Heteroskedasticity and its Solutions; Measurement Issues

Philip Leifeld

GV903: Advanced Research Methods, Week 7

University of Essex

1. OLS Estimation in Matrix Notation

Toy Example: Intercountry Life-Cycle Savings Data

sr	DV: People's life-time savings-to-income ratio.
pop15	Percentage of population less than 15 years old.
pop75	Percentage of the population over 75 years old.
dpi	Per capita disposable income.
ddpi	Percentage rate of change in disposable income.

Population equation:

$$sr = \beta_0 + \beta_1 pop15 + \beta_2 pop75 + \beta_3 dpi + \beta_4 ddpi + u$$

Estimation equation:

$$\widehat{sr} = \hat{\beta}_0 + \hat{\beta}_1 pop15 + \hat{\beta}_2 pop75 + \hat{\beta}_3 dpi + \hat{\beta}_4 ddpi$$

```
data("LifeCycleSavings")
fm1 <- lm(sr ~ pop15 + pop75 + dpi + ddpi, data = LifeCycleSavings)
```

```r
summary(fm1)
##
## Call:
## lm(formula = sr ~ pop15 + pop75 + dpi + ddpi, data = LifeCycleSavings)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.2422 -2.6857 -0.2488  2.4280  9.7509
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 28.5660865  7.3545161   3.884 0.000334 ***
## pop15       -0.4611931  0.1446422  -3.189 0.002603 **
## pop75       -1.6914977  1.0835989  -1.561 0.125530
## dpi         -0.0003369  0.0009311  -0.362 0.719173
## ddpi         0.4096949  0.1961971   2.088 0.042471 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.803 on 45 degrees of freedom
## Multiple R-squared:  0.3385,Adjusted R-squared:  0.2797
## F-statistic: 5.756 on 4 and 45 DF,  p-value: 0.0007904
```

Recap: Matrix Multiplication

Falk's scheme

			1	0	1	1
			0	0	1	1
			1	1	0	0
1	0	1	2	1	1	1
0	0	1	1	1	0	0
1	1	0	1	0	2	2
1	1	0	1	0	2	2

For each cell of the new matrix, calculate the dot product of the corresponding row of the first matrix and the column of the second matrix.

That is, compute the product of the first number in the row and the first number in the column, then the product of the second number in the row and the second number in the column etc, then add up the products.

Rewriting the Linear Model in Matrix Notation

$$
\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}_{n\times 1} = \begin{bmatrix} 1 & X_{11} & X_{21} & \cdots & X_{k1} \\ 1 & X_{12} & X_{21} & \cdots & X_{k2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{1n} & X_{2n} & \cdots & X_{kn} \end{bmatrix}_{n\times(k+1)} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}_{(k+1)\times 1} + \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}_{n\times 1}
$$

Note the column of ones in $\mathbf{X}$ for the intercept.

Note that the maximal index of $\boldsymbol{\beta}$ is k, not n.

In short:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$$

Creating the Data Structure in R

```r
Y <- LifeCycleSavings$sr
X <- cbind(1,
           LifeCycleSavings$pop15,
           LifeCycleSavings$pop75,
           LifeCycleSavings$dpi,
           LifeCycleSavings$ddpi)
head(X)
##      [,1]  [,2] [,3]    [,4] [,5]
## [1,]    1 29.35 2.87 2329.68 2.87
## [2,]    1 23.32 4.41 1507.99 3.93
## [3,]    1 23.80 4.43 2108.47 3.82
## [4,]    1 41.89 1.67  189.13 0.22
## [5,]    1 42.19 0.83  728.47 4.56
## [6,]    1 31.72 2.85 2982.88 2.43
```

$\mathbf{X}$ is also known as the *design matrix*.

Estimation of $\hat{\boldsymbol{\beta}}$

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}$$

Estimation of $\hat{\beta}$

$$\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}$$

Let's do this in R:

```
beta_hat <- solve(t(X) %*% X) %*% (t(X) %*% Y)
beta_hat
##                [,1]
## [1,] 28.5660865407
## [2,] -0.4611931471
## [3,] -1.6914976767
## [4,] -0.0003369019
## [5,]  0.4096949279
```

Estimation of $\hat{\beta}$

$$\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}$$

Let's do this in R:

```
beta_hat <- solve(t(X) %*% X) %*% (t(X) %*% Y)
beta_hat
##                [,1]
## [1,] 28.5660865407
## [2,] -0.4611931471
## [3,] -1.6914976767
## [4,] -0.0003369019
## [5,]  0.4096949279
```

And we can check if that was correct...

```
coef(fm1)
##   (Intercept)         pop15         pop75           dpi          ddpi
## 28.5660865407 -0.4611931471 -1.6914976767 -0.0003369019  0.4096949279
```

Fitted Values

Predict values by inserting values into the equation:

$$\widehat{sr} = \hat{\beta}_0 + \hat{\beta}_1 pop15 + \hat{\beta}_2 pop75 + \hat{\beta}_3 dpi + \hat{\beta}_4 ddpi$$

Fitted Values

Predict values by inserting values into the equation:

$$\widehat{sr} = \hat{\beta}_0 + \hat{\beta}_1 pop15 + \hat{\beta}_2 pop75 + \hat{\beta}_3 dpi + \hat{\beta}_4 ddpi$$

In R:

```r
pred1 <- X[, 1] * beta_hat[1] + X[, 2] * beta_hat[2] +
  X[, 3] * beta_hat[3] + X[, 4] * beta_hat[4] +
  X[, 5] * beta_hat[5]
```

Fitted Values

Predict values by inserting values into the equation:

$$\widehat{sr} = \hat{\beta}_0 + \hat{\beta}_1 pop15 + \hat{\beta}_2 pop75 + \hat{\beta}_3 dpi + \hat{\beta}_4 ddpi$$

In R:

```
pred1 <- X[, 1] * beta_hat[1] + X[, 2] * beta_hat[2] +
  X[, 3] * beta_hat[3] + X[, 4] * beta_hat[4] +
  X[, 5] * beta_hat[5]
```

Or we can do this with matrix algebra:

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta}$$

Fitted Values

Predict values by inserting values into the equation:

$$\widehat{sr} = \hat{\beta}_0 + \hat{\beta}_1 pop15 + \hat{\beta}_2 pop75 + \hat{\beta}_3 dpi + \hat{\beta}_4 ddpi$$

In R:

```
pred1 <- X[, 1] * beta_hat[1] + X[, 2] * beta_hat[2] +
  X[, 3] * beta_hat[3] + X[, 4] * beta_hat[4] +
  X[, 5] * beta_hat[5]
```

Or we can do this with matrix algebra:

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta}$$

```
pred2 <- X %*% beta_hat
```

Fitted Values

Predict values by inserting values into the equation:

$$\widehat{sr} = \hat{\beta}_0 + \hat{\beta}_1 pop15 + \hat{\beta}_2 pop75 + \hat{\beta}_3 dpi + \hat{\beta}_4 ddpi$$

In R:

```
pred1 <- X[, 1] * beta_hat[1] + X[, 2] * beta_hat[2] +
  X[, 3] * beta_hat[3] + X[, 4] * beta_hat[4] +
  X[, 5] * beta_hat[5]
```

Or we can do this with matrix algebra:

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta}$$

```
pred2 <- X %*% beta_hat
```

Note how this resembles the linear model equation, just without error term and based on estimated coefficients rather than the population parameters.

Fitted Values

Or we can use the `predict` function in `R`:

```
pred3 <- predict(fm1)
```

Fitted Values

Or we can use the `predict` function in `R`:

```
pred3 <- predict(fm1)
```

Let's compare the results:

```
head(data.frame(pred1, pred2, pred3))
##                pred1     pred2     pred3
## Australia 10.566420 10.566420 10.566420
## Austria   11.453614 11.453614 11.453614
## Belgium   10.951042 10.951042 10.951042
## Bolivia    6.448319  6.448319  6.448319
## Brazil     9.327191  9.327191  9.327191
## Canada     9.106892  9.106892  9.106892
```

Residuals

Residuals are the differences between observed and fitted values:

$$\hat{\mathbf{u}} = \hat{\mathbf{Y}} - \mathbf{Y}$$

Residuals

Residuals are the differences between observed and fitted values:

$$\hat{\mathbf{u}} = \hat{\mathbf{Y}} - \mathbf{Y} = \mathbf{X}\hat{\beta} - \mathbf{Y}$$

Residuals

Residuals are the differences between observed and fitted values:

$$\hat{\mathbf{u}} = \hat{\mathbf{Y}} - \mathbf{Y} = \mathbf{X}\hat{\beta} - \mathbf{Y}$$

Homoskedasticity posits that all errors have the same variance. We call this error variance σ^2 or mean squared error (MSE).

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \hat{u}_i^2}{n-k-1} = \frac{\text{SSR}}{df} = \frac{\hat{\mathbf{u}}^\top \hat{\mathbf{u}}}{n-k-1}$$

Residuals

Residuals are the differences between observed and fitted values:

$$\hat{\mathbf{u}} = \hat{\mathbf{Y}} - \mathbf{Y} = \mathbf{X}\hat{\beta} - \mathbf{Y}$$

Homoskedasticity posits that all errors have the same variance. We call this error variance σ^2 or mean squared error (MSE).

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \hat{u}_i^2}{n-k-1} = \frac{\text{SSR}}{df} = \frac{\hat{\mathbf{u}}^\top \hat{\mathbf{u}}}{n-k-1}$$

```
u <- pred1 - Y
sig2 <- (t(u) %*% u) / (nrow(X) - ncol(X))
sig2
##          [,1]
## [1,] 14.46029

sqrt(sig2) # Residual standard error (see summary(fm1) output)
##          [,1]
## [1,] 3.802669
```

Variance–Covariance Matrix

The covariance matrix shows the covariances between the different $\hat{\beta}$ estimates (and with themselves, i. e., the variances, on the diagonal):

$$\widehat{\mathrm{Var}}(\hat{\beta}) = \hat{\sigma}^2(\mathbf{X}^\top\mathbf{X})^{-1}$$

```
vcovmat <- sig2[1, 1] * solve(t(X) %*% X)
round(vcovmat, digits = 7)
##            [,1]       [,2]       [,3]       [,4]       [,5]
## [1,] 54.0889072 -1.0469276 -6.4480865 -0.0011359 -0.2716546
## [2,] -1.0469276  0.0209214  0.1199574  0.0000242  0.0029078
## [3,] -6.4480865  0.1199574  1.1741866 -0.0003703 -0.0116339
## [4,] -0.0011359  0.0000242 -0.0003703  0.0000009  0.0000467
## [5,] -0.2716546  0.0029078 -0.0116339  0.0000467  0.0384933
```

Standard Errors of Estimates

We can now extract the standard errors for the $\hat{\beta}$ estimates from the matrix:

```
sqrt(diag(vcovmat))
## [1] 7.3545161062 0.1446422248 1.0835989307 0.0009311072 0.1961971276
```

Standard Errors of Estimates

We can now extract the standard errors for the $\hat{\beta}$ estimates from the matrix:

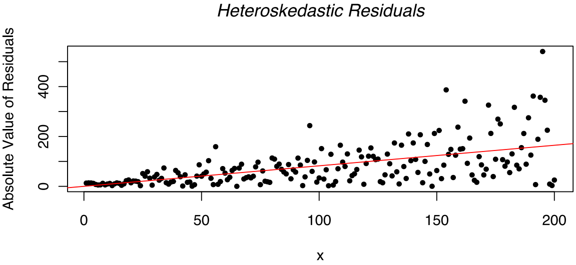
```
sqrt(diag(vcovmat))
## [1] 7.3545161062 0.1446422248 1.0835989307 0.0009311072 0.1961971276
```

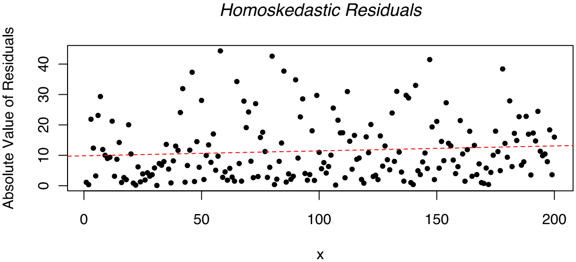
Let's see if this corresponds to the values from `summary(fm1)`:

```
round(summary(fm1)$coefficients, digits = 8)
##                Estimate Std. Error    t value   Pr(>|t|)
## (Intercept) 28.5660865 7.35451611  3.8841558 0.00033382
## pop15       -0.4611931 0.14464222 -3.1885098 0.00260302
## pop75       -1.6914977 1.08359893 -1.5609998 0.12552979
## dpi         -0.0003369 0.00093111 -0.3618293 0.71917316
## ddpi         0.4096949 0.19619713  2.0881801 0.04247114
```

2. Robust Standard Errors

The Problem: Heteroskedasticity





The Homoskedasticity Assumption

Remember the computation of the VCOV matrix (for coefficient SEs):

$$\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2(\mathbf{X}^\top\mathbf{X})^{-1}$$

Note how a constant residual variance σ^2 is assumed here.

The Homoskedasticity Assumption

Remember the computation of the VCOV matrix (for coefficient SEs):

$$\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2(\mathbf{X}^\top\mathbf{X})^{-1}$$

Note how a constant residual variance σ^2 is assumed here.

But this is sometimes unrealistic. We can fix this by replacing VCOV with a version based on varying variances instead of σ^2.

The Homoskedasticity Assumption

First, consider that the VCOV can be rewritten as follows:

$$
\begin{aligned}
\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}) &= \hat{\sigma}^2(\mathbf{X}^\top\mathbf{X})^{-1} \\
&= \hat{\sigma}^2(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1} \\
&= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\sigma^2\mathbf{I}\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1} \\
&= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top E(\hat{\mathbf{u}}\hat{\mathbf{u}}^\top|\mathbf{X})\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}
\end{aligned}
$$

where

$$
E(\mathbf{u}\mathbf{u}^\top|\mathbf{X}) = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix} = \sigma^2\mathbf{I}
$$

The Homoskedasticity Assumption

First, consider that the VCOV can be rewritten as follows:

$$
\begin{aligned}
\widehat{\text{Var}}(\hat{\boldsymbol{\beta}}) &= \hat{\sigma}^2(\mathbf{X}^\top\mathbf{X})^{-1} \\
&= \hat{\sigma}^2(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1} \\
&= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\sigma^2\mathbf{I}\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1} \\
&= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top E(\hat{\mathbf{u}}\hat{\mathbf{u}}^\top|\mathbf{X})\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}
\end{aligned}
$$

where

$$
E(\mathbf{u}\mathbf{u}^\top|\mathbf{X}) = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix} = \sigma^2\mathbf{I}
$$

$\mathbf{I}$ is an $n \times n$ identity matrix with ones on the diagonal.

The Homoskedasticity Assumption

First, consider that the VCOV can be rewritten as follows:

$$
\begin{aligned}
\widehat{\text{Var}}(\hat{\boldsymbol{\beta}}) &= \hat{\sigma}^2(\mathbf{X}^\top\mathbf{X})^{-1} \\
&= \hat{\sigma}^2(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1} \\
&= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\sigma^2\mathbf{I}\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1} \\
&= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top E(\hat{\mathbf{u}}\hat{\mathbf{u}}^\top|\mathbf{X})\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}
\end{aligned}
$$

where

$$
E(\mathbf{u}\mathbf{u}^\top|\mathbf{X}) = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix} = \sigma^2\mathbf{I}
$$

$\sigma^2\mathbf{I}$ thus produces the matrix you see above.

The Homoskedasticity Assumption

First, consider that the VCOV can be rewritten as follows:

$$
\begin{aligned}
\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}) &= \hat{\sigma}^2(\mathbf{X}^\top\mathbf{X})^{-1} \\
&= \hat{\sigma}^2(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1} \\
&= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\sigma^2\mathbf{I}\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1} \\
&= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top E(\hat{\mathbf{u}}\hat{\mathbf{u}}^\top|\mathbf{X})\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}
\end{aligned}
$$

where

$$
E(\mathbf{u}\mathbf{u}^\top|\mathbf{X}) = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix} = \sigma^2\mathbf{I}
$$

$E(\hat{\mathbf{u}}\hat{\mathbf{u}}^\top|\mathbf{X})$ is the expectation of $\mathbf{u}\mathbf{u}^\top$.

Heteroskedasticity-Consistent Standard Errors

We can now replace the constant error variances by varying error variances:

$$E(\mathbf{u}\mathbf{u}^\top|\mathbf{X}) = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix}$$

Heteroskedasticity-Consistent Standard Errors

We can now replace the constant error variances by varying error variances:

$$E(\mathbf{u}\mathbf{u}^\top|\mathbf{X}) = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix}$$

How do we design them?

Heteroskedasticity-Consistent Standard Errors

We can now replace the constant error variances by varying error variances:

$$E(\mathbf{u}\mathbf{u}^\top|\mathbf{X}) = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix}$$

How do we design them?

White (1980) shows that $\textbf{diag}\left[\hat{u}_i^2\right]$ is a consistent (but biased) estimator of $E(\mathbf{u}\mathbf{u}^\top|\mathbf{X})$:

$$\textbf{diag}\left[\hat{u}_i^2\right] = \begin{bmatrix} \hat{u}_1^2 & 0 & \cdots & 0 \\ 0 & \hat{u}_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \hat{u}_n^2 \end{bmatrix}$$

Heteroskedasticity-Consistent Standard Errors

Hence, for moderate to large samples, we can estimate VCOV as:

$$\widehat{\text{Var}}(\hat{\beta}) = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \text{diag}\left[\hat{u}_i^2\right] \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1}$$

We can also add a finite sample correction to make this not only consistent, but also unbiased:

$$\widehat{\text{Var}}(\hat{\beta}) = \frac{n}{df} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \text{diag}\left[\hat{u}_i^2\right] \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1}$$

The square root of the diagonal can be used to construct heteroskedasticity-consistent standard errors.

```
vcov_rob <- nrow(X) / (nrow(X) - ncol(X)) *
  solve(t(X) %*% X) %*% t(X) %*% diag(u^2) %*%
  X %*% solve(t(X) %*% X)
```

Default vs Robust Standard Errors

```r
library("texreg")
texreg(list(fm1, fm1), table = FALSE, dcolumn = TRUE, booktabs = TRUE,
       use.packages = FALSE, single.row = TRUE,
       override.se = list(NA, sqrt(diag(vcov_rob))))
```

	Model 1	Model 2
(Intercept)	$28.57\ (7.35)^{***}$	$28.57\ (6.72)^{***}$
pop15	$-0.46\ (0.14)^{**}$	$-0.46\ (0.13)^{**}$
pop75	$-1.69\ (1.08)$	$-1.69\ (1.07)$
dpi	$-0.00\ (0.00)$	$-0.00\ (0.00)$
ddpi	$0.41\ (0.20)^{*}$	$0.41\ (0.18)^{*}$
R^2	0.34	0.34
Adj. R^2	0.28	0.28
Num. obs.	50	50

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

We can also use the sandwich and lmtest packages...

```
library("sandwich")
library("lmtest")
hc <- vcovHC(fm1, type = "HC1") # without sample correction: "HC0"
coeftest(fm1, vcov = hc)        # hypothesis tests with new VCOV
##
## t test of coefficients:
##
##                Estimate  Std. Error t value  Pr(>|t|)
## (Intercept) 28.56608654  6.72441758  4.2481 0.0001069 ***
## pop15       -0.46119315  0.13272517 -3.4748 0.0011430 **
## pop75       -1.69149768  1.06956732 -1.5815 0.1207727
## dpi         -0.00033690  0.00055143 -0.6110 0.5442966
## ddpi         0.40969493  0.17953130  2.2820 0.0272679 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

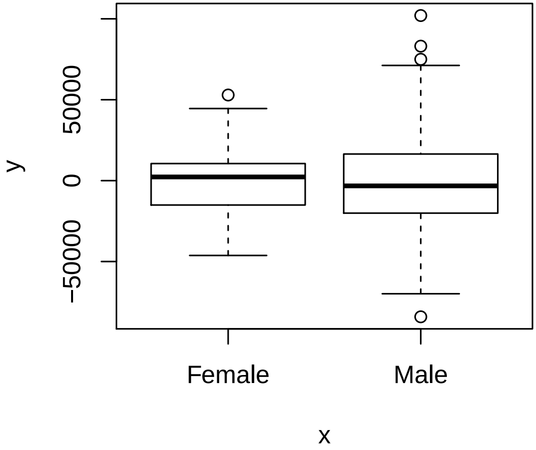
Another Example: Salaries and Years since PhD

```r
library("carData")
data("Salaries")
m <- lm(salary ~ yrs.since.phd + sex, data = Salaries)
```

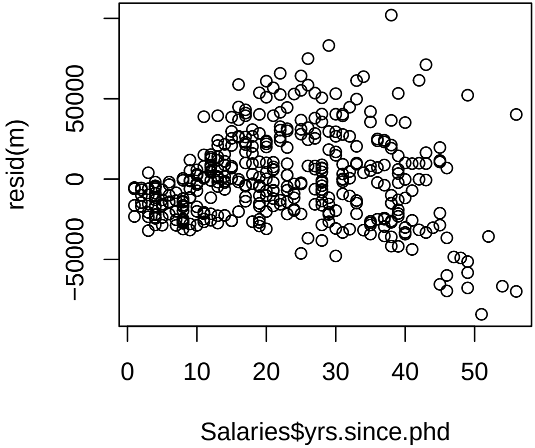
```r
coeftest(m)
##
## t test of coefficients:
##
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    85181.82    4748.32 17.9394  < 2e-16 ***
## yrs.since.phd    958.08     108.32  8.8450  < 2e-16 ***
## sexMale         7923.62    4684.08  1.6916  0.09151 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

coeftest(m, vcov = vcovHC(m, type = "HC1"))
##
## t test of coefficients:
##
##                Estimate Std. Error t value  Pr(>|t|)
## (Intercept)    85181.82    3886.71 21.9162 < 2.2e-16 ***
## yrs.since.phd    958.08     127.06  7.5402 3.265e-13 ***
## sexMale         7923.62    3689.38  2.1477   0.03235 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Yep, that looks like Heteroskedasticity. . .

```
plot(Salaries$sex, resid(m))
```

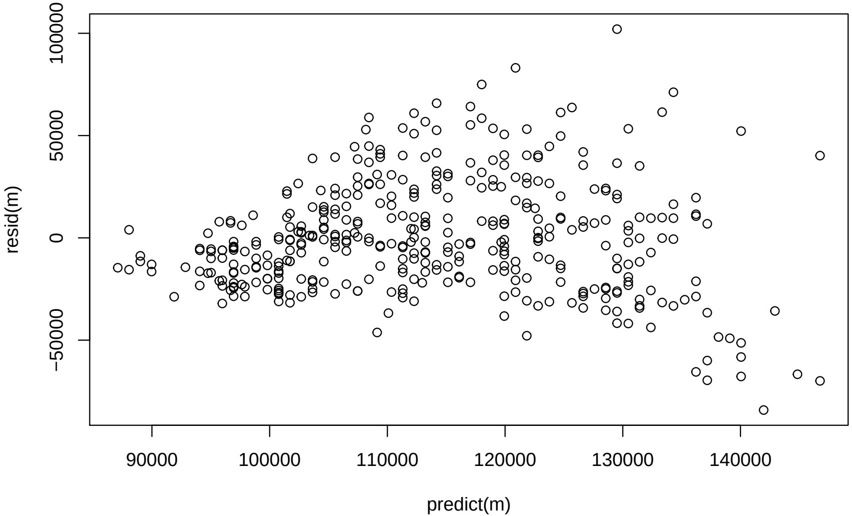



```
plot(Salaries$yrs.since.phd, resid(m))
```



Residual–vs–Fitted Plot

```r
plot(predict(m), resid(m))
```



Testing for Heteroskedasticity

Two tests that assess directly whether the residuals are correlated with the IVs:

Testing for Heteroskedasticity

Two tests that assess directly whether the residuals are correlated with the IVs:

Breusch-Pagan Test

1. Estimate model by OLS.
2. Re-estimate by OLS with squared residuals as DV.
3. Heteroskedasticity is present if F test is significant.

Testing for Heteroskedasticity

Two tests that assess directly whether the residuals are correlated with the IVs:

Breusch-Pagan Test

1. Estimate model by OLS.
2. Re-estimate by OLS with squared residuals as DV.
3. Heteroskedasticity is present if F test is significant.

White Test (Special Case)

1. Estimate model by OLS.
2. Estimate OLS with squared residuals as DV and fitted values + squared fitted values (= quadratic form) as IVs.
3. Heteroskedasticity is present if F test is significant.

```r
summary(lm(resid(m)^2 ~ Salaries$yrs.since.phd + Salaries$sex), digits = 2)
##
## Call:
## lm(formula = resid(m)^2 ~ Salaries$yrs.since.phd + Salaries$sex)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -1.500e+09 -6.256e+08 -2.069e+08  2.058e+08  9.129e+09
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -100515936  192367813  -0.523    0.602
## Salaries$yrs.since.phd  33209801    4388300   7.568 2.71e-13 ***
## Salaries$sexMale       120080705  189765505   0.633    0.527
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.113e+09 on 394 degrees of freedom
## Multiple R-squared:  0.133,Adjusted R-squared:  0.1286
## F-statistic: 30.22 on 2 and 394 DF,  p-value: 6.177e-13
```

Breusch–Pagan Test: Also Available in `lmtest` Package

```
bptest(m)
##
##  studentized Breusch-Pagan test
##
## data:  m
## BP = 52.797, df = 2, p-value = 3.43e-12
```

The Breusch–Pagan test has as many parameters as the original model. The White test has just three parameters no matter how many IVs.

Special Case of the White Test for Heteroskedasticity

```
pred <- predict(m)
pred_squared <- pred^2
summary(lm(resid(m)^2 ~ pred + pred_squared))
##
## Call:
## lm(formula = resid(m)^2 ~ pred + pred_squared)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -1.770e+09 -5.284e+08 -2.576e+08  2.300e+08  9.102e+09
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   8.890e+09  4.073e+09   2.183  0.02965 *
## pred         -1.765e+05  7.117e+04  -2.480  0.01357 *
## pred_squared  9.106e-01  3.081e-01   2.956  0.00331 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.102e+09 on 394 degrees of freedom
## Multiple R-squared:  0.1505,Adjusted R-squared:  0.1462
## F-statistic:  34.9 on 2 and 394 DF,  p-value: 1.112e-14
```

Concluding Thoughts about Robust Standard Errors

- Don't forget to also re-compute *p*-values etc.
- SEs are usually larger but can be smaller than original.
- Corrects *only* heteroskedasticity, no other problems!
- Computing these robust SEs usually does not do any harm.
- Make sure you use the sample correction ("HC1") with small samples.

3. WLS, GLS, and FGLS

From OLS to WLS

In OLS, we minimise the residual sum of squares (RSS or SSR):

$$\text{RSS} = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{i1} - \ldots - \hat{\beta}_k X_{ik})^2$$

From OLS to WLS

In OLS, we minimise the residual sum of squares (RSS or SSR):

$$\text{RSS} = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{i1} - \ldots - \hat{\beta}_k X_{ik})^2$$

In WLS, we minimise the weighted RSS:

$$\text{RSS} = \sum_{i=1}^{n} w_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{i1} - \ldots - \hat{\beta}_k X_{ik})^2$$

From OLS to WLS

In OLS, we minimise the residual sum of squares (RSS or SSR):

$$\text{RSS} = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{i1} - \ldots - \hat{\beta}_k X_{ik})^2$$

In WLS, we minimise the weighted RSS:

$$\text{RSS} = \sum_{i=1}^{n} w_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{i1} - \ldots - \hat{\beta}_k X_{ik})^2$$

OLS estimation:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}$$

From OLS to WLS

In OLS, we minimise the residual sum of squares (RSS or SSR):

$$\text{RSS} = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{i1} - \ldots - \hat{\beta}_k X_{ik})^2$$

In WLS, we minimise the weighted RSS:

$$\text{RSS} = \sum_{i=1}^{n} w_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{i1} - \ldots - \hat{\beta}_k X_{ik})^2$$

OLS estimation:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}$$

WLS estimation:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \text{diag}\,[\mathbf{W}]\,\mathbf{X})^{-1} \mathbf{X}^\top \text{diag}\,[\mathbf{W}]\,\mathbf{Y}$$

Weighted Least Squares: Introduction

- The WLS RSS equation shows how WLS is really just OLS where we standardise each model term by some weight (for each observation).

Weighted Least Squares: Introduction

- The WLS RSS equation shows how WLS is really just OLS where we standardise each model term by some weight (for each observation).
- So we can just transform each variable with the weight attached and use OLS for estimation:

$$WY = W\beta_0 + W\beta_1 X_1 + \ldots W\beta_k X_k + u$$

Weighted Least Squares: Introduction

- The WLS RSS equation shows how WLS is really just OLS where we standardise each model term by some weight (for each observation).
- So we can just transform each variable with the weight attached and use OLS for estimation:

$$WY = W\beta_0 + W\beta_1 X_1 + \dots W\beta_k X_k + u$$

- Then the SEs are also consistent and don't need any tweaks.

Weighted Least Squares: Introduction

- The WLS RSS equation shows how WLS is really just OLS where we standardise each model term by some weight (for each observation).
- So we can just transform each variable with the weight attached and use OLS for estimation:

$$WY = W\beta_0 + W\beta_1 X_1 + \ldots W\beta_k X_k + u$$

- Then the SEs are also consistent and don't need any tweaks.
- We can use WLS to model heteroskedastic data because data points with a smaller error variance are more informative and should receive a higher weight.

Weighted Least Squares: Introduction

- The WLS RSS equation shows how WLS is really just OLS where we standardise each model term by some weight (for each observation).
- So we can just transform each variable with the weight attached and use OLS for estimation:

$$WY = W\beta_0 + W\beta_1 X_1 + \dots W\beta_k X_k + u$$

- Then the SEs are also consistent and don't need any tweaks.
- We can use WLS to model heteroskedastic data because data points with a smaller error variance are more informative and should receive a higher weight.
- The question is: How do we translate those error variances into $\mathbf{W}$?

Weighted Least Squares: Specification of Weights

- We need a function that describes the “importance” of observations.

Weighted Least Squares: Specification of Weights

- We need a function that describes the "importance" of observations.
- For example, this could be the inverse of a variable that causes heteroskedasticity.

Weighted Least Squares: Specification of Weights

- We need a function that describes the “importance” of observations.
- For example, this could be the inverse of a variable that causes heteroskedasticity.
- Or this could be a separate variable that directly operationalises importance.

Weighted Least Squares: Specification of Weights

- We need a function that describes the "importance" of observations.
- For example, this could be the inverse of a variable that causes heteroskedasticity.
- Or this could be a separate variable that directly operationalises importance.
- Example: survey question on how much confidence the respondent has in his/her answer.

Weighted Least Squares: Specification of Weights

- We need a function that describes the “importance” of observations.
- For example, this could be the inverse of a variable that causes heteroskedasticity.
- Or this could be a separate variable that directly operationalises importance.
- Example: survey question on how much confidence the respondent has in his/her answer.
- This is a generalisation of OLS to weighted observations. OLS is a special case of WLS where all weights are equal.

Weighted Least Squares: Specification of Weights

- We need a function that describes the “importance” of observations.
- For example, this could be the inverse of a variable that causes heteroskedasticity.
- Or this could be a separate variable that directly operationalises importance.
- Example: survey question on how much confidence the respondent has in his/her answer.
- This is a generalisation of OLS to weighted observations. OLS is a special case of WLS where all weights are equal.
- One good estimate of importance could be the inverse of the residuals from the original OLS estimation. This is called *Feasible Least Squares* (FGLS). This is an example of a two-stage least squares (2SLS) estimator.

Weighted Least Squares: Specification of Weights

- We need a function that describes the "importance" of observations.
- For example, this could be the inverse of a variable that causes heteroskedasticity.
- Or this could be a separate variable that directly operationalises importance.
- Example: survey question on how much confidence the respondent has in his/her answer.
- This is a generalisation of OLS to weighted observations. OLS is a special case of WLS where all weights are equal.
- One good estimate of importance could be the inverse of the residuals from the original OLS estimation. This is called *Feasible Least Squares* (FGLS). This is an example of a two-stage least squares (2SLS) estimator.
- WLS: off-diagonal values of covariance matrix are zero. Generalised Least Squares (GLS): we can arbitrarily specify correlations between observed variances.

WLS Example

```
wls <- lm(salary ~ yrs.since.phd + sex, weights = 1 / yrs.since.phd, data = Salaries)
summary(wls)
##
## Call:
## lm(formula = salary ~ yrs.since.phd + sex, data = Salaries, weights = 1/yrs.since.phd)
##
## Weighted Residuals:
##      Min       1Q   Median       3Q      Max
## -13925.5  -4262.8    -77.1   4031.0  15583.9
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    74137.8     2960.3  25.044   <2e-16 ***
## yrs.since.phd   1494.3       87.8  17.018   <2e-16 ***
## sexMale         6902.2     3029.6   2.278   0.0232 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5730 on 394 degrees of freedom
## Multiple R-squared:  0.4332,Adjusted R-squared:  0.4303
## F-statistic: 150.6 on 2 and 394 DF,  p-value: < 2.2e-16
```

FGLS Example

Note how the functional form here is wrong because we assume that `sex` does not contribute to heteroskedasticity.

FGLS Example

Note how the functional form here is wrong because we assume that `sex` does not contribute to heteroskedasticity.

Let's do FGLS then:

1. Estimate original OLS and retain residuals.
2. Regress $\log(\hat{u}^2)$ on the original predictors and retain exponentiated predicted values.
3. Re-estimate original OLS/WLS and use the predicted values as weights.

```r
m <- lm(salary ~ yrs.since.phd + sex, data = Salaries)
u <- resid(m)
dv <- log(u^2)
fgls1 <- lm(dv ~ yrs.since.phd + sex, data = Salaries)
h_hat <- exp(predict(fgls1))
fgls2 <- lm(salary ~ yrs.since.phd + sex, weight = h_hat, data = Salaries)
```

```r
summary(fgls2)
##
## Call:
## lm(formula = salary ~ yrs.since.phd + sex, data = Salaries, weights = h_hat)
##
## Weighted Residuals:
##        Min         1Q     Median         3Q        Max
## -1.733e+09 -2.815e+08 -8.482e+07  2.302e+08  2.009e+09
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    98095.9     6402.4  15.322  < 2e-16 ***
## yrs.since.phd    452.6      122.2   3.703 0.000244 ***
## sexMale         8097.5     6214.4   1.303 0.193329
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 452600000 on 394 degrees of freedom
## Multiple R-squared:  0.04324,Adjusted R-squared:  0.03838
## F-statistic: 8.903 on 2 and 394 DF,  p-value: 0.0001653
```

Concluding Remarks on Heteroskedasticity

- Robust SEs are often considered more modern and have largely replaced WLS etc.
- However, in the example, OLS and FGLS yielded different coefficients and substantive results. This may point to a misspecified model. None of these methods help if other assumptions are violated!
- WLS requires knowledge of the functional form of heteroskedasticity, which is often hard to know.