Pol Sci 630: Problem Set 12: Heteroskedasticity, Autocorrelation

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

Due Date: Tue, Nov 17, 2015, 10 AM (Beginning of Class)

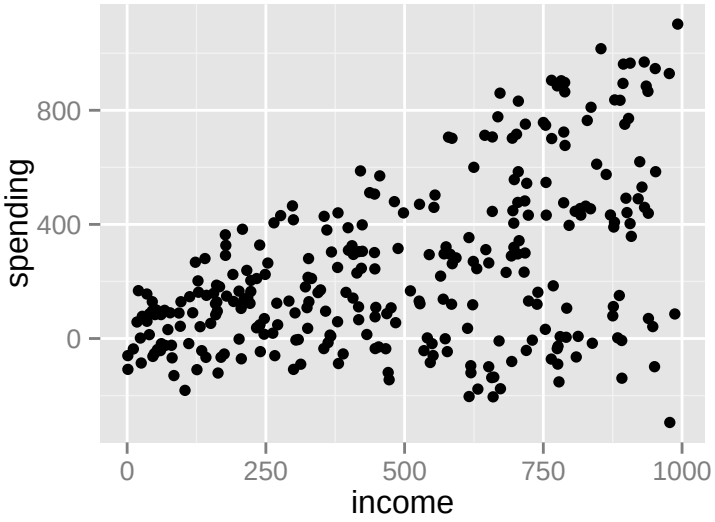
```
rm(list = ls())
library(ggplot2)
```

1 Heteroskedasticity (8 points)

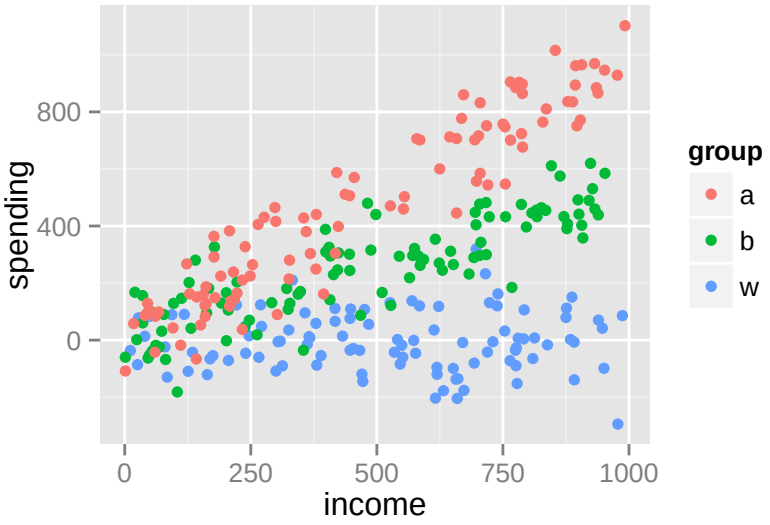
One common cause of heteroskedasticity is that our model does not take into account heterogenous effect across sub-populations. For example, we have a model of spending (dependent var) as a function of income (independent var), and the propensity to spend differs across ethnic groups. Formally,

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

where $\beta ethnic$ 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.

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?

1.4 Fixing: robust standard error

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

1.5 Fixing: FGLS

Conduct FGLS. Hint: For stability, log transform $residual^2$ in the auxiliary regression, then exponentiate the predicted value of the auxiliary regression to get the weight.

1.6 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.

2 Multicollinearity (4 points)

2.1 Diagnosing with VIF

Using dataset `Prestige`, run regression of prestiage against income, education, and women. Calculate VIF. Interpret the largest VIF.

2.2 Dealing with multicollinearity

If you are concerned that the VIF is causing your SEs to be pretty big. What should you do to address this issue?

3 Diagnosing autocorrelation (4 points)

3.1 Generating autocorrelated data

Similar to the lab, generate data (i.e. e, X, Y) that follow an AR(2) process, i.e.:

$$v(t) \sim N(0,1) \tag{2}$$

$$e(t) = a_1 e(t-1) + a_2 e(t-2) + v(t) \qquad \text{Important: } a_1 + a_2 < 1 \tag{3}$$

$$Y(t) = X(t) + e(t) \tag{4}$$

3.2 Diagnostics: Visual

Plot residual against time and against lagged , up to 4 lags (e.g. residual lag-1 residual, residual lag-2 residual, etc. up to 4 plots) How does the correlation look across the four plots?

3.3 Diagnostics: Hypothesis testing

Regress residuals against X and lag1 and lag2 residuals, and then doing an F test for joint significance in the lagged residuals.