Chapter 8 Code

Ryan Martin

February 28, 2018

I found an R version of Principles of Econometrics here

https://bookdown.org/ccolonescu/RPoE4/

It's great! Does all the examples in the text with R. May be very helpful if you want to do econometrics with R.

Chapter 8 - Heteroskedasticity

We learned before that heteroskedasticity is when the variance of the error terms changes with the covariates (x's). Heteroskedasticity violates (one of) the assumptions in SR1-SR5 and MR1-MR5. Note, in general, heteroskedasticity isn't the only way the variances could change. Correlation between the x's and residuals is not necessary or sufficient for heteroskedasticity. The residuals could vary by mean or have the same mean and variance, but still have a changing distribution! However, if they are normal, then the same mean and variance means not changing. Usually, changes in mean we hope to catch with "fixed effects" (as in chapter 7)

Heteroskedasticity is common in cross-sectional data. Cross-sectional data is common in economics (its expensive and takes patience to collect time series/panel data). Hence, heteroskedasticity is a well-known concern in economics.

Least Squares consequences

1. Under heteroskedasticity, (with the errors still uncorrelated with the covariates!) the OLS term is unbiased but no longer the variance minimizing estimate (remember, the variance minimizing estimate is called "best").
2. Standard errors will be incorrectly estimated, so CI and hypothesis tests will be misleading.

If $y_i = \beta_1 + \beta_2 x_i + e_i$ and $var(e_i) = \sigma^2$ (homoskedastic),

$$var(b_2) = \frac{\sigma^2}{\sum(x_i - \bar{x}^2)}$$

However, if $var(e_i) = \sigma_i^2$ (heteroskedastic),

$$var(b_2) = \frac{\sum(x_i - \bar{x})^2 sigma_i^2}{[\sum(x_i - \bar{x}^2)]^2}$$

Detecting Heteroskedasticity

1. Residual plots (like chapter 2)
2. Lagrange Multiplier Tests (a.ka. Breusch-Pagan test)
3. White test
4. Goldfeld-Quandt Test

Residual Plots

- like in chapter 2

Lagrange Multiplier Tests

Estimate the (extra) regression

$$\hat{e_i}^2 = \alpha_1 + \sum_{j=2}^{S} \alpha_j z_{ij} + v_i$$

Where z_{ij} are transforms of the x covariates. Turns out

$$\chi^2_{S-1} = N \times R^2$$

, where the R^2 is from the extra (not the original) regression. More details in Appendix 8B.

Turns out, the statistic from the linear one is valid for testing an alternative hypohtesis of heteroskedasticity where the variance is any function $h(\alpha_1 + \sum_{j=2}^{S} \alpha_j z_{ij})$

White Test

Specific version of Lagrange Multiplier tests. The z's are the x's, x's squared and interaction terms.

The Goldfeld-Quandt Test

For when can partition (separate everyone) into two classes.

If have two groups with different errors, σ_1^2 and σ_2^2, then under the null $\sigma_1^2 = \sigma_2^2$, $\frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} \sim F_{N_1-K_1, N_2-K_2}$

So, we can estimate the variance for the two groups and then do an F-test.

Can create groups from the x's if ther are no natural groups.

8.3 Estimating with Heteroskedasticity

Use White's heteroskedasticity-consistent standard errors a.k.a. heteroskedasticity robust standard errors a.k.a. robust standard errors.

For $y = \beta_1 + \beta_2 x + e$, then

$$\widehat{var}(b_2) = \frac{N}{N-2} \frac{\sum_{i=1}^{N} (x_i - \bar{x})^2 \hat{e_i}^2}{[\sum (x_i - \bar{x})^2]^2}$$

Note that this heteroskedasticity-consistent standard error *does not* find variance as a function of the covariates. Instead, it just comes up with a different variance estimate. A different way of averaging the variance.

There is matrix version too, for larger linear regressions.

GLS (Generalized Least Squares)

- Great if you actually know the variances, but when you don't it may be worse!

Cannot estimate n variances with n data points. Need to impose some structure. For example, can suppose

$$var(e_i) = \sigma_i^2 = \sigma^2 x_i$$

(What does this mean for our hypothesis tests?)

Remember that x_i are assumed fixed. Then consider rescaling by $\sqrt{x_i}$. Get (in bivariate regression case)

$$\frac{y_i}{\sqrt{x_i}} = \beta_1 \frac{1}{\sqrt{x_i}} + \beta_2 \frac{x_i}{\sqrt{x_i}} + \frac{e_i}{\sqrt{x_i}}$$

If define $y_i^\star = \frac{y_i}{\sqrt{x_i}}$, $x_{i1}^\star = \frac{1}{\sqrt{x_i}}$, $x_{i2}^\star = \frac{x_i}{\sqrt{x_i}}$ and $e_i^\star = \frac{e_i}{\sqrt{x_i}}$ then now our (interceptless!) regression can be run with homoskedasticity.

Heteroskedasticity in the Linear Probability Model - Covered in Problem 8.22 below

8.22

In Exercise 7.7 we considered a model designed to provide information to mortgage lenders. They want to determine borrower and loan factors that may lead to delinquency or foreclosure. In the file `lasvegas.dat`, there are 1000 observations on mortgages for single-family homes in Las Vegas, Nevada during 2008. The variable of interest is DELINQUENT, an indicator variable = 1 if the borrower missed at least three payments (90+ days late), but 0 otherwise. Explanatory variables are LVR = the ratio of the loan amount to the value of the property; REF = 1 if purpose of the loan was a "refinance" and 0 if the loan was for a purchase; INSUR = 1 if mortgage carries mortgage insurance, 0 otherwise; RATE = initial interest rate of the mortgage; AMOUNT = dollar value of mortgage (in $100,000); CREDIT = credit score, TERM = number of years between disbursement of the loan and the date it is expected to be fully repaid, ARM = 1 if mortgage has an adjustable rate, and 0 if the mortgage has a fixed rate.

a

Estimate the linear probability (regression) model explaining DELINQUENT as a function of the remaining variables. Use the White test with cross-product terms included to test for heteroskedasticity. Why did we include the cross-product terms?

Solution White's test is included below. Note that this is White's test for heteroskedasticity, not his heteroskedasticity corrected variance estimator. White has two things named after him in this chapter; he was a prolific academic with many interests.

The white test rejects the null at the .05 level. p-value is very small. Looks like heteroskedasticity is present.

Why include the cross-product terms? Well, in general because you want to leave room for the variance to depend on the combined presence of two variables. Note that, unfortunately, all of the packages I have found for R to do White's test do not generically include cross product terms or squared terms. So, we had to do it by hand. I showed the packaged regressions too (which are just errors squared on all the independent variables in the regression)

```r
my_wd <- "C:/Users/ryanj/Dropbox/TA/Econ 103/Winter 2018/Data/s4poe_statadata"
my_file <- paste(my_wd, "lasvegas.dta", sep = "/")
library(haven)
dat <- read_stata(my_file)
attach(dat)
#View(dat)

unique(delinquent) #only 0, 1. LHS is a dummy/bernoulli r.v.
```

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

```r
reg_out <- lm(data = dat, delinquent ~.)
  # ~. means regress on everything in data!
s <- summary(reg_out)
s
```

```
##
## Call:
## lm(formula = delinquent ~ ., data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.78293 -0.13202 -0.06188  0.07247  1.06843
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.6884913  0.2112456   3.259 0.001155 **
## lvr          0.0016239  0.0007846   2.070 0.038732 *
## ref         -0.0593237  0.0238299  -2.489 0.012957 *
## insur       -0.4815849  0.0236365 -20.375  < 2e-16 ***
## rate         0.0343761  0.0085999   3.997 6.88e-05 ***
## amount       0.0237680  0.0126699   1.876 0.060957 .
## credit      -0.0004419  0.0002018  -2.190 0.028782 *
## term        -0.0126195  0.0035386  -3.566 0.000379 ***
## arm          0.1283239  0.0318867   4.024 6.15e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3267 on 991 degrees of freedom
## Multiple R-squared:  0.3363, Adjusted R-squared:  0.3309
## F-statistic: 62.77 on 8 and 991 DF,  p-value: < 2.2e-16
```

```r
res.squared <- s$residuals^2
dat2 <- cbind(res.squared, dat[, -ncol(dat)]) #remove delinquent
  #and add res.squared
white.reg <- lm(data = dat2, res.squared ~ .*. + I(lvr^2) + I(ref^2) +
     I(insur^2) + I(rate^2) + I(amount^2) + I(credit^2) +
     I(term^2) + I(arm^2))
s2 <- summary(white.reg)
s2
```

```
##
## Call:
## lm(formula = res.squared ~ . * . + I(lvr^2) + I(ref^2) + I(insur^2) +
##     I(rate^2) + I(amount^2) + I(credit^2) + I(term^2) + I(arm^2),
```

```
##     data = dat2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.27531 -0.06393 -0.03537 -0.00238  1.11124
##
## Coefficients: (4 not defined because of singularities)
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    4.869e-01  1.408e+00   0.346  0.72959
## lvr           -4.207e-03  8.275e-03  -0.508  0.61128
## ref           -1.427e-01  2.561e-01  -0.557  0.57766
## insur         -2.074e-01  2.467e-01  -0.841  0.40056
## rate          -9.218e-02  9.866e-02  -0.934  0.35039
## amount         1.579e-01  1.652e-01   0.956  0.33935
## credit         1.092e-03  2.546e-03   0.429  0.66793
## term          -1.185e-02  3.950e-02  -0.300  0.76421
## arm           -1.216e-01  3.046e-01  -0.399  0.68987
## I(lvr^2)      -2.687e-05  2.399e-05  -1.120  0.26312
## I(ref^2)              NA         NA      NA       NA
## I(insur^2)            NA         NA      NA       NA
## I(rate^2)      2.369e-03  2.250e-03   1.053  0.29263
## I(amount^2)    3.195e-03  3.230e-03   0.989  0.32288
## I(credit^2)   -1.846e-07  1.405e-06  -0.131  0.89553
## I(term^2)      8.805e-05  4.571e-04   0.193  0.84730
## I(arm^2)              NA         NA      NA       NA
## lvr:ref        1.506e-03  1.138e-03   1.322  0.18633
## lvr:insur      6.097e-04  9.680e-04   0.630  0.52896
## lvr:rate       1.351e-04  3.910e-04   0.346  0.72973
## lvr:amount    -1.017e-05  5.779e-04  -0.018  0.98596
## lvr:credit    -7.105e-06  8.373e-06  -0.848  0.39638
## lvr:term       3.460e-04  1.296e-04   2.670  0.00771 **
## lvr:arm       -2.256e-04  1.428e-03  -0.158  0.87457
## ref:insur     -1.520e-02  2.799e-02  -0.543  0.58708
## ref:rate       1.280e-02  1.032e-02   1.240  0.21512
## ref:amount     1.472e-02  1.562e-02   0.942  0.34627
## ref:credit     4.638e-05  2.568e-04   0.181  0.85674
## ref:term      -4.563e-03  5.321e-03  -0.858  0.39134
## ref:arm        1.736e-02  3.794e-02   0.458  0.64732
## insur:rate     1.757e-02  1.021e-02   1.721  0.08550 .
## insur:amount   7.081e-03  1.557e-02   0.455  0.64935
## insur:credit   1.681e-05  2.321e-04   0.072  0.94228
## insur:term    -4.665e-03  4.191e-03  -1.113  0.26595
## insur:arm     -2.037e-02  4.023e-02  -0.506  0.61271
## rate:amount   -8.355e-03  6.123e-03  -1.365  0.17269
## rate:credit   -9.724e-06  8.964e-05  -0.108  0.91364
## rate:term      1.464e-03  1.659e-03   0.882  0.37779
## rate:arm       2.877e-02  1.622e-02   1.774  0.07641 .
## amount:credit  1.680e-04  1.258e-04   1.336  0.18192
## amount:term   -7.609e-03  4.146e-03  -1.835  0.06678 .
## amount:arm     5.936e-03  3.028e-02   0.196  0.84461
## credit:term   -1.993e-05  3.503e-05  -0.569  0.56948
## credit:arm    -6.591e-05  3.403e-04  -0.194  0.84648
## term:arm              NA         NA      NA       NA
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1667 on 959 degrees of freedom
## Multiple R-squared:   0.22,  Adjusted R-squared:  0.1874
## F-statistic: 6.761 on 40 and 959 DF,  p-value: < 2.2e-16
```

```
Chi.stat = nrow(dat2)* s2$r.squared
qchisq(.05, s2$df[1])
```

```
## [1] 27.32555
```

```
#reject the null, heteroskedasticity present


#to run white test can also use packages
#but these packages don't include squared or interaction terms
#Easier to just use package

library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
# just regresses errors squared on independent vars
# not squared or interaction terms
bpreg <- bptest(reg_out)
bpreg
```

```
##
##  studentized Breusch-Pagan test
##
## data:  reg_out
## BP = 184.99, df = 8, p-value < 2.2e-16
```

```
#alternative
# just regresses errors squared on independent vars
# not squared or interaction terms
library(olsrr)
```

```
##
## Attaching package: 'olsrr'
```

```
## The following object is masked from 'package:datasets':
##
##     rivers
```

```
ols_bp_test(reg_out)
```

```
##
##  Breusch Pagan Test for Heteroskedasticity
##  -----------------------------------------
##  Ho: the variance is constant
##  Ha: the variance is not constant
```

```
##
##                  Data
##  -------------------------------------
##  Response : delinquent
##  Variables: fitted values of delinquent
##
##          Test Summary
##  -----------------------------
##  DF            =    1
##  Chi2          =    272.8432
##  Prob > Chi2   =    2.724374e-61
```

b

Use the estimates from (a) to estimate the error variances for each observation. How many of these estimates are at least one? How many are at most 0? How many are less than .01?

Solution Note that we get one variance estimate per row in our data. That is, each of these are just the predicted variances per line of data. Note that since we have the LHS/y variable is a dummy (or Bernoulli) random variable, the variance of y_i should be $P(Y_i = 1) \times (1 - P(Y_i = 1))$ Thus, our variance estimate is just $\hat{y}_i \times (1 - \hat{y}_i)$. This is because our regression model is a linear probability model; $\hat{y}_i$ is the estimated probability our random variable is 1.

You may recall that the linear probability model has the problem that some probabilities will be below 0 or bigger than 1. These points variance estimates will be negative (look at formula for variance, $\hat{y}_i \times (1 - \hat{y}_i)$ to see why). We are counting how many of these problem points there are in this part of the problem

```
predicted_var <-  predict(reg_out)*(1 - predict(reg_out))
sum(predicted_var >1)
```

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

```
sum(predicted_var < 0)
```

```
## [1] 135
```

```
sum(predicted_var<.01) - sum(predicted_var < 0)
```

```
## [1] 23
```

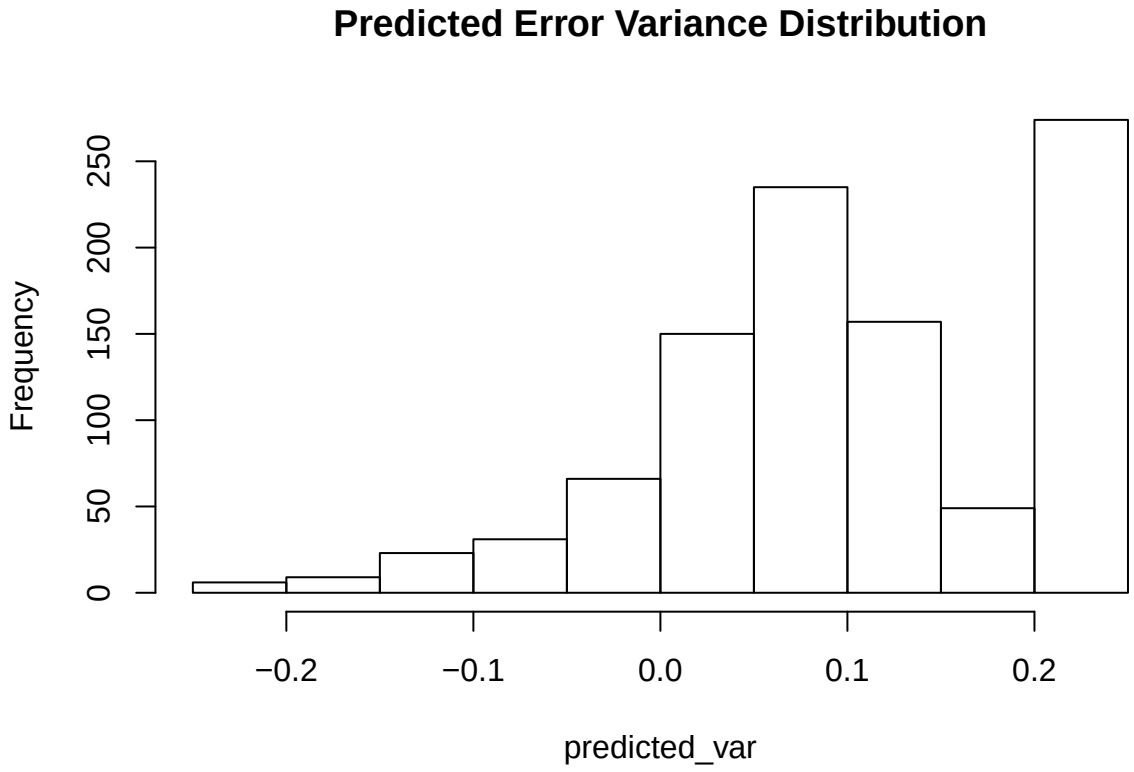
```
length(predicted_var) - sum(predicted_var<.01) #regular
```

```
## [1] 842
```

```
summary(predicted_var)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -0.24552  0.04137  0.09543  0.10579  0.22543  0.25000
```

```r
hist(predicted_var, main = "Predicted Error Variance Distribution")
```



c

Prepare a table containing estimates and standard errors from estimating the linear probability model in each of the following ways:

i. Least squares with conventional standard errors.

ii. Least squares with heteroskedasticity-robust standard errors

iii. Generalized least squares omitting observations with variance less than .01.

iv. Generalized least squares with variance less than .01 changed to .01

v Generalized least squares with variance less than .00001 changed to .00001.

Discuss and compare the different results.

Solution: Let me write the regression equation as

$$DELINQUENT_i = \beta_1 + \beta_2 x_{2i} + \cdots + \beta_K x_{Ki} + e_i$$

Since this is a linear probability model, the GLS solution is the regression

$$\frac{DELINQUENT_i}{\sqrt{\hat{DELINQUENT_i}(1 - \hat{DELINQUENT_i})}} = \frac{\beta_1}{\sqrt{\hat{DELINQUENT_i}(1 - \hat{DELINQUENT_i})}} +$$

$$\frac{\beta_2 x_{2i}}{\sqrt{DELIN\hat{Q}UENT_i(1-DELIN\hat{Q}UENT_i)}}+$$
$$\cdots+\frac{\beta_K x_{Ki}}{\sqrt{DELIN\hat{Q}UENT_i(1-DELIN\hat{Q}UENT_i)}}+\tilde{e}_i$$

Note that the hats are the OLS weights, not any other weights. We can do this with a weighted regression in R. Note that the weights should be $\frac{1}{\hat{y}_i(1-\hat{y}_i)}$ and not $\frac{1}{\sqrt{\hat{y}_i(1-\hat{y}_i)}}$. R applies the square root on its own.

```
summary(reg_out, robust = T)
```

```
##
## Call:
## lm(formula = delinquent ~ ., data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.78293 -0.13202 -0.06188  0.07247  1.06843
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.6884913  0.2112456   3.259 0.001155 **
## lvr          0.0016239  0.0007846   2.070 0.038732 *
## ref         -0.0593237  0.0238299  -2.489 0.012957 *
## insur       -0.4815849  0.0236365 -20.375  < 2e-16 ***
## rate         0.0343761  0.0085999   3.997 6.88e-05 ***
## amount       0.0237680  0.0126699   1.876 0.060957 .
## credit      -0.0004419  0.0002018  -2.190 0.028782 *
## term        -0.0126195  0.0035386  -3.566 0.000379 ***
## arm          0.1283239  0.0318867   4.024 6.15e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3267 on 991 degrees of freedom
## Multiple R-squared:  0.3363, Adjusted R-squared:  0.3309
## F-statistic: 62.77 on 8 and 991 DF,  p-value: < 2.2e-16
```

```
#putting in a table
library(stargazer)
```

```
##
## Please cite as:
```

```
##  Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.
```

```
##  R package version 5.2.1. https://CRAN.R-project.org/package=stargazer
```

```
#install.packages("stargazer")
library(sandwich)
cov <- vcovHC(reg_out, type = "HC") #White errors
robust.se <- sqrt(diag(cov))
library(nlme)

my_gls_weight = 1/predicted_var

#For case 3
case3 = predicted_var>.01
```

```r
datcase3 = dat[case3,]
my_gls_weight3 = my_gls_weight[case3]


#For case 4
my_gls_weight4 <- my_gls_weight
my_gls_weight4[predicted_var<.01] <- 1/.01
#my_gls_weight4 = sqrt(my_gls_weight4) #getting sd

#For case 5
my_gls_weight5 <- my_gls_weight
my_gls_weight5[predicted_var<.00001] <- 1/.00001
#my_gls_weight5 = sqrt(my_gls_weight5) #getting sd

reg3 <- lm(data = datcase3,
           delinquent ~. , weights = my_gls_weight3)

reg4 <- lm(data = dat, delinquent ~. , weights = my_gls_weight4)

reg5 <- lm(data = dat, delinquent ~. , weights = my_gls_weight5)

reg_gls1 <-  gls(data = datcase3, delinquent ~ ., method = "ML")
summary(reg_gls1)
```

```
## Generalized least squares fit by maximum likelihood
##   Model: delinquent ~ .
##   Data: datcase3
##        AIC      BIC    logLik
##   625.4539 672.8117 -302.7269
##
## Coefficients:
##                  Value  Std.Error    t-value p-value
## (Intercept)  0.8692091 0.24701664   3.518828  0.0005
## lvr          0.0019602 0.00103726   1.889769  0.0591
## ref         -0.0754943 0.02771082  -2.724362  0.0066
## insur       -0.5251805 0.02718903 -19.315901  0.0000
## rate         0.0431077 0.01006702   4.282071  0.0000
## amount       0.0277054 0.01403535   1.973972  0.0487
## credit      -0.0005796 0.00023894  -2.425791  0.0155
## term        -0.0227552 0.00474131  -4.799344  0.0000
## arm          0.2709861 0.05005686   5.413566  0.0000
##
##  Correlation:
##        (Intr) lvr    ref    insur  rate   amount credit term
## lvr    -0.034
## ref    -0.398 -0.026
## insur  -0.159 -0.105  0.050
## rate   -0.682  0.266  0.194 -0.095
## amount -0.048 -0.167 -0.078  0.162  0.115
## credit -0.805 -0.030  0.441  0.093  0.290 -0.037
## term   -0.463 -0.500  0.024  0.201  0.128 -0.020  0.135
## arm     0.043 -0.256  0.056 -0.218  0.094 -0.118  0.123 -0.380
##
## Standardized residuals:
```

```
##        Min         Q1        Med         Q3        Max
## -2.4884146 -0.4051950 -0.1723949  0.1018240  2.8512505
##
## Residual standard error: 0.3466619
## Degrees of freedom: 842 total; 833 residual
```

```
summary(reg3)
```

```
##
## Call:
## lm(formula = delinquent ~ ., data = datcase3, weights = my_gls_weight3)
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -2.0387 -0.4076 -0.2462  0.0743  4.9420
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.7156853  0.1952374   3.666 0.000262 ***
## lvr          0.0015858  0.0008091   1.960 0.050328 .
## ref         -0.0570920  0.0211370  -2.701 0.007053 **
## insur       -0.5015883  0.0292346 -17.157  < 2e-16 ***
## rate         0.0413004  0.0082189   5.025 6.16e-07 ***
## amount       0.0257738  0.0120687   2.136 0.033004 *
## credit      -0.0003825  0.0001845  -2.073 0.038465 *
## term        -0.0190100  0.0040514  -4.692 3.16e-06 ***
## arm          0.2088645  0.0406664   5.136 3.50e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9156 on 833 degrees of freedom
## Multiple R-squared:  0.3049, Adjusted R-squared:  0.2982
## F-statistic: 45.67 on 8 and 833 DF,  p-value: < 2.2e-16
```

```
summary(reg4)
```

```
##
## Call:
## lm(formula = delinquent ~ ., data = dat, weights = my_gls_weight4)
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -1.6374 -0.3242 -0.2344 -0.0032  9.8670
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.5586635  0.1316754   4.243 2.42e-05 ***
## lvr          0.0008577  0.0003790   2.263 0.023863 *
## ref         -0.0326787  0.0146471  -2.231 0.025900 *
## insur       -0.4769852  0.0296842 -16.069  < 2e-16 ***
## rate         0.0203535  0.0056992   3.571 0.000372 ***
## amount       0.0187387  0.0099430   1.885 0.059774 .
## credit      -0.0001617  0.0001183  -1.366 0.172146
## term        -0.0065417  0.0020826  -3.141 0.001733 **
## arm          0.0419310  0.0139866   2.998 0.002786 **
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9646 on 991 degrees of freedom
## Multiple R-squared:  0.2661, Adjusted R-squared:  0.2602
## F-statistic: 44.92 on 8 and 991 DF,  p-value: < 2.2e-16
```

```
summary(reg5)
```

```
##
## Call:
## lm(formula = delinquent ~ ., data = dat, weights = my_gls_weight5)
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -16.636  -0.413  -0.130  -0.049 308.704
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.616e-01  4.188e-01   1.341   0.1803
## lvr          5.436e-04  2.433e-04   2.234   0.0257 *
## ref         -2.665e-02  1.047e-02  -2.547   0.0110 *
## insur       -5.127e-01  4.086e-01  -1.255   0.2099
## rate         1.722e-04  4.841e-03   0.036   0.9716
## amount      -4.528e-03  8.923e-03  -0.507   0.6120
## credit      -2.350e-05  8.535e-05  -0.275   0.7831
## term        -1.231e-03  1.814e-03  -0.679   0.4975
## arm          1.875e-02  1.086e-02   1.727   0.0845 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.02 on 991 degrees of freedom
## Multiple R-squared:  0.01662,    Adjusted R-squared:  0.008682
## F-statistic: 2.094 on 8 and 991 DF,  p-value: 0.03386
```

```
dattest = dat*sqrt(my_gls_weight5)
dattest2 = cbind(dattest, new_const = sqrt(my_gls_weight5))
regtest5 <- lm(data = dattest2, delinquent ~. + 0)
summary(regtest5) #same as above
```

```
##
## Call:
## lm(formula = delinquent ~ . + 0, data = dattest2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.636  -0.413  -0.130  -0.049 308.704
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## lvr         5.436e-04  2.433e-04   2.234   0.0257 *
## ref        -2.665e-02  1.047e-02  -2.547   0.0110 *
## insur      -5.127e-01  4.086e-01  -1.255   0.2099
## rate        1.722e-04  4.841e-03   0.036   0.9716
## amount     -4.528e-03  8.923e-03  -0.507   0.6120
```

```
## credit    -2.350e-05  8.535e-05  -0.275   0.7831
## term      -1.231e-03  1.814e-03  -0.679   0.4975
## arm        1.875e-02  1.086e-02   1.727   0.0845 .
## new_const  5.616e-01  4.188e-01   1.341   0.1803
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.02 on 991 degrees of freedom
## Multiple R-squared:  0.03125,    Adjusted R-squared:  0.02245
## F-statistic: 3.552 on 9 and 991 DF,  p-value: 0.0002358
```

```
dattest3 = dat*sqrt(my_gls_weight4)
dattest4 = cbind(dattest3, new_const = sqrt(my_gls_weight4))
regtest4 <- lm(data = dattest4, delinquent ~. + 0)
summary(regtest4) #same as above
```

```
##
## Call:
## lm(formula = delinquent ~ . + 0, data = dattest4)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.6374 -0.3242 -0.2344 -0.0032  9.8670
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## lvr        0.0008577  0.0003790   2.263 0.023863 *
## ref       -0.0326787  0.0146471  -2.231 0.025900 *
## insur     -0.4769852  0.0296842 -16.069  < 2e-16 ***
## rate       0.0203535  0.0056992   3.571 0.000372 ***
## amount     0.0187387  0.0099430   1.885 0.059774 .
## credit    -0.0001617  0.0001183  -1.366 0.172146
## term      -0.0065417  0.0020826  -3.141 0.001733 **
## arm        0.0419310  0.0139866   2.998 0.002786 **
## new_const  0.5586635  0.1316754   4.243 2.42e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9646 on 991 degrees of freedom
## Multiple R-squared:  0.3028, Adjusted R-squared:  0.2964
## F-statistic: 47.82 on 9 and 991 DF,  p-value: < 2.2e-16
```

```
#this creates latex for a table. very convenient
#stargazer(reg_out, reg_out, reg3, reg4, reg5,
#    title = "Regression Comparisons",
#    se=list(NULL, robust.se, NULL,
#  NULL, NULL),  column.labels = c("default","robust", "Censored", #"Round .01", "Round .00001"), colum
#  font.size = "small", omit.stat = c("f"), digits = 4)
```

d

Using the results from (iv.), interpret each of the coefficients. Mention whether the signs are reasonable and whether they are significantly different from 0.

Solution For 4, only amount and credit are insignificant. It's a little surprising that credit history is not

Table 1: Regression Comparisons

	Dependent variable:				
	default	robust	delinquent Censored	Round .01	Round .00001
	(1)	(2)	(3)	(4)	(5)
lvr	0.0016**	0.0016**	0.0016*	0.0009**	0.0005**
	(0.0008)	(0.0007)	(0.0008)	(0.0004)	(0.0002)
ref	−0.0593**	−0.0593**	−0.0571***	−0.0327**	−0.0267**
	(0.0238)	(0.0239)	(0.0211)	(0.0146)	(0.0105)
insur	−0.4816***	−0.4816***	−0.5016***	−0.4770***	−0.5127
	(0.0236)	(0.0302)	(0.0292)	(0.0297)	(0.4086)
rate	0.0344***	0.0344***	0.0413***	0.0204***	0.0002
	(0.0086)	(0.0098)	(0.0082)	(0.0057)	(0.0048)
amount	0.0238*	0.0238*	0.0258**	0.0187*	−0.0045
	(0.0127)	(0.0144)	(0.0121)	(0.0099)	(0.0089)
credit	−0.0004**	−0.0004**	−0.0004**	−0.0002	−0.00002
	(0.0002)	(0.0002)	(0.0002)	(0.0001)	(0.0001)
term	−0.0126***	−0.0126***	−0.0190***	−0.0065***	−0.0012
	(0.0035)	(0.0035)	(0.0041)	(0.0021)	(0.0018)
arm	0.1283***	0.1283***	0.2089***	0.0419***	0.0188*
	(0.0319)	(0.0276)	(0.0407)	(0.0140)	(0.0109)
Constant	0.6885***	0.6885***	0.7157***	0.5587***	0.5616
	(0.2112)	(0.2275)	(0.1952)	(0.1317)	(0.4188)
Observations	1,000	1,000	842	1,000	1,000
R^2	0.3363	0.3363	0.3049	0.2661	0.0166
Adjusted R^2	0.3309	0.3309	0.2982	0.2602	0.0087
Residual Std. Error	0.3267 (df = 991)	0.3267 (df = 991)	0.9156 (df = 833)	0.9646 (df = 991)	14.0217 (df = 991)

Note: *p<0.1; **p<0.05; ***p<0.01

significant in predicting delinquent probabilities. It is negative and significant (at the .05 level) in all other regressions (as would be suspected). It is always weak, however. 1 unit increase in credit score decreases the average probability of default by no more than .0004 across all regressions.

The most practically significant term appears to be insurance. It is highly significant (at the .001 level) and has a large size. The interpretation is that a person with mortgage insurance will default with a probability 47.8% lower than a person without, all else constant.

Chapter 8 New Stata Code

```
estat imtest, white //does white test for heteroskedasticity
reg y x [aweight=1/var] //does weighted regression,
      // same as rescaling all by 1/sqrt(var)
reg y x, vce(robust) //use white's robust standard errors
estata hettest income, iid #BP test
```

Homework Problems: 8.4, 8.6, 8.8

Skills: 8.4 (data: `vacation.dta`) - Interpret residual plots. - Goldfeld-Quandt test - Compare OLS, OLS + White, GLS with $\sigma_i = \sigma \times INCOME$

8.6 Have table for regression

$$EHAT\_SQ = \alpha_1 + \alpha_2 ROOMS + \alpha_3 ROOMS^2 + \alpha_4 CRIME + \alpha_5 CRIME^2 + \alpha_6 DIST + \nu$$

where $EHAT\_SQ$ is the square of regression residuals. (i.e. this is a WHITE Heteroskedasticity test style regression) a. Read the table, note what seems to be influencing variance b. Test for heteroskedasticity.

8.8 (data: `stockton96.dta`)

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

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

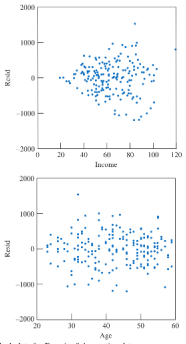


FIGURE 8.4 Residual plots for Exercise 8.4: vacation data.

Figure 1:

8.7 EXERCISES 323

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?

(b) Ordering the observations according to descending values of *INCOME*, and applying least squares to the first 100 observations, and again to the second 100 observations, yields the sums of squared errors

$$SSE_1 = 2.9471 \times 10^7 \quad SSE_2 = 1.0479 \times 10^7$$

Use the Goldfeld–Quandt test to test for heteroskedastic errors. Include specification of the null and alternative hypotheses.

(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 $\sigma_i^2 = \sigma^2 \times INCOME_i^2$.

(i) How do vacation miles traveled depend on income, age, and 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?

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

Figure 2:

Question Snips

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

Figure 3:

Table 8.4 **Estimated Variance Function for Exercise 8.6**

Dependent Variable: *EHAT_SQ*
Included observations: 506

Variable	Coefficient	Std. Error	*t*-value	*p*-value
C	1007.037	204.522	4.92	0.000
ROOMS	−305.311	63.088	−4.84	0.000
$ROOMS^2$	23.822	4.844	4.92	0.000
CRIME	2.285	1.242	1.84	0.067
$CRIME^2$	−0.039	0.019	−2.04	0.042
DIST	−4.419	2.466	−1.79	0.074

$R^2 = 0.08467$ $SSE = 5,038,458$ $SST = 5,504,525$

(b) Do you think heteroskedasticity is likely to be a problem?
(c) What misleading inferences are likely if the incorrect standard errors are used?

8.6 Continuing with the example in Exercise 8.5, Table 8.4 contains output for the following least squares regression

$$EHAT\_SQ = \alpha_1 + \alpha_2 ROOMS + \alpha_3 ROOMS^2 + \alpha_4 CRIME + \alpha_5 CRIME^2 + \alpha_6 DIST + v$$

where *EHAT_SQ* denotes the squares of the least squares residuals from the mean function estimated in Exercise 8.5.
(a) Discuss how each of the variables *ROOMS*, *CRIME*, and *DIST* influences the variance of house values.
(b) Test for heteroskedasticity.

Figure 4:

8.8 The file *stockton96.dat* contains 940 observations on home sales in Stockton, CA in 1996. They are a subset of the data in the file *stockton.dat* used for Exercise 7.4.
(a) Use least squares to estimate a linear equation that relates house price *PRICE* to the size of the house in square feet *SQFT* and the age of the house in years *AGE*. Comment on the estimates.
(b) Suppose that you own two houses. One has 1400 square feet; the other has 1800 square feet. Both are 20 years old. What price do you estimate you will get for each house.
(c) Use the White test (with cross-product term included) to test for heteroskedasticity.
(d) Estimate α_1 and α_2 in the variance function $\sigma_i^2 = \exp(\alpha_1 + \alpha_2 SQFT)$.
(e) Using the variance assumption from part (d), find generalized least squares estimates for the parameters of the equation estimated by least squares in part (a). Comment on the results.
(f) Use the results from part (e) to estimate the prices you will get for your two houses.

Figure 5: