第一次作业参考答案

1. 在一元线性回归方程中，在假设一到假设五成立的情况下，请推导 OLS 估计量 $\widehat{\beta}_0$ 的方差

$$Var\left(\widehat{\beta}_0\middle|x\right)=\frac{\left(\sigma^2/n\right)\sum_{i=1}^n x_i^2}{\sum_{i=1}^n\left(x_i-\bar{x}\right)^2}$$

证明. 除了假设一到假设五，我们还需要利用

$$\widehat{\beta}_0=\bar{y}-\widehat{\beta}_1\bar{x},$$
$$\mathrm{Var}\left(\widehat{\beta}_1\middle|x\right)=\frac{\sigma^2}{\sum_{i=1}^n\left(x_i-\bar{x}\right)^2}.$$

首先，注意到

$$\widehat{\beta}_1=\sum_{i=1}^n w_i\left(y_i-\bar{y}\right)=\sum_{i=1}^n w_i y_i,$$

其中，$w_i=\left(x_i-\bar{x}\right)/\sum_{j=1}^n\left(x_j-\bar{x}\right)^2$。这是因为

$$\sum_{i=1}^n w_i\bar{y}=\bar{y}\underbrace{\sum_{i=1}^n w_i}_{=0}=0.$$

那么，

$$\mathrm{Cov}\left(\bar{u},\widehat{\beta}_1\middle|x\right)=\frac{1}{n}\sum_{k=1}^n\sum_{i=1}^n w_i\mathrm{Cov}\left(u_k,y_i|x\right)=\frac{1}{n}\sum_{k=1}^n w_k\sigma^2=\frac{\sigma^2}{n}\underbrace{\sum_{k=1}^n w_k}_{=0}=0,$$

其中，$\mathrm{Cov}\left(u_k,y_i|x\right)=\sigma^2\mathbb{1}\{k=i\}$，$\mathbb{1}\{\cdot\}$ 是指示函数 (indicator function)，所以有

$$\sum_{i=1}^n w_i\mathrm{Cov}\left(u_k,y_i|x\right)=w_k\sigma^2.$$

因此，

$$\begin{aligned}
\operatorname{Var}\left(\widehat{\beta}_0\middle|x\right) &= \operatorname{Var}\left(\bar{y}-\widehat{\beta}_1\bar{x}\middle|x\right) \\
&= \operatorname{Var}\left(\beta_0+\beta_1\bar{x}+\bar{u}-\widehat{\beta}_1\bar{x}\middle|x\right) \\
&= \operatorname{Var}\left(\bar{u}-\widehat{\beta}_1\bar{x}\middle|x\right) \\
&= \operatorname{Var}\left(\bar{u}|x\right)+\bar{x}^2\operatorname{Var}\left(\widehat{\beta}_1\middle|x\right)-2\bar{x}\underbrace{\operatorname{Cov}\left(\bar{u},\widehat{\beta}_1\middle|x\right)}_{=0} \\
&= \frac{\sigma^2}{n}+\frac{\sigma^2\bar{x}^2}{\sum_{i=1}^n\left(x_i-\bar{x}\right)^2} \\
&= \frac{\left(\sigma^2/n\right)\sum_{i=1}^n\left(x_i-\bar{x}\right)^2+\left(\sigma^2/n\right)n\bar{x}^2}{\sum_{i=1}^n\left(x_i-\bar{x}\right)^2} \\
&= \frac{\left(\sigma^2/n\right)\sum_{i=1}^n x_i^2}{\sum_{i=1}^n\left(x_i-\bar{x}\right)^2}
\end{aligned}$$

□

2. 在给定二元线性回归方程 $y=\beta_0+\beta_1x_1+\beta_2x_2+u$，以 β_1 为例，第一种估计系数的办法是用最小二乘法直接估计：第二种办法分成两步进行。第一步先拿 x_1 对 x_2 做回归，得到残差，第二步拿 y 对第一步中得到的残差做回归。请从数学上证明两种办法得到的 β_1 的估计量 $\widehat{\beta}_1$ 是等价的。

证明. 在第一种方法中，

$$\min_{\widehat{\beta}_0,\widehat{\beta}_1,\widehat{\beta}_2}\sum_{i=1}^n\left[y_i-\left(\widehat{\beta}_0+\widehat{\beta}_1x_{1i}+\widehat{\beta}_2x_{2i}\right)\right]^2$$

一阶条件：

$$\sum_{i=1}^n\left(y_i-\widehat{\beta}_0-\widehat{\beta}_1x_{1i}-\widehat{\beta}_2x_{2i}\right)=0 \tag{1}$$

$$\sum_{i=1}^n\left(y_i-\widehat{\beta}_0-\widehat{\beta}_1x_{1i}-\widehat{\beta}_2x_{2i}\right)x_{1i}=0 \tag{2}$$

$$\sum_{i=1}^n\left(y_i-\widehat{\beta}_0-\widehat{\beta}_1x_{1i}-\widehat{\beta}_2x_{2i}\right)x_{2i}=0 \tag{3}$$

由（1）我们有

$$\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x}_1 - \widehat{\beta}_2 \bar{x}_2$$

把上式代入（2）和（3）有

$$\sum_{i=1}^{n} \left[(y_i - \bar{y}) - \widehat{\beta}_1 (x_{1i} - \bar{x}_1) - \widehat{\beta}_2 (x_{2i} - \bar{x}_2) \right] x_{1i} = 0$$
$$\sum_{i=1}^{n} \left[(y_i - \bar{y}) - \widehat{\beta}_1 (x_{1i} - \bar{x}_1) - \widehat{\beta}_2 (x_{2i} - \bar{x}_2) \right] x_{2i} = 0$$

因此有

$$\sum_{i=1}^{n} \left[(y_i - \bar{y}) - \widehat{\beta}_1 (x_{1i} - \bar{x}_1) - \widehat{\beta}_2 (x_{2i} - \bar{x}_2) \right] (x_{1i} - \bar{x}_1) = 0 \quad (4)$$
$$\sum_{i=1}^{n} \left[(y_i - \bar{y}) - \widehat{\beta}_1 (x_{1i} - \bar{x}_1) - \widehat{\beta}_2 (x_{2i} - \bar{x}_2) \right] (x_{2i} - \bar{x}_2) = 0 \quad (5)$$

定义 $Dy_i \equiv y_i - \bar{y}$, $Dx_{1i} \equiv x_{1i} - \bar{x}_1$, $Dx_{2i} \equiv x_{2i} - \bar{x}_2$，式（4）和（5）可以改写为

$$\sum_{i=1}^{n} \left[Dy_i - \widehat{\beta}_1 Dx_{1i} - \widehat{\beta}_2 Dx_{2i} \right] Dx_{1i} = 0 \quad (6)$$
$$\sum_{i=1}^{n} \left[Dy_i - \widehat{\beta}_1 Dx_{1i} - \widehat{\beta}_2 Dx_{2i} \right] Dx_{2i} = 0 \quad (7)$$

(6) 和（7）构成了包含两个未知数、两个方程的线性方程组，因此

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n} Dy_i Dx_{1i} \sum_{i=1}^{n} (Dx_{2i})^2 - \sum_{i=1}^{n} Dy_i Dx_{2i} \sum_{i=1}^{n} Dx_{1i} Dx_{2i}}{\sum_{i=1}^{n} (Dx_{1i})^2 \sum_{i=1}^{n} (Dx_{2i})^2 - (\sum_{i=1}^{n} Dx_{1i} Dx_{2i})^2}$$

在第二种方法中，我们在第一步考虑

$$x_1 = \alpha_0 + \alpha_1 x_2 + v$$

所以有

$$\widehat{\alpha}_1 = \frac{\sum_{i=1}^{n} (x_{2i} - \bar{x}_2)(x_{1i} - \bar{x}_1)}{\sum_{i=1}^{n} (x_{2i} - \bar{x}_2)^2} = \frac{\sum_{i=1}^{n} Dx_{2i} Dx_{1i}}{\sum_{i=1}^{n} (Dx_{2i})^2} \quad (8)$$
$$\widehat{\alpha}_0 = \bar{x}_1 - \widehat{\alpha}_1 \bar{x}_2 \quad (9)$$

将（9）代入残差 $\widehat{v}_i \equiv x_{1i} - \widehat{\alpha}_0 - \widehat{\alpha}_1 x_{2i}$, 那么

$$\widehat{v}_i = Dx_{1i} - \widehat{\alpha}_1 Dx_{2i} \tag{10}$$

注意到 $\bar{\widehat{v}} = 0$，因此在第二步中我们有

$$\widetilde{\beta}_1 = \frac{\sum_{i=1}^n \widehat{v}_i (y_i - \bar{y})}{\sum_{i=1}^n \widehat{v}_i^2} = \frac{\sum_{i=1}^n \widehat{v}_i Dy_i}{\sum_{i=1}^n \widehat{v}_i^2} \tag{11}$$

将 (8) 和（10）代入（11），我们有

$$\begin{aligned}
\widetilde{\beta}_1 &= \frac{\sum_{i=1}^n (Dx_{1i} - \widehat{\alpha}_1 Dx_{2i}) Dy_i}{\sum_{i=1}^n (Dx_{1i} - \widehat{\alpha}_1 Dx_{2i})^2} \\
&= \frac{\sum_{i=1}^n Dx_{1i} Dy_i - \frac{\sum_{i=1}^n Dx_{2i} Dx_{1i}}{\sum_{i=1}^n (Dx_{2i})^2} \sum_{i=1}^n Dx_{2i} Dy_i}{\sum_{i=1}^n (Dx_{1i})^2 + \frac{\left(\sum_{i=1}^n Dx_{2i} Dx_{1i}\right)^2}{\left(\sum_{i=1}^n (Dx_{2i})^2\right)^2} \sum_{i=1}^n (Dx_{2i})^2 - 2\frac{\sum_{i=1}^n Dx_{2i} Dx_{1i}}{\sum_{i=1}^n (Dx_{2i})^2} \sum_{i=1}^n Dx_{1i} Dx_{2i}} \\
&= \frac{\sum_{i=1}^n Dy_i Dx_{1i} \sum_{i=1}^n (Dx_{2i})^2 - \sum_{i=1}^n Dy_i Dx_{2i} \sum_{i=1}^n Dx_{1i} Dx_{2i}}{\sum_{i=1}^n (Dx_{1i})^2 \sum_{i=1}^n (Dx_{2i})^2 - (\sum_{i=1}^n Dx_{1i} Dx_{2i})^2}
\end{aligned}$$

因此 $\widehat{\beta}_1$ 和 $\widetilde{\beta}_1$ 是等价的。 □

3. 子虚国政府委托乌有大学进行一个项目，主要目的是研究劳动力市场上决定劳动收入的因素。乌有大学通过问卷调查。得到一个数据集。数据集包括如下变量：（1）*gender*：性别，其中 1 代表男性，2 代表女性；（2）*birthyear*：出生年份；（3）*marriage*：婚姻状况，其中 1 代表处于婚姻状态，0 代表处于非结婚状态（包括未婚，离异，丧偶等）；（4）*empjob_twage*：年总收入；（5）*schooling_yr*：受教育年数；

   在打开 Stata 后，首先使用 `cd` 把工作目录设定为存放有数据集的目录，再使用 `use` 打开数据集。例如，

```
cd "your_path" /* e.g. cd "/home/zhufeng/metrics" */
use "homework1_dataset.dta"
```

   (1) 创建两个新变量：(a) *male*：1 代表男性，0 代表女性；(b) *female*：1 代表女性，0 代表男性。

```
. * Question 3.1
. generate male = 1 if gender == 1
(892 missing values generated)
. replace  male = 0 if gender == 2
(892 real changes made)
.
. generate female = 1 if gender == 2
(1,948 missing values generated)
. replace  female = 0 if gender == 1
(1,948 real changes made)
.
```

(2) 给出以下变量的均值, 标准差, 最小值以及最大值: $female$, $male$, $birthyear$, $marriage$, $empjob\_twage$, $schooling\_yr$。

```
. * Question 3.2
. summarize female male birthyear marriage empjob_twage schooling_yr
```

Variable	Obs	Mean	Std. Dev.	Min	Max
female	2,840	.3140845	.464232	0	1
male	2,840	.6859155	.464232	0	1
birthyear	2,840	1974.865	11.2741	1914	1998
marriage	2,840	.6010563	.4897674	0	1
empjob_twage	2,840	4789.801	3361.39	176.1308	50725.66
schooling_yr	2,840	7.624296	2.9349	0	15

```
.
```

(3) 乌有大学的飘渺教授认为教育对于收入有着重要的影响，她建议估计下面这个回归方程式:

$$empjob\_twage = \beta_0 + \beta_1 \times schooling\_yr + u$$

请使用给定的数据集估计这个方程，给出回归结果。

```
. * Question 3.3
. regress empjob_twage schooling_yr
```

Source	SS	df	MS		
				Number of obs	= 2,840
				F(1, 2838)	= 66.25
Model	731763187	1	731763187	Prob > F	= 0.0000
Residual	3.1346e+10	2,838	11045077.6	R-squared	= 0.0228
				Adj R-squared	= 0.0225
Total	3.2078e+10	2,839	11298941	Root MSE	= 3323.4

empjob_twage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
schooling_yr	172.9853	21.25242	8.14	0.000	131.3136	214.6571
_cons	3470.91	173.6213	19.99	0.000	3130.473	3811.347

(4) 计算 $empjob\_twage$, $\widehat{empjob\_twage}$, 和 $\widehat{u}$ 的均值。他们之间有什么关系。

```
. * Question 3.4
. predict yhat
(option xb assumed; fitted values)
. predict ehat, residuals
. summarize empjob_twage yhat ehat
```

Variable	Obs	Mean	Std. Dev.	Min	Max
empjob_twage	2,840	4789.801	3361.39	176.1308	50725.66
yhat	2,840	4789.801	507.6947	3470.91	6065.689
ehat	2,840	.0000148	3322.828	-5300.15	45697.88

```
.
```

因此，$\overline{empjob\_twage} = \overline{\widehat{empjob\_twage}} + \overline{\widehat{u}}$。