Introduction

For this example, we have information on the consumption of beer for 500 individuals. We have the following information

- Number of bottles of beer bought in 2002
- Was the individual a male or female
- The price of a bottle of beer in their closest convenience store
- The price of a bottle of Coke in their closest convenience store

Introduction

Goal: Studying what determines beer consumption

- Which variable is the dependent variable? Number of bottles of beer
- We write y_i – number of bottles of beer fof individual i
- What are the explanatory variables? We have three explanatory variables:

  x_{i1} = 1 male, x_{i1}=0 female

  x_{i2}= price of a bottle of beer

  x_{i3}=price of a bottle of Coke

Before doing anything, we look at the data

```
proc means data=beer_cons;
run;
```

The MEANS Procedure

	N	Mean	Std Dev	Minimum	Maximum
beer	500	247.6600000	100.1217715	11.0000000	366.0000000
male	500	0.6180000	0.4863631	0	1.0000000
price_b	500	1.5046474	0.2871817	1.0025000	1.9987000
price_c	500	1.0027246	0.2886223	0.5067000	1.4989000

To study the y_i variable, we can regress it on a constant, x_{i1} , x_{i2}, and x_{i3}

The SAS code is

```
proc reg data=beer_cons;
  model beer = male price_b price_c / acov;
  title 'OLS for beer consumption';
  test price_b=0, price_c=0;
  output out=resdat_ols residual=uhat_ols predicted=yhat_ols;
run;
```

- The output is

The REG Procedure

Dependent Variable: beer

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	4771450	1590483	3419.35	<.0001
Error	496	230710	465.14212		
Corrected Total	499	5002160			

Root MSE	21.56715	R-Square	0.9539
Dependent Mean	247.66000	Adj R-Sq	0.9536
Coeff Var	8.70837		

Parameter Estimates

					Heteroscedasticity Consistent		
Variable	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standard Error	t Value	Pr > |t|
Intercept	197.12054	6.24628	31.56	<.0001	6.66570	29.57	<.0001
male	201.72346	1.99724	101.00	<.0001	2.47588	81.48	<.0001
price_b	-48.72603	3.37738	-14.43	<.0001	3.55793	-13.70	<.0001
price_c	-0.80795	3.35042	-0.24	0.8095	3.09636	-0.26	0.7942

- And the robust variances and covariance are

OLS for beer consumption
The REG Procedure
Dependent Variable: beer

Consistent Covariance of Estimates

variable	Intercept	male	price_b	price_c
intercept	44.431522734	-4.428401737	-17.69338802	-12.78407524
male	-4.428401737	6.1299754557	-2.222414478	2.0269925151
price_b	-17.69338802	-2.222414478	12.658896575	0.6057958513
price_c	-12.78407524	2.0269925151	0.6057958513	9.5874227615

With the test price_b=0, price_c=0 statement, you get a non-heteroskedasticity robust F test

OLS for beer consumption

The REG Procedure
Model: MODEL1

Test 1 Results for Dependent Variable beer				
Source	DF	Mean Square	F Value	Pr > F
Numerator	2	48414	104.09	<.0001
Denominator	496	465.14212		

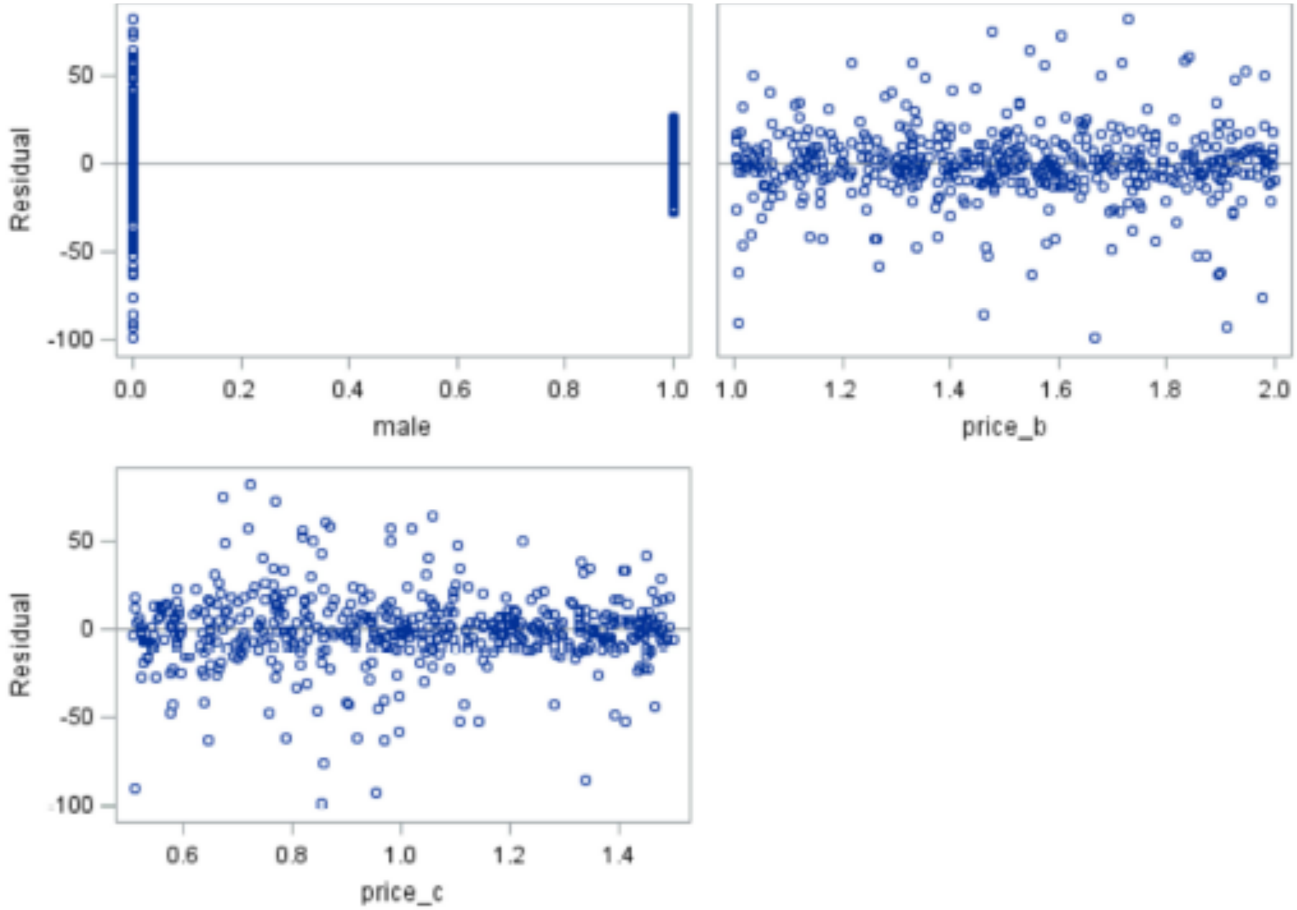
- And a heteroskedasticity robust Chi-square test

OLS for beer consumption

The REG Procedure
Model: MODEL1
Dependent Variable: beer

Test 1 Results using Heteroscedasticity Consistent Covariance Estimates		
DF	Chi-Square	Pr > ChiSq
2	187.80	<.0001

- Before doing any analysis/testing, we can check if there is something wrong with the regression model. We can look at the residuals. Is there something wrong with them? For example, is there a relationship between the explanatory variables and the dispersion of the residuals?


Residual
50
0
-50
-100
0.0
0.2
0.4
0.6
0.8
1.0
male
1.0
1.2
1.4
1.6
1.8
2.0
price_b
Residual
0.6
0.8
1.0
1.2
1.4
price_c

- There is more dispersion in the residuals for small values of the price_c
- There is more disperson for females than for males

Result: We can suspect that there is heteroskedasticity and that the variance of y_i depends on the sex of the individual and the price of a bottle of Coke

Testing for the presence of heteroskedasticity

- From the previous analysis, we think that the variance of the error term could vary with x_{i3} (price of a bottle of Coke) and x_{i1}(male)
- We do a formal test of heteroskedasticity

- We run the following regression:

$$\hat{u}_i^2 = \alpha_0 + \alpha_1 x_{1,i} + \alpha_3 x_{3,i} + v_i$$

and we test

$$H_0 \; : \; \alpha_1 = \alpha_3 = 0$$
$$H_1 \; : \; \alpha_1 \neq 0 \text{ and/or } \alpha_3 \neq 0$$

- The SAS code is

```
data resdat_ols;
 set resdat_ols;
 uhat2 = uhat_ols**2;
run;

proc reg data=resdat_ols;
 model uhat2 = male price_c;
 title 'Test for heteroskedasticity';
 test male=0, price_c=0;

run;
```

Test for heteroskedasticity

The REG Procedure

Dependent Variable: uhat2

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	115411724	57705862	54.36	<.0001
Error	497	527618317	1061606		
Corrected Total	499	643030041			

Root MSE	1030.34279	R-Square	0.1795
Dependent Mean	461.42098	Adj R-Sq	0.1762
Coeff Var	223.29778		

Variable	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1560.36736	179.94733	8.67	<.0001
male	-959.24578	94.98401	-10.10	<.0001
price_c	-504.75723	160.05940	-3.15	0.0017

F-value on joint test: 54.36,

Associated p-value<.0001

Reject null of homoskedasticity

Could also do LM

N*unadusted R-square = 0.1795*500 = 89.75

Chi-square, 2 degrees of freedom.005=10.597

Reject null of homoskedasticity

- We see that the variance of the residuals is affected by the dummy variable male. The effect is statistically significant at the 1% significance level since the p-value is below 1%

- We see that the variance of the residuals is affected by the price of a bottle of Coke. The effect is statistically significant at the 1% significance level since the p-value is below 1%

FGLS estimation instead of OLS

- Instead of simply using robust standard errors, we can treat the heteroskedasticity in the estimation. We do FGLS instead of OLS
- Earlier we saw that the variance depends on the gender and the price of a bottle of Coke.

- To make sure $\hat{\sigma}_1^2$ is positive we use the exponential:

$$Var[u_i] = \exp(\gamma_0 + \gamma_1 x_{1,i} + \gamma_3 x_{3,i})$$

- We estimate

$$\ln(\hat{u}_i^2) = \gamma_0 + \gamma_1 x_{1,i} + \gamma_3 x_{3,i} + e_i$$

- The predicted variance will be

$$\hat{\sigma}_i^2 = \exp(\hat{\gamma}_0 + \hat{\gamma}_1 x_{1,i} + \hat{\gamma}_3 x_{3,i})$$

SAS code to do FGLS with this functional form is

```
data resdat_ols;
  set resdat_ols;
  log_res_2 = log(uhat_ols**2);
run;

proc reg data=resdat_ols;
  model log_res_2 = male price_c;
  title 'Estimation and prediation of variance';
  output out=res_var predicted=log_h_hat;
run;

data res_var;
  set res_var;
  h_hat = exp(log_h_hat);
  one_over_h = 1/h_hat;
run;

proc reg data=res_var;
  model beer = male price_b price_c;
  weight one_over_h;
  title 'FGLS for beer consumption (using weight)';
run;
```

Estimation and prediation of variance

The REG Procedure

Dependent Variable: log_res_2

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	835.62789	417.81395	77.23	<.0001
Error	497	2688.93801	5.41034		
Corrected Total	499	3524.56590			

Root MSE	2.32601	R-Square	0.2371
Dependent Mean	4.03019	Adj R-Sq	0.2340
Coeff Var	57.71476		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	7.35871	0.40623	18.11	<.0001
male	1	-2.50243	0.21443	-11.67	<.0001
price_c	1	-1.77718	0.36134	-4.92	<.0001

FGLS for beer consumption (using weight)

The REG Procedure

Dependent Variable: beer

Weight: one_over_h

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	35808	11936	3000.66	<.0001
Error	496	1972.97750	3.97778		
Corrected Total	499	37781			

Root MSE	1.99444	R-Square	0.9478
Dependent Mean	313.83350	Adj R-Sq	0.9475
Coeff Var	0.63551		

Parameter Estimates

Variable	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	199.20201	4.05389	49.14	<.0001
male	202.14563	2.21808	91.14	<.0001
price_b	-50.35661	1.73218	-29.07	<.0001
price_c	-0.77552	1.82380	-0.43	0.6709

More analysis: Looking at the final output from the FGLS procedure

- Ceteris paribus, a male will buy 202 more bottles of beer than a female. This is statistically significant at the 1% level.
- Ceteris paribus, if the price of a bottle of beer goes up by $1, a consumer will reduce their annual consumption of beer by 50 bottles
- Variations in the price of a bottle of Coke does not have a statistically significant impact on the consumption of beer at the 10% level

The FGLS can also be performed manually

```
data res_var;
  set res_var;
  beer_star = beer / sqrt(h_hat);
  one_star = 1 / sqrt(h_hat);
  male_star = male / sqrt(h_hat);
  price_b_star = price_b / sqrt(h_hat);
  price_c_star = price_c / sqrt(h_hat);
run;


proc reg data=res_var;
  model beer_star = one_star male_star price_b_star price_c_star /noint;
  title 'FGLS for beer consumption (using star)';
run;
```

```
FGLS for beer consumption (using star)

The REG Procedure
Dependent Variable: beer_star

Parameter Estimates

                 Parameter  7  Standard
Variable          Estimate        Error    t Value    Pr > |t|
one_star         199.20201      4.05389      49.14      <.0001
male_star        202.14563      2.21808      91.14      <.0001
price_b_star     -50.35661      1.73218     -29.07      <.0001
price_c_star      -0.77552      1.82380      -0.43      0.6709
```

- As expected, we get the same estimated values and standard errors.

Why not combine the best of both worlds (FGLS + hetero-robust standard errors and statistics):

```
proc reg data=res_var;
  model beer = male price_b price_c / acov;
  weight one_over_h;
  title 'FGLS for beer consumption (using weight)';
run;
```