Gov 2000 - 10. Troubleshooting the Linear Model (II)

Matthew Blackwell
Harvard University
mblackwell@gov.harvard.edu

Where are we? Where are we going?

- Last week: finding and correcting violations of linearity and non-Normal errors
- This week: detecting and correcting violations of homoskedasticity
- Next week: panel data, fixed effects, maybe differences-in-differences

Review of the OLS assumptions

1. Linearity: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$
2. Random/iid sample: $(y_i, \mathbf{x}'_i)$ are a iid sample from the population.
3. No perfect collinearity: $\mathbf{X}$ is an $n \times (K+1)$ matrix with rank $K+1$
4. Zero conditional mean: $\mathbb{E}[\mathbf{u}|\mathbf{X}] = \mathbf{0}$
5. Homoskedasticity: $\text{var}(\mathbf{u}|\mathbf{X}) = \sigma^2_u \mathbf{I}_n$
6. Normality: $\mathbf{u}|\mathbf{X} \sim N(\mathbf{0}, \sigma^2_u \mathbf{I}_n)$

- 1-4 give us unbiasedness/consistency
- 1-5 are the Gauss-Markov, allow for large-sample inference
- 1-6 allow for small-sample inference

HETEROSKEDASTICITY

Review of Homoskedasticity

- Rewrite our estimator:

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} \\ &= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta}+\mathbf{u}) \\ &= \boldsymbol{\beta} + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{u}\end{aligned}$$

- Take the variance:

$$\begin{aligned}V[\hat{\boldsymbol{\beta}}|\mathbf{X}] &= V[\boldsymbol{\beta}|\mathbf{X}] + V[\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{u}|\mathbf{X}] \\ &= V[\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{u}|\mathbf{X}] \\ &= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'V[\mathbf{u}|\mathbf{X}](\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}')' \\ &= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'V[\mathbf{u}|\mathbf{X}]\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1} \\ &= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\sigma^2\mathbf{I}\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1} \text{ (by homoskedasticity)} \\ &= \sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}\end{aligned}$$

- where we used the fact that $\text{Var}[Au] = A\text{Var}[u]A'$
- Plug in $\widehat{\sigma}^2$ and you're all set.

Non-constant error variance

$$V[\mathbf{u}|\mathbf{X}] = \sigma^2\mathbf{I} = \begin{bmatrix} \sigma^2 & 0 & 0 & \ldots & 0 \\ 0 & \sigma^2 & 0 & \ldots & 0 \\ & & & \vdots & \\ 0 & 0 & 0 & \ldots & \sigma^2 \end{bmatrix}$$

$$V[\mathbf{u}|\mathbf{X}] = \begin{bmatrix} \sigma_1^2 & 0 & 0 & \ldots & 0 \\ 0 & \sigma_2^2 & 0 & \ldots & 0 \\ & & & \vdots & \\ 0 & 0 & 0 & \ldots & \sigma_n^2 \end{bmatrix}$$

- $\text{Cov}(u_i, u_j|\mathbf{X}) = 0$
- $\text{Var}(u_i|\mathbf{X}) = \sigma_i^2$

Consequences of Heteroskedasticity

- Standard $\widehat{\sigma}^2$ is biased and inconsistent for σ^2
- Standard error estimates incorrect:

$$\widehat{SE}[\widehat{\beta}_1] = \frac{\widehat{\sigma}^2}{\sum_i (X_i - \overline{X})^2}$$

- Test statistics won't have t or F distributions
- α-level tests, the probability of Type I error $\neq \alpha$
- Coverage of $1 - \alpha$ CIs $\neq 1 - \alpha$
- OLS is not BLUE
- $\widehat{\boldsymbol{\beta}}$ still unbiased and consistent for $\boldsymbol{\beta}$

Visual diagnostics

1. Plot of residuals versus fitted values
   - In R, `plot(mod, which = 1)`
2. Spread location plots
   - y-axis: Square-root of the absolute value of the residuals
   - x-axis: Fitted values
   - Usually has loess trend curve
   - In R, `plot(mod, which = 1)`

Exmaple: Buchanan votes

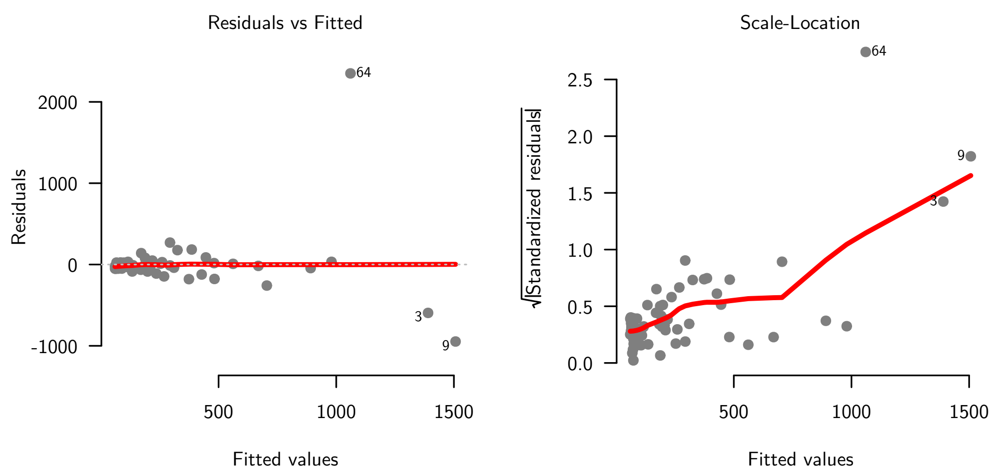
```
flvote <- foreign::read.dta("flbuchan.dta")
mod <- lm(edaybuchanan ~ edaytotal, data = flvote)
summary(mod)
```

```
##
## Call:
## lm(formula = edaybuchanan ~ edaytotal, data = flvote)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -947.05  -41.74  -19.47   20.20 2350.54
##
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5.423e+01  4.914e+01   1.104    0.274
## edaytotal   2.323e-03  3.104e-04   7.483 2.42e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 332.7 on 65 degrees of freedom
## Multiple R-squared:  0.4628, Adjusted R-squared:  0.4545
## F-statistic:    56 on 1 and 65 DF,  p-value: 2.417e-10
```

Diagnostics

```r
par(mfrow = c(1, 2))
plot(mod, which = 1, las = 1, bty = "n", pch = 19, col = "grey50", lwd = 3)
plot(mod, which = 3, las = 1, bty = "n", pch = 19, col = "grey50", lwd = 3)
```



Formal tests

- Plots are usually sufficient, but can use formal hypothesis test for heteroskedasticity
- $H_0 : \text{Var}[u_i|\mathbf{X}] = \sigma^2$
- Under zero conditional mean, this is equivalent to $H_0 : \mathbb{E}[u_i^2|\mathbf{X}] = \mathbb{E}[u_i^2] = \sigma^2$
- Under the null, the squared residuals should be unrelated to the independent variables
- **Breush-Pagan test:**

1. Regression y_i on $\mathbf{x}'_i$ and store residuals, $\widehat{u}_i$
2. Regress $\widehat{u}_i^2$ on $\mathbf{x}'_i$
3. Run F-test against null that all slope coefficients are 0

- In R, `bptest()` in the `lmtest` package

Breush-Pagan example

```
library(lmtest)
bptest(mod)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  mod
## BP = 12.59, df = 1, p-value = 0.0003878
```

Dealing with non-constant error variance

1. Transform the dependent variable
2. Model the heteroskedasticity using Weighted Least Squares (WLS)
3. Use an estimator of Var$[\widehat{\beta}]$ that is **robust to heteroskedasticity**
4. Admit we have the wrong model and use a different approach

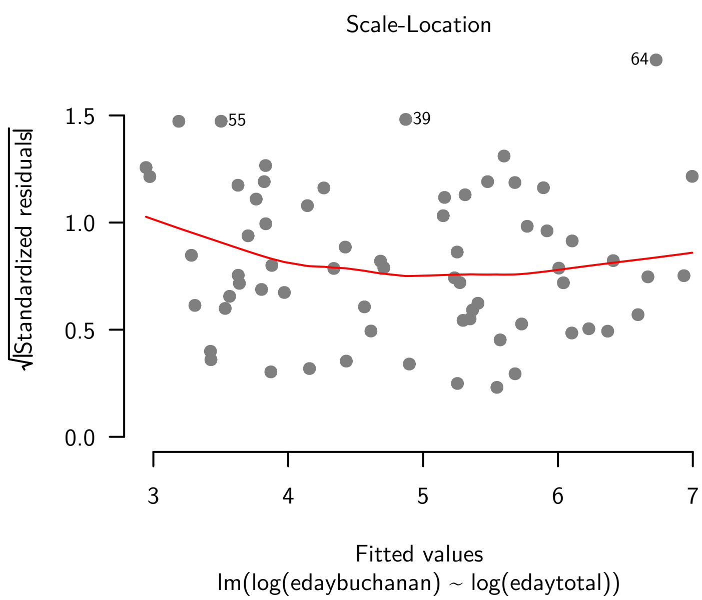
Example: Transforming Buchanan votes

```
mod2 <- lm(log(edaybuchanan) ~ log(edaytotal), data = flvote)
summary(mod2)
```

```
##
## Call:
## lm(formula = log(edaybuchanan) ~ log(edaytotal), data = flvote)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.02109 -0.25903  0.02473  0.29146  1.40713
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)    -2.72789    0.39956  -6.827  3.5e-09 ***
## log(edaytotal)  0.72853    0.03803  19.154  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4688 on 65 degrees of freedom
## Multiple R-squared:  0.8495, Adjusted R-squared:  0.8472
## F-statistic: 366.9 on 1 and 65 DF,  p-value: < 2.2e-16
```

```
plot(mod2, which = 3, las = 1, bty = "n", pch = 19, col = "grey50", lwd = 1)
```




```
library(lmtest)
bptest(mod, studentize = FALSE)
```

```
##
##  Breusch-Pagan test
##
## data:  mod
## BP = 250.07, df = 1, p-value < 2.2e-16
```

```
bptest(mod2, studentize = FALSE)
```

```
##
##  Breusch-Pagan test
##
## data:  mod2
## BP = 0.01105, df = 1, p-value = 0.9163
```

Weighted least squares

- Suppose that the heteroskedasticity is known up to a multiplicative constant:

$$\text{Var}[u_i|\mathbf{X}] = a_i\sigma^2$$

  where $a_i = a_i(\mathbf{x}_i')$ is a positive and known function of $\mathbf{x}_i'$

- WLS: multiply y_i by $1/\sqrt{a_i}$:

$$y_i/\sqrt{a_i} = \beta_0/\sqrt{a_i} + \beta_1 x_{i1}/\sqrt{a_i} + \cdots + \beta_k x_{ik}/\sqrt{a_i} + u_i/\sqrt{a_i}$$

- In matrix notation:

WLS intuition

- Rescales errors to $u_i/\sqrt{a_i}$, which maintains zero mean error
- But makes the error variance constant again:

$$\begin{aligned} \text{Var}\left[\frac{1}{\sqrt{a_i}}u_i|\mathbf{X}\right] &= \frac{1}{a_i}\text{Var}\left[u_i|\mathbf{X}\right] \\ &= \frac{1}{a_i}a_i\sigma^2 \\ &= \sigma^2 \end{aligned}$$

- If you know a_i, then you can use this approach to makes the model homoeskedastic and, thus, BLUE again
- When do we know a_i?

WLS procedure

- Define the weighting matrix:

$$\mathbf{W} = \begin{bmatrix} 1/\sqrt{a_1} & 0 & 0 & 0 \\ 0 & 1/\sqrt{a_2} & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 1/\sqrt{a_n} \end{bmatrix}$$

- Run the following regression:

$$\begin{aligned}\mathbf{W}\mathbf{y} &= \mathbf{W}\mathbf{X}\boldsymbol{\beta} + \mathbf{W}\mathbf{u} \\ \mathbf{y}^* &= \mathbf{X}^*\boldsymbol{\beta} + \mathbf{u}^*\end{aligned}$$

- Run regression of $\mathbf{y}^* = \mathbf{W}\mathbf{y}$ on $\mathbf{X}^* = \mathbf{W}\mathbf{X}$ and all Gauss-Markov assumptions are satisfied
- Plugging into the usual formula for $\widehat{\boldsymbol{\beta}}$:

$$\widehat{\boldsymbol{\beta}}_W = (\mathbf{X}'\mathbf{W}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}'\mathbf{W}\mathbf{y}$$

- **Challenge Question**: Compute $\mathbf{V} = \mathbf{W}\mathbf{W}$ and express the WLS estimator in terms of $\mathbf{V}$ instead of $\mathbf{W}$.

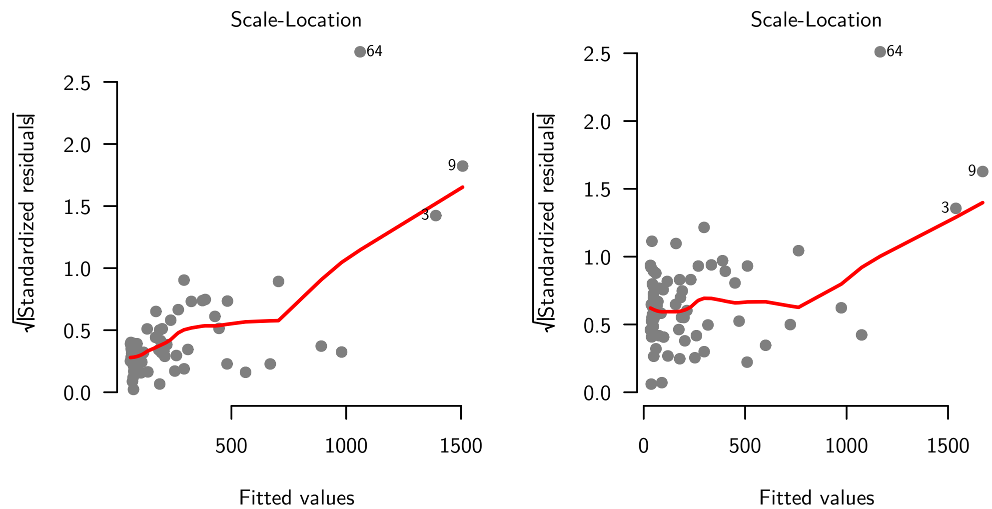
WLS example

- In R, use `weights =` argument to `lm` and give the weights squared: $1/a_i$
- With the Buchanan data, maybe we think that the variance is proportional to the total number of ballots cast:

```
mod.wls <- lm(edaybuchanan ~ edaytotal, weights = 1/edaytotal, data = flvote)
summary(mod.wls)
```

```
##
## Call:
## lm(formula = edaybuchanan ~ edaytotal, data = flvote, weights = 1/edaytotal)
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -1.4046 -0.2146 -0.0501  0.1971  3.4117
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.707e+01  8.507e+00   3.182  0.00225 **
## edaytotal   2.628e-03  2.502e-04  10.503 1.22e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5645 on 65 degrees of freedom
## Multiple R-squared:  0.6292, Adjusted R-squared:  0.6235
## F-statistic: 110.3 on 1 and 65 DF,  p-value: 1.22e-15
```

```
par(mfrow = c(1, 2))
plot(mod, which = 3, las = 1, bty = "n", pch = 19, col = "grey50", lwd = 2)
plot(mod.wls, which = 3, las = 1, bty = "n", pch = 19, col = "grey50", lwd = 2)
```



White's heteroskedasticity consistent estimator

$$\text{Var}[\mathbf{u}] = \mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & 0 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & 0 & \dots & 0 \\ & & & \vdots & \\ 0 & 0 & 0 & \dots & \sigma_n^2 \end{bmatrix}$$

then $\text{Var}[\hat{\boldsymbol{\beta}}|\mathbf{X}] = (\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}'\mathbf{\Sigma}\mathbf{X}\,(\mathbf{X}'\mathbf{X})^{-1}$ and White (1980) shows that

$$\widehat{\text{Var}[\hat{\boldsymbol{\beta}}|\mathbf{X}]} = (\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}' \begin{bmatrix} \hat{\mathbf{u}}_1^2 & 0 & 0 & \dots & 0 \\ 0 & \hat{\mathbf{u}}_2^2 & 0 & \dots & 0 \\ & & & \vdots & \\ 0 & 0 & 0 & \dots & \hat{\mathbf{u}}_n^2 \end{bmatrix} \mathbf{X}\,(\mathbf{X}'\mathbf{X})^{-1}$$

is a consistent estimator of $\text{Var}[\hat{\boldsymbol{\beta}}|\mathbf{X}]$ under any form of heteroskedasticity.

Computing HC/robust standard errors

1. Fit regression and obtain residuals $\widehat{\mathbf{u}}$

2. Construct the "meat" matrix $\widehat{\mathbf{\Sigma}}$ with squared residuals in diagonal:

$$\widehat{\mathbf{\Sigma}} = \begin{bmatrix} \hat{\mathbf{u}}_1^2 & 0 & 0 & \dots & 0 \\ 0 & \hat{\mathbf{u}}_2^2 & 0 & \dots & 0 \\ & & & \vdots & \\ 0 & 0 & 0 & \dots & \hat{\mathbf{u}}_n^2 \end{bmatrix}$$

3. Plug $\widehat{\mathbf{\Sigma}}$ into sandwich formula to obtain HC/robust estimator of the covariance matrix:

$$\text{Var}[\hat{\boldsymbol{\beta}}|\mathbf{X}] = \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{X}'\widehat{\mathbf{\Sigma}}\mathbf{X} \left(\mathbf{X}'\mathbf{X}\right)^{-1}$$

- There are various small-sample corrections to improve performance in (wait for it) small samples. The most common variant is called HC1:

$$V[\hat{\boldsymbol{\beta}}|\mathbf{X}] = \frac{n}{n-k-1} \cdot \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{X}'\widehat{\mathbf{\Sigma}}\mathbf{X} \left(\mathbf{X}'\mathbf{X}\right)^{-1}$$

Robust SEs in Florida data

```
library(sandwich)
library(lmtest)
coeftest(mod)
```

```
##
## t test of coefficients:
##
##               Estimate Std. Error t value  Pr(>|t|)
## (Intercept) 5.4229e+01 4.9141e+01  1.1035    0.2739
## edaytotal   2.3229e-03 3.1041e-04  7.4831 2.417e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
coeftest(mod, vcovHC(mod, type = "HC0"))
```

```
##
## t test of coefficients:
##
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5.4229e+01 4.0613e+01  1.3353  0.18644
## edaytotal   2.3229e-03 8.7048e-04  2.6685  0.00961 **
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
coeftest(mod, vcovHC(mod, type = "HC1"))
```

```
##
## t test of coefficients:
##
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5.4229e+01 4.1233e+01  1.3152  0.19306
## edaytotal   2.3229e-03 8.8377e-04  2.6284  0.01069 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

WLS vs. White's Estimator

- WLS:
    - With known weights, WLS is efficient
    - and $\widehat{SE}[\widehat{\boldsymbol{\beta}}_{WLS}]$ is unbiased and consistent
    - but weights usually aren't known
- White's Estimator:
    - Doesn't change estimate $\widehat{\boldsymbol{\beta}}$
    - Consistent for Var$[\widehat{\boldsymbol{\beta}}]$ under any form of heteroskedasticity
    - Because it relies on consistency, it is a large sample result, best with large n
    - For small n, performance might be poor

CLUSTERING

Clustered dependence: intuition

- Think back to the Gerber, Green, and Larimer (2008) social pressure mailer example.
- Their design: randomly sample households and randomly assign them to different treatment conditions
- But the measurement of turnout is at the individual level
- Violation of **iid/random sampling**:
    - errors of individuals within the same household are correlated

  - $\rightsquigarrow$ violation of homoskedasticity

- Called **clustering** or **clustered dependence**

Clustered dependence: notation

- Clusters: $j = 1, \ldots, m$
- Units: $i = 1, \ldots, n_j$
- n_j is the number of units in cluster j and $n = \sum_j n_j$ is the total number of units
- Units are (usually) belong to a single cluster:
  - voters in households
  - individuals in states
  - students in classes
  - rulings in judges
- Especially important when outcome varies at the unit-level, y_{ij} and the main independent variable varies at the cluster level, x_j.
- It's "cheating" to pretend that you have $n_j \times m$ iid observations when you might really only have m independent clusters.

Clustered dependence: definition

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \varepsilon_{ij} = \beta_0 + \beta_1 x_{ij} + v_j + u_{ij}$$

- $v_j \overset{iid}{\sim} N(0, \rho\sigma^2)$ cluster error component
- $u_{ij} \overset{iid}{\sim} N(0, (1-\rho)\sigma^2)$ unit error component
- v_{ij} and u_{ij} are assumed to be independent of each other
- $\rho \in (0, 1)$ is called the within-cluster correlation.
- What if we ignore this structure and just use ε_{ij} as the error?
- Variance of the composite error is σ^2:

$$\begin{aligned} \text{Var}[\varepsilon_{ij}] &= \text{Var}[v_j + u_{ij}] \\ &= \text{Var}[v_j] + \text{Var}[u_{ij}] \\ &= \rho\sigma^2 + (1-\rho)\sigma^2 = \sigma^2 \end{aligned}$$

- Covariance between two units i and s in the same cluster is $\rho\sigma^2$

$$\begin{aligned}\text{Cov}[\varepsilon_{ij}, \varepsilon_{sj}] &= \text{Cov}[v_j + u_{ij}, v_j + u_{sj}] \\ &= \text{Cov}[v_j, v_j] + \text{Cov}[v_j, u_{sj}] + \text{Cov}[u_{ij}, v_j] + \text{Cov}[u_{ij}, u_{sj}] \\ &= \text{Var}[v_j] + 0 + 0 + 0 = \rho\sigma^2\end{aligned}$$

- Correlation between units in the same group is just ρ:

$$\text{Cor}[\varepsilon_{ij}, \varepsilon_{sj}] = \frac{\text{Cov}[\varepsilon_{ij}, \varepsilon_{sj}]}{\sqrt{\text{Var}\varepsilon_{ij}\text{Var}[\varepsilon_{sj}]}} = \frac{\rho\sigma^2}{\sqrt{\sigma^2\sigma^2}} = \rho$$

- Covariance of two units i and s in different clusters j and k:

$$\begin{aligned}\text{Cov}[\varepsilon_{ij}, \varepsilon_{sk}] &= \text{Cov}[v_j + u_{ij}, v_k + u_{sk}] \\ &= \text{Cov}[v_j, v_k] + \text{Cov}[v_j, u_{sk}] + \text{Cov}[u_{ij}, v_k] + \text{Cov}[u_{ij}, u_{sk}] \\ &= 0 + 0 + 0 + 0 = 0\end{aligned}$$

Example covariance matrix

$$\boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_{1,1} & \varepsilon_{2,1} & \varepsilon_{3,1} & \varepsilon_{4,2} & \varepsilon_{5,2} & \varepsilon_{6,2} \end{bmatrix}'$$

$$\text{Var}[\boldsymbol{\varepsilon}] = \boldsymbol{\Sigma} = \begin{bmatrix} \sigma^2 & \sigma^2 \cdot \rho & \sigma^2 \cdot \rho & 0 & 0 & 0 \\ \sigma^2 \cdot \rho & \sigma^2 & \sigma^2 \cdot \rho & 0 & 0 & 0 \\ \sigma^2 \cdot \rho & \sigma^2 \cdot \rho & \sigma^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma^2 & \sigma^2 \cdot \rho & \sigma^2 \cdot \rho \\ 0 & 0 & 0 & \sigma^2 \cdot \rho & \sigma^2 & \sigma^2 \cdot \rho \\ 0 & 0 & 0 & \sigma^2 \cdot \rho & \sigma^2 \cdot \rho & \sigma^2 \end{bmatrix}$$

Error covariance matrix with clustering

- In general, we can write the covariance matrix as a **block diagonal**
- By independence, the errors are uncorrelated across clusters:

$$\text{Var}[\boldsymbol{\varepsilon}] = \boldsymbol{\Sigma} = \left[\begin{array}{c|c|c|c} \boldsymbol{\Sigma}_1 & 0 & \dots & 0 \\ \hline \mathbf{0} & \boldsymbol{\Sigma}_2 & \dots & \mathbf{0} \\ \hline & & \ddots & \\ \hline \mathbf{0} & \mathbf{0} & \dots & \boldsymbol{\Sigma}_m \end{array}\right]$$

Correcting for clustering

1. Including a dummy variable for each cluster (fixed effects, next week)
2. "Random effects" models (take above model as true and estimate ρ and σ^2)
3. Cluster-robust ("clustered") standard errors
4. Aggregate data to the cluster-level and use OLS $\overline{y}_j = \frac{1}{n_j}\sum_i y_{ij}$
   - If n_j varies by cluster, then cluster-level errors will have heteroskedasticity
   - Can use WLS with cluster size as the weights

Cluster-robust SEs

- First, let's write the within-cluster regressions like so:

$$\mathbf{y}_j = \mathbf{X}_j\boldsymbol{\beta} + \boldsymbol{\varepsilon}_j$$

  where $\mathbf{y}_j$ is the vector of responses for cluster j, $\mathbf{X}_j$ is the matrix of independent variables for cluster j, $\boldsymbol{\varepsilon}_j$ is the error for cluster j.
- We assume that respondents are independent across clusters, but possibly dependent within clusters. Thus, we have $\text{Var}[\boldsymbol{\varepsilon}_j|\mathbf{X}_j] = \boldsymbol{\Sigma}_j$ (but we'll leave this unspecified)
- Leads to this matrix:

$$\text{Var}[\hat{\boldsymbol{\beta}}|\mathbf{X}] = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\left(\sum_{j=1}^{m}\mathbf{X}_j'\boldsymbol{\Sigma}_j\mathbf{X}_j\right)\left(\mathbf{X}'\mathbf{X}\right)^{-1}$$

- Way to estimate this matrix: replace $\boldsymbol{\Sigma}_j$ with an estimate based on the within-cluster residuals, $\widehat{\boldsymbol{\varepsilon}}_j$:

$$\widehat{\boldsymbol{\Sigma}}_j = \widehat{\boldsymbol{\varepsilon}}_j\widehat{\boldsymbol{\varepsilon}}_j'$$

- Final expression for our cluster-robust covariance matrix estimate:

$$\widehat{\text{Var}}[\hat{\boldsymbol{\beta}}|\mathbf{X}] = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\left(\sum_{j=1}^{m}\mathbf{X}_j'\widehat{\boldsymbol{\varepsilon}}_j\widehat{\boldsymbol{\varepsilon}}_j'\mathbf{X}_j\right)\left(\mathbf{X}'\mathbf{X}\right)^{-1}$$

- With small-sample adjustment (which is what most software packages report):

$$\widehat{\text{Var}}_a[\hat{\boldsymbol{\beta}}|\mathbf{X}] = \frac{m}{m-1}\frac{n-1}{n-k-1}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\left(\sum_{j=1}^{m}\mathbf{X}_j'\widehat{\boldsymbol{\varepsilon}}_j\widehat{\boldsymbol{\varepsilon}}_j'\mathbf{X}_j\right)\left(\mathbf{X}'\mathbf{X}\right)^{-1}$$

Example

```r
load("gerber_green_larimer.RData")
social$voted <- 1 * (social$voted == "Yes")
social$treatment <- factor(social$treatment, levels = c("Control", "Hawthorne",
    "Civic Duty", "Neighbors", "Self"))
mod1 <- lm(voted ~ treatment, data = social)
```

```r
source("vcovCluster.R")
coeftest(mod1, vcov = vcovCluster(mod1, "hh_id"))
```

```
##
## t test of coefficients:
##
##                        Estimate Std. Error  t value  Pr(>|t|)
## (Intercept)           0.2966383  0.0013096 226.5172 < 2.2e-16 ***
## treatmentHawthorne    0.0257363  0.0032579   7.8997 2.804e-15 ***
## treatmentCivic Duty   0.0178993  0.0032366   5.5302 3.200e-08 ***
## treatmentNeighbors    0.0813099  0.0033696  24.1308 < 2.2e-16 ***
## treatmentSelf         0.0485132  0.0033000  14.7009 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Cluster-robust standard errors

- CRSE do not change our estimates $\widehat{\boldsymbol{\beta}}$, cannot fix bias
- CRSE is consistent estimator of $\text{Var}[\widehat{\boldsymbol{\beta}}]$ given clustered dependence
  - Relies on independence between clusters, dependence within clusters
  - Doesn't depend on the model we present
  - CRSEs usually $>$ conventional SEs—use when you suspect clustering
- Consistency of the CRSE are in the number of groups, not the number of individuals
  - CRSEs can be incorrect with a small (< 50 maybe) number of clusters
  - Block bootstrap can be a useful alternative (see Gov 2002)

SERIAL CORRELATION

Time dependence: serial correlation

- Sometimes we deal with data that is measured over time, $t = 1, \ldots, T$
- Examples: a country over several years or a person over weeks/months
- Often have **serially correlated**: errors in one time period are correlated with errors in other time periods
- Many different ways for this to happen, but we often assume a very limited type of dependence called AR(1).

AR(1) model

- Model for the mean:
$$y_t = \beta_0 + \beta_1 x_t + u_t$$
- Autoregressive error:
$$u_t = \rho u_{t-1} + e_t \quad \text{where} \quad |\rho| < 1$$
- $e_t \sim N(0, \sigma_e^2)$
- ρ is an unknown **autoregressive coefficient** and measures the dependence/correlation between the errors and lagged errors
- Just one of many possible time-series models: AR(2) has $u_t = \rho u_{t-1} + \delta u_{t-2} + e_t$
- Model could be wrong!

Error structure of the AR(1) model

$$V[\mathbf{u}] = \mathbf{\Sigma} = \sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 & \ldots & \rho^{T-1} \\ \rho & 1 & \rho & \ldots & \rho^{T-2} \\ \rho^2 & \rho & 1 & \ldots & \rho^{T-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{T-1} & \rho^{T-2} & \rho^{T-3} & \ldots & 1 \end{bmatrix}$$

- What is this saying?
  - Cov$[u_1, u_2] = \sigma^2 \rho$
  - Cov$[u_1, u_3] = \sigma^2 \rho^2$
  - Cov$[u_1, u_4] = \sigma^2 \rho^3$
  - Covariance/correlation decreases as time between errors grows (because $|\rho| < 1$)
- ρ is usually positive, which means we under estimate the variance

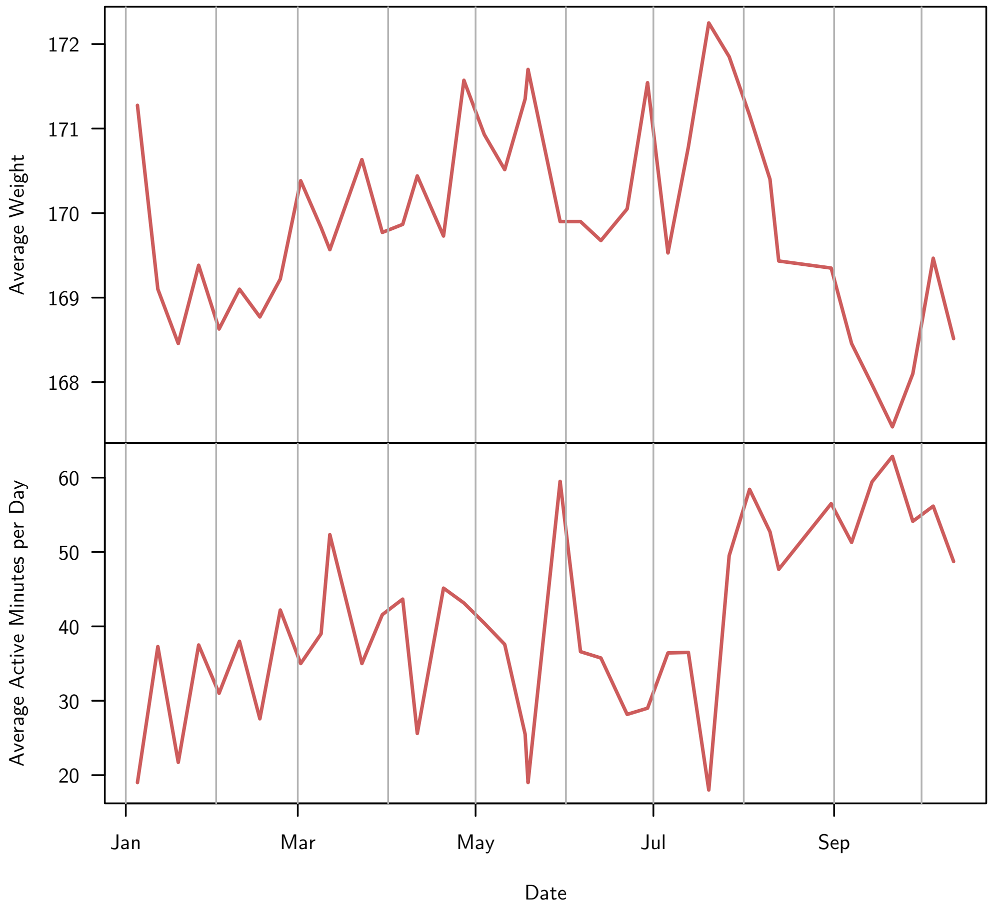
Detecting and fixing serial correlation

- Detection:
  - Plot residuals over time
  - Formal tests (Durbin-Watson statistics)
- Correction:
  - Use SEs that are robust to serial correlation
  - AR corrections (e.g. Prais-Winston, Cochrane-Orcutt, etc)
  - Lagged dependent variables or other dynamic models
  - First-differencing methods

Example: weight and activity

```
weight <- read.csv("weight_by_week.csv", stringsAsFactors = FALSE)
weight$date <- as.Date(weight$date)

par(mfrow = c(2, 1), mar = c(0, 4.1, 1.1, 1.1))
plot(weight$date, weight$weight, type = "l", lwd = 2, las = 1, col = "indianred",
    xaxt = "n", xlab = "", ylab = "Average Weight")
abline(v = seq.Date(from = as.Date("2014-01-01"), to = as.Date("2014-10-01"),
    by = "month"), col = "grey70")
par(mar = c(4.1, 4.1, 0, 1.1))
plot(weight$date, weight$active.mins, type = "l", lwd = 2, las = 1, col = "indianred",
    ylab = "Average Active Minutes per Day", xlab = "Date")
abline(v = seq.Date(from = as.Date("2014-01-01"), to = as.Date("2014-10-01"),
    by = "month"), col = "grey70")
```



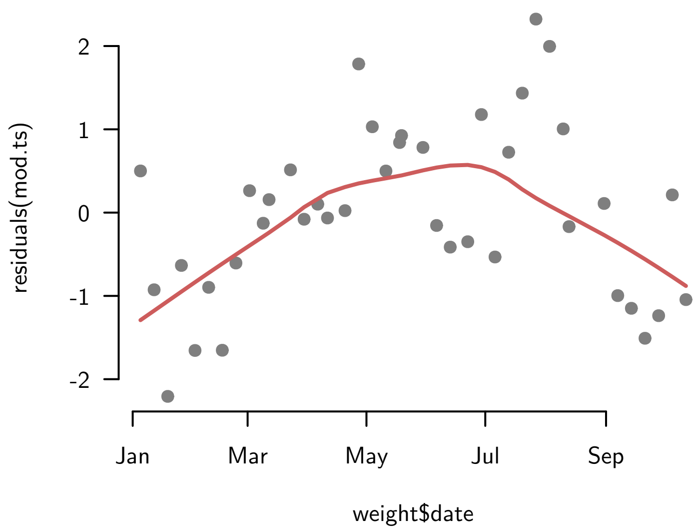

```
mod.ts <- lm(weight ~ active.mins, data = weight)
summary(mod.ts)
```

```
##
## Call:
## lm(formula = weight ~ active.mins, data = weight)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.20644 -0.70006 -0.02049  0.73918  2.32333
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 171.55208    0.58320 294.157  < 2e-16 ***
```

```
## active.mins  -0.04092    0.01384  -2.956  0.00534 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.056 on 38 degrees of freedom
## Multiple R-squared:  0.1869, Adjusted R-squared:  0.1655
## F-statistic: 8.735 on 1 and 38 DF,  p-value: 0.005336
```

Residuals over time

```r
plot(x = weight$date, y = residuals(mod.ts), pch = 19, col = "grey50", las = 1,
    bty = "n")
lines(lowess(x = weight$date, y = residuals(mod.ts)), col = "indianred", lwd = 2)
```



A formal test: Durbin-Watson

- Null, $H_0 : \rho = 0$
- Alternative, $H_a : \rho \neq 0$
- Durbin-Watson statistic:

$$DW = \frac{\sum_{t=2}^{n} (\widehat{u}_t - \widehat{u}_{t-1})^2}{\sum_{t=1}^{n} \widehat{u}_t^2} \quad \text{where} \quad DW \approx 2(1 - \widehat{\rho})$$

- If $DW \approx 2$ then $\widehat{\rho} \approx 0$
- $DW < 1$: strong evidence of positive serial correlation
- $DW > 3$: strong evidence of negative serial correlation

Durbin-Watson on weight

```r
dwtest(mod.ts)
```

```
##
##  Durbin-Watson test
##
## data:  mod.ts
## DW = 0.75138, p-value = 2.141e-06
## alternative hypothesis: true autocorrelation is greater than 0
```

Corrections: HAC standard errors

- We can generalize the HC/robust standard errors to be **heteroskedastic and autocorrelation consistent** (HAC) standard errors.
- Autocorrelation is just another term for serial correlation
- Very similar to HC/robust:
  - $\widehat{\beta}$ remain as our esitmates
  - HAC SEs are consistent for $\text{Var}[\widehat{\beta}]$ in the presence of heteroskedasticity and/or serial correlation
  - Can use the `sandwich` package in R, with covariance function `NeweyWest`

Example: Newey-West standard errors

```r
coeftest(mod.ts, vcov = NeweyWest)
```

```
##
## t test of coefficients:
##
##               Estimate Std. Error  t value Pr(>|t|)
## (Intercept) 171.552079   0.818649 209.5550  < 2e-16 ***
## active.mins  -0.040917   0.021210  -1.9291  0.06121 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```