# 社会统计学及SPSS软件应用

Instructor:王荣欣

Email: rxwang@qq.com

周一3-4节、单周周四3-4节，3A106-2

2020年11月2日

# CONTENTS

# CONTENTS

$\hat{Y}$=a+b（年龄）+c（年龄平方）

1 生成平方项 gen age2= age^2 or gen age2 = age*age
2 再进行回归 reg inc age age2

*教材第4章第71页
*教材第7章第133页

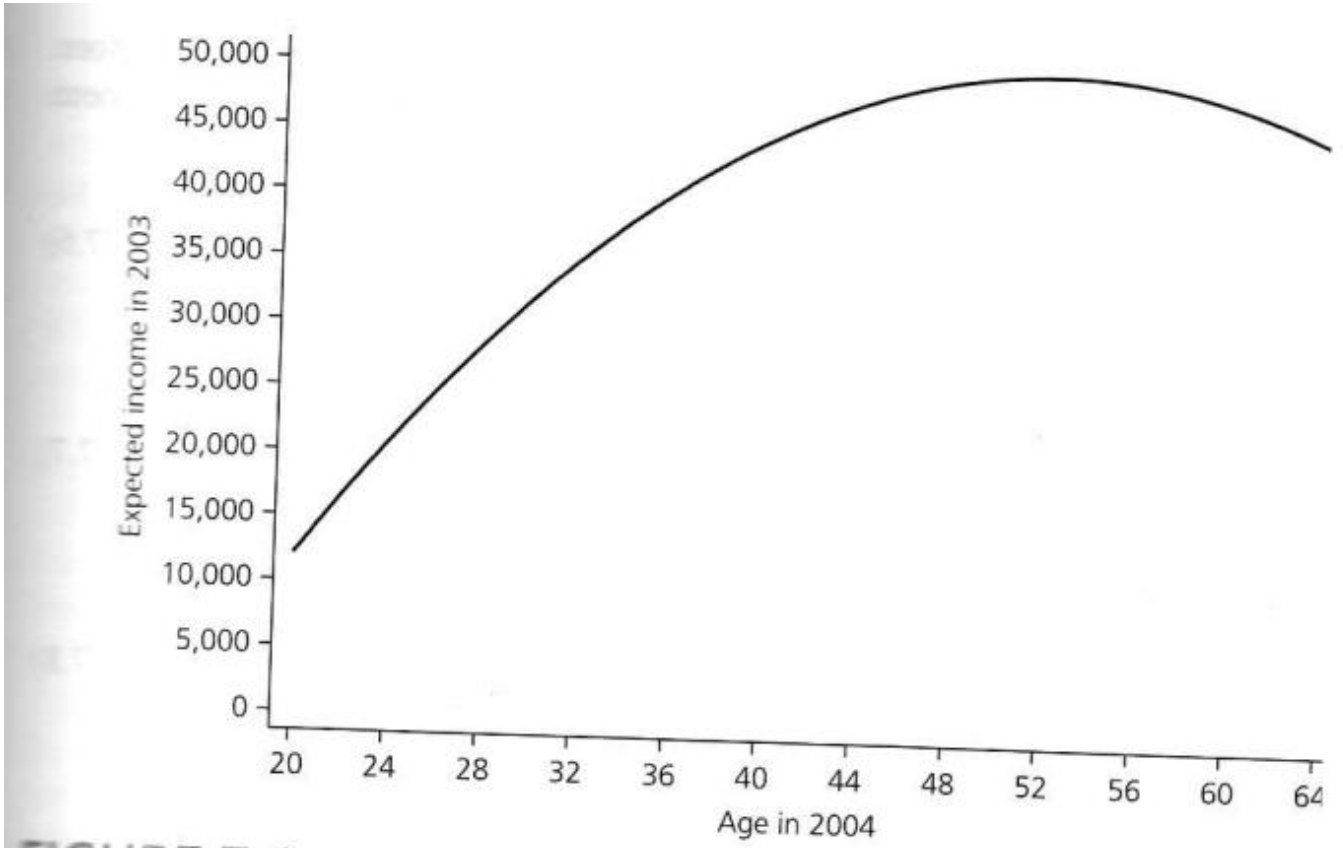



**FIGURE 7.1.** *The Relationship Between 2003 Income and Age, U.S. Adults Age Twenty to Sixty-Four in 2004 (N = 1,573).*

$ln\hat{Y}$=a+b（教育年限）+c（每周工作小时数）+d（男性）

1. 对收入取对数 gen lninc=ln(inc)
2. 再进行回归 reg lninc educ hrs male

*natural logarithm (abbreviated ln)

# 交互效应（INTERACTION EFFECTS）

1 理解方式一：两个自变量合在一起时，对因变量具有额外的效应。

eg: 在估计收入的模型中，“党员”和“男性”的效应可能都是正的，但两者结合在一起（即“男性党员”）可能有额外的效应。

2 理解方式二：一个变量的效应受到另外一个变量的值的影响，称这两个变量之间具有交互效应。

*教材第1章第25-27页

引入哑变量，以及哑变量与解释变量的“互动项”(interaction term)：

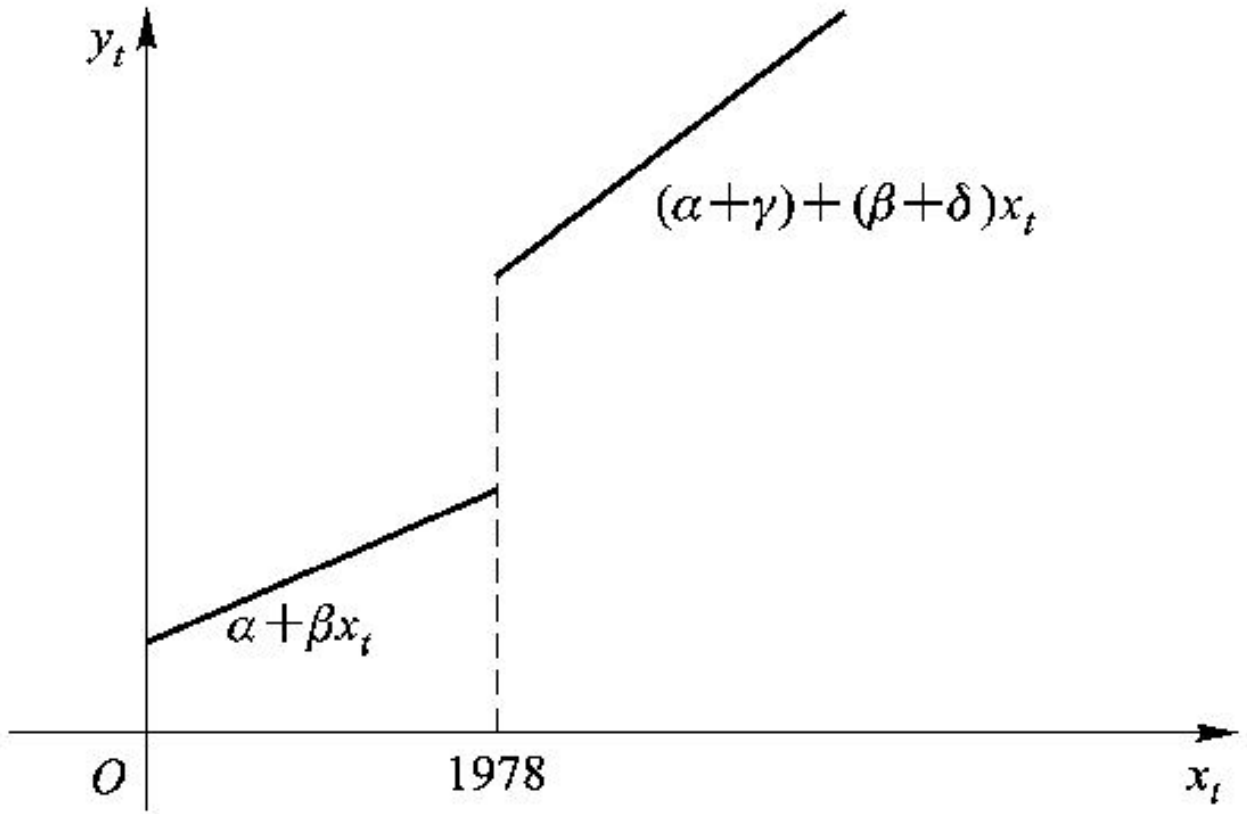
$$y_t = \alpha + \beta x_t + \gamma D_t + \delta D_t x_t + \varepsilon_t$$

该模型等价于

$$y_t = \begin{cases} \alpha + \beta x_t + \varepsilon_t, & \text{若 } t < 1978, \mathrm{D}t=0 \\ (\alpha + \gamma) + (\beta + \delta) x_t + \varepsilon_t, & \text{若 } t \geq 1978, \mathrm{D}t=1 \end{cases}$$

引入哑变量及其互动项，相当于在不同时期，使用不同的截距项与斜率。

如果仅仅引入互动项，则仅改变斜率。




引入哑变量及其互动项的效果

# 政治身份的承继效应是否存在性别差异？

1. 问题：因父辈的党员身份所带来的收入差异，对性别的影响是否一样？
2. 因变量：对数收入；交互项：父亲党员身份*性别

   $\hat{Y}$=a+b（父亲党员身份）+c（性别）+d（党员*性别）

   reg lninc party sex party*sex

3. 统计结果：父辈的政治身份对儿子的影响大于对女儿的影响。
4. 结论：父辈政治身份在性别上的不平等承继。

# 模型诊断

1. 因变量是否服从正态分布
2. 残差是否服从正态分布
   - quantile-quantile plot（Q-Q图）
   - residual-vs-fitted plot（残差VS拟合值图）
3. 异常观测值及诊断
   - outlier（异常值）：standardized residual（标准化残差）and studentized residual（学生化残差）
   - leverage points（杠杆值） or influential observations: Cook's distance（Cook距离）

# 模型诊断

1. 检查因变量
2. 检查模型：残差图
3. 检查样本：Cook距离
4. 检查自变量：VIF

# 模型诊断

## Example 4.1

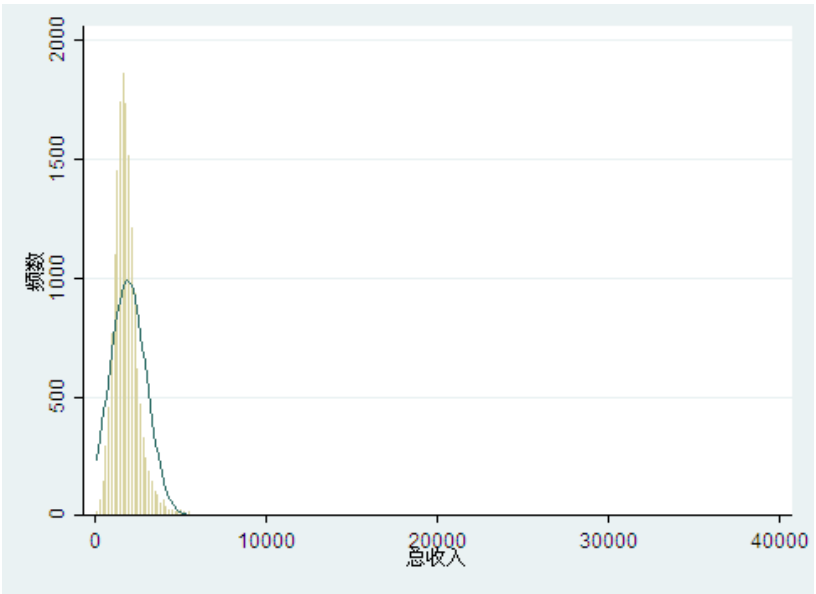
以*CHIP88*数据为例，介绍回归诊断。

$ln\hat{Y}$=a+b（教育年限）+c（工作年限）+d（党员）+e（性别）

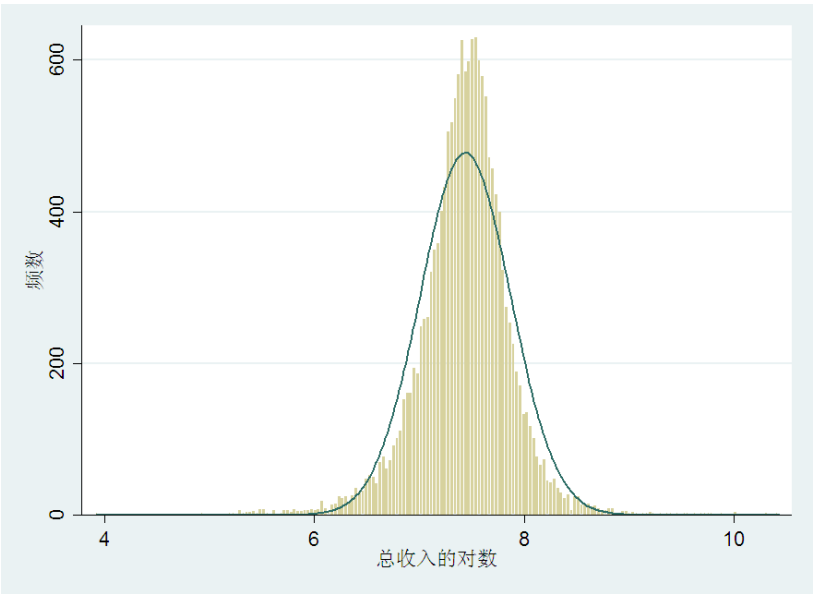
先进行回归 reg logearn edu exp cpc sex

1. 预测Y的拟合值 predict yhat
2. 预测残差 predict residual, residual
3. 画残差图 rvfplot （residual-versus-fitted plot）
4. 画Q-Q图 qnorm residual
5. 预测Cook距离 predict cookD, cooksd
6. 输出VIF estat vif

# 因变量是否服从正态分布




(a) 总收入earn的直方图

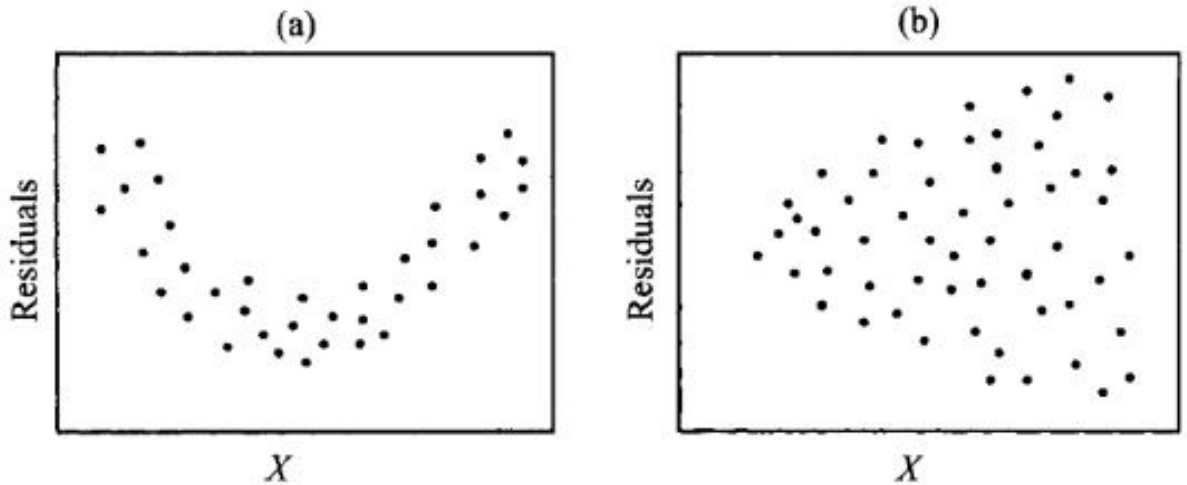



(b) 总收入对数logearn的直方图

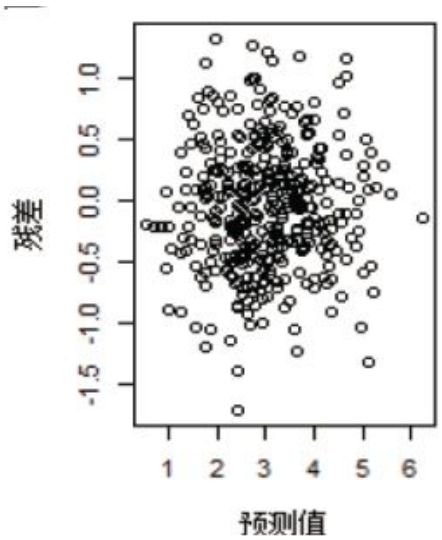
Figure 4.1: 总收入的对数转换

可使用标准化残差的相关图形来检验线性回归的模型假设。

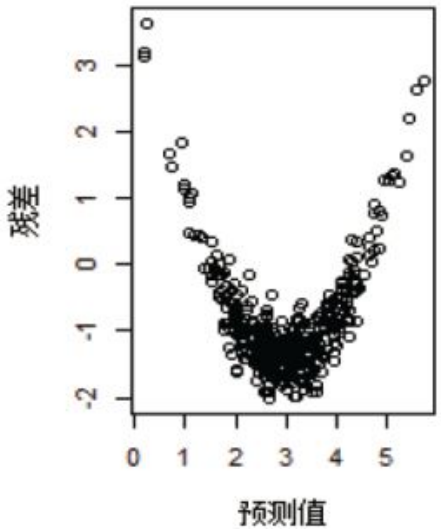
- 标准化残差图对每个自变量的散点图。在回归模型假设下，图中的点应随机散落而没有什么规律。下面图中什么假设被违反了？



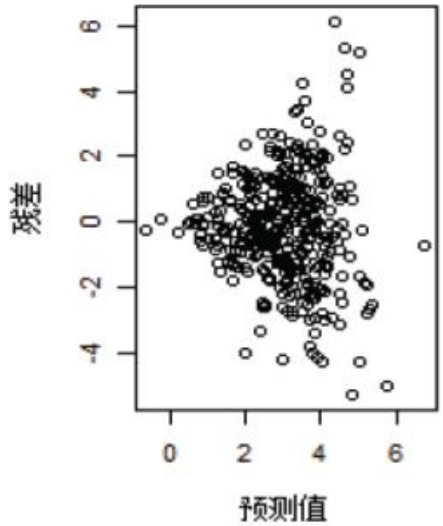

- 标准化残差图对因变量拟合值的散点图。在回归模型假设下，图中的点应随机散落而没有什么规律。




- **假设：误差是0均值、同方差的**
- 误差观测不到，残差可以
- **残差图**：横轴是预测值，Y轴是残差
- 假设成立，应观察到：**残差以0为平均水平，无规律的散乱分布**




- 残差并不以0为平均水平波动，呈现出**抛物线形状，遗漏**了重要的自变量，尤其是某些自变量的平方项
- 解决方案：加入新的自变量，或者考虑非线性模型




- 残差图呈现**喇叭状**，残差的波动随着预测值的增加变得剧烈
- 这是**异方差**的典型现象
- 解决方案：对因变量做**对数变换**（如果因变量取值为正），再建模

# 残差是否服从正态分布

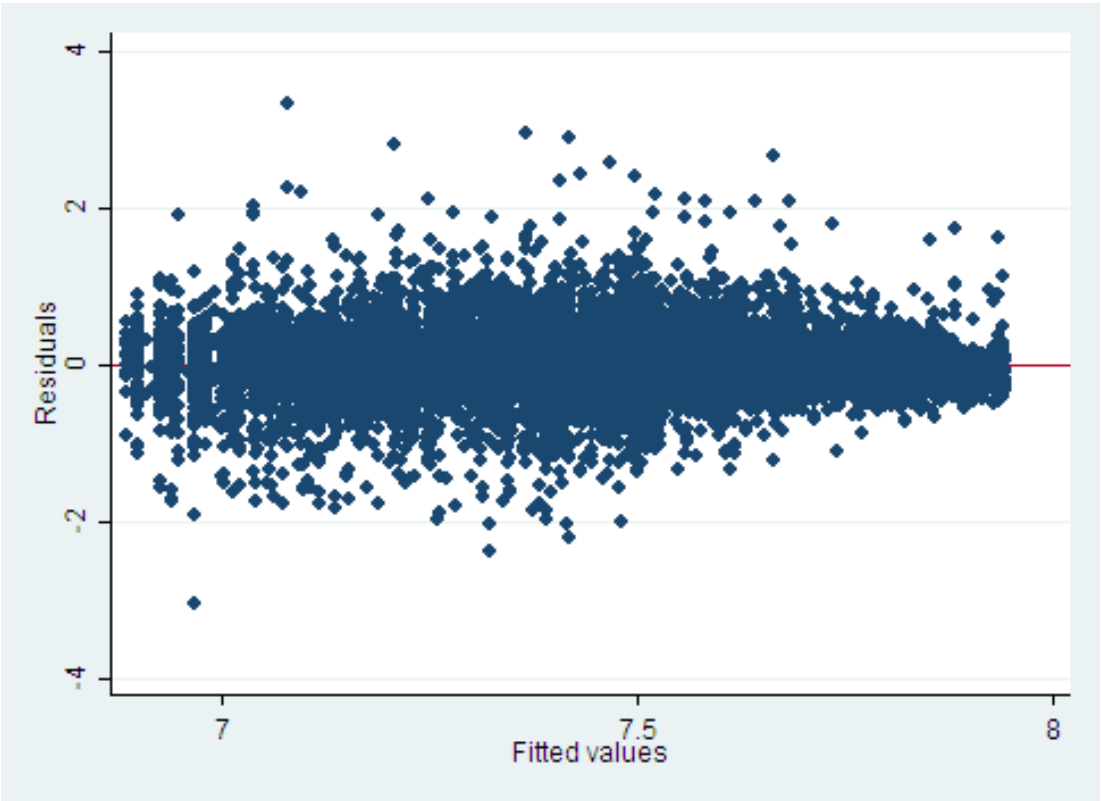
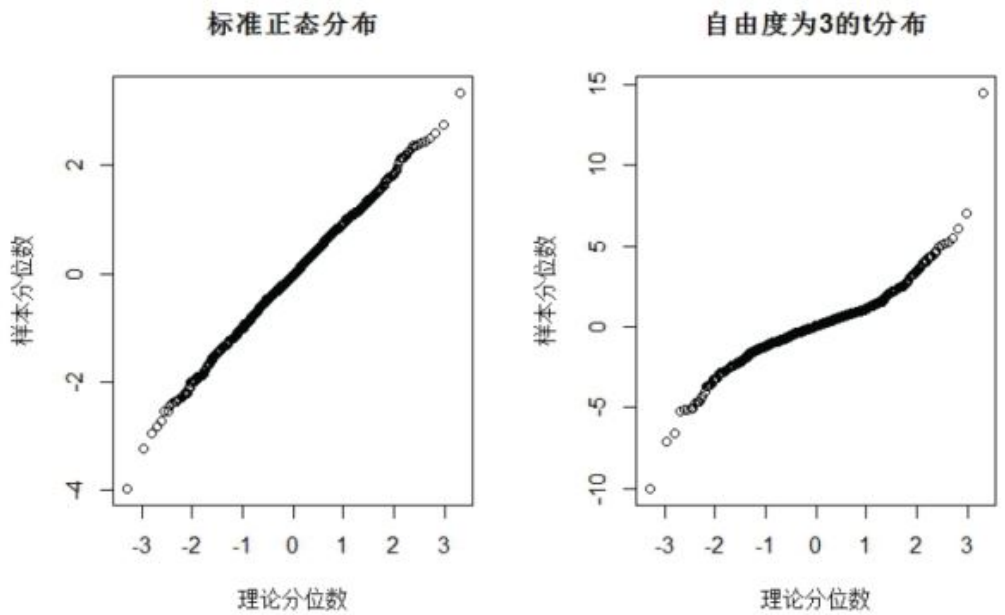



Figure 4.2: logearn的残差VS拟合值图

- **假设：误差服从正态分布**
- 误差观测不到，残差可以
- 通过检验残差的正态性来考察正态性假设是否成立，最常用的统计图是**QQ图**

- **横轴**：理论分位数
- **纵轴**：样本分位数
- 正态分布：QQ图的散点近似成**一条直线**
- **厚尾分布（t分布）**：QQ图两侧的“尾巴”偏离直线
- 解决方案：对因变量做对数变换



# 残差是否服从正态分布

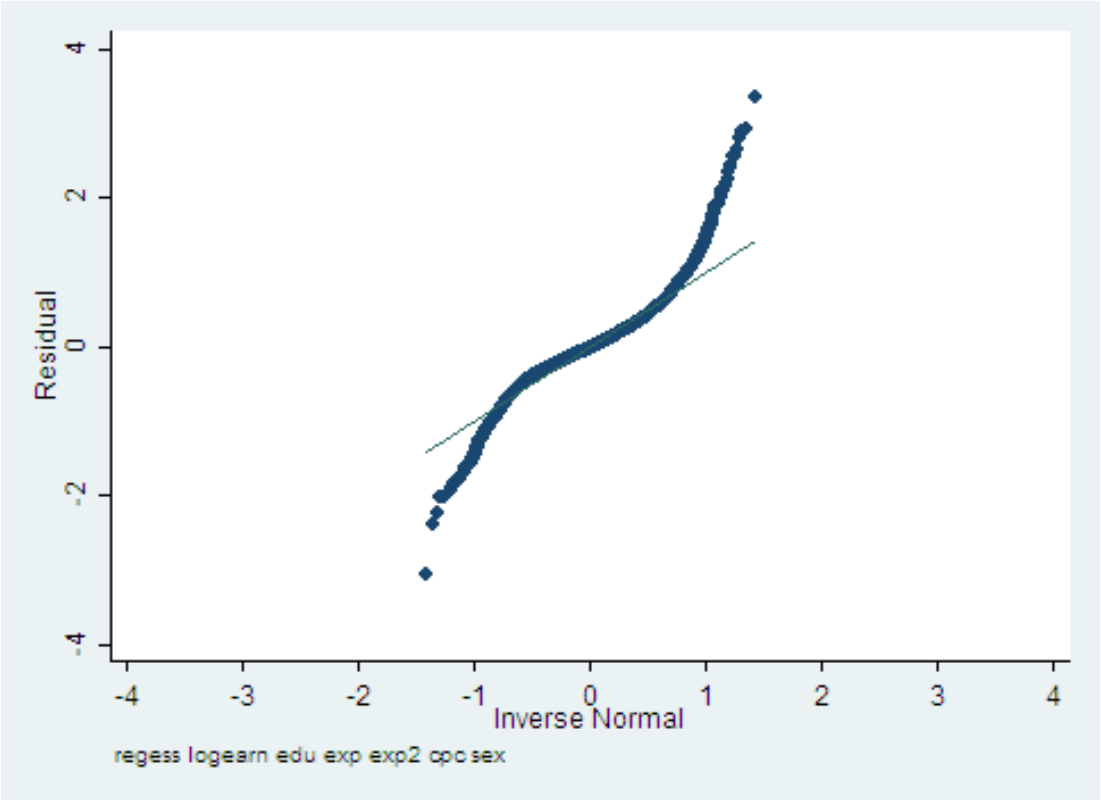



Figure 4.3: logearn回归残差QQ图

异常点可能来自两个方面。

- 因变量上的异常：表现为标准化残差$r_i$的绝对值很大。因为标准化残差应该近似服从均值为0、标准差为1的正态分布，所以其绝对值大于2或3的观测是异常点。
- 自变量上的异常：表现为杠杆值（leverage）　很大。

*教材第10章第218页

- **强影响点：**
  - 如果在计算某种指标的时候，包含和不包含某个样本点，对于结果影响很大，那么这个样本点就是强影响点。
  - 你能举个例子么？

- **Cook距离：**
  - **思想：**对于线性回归，如果包含和不包含某个样本点，对于回归系数的估计值影响很大，那么这个样本点就是强影响点。
  - 公式：别背公式，理解思想。
  - 注意：Cook距离是针对样本计算的，每个样本点都有一个Cook距离。
  - 思考：Cook距离多大才算强影响点。跑跑实际数据例子感受一下！

*教材第10章第219页

- Cook距离反映了使用完整数据和使用不包含第$i$个观测的数据所得到的拟合值的差异。
- 令$\hat{y}_{j(i)}$表示去掉第$i$个观测之后所得到的回归方程对$y_j$的拟合值。第$i$个观测的Cook距离的定义如下：

$$C_i = \frac{\sum_{j=1}^{N}(\hat{y}_j - \hat{y}_{j(i)})^2}{\hat{\sigma}^2(p+1)}$$
$$= \frac{r_i^2}{p+1} \times \frac{H_{ii}}{1-H_{ii}},\ i = 1, 2, \cdots, N。$$

- 可以看出，$r_i$的绝对值越大或者$H_{ii}$的值越大，Cook距离越大。

1 多重共线性是指多元回归中的自变量存在高度的相关。

2 variance inflation factor(VIF)：度量由某个自变量导致的多重共线性问题。

```
. estat vif
```

| Variable | VIF | 1/VIF |
|---|---|---|
| exp | 1.35 | 0.743321 |
| edu | 1.30 | 0.766980 |
| cpc | 1.28 | 0.779726 |
| sex | 1.09 | 0.920502 |
| Mean VIF | 1.25 | |

Figure 4.4: 方差膨胀因子VIF

*教材第6章第103-104页

- **多重共线性：**
  - 理解：某个自变量可以被另外一些自变量的线性组合所替代。
  - 后果：系数估计不可信，模型不能使用。
  - 怎么检查：计算自变量之间的相关系数，或者计算方差膨胀因子。
- **方差膨胀因子（VIF）：**
  - **思想：**用某个自变量作为因变量，其他自变量作为自变量，建立一个新的回归模型，并计算$R^2$。
  - 思考：如果新的回归模型的$R^2$很高，说明什么？
  - 公式：用1减去这个新的$R^2$，再取倒数。
  - 注意：方差膨胀因子是针对变量计算的，每个变量都有一个方差膨胀因子值。
  - 经验：如果某个变量的方差膨胀因子值超过　　就要引起注意。

- 方差膨胀因子的定义：

$$VIF_j = \frac{1}{1 - R_j^2}, \ j = 1, 2, \cdots, p.$$

其中，$R_j^2$是将第$j$个自变量当做因变量，对其他自变量做回归所得到的模型的$R^2$。

- 一般而言，方差膨胀因子大于10（等价于$R_j^2 > 0.9$），就认为存在多重共线性。

1 异方差（Heteroskedasticity）的解决：

稳健标准误

reg y x1 x2 x3, robust

2 多重共线性的解决：

逐步回归

stepwise, pr(.2): reg y x1 x2 x3