Pol Sci 630: Problem Set 12 Solutions: Heteroskedasticity, Autocorrelation

Prepared by: Anh Le (anh.le@duke.edu)

Due Date: Friday, Nov 20, 2015, 12 AM (Beginning of Lab)

```
rm(list = ls())
library(ggplot2)
```

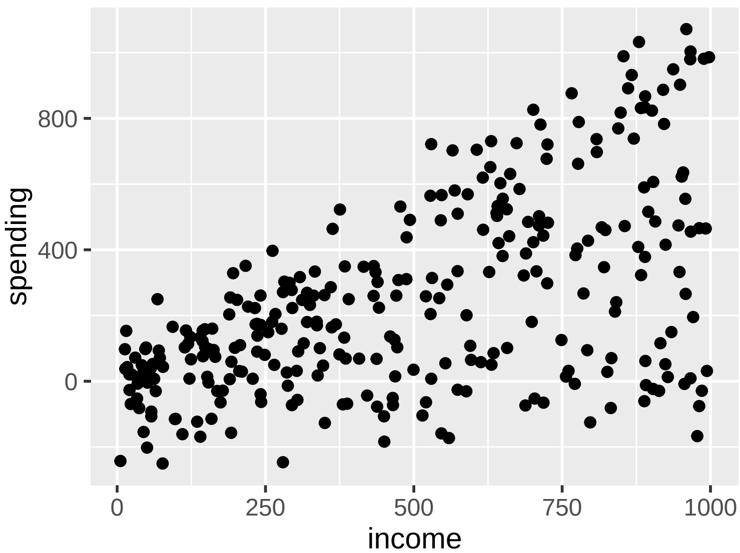
1 Heteroskedasticity

This exercise nudges you to think about heteroskedasticity as a theoretical / social science problem, not a mechanical / statistical issue to be blindly fixed.

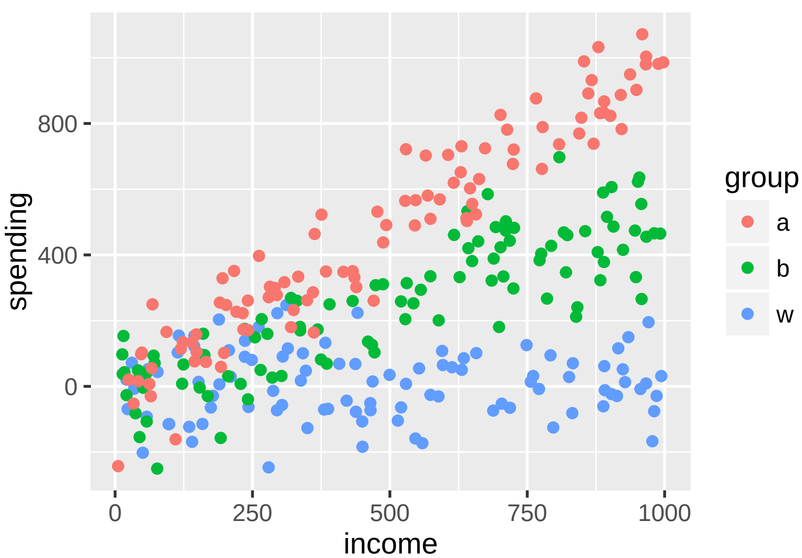
One common cause of heteroskedasticity is that our model does not take into account heterogenous effect across sub-populations. For example, we have a model of spending (dependent var) as a function of income (independent var), and the propensity to spend differs across ethnic groups. Formally,

$$spending = \beta_{ethnic} income + \epsilon \tag{1}$$

where β_{ethnic} takes a different value for white, black, and asian. If we don't know about this heterogeneity of propensity to spend across ethnic groups, the graph will show heteroskedasticity:



Buf if we are smart researcher, we'll realize the underlying cause of the heterogeneity, as shown in the following plot:



The take-home point is that heteroskedasticity could be a signal of underlying model specification, and we should think hard about the cause of heteroskedasticity instead of applying a quick fix.

1.1 Simulating

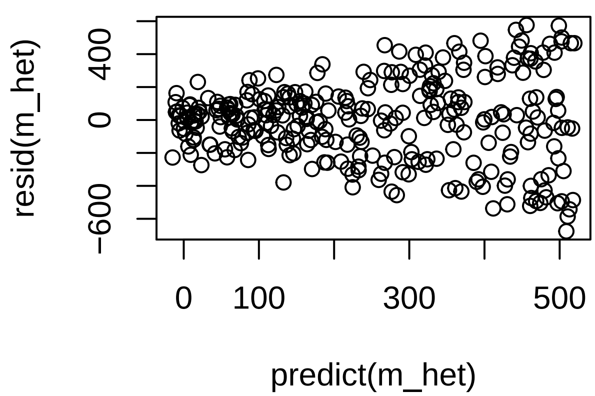
Simulate the spending and income pattern for three ethnic groups as described above. Re-create the two plots above. The numbers don't have to be the same – just make sure that your data has heteroskedasticity due to underlying heterogenous effect across ethnic groups as described in the example above. Note: Don't look at my code.

1.2 Diagnostics: Visual

Using the simulated data above, regress spending on income, plot the residual against the predicted value.

Solution

```
m_het <- lm(spending ~ income, data = d)
plot(predict(m_het), resid(m_het))
```



1.3 Diagonistics: Hypothesis test

Conduct BP test and White test. Why do the tests reach the same conclusion here, unlike in the lab tutorial?

Solution

```
library(AER)
bptest(m_het, varformula = ~ income, data = d)

##
```

```
##  studentized Breusch-Pagan test
##
## data:  m_het
## BP = 95.706, df = 1, p-value < 2.2e-16
```

```
bptest(m_het, varformula = ~ income + I(income^2), data = d)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  m_het
## BP = 103.16, df = 2, p-value < 2.2e-16
```

The test reaches the same conclusion because the variance of the error terms is a linear function of *income* (not of $income^2$, for example), so both the BP and the White tests are able to detect this.

1.4 Diagnostics: Repeat the White's test manually

Here's the instruction. Compare the result you get doing it by hand vs using R.

White test (Wooldridge "Introductory", Testing for heteroskedasticity)

1. Estimate the model `y ~ x_1 + x_2 + ... + x_k` by OLS, as usual. Obtain the OLS residual $\hat{u}$ and the fitted values $\hat{y}$. Compute $\hat{u}^2$ and $\hat{y}^2$.

2. Run the regression $\hat{u}^2 = \delta_0 + \delta_1 \hat{y} + \delta_2 \hat{y}^2$. Keep the R square.

3. **I want you to use the LM for this problem** Form either the F or LM statistic and compute the p-value (using the $F_{2,n-3}$ distribution in the former case and the χ^2_2 distribution in the latter case).

Solution

```
uhat <- resid(m_het) ; yhat <- predict(m_het)
m_white_stage2 <- lm(I(uhat^2) ~ yhat + I(yhat^2))
R_squared <- summary(m_white_stage2)$r.squared

n <- nrow(d) ; k <- 2 # k = 1 because we have 2 regressors, yhat and yhat squared
(LM_stat <- n * R_squared)
```

```
## [1] 103.157
```

```
1 - (pvalue <- pchisq(LM_stat, df = k))
```

```
## [1] 0
```

```
bptest(m_het, varformula = ~ yhat + I(yhat^2), data = d)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  m_het
## BP = 103.16, df = 2, p-value < 2.2e-16
```

The LM statistic and the p value are the same.

1.5 Fixing: robust standard error

Run hypothesis test without and with robust standard error. What's the conclusion?

Solution

```
summary(m_het)

##
## Call:
## lm(formula = spending ~ income, data = d)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -675.87 -149.60   19.05  135.80  576.52
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -17.82943   26.56899  -0.671    0.503
## income        0.53868    0.04639  11.611   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 243.8 on 298 degrees of freedom
## Multiple R-squared:  0.3115,Adjusted R-squared:  0.3092
## F-statistic: 134.8 on 1 and 298 DF,  p-value: < 2.2e-16

coeftest(m_het, vcov = vcovHC(m_het, type = "HC"))

##
## t test of coefficients:
##
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -17.829433  18.463054 -0.9657    0.335
## income        0.538684   0.051638 10.4319   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Both regressions show that income has a positive and significant impact on spending

1.6 Fixing: robust standard error by hand

Solution

```
m_robust1 <- lm(income ~ 1, data = d)
numerator <- sum((resid(m_robust1)**2) * (resid(m_het)**2))
SSR <- sum(resid(m_robust1)**2) ** 2

# Compare the two methods
(var_beta_x <- numerator / SSR)

## [1] 0.002666474

vcovHC(m_het, type = "HC")

##             (Intercept)       income
## (Intercept)   340.88438 -0.795780000
## income         -0.79578  0.002666474
```

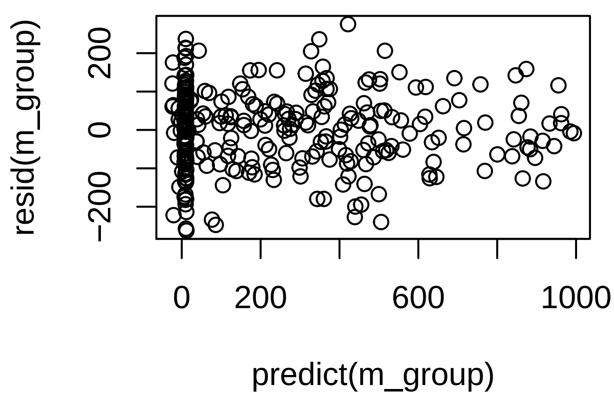
We see that the manual method and R's `vcovHC` gives the same robust standard error.

1.7 Fixing: Provide a correct model

Specify a regression model that takes into account heterogenous effect of income on spending across ethnic groups. Show that there's no longer heteroskedasticity.

Solution

```
m_group <- lm(spending ~ income + group + income:group, data = d)
plot(resid(m_group) ~ predict(m_group))
```



As shown in the diagnostics plot, there's no longer heteroskedasticity