1 第一问

源代码：

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
df=pd.read_excel("3-15.xlsx")
X=df.iloc[:,1]
y=df.iloc[:,2]
import statsmodels.api as sm
X=sm.add_constant(X)
ols=sm.OLS(y,X)
models=ols.fit()
models.params
```

out:

```
const   -0.788008
X        0.003619
dtype: float64
```

图 1: 最小二乘回归结果

统计分析：

导入数据后将每小时用电量 y 作为因变量，将每月总用电量 X 作为自变量。从图 1 可以读取最小二乘回归估计的结果，则经验回归方程为：

$$y = 0.0036x - 0.7880$$

2 第二问

源代码：

```
y_predict=models.predict()
outliers=models.get_influence()
ri=outliers.resid_studentized_internal
plt.plot(y_predict,ri,'b.')
plt.axhline(y=2,color="r",linestyle="--")
plt.axhline(y=-2,color="r",linestyle="--")
```

```
plt.xlabel("$\hat{y_i}$")
plt.ylabel("$\hat{r_i}$")
```

out:

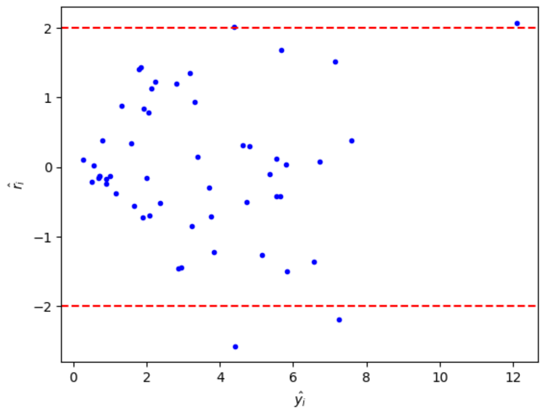

图 2: 学生化残差图

统计分析：

学生化残差：

$$r_i = \frac{\hat{e}}{\hat{\sigma}\sqrt{1-h_{ii}}}$$

其中，$\hat{e} = y - X\hat{\beta}$，$\hat{\sigma} = \frac{\hat{e}'\hat{e}}{n-1}$，$H = X(X'X)^{-1}X' \triangleq (h_{ij})$

从图 2 中可以看出，几乎所有点都落在 [-2,2] 的区间内，所以 Gauss-Markov 假设对本例适用。

3 第三问

源代码：

```
import numpy as np
X=df.iloc[:,1]
y=df.iloc[:,2]
u=np.sqrt(y)
import statsmodels.api as sm
X=sm.add_constant(X)
ols1=sm.OLS(u,X)
models1=ols1.fit()
models1.params
```

out:

```
const    0.589569
X        0.000940
dtype: float64
```

图 3: 最小二乘回归结果

```python
u_predict=models1.predict()
outliers1=models1.get_influence()
ri1=outliers1.resid_studentized_internal
plt.plot(u_predict,ri1,'b.')
plt.axhline(y=2,color="r",linestyle="--")
plt.axhline(y=-2,color="r",linestyle="--")
plt.xlabel("$\hat{y_i}$")
plt.ylabel("$\hat{r_i}$")
```

out:

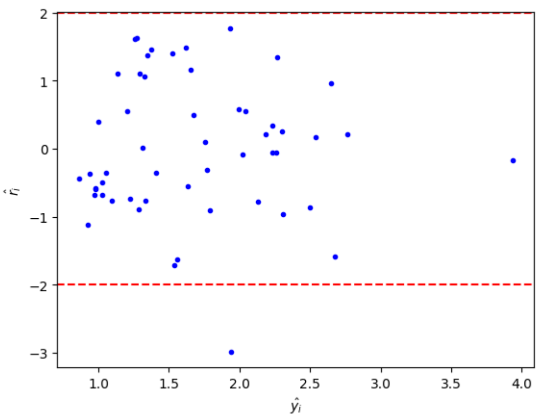

图 4: 学生化残差图

统计分析：

由图 3 可以读取最小二乘估计的结果，则经验回归方程为：

$$u = 0.0009x + 0.5896$$

学生化残差：

$$r_i = \frac{\hat{e}}{\hat{\sigma}\sqrt{1-h_{ii}}}$$

其中，$\hat{e} = u - X\hat{\beta}$，$\hat{\sigma} = \frac{\hat{e}'\hat{e}}{n-1}$，$H = X(X'X)^{-1}X' \triangleq (h_{ij})$

从图 4 中可以看出，几乎所有点都落在 [-2,2] 的区间内，所以 Gauss-Markov 假设对本例适用。

4 第四问

源代码：

```
library(xlsx)
df<-read.xlsx("3-15.xlsx",1)
X=as.matrix(df[,2])
y=as.matrix(df[,3])
library(MASS)
bc<-boxcox(Y~X,data=df,lambda=seq(0,1,0.01))
lambda<-bc$x[which.max(bc$y)]
lambda
```

out: 0.53

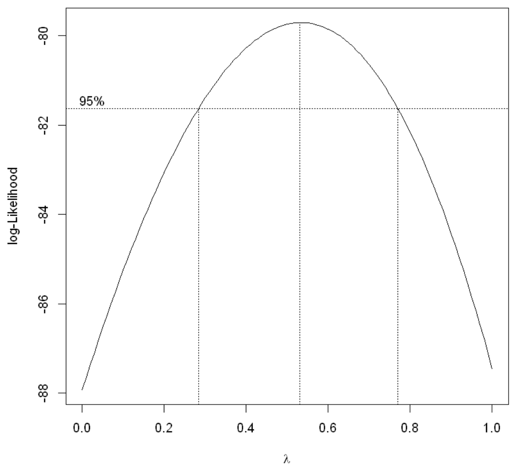

图 5: 用似然法估计 λ

统计分析：

从程序的输出结果可以看出，用似然法估计变换参数 λ 的结果为 $\hat{\lambda} \approx 0.53$

Box-Cox 变换：

$$Y^{(\lambda)} = \begin{cases} \dfrac{Y^{\lambda}-1}{\lambda}, \lambda \neq 0 \\ \ln Y, \lambda = 0 \end{cases}$$

用似然法估计变换参数 λ：

在完成 Box-Cox 变换后，有 $y^{(\lambda)} \sim N(X\beta, \sigma^2 I_n)$，可得到似然函数为 $L(\beta, \sigma^2) = \frac{1}{(\sqrt{2\pi}\sigma)^n} exp(-\frac{(y^{(\lambda)}-X\beta)'(y^{(\lambda)}-X\beta)}{2\sigma^2})J$，其中雅可比行列式 $J = \prod_{i=1}^{n} |\frac{\mathrm{d}y_i^{(\lambda)}}{\mathrm{d}y_i}|$。容易求得两个参数的 MLE 分别为：$\hat{\beta} = (X'X)^{-1}X'y^{(\lambda)}$ 和 $\hat{\sigma}^2(\lambda) = -\frac{1}{n}RSS(\lambda, y^{(\lambda)})$，记 $z^{(\lambda)} = \frac{y^{(\lambda)}}{J^{\frac{1}{n}}}$。代入

对数似然中有：$\ln L_{max}(\lambda) = -\frac{n}{2}\ln RSS(\lambda, z^{(\lambda)})$。综上，只需找到 λ 使得 $RSS(\lambda, z^{(\lambda)})$ 最小则可。

5 第五问

源代码：

```
cook=outliers.cooks_distance
plt.plot(cook[0],'b.')
plt.axvline(x=7,color="g",linestyle="--")
plt.axvline(x=25,color="g",linestyle="--")
plt.axvline(x=49,color="g",linestyle="--")
plt.axvline(x=51,color="g",linestyle="--")
```

out:

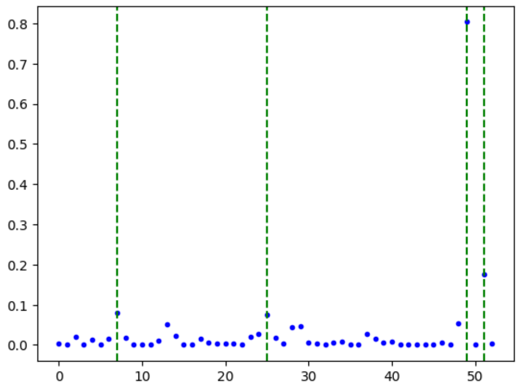

图 6: cook 距离图

统计分析：

从图 6 中可以看出，第 8、26、50、52 号点 cook 距离相对较大，尤其是第 50、52 号点。可以认为第 8、26、50、52 号点对应的数据是强影响点。

cook 距离求法：

$$D_i = \frac{1}{p}\left(\frac{h_{ii}}{1-h_{ii}}\right)r_i^2$$