Week 7: Standard Errors

Marc Meredith*

*Statistical Methods for Data Science

July 10, 2019

- Last week:
  - Developed the concept of a multivariate regression
  - Discussed the Gauss-Markov Theorem, which says that OLS is the best linear unbiased estimator when five assumptions hold
  - Detailed the full rank and function form assumptions of the Gauss-Markov Theorem
- This week:
  - How to estimate OLS standard errors
  - Detail the homoscadasticity and no autocorrelation assumptions of the Gauss-Markov Theorem
  - Present alternative ways to estimate standard errors when these assumption are unlikely to hold
  - Discuss considerations in how you present regression analysis

- Five assumptions about our data
  1. X has full rank
  2. The true model that generates our data is $Y_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik} + \epsilon_i$
  3. $\underline{E[\epsilon_i^2 \mid X] = \sigma^2}$
  4. $\underline{E[\epsilon_i \epsilon_j \mid X] = 0 \text{ if } i \neq j}$
  5. $E[\epsilon_i \mid X] = 0$

- Assumptions #3 and #4 are assumptions about the central tendencies of unobserved determinants, ϵ, of the dependent variable
- Whether or not these assumptions are proper affect how we should be calculating how much sampling error we expect to observe in regression coefficients
  - We often will underestimate the probability that we experience a large amount of sampling error when we improperly make these assumptions
  - Making us overly confident about what we can conclude about the underlying parameters from regression coefficients
- The two broad themes of the week are:
  1. How to assess whether it is proper to make these assumptions
  2. What you do when you decide that it isn't proper

Agenda:

- What are the standard errors of regression coefficients
- How to calculate standard errors when you think assumptions #3 and #4 are proper
- How to calculate standard errors when you do not think assumption #3 is proper
- How to calculate standard errors when you do not think assumption #4 is proper
- How to present regression results to communicate information about uncertainty

- Last week we showed that $\hat{\beta} = (X^T X)^{-1} X^T Y$, whe
    - $\hat{\beta}$ is our estimate of what the value of the parameter (i.e., a fixed number) β
    - Next week will focus on how to think about the value of β
- As with any estimate, sampling error likely causes $\hat{\beta} - \beta \neq 0$ in any sample size $n < \infty$
    - Even if $E[\hat{\beta} - \beta] = 0$
- The variance of a regression coefficient, $\sigma_{\hat{\beta}_j}{}^2$, summarizes the amount of variability we expect to observe in $\hat{\beta}_j - \beta_j$
    - The higher the variance, the greater the probability that β_j and $\hat{\beta}_j$ differ by more than some fixed value

- In week four, we talked about constructing confidence intervals for parameters on the basis the information that we observe in a sample
    - E.g., using the value of $\hat{\beta}_j$ to inform us about the likelihood that β_j lies in some range
- We established that we typically need to know the sampling distribution of an estimate to be able to say anything meaningful about the likelihood a parameter is contained in a range
- Thankfully, it is often the case that $\hat{\beta}_j \sim N(\beta_j, {\sigma_{\hat{\beta}_j}}^2)$, as it allows us to construct a confidence interval on β_j using information on $\hat{\beta}_j$ and ${\sigma_{\hat{\beta}_j}}^2$ using methods that we have previously discussed
    - Occurs either when $\epsilon \sim N(0, \sigma^2)$ or the sample is large enough to apply the Central Limit Theorem

- Because we rarely know $\sigma_{\hat{\beta}_j}$ but not β, we usually have to estimate $\sigma_{\hat{\beta}_j}$ to be able to construct a confidence interval for β
- The short hand $\hat{\sigma}_{\beta_j}$ to refer to our estimate of the $\sigma_{\hat{\beta}_j}$
- If:
  1. $\hat{\beta}_k \sim N(\beta_j, \sigma_{\hat{\beta}_j}{}^2)$
  2. $\hat{\sigma}_{\beta_j}$ is an unbiased estimate of $\sigma_{\hat{\beta}_j}$
- Then $\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\hat{\beta}_j}} \sim T_{n-k-1}$
  - Allows us to construct a confidence interval on β_j using information on $\hat{\beta}_j$ and $\hat{\sigma}_{\hat{\beta}_j}$ using methods that we have previously discussed

- $\hat{\sigma}_{\hat{\beta}_j}$ is referred to as the standard error of $\hat{\beta}_j$
- Standard errors need to accurately represent the amount potential sampling error to get accurate confidence intervals on parameters
  - When standard errors underestimate the amount potential sampling error, our confidence intervals will be too small
  - When standard errors overestimate the amount potential sampling error, our confidence intervals will be too small
- Thus, we are going to spend most of this class discussing:
  - Reasons why the assumptions required to justify OLS standard errors many not be proper
  - Alternative methods for estimating standard errors when these assumptions are not proper

The Gauss-Markov assumptions that affect how we estimate standard errors:

- Assumption #3: $E[\epsilon_i^2 \mid X] = \sigma^2$
  - Because $E[\epsilon_i^2 \mid X] = 0 \implies var(\epsilon_i \mid X) = \sigma^2$
  - This assumption implies that the variance of unobservables is independent of the observables
    - This is called homoscedasticity
    - In contrast, hetroscedasticity implies that the variance is related to the observables
- Assumption #4: $E[\epsilon_i \epsilon_j \mid X] = 0$ if $i \neq j$
  - This is called no autocorrelation
    - Errors are positively autocorrelated when $E[\epsilon_i \epsilon_j \mid X] > 0$ and negatively autocorrelated when $E[\epsilon_i \epsilon_j \mid X] < 0$
  - This assumption means that the level of the unobserved determinants of obs. i do not inform us about the expected level of the unobserved determinants of obs. j

- To give intuition about standard errors, I first present the standard error for a regression coefficient in the simplest situation:
  - Slope in bivariate regression
  - All five assumptions of the Gauss-Markov Theorem are proper
- In such circumstances, $\hat{\sigma}_{\hat{\beta}} = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{(n-2)\sum_{i=1}^{n}(X_i - \bar{X})^2}}$

- Why does $\hat{\sigma}_{\hat{\beta}} = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{(n-2)\sum_{i=1}^{n}(X_i - \bar{X})^2}}$?
- I show that:
  1. $\hat{\sigma}^2_{\hat{\beta}} = E[\frac{c\hat{ov}(X,\epsilon)}{v\hat{ar}(X)}^2]$
  2. Assumption #3 implies that $E[\frac{c\hat{ov}(X,\epsilon)}{v\hat{ar}(X)}^2] = \frac{\sigma_\epsilon{}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$
  3. Assumption #3 and #4 imply that $\sigma_\epsilon{}^2 = \frac{E[\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2]}{n-2}$
- Proof uses the Law of Iterated Expectation
  - $E[\epsilon] = E_x[\epsilon \mid X]$
    - I.e., the expected value of r.v. ϵ can be calculated by taking a weighted average of conditional expectation of ϵ given X

Showing $\hat{\sigma}^2_{\hat{\beta}} = E[\frac{\hat{cov}(X,\epsilon)}{\hat{var}(X)}^2]$:

- $\hat{\sigma}^2_{\hat{\beta}} =$

  $E[(\hat{\beta} - E[\hat{\beta}])^2] =$

  $E[(\frac{\hat{cov}(X,Y)}{\hat{var}(X)} - \beta)^2] =$

  - We derived $\hat{\beta} = \frac{\hat{cov}(X,Y)}{\hat{var}(X)}$ in week four
  - Gauss-Markov says $E[\hat{\beta}] = \beta$

  $E[(\frac{\hat{cov}(X,\alpha+\beta X+\epsilon)}{\hat{var}(X)} - \beta)^2] =$

  $E[(\frac{\hat{cov}(X,\alpha)+\hat{cov}(X,\beta X)+\hat{cov}(X,\epsilon)}{\hat{var}(X)} - \beta)^2] =$

  $E[(\frac{0+\beta\hat{var}(X)+\hat{cov}(X,\epsilon)}{\hat{var}(X)} - \beta)^2] = E[\frac{\hat{cov}(X,\epsilon)}{\hat{var}(X)}^2]$

Showing $E[\frac{\hat{cov}(X,\epsilon)}{\hat{var}(X)}^2] = \frac{\sigma_\epsilon{}^2}{\sum_{i=1}^n (X_i - \bar{X})^2}$:

- $E[\frac{\hat{cov}(X,\epsilon)}{\hat{var}(X)}^2] =$

  $E[\frac{\frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})(\epsilon_i - \bar{\epsilon})}{\frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2}^2] =$

  $E_x[\frac{\sum_{i=1}^n (X_i - \bar{X})^2 E[(\epsilon_i - \bar{\epsilon})^2 | X]}{\sum_{i=1}^n (X_i - \bar{X})^4}] =$

  $E_x[\frac{\sigma_\epsilon{}^2 \sum_{i=1}^n (X_i - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^4}] = \frac{\sigma_\epsilon{}^2}{\sum_{i=1}^n (X_i - \bar{X})^2}$

  - Assumption #3 of Gauss Markov is that $E[(\epsilon_i - \bar{\epsilon})^2 \mid X] = \sigma_\epsilon{}^2$ (e.g., homoscadasticity)

Show $\sigma_\epsilon{}^2 = \frac{E[\sum_{i=1}^n (Y_i - \hat{Y}_i)^2]}{n-2}$:

1. We know $Y_i - \hat{Y}_i = (Y_i - \bar{Y}) - \hat{\beta}(X_i - \bar{X})$
   - Derived from combining:
     - $\bar{Y} = \hat{\alpha} + \hat{\beta}\bar{X}$
     - $Y_i - \hat{Y}_i = Y_i - \hat{\alpha} - \hat{\beta}X_i$
2. We also know that $Y_i - \bar{Y} = \beta(X_i - \bar{X}) + (\epsilon_i - \bar{\epsilon})$
   - Derived from combining:
     - $Y_i = \alpha + \beta X_i + \epsilon_i$
     - $\bar{Y} = \alpha + \beta\bar{X} + \bar{\epsilon}$

Show ${\sigma_\epsilon}^2 = \frac{E[\sum_{i=1}^n (Y_i - \hat{Y}_i)^2]}{n-2}$ (continued):

- Combining 1 and 2 on the previous slide imply that $Y_i - \hat{Y}_i = (\beta - \hat{\beta})(X_i - \bar{X}) + (\epsilon_i - \bar{\epsilon}) \implies$ $\sum_{i=1}^n (Y_i - \hat{Y}_i)^2 =$ $\sum_{i=1}^n (\beta - \hat{\beta})^2 (X_i - \bar{X})^2 + 2(\beta - \hat{\beta})(X_i - \bar{X})(\epsilon_i - \bar{\epsilon}) + (\epsilon_i - \bar{\epsilon})^2$
  - $\sum_{i=1}^n (Y_i - \hat{Y}_i)^2 =$ $\sum_{i=1}^n ((\beta - \hat{\beta})(X_i - \bar{X}) + (\epsilon_i - \bar{\epsilon}))^2 =$ $\sum_{i=1}^n (\beta - \hat{\beta})^2 (X_i - \bar{X})^2 + 2(\beta - \hat{\beta})(X_i - \bar{X})(\epsilon_i - \bar{\epsilon}) + (\epsilon_i - \bar{\epsilon})^2$
- The next two slides show $E[\sum_{i=1}^n (Y_i - \hat{Y}_i)^2] = \sum_{i=1}^n (X_i - \bar{X})^2 E[(\beta - \hat{\beta})^2] +$ $2\sum_{i=1}^n E[(\beta - \hat{\beta})(X_i - \bar{X})(\epsilon_i - \bar{\epsilon})] + \sum_{i=1}^n E[(\epsilon_i - \bar{\epsilon})]^2 =$ ${\sigma_\epsilon}^2 - 2{\sigma_\epsilon}^2 + (n-1){\sigma_\epsilon}^2 = (n-2){\sigma_\epsilon}^2$

Show ${\sigma_\epsilon}^2 = \frac{E[\sum_{i=1}^n (Y_i - \hat{Y}_i)^2]}{n-2}$ (continued):

- $\sum_{i=1}^n (X_i - \bar{X})^2 E[(\beta - \hat{\beta})^2] = \sum_{i=1}^n (X_i - \bar{X})^2 var(\hat{\beta}) = \sum_{i=1}^n (X_i - \bar{X})^2 \frac{{\sigma_\epsilon}^2}{\sum_{i=1}^n (X_i - \bar{X})^2} = {\sigma_\epsilon}^2$
- $\sum_{i=1}^n E[(\epsilon_i - \bar{\epsilon})^2] = \sum_{i=1}^n E[{\epsilon_i}^2] - 2E[\epsilon_i \bar{\epsilon}] + E[\bar{\epsilon}^2] = n{\sigma_\epsilon}^2 - 2{\sigma_\epsilon}^2 + {\sigma_\epsilon}^2 = (n-1){\sigma_\epsilon}^2$
  - Assumption #4 of Gauss Markov is applied here (e.g., no autocorrelation)

Show ${\sigma_\epsilon}^2 = \frac{E[\sum_{i=1}^n (Y_i - \hat{Y}_i)^2]}{n-2}$ (continued):

- $2\sum_{i=1}^n E[(\beta - \hat{\beta})(X_i - \bar{X})(\epsilon_i - \bar{\epsilon})] = -2{\sigma_\epsilon}^2$
  - $\sum_{i=1}^n E[(\beta - \hat{\beta})(X_i - \bar{X})(\epsilon_i - \bar{\epsilon})] =$
    $\sum_{i=1}^n E[\frac{-\hat{cov}(X,\epsilon)}{\hat{var}(X)}(X_i - \bar{X})(\epsilon_i - \bar{\epsilon})] =$
    $E[\frac{-\hat{cov}(X,\epsilon)}{\hat{var}(X)}\sum_{i=1}^n (X_i - \bar{X})(\epsilon_i - \bar{\epsilon})] =$
    $E[\frac{-(\sum_{i=1}^n (X_i - \bar{X})(\epsilon_i - \bar{\epsilon}))^2}{\sum_{i=1}^n (X_i - \bar{X})^2}] =$
    $E_x[\frac{-\sum_{i=1}^n (X_i - \bar{X})^2 E[(\epsilon_i - \bar{\epsilon})^2 | X]}{\sum_{i=1}^n (X_i - \bar{X})^2}] = -{\sigma_\epsilon}^2$
    - Assumption #3 of Gauss Markov is that $E[(\epsilon_i - \bar{\epsilon})^2 \mid X] = {\sigma_\epsilon}^2$ (e.g., homoscadasticity)

- What are the implications of $\hat{\sigma}_{\hat{\beta}} = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{(n-2)\sum_{i=1}^{n}(X_i - \bar{X})^2}}$?
- Uncertainty about β decreases as
  - The model fits better (i.e., lower $\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$)
  - The sample size gets bigger (i.e., higher n)
  - There is more variability in X (i.e., higher $\sum_{i=1}^{n}(X_i - \bar{X})^2$)
- Even when we calculate standard errors in some other way or are running multivariate regressions, these properties usually remain true

- We can generalize a formula for the variance for a vector $\hat{\beta} = (X^TX)^{-1}X^TY$ in a multivariate OLS regression with k independent variables
- $var(\hat{\beta}) =$
  $E[(\hat{\beta} - E[\hat{\beta}])(\hat{\beta} - E[\hat{\beta}])^T] =$
  $E[(X^TX)^{-1}X^T\epsilon\epsilon^TX(X^TX)^{-1}] =$
  - $\hat{\beta} = (X^TX)^{-1}X^TY = (X^TX)^{-1}X^T(X\beta + \epsilon). = \beta. + (X^TX)^{-1}X^T\epsilon$
  - $E[\hat{\beta}] = \beta$
  - So $\hat{\beta} - E[\hat{\beta}] = .(X^TX)^{-1}X^T\epsilon$

  $E_X[(X^TX)^{-1}X^TE[\epsilon\epsilon^T \mid X]X(X^TX)^{-1}]$

- What is $E[\epsilon\epsilon^T \mid X]$?

$$E[\begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \dots \\ \epsilon_n \end{pmatrix} \begin{pmatrix} \epsilon_1 \epsilon_2 \dots \epsilon_n \end{pmatrix} \mid X] =$$

$$E\begin{pmatrix} \epsilon_1\epsilon_1 & \epsilon_1\epsilon_2 & \dots & \epsilon_1\epsilon_n \\ \epsilon_2\epsilon_1 & \epsilon_2\epsilon_2 & \dots & \epsilon_2\epsilon_n \\ \dots & \dots & \dots & \dots \\ \epsilon_n\epsilon_1 & \epsilon_n\epsilon_2 & \dots & \epsilon_n\epsilon_n \end{pmatrix} \mid X) =$$

$$\begin{pmatrix} E[\epsilon_1\epsilon_1 \mid X] & E[\epsilon_1\epsilon_2 \mid X] & \dots & E[\epsilon_1\epsilon_n \mid X] \\ E[\epsilon_2\epsilon_1 \mid X] & E[\epsilon_2\epsilon_2 \mid X] & \dots & E[\epsilon_2\epsilon_n \mid X] \\ \dots & \dots & \dots & \dots \\ E[\epsilon_n\epsilon_1 \mid X] & E[\epsilon_n\epsilon_2 \mid X] & \dots & E[\epsilon_n\epsilon_n \mid X] \end{pmatrix}$$

- Without assumptions #3 and #4, $E[\epsilon_i \epsilon_j \mid X] = \sigma^2 w_{ij} \implies$
  $E[\epsilon_i \epsilon_j \mid X] =$
  $$\begin{pmatrix} E[\epsilon_1\epsilon_1 \mid X] & E[\epsilon_1\epsilon_2 \mid X] & \dots & E[\epsilon_1\epsilon_n \mid X] \\ E[\epsilon_2\epsilon_1 \mid X] & E[\epsilon_2\epsilon_2 \mid X] & \dots & E[\epsilon_2\epsilon_n \mid X] \\ \dots & \dots & \dots & \dots \\ E[\epsilon_n\epsilon_1 \mid X] & E[\epsilon_n\epsilon_2 \mid X] & \dots & E[\epsilon_n\epsilon_n \mid X] \end{pmatrix} =$$
  $$\sigma^2 \begin{pmatrix} w_{11} & w_{12} & \dots & w_{1n} \\ w_{21} & w_{22} & \dots & w_{2n} \\ \dots & \dots & \dots & \dots \\ w_{n1} & w_{n2} & \dots & w_{nn} \end{pmatrix} = \sigma^2 \Omega$$
  - Note that when $i \neq j$, $E[\epsilon_i \epsilon_j \mid X] = E[\epsilon_j \epsilon_i \mid X] \implies w_{ij} = w_{ji}$

- How should we think about w_{ij}?
  - When $i = j$
    - The relative variability of the unobserved determinants of observation i
    - $w_{ii} > w_{jj}$ means that we expect more error for observation i than observation j
  - When $i \neq j$
    - The relative covariation of the unobserved determinants of observation i and observation j
    - $w_{ij} > 0$ means that we expect the error for observation i and observation j to positively associate
    - $w_{ij} < 0$ means that we expect the error for observation i and observation j to negatively associate

- Given that $var(\hat{\beta}) = \sigma^2(X^TX)^{-1}X^T\Omega X(X^TX)^{-1}$, how can we solve for Ω?
  1. Make assumptions about the values of w_{ij}
     - When we make assumption #3 we are assuming that $w_{ii} = w_{jj}$ for all i and j
     - When we make assumption #4 we are assuming that $w_{ij} = 0$ for all $i \neq j$
  2. Use data to estimate the values of w_{ij}
     - But e_ie_j is the only piece of data we have with which to estimate w_{ij}
     - So we cannot apply something like the Law of Large Numbers to estimate w_{ij} no matter how many observations we have in our dataset (i.e., the dimensionality problem)

- Assumptions #3 and #4 mean that $E[\epsilon\epsilon^T \mid X] = \sigma^2\Omega = \sigma^2 I_n$

$$I_n = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1 \end{pmatrix}$$

- When this happens then $var(\hat{\beta}) = \sigma^2(X^TX)^{-1}$
  - $E_X[(X^TX)^{-1}X^TE[\epsilon\epsilon^T \mid X]X(X^TX)^{-1}]. = E_X[(X^TX)^{-1}X^T\sigma^2 I_n X(X^TX)^{-1}] = \sigma^2(X^TX)^{-1}$
  - Where we can estimate $\hat{\sigma}^2 = \frac{1}{n-k-1}(Y_i - \hat{Y}_i)^2$

```
> reg3 <- lm(obamaapp ~ cats + partisanship, data = mydata)
> summary(reg3)

Call:
lm(formula = obamaapp ~ cats + partisanship, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max
-0.8698 -0.2019 -0.1228  0.2092  0.8772

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.45682    0.01685   27.11   <2e-16 ***
cats          0.07904    0.03576    2.21   0.0274 *
partisanship  0.33398    0.01822   18.34   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.4161 on 783 degrees of freedom
  (235 observations deleted due to missingness)
Multiple R-squared:  0.3099,       Adjusted R-squared:  0.3082
F-statistic: 175.8 on 2 and 783 DF,  p-value: < 2.2e-16
```

```
> temp <- vcov(reg3)
> print(temp)
               (Intercept)          cats  partisanship
(Intercept)   2.839481e-04 -2.828664e-04 -5.439204e-06
cats         -2.828664e-04  1.278735e-03 -6.054205e-05
partisanship -5.439204e-06 -6.054205e-05  3.317914e-04
> print(temp[1, 1]^(1/2))
[1] 0.01685076
> print(temp[2, 2]^(1/2))
[1] 0.03575941
> print(temp[3, 3]^(1/2))
[1] 0.01821514
```

Quiz question: Which of these intervals approximately represents the 95% confidence interval on the true increase in Obama approval among people who prefer a cat relative to people who prefer dogs if we thought all five Gauss-Markov assumption were proper?

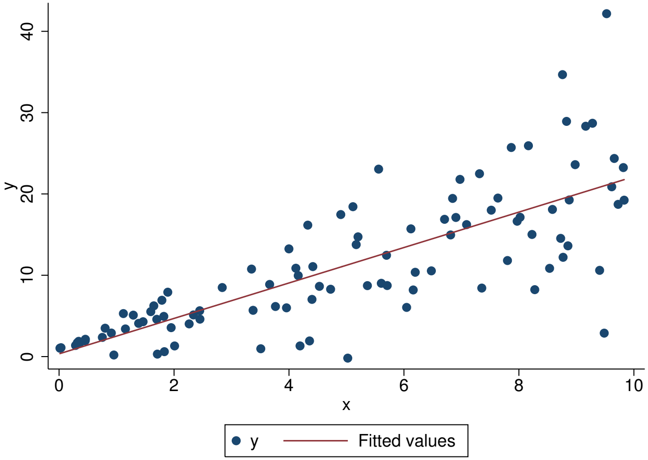
```
Call:
lm(formula = obamaapp ~ cats + partisanship, data = mydata)

Residuals:
    Min       1Q   Median       3Q      Max
-0.8698 -0.2019 -0.1228  0.2092  0.8772

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.45682    0.01685   27.11   <2e-16 ***
cats          0.07904    0.03576    2.21   0.0274 *
partisanship  0.33398    0.01822   18.34   <2e-16 ***
```

A. 4.3 and 11.5 percentage point increase
B. 2.5 and 13.3 percentage point increase
C. 0.8 and 15.1 percentage point increase
D. -2.8 and 18.7 percentage point increase

- What does it mean for $E[\epsilon_i^2 \mid X_i] \neq \sigma^2$?
- Example of where $E[\epsilon_i^2 \mid X_i]$ increasing in X_i:



When does $E[\epsilon_i^2 \mid X] \neq \sigma^2$?

1. We have a variable in our regression that affects how good our regression is at explaining the dependent variable

- Example:
    - Dependent variable is economic growth
    - And one of the explanatory variables is percent of GDP that comes from oil revenues (potentially interacted with oil prices)
    - Oil revenues will affect both GDP growth and the amount of prediction error in GDP growth

When does $E[\epsilon_i^2 \mid X] \neq \sigma^2$?

2. When we have a binary dependent variable

$$Y_i = \begin{cases} 1 & \text{with prob. } X_i\beta \\ 0 & \text{with prob. } 1 - X_i\beta \end{cases}$$

- So it is the case that:

$$\epsilon_i = \begin{cases} 1 - X_i\beta & \text{with prob. } X_i\beta \\ -X_i\beta & \text{with prob. } 1 - X_i\beta \end{cases}$$

- $E[{\epsilon_i}^2 \mid X_i] =$
$(1 - X_i\beta)^2 X_i\beta + (-X_i\beta)^2(1 - X_i\beta) =$
$(1 - X_i\beta)X_i\beta(1 - X_i\beta + X_i\beta) =$
$(1 - X_i\beta)X_i\beta$

When does $E[\epsilon_i^2 \mid X] \neq \sigma^2$?

3. When our dependent variable is systematically measured with more error in some cases than others

- Examples:
  - Precinct-level election data
  - County-level unemployment data
  - Wyoming / California in a nationally representative survey
- We think that our dependent variable will be measured more accurately when more individuals are aggregated together
  - E.g., We get a more accurate read of public opinion in California than in Wyoming on a nationally representative survey

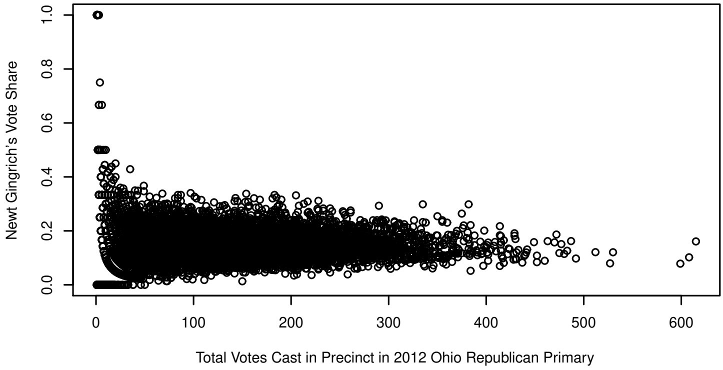
- Suppose:
  - We are unwilling to make assumption #3: $E[\epsilon_i^2 \mid X] = \sigma^2$
  - We are willing to make assumption #4: $E[\epsilon_i \epsilon_j \mid X] = 0$ if $i \neq j$
- In this case: $\sigma^2 \Omega = \sigma^2 \begin{pmatrix} w_{11} & 0 & \dots & 0 \\ 0 & w_{22} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & w_{nn} \end{pmatrix}$
  - We need to think of ways to account of heteroscadasticity that don't require us to estimate more parameters than observations in our dataset (i.e., the dimensionality problem)

- There are three general techniques we can use when there is heteroscadasticity:
  - Weighting
  - Huber-White (or robust) standard errors
  - Model it (beyond the scope of this course)
- Our goal with all three approaches is use the information from observations that we believe have less variance more than the information from observations that have more variance
  - An observation k with an extremely large or small ϵ_k is more likely to cause us to misestimate the relationship between Y and X
  - There is a greater chance that ϵ_j is extremely large or small than ϵ_i when $w_{jj} > w_{ii}$
- The choice of whether to use weighting or Huber-White standard errors to account for heteroscadasticity depends on whether we know the values of w_{ii}

- When we know the values of w_{ii}, we want to weight to account for heteroscadasticity
- Weighting means that assign a value v_i to each observation i that says how much relative impact it should have on the estimated regression coefficients
    - When $v_1 = 2$ and $v_2 = 1$ then observation 1 has twice the relative impact as observation 2 on our regression coefficients as it would if we used OLS
    - Note that this would also be true if $v_1 = 4$ and $v_2 = 2$
- The goal of weighting is set $v_i = \frac{1}{w_{ii}}$
    - E.g., If $w_{ii} = 3w_{jj}$ then $v_j = 3$ and $v_i = 1$

- Aggregated data is a common place where we know the w_{ii} for every observation i
- Suppose that our dependent variable $\bar{Y}_i$ is constructed by observing N_i observations from unit i in which:
  - $Y_{i,n} = \alpha + \beta X_i + \epsilon_{i,n}$
  - $\bar{Y}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} Y_{i,j}$
- Then $var(\bar{Y}_i) = \frac{\sigma_\epsilon{}^2}{N_i}$ when $\epsilon_{i,j}$ and $\epsilon_{i,k}$ are independent for all j and k
- Implication is that if $v_j = 1$ then $v_k = \frac{N_j}{N_k}$
  - Because $\frac{v_k}{v_j} = \frac{var(\bar{Y}_j)}{var(\bar{Y}_k)} = \frac{\frac{\sigma_\epsilon{}^2}{N_j}}{\frac{\sigma_\epsilon{}^2}{N_k}} = \frac{N_k}{N_j}$

- To highlight how we can use weighting to account for heteroscadasticity, I run a regression that examines how Newt Gingrich's ballot position affected his vote share in the 2012 Ohio Republican primary
  - Because it is thought that being listed first on the ballot causes a candidate to receive more votes, Ohio uses different ballot orderings in different precincts
- The dependent variable in this regression is Gingrich's vote share in a precinct and the independent variable is whether Gingrich was listed first on the ballot in this precinct
- I weight each precinct by the number of ballots cast in the precinct because I expect that the variance in Gingrich's vote share will decrease proportionately to the number of ballots cast


Newt Gingrich's Vote Share
0.0
0.2
0.4
0.6
0.8
1.0
0
100
200
300
400
500
600
Total Votes Cast in Precinct in 2012 Ohio Republican Primary

```
> # OLS bivariate regression
> output1 <- lm(voteshare ~ First, data = Gingrich)
> summary(output1)

Call:
lm(formula = voteshare ~ First, data = Gingrich)

Residuals:
     Min       1Q   Median       3Q      Max
-0.15079 -0.03894 -0.00534  0.03217  0.85439

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.145607   0.000853 170.706   <2e-16 ***
First       0.005184   0.002066   2.509   0.0121 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.07584 on 9528 degrees of freedom
Multiple R-squared:  0.0006601,        Adjusted R-squared:  0.0005552
F-statistic: 6.294 on 1 and 9528 DF,  p-value: 0.01213
```

```
> # Weighted regression
> output2 <- lm(voteshare ~ First, weight = tot_votes,  data = Gingrich)
> summary(output2)

Call:
lm(formula = voteshare ~ First, data = Gingrich, weights = tot_votes)

Weighted Residuals:
     Min       1Q   Median       3Q      Max
-1.93041 -0.35735 -0.05133  0.32497  2.99738

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.1450700  0.0005405 268.375  < 2e-16 ***
First       0.0055244  0.0013118   4.211 2.56e-05 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.5418 on 9528 degrees of freedom
Multiple R-squared:  0.001858,       Adjusted R-squared:  0.001753
F-statistic: 17.74 on 1 and 9528 DF,  p-value: 2.561e-05
```

- The regressions reported on the previous two slides show that
  - We can reduce uncertainty in our coefficient estimates by putting more weight on observations with less variance relative to observations with more variance
    - The standard error on the effect of being listed first on the ballot decreases from 0.21 to 0.14 when we weight observations in inverse proportion to the variance
  - Weighting affects the coefficients that we estimate, even when both the weighted and unweighted regressions produce unbiased estimates
    - The estimate of the increase in vote share from being listed first relative to not being listed first is 0.52 and 0.55 in the unweighted and weighted regression, respectively

- Heteroscadasticity is just one reason why you may want to run a weighted regression
- Another reason that you apply weights is to make your sample more representative of the underlying population
- While the syntax within R for applying these weights is exactly the same, it is important that you are clear in your mind about why you are using weights

```
> reg3 <- lm(obamaapp ~ cats + partisanship, data = mydata)
> summary(reg3)

Call:
lm(formula = obamaapp ~ cats + partisanship, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max
-0.8698 -0.2019 -0.1228  0.2092  0.8772

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.45682    0.01685   27.11   <2e-16 ***
cats          0.07904    0.03576    2.21   0.0274 *
partisanship  0.33398    0.01822   18.34   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.4161 on 783 degrees of freedom
  (235 observations deleted due to missingness)
Multiple R-squared:  0.3099,      Adjusted R-squared:  0.3082
F-statistic: 175.8 on 2 and 783 DF,  p-value: < 2.2e-16
```

```
> reg15 <- lm(obamaapp ~ cats + partisanship, weight = WGHT, data = mydata)
> summary(reg15)

Call:
lm(formula = obamaapp ~ cats + partisanship, data = mydata, weights = WGHT)

Weighted Residuals:
     Min       1Q   Median       3Q      Max
-1.69823 -0.22426 -0.05687  0.23344  1.76045

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.45287    0.01704  26.575   <2e-16 ***
cats          0.09960    0.04100   2.429   0.0154 *
partisanship  0.31307    0.01957  15.999   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.4329 on 783 degrees of freedom
  (235 observations deleted due to missingness)
Multiple R-squared:  0.2521,        Adjusted R-squared:  0.2502
F-statistic:   132 on 2 and 783 DF,  p-value: < 2.2e-16
```

- The previous two slides highlight patterns that often play out when weighting to make samples more representative of the population
  - Estimate similar, but not identical, coefficients in the weighted and unweighted regressions
  - Estimate slightly higher standard errors in the weighted regression than the unweighted regression
- When you observe different patterns, suggest that there may be some substantial heterogeneity in how one or more of your explanatory variables relates to the dependent variables
  - See our discussion last week about interaction terms

- We use Huber-White standard errors when we think there is heteroscadasticity in our data, but don't know the exact form
- Huber-White gets around the dimensionality problem we discussed earlier by estimating $\frac{X^T \Omega X}{n}$ instead of Ω
  - Because X is a $nx(k+1)$ matrix and Ω is a nxn matrix, this implies $\frac{X^T \Omega X}{n}$ is $(k+1)x(k.+1)$
- White-Huber (or robust) heteroscadastic consistent estimator:
$var(\hat{\beta}) = \frac{1}{n}(\frac{X^T X}{n})^{-1} \frac{\sum_{i=1}^{n} e_i^2 X_i^T X_i}{n-k} (\frac{X^T X}{n})^{-1}$

- To illustrate how Huber-White standard errors work consider the bivariate regression: $Y_i = \alpha + X_i\beta + \epsilon_i$
- In this case $X^T\sigma^2\Omega X =$

$$\begin{pmatrix} 1 & 1 & \dots & 1 \\ X_1 & X_2 & \dots & X_n \end{pmatrix} \begin{pmatrix} w_{11} & 0 & \dots & 0 \\ 0 & w_{22} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & w_{nn} \end{pmatrix} \begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ \dots & \dots \\ 1 & X_n \end{pmatrix}$$

$$\sigma^2 \begin{pmatrix} \sum_{i=1}^n w_{ii} & \sum_{i=1}^n w_{ii}X_i \\ \sum_{i=1}^n w_{ii}X_i & \sum_{i=1}^n w_{ii}X_i^2 \end{pmatrix}$$

- It can be shown that:
  1. $E[\sum_{i=1}^{n}(Y_i - X_i\hat{\beta})^2] = \sigma^2 \sum_{i=1}^{n} w_{ii}$
  2. $E[\sum_{i=1}^{n}(Y_i - X_i\hat{\beta})^2 X_i] = \sigma^2 \sum_{i=1}^{n} w_{ii} X_i$
  3. $E[\sum_{i=1}^{n}(Y_i - X_i\hat{\beta})^2 X_i^2] = \sigma^2 \sum_{i=1}^{n} w_{ii} X_i^2$
- Thus in the bivariate case, Huber-White standard errors estimate $X^T \hat{\sigma^2} \Omega X = \begin{pmatrix} \sum_{i=1}^{n}(Y_i - X_i\hat{\beta})^2 & \sum_{i=1}^{n}(Y_i - X_i\hat{\beta})^2 X_i \\ \sum_{i=1}^{n}(Y_i - X_i\hat{\beta})^2 X_i & \sum_{i=1}^{n}(Y_i - X_i\hat{\beta})^2 X_i^2 \end{pmatrix}$

- Proof that $E[\sum_{i=1}^{n}(Y_i - X_i\hat{\beta})^2] = \sigma^2 \sum_{i=1}^{n} w_{ii}$:
  - $E[\sum_{i=1}^{n}(Y_i - X_i\hat{\beta})^2 X_i] =$
    $E_X[\sum_{i=1}^{n} E[(Y_i - X_i\hat{\beta})^2 \mid X_i]X_i] =$
    $E_X[\sum_{i=1}^{n} \sigma^2 w_{ii} X_i] =$
    $\sigma^2 E_X[\sum_{i=1}^{n} w_{ii} X_i] = \sigma^2 \sum_{i=1}^{n} w_{ii} X_i$
- The proofs of 1 and 3 on the previous slide is the similar
- Similar logic holds if we expand upon this method for a regression with k explanatory variables

- It can be shown that when errors are heteroscadastic $plim_{n \to \infty} \sum_{i=1}^{n} \frac{1}{n} e_i^2 X_i^T X_i \to \frac{X^T \sigma^2 \Omega X}{n}$
- But sampling error will cause differences between $\sum_{i=1}^{n} \frac{1}{n} e_i^2 X_i^T X_i$ and $\frac{X^T \sigma^2 \Omega X}{n}$ within normally sized samples
- Implications for using Huber-White standard errors in practice
  - We need a large enough sample before we can apply robust standard errors
  - Multiple methods have been developed to account for the finite sample bias that generally produce similar, but not identical results

- Estimating Huber-White standard errors in R is straightforward
- To implement:
  1. Run an OLS regression in R using the lm() function
  2. Use the vcovHC() function to estimate the Huber-White standard errors
     - Part of the sandwich library
     - Requires specifying a method for accounting for small sample biases
  3. Use the coeftest() to output the coefficients and standard errors
     - Part of the lmtest library

```
> reg3 <- lm(obamaapp ~ cats + partisanship, data = mydata)
> summary(reg3)

Call:
lm(formula = obamaapp ~ cats + partisanship, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max
-0.8698 -0.2019 -0.1228  0.2092  0.8772

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.45682    0.01685   27.11   <2e-16 ***
cats          0.07904    0.03576    2.21   0.0274 *
partisanship  0.33398    0.01822   18.34   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.4161 on 783 degrees of freedom
  (235 observations deleted due to missingness)
Multiple R-squared:  0.3099,        Adjusted R-squared:  0.3082
F-statistic: 175.8 on 2 and 783 DF,  p-value: < 2.2e-16
```

```
> reg15 <- lm(obamaapp ~ cats + partisanship, weight = WGHT, data = mydata)
> summary(reg15)

Call:
lm(formula = obamaapp ~ cats + partisanship, data = mydata, weights = WGHT)

Weighted Residuals:
     Min       1Q   Median       3Q      Max 
-1.69823 -0.22426 -0.05687  0.23344  1.76045 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)   0.45287    0.01704  26.575   <2e-16 ***
cats          0.09960    0.04100   2.429   0.0154 *  
partisanship  0.31307    0.01957  15.999   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1 

Residual standard error: 0.4329 on 783 degrees of freedom
  (235 observations deleted due to missingness)
Multiple R-squared:  0.2521,Adjusted R-squared:  0.2502 
F-statistic:   132 on 2 and 783 DF,  p-value: < 2.2e-16
```

```
> library(sandwich)
> library(lmtest)
> coeftest(reg15, vcov = vcovHC(reg15, type="HC1"))

t test of coefficients:

              Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.452869   0.025943 17.4564  < 2e-16 ***
cats          0.099600   0.047704  2.0879  0.03713 *
partisanship  0.313073   0.022060 14.1918  < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

> coeftest(reg15, vcov = vcovHC(reg15, type="HC3"))

t test of coefficients:

              Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.452869   0.026142 17.3233  < 2e-16 ***
cats          0.099600   0.048242  2.0646  0.03929 *
partisanship  0.313073   0.022224 14.0869  < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Quiz question: Which of these statements summarizes how you should be thinking about heteroscadasticity?

A. Heteroscadasticity is not much of a problem, particularly when we have a sizable sample, because the solutions to deal it with are pretty straightforward.
B. We generally lack a solution to account for heteroscadasticity, so we cannot really use regression analysis when we suspect it is present
C. We should adjust how we run regressions to account for heteroscadasticity, but it is not a big deal if we don't because it doesn't affect out ability to get unbiased estimates of parameters.
D. While in theory we should adjust how we run regressions to account for heteroscadasticity, the solutions to deal with it are so complex that we should just stick with OLS most of the time.

Final comments on heteroscadisticity:

- Accounting for heteroscadsticity is relatively straightforward
  - If we have a large enough sample
- Consequences are fairly predictable
  - Usually makes our standard errors get a little bit bigger
  - Unless weighting is allowing us to use our information more effectively
- Caution should be taken when correcting for heteroscadascitiy leads to substantial, unexpected change in results

- Autocorrelation occurs when we think that there are common unobservable variables that affect two or more observations in our dataset
  - E.g., $E[\epsilon_i \epsilon_j \mid X] \neq 0$
- Some reasons why this could happen:
  1. When multiple observations necessarily share the same explanatory variable by construction
  2. When we observe the same unit-of-observations over time (i.e., panel data or time-series data)
- I'll present a strategy that can be used to get more accurate estimates of the uncertainty in regression coefficients in each of these circumstances

- Sometimes two or more observations in a dataset share the same explanatory variable, X_j, by construction
    - X_j is the unemployment rate in the state and you have multiple observations from the same state
    - X_j is the mother's education and you have multiple observations from the same family
- Autocorrelation is present in such circumstances because
    - Any effect of $\hat{\beta}_j - \beta_j$ will necessarily affect these observations in the same way
    - Which is compounded if there are other (unmodeled) determinents that these observations also share in common
- Failure to account for this autocorrelation can cause substantial underestimation of the amount of uncertainty in coefficient estimates

- To highlight the issues of common explanatory variables we'll analyze some of the data generated by the Tennessee Star Experiment
  - An experiment in which students were randomly assigned to be in a smaller or larger kindergarten classroom
- To explore whether students learn better when the classroom is small, we'll run a regression in which the dependent variable is a kid's score on a standardized test and the key explanatory variable is an indicator for whether the kid was randomly assigned to small classroom
  - All kids in the same classroom will have the same value for this explanatory variable

```
> # OLS regression
> output1 <- lm(gktreadss ~ small + poor + female, data = classsizedata2)
> summary(output1)

Call:
lm(formula = gktreadss ~ small + poor + female, data = classsizedata2)

Residuals:
    Min      1Q  Median      3Q     Max 
-47.017 -11.942  -1.352  10.061  53.135 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  45.3519     0.3936 115.230  < 2e-16 ***
smallTRUE     3.0773     0.4719   6.521 7.56e-11 ***
poorTRUE     -9.4871     0.4329 -21.913  < 2e-16 ***
femaleTRUE    3.5875     0.4328   8.289  < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 16.43 on 5767 degrees of freedom
Multiple R-squared:  0.09297,	Adjusted R-squared:  0.0925 
F-statistic:   197 on 3 and 5767 DF,  p-value: < 2.2e-16
```

- My expectation is that if kid i and kid j were in the same classroom then $E[\epsilon_i \epsilon_j \mid X] > 0$
  - Either because both are positive or both are negative
- Kids in the same classroom will have a similar unobservable experience that affects how much they learn because of:
  - Quality of the instruction
  - Temperature in the classroom
  - Disruptive kids

- Clustered standard errors are a statistical technique that we can use to get more accurate measures of uncertainty when analyzing data in which some observations share a common independent variable by construction
  - Logic is quite similar to the Huber-White errors
- The two most important thing to know are that:
  1. It only changes the standard errors, not the coefficients, from an OLS regression
  2. It is applying asymptotics w.r.t. the number of clusters (rather than the number of observations)

```
> # OLS regression
> output1 <- lm(gktreadss ~ small + poor + female, data = classsizedata2)
> olsvariance <- vcov(output1)
> olsstderrors <- sqrt(diag(olsvariance))
> summary(output1)

Call:
lm(formula = gktreadss ~ small + poor + female, data = classsizedata2)

Residuals:
    Min      1Q  Median      3Q     Max
-47.017 -11.942  -1.352  10.061  53.135

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  45.3519     0.3936 115.230  < 2e-16 ***
smallTRUE     3.0773     0.4719   6.521 7.56e-11 ***
poorTRUE     -9.4871     0.4329 -21.913  < 2e-16 ***
femaleTRUE    3.5875     0.4328   8.289  < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 16.43 on 5767 degrees of freedom
Multiple R-squared:  0.09297,       Adjusted R-squared:  0.0925
F-statistic:   197 on 3 and 5767 DF,  p-value: < 2.2e-16
```

```
> # Clustered standard errors
> library(sandwich)
> print(length(classsizedata2$gktchid))
[1] 5771
> print(length(unique(classsizedata2$gktchid)))
[1] 325
> clusteredvariance <- vcovCL(output1, classsizedata2$gktchid)
> clusteredstderrors <- sqrt(diag(clusteredvariance))
> print(olsstderrors)
(Intercept)   smallTRUE    poorTRUE  femaleTRUE
     0.3936      0.4719      0.4329      0.4328
> print(clusteredstderrors)
(Intercept)   smallTRUE    poorTRUE  femaleTRUE
     0.7531      1.1191      0.7565      0.4550
```

- Notice that clustered standard errors result in the standard error increasing from:
  - 0.472 to 1.129 on the small-classroom indicator
    - By construction people in the same classroom have the same classroom size
  - 0.433 to 0.757 on the poor-student indicator
    - People in the same classrooms tend to have similar incomes
  - 0.433 to 0.455 on the female-student indicator
    - The gender makeup of students tends to be pretty similar over classrooms
- Highlights that the effective sample size depends on how much variation there is in the variable within clusters

- A multi-level model is another statistical technique that can be used to get more accurate measures of uncertainty in regression coefficients when analyzing data in which some observations necessarily share a common independent variable
  - I'll present the logic shortly when talking about random effects for panel data
- Things to keep in mind about multi-level model
  1. It changes the coefficients and standard errors from an OLS regression
  2. It also is applying asymptotics w.r.t. the number of groups (rather than the number of observations)
  3. But it may work better than clustered standard errors when you have a medium number of groups

```
> library(lme4)
Loading required package: Matrix
> model <- lmer(gktreadss ~ small + poor + female + (1 | gktchid), data=classsizedata2)
> summary(model)
Linear mixed model fit by REML ['lmerMod']
Formula: gktreadss ~ small + poor + female + (1 | gktchid)
   Data: classsizedata2

REML criterion at convergence: 47317.5

Scaled residuals:
    Min      1Q  Median      3Q     Max
-4.3461 -0.6538 -0.0605  0.5848  4.0191

Random effects:
 Groups   Name        Variance Std.Dev.
 gktchid  (Intercept)  81.99    9.055
 Residual             189.02   13.748
Number of obs: 5771, groups:  gktchid, 325

Fixed effects:
            Estimate Std. Error t value
(Intercept)  45.5241     0.7337  62.048
smallTRUE     3.0824     1.1040   2.792
poorTRUE     -8.8576     0.4328 -20.464
femaleTRUE    2.8544     0.3710   7.693
```

- Autocorrelation also is almost always present is when a regression includes more that one observation from the same unit of analysis
- Types of such regressions:
  - Time-series regressions: Relate over-time variation in a dependent and independent variable(s) from a single unit of analysis
  - Time-series, cross-sectional or panel regressions: Relate over-time and across-unit variation in a dependent and independent variable(s) from a multiple units of analysis
    - Time-series, cross-sectional regressions have more time periods than units of analysis
    - Panel regressions have more units of analysis than time periods
- Often even more severe problems with OLS understating uncertainty in such regressions than in regressions with a grouped explanatory variable

- Time-series regressions estimate whether a variable tends to be higher or lower in a time period based on the value of other variable(s) in that same time period
  - E.g., $Y_t = \alpha + \beta X_t + \epsilon_t$
    - For example, examining how a variable like presidential approval in month m associates with the unemployment rate in that same month m
- It often is the case that there are similar unobservable determinants of the dependent variable in proximate time periods
  - E.g., $E[\epsilon_t \epsilon_{t+1} \mid X] > 0$?
    - The same factors that cause us to over predict (or under predict) presidential approval in month m also are likely to cause us to over predict (or under predict) presidential approval in month $m+1$
- It also often is the case that the value of an explanatory variable is similar in proximate time periods

- The econometrics of time-series data become extremely challenging, and are largely outside the scope of this course
  - Clustered standard errors or random effects aren't an option with time-series data because we only observe a single cluster
- One of the most widely used, and simplest, time-series models assumes that $\epsilon_t = \rho\epsilon_{t-1} + \nu_t$
  - Called the auto-regressive one (or AR(1)) model
  - The closer the estimated value of ρ is to one, the more autocorrelated the unobservables in proximate time periods
- The package "prais" in R allows you to estimate a AR(1) model in R
  - If this model is incorrect, both the coefficients and standard errors may be incorrect

```
> # OLS regression
> output1 <- lm(pres_approve ~ unemploy_rate, data = presapprove)
> summary(output1)

Call:
lm(formula = pres_approve ~ unemploy_rate, data = presapprove)

Residuals:
    Min      1Q  Median      3Q     Max
-35.960  -8.493  -1.324   8.877  39.489

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)      6.418     13.609   0.472  0.63831
unemploy_rate    8.131      2.553   3.185  0.00197 **
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 15.46 on 92 degrees of freedom
Multiple R-squared:  0.09933,       Adjusted R-squared:  0.08954
F-statistic: 10.15 on 1 and 92 DF,  p-value: 0.001974
```

```
> # PRAIS regression
> output2 <- prais.winsten(pres_approve ~ unemploy_rate, data = presapprove)
> getAnywhere(output2)

[[1]]

Call:
lm(formula = fo)

Residuals:
    Min      1Q  Median      3Q     Max
-9.4844 -2.6079 -0.6596  1.8846 29.3556

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
Intercept      44.0271    20.1574   2.184   0.0315 *
unemploy_rate   0.6467     3.2289   0.200   0.8417
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 4.987 on 92 degrees of freedom
Multiple R-squared:  0.2013,      Adjusted R-squared:  0.184
F-statistic:  11.6 on 2 and 92 DF,  p-value: 3.225e-05

[[2]]
      Rho Rho.t.statistic Iterations
 0.9571359        29.81672          5
```

- The previous two slides shows that using OLS to conduct a time-series regression can generate misleading results
  - It too easily misattributes the relationship between a dependent variable and an autocorrelated unobserved determinant to whatever observed determinant happens to be higher / lower than average when the unobserved determinant is higher or lower than normal
    - Presidential approval in 2017 was lower than expected because of Trump
    - Risk misattributing this "Trump effect" to some independent variable that also happened to be higher or lower than normal in 2017
- Thus, OLS often is inappropriate even for exploratory analysis when doing a time-series regression

- Autocorrelation is also almost always present when we run regressions using panel data
  - Panel data refer to the case where there are multiple units of observation i that we observe at multiple points in time t
- Let $Y_{i,t}$ be some outcome of interest for unit i at time t
  - Where we observe $1, 2, ..., N$ units at $1, 2, ..., T$
  - Where typically $N >> T$
- Let $X_{j,i,t}$ be the jth explanatory variable for unit i at time t that affects $Y_{i,t}$
  - Just as before we observe k different explanatory variables

fips	year	state	county	demshare	unemp
1001	2000	Alabama	Autauga	0.287192	4
1001	2004	Alabama	Autauga	0.2369404	4.8
1001	2008	Alabama	Autauga	0.2577302	5.1
1001	2012	Alabama	Autauga	0.2658783	6.9
1001	2016	Alabama	Autauga	0.2376967	5.1
1003	2000	Alabama	Baldwin	0.2478222	3.7
1003	2004	Alabama	Baldwin	0.2250289	5.2
1003	2008	Alabama	Baldwin	0.2381192	4.6
1003	2012	Alabama	Baldwin	0.2158944	7.5
1003	2016	Alabama	Baldwin	0.193856	5.3
1005	2000	Alabama	Barbour	0.4990861	5.5
1005	2004	Alabama	Barbour	0.4483623	7.2
1005	2008	Alabama	Barbour	0.4898538	8.8
1005	2012	Alabama	Barbour	0.5136849	11.5
1005	2016	Alabama	Barbour	0.4652784	8.3

- Our baseline model is that: $Y_{i,t} = \beta_1 X_{1,i,t} + \beta_2 X_{2,i,t} + \cdots + \beta_k X_{k,i,t} + \epsilon_{i,t}$
- We can redefine $\epsilon_{i,t} = \alpha_i + \nu_{i,t}$
  - α_i are unobserved characteristics that affect unit i the same in all time periods
  - $\nu_{i,t}$ are unobserved characteristics that affect unit i specifically at time t
- The next slide shows that $E[\epsilon_{i,t}\epsilon_{i,t'} \mid X] \neq 0$ if ${\sigma_\alpha}^2 > 0$
  - ${\sigma_\alpha}^2 = E[\alpha_i^2 \mid X] - E[\alpha_i \mid X]^2$

- $E[\epsilon_{i,t}\epsilon_{i,t'} \mid X] =$
  $E[(\alpha_i + \nu_{i,t})(\alpha_i + \nu_{i,t'}) \mid X] =$
  $E[\alpha_i^2 \mid X] + E[\nu_{i,t}\alpha_i \mid X] + E[\nu_{i,t'}\alpha_i \mid X] + E[\nu_{i,t}\nu_{i,t'} \mid X] =$
  $.\sigma_\alpha{}^2 + E[\alpha_i \mid X]^2 + [\nu_{i,t}\alpha_i \mid X] + E[\nu_{i,t'}\alpha_i \mid X] + E[\nu_{i,t}\nu_{i,t'} \mid X]$

- Two approaches for dealing with this issue
  - Clustered standard errors by unit of observation
  - Random effects by unit of observation
- Comparative advantages of the different approaches:
  - Clustering only adjusts the standard errors, but does not change the OLS coefficients
  - Random effects may be more efficient, and have better sample sample properties, when N is neither small or large
- Neither work well when N is small

- When running random effects generally assume that $E[\alpha_i \alpha_{i'}] =$
  - ${\sigma_\alpha}^2$ when $i = i'$
  - 0 when $i \neq i'$
- And that $E[\nu_{i,t}\nu_{i',t'}] =$
  - ${\sigma_\nu}^2$ when $i = i'$ and $t = t'$
  - 0 otherwise
    - Often more plausible that $E[\nu_{i,t}\nu_{i',t}] = 0$ when the regression model includes time period dummy variables
    - Also note that this is assuming no within unit serial correlation (e.g. $E[\nu_{i,t}\nu_{i,t+1}] = 0$)

- Suppose we have two units and two time periods, so:

$$\begin{pmatrix} Y_{1,1} \\ Y_{1,2} \\ Y_{2,1} \\ Y_{2,2} \end{pmatrix} = \begin{pmatrix} X_{1,1} \\ X_{1,2} \\ X_{2,1} \\ X_{2,2} \end{pmatrix} \beta + \begin{pmatrix} \epsilon_{1,1} \\ \epsilon_{1,2} \\ \epsilon_{2,1} \\ \epsilon_{2,2} \end{pmatrix}$$

- $\Omega = E[\epsilon\epsilon^T \mid X] =$

$$\begin{pmatrix} E[\epsilon_{1,1}\epsilon_{1,1} \mid X] & E[\epsilon_{1,1}\epsilon_{1,2} \mid X] & E[\epsilon_{1,1}\epsilon_{2,1} \mid X] & E[\epsilon_{1,1}\epsilon_{2,2} \mid X] \\ E[\epsilon_{1,2}\epsilon_{1,1} \mid X] & E[\epsilon_{1,2}\epsilon_{1,2} \mid X] & E[\epsilon_{1,2}\epsilon_{2,1} \mid X] & E[\epsilon_{1,2}\epsilon_{2,2} \mid X] \\ E[\epsilon_{2,1}\epsilon_{1,1} \mid X] & E[\epsilon_{2,1}\epsilon_{1,2} \mid X] & E[\epsilon_{2,1}\epsilon_{2,1} \mid X] & E[\epsilon_{2,1}\epsilon_{2,2} \mid X] \\ E[\epsilon_{2,2}\epsilon_{1,1} \mid X] & E[\epsilon_{2,2}\epsilon_{1,2} \mid X] & E[\epsilon_{2,2}\epsilon_{2,1} \mid X] & E[\epsilon_{2,2}\epsilon_{2,2} \mid X] \end{pmatrix} =$$

$$\begin{pmatrix} {\sigma_\alpha}^2 + {\sigma_\nu}^2 & {\sigma_\alpha}^2 & 0 & 0 \\ {\sigma_\alpha}^2 & {\sigma_\alpha}^2 + {\sigma_\nu}^2 & 0 & 0 \\ 0 & 0 & {\sigma_\alpha}^2 + {\sigma_\nu}^2 & {\sigma_\alpha}^2 \\ 0 & 0 & {\sigma_\alpha}^2 & {\sigma_\alpha}^2 + {\sigma_\nu}^2 \end{pmatrix}$$

- The logic from the previous slide generalizes to a model with N units and T time periods
- Define Σ as a T X T matrix such that:
  - The diagonal elements of Σ equal ${\sigma_\alpha}^2 + {\sigma_\nu}^2$
  - The non-diagonal elements of Σ equal ${\sigma_\alpha}^2$
- Define $\bar{0}_T$ as a T X T matrix of zeros
- Then Ω is a T*N X T*N matrix such that

$$\Omega = E[\epsilon\epsilon^T \mid X] = \begin{pmatrix} \Sigma & \bar{0}_T & \dots & \bar{0}_T \\ \bar{0}_T & \Sigma & \dots & \bar{0}_T \\ \dots & \dots & \dots & \dots \\ \bar{0}_T & \bar{0}_T & \bar{0}_T & \Sigma \end{pmatrix}$$

- To implement a random effects regression:
  1. Estimate $\hat{\beta}$ using OLS and construct $e_{i,t}$
  2. Estimate $\hat{\sigma_\alpha}^2 + \hat{\sigma_\nu}^2 = \frac{1}{NT-k-1} \sum_{i=1}^{N} \sum_{t=1}^{T} {e_{i,t}}^2$
  3. Estimate $\hat{\sigma_\alpha}^2 = \frac{1}{\frac{N(T^2-T)}{2} - k - 1} \sum_{i=1}^{N} \sum_{t=1}^{T-1} \sum_{t'=t+1}^{T} e_{i,t} e_{i,t'}$
  4. Back out $\hat{\sigma_\nu}^2 = \hat{\sigma_\alpha}^2 + \hat{\sigma_\nu}^2 - \hat{\sigma_\alpha}^2$
  5. Construct $\hat{\Omega}$ and reestimate $\hat{\beta}$ assuming that $\Omega = \hat{\Omega}$
- Unlike with clustered standard errors, will not get the same coefficients as when we use OLS

- The same syntax is used in R to estimate a baseline random effects regression that we used to estimate a baseline multi-level model
- To show how to implement, I run a regression in which the dependent variable is the Democratic presidential candidate's vote share in a county and the independent variable is the unemployment rate in the county in year of the election
- Two potential sources of autocorrelation:
  1. We are repeatedly measuring vote share in the same set of counties
  2. Autocorrelation in the unemployment rate across counties within a year
- The random effects specification that I developed on the previous slides deals with the first, but not the second, of these sources of autocorrelation

```
> # Indicator for whether the incumbent president is a Democrat
> mydata$dem <- 0
> mydata$dem[mydata$year == 2000] <- 1
> mydata$dem[mydata$year == 2012] <- 1
> mydata$dem[mydata$year == 2016] <- 1
> # OLS regression
> output1 <- lm(demshare ~ unemp*dem, data = mydata)
> summary(output1)

Call:
lm(formula = demshare ~ unemp * dem, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max
-0.3832 -0.1008 -0.0098  0.0881  0.5405

Coefficients:
             Estimate Std. Error t value             Pr(>|t|)
(Intercept)  0.280295   0.005456   51.37 <0.0000000000000002 ***
unemp        0.021243   0.000902   23.55 <0.0000000000000002 ***
dem          0.011553   0.006487    1.78                0.075 .
unemp:dem   -0.008740   0.001055   -8.28 <0.0000000000000002 ***
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 0.136 on 15202 degrees of freedom
Multiple R-squared:  0.0816,	Adjusted R-squared:  0.0814
F-statistic:  450 on 3 and 15202 DF,  p-value: <0.0000000000000002
```

```
> # Random effects
> options(scipen=999)
> library(lme4)
> output2 <- lmer(demshare ~ unemp*dem + (1 | fips), data=mydata)
> summary(output2)
Linear mixed model fit by REML ['lmerMod']
Formula: demshare ~ unemp * dem + (1 | fips)
   Data: mydata

REML criterion at convergence: -34261

Scaled residuals:
   Min     1Q Median     3Q    Max
-4.711 -0.524  0.051  0.551  4.650

Random effects:
 Groups   Name        Variance Std.Dev.
 fips     (Intercept) 0.01598  0.1264
 Residual             0.00316  0.0562
Number of obs: 15206, groups:  fips, 3111

Fixed effects:
              Estimate Std. Error t value
(Intercept)  0.3774666  0.0037640  100.28
unemp        0.0040610  0.0005094    7.97
dem         -0.0366102  0.0028655  -12.78
unemp:dem   -0.0000518  0.0004712   -0.11
```

- The OLS and random-effects regression show substantively different findings about how a one percentage point increase in the unemployment rate associates with Democratic party vote share in a county
- OLS regression:
  - 2.1 percentage point increase in Democratic vote share when the incumbent is a Republican
  - 1.3 percentage point increase in Democratic vote share when the incumbent is a Democrat
- Random-effects regression:
  - 0.4 percentage point increase in Democratic vote share whether the incumbent is a Democrat or Republican
- The null hypothesis of no relationship between unemployment and presidential voting when a Republican in incumbent is rejected in both

- A second random effect can be included to account for autocorrelation in the unemployment rate across counties within a year
- In theory, we adjust our Ω so that $E[\epsilon_{i,t}\epsilon_{j,t}] = \sigma_{\tau}{}^{2}$ (instead of zero)
- In practice, this means using six years of data to estimate $\hat{\sigma_{\tau}}^{2}$

```
> output3 <- lmer(demshare ~ unemp*dem + (1 | fips) + (1 | year), data=mydata)
> summary(output3)
Linear mixed model fit by REML ['lmerMod']
Formula: demshare ~ unemp * dem + (1 | fips) + (1 | year)
   Data: mydata

REML criterion at convergence: -39066

Scaled residuals:
   Min      1Q Median      3Q     Max
-4.717 -0.533 -0.008  0.494  4.826

Random effects:
 Groups   Name        Variance Std.Dev.
 fips     (Intercept) 0.01644  0.1282
 year     (Intercept) 0.00140  0.0374
 Residual             0.00211  0.0459
Number of obs: 15206, groups:  fips, 3111; year, 5

Fixed effects:
             Estimate Std. Error t value
(Intercept)  0.388752   0.026689   14.57
unemp        0.002064   0.000470    4.39
dem         -0.035793   0.034226   -1.05
unemp:dem    0.000084   0.000394    0.21
```

- Including a second random effect dramatically changed the standard error on the coefficient on the indicator for whether the incumbent president is a Democrat
  - Increased almost by a factor of 10
  - Causing the t-value to go from -12.78 to -1.04
- Highlights how we can be very overconfident when rejecting null hypotheses if we are not careful about accounting for autocorrelation

- We also can use clustering to account for autocorrelation
- Just as with random effects, can either use one-way clustering (to account for repeatedly measuring outcomes within the same counties) or two-way clustering (to also account for autocorrelation of unemployment over counties within the same year)
- Unlike with random effects, only affects standard errors and not the estimate coefficients

```
> # Compare ols, one-way, and two-way clustered standard errors
> olsstderrors <- sqrt(diag(vcov(output1)))
> print(olsstderrors)
(Intercept)       unemp         dem   unemp:dem
   0.005456    0.000902    0.006487    0.001055
> library(sandwich)
> clusteredstderrors <- sqrt(diag(vcovCL(output1, mydata$fips)))
> print(clusteredstderrors)
(Intercept)       unemp         dem   unemp:dem
  0.0073048   0.0011943   0.0041485   0.0007362
> clusteredstderrors2 <- sqrt(diag(vcovCL(output1, cbind(mydata$year, mydata$fips))))
> print(clusteredstderrors2)
(Intercept)       unemp         dem   unemp:dem
   0.018461    0.001645    0.049556    0.004125
```

- The standard errors generated using random effects and clustering are similar
- But substantively the two approaches produce quite different results because the estimated coefficient on the unemployment rate is so different
- Broad takeaway is that:
  - While random effects and clustering should produce similar results when the effective sample size is large, results can be quite sensitive when the effective sample size is small
  - Autocorrelation can mean that the effective sample size can be small even when there are many observations in the regression

Quiz question: Which of these statements summarizes how you should be thinking about autocorrelation relative to heteroscadasiticty?

A. Autocorrelation is less of an issue than heteroscadasticity. Failing to account for autocorrelation is less consequential than failing to account for heteroscadasiticty and the solutions for accounting for autocorrelation are easier to implement and more feasible than the solutions for accounting for heteroscadasiticty.

B. Autocorrelation is similar issue to heteroscadasticity. The consequences of failing to account for heteroscadasiticty and autocorrelation are similar and the solutions for accounting for autocorrelation are equally easy to implement and feasible as the solutions for accounting for heteroscadasiticty.

C. Autocorrelation is more of an issue than heteroscadasticity. Failing to account for autocorrelation is more consequential than failing to account for heteroscadasiticty and the solutions for accounting for autocorrelation are more difficult to implement and less feasible than the solutions for accounting for heteroscadasiticty.

Final comments on autocorrelation:

- Failure to account for autocorrelation can lead to OLS regression being extremely misleading
- Accounting for autocorrelation can be challenging
  - Particularly when we have few units of observation
- Which should inform how you approach data collection
  - Almost always reduce standard errors more by bringing more units of observation into the study than by bringing in more observations from the same units of observation

- We will conclude this week by talking about how to present regression coefficients
- Two takeaways from the previous section:
  1. We need to be able to report regression coefficients in a way that highlights both the estimate and the degree of uncertainty
  2. We need to be able to compare regression output over specifications
- We'll talk about ways to generate tables and figures that allow us to do this

- We'll use the stargazer package to output regression tables
- Stargazer allows users to group regression results into tables so that
  - Each column reports the results of a separate regression
  - Each pair of row reports the coefficient and standard error for a given variable
- Generating such tables is an automated way is an important safeguard in limiting mistakes

```
> # Illustrates stargazer
> library(sandwich)
> library(stargazer)
>
> # Regressions
> output1 <- lm(gktreadss ~ small, data = classsizedata2)
> output1se <- sqrt(diag(vcov(output1)))
> output2 <- lm(gktreadss ~ small + poor, data = classsizedata2)
> output2se <- sqrt(diag(vcov(output2)))
> output3 <- lm(gktreadss ~ small + poor + female, data = classsizedata2)
> output3se <- sqrt(diag(vcov(output3)))
> output3clusterse <- sqrt(diag(vcovCL(output3, classsizedata2$gktchid)))
> output4 <- lmer(gktreadss ~ small + poor + female + (1 | gktchid), data=classsizedata2)
> output4se <- sqrt(diag(vcov(output4)))
>
> # Baseline table
> stargazer(output1, output2, output3, output4, type = "text")
```

	Dependent variable:			
	gktreadss			
	OLS			linear mixed-effects
	(1)	(2)	(3)	(4)
small	3.214***	3.074***	3.077***	3.082***
	(0.494)	(0.475)	(0.472)	(1.104)
poor		-9.445***	-9.487***	-8.858***
		(0.435)	(0.433)	(0.433)
female			3.587***	2.854***
			(0.433)	(0.371)
Constant	42.473***	47.079***	45.352***	45.524***
	(0.271)	(0.336)	(0.394)	(0.734)
Observations	5,771	5,771	5,771	5,771
R2	0.007	0.082	0.093	
Adjusted R2	0.007	0.082	0.092	
Log Likelihood				-23,658.760
Akaike Inf. Crit.				47,329.530
Bayesian Inf. Crit.				47,369.490
Residual Std. Error	17.188 (df = 5769)	16.529 (df = 5768)	16.433 (df = 5767)	
F Statistic	42.417*** (df = 1; 5769)	258.165*** (df = 2; 5768)	197.032*** (df = 3; 5767)	
Note:				*p<0.1; **p<0.05; ***p<0.01

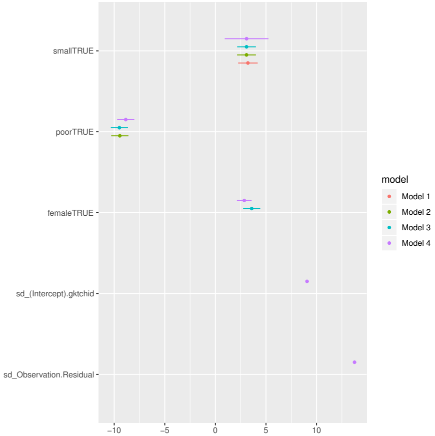
- Advice for formatting tables:
  - Use words, rather than variables names
  - Present the "right" number of digits
  - Make adjacent comparisons whenever possible
- The goal is to make the table as self contained as possible

```
> # Full table
> stargazer(output1, output2, output3, output3, output4, type = "text",
+           digits = 2, model.names = FALSE,
+           se  = list(output1se, output2se, output3se,
+                      output3clusterse, output4se),
+           omit.stat = c("f", "adj.rsq", "ser", "ll", "aic", "bic"),
+           dep.var.labels=c("Student's Reading Test Score"),
+           covariate.labels = c("Small Classroom",
+                                "Student is Eligible for Free and Reduced Priced Lunch",
+                                "Student is Female"),
+           column.labels = c("OLS", "OLS", "OLS", "Cluster", "Ran. Eff."))
```

	Dependent variable:				
	Student's Reading Test Score				
	OLS	OLS	OLS	Cluster	Ran. Eff.
	(1)	(2)	(3)	(4)	(5)
Small Classroom	3.21***	3.07***	3.08***	3.08***	3.08***
	(0.49)	(0.47)	(0.47)	(1.12)	(1.10)
Student is Eligible for Free and Reduced Priced Lunch		-9.44***	-9.49***	-9.49***	-8.86***
		(0.44)	(0.43)	(0.76)	(0.43)
Student is Female			3.59***	3.59***	2.85***
			(0.43)	(0.45)	(0.37)
Constant	42.47***	47.08***	45.35***	45.35***	45.52***
	(0.27)	(0.34)	(0.39)	(0.75)	(0.73)
Observations	5,771	5,771	5,771	5,771	5,771
R2	0.01	0.08	0.09	0.09	
Note:			*p<0.1; **p<0.05; ***p<0.01		

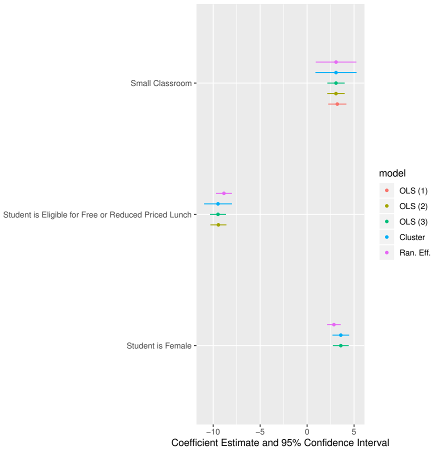
- Other times regression coefficients are more effectively presented using a figure instead of a table
  - Often easier to evaluate the stability of regression coefficients across many models while accounting for the uncertainty
- A number of packages have been developed to visualize regression coefficients
- I'm going to show you how to use the dwplot function, which is part of the dotwhisker library

```
# Baseline figure
library(dotwhisker)
p <- dwplot(list(output1, output2, output3, output4))
```


smallTRUE
poorTRUE
femaleTRUE
sd_(Intercept).gktchid
sd_Observation.Residual
−10
−5
0
5
10
model
Model 1
Model 2
Model 3
Model 4

```
# Full figure
library(dplyr)
library(broom)

output1tidy <- tidy(output1) %>% mutate(model = "OLS (1)")
output1tidy <- output1tidy %>% select(-c(statistic, p.value))
output2tidy <- tidy(output2) %>% mutate(model = "OLS (2)")
output2tidy <- output2tidy %>% select(-c(statistic, p.value))
output3tidy <- tidy(output3) %>% mutate(model = "OLS (3)")
output3tidy <- output3tidy %>% select(-c(statistic, p.value))
output4tidy <- output3tidy %>% mutate(model = "Cluster")
output4tidy$std.error <- output3clusterse
output5tidy <- tidy(output4)  %>% mutate(model = "Ran. Eff.")
output5tidy <- output5tidy %>% select(-c(statistic, group))
output5tidy <- output5tidy %>% filter(term != "sd_(Intercept).gktchid")
output5tidy <- output5tidy %>% filter(term != "sd_Observation.Residual")
all_models <- rbind(output1tidy, output2tidy, output3tidy, output4tidy, output5tidy)
p <- dwplot(all_models) %>%
  relabel_predictors(c(smallTRUE = "Small Classroom",
                       poorTRUE = "Student is Eligible for Free or Reduced Priced Lunch",
                       femaleTRUE = "Student is Female")) +
  xlab("Coefficient Estimate and 95% Confidence Interval")
```


Small Classroom
Student is Eligible for Free or Reduced Priced Lunch
Student is Female
−10
−5
0
5
Coefficient Estimate and 95% Confidence Interval
model
OLS (1)
OLS (2)
OLS (3)
Cluster
Ran. Eff.

Key takeaways:

- Good measures of uncertainty are neeed to construct accurate confidence intervals about parameters
- Violations of assumptions #3 and/or #4 of the Gauss-Markov Theorem are more likely to results in the understatement than overstatement of uncertainty the relationship between independent variable(s) and the dependent variable
- Reasons why we might violate these assumptions:
  - Our explanatory variables affect the expected variance of the unmodeled variables
  - Our dependent variable is aggregating an inconsistent number of individual-level outcomes
  - Some of our observations necessarily share common explanatory variables
  - We are repeatedly measuring the same outcome for the same unit of observation(s)

Key takeaways (continued):

- We have some decent ways to deal with violations of assumptions #3 and assumption #4 of the Gauss-Markov Theorem when we have a large effective sample
  - Just because we have a lot of observations in our regression doesn't mean we have a large effective sample
- In addition to making sure we have good measure of uncertainty, it is important that we construct tables and figures in a way that communicates information about uncertainty effectively
  - Effective communication generally means that your tables and figures don't require much additional context to understand