1. （20分）为了研究汽车汽油销售量与汽油税率的关系，一项研究对2001 年美国51个州或地区的数据集fuel2001进行了回归分析。该数据的变量包括Fuel (人均汽油销售量, 单位: 加仑)、Tax (汽油税率, 单位: 美分/加仑)、 Drivers (1000 人中持驾照的平均人数)、Income (人均收入, 单位1000美元)、 Miles (国有公路长度, 单位英里)。 下面是R软件的部分输出结果汇总 (含截距项，但下述结果中没有列出):

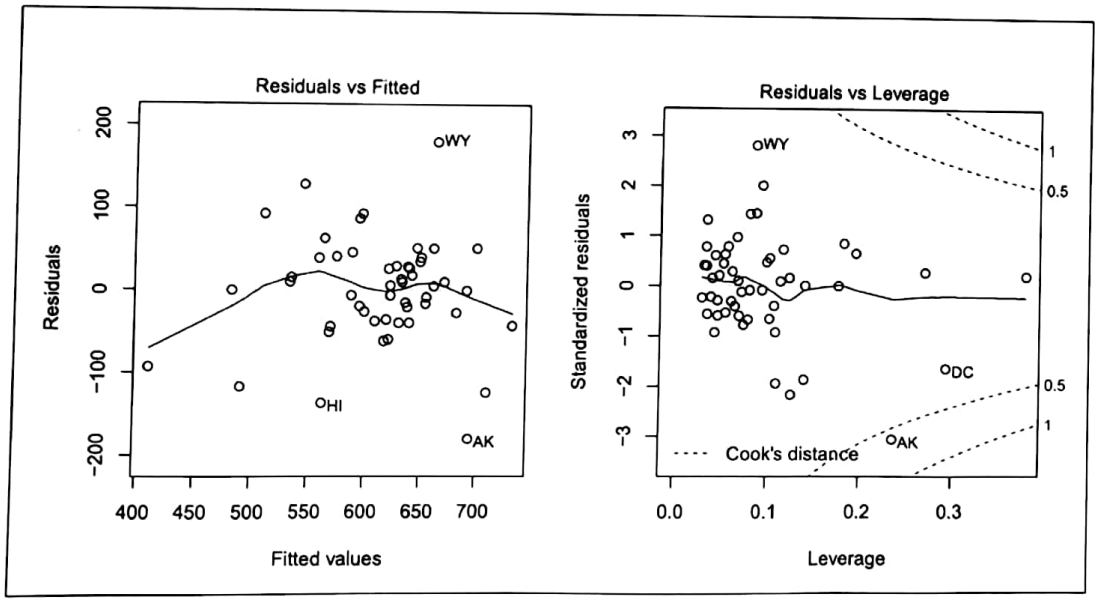
Call: lm(formula = Fuel ~ Tax + Drivers + Income + Miles, data = fuel2001)
Coefficients:

	Estimate	Std. Error	t value	Pr(> |t|)
(Intercept)	-	-	-	-
Tax	−4.200	2.100	①	0.0460
Drivers	0.535	②	3.898	0.0003
Income	③	2.205	−3.236	0.0020
Miles	0.402	0.187	④	⑤

Residual standard error: 67.17
Multiple R-Squared: 0.476
F-statistic: ⑥ on ⑦ and ⑧ degrees of freedom

(a) 请填写①-⑧处的数字（其中⑤处填写大于还是小于0.05, ⑥为F 检验的值, ⑦ 和⑧处为F 检验的两个自由度）。 (b) 试解释Tax的回归系数估计值 −4.2 的含义。

2. （15分）上题的部分回归诊断图如下。左图中标出了残差绝对值最大的3个异常点: WY (Wyoming, 怀俄明), AK (Alaska,阿拉斯加) 和 HI (Hawaii, 夏威夷); 右图标出了Cook距离最大的三个州或地区: AK, DC (Washington DC, 华盛顿特区) 和 WY。



(a) 从残差图(左图)来看，线性模型的高斯-马尔可夫(Gauss-Markov) 假设是否满足？如果你认为满足，说明理由；如果不满足，你拟采取什么措施？

(b) 左图表明HI和AK的人均汽油销售量偏低, 右图表明AK是高影响点, 而HI不是, 为什么?

(c) 根据右图所标的DC的位置，说明DC的自变量和响应变量各有什么特点， 在左图中它大概在哪个位置（边缘还是中间、上方还是下方）？

3. （20分） 假设独立样本$(x_i, y_i) \in R^2, i = 1, ..., n$ 满足下述模型

$$y = f(x) + \epsilon, \epsilon \sim (0, \sigma^2), \quad f(x) = \begin{cases} a, & \text{若} x \le t \\ a + b(x - t), & \text{若} x > t \end{cases},$$

假设 t 已知，且$x_1 \le ... \le x_m \le t < x_{m+1} \le ... \le x_n$。

(a) 基于 t 之后的数据 $(x_i, y_i), i = m + 1, ..., n$，求出 b 的LS估计 $\widetilde{b}$ 及其方差。

(b) 写出所有数据 $(x_i, y_i), i = 1, ..., n$ 满足的矩阵-向量形式的线性模型 $\mathbf{y} = X\boldsymbol{\beta} + \epsilon$，特别地，写出设计阵 X 的具体形式。

(c) 基于上述模型，求解 b 的LS估计 $\widehat{b}$ 及其方差，简单解释为什么它们与 t 之前的数据有关。

4. （20分） 假设数据 $(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n)$ 满足如下线性回归模型

$$y_i = \mathbf{x}_i^\top \beta + \epsilon_i, \quad \epsilon_i, \ i = 1, ..., n \text{ iid } \sim (0, \sigma^2),$$

其中 $\mathbf{x}_i$ 为 $p \times 1$ 自变量，其第一个元素为1. 设 $\hat{\beta}$ 为 β 的最小二乘估计。

(a) 假设对某个 $1 \le i \le n$, $\mathbf{x}_i = \bar{\mathbf{x}}$, 其中 $\bar{\mathbf{x}} = \sum_{j=1}^n \mathbf{x}_j / n$为自变量的样本平均值。 证明拟合值 $\hat{y}_i = \bar{y}$。

(b) 假设 $\mathbf{x}_i$ 作为设计阵X的行向量共出现了m次 $(m \ge 1)$，假设$X_{(-i)}^\top X_{(-i)}$可逆，其中$X_{(-i)}$是删除X第i行后的矩阵， 则 杠杆值 $h_{ii} < 1/m$。

5. （25分） 假设模型

$$\mathbf{y}_{n\times 1} = X_{n\times p}\boldsymbol{\beta}_{p\times 1} + \epsilon_{n\times 1} = \sum_{j=1}^{p} \mathbf{x}_j \beta_j + \epsilon, \ \epsilon \sim (\mathbf{0}, \sigma^2 I_n),$$

其中 $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^\top$ 为回归系数, 设计阵 $X = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_p)$ 满足条件 $X^\top X = I_p$。

(a) 基于数据$\mathbf{y}$和X, 求 β_j 的最小二乘估计 $\widehat{\beta}_j$ 及其方差。

(b) 对任一下标集合 $A \subseteq \Omega = \{1, 2, ..., p\}$, 定义

$$\widetilde{\mathbf{y}}^{(A)} = \begin{cases} \sum_{j\in A} \mathbf{x}_j \widehat{\beta}_j, & \text{若 } A \ne \phi \text{ (空集)} \\ \mathbf{0}, & \text{若 } A = \phi \end{cases}.$$

其均方误差定义为 $m(A) = E||\widetilde{\mathbf{y}}^{(A)} - X\boldsymbol{\beta}||^2$。证明: $m(A) = |A|\sigma^2 + \sum_{j\notin A} \beta_j^2$, 其中 $|A|$ 为集合 A 中元素的个数。

(c) 证明如果 $||\boldsymbol{\beta}||^2 \le \sigma^2$, 则对任何$A \subseteq \Omega$ 有 $m(\phi) \le m(A) \le m(\Omega)$.

(d) 令 $\widehat{m}(A) = (2|A| - p)\widehat{\sigma}^2 + \sum_{j\notin A} \widehat{\beta}_j^2$, 其中 $\widehat{\sigma}^2 = \mathbf{y}^\top (I_n - XX^\top)\mathbf{y}/(n - p)$, 证明 $\widehat{m}(A)$是 $m(A)$ 的无偏估计。

(e) 最优子集 A_{opt} 是 Ω 所有 2^p 个子集中使得$\widehat{m}(A)$ 达到最小的子集。试设计一种算法，只需搜索至多 $p + 1$ 个子集即可求出最优子集A_{opt}。