实验报告

实验目的

对 100 个男女成人身高和体重的数据进行描述性分析和画图探索性分析；将其中某个变量作为响应变量，其他变量作为协变量，提出有意义的研究问题和假设，并回答这些问题和假设，解释模型中参数的含义；解读 R 命令，说明结果是否合理，说明模型是否很好的描述了变量之间的关系，如果不够好给出改进方法。

实验步骤

（1）对数据做描述性分析和画图做探索性分析 (30 分)

首先计算总体身高，体重的均值，标准差，再分别计算男性和女性的身高和体重的均值与标注差，如表 (0.1) 所示。

	身高 (cm)	体重 (kg)
总体均值	174.20	119.63
总体标注差	10.95	42.30
男性均值	180.96	78.78
男性标准差	9.79	11.74
女性均值	167.50	160.48
女性标准差	7.35	8.44

Table 0.1: 数据的描述性分析

可以看到男生和女生平均身高和平均体重都有差异，而且数据呈现异方差性。

查看数据的最大最小值以及分位点，如表 (0.2) 所示。

x	y	sex
Min. :155.0	Min. : 61.0	female:50
1st Qu.:167.8	1st Qu.: 78.0	male :50
Median :173.0	Median :122.5	
Mean :174.2	Mean :119.6	
3rd Qu.:180.0	3rd Qu.:161.2	
Max. :199.0	Max. :176.0	

Table 0.2: 分为点

画出数据散点图如下 (图 1)。可以看到数据分为两部分，且体重和身高具有明显的线性关系。

分别画出总体男性女性身高箱线图 (图 2)。

总体男性女性体重箱线图 (图 2)。

分别画出总体，男性，女性身高直方图，体重直方图 (图 2)。

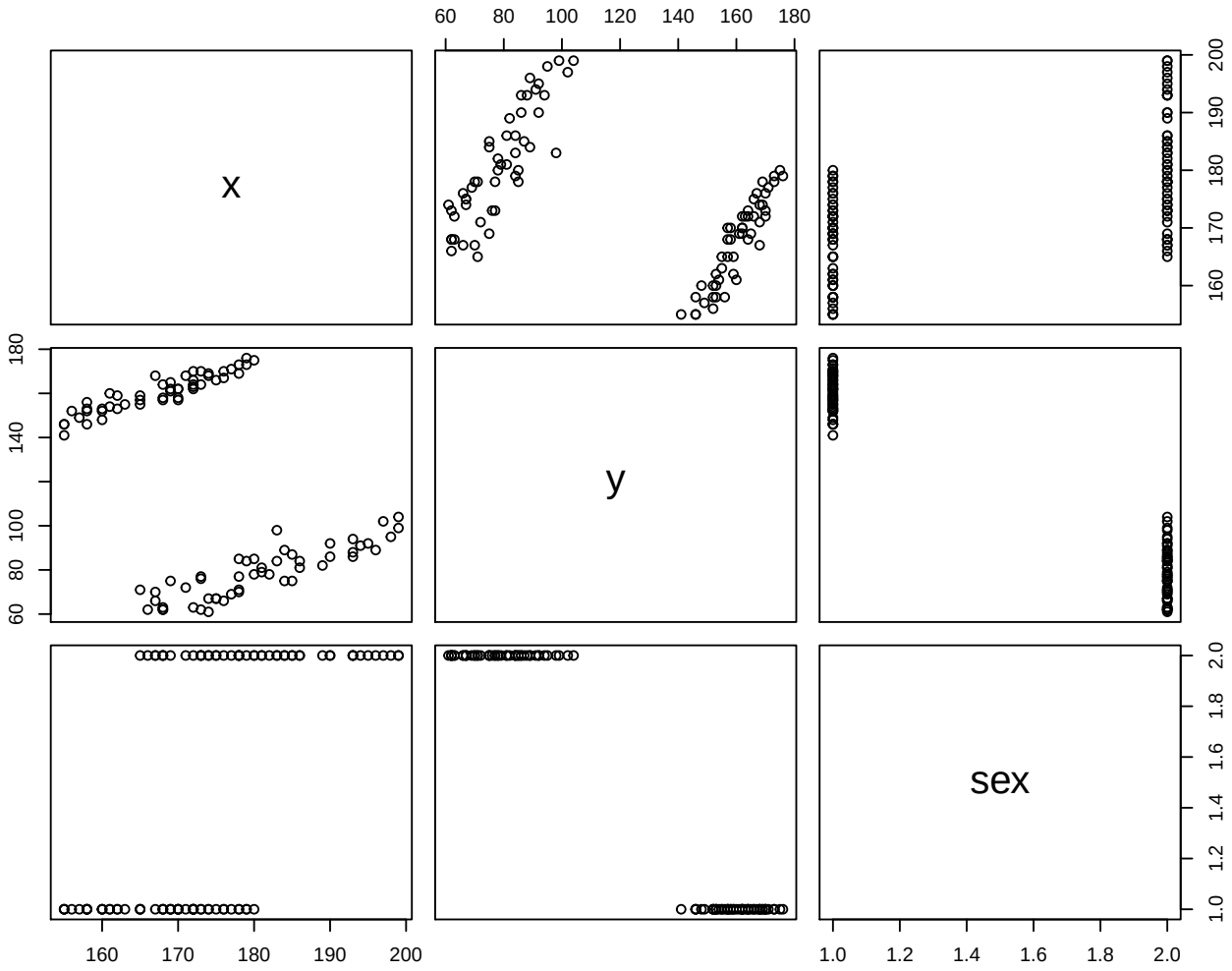


Figure 1: 数据散点图

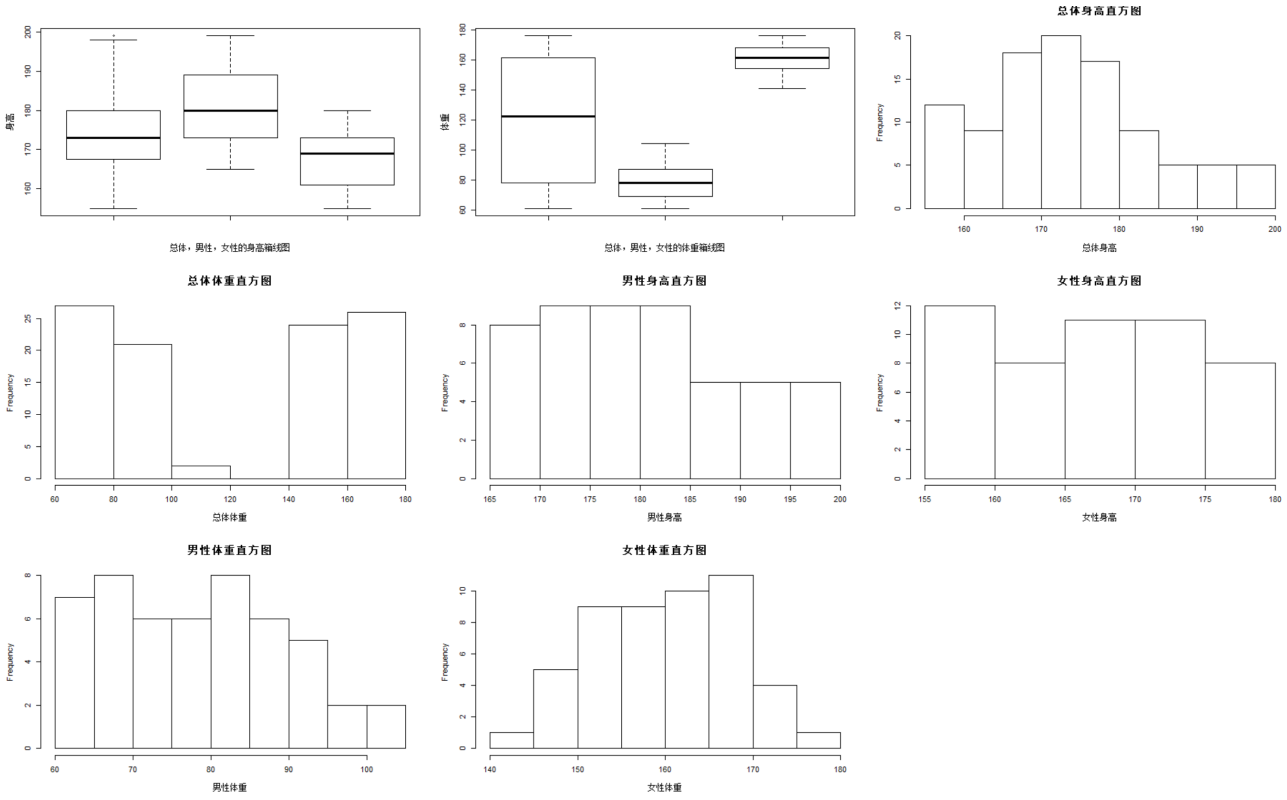


Figure 2: 箱线图和直方图

(2) 根据常识，你认为在变量 *x*, *y*, 和 *sex* 中，哪个变量可以作为响应变量，哪些变量可以作为协变量？并说明理由。（20 分）

从数据看可以看出此数据描述的是成年人的身高体重，所以可以选择体重作为响应变量，身高与性别作为协变量。

因为通常人的体重受到身高与性别的影响，而且成年人身高变化不会太大，体重会有较大变化，所以选择体重作为响应变量，身高与性别作为协变量。

(3) 令

$$\begin{cases} Z = 1 & male \\ Z = 0 & female. \end{cases}$$

我们可以建立如下的线性回归模型：

$$Y = \beta_0 + \beta_1 X + \alpha_0 Z + \alpha_1 XZ + \epsilon \tag{1}$$

其中，Y 表示身高，X 表示体重，Z 为性别变量，其定义方法如 (1).ϵ 是随机误差，通常我们假定随机误差为零均值，方差为 σ^2 的独立同分布。但是在这个数据上面，我们通过 (1) 和 (2) 中的描述性分析，对于男生和女生，其方差很有可能是不一样的，因此我们要在下面对于对随机误差是否异方差进行检验。

事实上, 我们上面这个模型是对于男性和女性分别进行建模：

$$\text{男性:} \quad Y = (\beta_0 + \alpha_0) + (\beta_1 + \alpha_1) X + \epsilon$$
$$\text{女性:} \quad Y = \beta_0 + \beta_1 X + \epsilon$$

我们首先对这个模型在 R 上简单做个回归，进行分析。

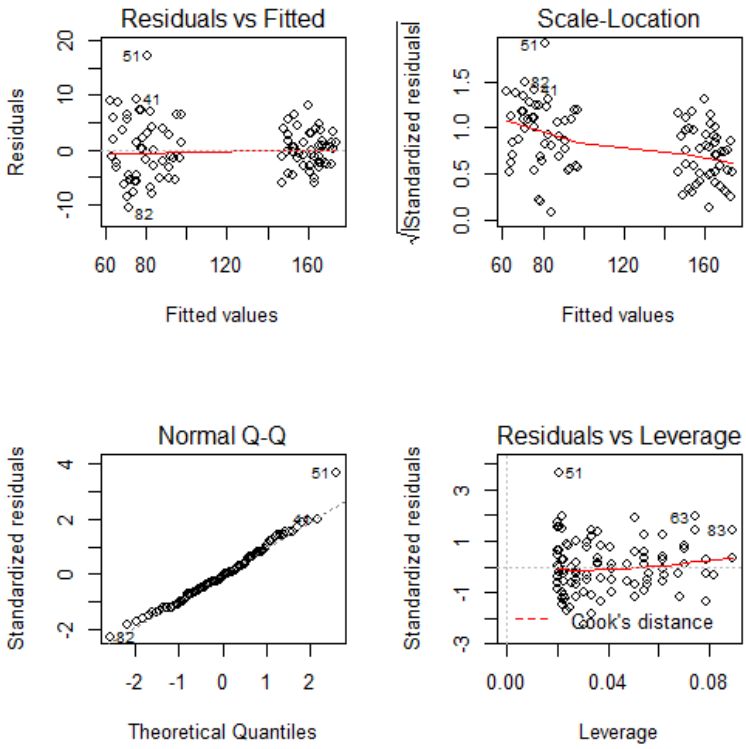


Figure 3: 残差图

(i) 我们首先做下面的检验

$$H_0 : \beta_1 = \alpha_0 = \alpha_1 = 0 \quad v.s. \quad H_1 : \beta_1, \alpha_0, \alpha_1 \text{中至少有一个非零}$$

这个检验是检验我们的数据是否有线性回归的关系，我们可以利用 F 统计量来进行检验，注意，在 H_0 下，我们的模型 (reduced model) 是

$$Y = \beta_0 + \epsilon$$

记 $SSE(R) = \sum \left(Y_i - \hat{Y}_i\right)^2$(SSE 即 sum of square error) 其中 $\hat{Y}_i$ 是通过上述模型得到的拟合值. 我们在全模型 (2) 中 (full model)，也有 $SSE(F) = \sum \left(Y_i - \hat{Y}_i\right)^2$(注意，这里的 $\hat{Y}$ 是在全模型下得到的拟合值)，我们的 F 统计量为

$$F^\star = \frac{SSE(R) - SSE(F)}{(n-1) - (n-4)} \div \frac{SSE(F)}{n-4}$$

上述检验统计量在 H_0 下服从自由度为 (3,n-4=96) 的 F 分布，我们直接使用 R 中的 lm() 函数，就可以做上述的检验, 在回归结果的最后一行即是 F 统计量的值，我们可以看到此时 p-value 远远小于显著性水平 $\alpha = 0.05$, 因此我们拒绝原假设。

(ii) 接下来我们关心的问题是不同的性别，身高对于体重的影响是否会不一样，即我们关心 α_1 是否等于 0.

$$H_0 : \alpha_1 = 0 \ \ v.s. \ \ H_1 : \alpha_1 \neq 0$$

对于这个问题，我们可以直接用 t 检验来做，看到回归结果中 coefficients 那里，对应于 X:ID01 的那一项的系数即是 α_1 的估计，可以看到其 p-value 达到 0.8272, 所以我们不能拒绝原假设，因此我们得到一个结论，即，不同的性别，身高对于体重的影响是一样的。

(iii) 同样我们关心性别因素是否会对体重造成影响，即

$$H_0 : \alpha_0 = 0 \ \ v.s. \ \ H_1 : \alpha_0 \neq 0$$

对于这个问题，我们也可以直接用 t 检验来做，看到回归结果中中 coefficients 那里，对应于 ID01 的那一项的系数即是 α_0 的估计，可以看到其 p-value 达到 1.2e-05, 远远小于显著性水平 $\alpha = 0.05$, 所以我们拒绝原假设，因此我们得到一个结论，即，性别因素对于体重是有影响的。

(iv) 我们还关心身高因素是否会对体重造成影响，即

$$H_0 : \beta_1 = 0 \ \ v.s. \ \ H_1 : \beta_1 \neq 0$$

对于这个问题，我们也可以直接用 t 检验来做，看到回归结果中 coefficients 那里，对应于 x 的那一项的系数即是 β_1 的估计，可以看到其 p-value<1.2e-015, 远远小于显著性水平 $\alpha = 0.05$, 所以我们拒绝原假设，因此我们得到一个结论，即，身高因素对于体重是有影响的。

(v) 最后，我们要进行模型检验，即我们的模型假设是不是合理的。请注意，我们做最小二乘法的时候，一般假定随机误差的方差是相同的，但我们从上面的残差图 (Figure3) 来看，男性和女性的随机误差的方差很有可能是不一样的，因此我们要做异方差检验，我们令男女性的随机误差的方差分别为 σ_M^2 和 σ_F^2, 则我们的假设检验是：

$$H_0 : \sigma_M^2 = \sigma_F^2, \ \ v.s. \ \ H_1 : \sigma_M^2 \neq \sigma_F^2$$

我们可以用 Breusch-Pegan 检验来做，这个检验的想法是假定随机误差的方差与性别有关：满足

$$log_\epsilon \sigma_i^2 = \gamma_0 + \gamma_1 Z_i$$

因此上述检验可以转化为检验 γ_1 是否等于 0，我们要先对模型 (2) 做回归，并用回归得到的残差带入上式进行回归，再检验 γ_1 是否等于 0。R 中的 lmtest 宏包提供了检验函数 bptest(), 我们在 R 中做检验得到 p-value 远远小于显著性水平 $\alpha = 0.05$, 因此我们拒绝原假设. 因此我们的随机误差存在异方差。

当随机误差存在异方差时，我们可以使用加权最小二乘法来估计，即我们假定 $\epsilon \sim N(\mathbf{0}_{n\times 1}, \Sigma)$, 并记 $U = (\mathbf{1}, X, Z, XZ)$ 为我们的设计矩阵 (其中 $\mathbf{1} = (1, 1, \cdots, 1)^T \in \mathbb{R}^{n\times 1}$), 记 $\theta = \left(\beta_0, \beta_1, \alpha_0, \alpha_1\right)^T$ 为

真实的回归系数. 特别的，在我们的模型中，Σ 是对角矩阵，对角元只能是 σ_M^2 或者 σ_F^2. 我们考虑线性回归模型

$$\Sigma^{-1/2}Y = \Sigma^{-1/2}U\theta + \Sigma^{-1/2}\epsilon$$

则此时，$\Sigma^{-1/2}\epsilon \sim N(\mathbf{0}_{n\times 1}, \mathbf{I}_{n\times n})$, 我们就解决了异方差的问题。

但现在我们并不知道 σ_M^2 和 σ_F^2 的真实值，因此我们用模型 (2) 中回归得到的残差来得到这两者的估计 $\hat{\sigma}_M^2$ 和 $\hat{\sigma}_F^2$, 并因此得到矩阵 $\hat{\Sigma}$, 我们考虑对下面的模型进行线性回归：

$$\hat{\Sigma}^{-1/2}Y = \hat{\Sigma}^{-1/2}U\theta + \hat{\Sigma}^{-1/2}\epsilon \tag{2}$$

在 R 中我们进行检验，并得到对应的结果。

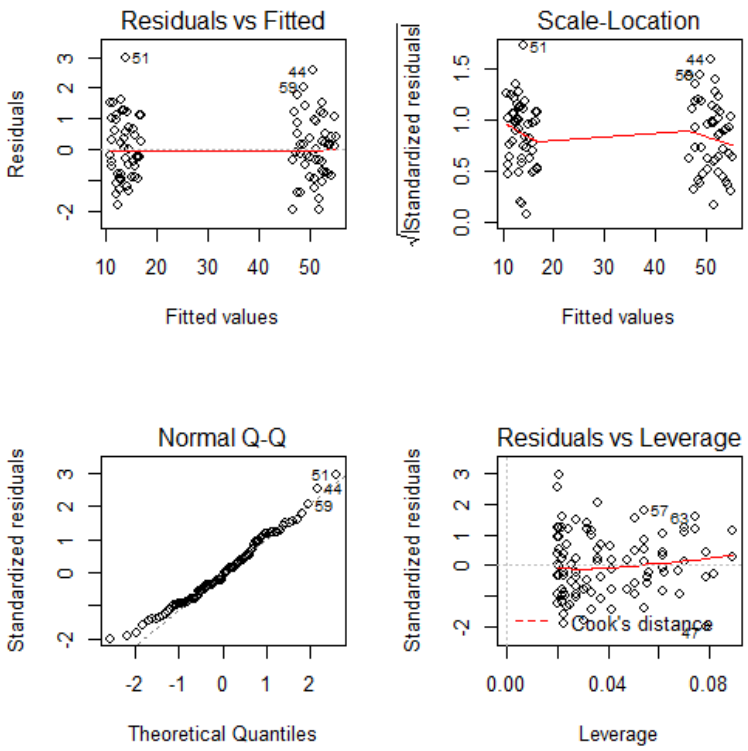


Figure 4: 残差图

我们来看我们是否解决了异方差的问题，从残差图来看，我们的随机误差应该没有异方差，我们还可以用上面提到的 BP test 来检验，并得到如下结果：由 r 程序中 BP test 的结果来看，p-value=1, 因此我们认为这时候我们的随机误差是同方差的。

(4)(20 分) 解读下面 R 命令的含义：

(i) datM=dat[dat$sex==’male’]

解：提取性别为男性的人的身高和体重数据。

(ii) datF=dat[dat$sex==’female’]

解：提取性别为女性的人的身高和体重数据。

(iii)

解：通过在 r 中查看回归结果，即 summary(L01),summary(L12),summary(Lfactor) 可以看出 (a) 与 (c) 的回归结果没有差别，而 (b) 的回归结果与上述两者有差别。

(a) 与 (c) 之所以没有差别，是因为它们对于性别这一项采取了相同的编码模式，在 (c) 中，factor(male)=1 而 factor(female)=0. 而 (b) 之所以不同，是因为它对性别这一项采用了不同的编码方式，(b) 将 male 编码为 1 而将 female 编码为 2.

对于 L01, 男性与女性的这两者的线性回归输出结果分别为：

$$\begin{aligned} &\text{男性: } \hat{y} = \hat{\alpha}_{01} + \hat{\beta}_{01} x + \hat{\gamma}_{01} \\ &\text{女性: } \hat{y} = \hat{\alpha}_{01} + \hat{\beta}_{01} x \end{aligned} \tag{3}$$

而对于 L12, 模型的描述有不同

$$\begin{aligned} &\text{男性: } \hat{y} = \hat{\alpha}_{12} + \hat{\beta}_{12} x + \hat{\gamma}_{12} \\ &\text{女性: } \hat{y} = \hat{\alpha}_{01} + \hat{\beta}_{01} x + 2\hat{\gamma}_{12} \end{aligned} \tag{4}$$

在 L01 中，$\hat{\gamma}_{01}$ 描述的是男性的体重比女性的体重平均多的程度，而在 L12 中，$\hat{\gamma}_{12}$ 描述的是男性的体重比女性的体重平均少的程度，因此我们有

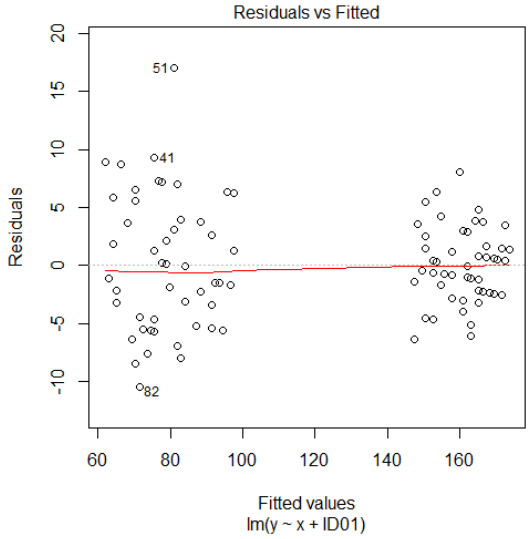
$$\begin{aligned} &\hat{\gamma}_{01} = -\hat{\gamma}_{12} \\ &\hat{\alpha}_{01} + \hat{\gamma}_{01} = \hat{\alpha}_{12} + \hat{\gamma}_{12} \\ &\hat{\beta}_{01} = \hat{\beta}_{12} \end{aligned} \tag{5}$$

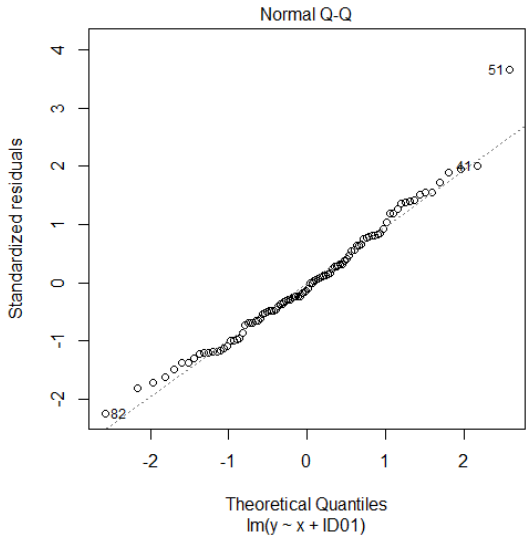
我们可以看到 summary(L01),summary(L12),summary(Lfactor) 中也和我们的分析结果是一样的。

我们认为这三个模型都是合理的，因为不同的编码模式不应该对模型造成本质上的区别。

(iv)

解：Lx 的回归模型为 $Y = \alpha + \beta x + \epsilon$,Lx 描述的是身高和体重的关系，Lsex 的回归模型为 $Y = \alpha + \epsilon$ 描述的是平均身高,Lx 意味着我们假定身高只与体重有关，与性别无关，Lsex 意味着我们假定身高与体重和性别都没有关系。





从 QQ 图来看，基本是一条直线，这意味着基本我们可以认为误差分布符合我们线性回归模型的假定服从正态分布。但是从残差图来看，我们明显观察到存在异方差的现象。L01,L12,Lfactor 这三个模型可以认为还算比较好的描述了 y,x 和 sex 之间的关系，Lx 和 Lsex 则描述的不够好。我们还可以做这样的改进，正如我们在第 (3) 问提到的，我们可以对模型做加权最小二乘拟合，这样子就可以解决异方差的问题