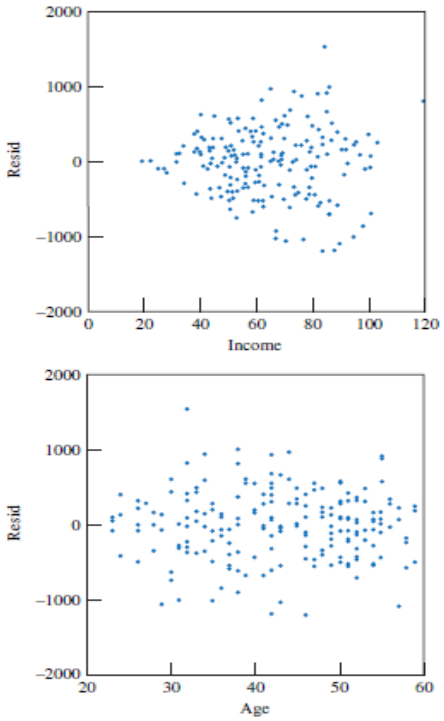
Jeffrey Chang
jlc190004

Question 8.4

A sample of 200 Chicago households was taken to investigate how far American households tend to travel when they take vacation. Measuring distance in miles per year, the following model was estimated

$$MILES = \beta_1 + \beta_2 INCOME + \beta_3 AGE + \beta_4 KIDS + e$$

The variables are self-explanatory except perhaps for AGE, the average age of the adult members of the household. The data are in the file *vacation.dat*.



(a) The equation was estimated by least squares and the residuals are plotted against age and income in Figure 8.4. What do these graphs suggest to you?

The plots on the right show the relation between residuals against INCOME and residuals against AGE.

- You can observe the absolute value of the residuals increases as income increases.
- There is no obvious relationship between the residuals and age. it looks like the discrete distribution.
- The plots suggest that the error variance depends on income, but not age.

(c) Table 8.2 contains three sets of estimates: those from least squares, those from least squares with White's standard errors, and those from generalized least squares under the assumption $\sigma_i^2 = \sigma^2 \times INCOME^2$.

Table 8.2 **Output for Exercise 8.4**

Variable	Coefficient	Std. Error	*t*-value	*p*-value
Least squares estimates				
C	−391.55	169.78	−2.31	0.022
INCOME	14.20	1.80	7.89	0.000
AGE	15.74	3.76	4.19	0.000
KIDS	−81.83	27.13	−3.02	0.003
Least squares estimates with White standard errors				
C	−391.55	142.65	−2.74	0.007
INCOME	14.20	1.94	7.32	0.000
AGE	15.74	3.97	3.97	0.000
KIDS	−81.83	29.15	−2.81	0.006
Generalized least squares estimates				
C	−425.00	121.44	−3.50	0.001
INCOME	13.95	1.48	9.42	0.000
AGE	16.72	3.02	5.53	0.000
KIDS	−76.81	21.85	−3.52	0.001

Jeffrey Chang
jlc190004

(i) How do vacation miles traveled depend on income, age, and the number of kids in the household?

MILES = β_1 + β_2(+) *INCOME* + β_3(+) *AGE* + β_4(-) *KIDS* + *e*

According to Table 8.2 with three sets of estimates, miles has a positive and significant (p-value is <0.05) relationship with INCOME and AGE. but negatively related to the number of KIDS in the household.

(ii) How do White's standard errors compare with the least squares standard errors? Do they change your assessment of the precision of estimation?

The White std. errors are higher but not much than std. errors of Least-squares estimates, (1.94 > 1.8, 3.97 > 3.76, 29.15 > 27.13). Therefore, White's std. errors' estimations are less precise, but this doesn't have a huge impact on the conclusion of the precision of estimation.

(iii) Is there evidence to suggest the generalized least squares estimates are better estimates?

The generalized least squares std. errors are less than the White std. errors, suggesting that generalized least squares is a better estimation. Weighted Least Squares estimator is more efficient than the least squares estimator if we know the form of the variance

Question 8.12

In the file *pubexp.dat* there are data on public expenditure on education (EE), gross domestic product (GDP), and population (P) for 34 countries in the year 1980. It is hypothesized that per capita expenditure on education is linearly related to per capita GDP. That is,

$$y_i = \beta_1 + \beta_2 x_i + e_i$$

where

$$y_i = \left(\frac{EE_i}{P_i}\right) \quad \text{and} \quad x_i = \left(\frac{GDP_i}{P_i}\right)$$

It is suspected that e_i may be heteroskedastic with a variance related to x_i.

(a) Why might the suspicion about heteroskedasticity be reasonable?

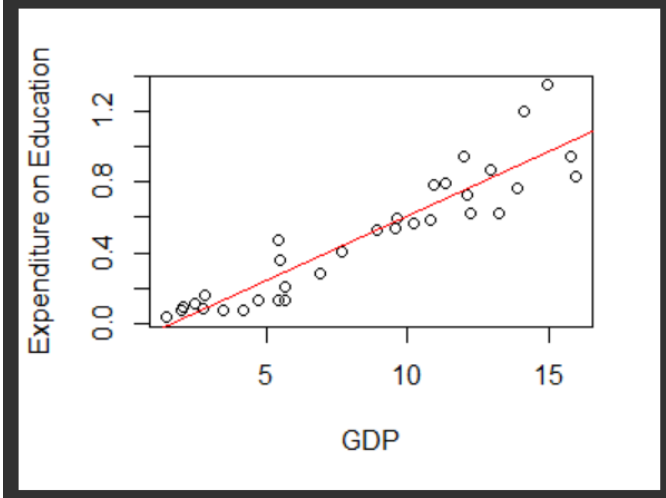
We expect that the richer countries with a higher GDP per capita have more money to use. It is more flexible for richer countries people to decide how much they want on education spending. However, for countries with smaller GDP, their people might only put a limit budget on education without many choices.

Jeffrey Chang
jlc190004

(b) Estimate the equation using least squares; plot the least squares function and the residuals. Is there any evidence of heteroskedasticity?

Plot the fitted regression line and data points as following:

```r
```{r, fig.width = 4, fig.height = 3}
mod1 <- lm(ee_p~gdp_p, data=pubexp)
plot(pubexp$gdp_p,pubexp$ee_p, type="p",
xlab="GDP", ylab="Expenditure on Education")
abline(mod1, col="red")
```
```



The regression results:

term	estimate	std.error	statistic	p.value
(Intercept)	-0.1246	0.0485	-2.5673	0.0151
gdp_p	0.0732	0.0052	14.1275	0.0000

Evidence of heteroskedasticity: The plotted values are more spread about the red regression line for bigger values of GDP per capita. This point out that heteroskedasticity is existence, and that the variance of the error terms is increasing with GDP per capita.

Jeffrey Chang
jlc190004

(c) Test for the existence of heteroskedasticity using a White test.

The White test equation:

$\widehat{e_i} = \alpha_1 + \alpha_2(GDP_i/P_i) + \alpha_3(GDP_i/P_i)^2 + v_i$

So we run the $\widehat{e_i}$ regression, the result as following:

```
```{r}
ressq <- resid(mod1)^2
modres <- lm(ressq~gdp_p+I(gdp_p^2), data=pubexp)
summary(modres)
```
```

```
Call:
lm(formula = ressq ~ gdp_p + I(gdp_p^2), data = pubexp)

Residuals:
      Min        1Q    Median        3Q       Max
-0.048773 -0.011263 -0.005239  0.005350  0.097664

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.0176770  0.0161120   1.097   0.2810
gdp_p       -0.0052062  0.0045479  -1.145   0.2611
I(gdp_p^2)   0.0004840  0.0002638   1.835   0.0762 .
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.02513 on 31 degrees of freedom
Multiple R-squared:  0.293,     Adjusted R-squared:  0.2474
F-statistic: 6.423 on 2 and 31 DF,  p-value: 0.004636
```

This regression returns an R^2 = 0.293.

White test we are testing the hypothesis H_0: $\alpha_1 = \alpha_2 = 0$, H_1: $\alpha_1 \neq 0$ and/or $\alpha_2 \neq 0$.

```
```{r}
N <- 34
gmodres <- glance(modres)
chisq1<-((gmodres$r.squared)*N)
print( paste0("The White test statistic is = N*R^2 =  ",
              round(chisq1, 4)))

print( paste0("The critical chi-squared value for the White test at a = 0.05 is = ",
              round(qchisq(.95, df=2), 4)))

```
```

```
[1] "The White test statistic is = N*R^2 =  9.9614"
[1] "The critical chi-squared value for the White test at a = 0.05 is = 5.9915"
```

Since 9.961 > 5.991, we reject the null hypothesis and conclude that heteroskedasticity exists.

Jeffrey Chang
jlc190004

(d) Use White's formula for least squares variance estimates to find some alternative standard errors for the least squares estimates obtained in part (b). Use these standard errors and those obtained in part (b) to construct two alternative 95% confidence intervals for β_2. What can you say about the confidence interval that ignores the heteroskedasticity?

```r
# Regular standard errors in the 'MILES' equation
kable(tidy(mod1),digits = 6)
```

term	estimate	std.error	statistic	p.value
(Intercept)	-0.124573	0.048523	-2.567308	0.015126
gdp_p	0.073173	0.005179	14.127545	0.000000

```r
# needs package 'car'
# Robust (HC1) standard errors in the 'MILES' equation
cov1 <- hccm(mod1, type="hc1")
mod1.HC1 <- coeftest(mod1, vcov.=cov1)
kable(tidy(mod1.HC1), digits = 6)
```

term	estimate	std.error	statistic	p.value
(Intercept)	-0.124573	0.040414	-3.08242	0.004204
gdp_p	0.073173	0.006212	11.78005	0.000000

```r
t <- qt(c(0.975), df=32)
round(t, 4)
```

```
[1] 2.0369
```

$t_{(0.975,32)}$ = 2.0369

The 95% confidence interval for β_2 using the conventional least squares standard errors:
$b_2 \pm t_{(0.975,32)}$*se (b_2) = 0.73173 ± 2.0369*0.005179 => [0.0626, 0.0837]

The 95% confidence interval for β_2 using White's standard errors:
$b_2 \pm t_{(0.975,32)}$*se (b_2) = 0.73173 ± 2.0369*0.006212 => [0.0605, 0.0858]

According to the dataset, the confidence interval from White's standard errors is wider. If we ignore the heteroskedasticity, it looks like the precision of least squares estimation will be exaggerated.

Jeffrey Chang
jlc190004

(e) Re-estimate the equation under the assumption that var (e_i)= $\sigma^2 x_i$. Report the results. Construct a 95% confidence interval for β_2. Comment on its width relative to that of the confidence intervals found in part (d).

Re-estimating the equation under the assumption that var (e_i) = $\sigma^2 x_i$

```{r}
# part E
w <- 1/pubexp$gdp_p
pubexp.wls <- lm(ee_p~gdp_p, weights=w, data=pubexp)
kable(tidy(pubexp.wls),
caption="WLS estimates for the 'EE' equation" , digits = 6)
```

term	estimate	std.error	statistic	p.value
(Intercept)	-0.092921	0.028904	-3.214853	0.002978
gdp_p	0.069321	0.004412	15.713067	0.000000

The 95% confidence interval for β_2 is

$b_2 \pm t_{(0.975,32)}$*se (b_2) = 0.069321 ± 2.0369*0.004412 => [0.0603, 0.0783]

Under the assumption of var (e_i)= $\sigma^2 x_i$, confidence interval [0.0603, 0.0783] is less than both confidence intervals calculated in part (d) => [0.0626, 0.0837] and [0.0605, 0.0858].

Weighted Least Squares estimator is more efficient than the least squares estimator if we know the form of the variance and when heteroskedasticity is happening. However, if our assumption regarding the form of the variance is incorrect, then not only will the estimator be inefficient - the SE will also be incorrect. If there is no heteroskedasticity, a direct comparison of the Weighted Least Squares interval with that obtained using the conventional least squares standard errors is not meaningful.