作业1：回归函数的推导：

样本 $\langle x_i, y_i \rangle$，考虑 m 个样本

$y = \beta_0 + \beta_1 x$，求最佳拟合直.

OLSE：(最小二乘)

对于单个样本 $\langle x_i, y_i \rangle$

预测值 $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

对于 m 个样本，即：$Q = \sum^{m} (\hat{y}_i - y_i)^2$

则转换为求 Q 的最小值：

i.e. $\sum (\hat{\beta}_0 + \hat{\beta}_1 x_i - y_i)^2 = \min (\beta_0 + \beta_1 x_i - y_i)^2$，$\beta_0$ 在 $\beta_0 = \hat{\beta}_0$；β_1 在 $\beta_1 = \hat{\beta}_1$ 取最小值

对 Q：求 β_1 和 β_0 偏导：

$$\begin{cases} \dfrac{\partial \sum^{m} (\beta_0 + \beta_1 x_i - y_i)^2}{\partial \beta_0} = 0 \\ \dfrac{\partial \sum^{m} (\beta_0 + \beta_1 x_i - y_i)^2}{\partial \beta_1} = 0 \end{cases} \Rightarrow \begin{cases} 2\sum^{m} (\beta_0 + \beta_1 x_i - y_i) = 0 \quad ① \\ 2\sum^{m} (\beta_0 + \beta_1 x_i - y_i) x_i = 0 \quad ② \end{cases}$$

由方程组①、②，可得：

$$m\beta_0 + \sum^{m} x_i \beta_1 - \sum^{m} y_i = 0 \quad ③$$

对③，等号左右同除 m：

$$\beta_0 + \frac{\sum^{m} x_i}{m} \beta_1 - \frac{\sum^{m} y_i}{m} = 0 \quad ④$$

故 $\beta_0 = \bar{y} - \beta_1 \bar{x}$ ⑤

由②，

$$\beta_0 \sum^{m} x_i + \beta_1 \sum^{m} x_i^2 - \sum^{m} y_i x_i = 0 \quad ⑥$$

将⑤代入⑥，

$$\bar{y} \sum^{m} x_i - \beta_1 \bar{x} \sum^{m} x_i + \beta_1 \sum^{m} x_i^2 - \sum^{m} y_i x_i = 0$$

$$\bar{y} \sum^{m} x_i - \sum^{m} y_i x_i = \beta_1 (\bar{x} \sum^{m} x_i - \sum^{m} x_i^2)$$

$$\begin{cases} \bar{x} \sum^{m} (\bar{x} - x_i) = 0 \\ \bar{x} \sum^{m} (\bar{y} - y_i) = 0 \end{cases}$$

$$\Rightarrow \beta_1 = \frac{\bar{y} \sum^{m} x_i - \sum^{m} y_i x_i}{\bar{x} \sum^{m} x_i - \sum^{m} x_i^2} = \frac{\sum^{m} (\bar{y} - y_i)(\bar{x} - x_i)}{\sum^{m} (\bar{x} - x_i)(\bar{x} - x_i)}$$

所以 $\hat{\beta}_1 = \dfrac{l_{xy}}{l_{xx}}$；$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

例2：R^2 和 corr 的关系推导

$$\rho(y_i, \hat{y}_i) = \frac{\mathrm{Cov}(y_i, \hat{y}_i)}{\sqrt{\mathrm{Var}(y_i)\mathrm{Var}(\hat{y}_i)}}$$

$$= \frac{\sum^n (y_i - \bar{y})(\hat{y}_i - \bar{y})}{\sqrt{\sum^n (y_i - \bar{y})^2 \sum^n (\hat{y}_i - \bar{y})^2}}$$

$$= \frac{\sum^n (y_i - \hat{y} + \hat{y} - \bar{y})(\hat{y}_i - \bar{y})}{\sqrt{\sum^n (y_i - \bar{y})^2 \sum^n (\hat{y}_i - \bar{y})^2}}$$

$$= \frac{\sum^n (y_i - \hat{y})(\hat{y}_i - \bar{y}) + \sum^n (\hat{y}_i - \bar{y})^2}{\sqrt{\sum^n (y_i - \bar{y})^2 \sum^n (\hat{y}_i - \bar{y})^2}}$$

$$= \frac{\sum^n (\hat{y}_i - \bar{y})^2}{\sqrt{\sum^n (y_i - \bar{y})^2 \sum^n (\hat{y}_i - \bar{y})^2}}$$

$$= \sqrt{\frac{\left[\sum^n (\hat{y}_i - \bar{y})^2\right]^2}{\sum^n (y_i - \bar{y})^2 \sum^n (\hat{y}_i - \bar{y})^2}}$$

$$= \sqrt{\frac{\sum^n (\hat{y}_i - \bar{y})^2}{\sum^n (y_i - \bar{y})^2}}$$

$$= \sqrt{R^2}$$

作业 3:

R^2 很高时，假设检验一定会是显著的吗?

从 P 值看假设检验显著时，R^2 一定高吗?

$$t = \frac{r - 0}{\sqrt{\frac{1 - r^2}{n - 2}}}$$

在一元回归中 F 检验和 T 检验等价。

R^2 衡量了模型的拟合优度，而假设检验(t 检验)则是从参数是否显著的角度出发（看 x 对 y 是否有明显的影响)。

所以 R^2 的假设检验的 p 值没有必然关系。

作业 4:

为什么数据被复制了一份，对假设检验的 p 值有影响，对假设检验的哪里有影响呢?

样本量越大，显著性水平需要降低设置(变严格)。

样本量变大,p 值变小，线性回归拟合的值不会变。

这题不是特别理解，猜测是跟中心极限定理大样本量类似的思想有关。样本量越大，我提出的原假设在更大样本量的情况下，由 p 值的变小说明我提出的原假设更有力度，即偶然性，拒绝原假设的概率会变小。

在数据复制一份的情况下，样本方差也会变小。

```
[1]
import numpy as np

[2]
x=[1,2,3,4,5]
x_duplicate=[1,2,3,4,5,1,2,3,4,5]

[3]
np.var(x, ddof = 1)

2.5

[4]
np.var(x_duplicate, ddof = 1)

2.2222222222222223
```