回归分析-第二次上机

南开大学统计与数据科学学院，马东升

2024 年 11 月 12 日

1 第（1）问

1.1 题目

研究高峰时期居民家庭每小时的用电量 Y 与每月总用电量 X 之间的关系。

（1）用最小二乘法求经验回归方程。

1.2 源代码（R）

```r
library(readxl)
data <- read_excel("3-15.xlsx", sheet = 1)
head(data)

model <- lm(data[[3]] ~ data[[2]], data = data)
summary(model)
```

1.3 统计分析结论

代码输出的结果为：

$$\hat{Y} = -0.7880079 + 0.0036186X \tag{1}$$

我们考虑显著性，方程整体的 F 值是 114.9，对应 p 值为 $1.146e-14$，这说明回归方程整体是显著的。

而对于 X 的系数来说，t 值是 10.719，对应 p 值为 $1.15e-14$，显然也是显著的。但其截距项 t 值为 -1.751，对应 p 值为 0.0859，在 $\alpha = 0.05$ 的意义下不显著。

2 第（2）问

2.1 题目

（2）以拟合值 $\hat{y}_i$ 为横坐标，学生化残差 r_i 为纵坐标，作残差图，分析高斯-马尔克斯假设对本例的适用性。

2.2 源代码（R）

```r
fitted_values <- fitted(model)          # 拟合值 yi_hat
stud_residuals <- rstudent(model)       # 学生化残差

plot(fitted_values, stud_residuals,
```

```
        xlab = expression(hat(y)[i]),
        ylab = "Studentized Residuals",
        main = "Fitted Values vs. Studentized Residuals")
abline(h = 2, col = "black", lty = 2)
abline(h = -2, col = "black", lty = 2)
abline(h = 0, col = "red", lty = 2)
```

2.3 统计分析结论

我们作出的残差图如下图 1 所示：

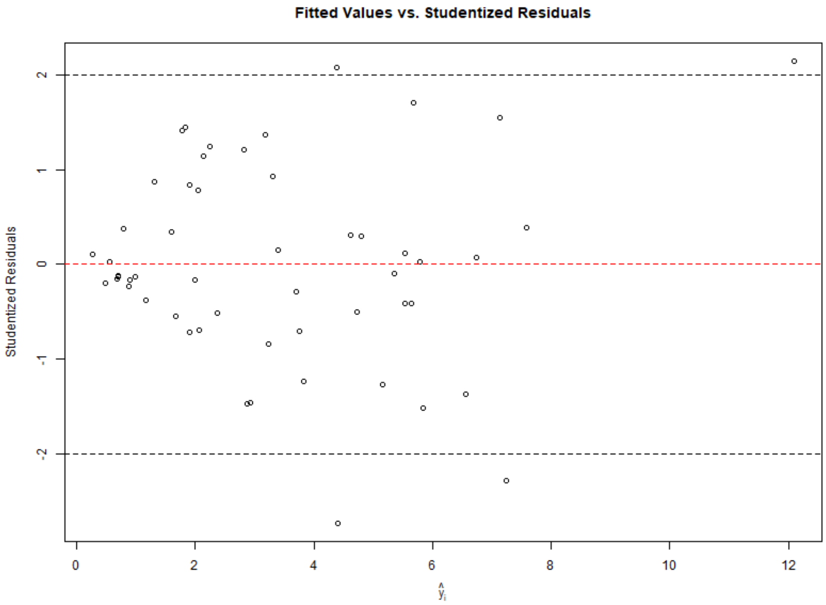

图 1: 直接回归所得的学生化残差图

我们需要考虑以下两点：

1. r_i 的绝对值似乎与 $\hat{y}_i$ 正相关：特别是在 $\hat{y} \in [0, 2]$ 的区间内，r_i 似乎特别的小；而其后的区间，有一种 r_i 的绝对值随 $\hat{y}_i$ 增大的趋势。
2. 大概有 4 个点（约占总体的 7.5%）的 r_i 的绝对值大于 2，相对于理论值（4.6%）还是偏大。

则这两点可以反映出，模型 (1) 可能**不太符合高斯-马尔可夫假设**，所以我们可能还是需要做一些模型的修改。

3 第（3）问

3.1 题目

考虑 $U = Y^{1/2}$，再对 U 和 X 做（1）（2）的统计分析。

3.2 源代码（R）

```
data$V4 <- sqrt(data[[3]])

```

```r
model2 <- lm(data[[4]] ~ data[[2]], data = data)
summary(model2)

fitted_values2 <- fitted(model2)          # 拟合值 yi_hat
stud_residuals2 <- rstudent(model2)       # 学生化残差

plot(fitted_values2, stud_residuals2,
     xlab = expression(hat(y)[i]),
     ylab = "Studentized Residuals",
     main = "Fitted Values vs. Studentized Residuals")
abline(h = 2, col = "black", lty = 2)
abline(h = -2, col = "black", lty = 2)
abline(h = 0, col = "red", lty = 2)
```

3.3 统计分析结论

代码输出的结果为：

$$\hat{U} = 0.5896 + 0.0009396X \tag{2}$$

这一模型的 F 值、两个系数的 t 值所对应的 p 值都极小，非常显著，在此不再赘述。

我们将 Y 回代，得到 Y 与 X 之间的经验回归方程：

$$\hat{Y} = (0.5896 + 0.0009396X)^2 \tag{3}$$

我们做出的残差图如下图 2 所示：

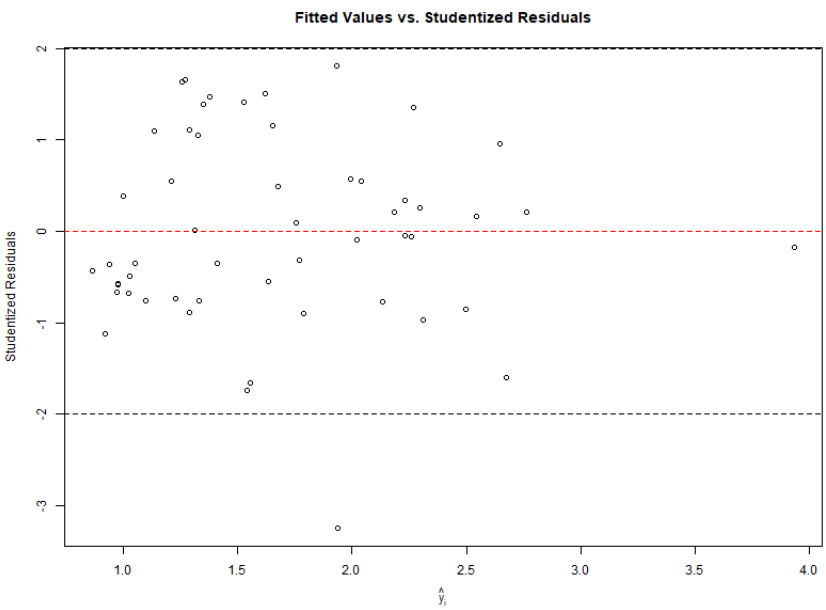

图 2: 变换后所得的学生化残差图

我们发现这张图上，r_i 与 $\hat{y}_i$ 几乎没有任何趋势，而且只有 1 个点的 r_i 的绝对值大于 2（约占 1.9%），所以相对于图 1 所对应的模型，我们认为**变换后的模型更符合高斯-马尔科夫假设**。

4 第（4）问

4.1 题目

将 Box-Cox 变换应用到本例，计算变换参数 λ 的值，并作讨论。

4.2 源代码（R）

```r
library(MASS)
model <- lm(data[[3]] ~ data[[2]], data = data)

# 获取 lambda 值和对应的对数似然值
bc <- boxcox(model, lambda = seq(-2, 2, by = 0.01))
lambda_values <- bc$x           # lambda 值
log_likelihoods <- bc$y         # 对应的对数似然值

best_lambda <- lambda_values[which.max(log_likelihoods)]

# 输出结果
cat("最大对数似然值对应的 lambda =", best_lambda, "\n")

# 将 lambda 值和对数似然值组合成数据框
lambda_loglik_table <- data.frame(
  Lambda = lambda_values,
  Log_Likelihood = log_likelihoods
)

# 查看表格
print(lambda_loglik_table)

model3 <- lm(data[[3]]^0.53 ~ data[[2]], data = data)
summary(model3)
```

4.3 统计分析结论

在步长为 0.01 的情况下，代码输出的最佳 λ 值为 0.53。由于代码自带的是对应的对数似然值，因而应该是对数似然值越大，λ 越好，图像见图 3。

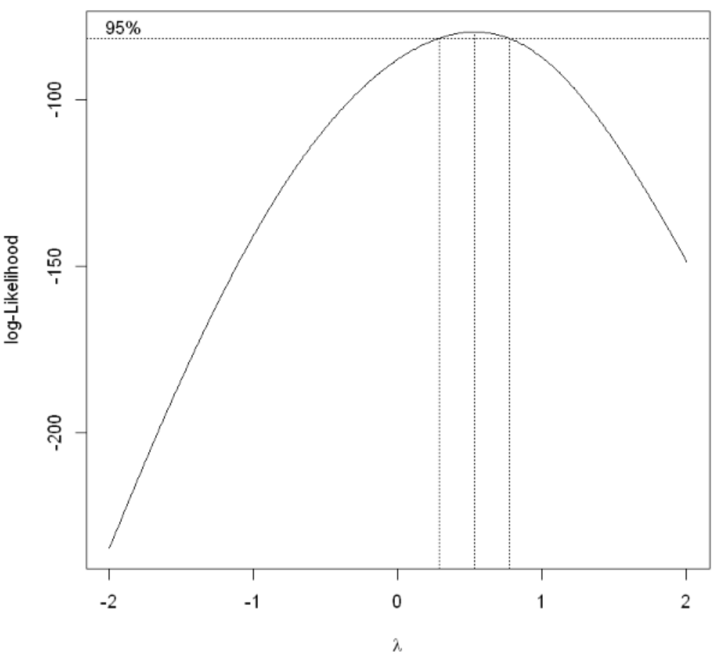

图 3: Box-Cox 中，不同 λ 对应的对数似然值

我们用 $\lambda = 0.53$ 得到下述经验回归方程：

$$\hat{Y}^{0.53} = 0.5483472 + 0.0010311X \tag{4}$$

各部分都显著，这里不多赘述。

但我们发现，(2) 式与 (4) 式中，X 的系数非常接近，截距项也差距不大，且 0.5 相对 0.53 简单且易解释，**所以我们选择 $\lambda = 0.5$ 更佳。**

最后我们附上不同 λ 对应的对数似然值，如表 1 所示。我们发现 0.5 和 0.53 差距确实不大，那么选择 0.5 也比较合理。

表 1: 不同 λ 对应的对数似然值

λ	-2	-1.5	-1	-0.5	0	0.25
lnL	-234.59876	-183.77788	-140.64887	-108.36754	-87.92431	-82.16291
λ	0.5	0.53	0.6	1	1.5	2
lnL	-79.74882	-79.71255	-79.85804	-87.44547	-112.57258	-148.53872

5 第（5）问

5.1 题目

作影响分析，找出强影响点。

5.2 源代码（R）

```
cooks_d <- cooks.distance(model)
cooks_d2 <- cooks.distance(model2)

```

```
data$r1 <- stud_residuals
data$r2 <- stud_residuals2
data$cook1 <- cooks_d
data$cook2 <- cooks_d2

data_sorted <- data[order(data[[7]], decreasing = TRUE), ]
data_sorted

data_sorted <- data[order(data[[8]], decreasing = TRUE), ]
data_sorted
```

5.3 统计分析结论

如果我们使用 (1) 式对应的模型，那么我们得出按 Cook 统计量大小排序前 6 的数据如表 2 所示，注意从第 7 开始均小于 0.05，故不予列出。

表 2: 原模型按 Cook 统计量大小排序前 6

用户	X	Y	D_i
50	3560	14.94	$8.037786e-01$
52	2221	3.85	$1.756379e-01$
8	2189	9.50	$8.098912e-02$
26	1434	0.31	$7.487420e-02$
49	1787	8.33	$5.278705e-02$
14	2030	4.43	$5.071242e-02$

注意到**用户 50** 对应的数据，$D_i = 0.8037786$，远远大于其它数据所得出的 Cook 统计量，则其是强影响点，需格外注意。当然表 2 剩余的数据，影响也比较大，可以稍加注意。

如果我们使用 (3) 式对应的模型，那么我们得出按 Cook 统计量大小排序前 5 的数据如表 3 所示。

表 3: 原模型按 Cook 统计量大小排序前 5

用户	X	Y	D_i
26	1434	0.31	$1.005740e-01$
52	2221	3.85	$9.153045e-02$
38	724	4.10	$3.745825e-02$
25	710	4.00	$3.707734e-02$
30	1428	7.58	$3.542935e-02$

注意到**用户 26、52** 对应的 D_i较大， 远大于其它数据所得出的 Cook 统计量，则这两个点是强影响点，需格外注意。其余的数据均影响较小。