线性回归案例：中老年人抑郁水平研究

吴翔

2019-03-14

概述

我们通过案例来阐述如何得到可靠的回归分析结果。

本案例源自CHARLS数据集，我们非常感谢CHARLS团队。若非如此，我们无法在本次教学中给出这个合适的案例。

```
# clean the work directory
rm(list = ls())

# set seeds
set.seed(123)

# read dataset
suppressMessages(library(tidyverse))
suppressMessages(library(broom))
suppressMessages(library(stargazer))
load("charlswh.RData")
charlswh <- charlswh %>%
    rename(hukou = r4hukou) %>%
    filter(hukou < 2) %>%
    mutate(income = income / 10000)
```

可以看到，数据集包含488个样本和5个变量。

```
# display variables
str(charlswh)
```

```
## 'data.frame':    488 obs. of  5 variables:
##  $ ID    : chr  "207428209002" "242702231002" "072428332001" "014051314002" ...
##  $ cesd10: int  1 15 2 3 2 5 11 0 1 1 ...
##  $ income: num  5 3 1.2 2 3.5 ...
##  $ hukou : num  0 0 1 1 0 1 0 1 0 0 ...
##  $ educ  : num  0 1 1 1 1 1 1 1 1 1 ...
```

各变量含义如下：

- 抑郁水平 `cesd10`：采用CESD-10抑郁量表测量得到的结果
- 收入 `income`：个人年收入，以万元计
- 教育水平 `educ`：虚拟变量，`educ = 0` 表示小学及以下教育程度，`educ = 1` 表示初中及以上教育程度
- 户口 `hukou`：虚拟变量，`hukou = 0` 表示农村户口，`hukou = 1` 表示城市户口

各个变量分布情况如下：

```
# depression
summary(charlswh$cesd10)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   2.000   5.000   6.617   9.000  30.000
```

```
# income
summary(charlswh$income)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   0.010   0.700   1.690   2.116   3.000  20.000
```

```
# hukou
table(charlswh$hukou)
```

```
## 
##   0   1 
## 352 136
```

```
# education
table(charlswh$educ)
```

```
## 
##   0   1 
## 116 372
```

初步分析

首先，分别估计三个模型。

```
# estimate three models
fit1 <- lm(cesd10 ~ income, data = charlswh)
fit2 <- lm(cesd10 ~ income + educ, data = charlswh)
fit3 <- lm(cesd10 ~ income + educ + hukou, charlswh)

# summary of results
summary(fit1)
```

```
## 
## Call:
## lm(formula = cesd10 ~ income, data = charlswh)
## 
## Residuals:
##    Min     1Q Median     3Q    Max 
## -7.659 -4.174 -1.475  2.200 24.073 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)   7.6717     0.3753  20.442  < 2e-16 ***
## income       -0.4985     0.1254  -3.975 8.11e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 5.862 on 486 degrees of freedom
## Multiple R-squared:  0.03149,    Adjusted R-squared:  0.02949 
## F-statistic:  15.8 on 1 and 486 DF,  p-value: 8.109e-05
```

```
summary(fit2)
```

```
##
## Call:
## lm(formula = cesd10 ~ income + educ, data = charlswh)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.087 -3.958 -1.282  2.290 24.445
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.0967     0.5612  16.211  < 2e-16 ***
## income       -0.3983     0.1275  -3.123 0.001896 **
## educ         -2.1472     0.6340  -3.387 0.000765 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.8 on 485 degrees of freedom
## Multiple R-squared:  0.05386,    Adjusted R-squared:  0.04996
## F-statistic: 13.81 on 2 and 485 DF,  p-value: 1.475e-06
```

```
summary(fit3)
```

```
##
## Call:
## lm(formula = cesd10 ~ income + educ + hukou, data = charlswh)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.090 -3.931 -1.099  2.379 24.019
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.0986     0.5601  16.245  < 2e-16 ***
## income       -0.3413     0.1318  -2.590  0.00988 **
## educ         -1.9233     0.6467  -2.974  0.00309 **
## hukou        -1.0529     0.6266  -1.680  0.09356 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.789 on 484 degrees of freedom
## Multiple R-squared:  0.05935,    Adjusted R-squared:  0.05352
## F-statistic: 10.18 on 3 and 484 DF,  p-value: 1.64e-06
```

相应地，可以将三个模型放置在同一个表格中，便于对比。

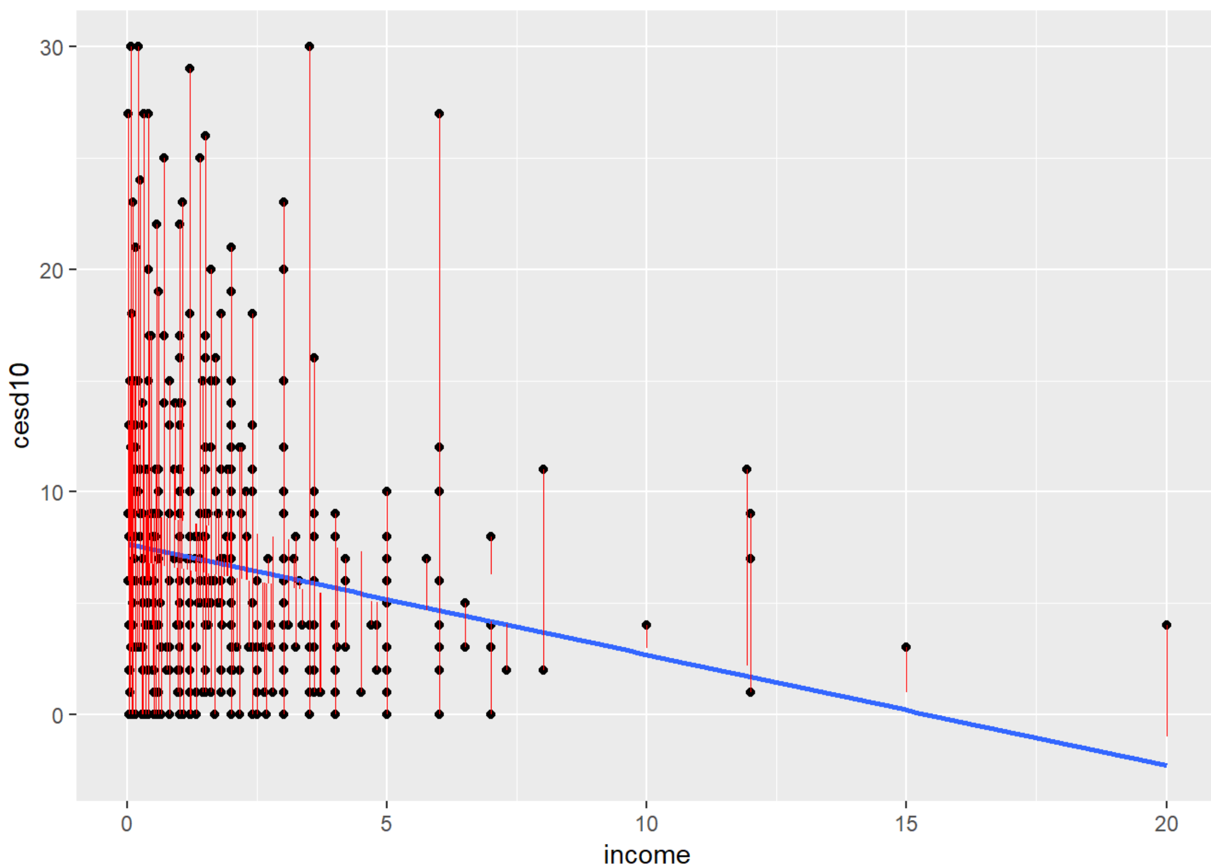
```
# output as a table
stargazer(fit1, fit2, fit3, type = "html")
```

	Dependent variable:		
	cesd10		
	(1)	(2)	(3)
income	-0.498***	-0.398***	-0.341***
	(0.125)	(0.128)	(0.132)

educ		-2.147***	-1.923***
		(0.634)	(0.647)
hukou			-1.053*
			(0.627)
Constant	7.672***	9.097***	9.099***
	(0.375)	(0.561)	(0.560)
Observations	488	488	488
R^2	0.031	0.054	0.059
Adjusted R^2	0.029	0.050	0.054
Residual Std. Error	5.862 (df = 486)	5.800 (df = 485)	5.789 (df = 484)
F Statistic	15.800*** (df = 1; 486)	13.805*** (df = 2; 485)	10.179*** (df = 3; 484)
Note:			*p<0.1;* ***p<0.05;*** p<0.01

我们选择了模型2，并展示回归结果图。

```
# calculate regression diagnostics
model.diag.metrics <- augment(fit2)
# plot the fitted values
ggplot(model.diag.metrics, aes(income, cesd10)) +
  geom_point() +
  stat_smooth(method = lm, se = FALSE) +
  geom_segment(aes(xend = income, yend = .fitted), color = "red", size = 0.3)
```

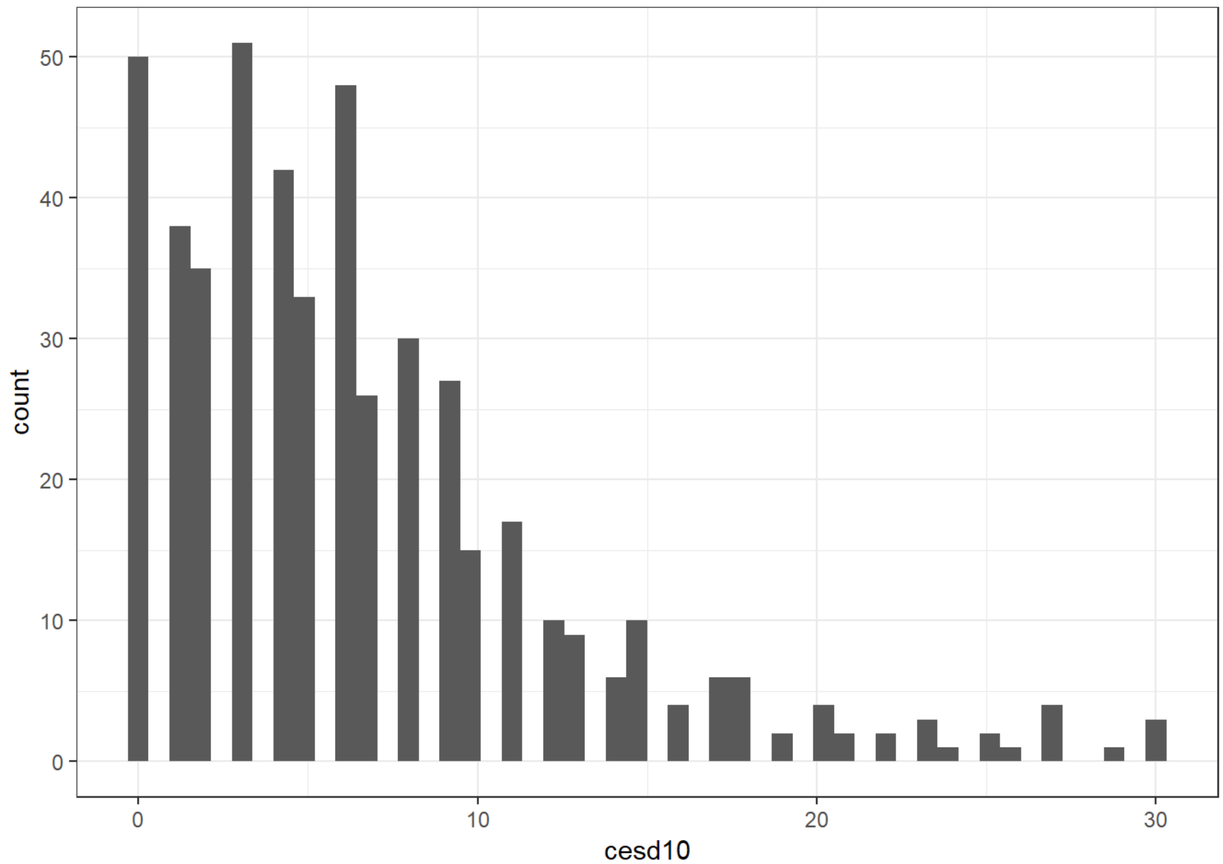


回归诊断

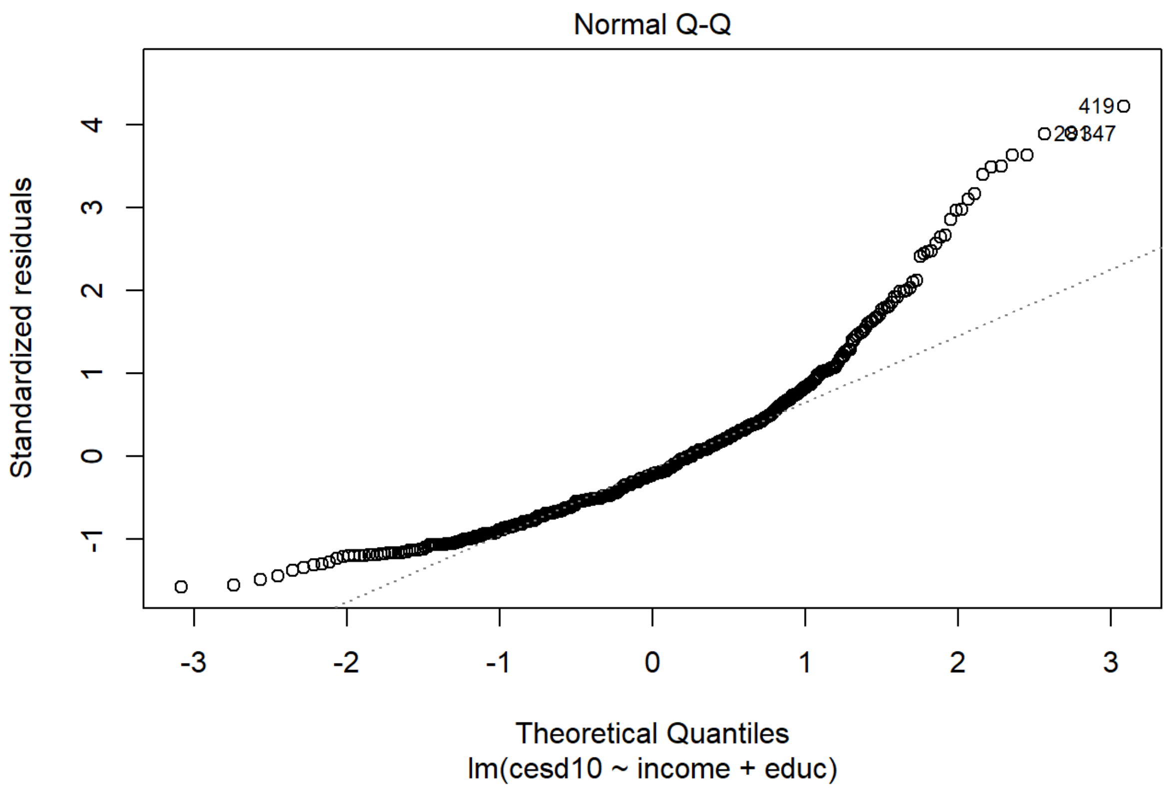
残差项的正态分布

因变量很明显不服从正态分布，而QQ图也显示，残差项也明显不服从正态分布。

```
# plot
ggplot(charlswh, aes(x = cesd10)) + geom_histogram(bins = 50) + theme_bw()
```

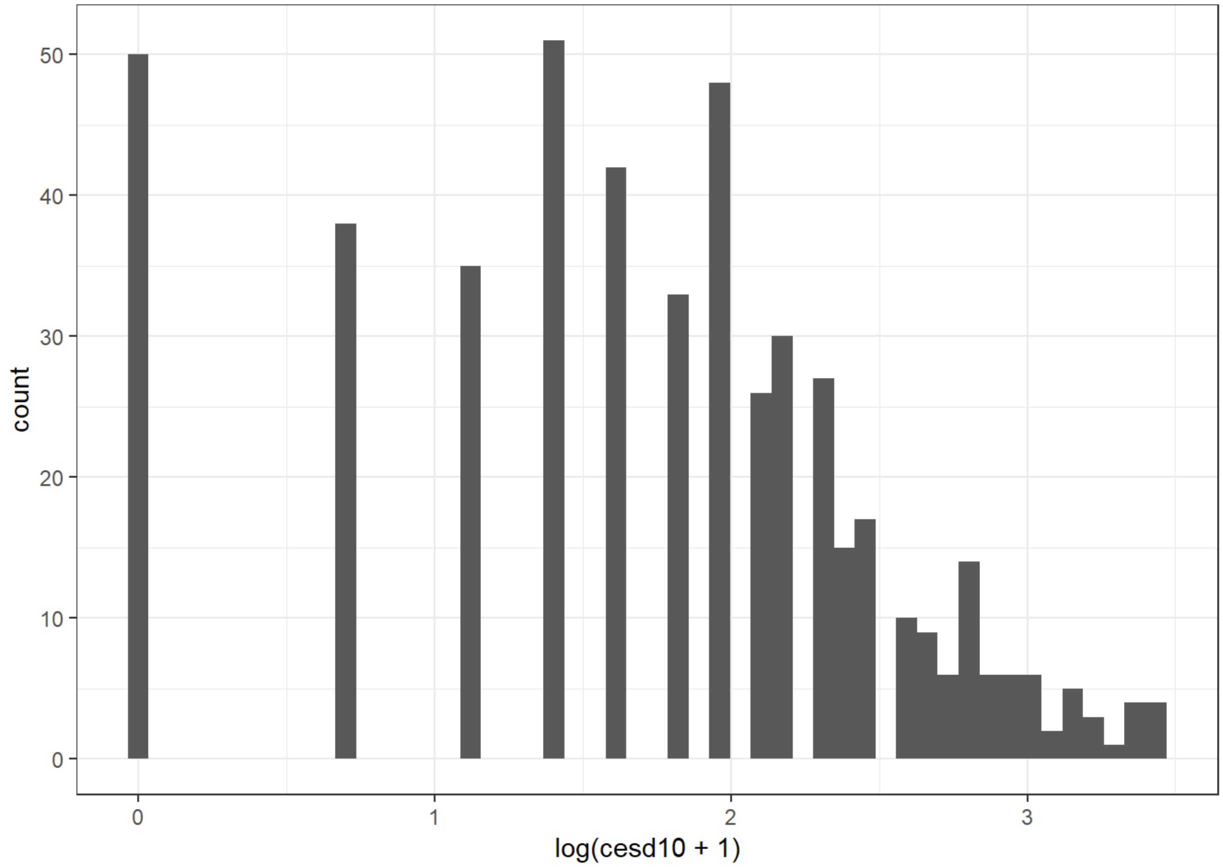



```
# residual
plot(fit2, 2)
```



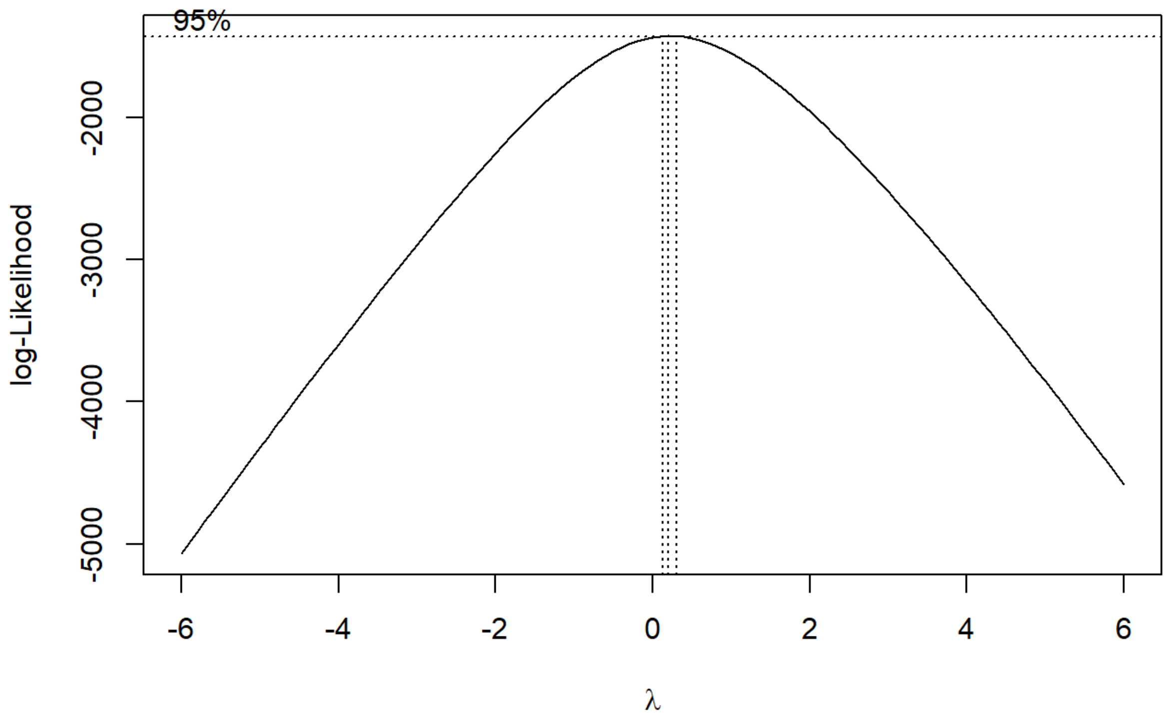
此时，可以采用简单的对数变换。考虑到有零值，我们采用$log(\alpha + \text{cesd10})$的方式完成变换。可以看到，对数变换使结果变量更加接近正态分布。

```
# plot
ggplot(charlswh, aes(x = log(cesd10 + 1))) + geom_histogram(bins = 50) + theme_bw()
```

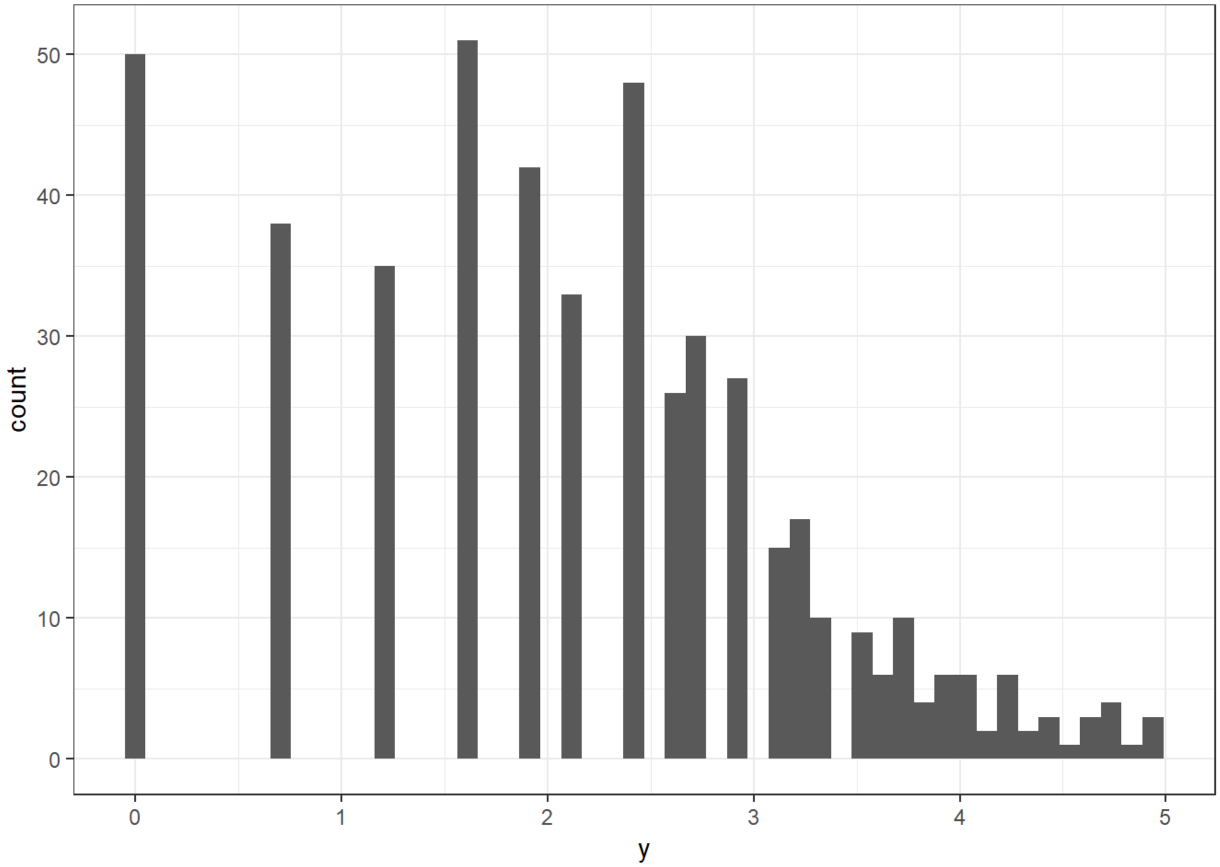


此外，我们也可以采用Box-Cox变换，得到新的结果变量y。

```
suppressMessages(library(MASS))
# box-cox transformation
a <- boxcox(I(cesd10 + 1) ~ 1, data = charlswh, lambda = seq(-6, 6, 1/10)) %>%
    as.data.frame()
```




```
# get the lambda value with the largest likelihood
lambda <- a$x[which.max(a$y)]
# get new response variable
charlswh$y <- ((charlswh$cesd10 + 1) ^ lambda - 1)/lambda
# plot
ggplot(charlswh, aes(x = y)) + geom_histogram(bins = 50) + theme_bw()
```



计算三者的偏度，可以看到，Box-Cox变换效果最好。

```
suppressMessages(library(e1071))
# original variable
skewness(charlswh$cesd10)
```

```
## [1] 1.495881
```

```
# log transformation
skewness(log(charlswh$cesd10 + 1))
```

```
## [1] -0.4400588
```

```
# box-cox transformation
skewness(charlswh$y)
```

```
## [1] -0.0618485
```

那么，我们采用Box-Cox变换，并重新检视几个回归模型。

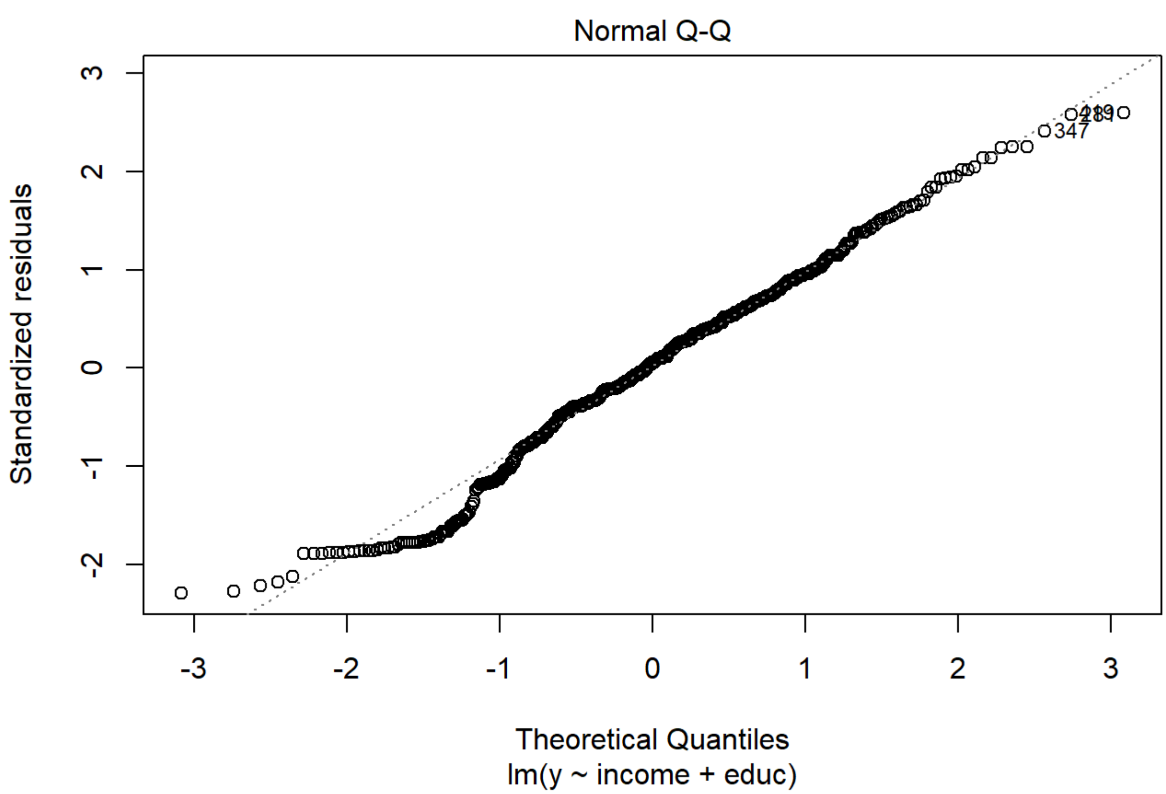
```
# estimate three models
fit4 <- lm(y ~ income, data = charlswh)
fit5 <- lm(y ~ income + educ, data = charlswh)
fit6 <- lm(y ~ income + educ + hukou, charlswh)

# output as a table
stargazer(fit4, fit5, fit6, type = "html")
```

	Dependent variable:		
	y		
	(1)	(2)	(3)
income	-0.088***	-0.067***	-0.055**
	(0.025)	(0.025)	(0.026)
educ		-0.453***	-0.405***
		(0.126)	(0.128)
hukou			-0.227*
			(0.125)
Constant	2.331***	2.631***	2.632***
	(0.075)	(0.112)	(0.111)
Observations	488	488	488
R^2	0.025	0.050	0.057
Adjusted R^2	0.023	0.046	0.051
Residual Std. Error	1.167 (df = 486)	1.153 (df = 485)	1.150 (df = 484)
F Statistic	12.497*** (df = 1; 486)	12.858*** (df = 2; 485)	9.721*** (df = 3; 484)
Note:			*p<0.1;* **p<0.05;** p<0.01

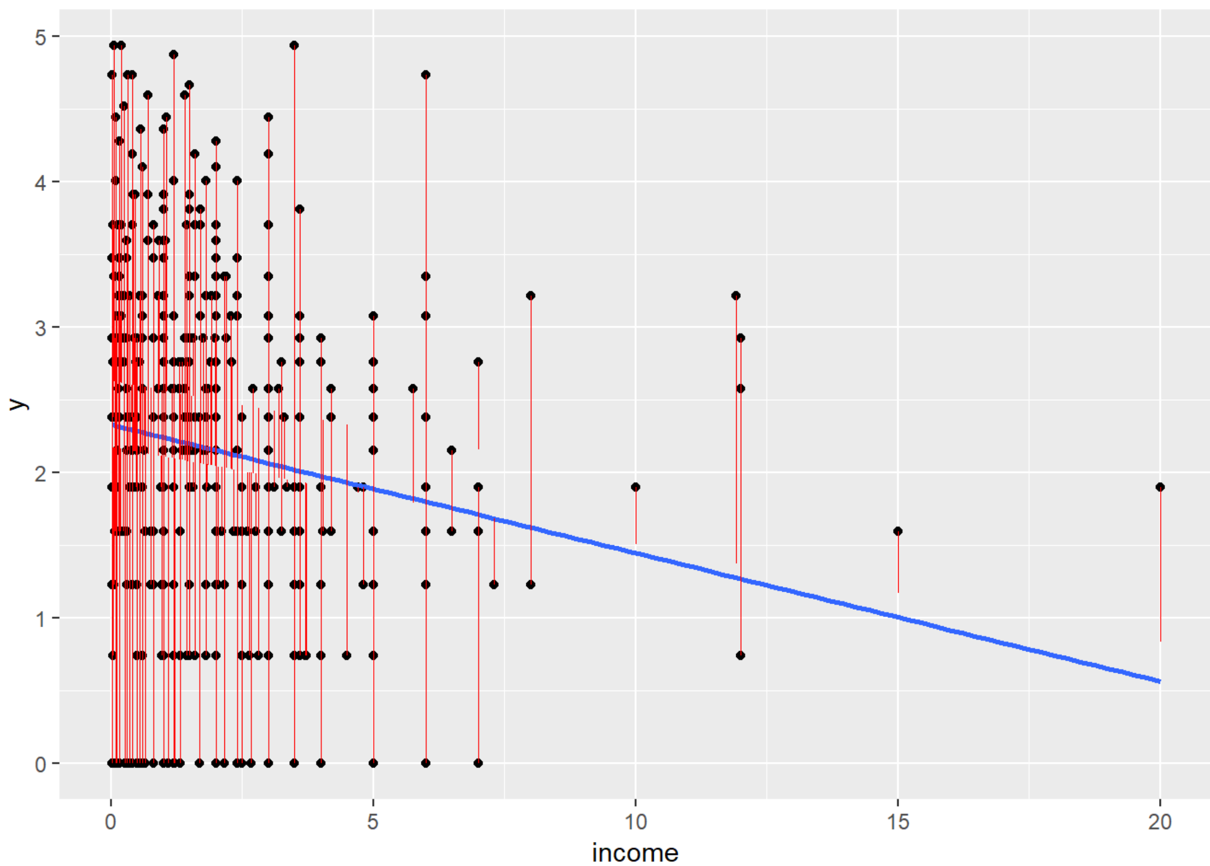
类似地，我们选择模型5，并再一次检视残差项的分布情况。可以看到，误差项已经接近正态分布了。

```
# residual
plot(fit5, 2)
```

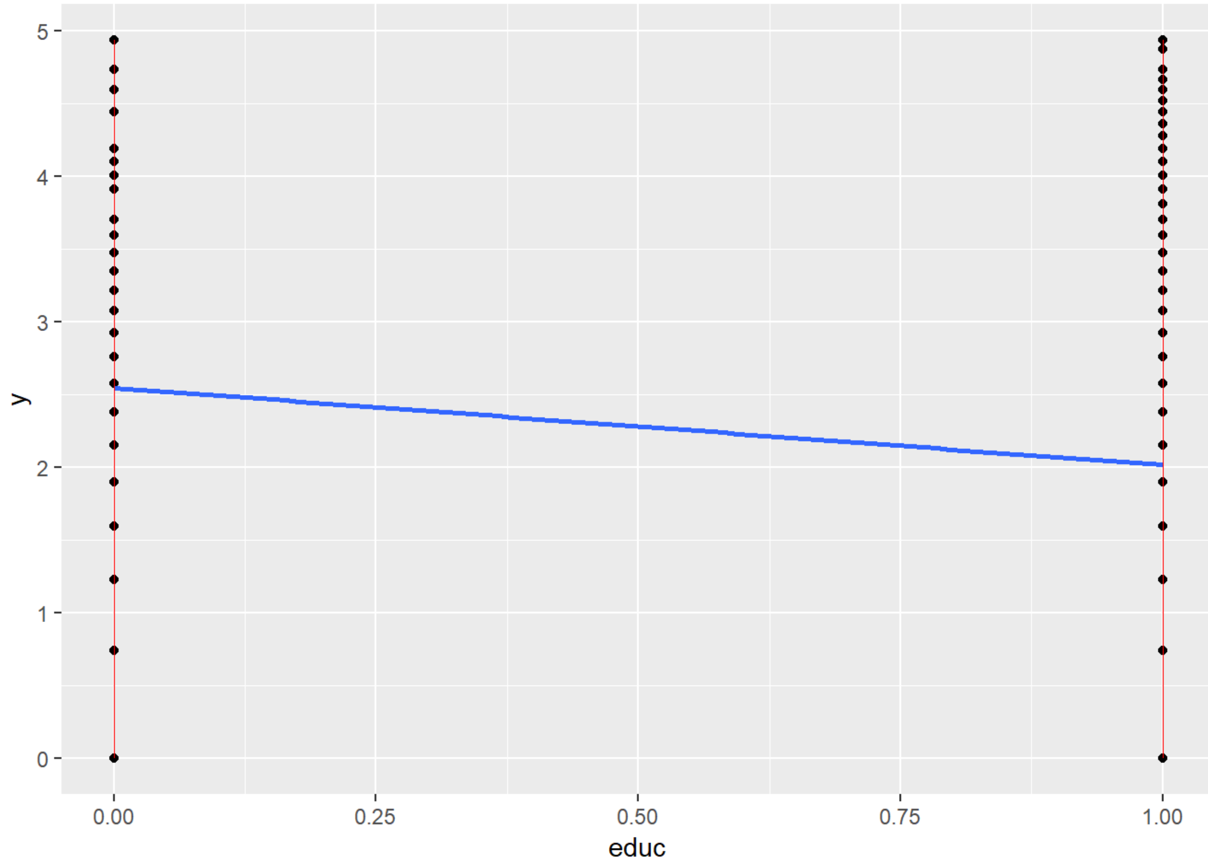


类似地，展示散点图和回归线，分别查看 `income` 和 `educ` 和 `y` 的关系（似乎有异方差问题？）。

```
# calculate regression diagnostics
model.diag.metrics <- augment(fit5)
# plot the fitted values ~ income
ggplot(model.diag.metrics, aes(income, y)) +
  geom_point() +
  stat_smooth(method = lm, se = FALSE) +
  geom_segment(aes(xend = income, yend = .fitted), color = "red", size = 0.3)
```



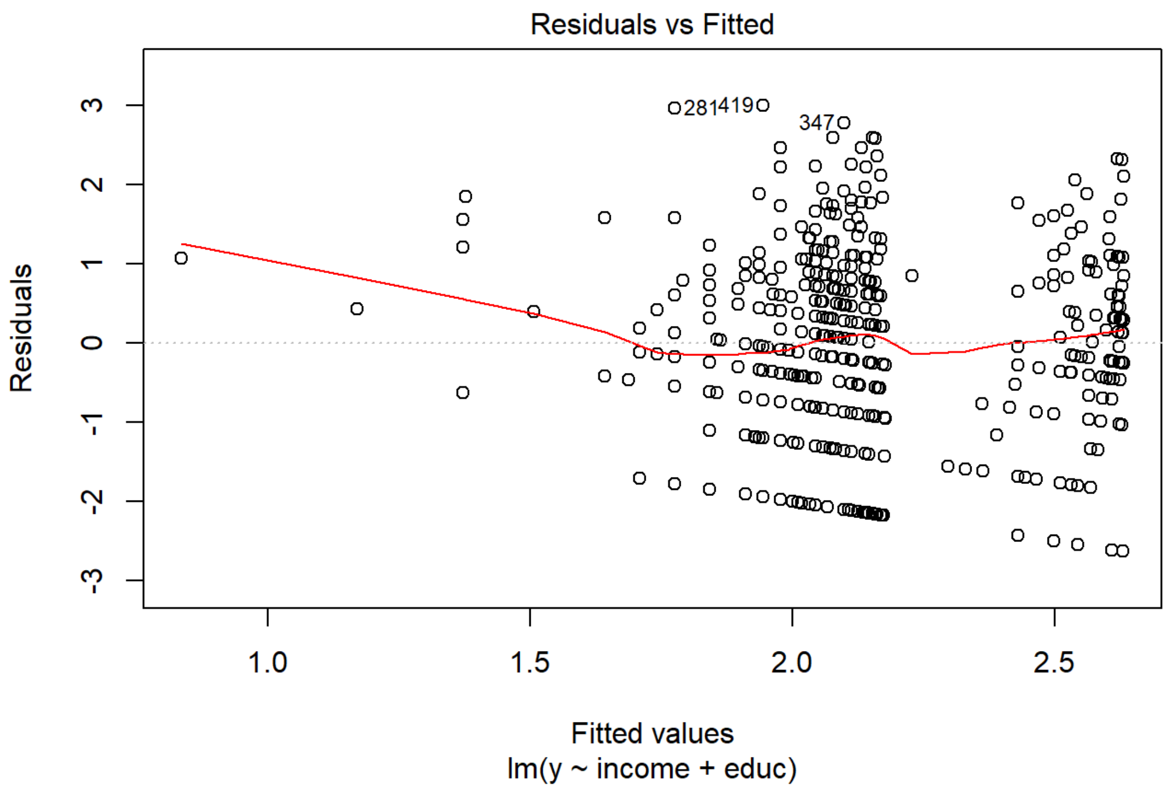

```
# plot the fitted values ~ educ
ggplot(model.diag.metrics, aes(educ, y)) +
  geom_point() +
  stat_smooth(method = lm, se = FALSE) +
  geom_segment(aes(xend = educ, yend = .fitted), color = "red", size = 0.3)
```



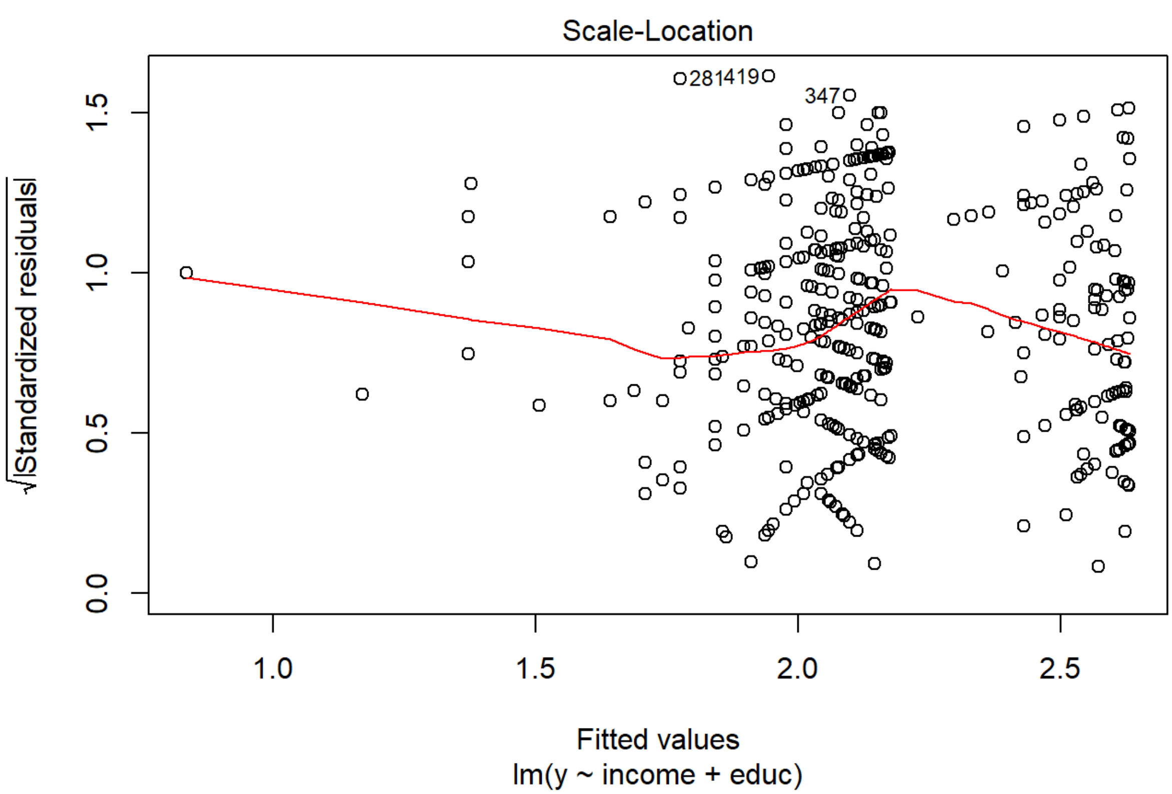
模型的非线性及异方差

从拟合值和（标准化）残差项来看，可能存在异方差和非线性问题，但是需要更多的检测以便进一步确定问题所在。

```
# residual - fitted value
plot(fit5, 1)
```

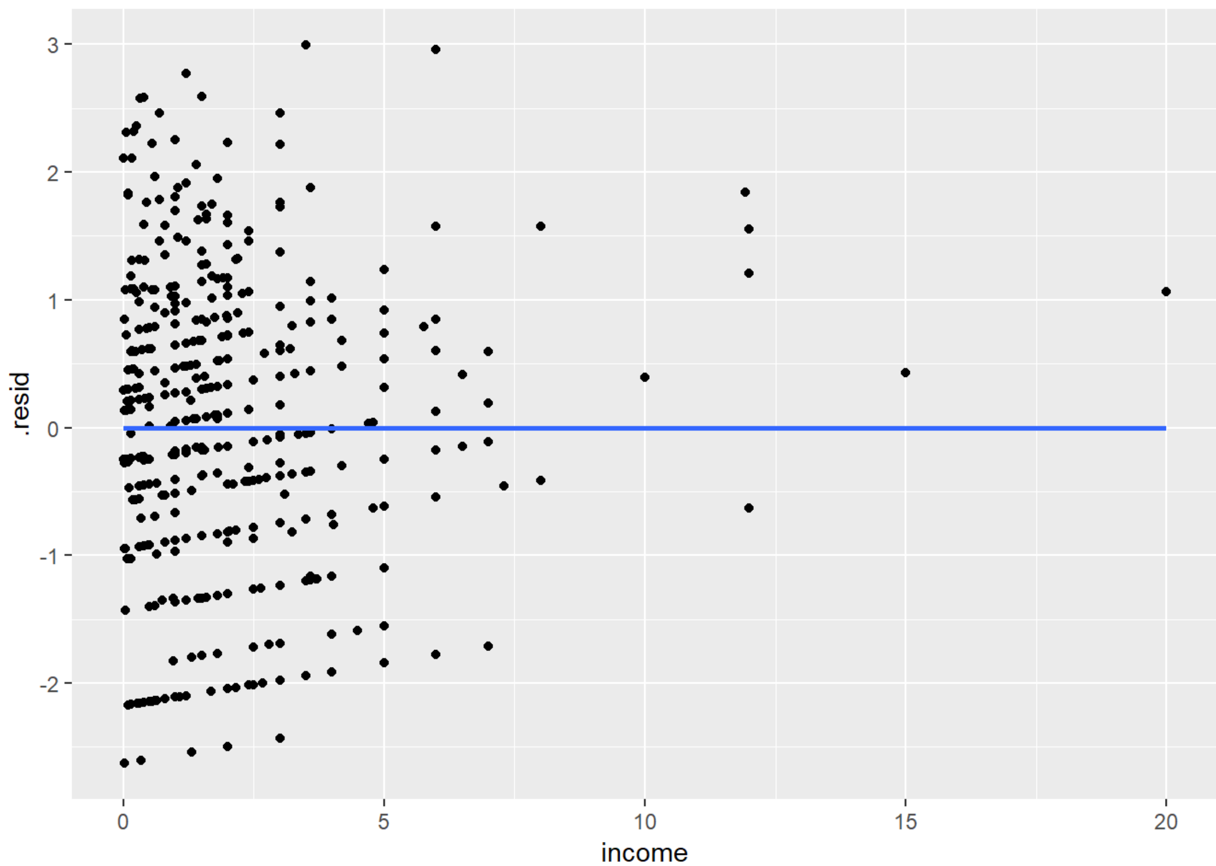



```r
# standardized residual - fitted value
plot(fit5, 3)
```

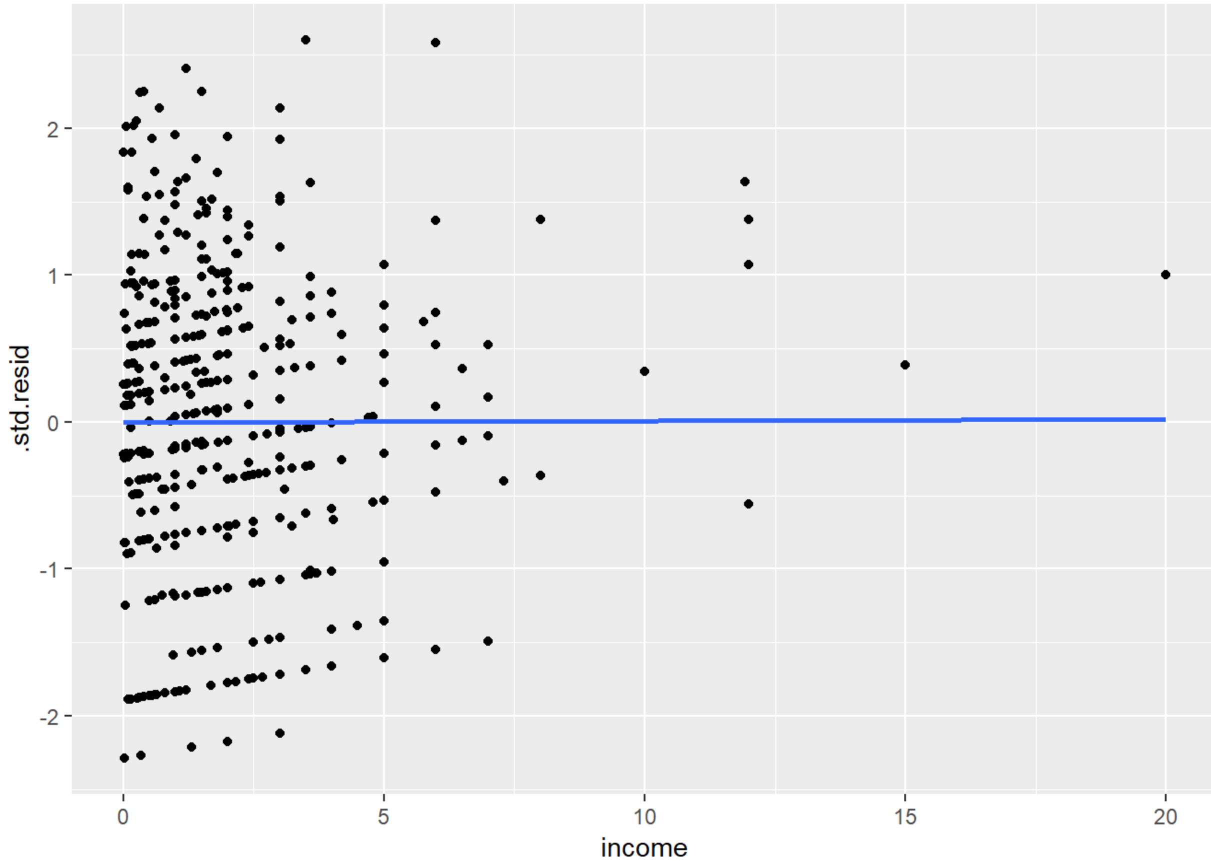


更进一步，我们检测两个解释变量和（标准化）残差项的关系。

```
# plot income ~ residual
ggplot(model.diag.metrics, aes(income, .resid)) +
  geom_point() + stat_smooth(method = lm, se = FALSE)
```

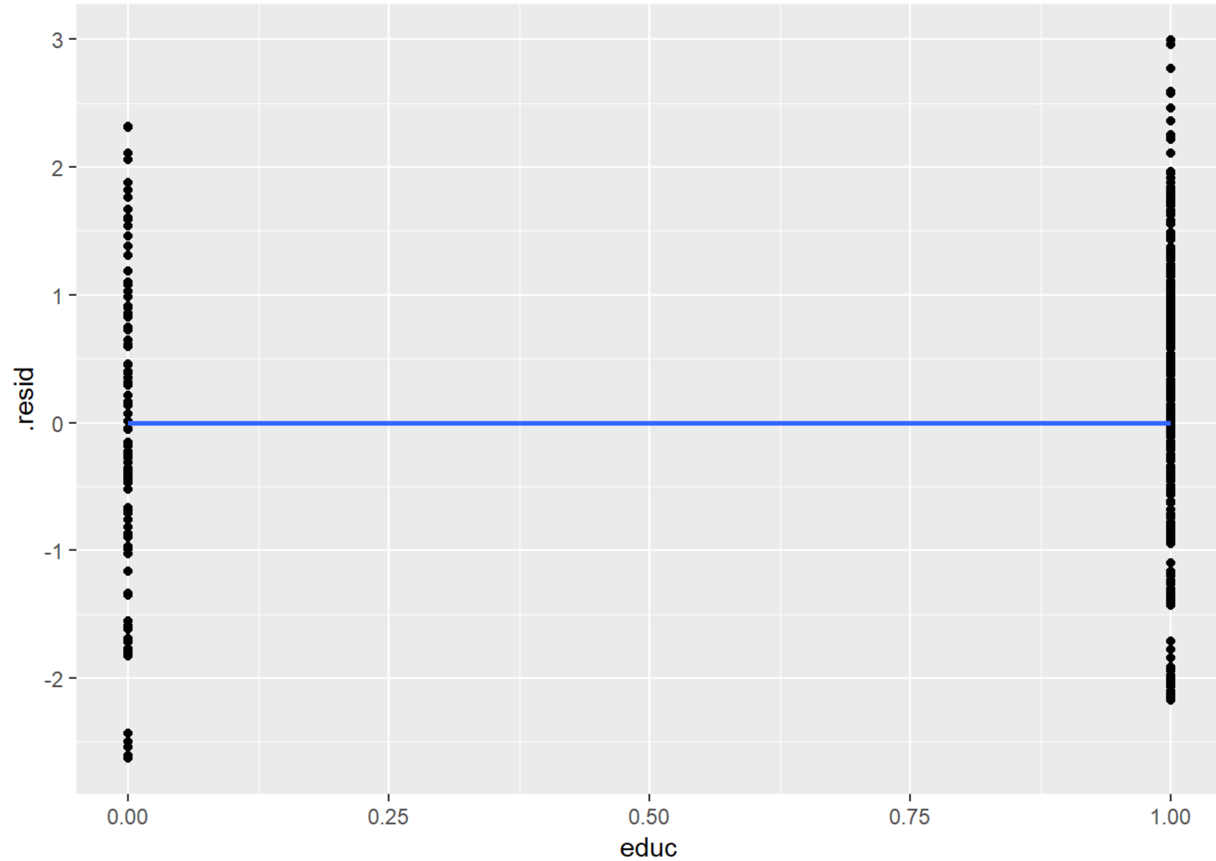



```
# plot income ~ standardized residual
ggplot(model.diag.metrics, aes(income, .std.resid)) +
  geom_point() + stat_smooth(method = lm, se = FALSE)
```

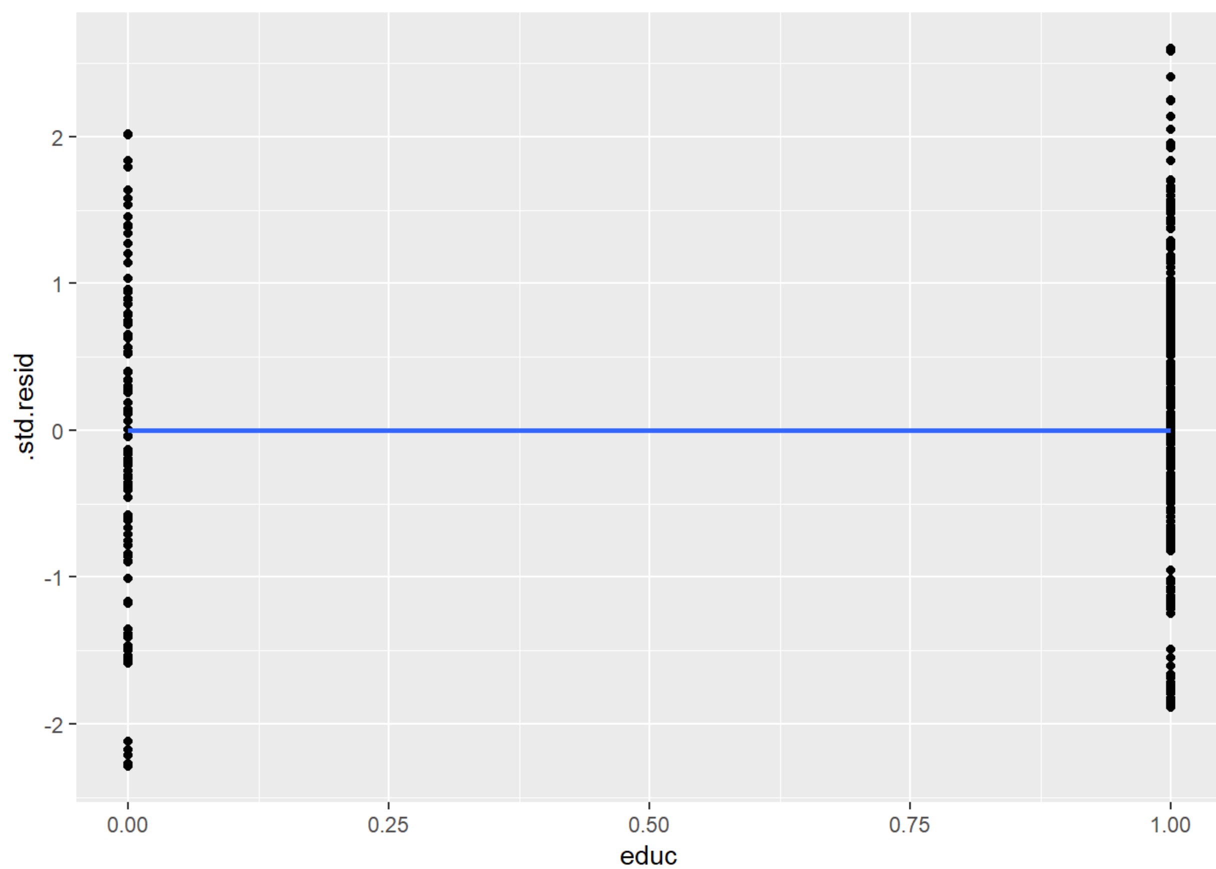


可以看到，似乎存在如下非线性和异方差问题： `income` 越大，抑郁水平的方差就越小。

```
# plot educ ~ residual
ggplot(model.diag.metrics, aes(educ, .resid)) +
  geom_point() + stat_smooth(method = lm, se = FALSE)
```




```
# plot educ ~ standardized residual
ggplot(model.diag.metrics, aes(educ, .std.resid)) +
  geom_point() + stat_smooth(method = lm, se = FALSE)
```



教育程度则未发现显著的异方差问题。

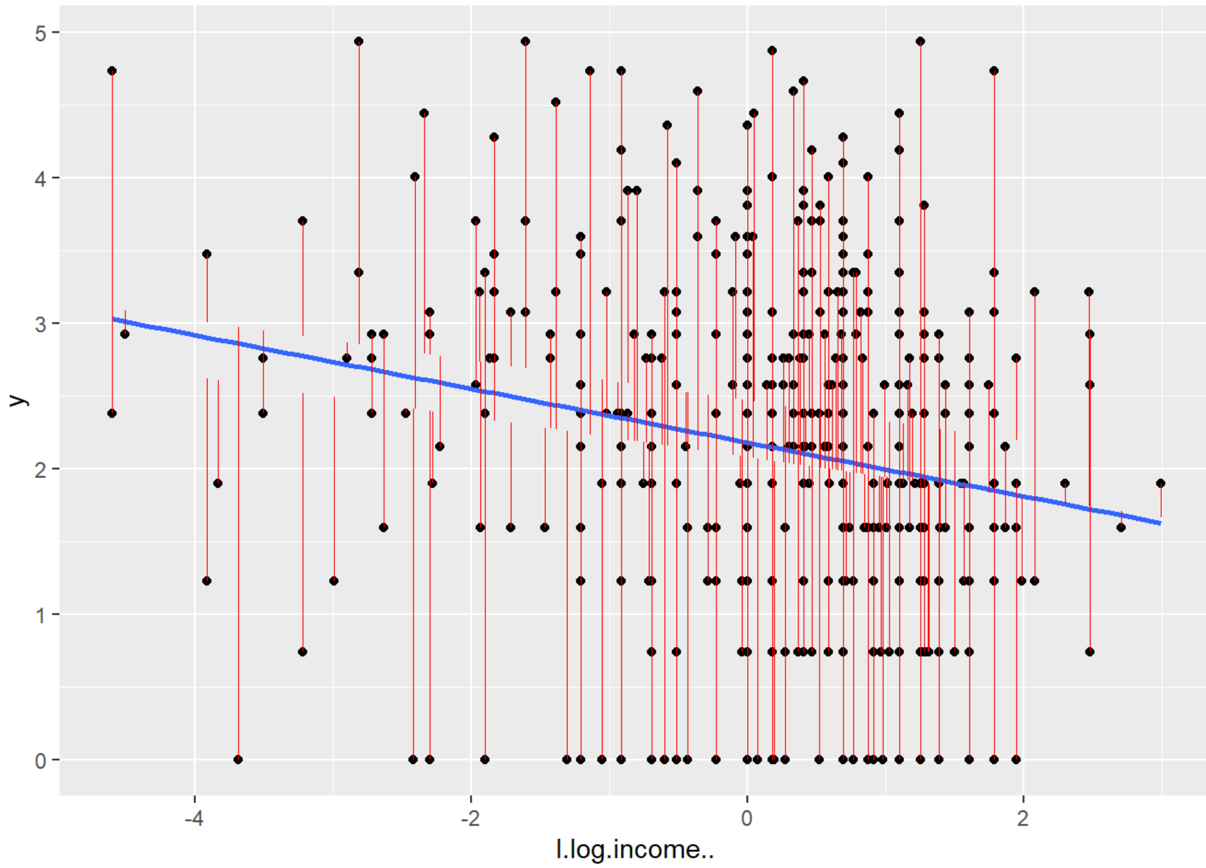
为了解决以上问题，我们再次进行变换，即对收入取对数，从而希望消除非线性和异方差问题。

```
# fit the new model
fit7 <- lm(y ~ I(log(income)) + educ, data = charlswh)
summary(fit7)
```

```
##
## Call:
## lm(formula = y ~ I(log(income)) + educ, data = charlswh)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.97578 -0.70341  0.03177  0.77321  3.02784
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     2.46860    0.10984  22.474  < 2e-16 ***
## I(log(income)) -0.13749    0.04464  -3.080  0.00219 **
## educ           -0.38750    0.13078  -2.963  0.00320 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.15 on 485 degrees of freedom
## Multiple R-squared:  0.0551, Adjusted R-squared:  0.0512
## F-statistic: 14.14 on 2 and 485 DF,  p-value: 1.074e-06
```

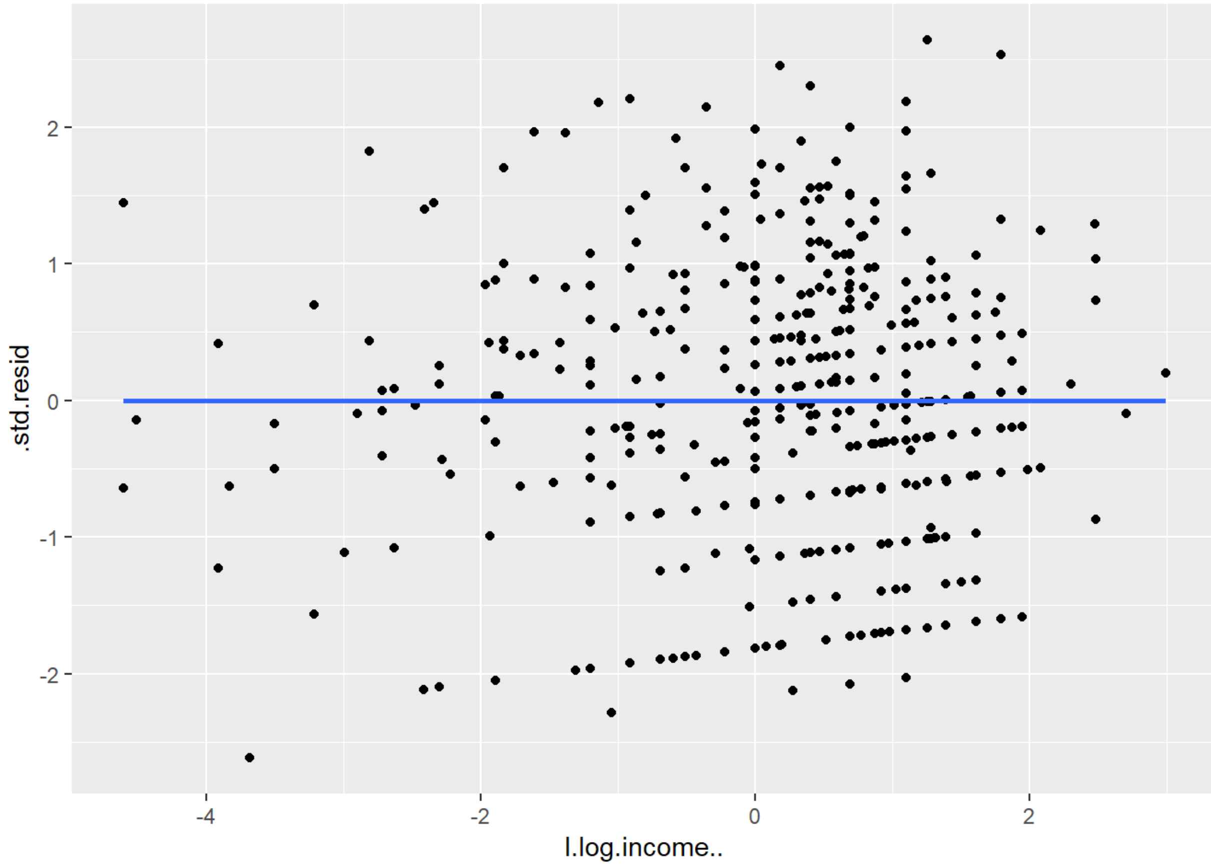
再来看散点图和回归线。

```
# calculate regression diagnostics
model.diag.metrics <- augment(fit7)
# plot the fitted values ~ income
ggplot(model.diag.metrics, aes(I.log.income.., y)) +
  geom_point() + stat_smooth(method = lm, se = FALSE) +
  geom_segment(aes(xend = I.log.income.., yend = .fitted), color = "red", size = 0.3)
```



以及收入的对数与标准化残差的关系。

```
# plot log educ ~ standardized residual
ggplot(model.diag.metrics, aes(I.log.income.., .std.resid)) +
  geom_point() + stat_smooth(method = lm, se = FALSE)
```

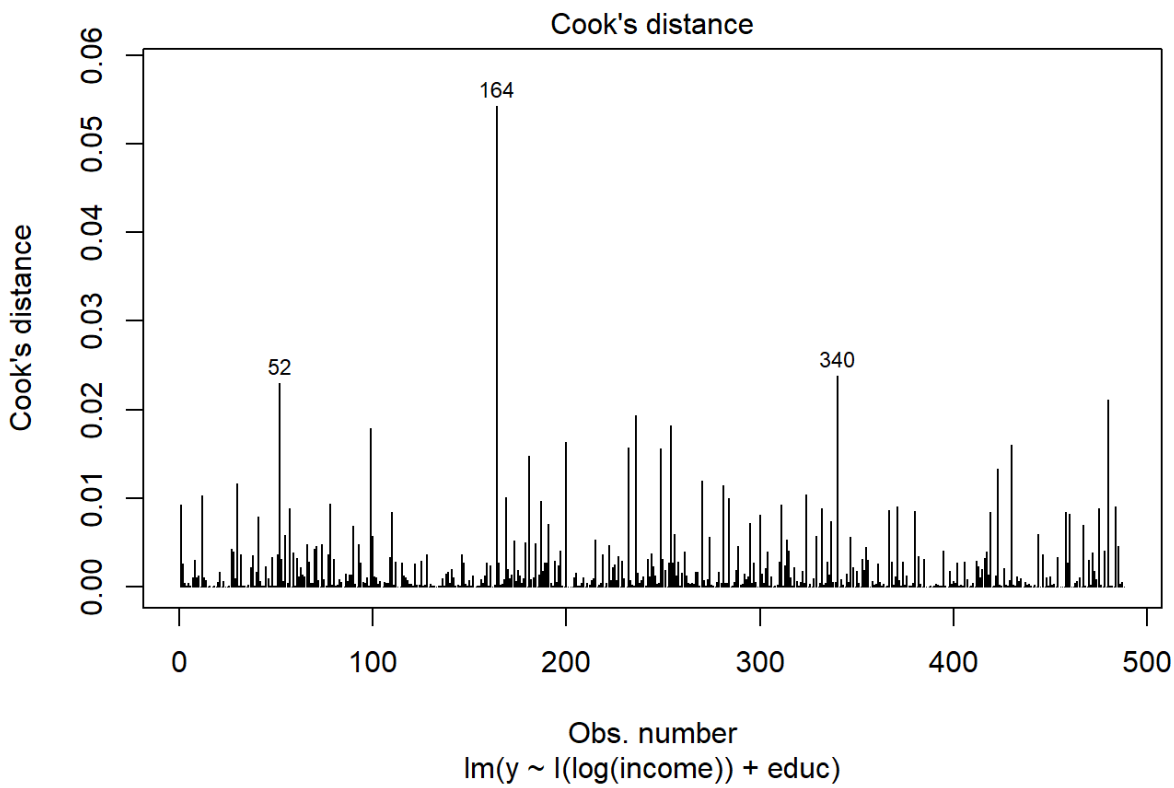


可以看到，采用 `log(income)` 之后，可以认为并不存在明显的非线性和异方差问题。

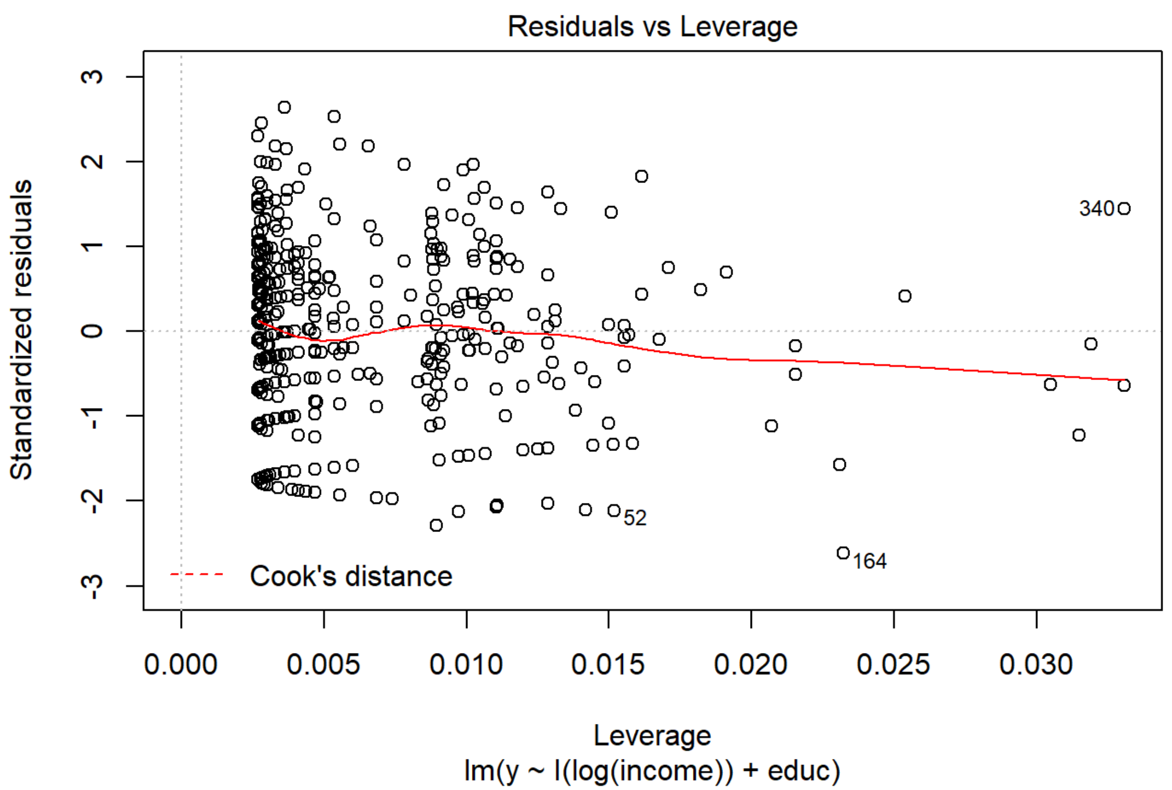
高影响点与异常值

从图上可以看出，有三个高影响点。标准化残差绝对值未超过3，因此可以认为没有极端偏离回归线的异常值。

```
# Cook's distance
plot(fit7, 4)
```




```
# Leverage
plot(fit7, 5)
```



此外，可以看到，对回归系数影响最大的是第164号样本。

```
# remove the 164th sample and refit the model
fit8 <- lm(y ~ I(log(income)) + educ, data = charlswh[-164, ])
# output as a table
stargazer(fit7, fit8, type = "html")
```

	Dependent variable:	
	y	
	(1)	(2)
I(log(income))	-0.137***	-0.152***
	(0.045)	(0.045)
educ	-0.388***	-0.399***
	(0.131)	(0.130)
Constant	2.469***	2.487***
	(0.110)	(0.109)
Observations	488	487
R^2	0.055	0.062
Adjusted R^2	0.051	0.058
Residual Std. Error	1.150 (df = 485)	1.143 (df = 484)
F Statistic	14.141*** (df = 2; 485)	16.021*** (df = 2; 484)
Note:	*p<0.1; **p<0.05;** p<0.01	

可以看到，删除这个样本之后，回归系数发生了较大变化，并且R^2有了显著增加。

那么这个样本是否真的“异常”呢？

```
# the 164th sample
charlswh[164, ]
```

```
##               ID cesd10 income hukou educ y
## 164 270402213001      0  0.025     0    0 0
```

可以看到，第164个样本是：年收入0.025万元、农村户口、小学及以下教育程度，但是抑郁程度为0！

换言之，这位低收入、低教育程度、农村户口的中老年受访者，拥有整个样本中最佳的精神健康状态，丝毫无抑郁的表现。

那么，这个样本是否真的“异常”呢？应该来讲，我们并无充分的理由认为这是异常样本。更可能的情况是，这是真实的存在。因此，我们不宜排除这个观测样本。

最终模型

经过反复地回归诊断，我们选择了如下模型：

$$\text{BoxCox}(cesd10_i) = \alpha + \beta_1 \log(income_i) + \beta_2 educ_i + \epsilon_i.$$

相应的回归结果为：

```
# display the results
summary(fit7)
```

```
##
## Call:
## lm(formula = y ~ I(log(income)) + educ, data = charlswh)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.97578 -0.70341  0.03177  0.77321  3.02784
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     2.46860    0.10984  22.474  < 2e-16 ***
## I(log(income)) -0.13749    0.04464  -3.080  0.00219 **
## educ           -0.38750    0.13078  -2.963  0.00320 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.15 on 485 degrees of freedom
## Multiple R-squared:  0.0551, Adjusted R-squared:  0.0512
## F-statistic: 14.14 on 2 and 485 DF,  p-value: 1.074e-06
```

```
# save the results
save(fit7, charlswh, file = "../ch3/charlswh.RData")
```