SOC-GA 2332 Intro to Stats Lab 7

Risa Gelles-Watnick

10/14/2025

```
# load packages
pacman::p_load(
  tidyverse,
  stargazer,
  coefplot,
  sandwich,
  lmtest,
  ggpubr,
  psych
)
```

Part 0: Housekeeping

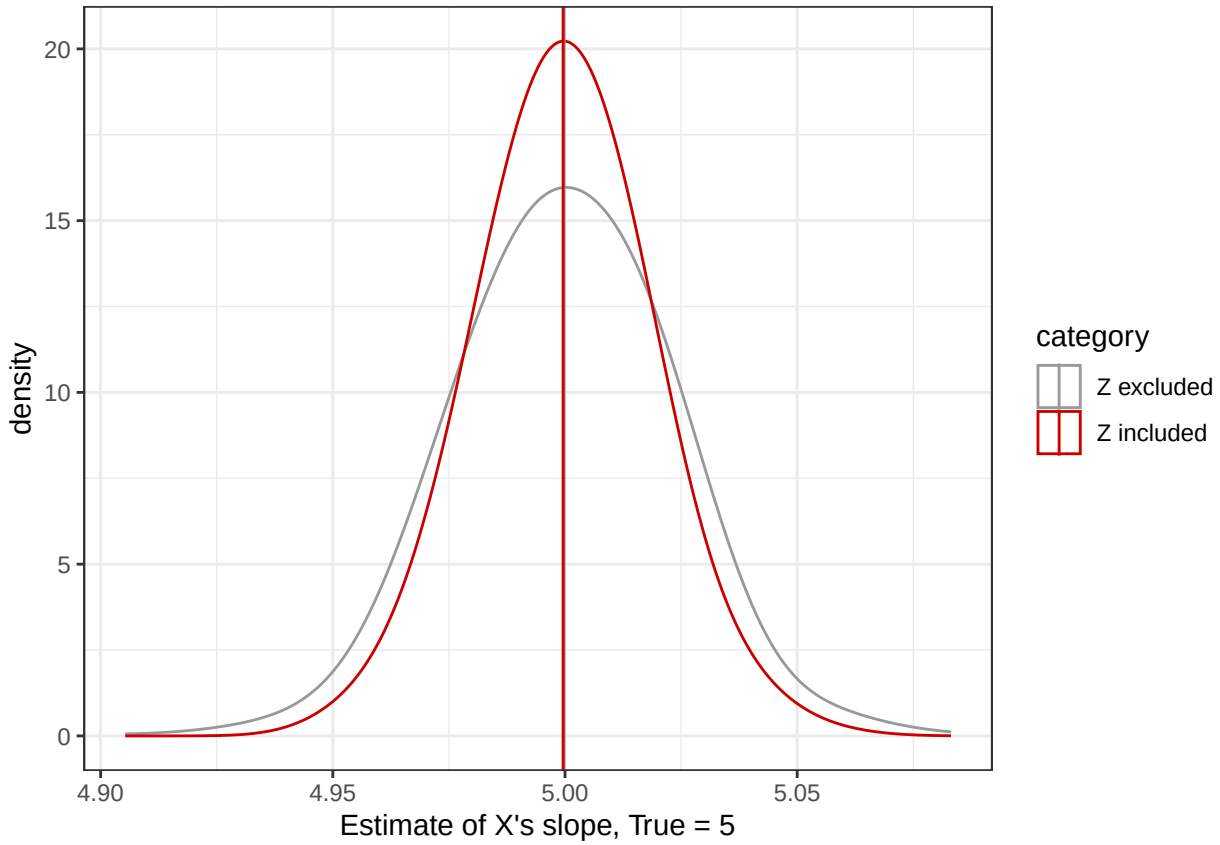
- **Problem Set 2** is due on Friday Oct 17th, 11:59 pm (tonight!)
- Lab next week **(10/24)** will be in the NYU Academic Resource Center, ARC_LL01
- Lab will end at 12:15pm on **11/14** for Prosem TA practice

Part 1: How Does Multivariate Relationships Affect Regression Estimates

- Multiple Causes

```
set.seed(3636)

## empty results
woz <- c()
wz <- c()

for (i in 1:1000){

 ## create hypothetical variables
  X <- runif(500, min=1, max=10)
  Z <- runif(500, min=2, max=5)
  Y <- 10 + 5*X + Z + rnorm(500,0,1)

  ## data
  data <- data.frame(X=X,Y=Y,Z=Z)

  ## regression
  lm1 <- lm(Y ~ X, data)
```

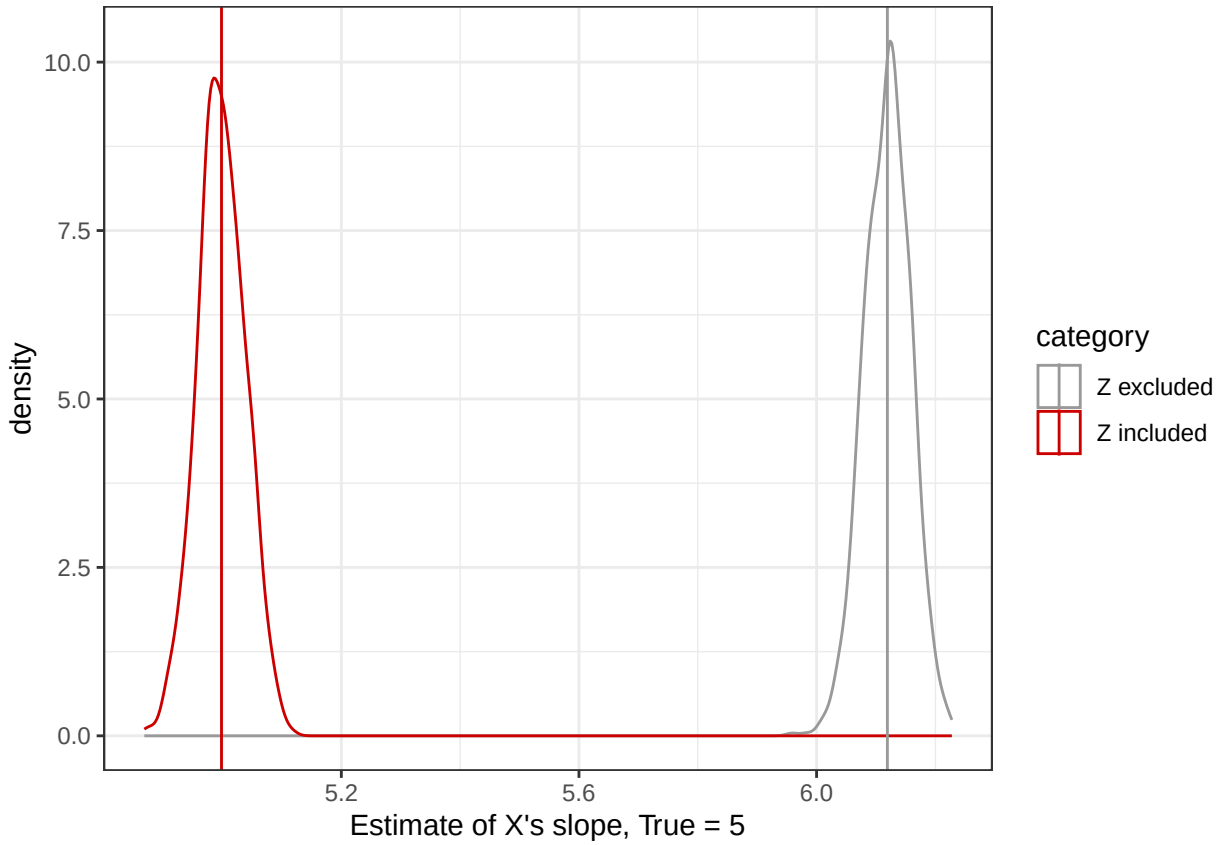
```r
  lm2 <- lm(Y ~ X + Z, data)

  ## extract results
  woz <- c(woz, summary(lm1)$coef[2,1])
  wz <- c(wz, summary(lm2)$coef[2,1])
}

## combine results
results <-
  data.frame(estimate = c(woz,wz),
             category = c(rep("Z excluded",1000),rep("Z included",1000)))

mean <- results %>%
  group_by(category) %>%
  summarize(mean = mean(estimate))

## plot
results %>%
  ggplot(aes(x=estimate,group=category)) +
  geom_density(aes(color=category),bw=0.01) +
  geom_vline(data = mean, aes(xintercept = mean, color = category)) +
  scale_color_manual(values=c("#999999", "red3")) +
  xlab("Estimate of X's slope, True = 5") +
  theme_bw()
```



- Confounding

Question: What is the difference between multiple causes and confounding?

```r
set.seed(2023)

## empty results
woz <- c()
wz <- c()

for (i in 1:1000){

 ## create hypothetical variables
  Z <- runif(500, min=1, max=10)
  X <- runif(500, min=1, max=5) + 0.5*Z
  Y <- 10 + 5*X + Z + rnorm(500,0,1)

  ## data
  data <- data.frame(X=X,Y=Y,Z=Z)

  ## regression
  lm1 <- lm(Y ~ X, data)
  lm2 <- lm(Y ~ X + Z, data)

  ## extract results
  woz <- c(woz, summary(lm1)$coef[2,1])
  wz <- c(wz, summary(lm2)$coef[2,1])
}

## combine results
results <-
  data.frame(estimate = c(woz,wz),
             category = c(rep("Z excluded",1000),rep("Z included",1000)))

mean <- results %>%
  group_by(category) %>%
  summarize(mean = mean(estimate))

## plot
results %>%
  ggplot(aes(x=estimate,group=category)) +
  geom_density(aes(color=category),bw=0.01) +
  geom_vline(data = mean, aes(xintercept = mean, color = category)) +
  scale_color_manual(values=c("#999999", "red3")) +
  xlab("Estimate of X's slope, True = 5") +
  theme_bw()
```

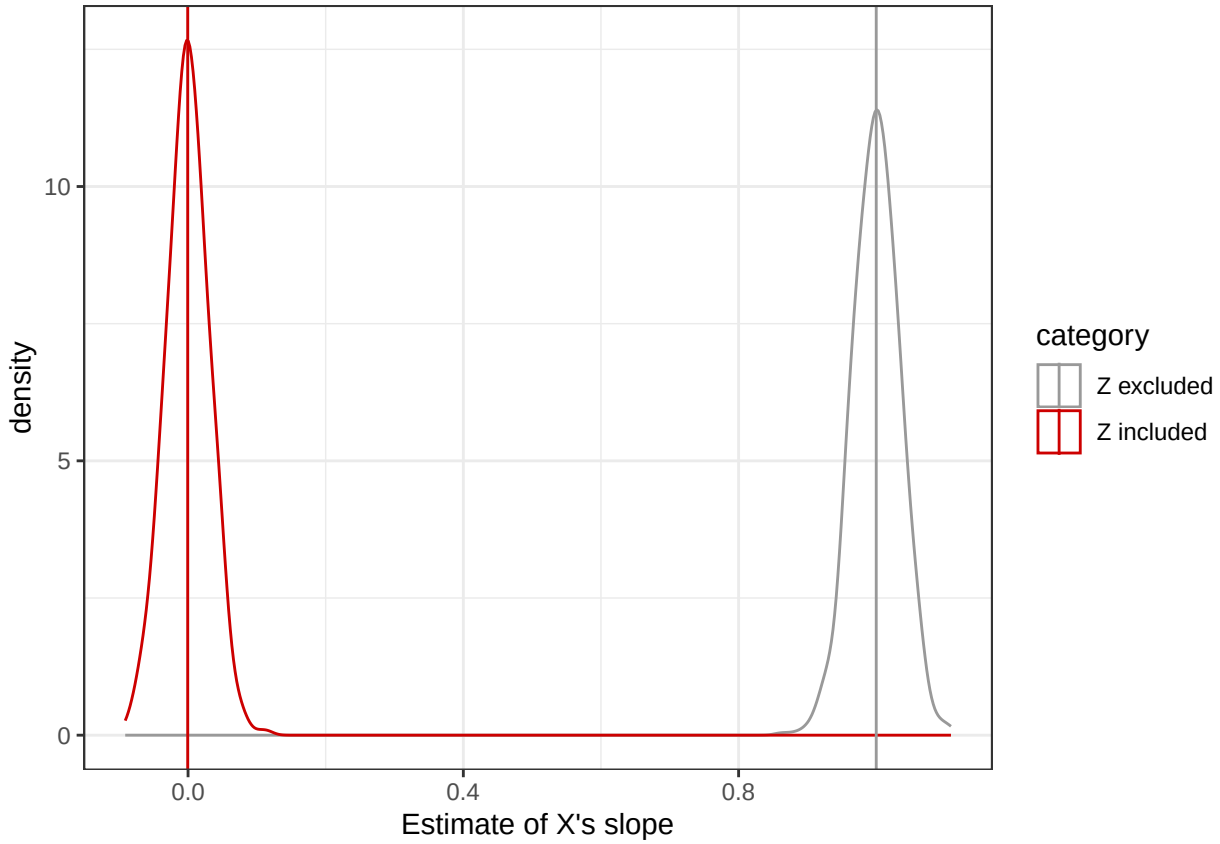


- Mediation

Question: What is mediation?

```
set.seed(3636)

## empty results
woz <- c()
wz <- c()

for (i in 1:1000){

 ## create hypothetical variables
  X <- runif(500, min=1, max=10)
  Z <- runif(500, min=2, max=5) + 0.5*X
  Y <- 2*Z + rnorm(500,0,1)

  ## data
  data <- data.frame(X=X,Y=Y,Z=Z)

  ## regression
  lm1 <- lm(Y ~ X, data)
  lm2 <- lm(Y ~ X + Z, data)

  ## extract results
  woz <- c(woz, summary(lm1)$coef[2,1])
  wz <- c(wz, summary(lm2)$coef[2,1])
}
```

```r
## combine results
results <-
  data.frame(estimate = c(woz,wz),
             category = c(rep("Z excluded",1000),rep("Z included",1000)))

mean <- results %>%
  group_by(category) %>%
  summarize(mean = mean(estimate))

## plot
results %>%
  ggplot(aes(x=estimate,group=category)) +
  geom_density(aes(color=category),bw=0.01) +
  geom_vline(data = mean, aes(xintercept = mean, color = category)) +
  scale_color_manual(values=c("#999999", "red3")) +
  xlab("Estimate of X's slope") +
  theme_bw()
```



Question: What is the "true" effect of X on Y in this model, given our known data generating process? Does the regression calculate this "true" slope for X when controlling for Z? Why or why not?

$$Y = 2 * Z + \epsilon$$

$$Y = 2 * (U_z + 0.5 * X) + \epsilon$$

$$Y = 2U_z + 1 * X + \epsilon$$

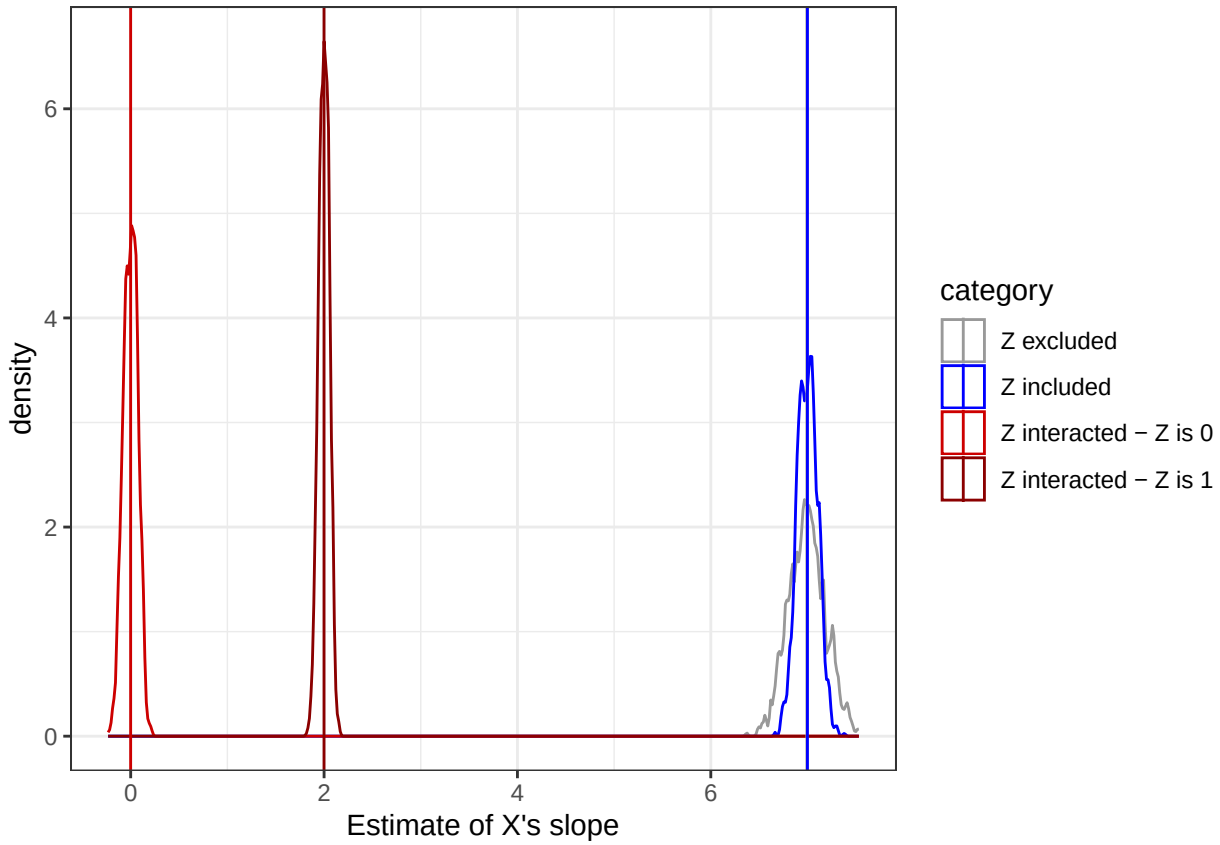
So the "true" effect of X on Y is 1. We don't get a slope of 1 for X in our regression because we're controlling for Z. Since X is entirely moderated through Z, that is, X only affects Y through its effect on Z, when we control for Z the effect of X on Y is 0.

- Moderation

Question: What is moderation? How do we control for it in regressions?

```
set.seed(2023)

## empty results
woz <- c()
wz <- c()
wzintz0 <- c()
wzintz1 <- c()

for (i in 1:1000){

 ## create hypothetical variables
  X <- runif(500, min=1, max=10)
  Z <- runif(500, min=2, max=5)
  Y <- 2*X*Z + rnorm(500,0,1)

  ## data
  data <- data.frame(X=X,Y=Y,Z=Z)

  ## regression
  lm1 <- lm(Y ~ X, data)
  lm2 <- lm(Y ~ X + Z, data)
  lm3 <- lm(Y ~ X*Z, data) # this is equivalent to lm(Y ~ X + Z + X*Z, data)

  ## extract results
  woz <- c(woz, summary(lm1)$coef[2,1])
  wz <- c(wz, summary(lm2)$coef[2,1])
  wzintz0 <- c(wzintz0, summary(lm3)$coef[2,1])
  wzintz1 <- c(wzintz1, summary(lm3)$coef[2,1] + summary(lm3)$coef[4,1])
}

## combine results
results <-
  data.frame(estimate = c(woz, wz, wzintz0, wzintz1),
             category = c(
               rep("Z excluded", 1000),
               rep("Z included", 1000),
               rep("Z interacted - Z is 0", 1000),
               rep("Z interacted - Z is 1", 1000))
  )

mean <- results %>%
  group_by(category) %>%
  summarize(mean = mean(estimate))

## plot
results %>%
  ggplot(aes(x=estimate,group=category)) +
```

```
  geom_density(aes(color=category),bw=0.01) +
  geom_vline(data = mean, aes(xintercept = mean, color = category)) +
  scale_color_manual(values=c("#999999", "blue", "red3", "red4")) +
  xlab("Estimate of X's slope") +
  theme_bw()
```



Question: What's a good way to check whether you should be including an interaction term in your regression model?

Part 2: Additional Topics in Regression

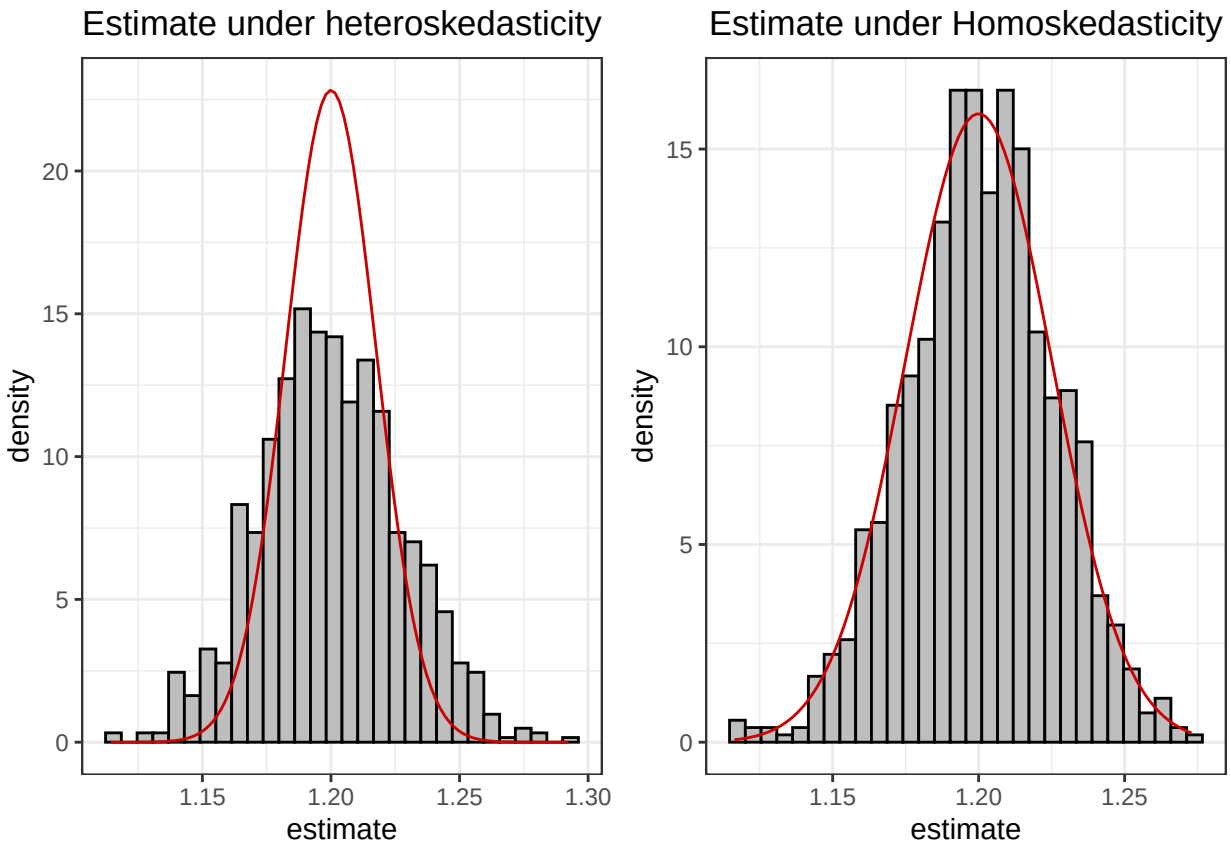
1. Heteroskedasticity and Robust Standard Errors

- Heteroskedasticity occurs when the **variance of the error term changes across different values of the explanatory variables**; $Var(\epsilon_i|X) \neq Var(\epsilon_i)$, or, as we see in lecture, we assume that $Var(\epsilon_i|X) = \sigma^2 h(X)$
- Heteroskedasticity violates the basic assumption of OLS, in which the variance of the error term should be constant across different values of the explanatory variables.
- *Question*: Will heteroskedasticity make estimates biased and inconsistent?
- In OLS estimation, the standard error of $\hat{\beta}_1$, $se_{\hat{\beta}_1}$ is derived by assuming homoskedasticity. Specifically, given known X and the uncertainty coming from sampling the same X but with different ϵ_i from the population, we assume $Var(y|X) = Var(\epsilon|X) = Var(\epsilon) = \sigma^2$

$$
\begin{aligned}
\hat{\beta} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\
\text{Var}(\hat{\beta}|\mathbf{X}) &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\ \text{Var}(\mathbf{y}|\mathbf{X})\ \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\ \sigma^2\mathbf{I}\ \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
&= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1} \\
&= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}
\end{aligned}
$$

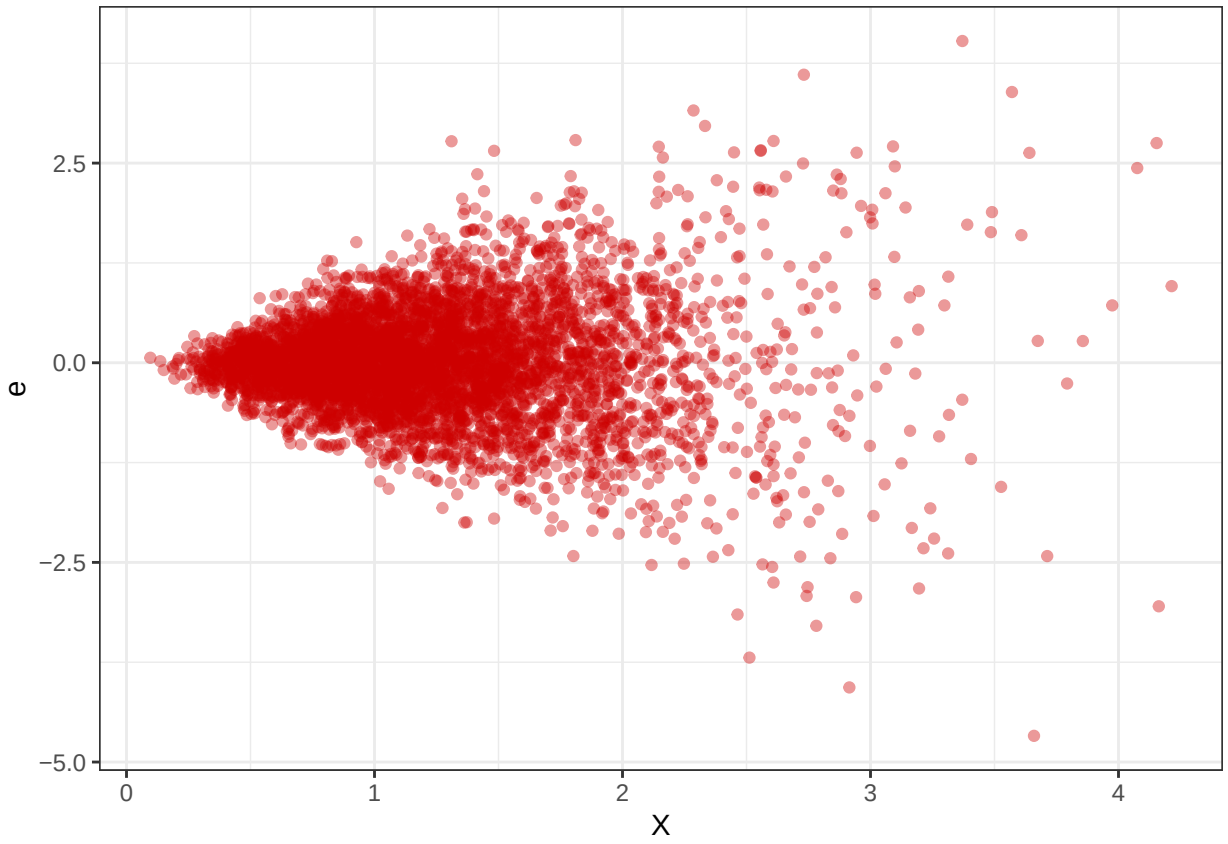
```r
set.seed(2023)

## empty results
homo <- c()
hetero <- c()

## create x
X <- rgamma(5000, 5, 4)

for (i in 1:1000){

 ## create hypothetical data
  homo_Y <- -0.25 + 1.2*X + rnorm(5000,0,1)
  hetero_Y <- -0.25 + 1.2*X + rnorm(5000,0,0.5*X)

  ## data
  homo_data <- data.frame(X=X,homo_Y)
  hetero_data <- data.frame(X=X,hetero_Y)

  ## regression
  lm1 <- lm(homo_Y ~ X, homo_data)
  lm2 <- lm(hetero_Y ~ X, hetero_data)

  ## extract results
  homo <- c(homo, summary(lm1)$coef[2,1])
  hetero <- c(hetero, summary(lm2)$coef[2,1])
}

## combine results
results <-
  data.frame(estimate = c(homo,hetero),
             category = c(rep("homoskedasticity",1000),rep("heteroskedasticity",1000)))

## plot
hetero <-
  results %>%
  filter(category=="heteroskedasticity") %>%
  ggplot(aes(x=estimate)) +
  geom_histogram(aes(y=..density..),fill="grey",color="black") +
  stat_function(fun = dnorm,
                args = list(mean = 1.2,
                            sd = summary(lm2)$coef[2,2]),
                color = "red3") +
  ggtitle("Estimate under heteroskedasticity") +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5))
```

```r
homo <-
  results %>%
  filter(category=="homoskedasticity") %>%
  ggplot(aes(x=estimate)) +
  geom_histogram(aes(y=..density..),fill="grey",color="black") +
  stat_function(fun = dnorm,
                args = list(mean = 1.2,
                            sd = summary(lm1)$coef[2,2]),
                color = "red3") +
  ggtitle("Estimate under Homoskedasticity") +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5))

ggarrange(hetero, homo)
```



- Detecting heteroskedasticity by plotting $Y - \hat{Y}$ over X

```r
data.frame(X = X, e = hetero_Y - predict.lm(lm2,hetero_data)) %>%
  ggplot(aes(x=X,y=e)) +
  geom_point(color="red3",alpha=0.4) +
  theme_bw()
```



- Robust standard error
- If you have a reason to believe that your dataset violates the assumption of homoskedasticity, you can use the packages `sandwich` and `lmtest` to get robust standard errors.

```
## original SE
summary(lm2)$coef
```

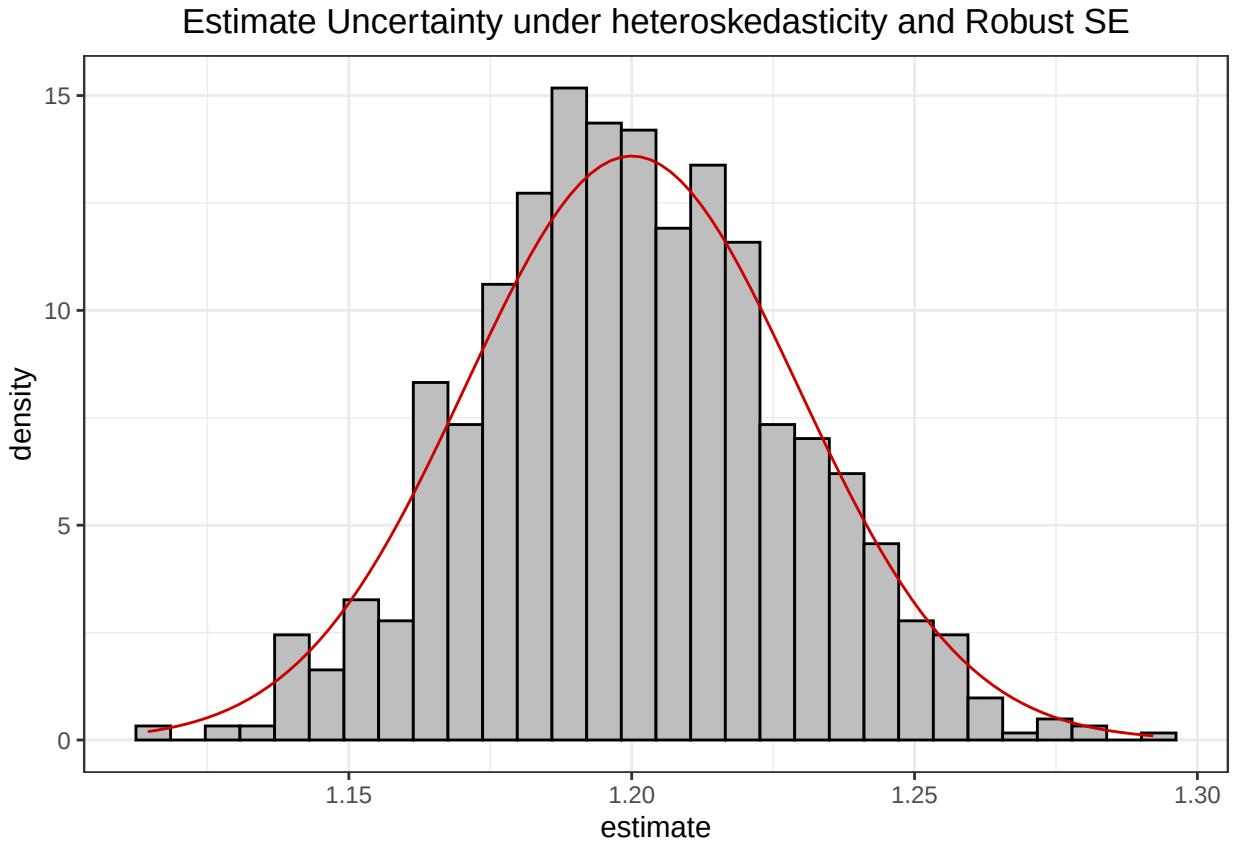
```
##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) -0.2255046 0.02382858 -9.463619 4.461315e-21
## X            1.1793730 0.01747531 67.487975 0.000000e+00
```

```
## robust SE
coeftest(lm2, vcov = vcovHC(lm2, type="HC1")) # heteroskedasticity-consistent variance-covariance
```

```
##
## t test of coefficients:
##
##              Estimate Std. Error t value  Pr(>|t|)
## (Intercept) -0.225505   0.031452 -7.1698 8.611e-13 ***
## X            1.179373   0.029350 40.1837 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
results %>%
  filter(category=="heteroskedasticity") %>%
  ggplot(aes(x=estimate)) +
  geom_histogram(aes(y=..density..),fill="grey",color="black") +
  stat_function(fun = dnorm,
                args = list(mean = 1.2,
```

```
                          sd = coeftest(lm2, vcov = vcovHC(lm2, type="HC1"))[2,2]),
              color = "red3") +
ggtitle("Estimate Uncertainty under heteroskedasticity and Robust SE") +
theme_bw() +
theme(plot.title = element_text(hjust = 0.5))
```



Question: What is an example of a research question where we might expect heteroskedasticity of errors?

2. Clustering of Errors

- One of the basic OLS assumptions is that the error term is independently distributed across observations. i.e.:

$$\text{Corr}(\epsilon_i, \epsilon_j | X) = 0 \ \forall \ i \neq j$$

- This assumption could be violated when your data have a "nested" structure, or your data is ordered by time and the trend is highly correlated between time unit. In such cases, you should employ other modeling techniques to address correlated errors. For example, you can use multilevel modeling for nested data, and longitudinal data analysis techniques for time-series data.

Part 3: F-test for Nested Models

- We can use F-test to compare two regression models. The idea behind the F-test for nested models is to check **how much errors are reduced after adding additional predictors**. A relatively large reduction in error yields a large F-test statistic and a small P-value. The P-value for F statistics is the right-tail probability.

- If the F's p-value is significant (smaller than 0.05 for most social science studies), it means that at least one of the additional β_j in the full model is not equal to zero.
- The F test statistic for nested regression models is calculated by:

$$F = \frac{(SSE_{\text{restricted}} - SSE_{\text{full}})/df_1}{SSE_{\text{full}}/df_2}$$

where df_1 is the number of **additional** predictors added in the full model and df_2 is the **residual degrees of freedom for the full model**, which equals ($n - 1 -$ number of IVs in the complete model). The df of the F test statistic is (df_1, df_2).

For example, let's look at the earnings data set we used last class.

Performing the same cleaning operations as last time:

```
## read data
earnings_df <- read.csv("data/earnings_df.csv", stringsAsFactors = F)

## recode age
earnings_df <-
  earnings_df %>%
  mutate(age = case_when(
    age > 9000 ~ NA,
    .default = age
  ))

## recode female
earnings_df <- earnings_df %>%
  mutate(female = case_when(
    sex == "female" ~ 1,
    .default = 0))

## base R way of doing it
earnings_df$female <- 0
earnings_df[earnings_df$sex=="female", "female"] <- 1

## create black and other
earnings_df <-
  earnings_df %>%
  mutate(black = case_when(
    race == "black" ~ 1,
    .default = 0
  )) %>%
  mutate(other = case_when(
    race == "other" ~ 1,
    .default = 0
  ))
```

And running models #3 and #4 from last class:

(3) Model 3: earn ~ age + edu + female
(4) Model 4: earn ~ age + edu + female + race

```
m3 <- lm(earn ~ age + edu + female,
         data = earnings_df)

m4 <- lm(earn ~ age + edu + female + black + other,
```

```
         data = earnings_df)

stargazer(m3, m4,
          type = "latex")
```

% Table created by stargazer v.5.2.3 by Marek Hlavac, Social Policy Institute. E-mail: marek.hlavac at gmail.com % Date and time: Wed, Oct 15, 2025 - 17:39:30

Table 1:

	Dependent variable:	
	earn	
	(1)	(2)
age	0.160***	0.158***
	(0.022)	(0.022)
edu	4.500***	4.477***
	(0.112)	(0.112)
female	−20.528***	−20.572***
	(0.568)	(0.565)
black		−2.307***
		(0.623)
other		−0.767
		(1.137)
Constant	25.439***	26.429***
	(1.207)	(1.230)
Observations	980	980
R^2	0.744	0.747
Adjusted R^2	0.743	0.746
Residual Std. Error	8.869 (df = 976)	8.817 (df = 974)
F Statistic	943.551*** (df = 3; 976)	575.667*** (df = 5; 974)
Note:		*p<0.1; **p<0.05; ***p<0.01

According to the equation we just wrote out above, we can hand-calculate the F value for `m3` vs `m4`:

```
# SSE_restricted:
sse_m3 <- sum(m3$residuals^2)

# SSE_full:
sse_m4 <- sum(m4$residuals^2)

# We add one additional IV, so:
df1 <- 2

# Residual df for the full model (m5):
df2 <- m4$df.residual
```

```r
# Calculate F:
F_stats <- ((sse_m3 - sse_m4)/df1)/(sse_m4/df2)
F_stats
```

```
## [1] 6.855912
```

```r
# Check tail probability using `1 - pf()`
1 - pf(F_stats, df1, df2)
```

```
## [1] 0.001104788
```

- *Question:* What is your null and alternative hypotheses? What's your decision given the F-test result?
- You can also use `anova()` to perform a F-test in R.

```r
anova(m3, m4)
```

```
## Analysis of Variance Table
##
## Model 1: earn ~ age + edu + female
## Model 2: earn ~ age + edu + female + black + other
##   Res.Df   RSS Df Sum of Sq      F   Pr(>F)
## 1    976 76776
## 2    974 75711  2    1065.8 6.8559 0.001105 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- *Question:* In your own words, what is an ANOVA test doing? How is this different from a t-test or a linear regression?

Part 4: Exercise (20 minutes)

How does the number of theaters a movie opens in affect the total box office earnings of that movie? How much of an increase in box office earnings can producers expect for each theater their movie plays in? Using your knowledge of linear regressions and this (fake!) dataset, model the relationship between number of theaters and box office earnings. Explore the dataset and try to create the best model you can for isolating the effect of theaters on box office earnings. Make sure to check that linear regression assumptions are satisfied and compare models using the tests we learned about in lecture. Report the results of your best regression model in a clean table.

Variables:

theaters: # of theaters that showed the movie on opening weekend

reviews: average score reviewers gave the movie (1-10 with 10 being most positive)

series: dummy variable for whether the movie is part of a series (1 if yes)

budget: budget of movie in 1000s of dollars

studio: studio that released the movie

boxoffice_earnings: how much money the movie made on opening weekend in 1000s of dollars

- Explore the data

```r
# reading in data
boxoffice <- read.csv("data/boxoffice.csv")
```

- Check regression assumptions are met
- Set up several candidates for regression models

- Compare your models using the F-test or ANOVA testing and pick the best one (best meaning the simplest model that describes the data best).
- Display final model

Part 5: Replication Project Tips

1 Samples

- For the 1970 sample, use the 1% state sample
- For the 1990 sample, use the 1% metro sample
  - The reason is that it is the only 1970 sample that provides non-missing `METRO` information
- For the 2010 sample, use the single-year ACS sample, not 3- or 5-year pooled sample
- Some of you may encounter memory issues.
  - A. Try restarting your PC. This will release some used memories.
  - B. Try restricting the sample first before doing any operations. Specifically, you should drop Rs who are younger than 25 and older than 59 (confirm if this is the case in the paper); keep only non-Hispanic White and Black Rs; Rs who are in the workforce (variable `LABFORCE`) and have valid occupation (variable `OCC1990`); and who are economically active (`INCWAGE>0`; pp.1046)
- If you read pp.1046 carefully, you will notice that Rs with the top and bottom earning percentile are excluded
  - You can create percentiles using `quantile(ma$WEEKEARN, seq(0.01,1,0.01))`, suppose your dataframe is `ma`, and the weekly earning variable is `WEEKEARN`
- You should also drop Rs who have missing values for any of the used variables
  - You may consider using the function `complete.cases()` to enable this feature. Reference

2 Variables

- Use `BPL` rather than `NATIVITY`
- The latter has no valid values for most samples
- Use `HISPAN` to exclude Hispanic Whites and Hispanic Blacks
- Use `CLASSWKR` to determine whether R is in a public sector or not
  - You should look at CLASSWKRD, which gives detailed classification of `CLASSWKR`
- Use CPI99 to adjust inflation for `INCWAGE`
- The main dependent variable is the logged form of **weekly earnings**
  - You will need `WKSWORK1` and `WKSWORK2` to measure the number of weeks worked last year. `WKSWORK1` always gives the best continuous estimate, but when `WKSWORK1` is not available, you should turn to `WKSWORK2`
  - `WKSWORK2` is coded in intervals. For example, `WKSWORK2` = 1 means R worked for 1-13 weeks. Use the middle number as a proxy, that is, 7 weeks.
- Recoding weekly working hours has a similar process. You will need `UHRSWORK` and `HRSWORK2` to construct the measure. `UHRSWORK` always gives the best continuous estimate, but when `UHRSWORK` is not available, you should turn to `HRSWORK2` using the middle number as a proxy for the interval estimate.
- To estimate potential years of experience, the formula is given by `LMEXP = AGE - EDUYEAR - 6`
  - `EDUYEAR` needs to be estimated
  - Codes for this process are available in the `code` folder

3 Duncan's Dissimilarity Index

- In Table A1a and A1b, you will notice that there is a dissimilarity index. This is a very commonly used measure of occupational segregation.
  - Check Martin-Caughey (2022) on within-occupation variation and gender segregation using job titles and verbatim texts in GSS that describe jobs
  - The standard Duncan's Dissimilarity/Segregation Index is given by:

$$D = \frac{1}{2}\sum_{i=1}^{n}\left|\frac{a_i}{A} - \frac{b_i}{B}\right|$$

where a_i and b_i is the number of White and Black workers in occupation group i. A and B represents the total number of White and Black workers. * Instead of using hundreds of `OCC1990` categories, you will use 2-digit aggregated categories of `OCC1990` + Use the `merge()` function

4 General Instructions

- It is totally okay if you cannot get exactly the same numbers! I also couldn't.
- But they should be close enough. If they deviate a lot, you need to explain your speculations why the numbers differ this much.
- The total number of observation N may give you some hints (e.g., you did not restrict your sample as much as the original paper).

Part 4: Kitagawa-Oaxaca-Blinder (KOB) Decomposition

- KOB Decomposition is a common practice to decompose group-wise difference into 1. endowment difference, 2. coefficient (slope) difference, and 3. unexplained portion.
  - The group-wise difference is measured at the mean level, e.g. *mean* weakly income
  - For example, gender difference in mean weekly income can be decomposed into:
    * 1. the gender difference in *mean* education, years of experience, etc. Women once lagged behind men in education, but now has surpassed men in mean education
    * 2. the gender difference in *returns* to education, years of experience, etc. Women typically have lower returns to education, but the trend is recently reversed, especially at the lower end (e.g. women with HS or some college are less penalized than men).
    * 3. the portion that cannot be explained by above observed characteristics
- KOB starts from OLS regression. Now we focus on racial (White-Black) gap in mean weakly income
  - OLS by race, White: $Y_{iw} = \alpha_w + \sum_{k=1}^{\ell}\beta_{kw}X_{ikw} + \epsilon_{iw}$
  - OLS by race, Black: $Y_{ib} = \alpha_b + \sum_{k=1}^{\ell}\beta_{kb}X_{ikb} + \epsilon_{ib}$
- The mean value of Y_i, $\bar{Y}_w$ or $\bar{Y}_b$, according to the properties of OLS, is:

$$\bar{Y}_w = \alpha_w + \sum_{k=1}^{\ell}\beta_{kw}\bar{X}_{kw} \quad \bar{Y}_b = \alpha_b + \sum_{k=1}^{\ell}\beta_{kb}\bar{X}_{kb}$$

- The mean racial pay gap:

$$\begin{aligned}\bar{Y}_w - \bar{Y}_b &= \alpha_w + \sum_{k=1}^{\ell}\beta_{kw}\bar{X}_{kw} - \alpha_b - \sum_{k=1}^{\ell}\beta_{kb}\bar{X}_{kb} \\ &= \underbrace{\sum_{k=1}^{\ell}(\bar{X}_{kw} - \bar{X}_{kb})\beta_{kw}}_{\text{endowment difference}} + \left[\underbrace{\sum_{k=1}^{\ell}\bar{X}_{kb}(\beta_{kw} - \beta_{kb})}_{\text{coefficient difference}} + \underbrace{(\alpha_w - \alpha_b)}_{\text{intercept difference}}\right]\end{aligned}$$

- When we focus on one of the explanatory features, such as education, we may visualize the decomposition as follows:

```
knitr::include_graphics("graph/KOB.png")
```

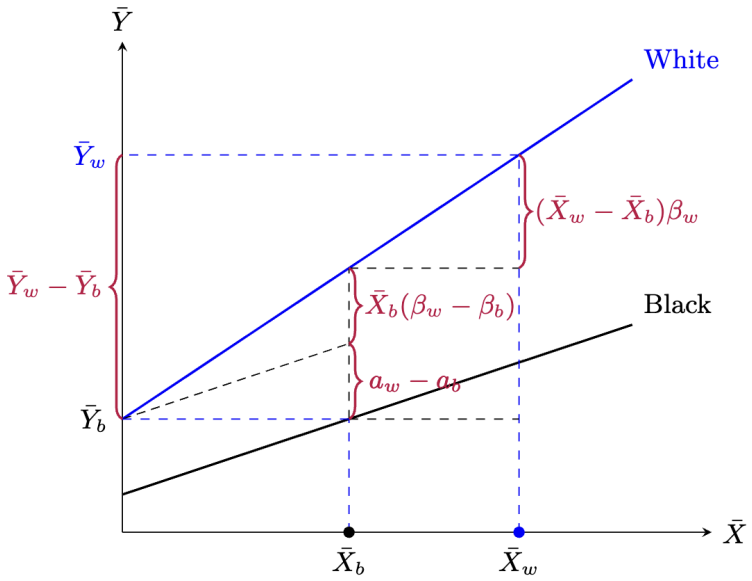


Figure 1: Graphical Demonstration of KOB Decomposition

- Although KOB Decomposition is clean and intuitive, it also has some clear shortcomings. One of the key problems of KOB is that the decomposition results are sensitive to the reference group. E.g., when analyzing the education component in the racial pay gap, whether HS or graduate education is used as the reference group will change the results.
  - For a detailed demonstration of this problem, see Jones and Kelley (1984)
  - Recent work used normalized coefficients to address the issue. Check this well-cited SMR paper: Kim (2010)
  - It also suffers from the typical problem of OLS. Without a causal inference design, the "discrimination" part of the model may be due to unmeasured characteristics (e.g., years of education vs field of studies).