = \sqrt{inc}$。因此：
$\mathrm{Var}(u \mid inc) = \mathrm{Var}(\sqrt{inc} \cdot e \mid inc) = (\sqrt{inc})^2 \cdot \mathrm{Var}(e \mid inc) = inc \cdot \mathrm{Var}(e \mid inc)$

由于假设 e 和 inc 独立，e 的条件方差等于其无条件方差：
$\mathrm{Var}(e \mid inc) = \mathrm{Var}(e) = \sigma_e^2$

将此结果代入，我们得到：
$\mathrm{Var}(u \mid inc) = inc \cdot \sigma_e^2$

由于 $\mathrm{Var}(u \mid inc)$ 的值依赖于 inc，它不是一个常数。特别地，随着 inc 的增加，$\mathrm{Var}(u \mid inc)$ 也随之增加。这就违反了同方差假设（SLR.5），该假设要求误差项的方差对于所有的解释变量的取值都保持不变。这种现象被称为异方差性。

由于 $\mathrm{Var}(sav \mid inc) = \mathrm{Var}(\beta_0 + \beta_1 inc + u \mid inc) = \mathrm{Var}(u \mid inc)$，因此 sav 的方差也随着 inc 的增加而增加。

Solution (iii)

支持储蓄方差随着家庭收入递增的证据可以从以下几个方面进行讨论：

- **储蓄行为的多样性：** 高收入家庭在满足基本生活需求后，拥有更多的可自由支配收入。这使得他们的储蓄决策更加多样化和灵活。他们可以选择进行风险较高的投资，如股票、基金等，这些投资的收益波动性较大，从而导致储蓄的方差也较大。相比之下，低收入家庭的收入主要用于基本消费，储蓄的额度和方式都相对固定和有限，因此方差较小。
- **预防性储蓄动机的差异：** 随着收入的增加，家庭对未来不确定性的担忧（如失业、疾病等）可能会有不同的反应。虽然高收入家庭有更强的能力应对风险，但他们也可能为了维持较高的生活水平而进行更多的预防性储蓄。这种预防性储蓄的规模和形式可能因家庭而异，从而增加了储蓄的波动性。
- **生命周期阶段的影响：** 年轻的高收入家庭可能为了未来的大额支出（如购房、子女教育）而积极储蓄，而年长的高收入家庭可能已经积累了大量财富，储蓄行为会转向财富管理和传承，这两种情况下的储蓄行为差异很大。这种在不同生命周期阶段储蓄行为的差异在低收入家庭中则不那么明显。
- **投资机会的差异：** 高收入家庭通常能够接触到更多样化的金融产品和投资机会，这些机会往往伴随着不同的风险和回报水平。他们可以根据自身的风险偏好进行资产配置，从而导致储蓄总额的方差

增大。而低收入家庭的投资渠道相对有限，大多选择银行存款等低风险方式，储蓄的稳定性更高。

综上所述，理论和经验证据都倾向于支持储蓄的方差随着家庭收入的增加而增加的观点。这种异方差性的存在对于计量经济学模型的估计和推断具有重要意义，在进行回归分析时需要采用适当的方法（如加权最小二乘法）来处理。

Chapter 2 Problem 10

Problem

令 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 分别为 OLS 截距和斜率估计量，并令 $\bar{u}$ 为误差（不是残差）的样本均值。
（i）证明：$\hat{\beta}_1$ 可写成 $\hat{\beta}_1 = \beta_1 + \sum_{i=1}^{n} w_i u_i$ ，其中 $w_i = d_i / \mathrm{SST}_x, d_i = x_i - \bar{x}$ 。
（ii）利用第（i）部分及 $\sum_{i=1}^{n} w_i = 0$ ，证明：$\hat{\beta}_1$ 和 $\bar{u}$ 不相关。[提示：要求你证明 $\mathrm{E}\left[\left(\hat{\beta}_1 - \beta_1\right)\bar{u}\right] = 0$ 。]
（iii）证明 $\hat{\beta}_0$ 可写成 $\hat{\beta}_0 = \beta_0 + \bar{u} - \left(\hat{\beta}_1 - \beta_1\right)\bar{x}$ 。
（iv）利用第（ii）部分和第（iii）部分证明：$\mathrm{Var}\left(\hat{\beta}_0\right) = \sigma^2/n + \sigma^2(\bar{x})^2/\mathrm{SST}_x$ 。
（v）第（iv）部分中的表达式能简化成方程（2．58）吗？ [提示：$\mathrm{SST}_x/n = n^{-1}\sum_{i=1}^{n} x_i^2 - (\bar{x})^2$.]

Solution (i)

OLS 斜率估计量 $\hat{\beta}_1$ 的定义为：
$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\mathrm{SST}_x}$
真实的模型为 $y_i = \beta_0 + \beta_1 x_i + u_i$。对该式在样本中求平均，可得 $\bar{y} = \beta_0 + \beta_1\bar{x} + \bar{u}$。
因此，$y_i - \bar{y} = (\beta_0 + \beta_1 x_i + u_i) - (\beta_0 + \beta_1\bar{x} + \bar{u}) = \beta_1(x_i - \bar{x}) + (u_i - \bar{u})$。
将上式代入 $\hat{\beta}_1$ 的分子中：
$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n}(x_i - \bar{x})[\beta_1(x_i - \bar{x}) + (u_i - \bar{u})]$
$= \beta_1\sum_{i=1}^{n}(x_i - \bar{x})^2 + \sum_{i=1}^{n}(x_i - \bar{x})(u_i - \bar{u})$
$= \beta_1\mathrm{SST}_x + \sum_{i=1}^{n}(x_i - \bar{x})u_i - \bar{u}\sum_{i=1}^{n}(x_i - \bar{x})$
由于 $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$，上式简化为：
$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \beta_1\mathrm{SST}_x + \sum_{i=1}^{n}(x_i - \bar{x})u_i$
将此结果代回 $\hat{\beta}_1$ 的表达式中：
$\hat{\beta}_1 = \frac{\beta_1\mathrm{SST}_x + \sum_{i=1}^{n}(x_i-\bar{x})u_i}{\mathrm{SST}_x} = \beta_1 + \frac{\sum_{i=1}^{n}(x_i-\bar{x})u_i}{\mathrm{SST}_x}$

令 $d_i = x_i - \bar{x}$ 且 $w_i = d_i / \mathrm{SST}_x$，则上式可以写为：
$\hat{\beta}_1 = \beta_1 + \sum_{i=1}^{n} w_i u_i$

□

Solution (ii)

为了证明 $\hat{\beta}_1$ 和 $\bar{u}$ 不相关，我们需要证明它们的协方差为零，即 $\mathrm{Cov}(\hat{\beta}_1, \bar{u}) = 0$。
根据协方差的定义，$\mathrm{Cov}(\hat{\beta}_1, \bar{u}) = \mathrm{E}[(\hat{\beta}_1 - \mathrm{E}[\hat{\beta}_1])(\bar{u} - \mathrm{E}[\bar{u}])]$。
在标准 OLS 假设下，$\mathrm{E}[\hat{\beta}_1] = \beta_1$ 且 $\mathrm{E}[\bar{u}] = \mathrm{E}[\frac{1}{n}\sum u_i] = \frac{1}{n}\sum \mathrm{E}[u_i] = 0$。
因此，我们需要证明 $\mathrm{E}[(\hat{\beta}_1 - \beta_1)\bar{u}] = 0$。
从第（i）部分可知，$\hat{\beta}_1 - \beta_1 = \sum_{i=1}^{n} w_i u_i$。
所以，
$\mathrm{E}[(\hat{\beta}_1 - \beta_1)\bar{u}] = \mathrm{E}\left[(\sum_{i=1}^{n} w_i u_i)\left(\frac{1}{n}\sum_{j=1}^{n} u_j\right)\right] = \frac{1}{n}\mathrm{E}\left[\sum_{i=1}^{n}\sum_{j=1}^{n} w_i u_i u_j\right]$
$= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} w_i \mathrm{E}[u_i u_j]$
根据 OLS 假设，误差项是相互独立的（随机抽样），且具有相同的方差 σ^2。这意味着：
$\mathrm{E}[u_i u_j] = 0$ 当 $i \neq j$
$\mathrm{E}[u_i u_j] = \mathrm{E}[u_i^2] = \mathrm{Var}(u_i) = \sigma^2$ 当 $i = j$
因此，双重求和中只有当 $i = j$ 时项才不为零：
$\mathrm{E}[(\hat{\beta}_1 - \beta_1)\bar{u}] = \frac{1}{n}\sum_{i=1}^{n} w_i \mathrm{E}[u_i^2] = \frac{1}{n}\sum_{i=1}^{n} w_i \sigma^2 = \frac{\sigma^2}{n}\sum_{i=1}^{n} w_i$
根据提示 $\sum_{i=1}^{n} w_i = \sum_{i=1}^{n} \frac{x_i - \bar{x}}{\mathrm{SST}_x} = \frac{1}{\mathrm{SST}_x}\sum_{i=1}^{n}(x_i - \bar{x}) = 0$。
所以，
$\mathrm{E}[(\hat{\beta}_1 - \beta_1)\bar{u}] = \frac{\sigma^2}{n} \cdot 0 = 0$
因此，$\hat{\beta}_1$ 和 $\bar{u}$ 不相关。

□

Solution (iii)

OLS 截距估计量 $\hat{\beta}_0$ 的定义为 $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$。
我们已知 $\bar{y} = \beta_0 + \beta_1 \bar{x} + \bar{u}$。
将 $\bar{y}$ 的表达式代入 $\hat{\beta}_0$ 的定义式中：
$\hat{\beta}_0 = (\beta_0 + \beta_1 \bar{x} + \bar{u}) - \hat{\beta}_1 \bar{x}$
重新整理上式：
$\hat{\beta}_0 = \beta_0 + \bar{u} + \beta_1 \bar{x} - \hat{\beta}_1 \bar{x}$
$\hat{\beta}_0 = \beta_0 + \bar{u} - (\hat{\beta}_1 - \beta_1)\bar{x}$

□

Solution (iv)

我们要求 $\mathrm{Var}(\hat{\beta}_0)$。根据第（iii）部分的结论：
$\mathrm{Var}(\hat{\beta}_0) = \mathrm{Var}(\beta_0 + \bar{u} - (\hat{\beta}_1 - \beta_1)\bar{x})$
由于 β_0，β_1 是常数，$\bar{x}$ 在给定样本的条件下也是常数，因此：
$\mathrm{Var}(\hat{\beta}_0) = \mathrm{Var}(\bar{u} - (\hat{\beta}_1 - \beta_1)\bar{x})$
利用方差公式 $\mathrm{Var}(A - B) = \mathrm{Var}(A) + \mathrm{Var}(B) - 2\,\mathrm{Cov}(A, B)$，可得：
$\mathrm{Var}(\hat{\beta}_0) = \mathrm{Var}(\bar{u}) + \mathrm{Var}((\hat{\beta}_1 - \beta_1)\bar{x}) - 2\,\mathrm{Cov}(\bar{u}, (\hat{\beta}_1 - \beta_1)\bar{x})$
我们分别计算这三项：

1. $\mathrm{Var}(\bar{u}) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} u_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(u_i) = \frac{1}{n^2}(n\sigma^2) = \frac{\sigma^2}{n}$。
2. $\mathrm{Var}((\hat{\beta}_1 - \beta_1)\bar{x}) = (\bar{x})^2\,\mathrm{Var}(\hat{\beta}_1 - \beta_1) = (\bar{x})^2\,\mathrm{Var}(\hat{\beta}_1)$。已知 $\mathrm{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\mathrm{SST}_x}$，所以该项为 $(\bar{x})^2\frac{\sigma^2}{\mathrm{SST}_x}$。
3. $\mathrm{Cov}(\bar{u}, (\hat{\beta}_1 - \beta_1)\bar{x}) = \bar{x}\,\mathrm{Cov}(\bar{u}, \hat{\beta}_1 - \beta_1) = \bar{x}\,\mathrm{Cov}(\bar{u}, \hat{\beta}_1)$。根据第（ii）部分的结论，$\mathrm{Cov}(\bar{u}, \hat{\beta}_1) = 0$，所以该项为 0。

将这三项结果合并：
$\mathrm{Var}(\hat{\beta}_0) = \frac{\sigma^2}{n} + \frac{\sigma^2(\bar{x})^2}{\mathrm{SST}_x} - 0$
$\mathrm{Var}(\hat{\beta}_0) = \frac{\sigma^2}{n} + \frac{\sigma^2(\bar{x})^2}{\mathrm{SST}_x}$

□

Solution (v)

是的，第（iv）部分中的表达式可以被简化。方程（2.58）通常指 $\mathrm{Var}(\hat{\beta}_0)$ 的另一种常见形式。我们从第（iv）部分的结果开始：
$\mathrm{Var}(\hat{\beta}_0) = \frac{\sigma^2}{n} + \frac{\sigma^2(\bar{x})^2}{\mathrm{SST}_x} = \sigma^2\left(\frac{1}{n} + \frac{(\bar{x})^2}{\mathrm{SST}_x}\right)$
将括号内的两项通分：
$\mathrm{Var}(\hat{\beta}_0) = \sigma^2\left(\frac{\mathrm{SST}_x + n(\bar{x})^2}{n\cdot\mathrm{SST}_x}\right)$
根据提示，$\mathrm{SST}_x = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n(\bar{x})^2$。
将这个关系代入上式分子的 SST_x：
$\mathrm{SST}_x + n(\bar{x})^2 = \left(\sum_{i=1}^{n} x_i^2 - n(\bar{x})^2\right) + n(\bar{x})^2 = \sum_{i=1}^{n} x_i^2$
所以， $\mathrm{Var}(\hat{\beta}_0)$ 的表达式可以简化为：

$\mathrm{Var}(\hat{\beta}_0) = \sigma^2 \left(\frac{\sum_{i=1}^n x_i^2}{n \cdot \mathrm{SST}_x} \right) = \frac{\sigma^2 \sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{x})^2}$
这个形式通常就是指代号为（2.58）的方程。

Chapter 2 Problem 11

Problem

假设你对估计学生花在学习 SAT 预备课程上的小时数（hours）对 SAT 最终成绩（sat）的影响感兴趣．样本整体是某一年即将上大学的高三学生．
（i）假设你被允许进行一个对照试验。解释你将如何设计实验从而估计 hours 对 sat 的因果效应。
（ii）考虑一个更加实际的情形，即由学生选择在学习预备课程上花多少时间，而你只能随机地从总体中抽出 sat 和 hours 两个变量。将总体模型写作如下形式：
$sat = \beta_0 + \beta_1 \text{ hours } + u$

式中，与通常带截距的模型一样，我们可以假定 $\mathrm{E}(u) = 0$ 。列举出至少两个 u 中包含的因素．这些因素是否与 hours 正相关或负相关?
（iii）如果上一问等式中的预备课程有效果，那么在第（ii）部分的方程中，β_1 的符号应该是正还是负?
（iv）在第（ii）部分的方程中，β_0 应该如何解释?

Solution (i)

为了通过对照试验来估计 hours 对 sat 的因果效应，我们可以设计如下一个随机对照试验（Randomized Controlled Trial, RCT）：

1. **招募被试：** 首先，从目标总体（某一年即将上大学的高三学生）中招募一大批愿意参与此项研究的学生。
2. **随机分组：** 将这些学生完全随机地分配到不同的组中。至少需要一个**控制组**（Control Group）和一个**处理组**（Treatment Group）。为了更精确地估计效果，可以设置多个处理组。例如：
   - **控制组 (Group A):** 被分配到该组的学生不参加任何 SAT 预备课程（hours = 0）。
   - **处理组 1 (Group B):** 被要求参加一个为期 20 小时的 SAT 预备课程。
   - **处理组 2 (Group C):** 被要求参加一个为期 40 小时的 SAT 预备课程。
   - **处理组 3 (Group D):** 被要求参加一个为期 60 小时的 SAT 预备课程。
3. **实施处理：** 确保每个组的学生都严格遵守分配的学习小时数。课程的内容和教师对于所有处理组都应保持一致，以确保唯一的区别是学习时长。

4. **收集数据：** 待所有学生完成指定时长的课程后，让他们参加同一次 SAT 考试，并收集他们的最终成绩（sat）。
5. **估计效应：** 比较不同组别的平均 SAT 成绩。例如，处理组 B 的平均分与控制组 A 的平均分之差，就是 20 小时课程带来的平均因果效应。通过对收集到的实验数据进行 `sat` 对 `hours` 的 OLS 回归，得到的斜率系数 $\hat{\beta}_1$ 将是 `hours` 对 `sat` 的因果效应的一个无偏估计。

关键在于随机化。随机分配确保了（在样本量足够大的情况下）各个组的学生在所有其他潜在影响 SAT 成绩的因素上（如个人能力、家庭背景、学习动机等）在统计上是相同的。因此，各组之间 SAT 成绩的任何系统性差异都可以归因于预备课程时长的不同，从而分离出因果效应。

Solution (ii)

在这种观测性研究的情形下，误差项 u 包含了除 `hours` 以外所有可能影响 `sat` 成绩的因素。以下是其中两个主要因素及其与 `hours` 的可能关系：

1. **个人能力（Innate Ability）：** 这包括学生的智商、天赋、以及在高中阶段已经掌握的知识基础。个人能力强的学生通常在 SAT 考试中表现更好。
   - **与 hours 的关系：** 这种关系是不明确的，可能是负相关。能力较差的学生可能会觉得需要更多地依赖预备课程来弥补自己的不足，因此会投入更多的时间（ `hours` 较高）。反之，天资聪颖的学生可能认为自己不需要花太多时间在预备课程上，因此学习时间较短（ `hours` 较低）。如果这种情况成立，则个人能力与 `hours` **负相关**。
2. **学习动机（Motivation）：** 指学生努力学习并取得好成绩的内在驱动力。动机更强的学生通常会更认真地准备考试，也可能在考试中有更好的发挥，从而获得更高的 `sat` 成绩。
   - **与 hours 的关系：** 学习动机强的学生，不仅会在学校课程上更努力，也更可能主动花更多时间在 SAT 预备课程上，以期获得理想的大学录取。因此，学习动机与 `hours` **正相关**。

其他可能的因素还包括：**家庭收入和父母的教育水平**（通常与 `hours` 正相关，因为富裕家庭更能负担课程费用，且受教育程度高的父母更重视此事）、**高中教学质量**等。这些因素存在于 u 中，并且都与 `hours` 相关，这将导致 OLS 估计量 $\hat{\beta}_1$ 存在遗漏变量偏误（Omitted Variable Bias），无法准确反映 `hours` 对 `sat` 的真实因果效应。

Solution (iii)

如果 SAT 预备课程确实有效果，这意味着在其他所有条件相同的情况下（ceteris paribus），增加学习预备课程的时间应该会导致 SAT 成绩的提高。因此， `hours` 和 `sat` 之间存在一个正向的因果关系。

在总体模型 $sat = \beta_0 + \beta_1 \text{hours} + u$ 中，β_1 衡量的是 `hours` 每增加一个单位（一小时）， `sat` 成绩的期望变化量。所以，如果课程有效，β_1 **的符号应该是正的**。

Solution (iv)

在方程 $sat = \beta_0 + \beta_1 \text{ hours } + u$ 中，截距项 β_0 的解释是：当解释变量 `hours` 取值为 0 时，被解释变量 `sat` 的期望值。

具体来说，β_0 **代表了那些没有花任何时间在 SAT 预备课程上的学生的平均 SAT 成绩**。在这个应用场景中， `hours = 0` 是一个非常现实且有意义的取值（因为很多学生根本不参加预备课程），所以对 β_0 的这种解释是完全合理和有价值的。它为我们提供了一个比较基准，即不参加预备课程的学生群体的平均水平。

Chapter 2 Problem C2

Problem

数据集 CEOSAL2 包含了美国公司首席执行官的信息。变量 salary 是以千美元计的年辞，ceaten是已担任公司 CEO 的年数。
（i）求出样本中的平均年薪和平均任期.
（ii）有多少位 CEO 尚处于担任 CEO 的第一年（也就是说，ceoten $= 0$ ）？ 最长的 CEO 任期是多少?
（iii）估计简单回归模型

$$\log(\text{ salary }) = \beta_0 + \beta_1 \text{ ceoten } + u$$

以通常格式报告结果。多担任一年 CEO，预计年薪增长（近做）的百分数是多少?

Solution (i)

根据对CEOSAL2数据集的分析，样本中包含的177位CEO的平均信息如下：

- **平均年薪**: 865,900美元 (即865.9千美元)。
- **平均任期**: 约7.95年。

Solution (ii)

在样本数据中：

- 有 **5** 位CEO正处于他们担任CEO的第一年（ `ceoten = 0`）。

- 最长的CEO任期为 **37** 年。

Solution (iii)

为了探究CEO任期对其薪酬的影响，我们估计了以下简单回归模型：
$\log(\text{ salary }) = \beta_0 + \beta_1 \text{ ceoten } + u$
其中 `salary` 以千美元为单位。使用普通最小二乘法（OLS）得到的估计结果如下：

$$\widehat{\log(\text{ salary })} = 6.51 + 0.0097 \cdot \text{ ceoten}$$
$$n = 177, \quad R^2 = 0.013$$

结果解释:
该模型的斜率系数 $\hat{\beta}_1 = 0.0097$ 表明，CEO的任期（ceoten）与对数薪酬之间存在正相关关系。由于因变量是对数形式，我们可以这样解释系数：

每多担任一年CEO，其年薪预计增长约0.97%。

这个效应虽然在统计上是正的，但从 $R^2 = 0.013$ 来看，CEO的任期年限本身只能解释薪酬对数变化中非常小的一部分（约1.3%）。这表明还有许多其他更重要的因素会影响CEO的薪酬。

Chapter 3 Problem 5

Problem

在一项调查大学 GPA 与在各种活动中所耗费时间之间关系的研究中．你对几个学生分发了训查问卷。学生被问道，他们每周在学习、睡觉、工作和休闲这四种活动中各花多少小时。任何活动都被列为这四种活动之一，所以对每个学生来说，这四种活动的小时数之和都是 168 。

（i）在模型
$GPA = \beta_0 + \beta_1 study + \beta_2 sleep + \beta_3 work + \beta_4 leisure + u$
中，保持 sleep、work 和leisure 不变而改变 study 是否有意义?

（ii）解释为什么这个模型违背了假设 MLR.3。

（iii）你如何重新构建这个模型，才能使得它的参数具有一个有用的解释，而又不违肯假设 MLR.3?

Solution (i)

没有意义。在多元回归分析中，对某个自变量系数（如 β_1）的解释是在“保持其他所有自变量不变”（ceteris paribus）的条件下进行的。然而，根据问题的设定，这四项活动的时间总和是固定的，即：
$study + sleep + work + leisure = 168$
因此，如果 `study` 的时间增加一个小时，那么其他三项活动（ `sleep` ， `work` ， `leisure` ）的总时间就必须减少一个小时，以维持总和为168小时。我们不可能在改变 `study` 的同时，保持其他所有变量完全不变。所以，这个模型中对单个系数的常规解释是没有实际意义的。

Solution (ii)

假设 MLR.3 指的是“不存在完全共线性”，即任何一个自变量都不能是其他自变量的完全线性函数。

在这个模型中，由于四个自变量的和是一个常数168，它们之间存在一个完全的线性关系。例如，我们可以把任何一个变量表示为其他变量的线性组合：
$leisure = 168 - study - sleep - work$
这就意味着自变量之间存在完全共线性。当自变量之间存在这种关系时，我们无法唯一地估计出模型中的参数（$\beta_1, \beta_2, \beta_3, \beta_4$），因为有无穷多组系数可以满足最小二乘法的要求。从数学上讲，这会导致设计矩阵 $(X'X)$ 不可逆，从而无法计算出 OLS 估计量。这种情况通常被称为“虚拟变量陷阱”(dummy variable trap)，尽管这里的变量是定量的而非虚拟的，但原理是完全相同的。

Solution (iii)

为了解决完全共线性问题并使模型的参数具有有用的解释，我们可以从模型中剔除其中一个自变量。例如，我们可以剔除 `leisure` ，将其作为基准组或参照活动。

重新构建的模型如下：
$GPA = \beta_0 + \beta_1 study + \beta_2 sleep + \beta_3 work + u$
现在，这个模型不再违背假设 MLR.3，因为剩下的自变量之间不再存在完全的线性关系。并且，它的参数具有了非常有用的解释：

- β_1：衡量的是在保持 `sleep` 和 `work` 时间不变的情况下，将**一小时的 `leisure` 时间用于 `study`** 对 GPA 的影响。
- β_2：衡量的是在保持 `study` 和 `work` 时间不变的情况下，将**一小时的 `leisure` 时间用于 `sleep`** 对 GPA 的影响。
- β_3：衡量的是在保持 `study` 和 `sleep` 时间不变的情况下，将**一小时的 `leisure` 时间用于 `work`** 对 GPA 的影响。
- β_0：截距项现在表示一个学生将所有168小时都用于 `leisure` 时的预测 GPA 值。

通过这种方式重新构建模型，我们不仅解决了技术上的估计问题，而且得到的系数也反映了不同活动之间时间分配的“权衡”(trade-off）对学业成绩的影响，这使得解释变得直观且有意义。

Chapter 3 Problem 6

Problem

6．考虑含有三个自变量的多元回归模型，并满足假设 MLR． 1 到 MLR．4，

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u$$

你对估计 x_1 和 x_2 的参数之和感兴趣；把二者之和记为 $\theta_1 = \beta_1 + \beta_2$ 。
（i）证明 $\hat{\theta}_1 = \hat{\beta}_1 + \hat{\beta}_2$ 是 θ_1 的一个无偏估计量。
（ii）求出用 $\mathrm{Var}\left(\hat{\beta}_1\right)$ 、 $\mathrm{Var}\left(\hat{\beta}_2\right)$ 和 $\mathrm{Corr}\left(\hat{\beta}_1, \hat{\beta}_2\right)$ 表示的 $\mathrm{Var}\left(\hat{\theta}_1\right)$ 。

Solution (i)

根据高斯-马尔可夫假设（MLR.1 到 MLR.4），普通最小二乘法（OLS）估计量是无偏的。即对于模型中的任意参数 β_j ，都有：
$E\left(\hat{\beta}_j\right) = \beta_j, \quad j = 0, 1, 2, 3$

我们需要证明 $\hat{\theta}_1$ 是 θ_1 的无偏估计量，即证明 $E\left(\hat{\theta}_1\right) = \theta_1$ 。
根据期望的线性性质：
$E\left(\hat{\theta}_1\right) = E\left(\hat{\beta}_1 + \hat{\beta}_2\right) = E\left(\hat{\beta}_1\right) + E\left(\hat{\beta}_2\right)$

将无偏性结果代入上式：
$E\left(\hat{\theta}_1\right) = \beta_1 + \beta_2$

因为我们定义 $\theta_1 = \beta_1 + \beta_2$ ，所以：
$E\left(\hat{\theta}_1\right) = \theta_1$

因此， $\hat{\theta}_1 = \hat{\beta}_1 + \hat{\beta}_2$ 是 θ_1 的一个无偏估计量。

Solution (ii)

我们需要计算 $\hat{\theta}_1$ 的方差：

$$\operatorname{Var}\left(\hat{\theta}_1\right) = \operatorname{Var}\left(\hat{\beta}_1 + \hat{\beta}_2\right)$$

根据和的方差公式 $\operatorname{Var}(X+Y) = \operatorname{Var}(X) + \operatorname{Var}(Y) + 2\operatorname{Cov}(X,Y)$：

$$\operatorname{Var}\left(\hat{\theta}_1\right) = \operatorname{Var}\left(\hat{\beta}_1\right) + \operatorname{Var}\left(\hat{\beta}_2\right) + 2\operatorname{Cov}\left(\hat{\beta}_1, \hat{\beta}_2\right)$$

利用相关系数的定义 $\operatorname{Corr}(X,Y) = \frac{\operatorname{Cov}(X,Y)}{\sqrt{\operatorname{Var}(X)}\sqrt{\operatorname{Var}(Y)}}$，我们可以将协方差表示为：

$$\operatorname{Cov}\left(\hat{\beta}_1, \hat{\beta}_2\right) = \operatorname{Corr}\left(\hat{\beta}_1, \hat{\beta}_2\right) \cdot \sqrt{\operatorname{Var}\left(\hat{\beta}_1\right)} \cdot \sqrt{\operatorname{Var}\left(\hat{\beta}_2\right)}$$

代回方差公式，得到最终表达式：

$$\operatorname{Var}\left(\hat{\theta}_1\right) = \operatorname{Var}\left(\hat{\beta}_1\right) + \operatorname{Var}\left(\hat{\beta}_2\right) + 2\operatorname{Corr}\left(\hat{\beta}_1, \hat{\beta}_2\right) \sqrt{\operatorname{Var}\left(\hat{\beta}_1\right) \operatorname{Var}\left(\hat{\beta}_2\right)}$$

Chapter 3 Problem C1

Problem

健康官员（和其他人）关心的一个问题是：孕妇在怀孕期间抽烟对婴儿健康的影响。对婴儿健康的度量方法之一是婴儿出生时的体重；过低的出生体重会使婴儿有感架各种疾病的危险。由于除了抽烟之外，其他影响婴儿出生体重的因索可能与抽烟相关，所以我们应该考虑这些因索。比如，高收人通常会使母亲得到更好的产前照硕和更好的菅养。表达这一点的一个方程是

$$bwght = \beta_0 + \beta_1 cigs + \beta_2 faminc + u$$

（i）
β_2
的符号最可能是什么?

（ii）你是否认为 cigs 和 faminc 可能相关？解释为什么可能是正或负相关.

（iii）现在利用 BWGHT 中的数据分别估计包含和不包含 faminc 的方程。以方程的形式报告结论，包括样本容量和
R^2
。讨论你的结论，主要看增加 faminc 是否会显著改变 ags 对 bught 的估计影响。

Solution (i)

β_2
衡量的是在保持吸烟数量（cigs）不变的情况下，家庭收入（faminc）对婴儿出生体重（bwght）的影响。普遍认为，较高的家庭收入能为孕妇带来更好的产前护理、更优质的营养和更健康的生活环境。这些因素都有利于胎儿的成长和发育，从而增加婴儿的出生体重。因此，β_2
的符号最可能是正（+）的。研究也表明，较低的家庭收入与较高的低出生体重风险相关。

Solution (ii)

变量cigs（怀孕期间每日吸烟量）和faminc（家庭收入）很可能存在相关性。

负相关（最可能的情况）：
通常，社会经济地位较低的人群吸烟率更高。家庭收入较低的个体可能面临更大的生活压力、较少的健康教育资源，并且更容易生活在鼓励吸烟的社会环境中。因此，家庭收入（faminc）越高，吸烟量（cigs）可能越低，两者之间呈现负相关关系。多项研究数据表明，家庭收入低于2万美元的家庭中吸烟率约为32.2%，而收入超过10万美元的家庭中吸烟率则降至12.1%。

正相关（可能性较低）：
在某些特定情况下，也可能存在正相关关系。例如，有人可能认为收入较高的家庭有更强的经济能力购买香烟。然而，考虑到健康意识和教育水平通常随收入增加而提高，这种正相关关系在总体人口中出现的可能性较小。一项针对欧洲六个城市青少年的研究发现，个人收入较高的青少年吸烟的可能性也更高，但这可能不适用于怀孕的成年女性群体。

综上所述，cigs和faminc之间最可能存在负相关关系。

Solution (iii)

利用 BWGHT 数据集，我们分别估计包含和不包含家庭收入（faminc）的两个模型。

模型 1：不包含 faminc

此模型考察每日吸烟量对婴儿出生体重的影响，不控制家庭收入。
估计的方程为：
$\hat{bwght} = 119.77 - 0.514 \cdot cigs$
样本容量 (n) = 1388
$R^2 = 0.023$

这个结果表明，平均而言，孕妇每天多抽一支烟，婴儿出生体重会降低约0.514盎司。

模型 2：包含 faminc

此模型同时考察每日吸烟量和家庭收入对婴儿出生体重的影响。
估计的方程为：
$\hat{bwght} = 116.97 - 0.463 \cdot cigs + 0.093 \cdot faminc$
样本容量 (n) = 1388
R^2 = 0.030

结论与讨论

1. **系数变化**: 将家庭收入（faminc）加入到回归模型后，吸烟量（cigs）的系数从-0.514变为-0.463。虽然影响方向仍然为负，但其绝对值变小了。这表明，在控制了家庭收入后，吸烟对出生体重的负面影响估计值有所减弱。
2. **遗漏变量偏误**: 这种系数的变化是典型的遗漏变量偏误（omitted variable bias）的体现。在模型1中，由于没有控制与吸烟量（cigs）负相关且与出生体重（bwght）正相关的家庭收入（faminc），导致模型将一部分本应归因于低收入的负面影响错误地归因给了吸烟。因此，模型1高估了吸烟对出生体重的负面影响。
3. **faminc的显著性**: 在模型2中，faminc的系数为+0.093，符号与预期一致，即家庭年收入每增加1000美元，婴儿出生体重平均增加约0.093盎司。这个系数在统计上是显著的，说明家庭收入确实是影响婴儿出生体重的一个重要因素。
4. **模型解释力**: 加入faminc后，模型的R^2从0.023增加到0.030，说明模型对婴儿出生体重变异的解释能力略有提升。

综上所述，增加faminc确实显著改变了cigs对bwght的估计影响。简单的回归模型（模型1）可能夸大了吸烟的危害，因为它忽略了与吸烟和出生体重都有关的潜在因素，如家庭收入。更全面的模型（模型2）提供了对吸烟影响的更准确估计。

Chapter 3 Problem C6

Problem

本题利用 WAGE2 中的数据．照常保证如下所有回归都含有截距．

（i）将 IQ 对 educ 进行简单回归，并得到斜率系数 $\tilde{\delta}_1$。

（ii）将 log(wage) 对 educ 进行简单回归，并得到斜率系数 $\tilde{\beta}_1$。

（iii）将 log(wage) 对 educ 和 IQ 进行多元回归，并分别得到斜率系数 $\hat{\beta}_1$ 和 $\hat{\beta}_2$。

（iv）验证 $\tilde{\beta}_1 = \hat{\beta}_1 + \hat{\beta}_2\tilde{\delta}_1$。

Solution (i)

我们将 IQ 作为因变量，教育年限（educ）作为自变量进行简单线性回归。回归模型如下：
$IQ = \delta_0 + \delta_1 educ + v$

利用 WAGE2 数据集（样本容量 n=935）进行估计，得到的结果如下：
$\widehat{IQ} = 53.9 + 3.53 \cdot educ$
$n = 935, \quad R^2 = 0.252$

因此，斜率系数 $\tilde{\delta}_1 = 3.53$。这个结果表明，教育年限每增加一年，IQ分数平均提高约3.53分。

Solution (ii)

我们将对数工资 log(wage) 作为因变量，教育年限（educ）作为自变量进行简单线性回归。回归模型如下：
$log(wage) = \beta_0 + \beta_1 educ + w$

利用 WAGE2 数据集进行估计，得到的结果如下：
$\widehat{log(wage)} = 5.97 + 0.0598 \cdot educ$
$n = 935, \quad R^2 = 0.115$

因此，斜率系数 $\tilde{\beta}_1 = 0.0598$。这个结果意味着，在不考虑其他因素的情况下，每增加一年教育，工资大约增长5.98%。

Solution (iii)

我们将对数工资 log(wage) 作为因变量，教育年限（educ）和IQ分数作为自变量进行多元线性回归。回归模型如下：
$log(wage) = \beta_0 + \beta_1 educ + \beta_2 IQ + u$

利用 WAGE2 数据集进行估计，得到的结果如下：
$\widehat{log(wage)} = 5.65 + 0.0391 \cdot educ + 0.0059 \cdot IQ$
$n = 935, \quad R^2 = 0.136$

因此，斜率系数分别为 $\hat{\beta}_1 = 0.0391$ 和 $\hat{\beta}_2 = 0.0059$。
这里，$\hat{\beta}_1$ 的含义是：在保持IQ分数不变的情况下，每增加一年教育，工资大约增长3.91%。

Solution (iv)

本部分旨在验证遗漏变量偏误的公式。该公式表明，简单回归中的系数（$\tilde{\beta}_1$）等于多元回归中该变量的系数（$\hat{\beta}_1$）加上一个偏误项，该偏误项等于多元回归中被遗漏变量的系数（$\hat{\beta}_2$）与被遗漏变量对包含变量的回归系数（$\tilde{\delta}_1$）的乘积。

公式为：$\tilde{\beta}_1 = \hat{\beta}_1 + \hat{\beta}_2\tilde{\delta}_1$

我们将前三步中得到的系数值代入公式右侧：
$\hat{\beta}_1 + \hat{\beta}_2\tilde{\delta}_1 = 0.0391 + (0.0059 \cdot 3.53)$
$= 0.0391 + 0.020827$
$= 0.059927$

将计算结果与（ii）中得到的 $\tilde{\beta}_1$ 进行比较：
$0.059927 \approx 0.0598$

结果基本相等（微小的差异源于系数的四舍五入）。这验证了遗漏变量偏误的公式。简单回归中教育对工资的回报（5.98%）被高估了，因为它包含了与教育正相关的IQ对工资的积极影响。在控制了IQ之后，教育的真实回报（3.91%）就显现出来了。