

450A Political Methodology I

Lecture 6: Multiple Regression

3. Heteroskedasticity, Clustering and Serial Correlation

Stanford University

Diagnosing and Fixing Problems

- In the previous lecture, we have assumed homoskedasticity ($V[u|X] = \sigma^2 I$)
- This assumption can be violated in two ways:
 - (1) Variances are not equal across observations ($V[u_i|X] \neq V[u_j|X]$)
 - (2) Covariances are not zero between observations ($Cov[u_i, u_j] \neq 0$)
- We call (1) **heteroskedasticity** and (2) **clustering** or **serial correlation**, depending on the kind of dependence across observations.
- It turns out that, even if homoskedasticity is violated, we can obtain **robust standard errors** that account for the violation and make statistical inference possible

Heteroskedasticity

Clustering

Serial Correlation

Summary

Review: Variance of $\hat{\beta}$ under Homoskedasticity

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

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

Replacing σ^2 with our estimator $\hat{\sigma}^2$ gives us our estimator for the $(k+1) \times (k+1)$ variance-covariance matrix for the vector of regression coefficients:

$$\widehat{V[\hat{\beta}|\mathbf{X}]} = \hat{\sigma}^2 (\mathbf{X}'\mathbf{X})^{-1}$$

Non-constant Error Variance

Homoskedasticity implies the following variance-covariance matrix of the errors:

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

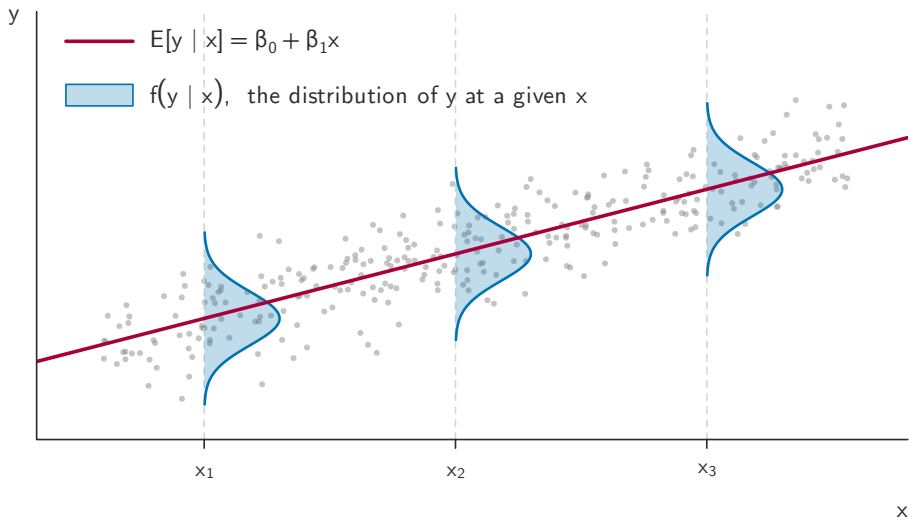
We now allow for violations of homoskedasticity. We may assume:

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

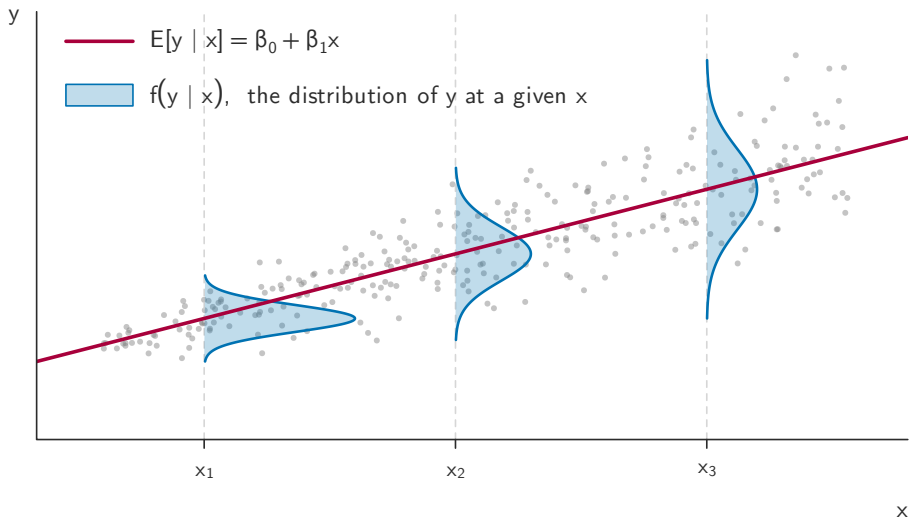
Errors exhibit **heteroskedasticity of unknown form**:

- $\text{Cov}[u_i, u_j | \mathbf{X}] = 0 \implies$ We still assume independence of errors across observations
- $V[u_i | \mathbf{X}] = \sigma_i^2 \implies$ We however allow error variances to vary across observations

Example: $V[u_i|\mathbf{X}] = \sigma^2$ Homoskedasticity



Example: $V[u_i|\mathbf{X}] = \sigma_i^2$ Heteroskedasticity



Consequences of Heteroskedasticity

- The usual estimator $\hat{\sigma}^2 (\mathbf{X}'\mathbf{X})^{-1}$ is no longer unbiased or consistent for $V[\hat{\beta}|\mathbf{X}]$. This implies that the usual standard errors are no longer valid
- Test statistics, which are functions of the standard errors, will no longer have t and F distributions.
- For “ α ” level tests, probability of Type I error will not be α .
- “ $1 - \alpha$ ” confidence intervals will not have $1 - \alpha$ coverage probability.
- The OLS estimator is no longer BLUE.

However,

- The degree of the problem depends on how serious the heteroskedasticity is.
- If the other assumptions still hold (including zero conditional mean), $\hat{\beta}$ is still unbiased and consistent for β . Why?

Visual Diagnostics of Non-constant Error Variances

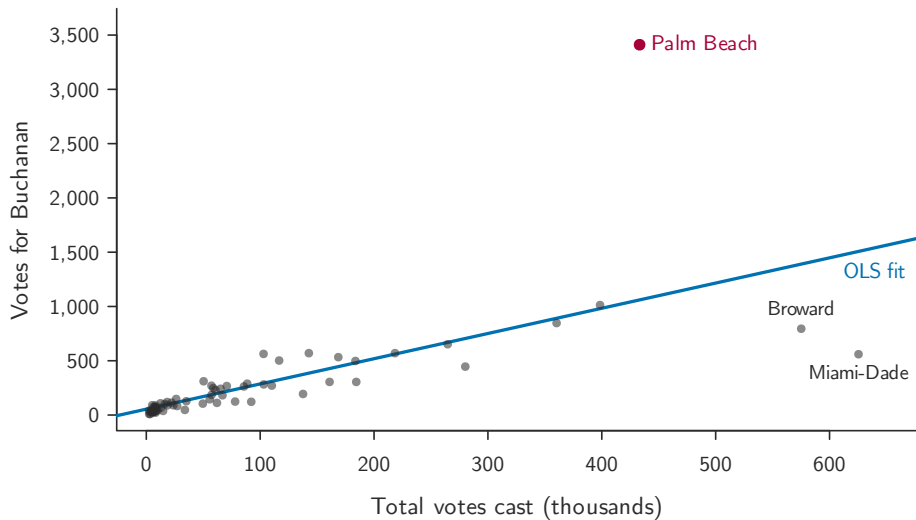
Because we cannot directly observe the errors, we investigate **residuals** to see if there is evidence for heteroskedasticity:

- Under homoskedasticity, we expect the variance of the residuals to be constant across the fitted values and all the values of the explanatory variables
- We can therefore diagnose the plausibility of homoskedasticity by plotting:
 - Residuals against fitted values ($\hat{y}_i$)
 - Residuals against explanatory variables (x_{ij})

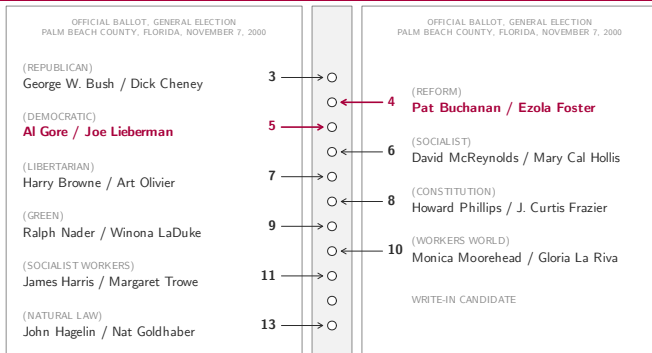
Example: Buchanan Votes in Florida

- We examine data from the 2000 presidential election in Florida used in Wand et al. (2001) *APSR*.
- Our analysis takes place at the county level, and we will regress the number of Buchanan votes in each county on the total number of votes in each county.
- Nationwide Buchanan finished fourth in the presidential election with 0.4% of the popular vote.

Buchanan Votes and Total Votes

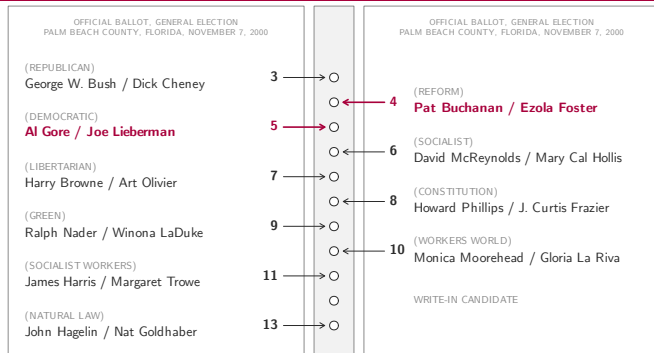


The Butterfly Ballot



Bush spokesman Ari Fleischer stated, "Palm Beach county is a Pat Buchanan stronghold and that's why Pat Buchanan received 3,407 votes there." Buchanan said "it's very easy for me to see how someone could have voted for me in the belief they voted for Al Gore"

The Butterfly Ballot



Wand et al. (2001) conclude that the ballot caused more than 2,000 Democratic voters to vote by mistake for Buchanan, more than Bush's certified margin of 537 votes in Florida. Other things equal, Gore would have won Florida's 25 electoral votes and the presidency.

Buchanan Votes and Total Votes

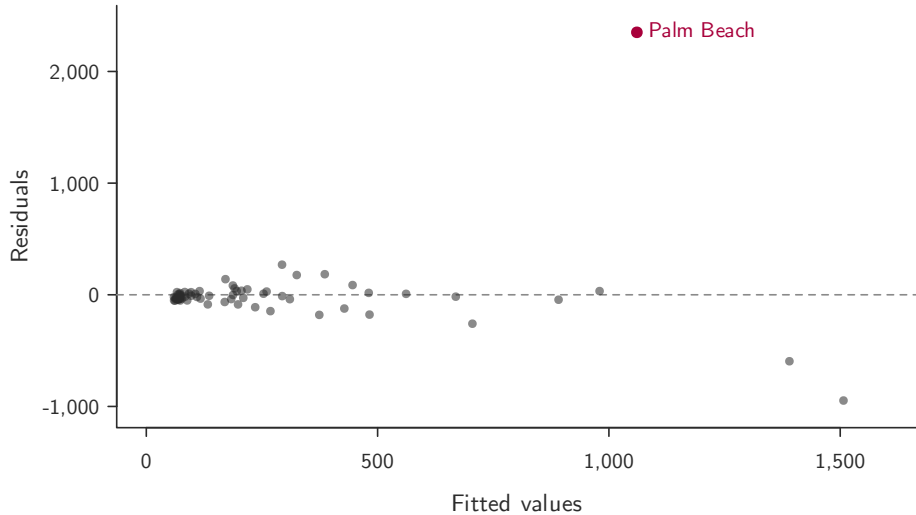
R Code

```
> mod1 <- lm(buchanan00~TotalVotes00,data=dta)
> summary(mod1)
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
TotalVotes00 2.323e-03  3.104e-04   7.483 2.42e-10 ***
---
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

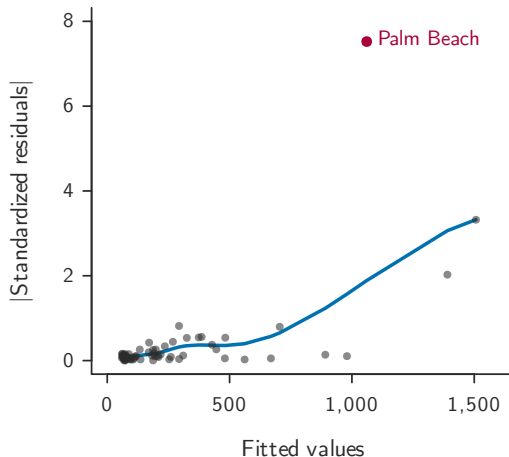
> residuals           <- resid(mod1)
> standardized_residuals <- rstandard(mod1)
> studentized_residuals <- rstudent(mod1)
> plot(fitted(mod1), residuals)
```

Example: Buchanan Votes in Florida – Residuals



Example: Buchanan Votes in Florida – Magnitude of Residuals

- There is a positive relationship between total votes and votes for Buchanan.
- Plotting the residuals against the fitted values gives us some sense of the variance of the residuals
- We can do better by constructing the so-called **spread-location plot**:
 - Standardize the residuals and take absolute values to fold the plot in half
 - Add a smooth (loess) line to check if the variance varies with the fitted values



Buchanan Votes: Why the Spread Grows

- Buchanan votes is a **count**, and the spread of a count grows with its size: for a pure count, roughly with the square root (ten times the voters, about three times the spread). So the spread of the residuals grows with county size.
- The residual SD is 25 votes in the smallest third of counties, 60 in the middle third and 191 in the largest third (refit without Palm Beach).
- Three big counties drive the fit in levels. Miami-Dade, Broward and Palm Beach cast 27% of the votes. Their leverage h_{ii} (the i th diagonal entry of the hat matrix $\mathbf{P}$) is .27, .22 and .12. The average is .03.
- They pull the line in opposite directions. The slope is 2.32 Buchanan votes per 1,000 votes cast. It is 1.52 without Palm Beach and 3.67 without Miami-Dade and Broward.

Formal Tests for Non-constant Error Variances

- Plots are usually sufficient, but formal tests for heteroskedasticity exist (e.g. Breusch-Pagan, Cook-Weisberg, White, etc.).
- They are all roughly based on the same idea:
 - $H_0: V[u_i|\mathbf{X}] = \sigma^2$
 - Under the zero conditional mean assumption, this is equivalent to $H_0: E[u_i^2|\mathbf{X}] = E[u_i^2] = \sigma^2$, a constant unrelated to $\mathbf{X}$
 - This implies that, under H_0 , the **squared residuals** should also be unrelated to the explanatory variables
- The **Breusch-Pagan test**:
 - 1 Regress y_i on $\mathbf{x}_i$ and store residuals $\hat{u}_i$
 - 2 Regress $\hat{u}_i^2$ on $\mathbf{x}_i$
 - 3 Test that all slope coefficients are zero (an F test, or the χ^2 statistic that `bptest` reports)
 - In R, `bptest` in the `lmtest` package

Dealing with Non-constant Error Variances

Once we detect a likely heteroskedasticity, how should we deal with it?

- ① Use a **variance stabilizing transformation** (this will affect the other model assumptions).
- ② Model the heteroskedasticity (e.g. **Weighted Least Squares, WLS**). This needs a good “weighting” model. Not common in applied work.
- ③ Use an estimator of $V[\hat{\beta}]$ that is **robust to heteroskedasticity** (e.g. White’s heteroskedasticity consistent estimator and its descendants).
- ④ Admit that we have the wrong model and use an entirely different approach (e.g. nonlinear regression).

Variance Stabilizing Transformations

Let σ_i^2 be the conditional variance of Y_i . If σ_i^2 is proportional to some function of the mean $\mu_i = \mathbf{x}_i\boldsymbol{\beta}$, then a variance stabilizing transformation may be appropriate.

Note: Transformations will affect the other regression assumptions, as well as interpretation of the regression coefficients.

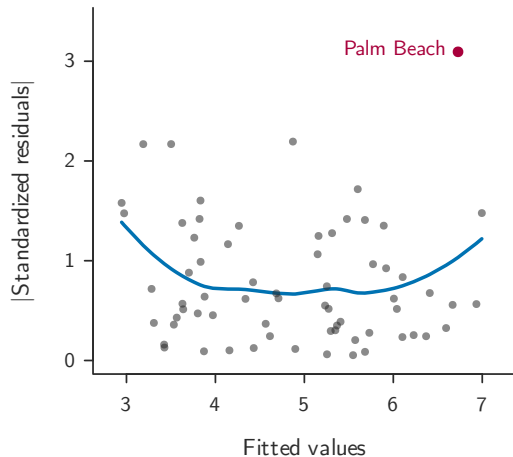
Transformation	Mean/Variance Relationship
$\sqrt{Y}$	$\sigma_i^2 \propto \mu_i$
$\log Y$	$\sigma_i^2 \propto \mu_i^2$
$1/Y$	$\sigma_i^2 \propto \mu_i^4$

For example, if $\sigma_i^2 \propto \mu_i$, then by the Delta method,

$$\text{var}(\sqrt{Y_i} \mid \mathbf{X}) \approx \frac{1}{4\mu_i} \text{var}(Y_i \mid \mathbf{X}) = \text{const.}$$

Example: Buchanan Votes in Florida

- Things tend to look better after a log transformation of Y and X
- We are much closer to a flat line now indicating that we approximately have a constant variance
- We could probably improve further by removing Palm Beach



Example: Buchanan Votes in Florida

The Breusch-Pagan test rejects constant variance in levels but not in logs (not rejecting does not prove constant variance):

```
R Code
> library(lmtest)
> mod1 <- lm(buchanan00 ~ TotalVotes00, data=dta)
> bptest(mod1, studentize = FALSE)

Breusch-Pagan test

data:  mod1
BP = 250.0671, df = 1, p-value < 2.2e-16
> dta$lnbuchanan00 <- log(dta$buchanan00)
> dta$lnTotalVotes00 <- log(dta$TotalVotes00)
> mod2 <- lm(lnbuchanan00 ~ lnTotalVotes00, data=dta)
> bptest(mod2, studentize = FALSE)

Breusch-Pagan test

data:  mod2
BP = 0.0111, df = 1, p-value = 0.916
```

With bptest's default (Koenker's studentized version): BP = 12.6, $p = 0.0004$ in levels; BP = 0.010, $p = 0.92$ in logs.

White's Heteroskedasticity Consistent Estimator

Let's go back to our derivation of the variance of the regression estimator, when the error structure, Σ , may be heteroskedastic:

$$\begin{aligned}V[\hat{\beta}|\mathbf{X}] &= V[\beta|\mathbf{X}] + V[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}|\mathbf{X}] \\&= V[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}|\mathbf{X}] \\&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'V[\mathbf{u}|\mathbf{X}](\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \text{ (note: } \mathbf{X} \text{ nonrandom } |\mathbf{X}) \\&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'V[\mathbf{u}|\mathbf{X}]\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\Sigma\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\end{aligned}$$

Idea: one squared residual cannot estimate each σ_i^2 well, but we only need the $(k+1) \times (k+1)$ average $\frac{1}{n}\mathbf{X}'\Sigma\mathbf{X}$, and $\frac{1}{n}\mathbf{X}'\hat{\Sigma}\mathbf{X}$ estimates it consistently.

White's Heteroskedasticity Consistent Estimator

Suppose we have **heteroskedasticity of unknown form**:

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

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

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

is a consistent estimator of $V[\hat{\beta}|\mathbf{X}]$ **under any form of heteroskedasticity**, as long as the observations are independent.

The estimate based on the above is called the **heteroskedasticity consistent (HC)** or **robust standard errors**.

White's Heteroskedasticity Consistent Estimator

Robust standard errors are easily computed with the “sandwich” formula:

- 1 Fit the regression and obtain the residuals $\hat{u}$
- 2 Construct $\hat{\Sigma}$ with the squared residuals on the diagonal (the “meat” of the sandwich is $\mathbf{X}'\hat{\Sigma}\mathbf{X}$; the “bread” is $(\mathbf{X}'\mathbf{X})^{-1}$):

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

- 3 Plug $\hat{\Sigma}$ into the sandwich formula to obtain the robust estimator of the variance-covariance matrix

$$\hat{V}_{HC0}[\hat{\beta}|\mathbf{X}] = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\hat{\Sigma}\mathbf{X} (\mathbf{X}'\mathbf{X})^{-1}$$

- There are various **small sample corrections** to improve performance when sample size is small. The most common variant (sometimes labeled HC1) is:

$$\hat{V}_{HC1}[\hat{\beta}|\mathbf{X}] = \frac{n}{n-k-1} \cdot (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\hat{\Sigma}\mathbf{X} (\mathbf{X}'\mathbf{X})^{-1}$$

Regular & Robust Standard Errors in Florida Example

R Code

```
> library(sandwich)
> library(lmtest)
> coeftest(mod1) # homoskedasticity
t test of coefficients:
      Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.4231e+01 4.9141e+01  1.1036  0.2738
TotalVotes00 2.3229e-03 3.1041e-04  7.4831 2.417e-10 ***

> coeftest(mod1,vcov = vcovHC(mod1, type = "HC0")) # classic White
t test of coefficients:
      Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.4231e+01 4.0612e+01  1.3353  0.18642
TotalVotes00 2.3229e-03 8.7047e-04  2.6685  0.00961 **

> coeftest(mod1,vcov = vcovHC(mod1, type = "HC1")) # small sample correction
t test of coefficients:
      Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.4231e+01 4.1232e+01  1.3153  0.19304
TotalVotes00 2.3229e-03 8.8376e-04  2.6284  0.01069 *
```

Florida: HC3, and Without Palm Beach

- **HC3** puts $\hat{u}_i^2 / (1 - h_{ii})^2$ on the diagonal of $\hat{\Sigma}$ instead of $\hat{u}_i^2$.
- This matters with high-leverage points. OLS pulls the line toward them, so their residuals understate their errors. HC3 scales these residuals back up. Miami-Dade's squared residual counts 1.85 times.

Slope and SEs (per 1,000 votes cast)	All 67 counties	Without Palm Beach
Slope	2.32	1.52
Classical SE	0.31	0.12
HC0	0.87	0.29
HC1	0.88	0.30
HC3	1.05	0.40

- The unequal spread is not only Palm Beach. Without it, the Breusch-Pagan test still rejects ($p < .001$), and the robust SE (HC1) is 2.5 times the classical SE.

White's estimator:

$$\hat{V}_{HC0} \equiv \hat{V}[\hat{\beta}|\mathbf{X}] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\hat{\Sigma}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$$

- White's estimator does not change our point estimate $\hat{\beta}$
- White's estimator for $V[\hat{\beta}]$ is consistent for any form of heteroskedasticity.
- If n is large, you can get approximately correct CI coverage and test levels.
- For small n , sometimes a correction factor is utilized to improve performance, but this is often insufficient to address the problem.

Worksheet Q1 The classic variance estimator is

$$\hat{V}_0 = \hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}.$$

The White estimator is

$$\hat{V}_{HC0} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\text{diag}(\hat{u}_1^2, \dots, \hat{u}_n^2)\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}.$$

Assume (1) $E[u_i | \mathbf{X}] = 0$; (2) $\text{Var}(u_i | \mathbf{X}) = \sigma^2$; (3) $\frac{1}{n}\mathbf{X}'\mathbf{X} \rightarrow Q \succ 0$; and (4) standard regularity conditions. Show that

$$\text{plim}(n \hat{V}_0) = \text{plim}(n \hat{V}_{HC0}) = \sigma^2 Q^{-1}.$$

Hint:

$$\frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i' \hat{u}_i^2 = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i' u_i^2 + o_p(1) \xrightarrow{p} \sigma^2 E[\mathbf{x}_i \mathbf{x}_i'] = \sigma^2 Q.$$

Heteroskedasticity

Clustering

Serial Correlation

Summary

Cluster Dependence: Intuition

- Assume we draw a random sample of households and randomly assign half of them to an encouragement to vote
- We measure turnout for all individuals and regress this outcome on a treatment indicator
- How does this violate our regression assumptions?
 - Regression assumes that individuals are sampled independently, but this is not true because sampling and the intervention occurred at the household level, not at the level of individuals
 - Errors/outcomes of individuals in the same household are correlated so by assuming independence we overstate the amount of information
 - Because the treatment is the same for everyone in a household and the errors are positively correlated within households, conventional standard errors will be biased downward
- Scenario is called **clustering** or **cluster dependence** and this problem is very common in many social science data applications.

Cluster Dependence: Intuition

- Often in social research units $i = 1, \dots, N$ are nested within clusters $j = 1, \dots, M$
- Clusters could be any grouping that could be expected to lead to dependencies:
 - voters in households
 - individuals in states
 - students in classes
 - rulings by judges
 - yearly inflation rates of countries
- Two types of clustering: clustering in sampling and clustering in assignment
- “Randomization by cluster accompanied by an analysis appropriate to randomization by individual is an exercise in self-deception” (Cornfield 1978, pp. 101–102)
- “making a data set larger by copying a smaller one n times generates no new information” (Angrist and Pischke 2009)

Cluster Dependence: A Formal Definition

Assume **units** $i = 1, \dots, N$ are nested within **clusters** $j = 1, \dots, M$ (e.g. voters within households, students within classes, etc.).

We assume a variance structure that allows the errors to be:

- potentially correlated for units within the same cluster (correlation within)
- uncorrelated for units across clusters (independence across)

Our model with cluster dependence is:

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \varepsilon_{ij} = \beta_0 + \beta_1 x_{ij} + \mathbf{v}_j + \mathbf{u}_{ij},$$

where

- $\mathbf{v}_j \stackrel{\text{i.i.d.}}{\sim} N(0, \rho\sigma^2)$ cluster error component
- $\mathbf{u}_{ij} \stackrel{\text{i.i.d.}}{\sim} N(0, (1 - \rho)\sigma^2)$ unit error component
- Note that $\mathbf{v}_j$ and $\mathbf{u}_{ij}$ are assumed to be independent of each other and $x_{i,j}$
- ρ is called the **within-cluster correlation**

Cluster Dependence: Example

Worksheet Q2

Suppose we have $M = 2$ clusters and $N = 6$ units: cluster 1 contains units $i = 1, 2, 3$ and cluster 2 contains units $i = 4, 5, 6$. Stack the errors in the order

$$\boldsymbol{\varepsilon} = (\varepsilon_{1,1}, \varepsilon_{2,1}, \varepsilon_{3,1}, \varepsilon_{4,2}, \varepsilon_{5,2}, \varepsilon_{6,2})'.$$

- ❶ Show that $V[\varepsilon_{ij}] = \sigma^2$ for all i, j .
- ❷ For two units $i \neq \ell$ in the *same* cluster j , compute $\text{Cov}(\varepsilon_{ij}, \varepsilon_{\ell j})$.
- ❸ For units in *different* clusters $j \neq k$, compute $\text{Cov}(\varepsilon_{ij}, \varepsilon_{\ell k})$.
- ❹ Using your answers above, write out the 6×6 covariance matrix $\Sigma = V[\boldsymbol{\varepsilon}]$ for the given ordering $(\varepsilon_{1,1}, \varepsilon_{2,1}, \varepsilon_{3,1}, \varepsilon_{4,2}, \varepsilon_{5,2}, \varepsilon_{6,2})'$.
- ❺ Compute the correlation $\text{Cor}(\varepsilon_{ij}, \varepsilon_{\ell j})$ for $i \neq \ell$ in the same cluster j and verify that it equals ρ .

Cluster Dependence: Example

$$\epsilon = \begin{bmatrix} \epsilon_{1,1} & \epsilon_{2,1} & \epsilon_{3,1} & \epsilon_{4,2} & \epsilon_{5,2} & \epsilon_{6,2} \end{bmatrix}'$$

$$V[\epsilon] = \Sigma = \begin{bmatrix} V[\epsilon_{1,1}] & \text{Cov}[\epsilon_{2,1}, \epsilon_{1,1}] & \text{Cov}[\epsilon_{3,1}, \epsilon_{1,1}] & . & . & . \\ \text{Cov}[\epsilon_{1,1}, \epsilon_{2,1}] & V[\epsilon_{2,1}] & \text{Cov}[\epsilon_{3,1}, \epsilon_{2,1}] & . & . & . \\ \text{Cov}[\epsilon_{1,1}, \epsilon_{3,1}] & \text{Cov}[\epsilon_{2,1}, \epsilon_{3,1}] & V[\epsilon_{3,1}] & . & . & . \\ \text{Cov}[\epsilon_{1,1}, \epsilon_{4,2}] & \text{Cov}[\epsilon_{2,1}, \epsilon_{4,2}] & \text{Cov}[\epsilon_{3,1}, \epsilon_{4,2}] & V[\epsilon_{4,2}] & . & . \\ \text{Cov}[\epsilon_{1,1}, \epsilon_{5,2}] & \text{Cov}[\epsilon_{2,1}, \epsilon_{5,2}] & \text{Cov}[\epsilon_{3,1}, \epsilon_{5,2}] & \text{Cov}[\epsilon_{4,2}, \epsilon_{5,2}] & V[\epsilon_{5,2}] & . \\ \text{Cov}[\epsilon_{1,1}, \epsilon_{6,2}] & \text{Cov}[\epsilon_{2,1}, \epsilon_{6,2}] & \text{Cov}[\epsilon_{3,1}, \epsilon_{6,2}] & \text{Cov}[\epsilon_{4,2}, \epsilon_{6,2}] & \text{Cov}[\epsilon_{5,2}, \epsilon_{6,2}] & V[\epsilon_{6,2}] \end{bmatrix}$$

$$= \begin{bmatrix} \sigma^2 & \sigma^2 \cdot \rho & \sigma^2 \cdot \rho & 0 & 0 & 0 \\ \sigma^2 \cdot \rho & \sigma^2 & \sigma^2 \cdot \rho & 0 & 0 & 0 \\ \sigma^2 \cdot \rho & \sigma^2 \cdot \rho & \sigma^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma^2 & \sigma^2 \cdot \rho & \sigma^2 \cdot \rho \\ 0 & 0 & 0 & \sigma^2 \cdot \rho & \sigma^2 & \sigma^2 \cdot \rho \\ 0 & 0 & 0 & \sigma^2 \cdot \rho & \sigma^2 \cdot \rho & \sigma^2 \end{bmatrix}$$

Error Variance Matrix with Cluster Dependence

The variance-covariance matrix of the error, Σ , is **block diagonal**:

- By independence, the errors are uncorrelated across clusters:

$$V[\epsilon] = \Sigma = \begin{bmatrix} \Sigma_1 & 0 & \dots & 0 \\ 0 & \Sigma_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \Sigma_M \end{bmatrix}$$

- But the errors may be correlated for units within the same cluster:

$$\Sigma_j = \begin{bmatrix} \sigma^2 & \sigma^2 \cdot \rho & \dots & \sigma^2 \cdot \rho \\ \sigma^2 \cdot \rho & \sigma^2 & \dots & \sigma^2 \cdot \rho \\ \vdots & \vdots & \ddots & \vdots \\ \sigma^2 \cdot \rho & \sigma^2 \cdot \rho & \dots & \sigma^2 \end{bmatrix}$$

Estimating Cluster Robust Standard Errors

We can relax this further using a more flexible cluster “corrected” estimator that allows for heteroskedasticity and arbitrary cluster specific within correlations of the errors

$$\widehat{\text{Var}}(\hat{\beta} \mid \mathbf{X}) = (\mathbf{X}'\mathbf{X})^{-1} \left(\sum_{j=1}^M \mathbf{x}'_j \hat{\Sigma}_j \mathbf{x}_j \right) (\mathbf{X}'\mathbf{X})^{-1},$$

where M is the total number of clusters, $\mathbf{X}_j$ holds the rows of $\mathbf{X}$ (the predictors, including the intercept) for the units in cluster j , and $\hat{\Sigma}_j$ estimates the covariance matrix of the errors in cluster j (next slide).

Justification for this estimator is based on a central limit theorem in the number of clusters so M needs to be large. If M is small the variance estimates can be severely downward biased. With few clusters, a wild cluster bootstrap-t does better (Cameron, Gelbach and Miller 2008).

Worksheet Q3

The general cluster-robust variance estimator can be written as

$$\widehat{V}_{CR}[\hat{\beta}] = (\mathbf{X}'\mathbf{X})^{-1} \left(\sum_{j=1}^M \mathbf{X}'_j \hat{\mathbf{u}}_j \hat{\mathbf{u}}'_j \mathbf{X}_j \right) (\mathbf{X}'\mathbf{X})^{-1},$$

where $\hat{\mathbf{u}}_j$ is the vector of OLS residuals in cluster j (they estimate the full errors $\varepsilon_{ij} = v_j + u_{ij}$, not only u_{ij}).

Explain how this generalizes the White estimator.

Estimating Cluster Robust Standard Errors

- The more flexible cluster “corrected” estimator can be found in the `estimatr` package

```
> library("estimatr")  
> lm_robust(y ~ x1 + x2, data = dta, clusters = district, se_type = "stata")
```

- Or alternatively:

```
> library(sandwich)  
> vcovCL.2 <- vcovCL(model, cluster = ~ district)  
> coeftest(model, vcov = vcovCL.2)
```

Example: Social Pressure Experiment

- Gerber, Green and Larimer (2008) conducted a “social pressure” experiment in Michigan where a randomly selected group of voters received a letter encouraging them to vote.
- The outcome, voted or not, is measured at the **individual level**. The treatment was assigned by **household**. Everyone in a household got the same letter, or none.
- Individuals are nested within households and we may expect a fair amount of within-household correlation
- Let's look at a random subset of the data

Example: Social Pressure Experiment

Neighbors mailing

3 0 4 2 3 - 3 ||| || ||| |||

For more information: (517) 351-1975
email: ctow@gnbner.com
Practical Political Consulting
P. O. Box 6249
East Lansing, MI 48826

PSRST STD
U.S. Postage
PAID
Lansing, MI
Permit # 444

ECRL0T **C050
THE JACKSON FAMILY
9999 MAPLE DR
FLINT MI 48507

Dear Registered Voter:

WHAT IF YOUR NEIGHBORS KNEW WHETHER YOU VOTED?

Why do so many people fail to vote? We've been talking about the problem for years, but it only seems to get worse. This year, we're taking a new approach. We're sending this mailing to you and your neighbors to publicize who does and does not vote.

The chart shows the names of some of your neighbors, showing which have voted in the past. After the August 8 election, we intend to mail an updated chart. You and your neighbors will all know who voted and who did not.

DO YOUR CIVIC DUTY — VOTE!

MAPLE DR	Aug 04	Nov 04	Aug 05
9995 JOSEPH JAMES SMITH	Voted	Voted	_____
9995 JENNIFER KAY SMITH		Voted	_____
9997 RICHARD B JACKSON		Voted	_____
9999 KATHY MARIE JACKSON		Voted	_____
9999 BRIAN JOSEPH JACKSON		Voted	_____
9991 JENNIFER KAY THOMPSON		Voted	_____
9991 BOB R THOMPSON		Voted	_____
9993 BILL S SMITH			_____
9989 WILLIAM LUKE CASPER		Voted	_____
9989 JENNIFER SUE CASPER		Voted	_____
9987 MARIA S JOHNSON	Voted	Voted	_____
9987 TOM JACK JOHNSON	Voted	Voted	_____
9987 RICHARD TOM JOHNSON		Voted	_____
9985 ROSEMARY S SUE		Voted	_____
9985 KATHRYN L SUE		Voted	_____
9985 HOWARD BEN SUE		Voted	_____
9983 NATHAN CHAD BERG		Voted	_____
9983 CARRIE ANN BERG		Voted	_____
9981 EARL JOEL SMITH			_____
9979 DEBORAH KAY WAYNE		Voted	_____
9979 JOEL R WAYNE		Voted	_____

Source: The "Neighbors" mailing, Gerber, Green & Larimer (2008) replication archive, doi:10.60600/YU/CGMWNW (CC0)

Example: Social Pressure Experiment

R Code

```
> mod1 <- lm(voted~Tr_Hawthorne+Tr_CivicDuty+Tr_Neighbors+Tr_Self,data=dta)
> coeftest(mod1) # homoskedastic
t test of coefficients:

            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.3076541  0.0052062  59.0935 < 2.2e-16 ***
Tr_Hawthorne  0.0237709  0.0128791   1.8457  0.06496 .
Tr_CivicDuty -0.0044725  0.0127744  -0.3501  0.72626
Tr_Neighbors  0.0718986  0.0129031   5.5722 2.561e-08 ***
Tr_Self       0.0502013  0.0127711   3.9308 8.505e-05 ***
```

In this subset (14,392 voters), the intra household error correlation is estimated at $\hat{\rho} = .50$ (.53 in the full sample of 344,084 voters)

This suggests that treating voters within the same household as independent is unrealistic (i.e., about half of the error variance comes from a component shared by everyone in the household)

Example: Social Pressure Experiment

R Code

```
> mod1 <- lm(voted~Tr_Hawthorne+Tr_CivicDuty+Tr_Neighbors+Tr_Self,data=dta)
> coeftest(mod1) # homoskedastic
t test of coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.3076541  0.0052062  59.0935 < 2.2e-16 ***
Tr_Hawthorne  0.0237709  0.0128791   1.8457  0.06496 .
Tr_CivicDuty -0.0044725  0.0127744  -0.3501  0.72626
Tr_Neighbors  0.0718986  0.0129031   5.5722 2.561e-08 ***
Tr_Self       0.0502013  0.0127711   3.9308 8.505e-05 ***
>
> coeftest(mod1, vcov = vcovCL(mod1, cluster = ~ hh_id)) # clustered
t test of coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.3076541  0.0063711  48.2894 < 2.2e-16 ***
Tr_Hawthorne  0.0237709  0.0157746   1.5069  0.131857
Tr_CivicDuty -0.0044725  0.0154991  -0.2886  0.772917
Tr_Neighbors  0.0718986  0.0164533   4.3699 1.252e-05 ***
Tr_Self       0.0502013  0.0162195   3.0951  0.001971 **
```

CR standard errors are 21–28% larger than homoskedastic

Example: Large Clusters, a Regressor Assigned by Region

- British Election Study, May–June 2016: 16,331 respondents in 167 NUTS-3 regions (98 per region on average).
- $Y = 1$ if the respondent intends to vote Leave. X = the Chinese import shock of the respondent's region, the same for everyone in it. Both models control for age and gender.

	$\hat{\beta}_{shock}$	Classical SE	Clustered SE	Ratio	$\hat{\rho}$
Without region dummies	0.279	0.029	0.067	2.3	.03
With NUTS-1 region dummies (11 regions)	0.154	0.032	0.039	1.2	.01

- $\hat{\rho} = .03$ is small, but the clusters are large. With 98 respondents per region, it is enough to more than double the SE.
- The ratio falls because people in one part of Britain share regional factors and shocks (for example, Scotland's political culture). The first model leaves them in the error. The region dummies absorb them, so little correlation is left within NUTS-3 regions.

Source: Colantone & Stanig (2018), *APSR* 112(2), individual data (BES wave 8, 2016); linear probability model, SEs clustered by NUTS-3 region

Cluster Robust Standard Errors (CRSE)

- CRSE do not change our estimate $\hat{\beta}$ and do not fix potential bias in $\hat{\beta}$
- CRSE are consistent estimators for $V[\hat{\beta}]$ given cluster dependence (i.e. independence across clusters, dependence within)
 - How much larger CRSE are depends on the within-cluster correlation of both the errors and the regressor, and on cluster size (21–28% for the letters, 2.3 times for the Brexit survey)
 - Always use CRSE if you suspect cluster dependence, unless number of clusters is small
- Performance of CRSE depends on number of clusters. With too few clusters, CRSE are not valid and often downward biased
 - There is no safe cutoff: depending on the setting, “few” can mean fewer than 20 to fewer than 50 balanced clusters (Cameron and Miller 2015).
 - A simulation-based method called the wild cluster bootstrap can be used if number of clusters is too small (450C and D)

Heteroskedasticity

Clustering

Serial Correlation

Summary

Time Dependence: Serial Correlation

Suppose we observe a unit at multiple times $t = 1, \dots, T$ (e.g. a country over several years, an individual over several months, etc.).

Such observations are often **serially correlated** (not independent across time). We can model this with the following **AR(1) model**:

$$y_t = \beta_0 + \beta_1 x_t + u_t$$

where the **autoregressive error** is

$$u_t = \rho u_{t-1} + e_t \quad \text{where} \quad |\rho| < 1$$

- $e_t \sim N(0, \sigma_e^2)$
- ρ is an unknown **autoregressive coefficient** (note if $\rho = 0$ we have classic errors used before)
- Typically assume stationarity meaning that $V[u_t]$ and $\text{Cov}[u_t, u_{t+h}]$ are independent of t
- Generalizes to higher order serial correlation (e.g. an AR(2) model is given by $u_t = \rho u_{t-1} + \delta u_{t-2} + e_t$).

The Error Structure for the AR(1) Model

We have $\mathbf{u} = [u_1 \quad u_2 \quad \dots \quad u_T]'$ and the AR(1) model implies the following error structure:

$$V[\mathbf{u}] = \Sigma = \frac{\sigma_e^2}{(1 - \rho^2)} \begin{bmatrix} 1 & \rho & \rho^2 & \dots & \rho^{T-1} \\ \rho & 1 & \rho & \dots & \rho^{T-2} \\ \rho^2 & \rho & 1 & \dots & \rho^{T-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{T-1} & \rho^{T-2} & \rho^{T-3} & \dots & 1 \end{bmatrix}$$

That is, the covariance between errors in $t = 1$ and $t = 2$ is $\frac{\sigma_e^2}{(1 - \rho^2)}\rho$, between errors in $t = 1$ and $t = 3$ is $\frac{\sigma_e^2}{(1 - \rho^2)}\rho^2$, etc.

This implies that the correlation between the errors decays exponentially with the number of periods separating them.

ρ is usually positive. If the regressor is also positively autocorrelated, as most time series are, ignoring serial correlation makes the conventional standard errors too small.

Derivation of Error Structure for the AR(1) Model

We have

$$V[u_t] = V[\rho u_{t-1} + e_t] = \rho^2 V[u_{t-1}] + \sigma_e^2$$

with stationarity, $V[u_t] = V[u_{t-1}]$, and so

$$V[u_t](1 - \rho^2) = \sigma_e^2 \Rightarrow V[u_t] = \frac{\sigma_e^2}{(1 - \rho^2)}$$

also

$$\text{Cov}[u_t, u_{t-1}] = E[u_t u_{t-1}] = E[(\rho u_{t-1} + e_t) u_{t-1}] = \rho V[u_{t-1}] = \rho \frac{\sigma_e^2}{(1 - \rho^2)}$$

or generally

$$\text{Cov}[u_t, u_{t-h}] = \rho^h \frac{\sigma_e^2}{(1 - \rho^2)}$$

How to Detect and Fix Serially Correlated Errors

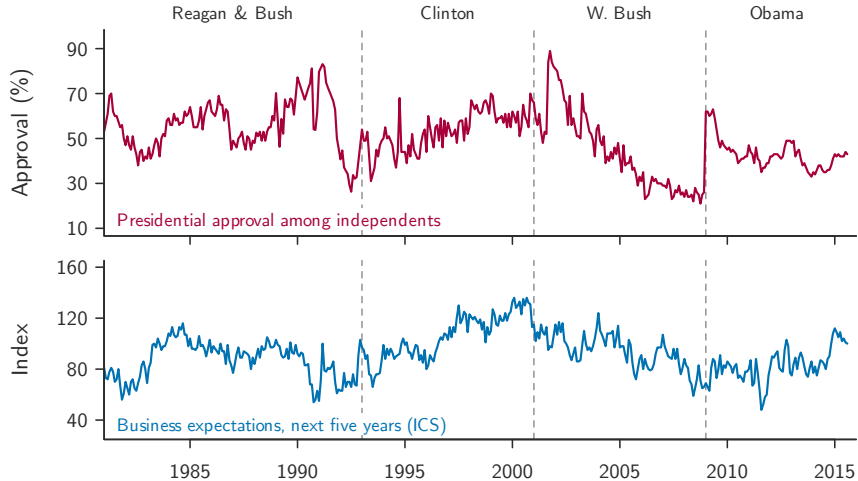
Detection:

- Plot residuals over time (or more fancy “autocorrelation” plots)
- Formal tests (e.g. Durbin-Watson statistics)

Possible Corrections:

- Use standard errors that are robust to serial correlation (e.g. Newey-West)
- AR corrections (e.g. Prais-Winsten, Cochrane-Orcutt, etc.)
- Lagged dependent variables or other dynamic panel models
- First-differencing the data

Monthly Approval and Business Expectations



Source: Donovan, Kellstedt, Key & Lebo (2020), *Political Behavior* 42(4); monthly, Jan. 1981–Aug. 2015. Dashed lines: a new administration.

Monthly Approval and Business Expectations

R Code

```
> mod1 <- lm(ind ~ bus5yr + clinton + gwb + obama, data=dta)
> coeftest(mod1)
```

t test of coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	27.615771	3.319727	8.3187	1.323e-15	***
bus5yr	0.324423	0.036757	8.8262	< 2.2e-16	***
clinton	-8.299574	1.597577	-5.1951	3.228e-07	***
gwb	-14.148390	1.465549	-9.6540	< 2.2e-16	***
obama	-11.392846	1.531328	-7.4398	5.943e-13	***

ind: approval among independents (%); bus5yr: expected business conditions over the next five years (Index of Consumer Sentiment);
clinton, gwb, obama: administration dummies (Reagan and G. H. W. Bush are the baseline).

Tests for Serial Correlation: Durbin-Watson

Recall our AR(1) model is:

$$y_t = \beta_0 + \beta_1 x_t + u_t$$

where $u_t = \rho u_{t-1} + e_t$, $e_t \sim N(0, \sigma_e^2)$, and ρ is our unknown **autoregressive coefficient** (with $|\rho| < 1$).

The null hypothesis (no serial correlation) is: $H_0: \rho = 0$

The alternative (positive serial correlation): $H_1: \rho > 0$

One common test for serial correlation is the **Durbin-Watson statistic**

(Optional) Tests for Serial Correlation: Durbin-Watson

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

$$DW - 2(1 - \hat{\rho}) = -\frac{\hat{u}_1^2 + \hat{u}_T^2}{\sum_{t=1}^T \hat{u}_t^2} \xrightarrow{p} 0,$$

where $\hat{\rho} = \frac{\sum_{t=2}^T \hat{u}_t \hat{u}_{t-1}}{\sum_{t=1}^T \hat{u}_t^2}$ is the first-order sample autocorrelation of the residuals. Since $\hat{\rho} \xrightarrow{p} \rho$, the Durbin–Watson statistic converges in probability to $2(1 - \rho)$ as $T \rightarrow \infty$.

Rules of thumb (the test's exact critical values depend on T and k):

- If $DW \approx 2$ then $\hat{\rho} \approx 0$
- If $DW < 1$ we have serious positive serial correlation
- If $DW > 3$ we have serious negative serial correlation

R Code

```
> library(lmtest)
> dwtest(mod1)

Durbin-Watson test

data:  mod1
DW = 0.32152, p-value < 2.2e-16
alternative hypothesis: true autocorrelation is greater than 0
```

The test rejects no serial correlation: the residuals are strongly positively autocorrelated ($\hat{\rho} \approx 0.84$). Business expectations are also strongly autocorrelated (0.90 from one month to the next), so the conventional standard errors are likely far too small.

Corrections: HAC Standard Errors

- A common way to correct for serial correlation is to use OLS but to estimate the variances using an estimator that is **heteroskedasticity and autocorrelation consistent (HAC)** (Newey and West 1987).
- The theory behind the HAC variance estimator is somewhat complicated, but the interpretation is similar to our usual OLS robust standard errors.
 - HAC standard errors leave estimate of $\hat{\beta}$ unchanged and do not fix potential bias in $\hat{\beta}$
 - HAC estimators are consistent for $V[\hat{\beta}]$ under heteroskedasticity and/or autocorrelation as T grows, if the number of lags used grows slowly with T
 - The `sandwich` package in R implements a variety of HAC estimators
 - A common option is `NeweyWest`; by default it chooses the number of lags from the data and applies a prewhitening step to the regressor-times-residual terms (see `?NeweyWest`). Without that step, the slope's SE on the next slide would be 0.075 instead of 0.079

```
> coeftest(mod1) # homoskedastic errors
t test of coefficients:
      Estimate Std. Error t value Pr(>|t|)
(Intercept)  27.615771   3.319727  8.3187 1.323e-15 ***
bus5yr        0.324423   0.036757  8.8262 < 2.2e-16 ***
clinton      -8.299574   1.597577 -5.1951 3.228e-07 ***
gwb          -14.148390   1.465549 -9.6540 < 2.2e-16 ***
obama        -11.392846   1.531328 -7.4398 5.943e-13 ***

> coeftest(mod1, vcov = NeweyWest) # HAC errors
t test of coefficients:
      Estimate Std. Error t value Pr(>|t|)
(Intercept)  27.615771   8.791195  3.1413 0.001804 **
bus5yr        0.324423   0.078904  4.1116 4.747e-05 ***
clinton      -8.299574   2.927949 -2.8346 0.004815 **
gwb          -14.148390   6.244890 -2.2656 0.023996 *
obama        -11.392846   4.647416 -2.4514 0.014644 *
```

Once we correct for autocorrelation, the slope's standard error is 2.1 times as large (0.079 instead of 0.037). The slope is still clearly positive ($t = 4.1$).

First Differences: Is the Link Month to Month?

- HAC fixes the standard errors, not a spurious trend. Two series that drift slowly together give a large slope in levels even if their month-to-month changes are unrelated.
- First differences check whether the link is month to month. We regress the change in approval on the change in business expectations.

R Code

```
> mod_fd <- lm(d_ind ~ d_bus + d_clinton + d_gwb + d_obama, data = fd)
> coeftest(mod_fd)
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-0.130212	0.277004	-0.4701	0.63855
d_bus	0.077375	0.036214	2.1366	0.03322 *
...				

- The slope shrinks from 0.32 to 0.077. With the usual SE, $t = 2.1$. With the Newey-West SE (computed without prewhitening, which fails for this model), $t = 1.9$, just short of the 5% cutoff (1.96). So the month-to-month link is positive but small, and not clearly different from zero. It is much weaker than the slope in levels. The data do not tell us why: the two series may drift together for other reasons, approval may take more than a month to respond, or monthly changes in a survey index may be mostly noise, which pulls the slope toward zero.

Heteroskedasticity

Clustering

Serial Correlation

Summary

Summary

- Non-constant error variance and correlated errors can make conventional standard errors badly wrong, even if the estimates of the regression coefficients are unbiased
 - Heteroskedasticity
 - Clustering
 - Serial correlation
- Both can be detected with plots or formal tests.
- A common approach is simply to always use robust standard errors unless n or the number of clusters M is too small
 - White standard errors
 - Cluster robust standard errors
 - Newey-West HAC standard errors