R Coding Demonstration Week 6: Race and Voter Turnout

Matthew Blackwell

Gov 51 (Harvard)

Berkeley gender bias

- Graduate admissions data from Berkeley, 1973
- Acceptance rates:
  - Men: 8442 applicants, 44% admission rate
  - Women: 4321 applicants, 35% admission rate
- Evidence of discrimination toward women in admissions?
- This is a **marginal relationship**.
- What about the **conditional relationship** within departments?

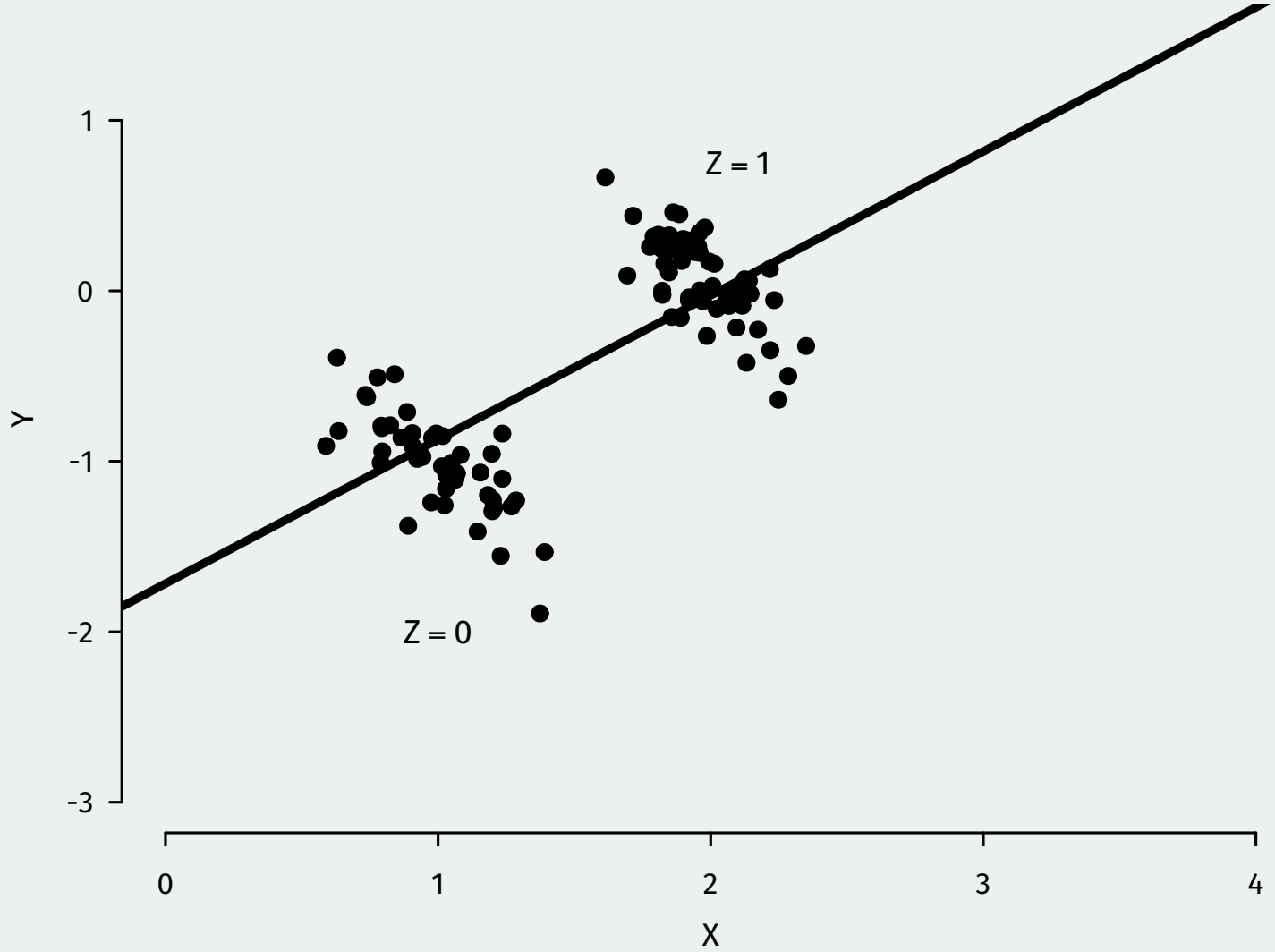
Berkeley gender bias, II

- Within departments:

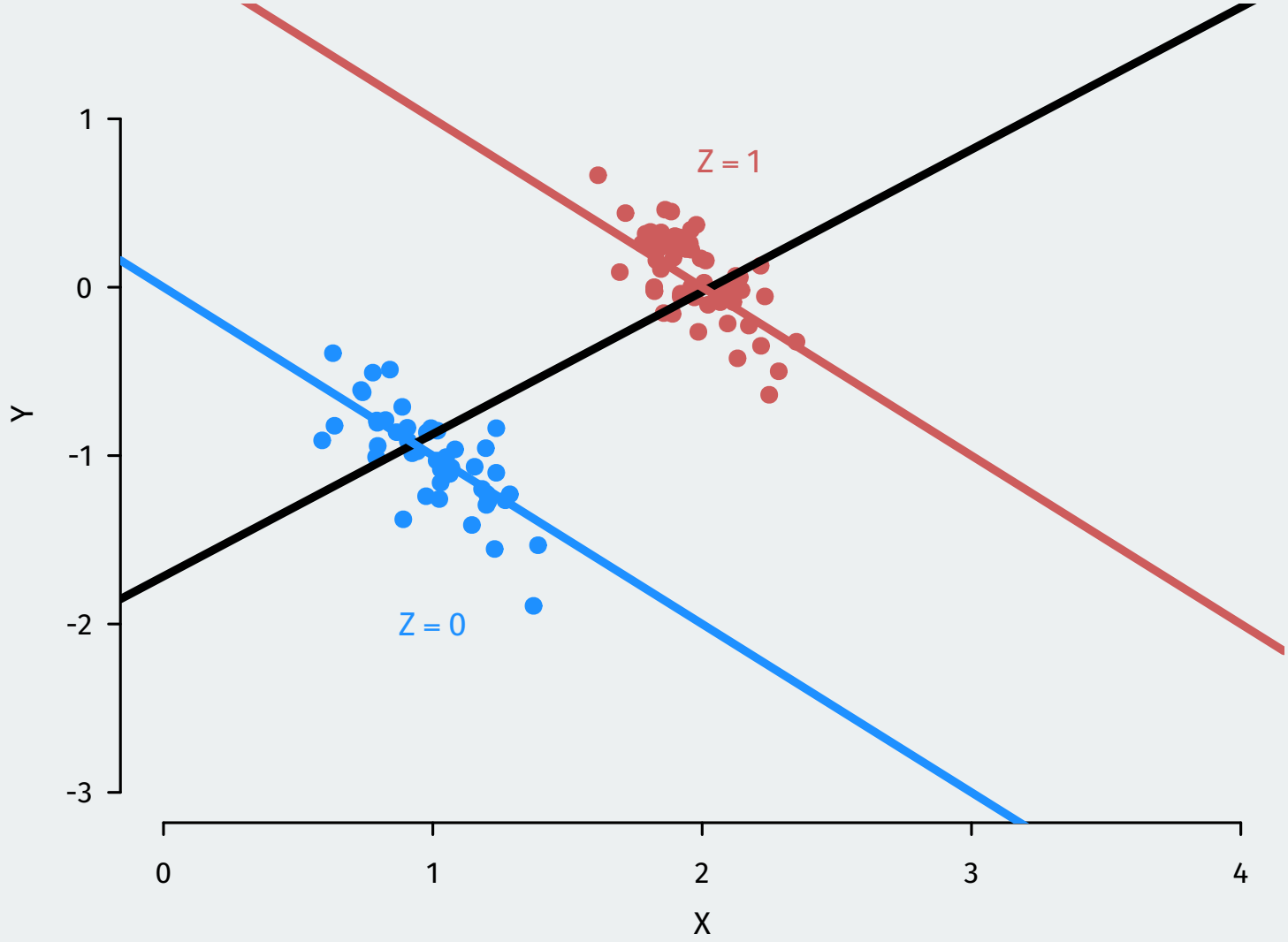
	Men		Women	
Dept	Applied	Admitted	Applied	Admitted
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%
D	417	33%	375	35%
E	191	28%	393	24%
F	373	6%	341	7%

- Within departments, women do somewhat better than men!
- Women apply to more challenging departments.
- Marginal relationships (admissions and gender) $\neq$ conditional relationship given third variable (department).

Simpson's paradox



Simpson's paradox



Why control for another variable

- Descriptive
  - Get a sense for the relationships in the data.
  - Conditional on the number of steps I've taken, does higher activity levels correlate with less weight?
- Predictive
  - We can usually make better predictions about the dependent variable with more information on independent variables.
- Causal
  - Alternative form of **statistical control** to block potential **confounding**.

Data Example

- Do co-ethnic candidates mobilizing voters?
  - e.g., Black voters turnout at higher rates for Black candidates? Higher Latino turnout for Latino candidates?
  - Lots of studies show this basic relationship.
- But what about confounders?
  - Black candidates more likely to run in districts with higher share of Black voters.
- Study by Bernard Fraga get better data on racial/ethnic make-up of voters:

  Fraga, Bernard. (2015) "Candidates or Districts? Reevaluating the Role of Race in Voter Turnout," American Journal of Political Science, Vol. 60, No. 1, pp. 97–122.

Data

- We'll focus on a subset of the Fraga data that focuses on Black voter turnout:

```
blackturnout <- read.csv("data/blackturnout.csv")
```

- Variables:

Name	Description
`year`	Year the election was held
`state`	State in which the election was held
`district`	District in which the election was held
`black_turnout`	Prop. of the black voting-age population that voted in general election
`black_share`	Prop. of a district's voting-age population that is black
`black_candidate`	1 if election includes a black candidate; 0 otherwise

Question 1

Run a regression with Black voter turnout (`black_turnout`) as the dependent variable and Black share of the voting-age population (`black_share`) as the independent variable.

Provide an interpretation of each coefficient and calculate the R^2 and interpret it.

Answer 1

```
cvap_fit <- lm(black_turnout ~ black_share, data = blackturnout)
cvap_fit
```

```
##
## Call:
## lm(formula = black_turnout ~ black_share, data = blackturnout)
##
## Coefficients:
## (Intercept)  black_share
##       0.376         0.196
```

```
summary(cvap_fit)$r.squared
```

```
## [1] 0.0284
```

Question 2

Run a regression with Black voter turnout (`black_turnout`) as the dependent variable and there being a Black candidate in the election (`black_candidate`) as the independent variable.

Provide an interpretation of each coefficient and calculate the R^2 and interpret it.

If you have time: calculate the RMSE for this model and the previous one and determine which variable better predicts turnout.

Answer 2

```
cand_fit <- lm(black_turnout ~ black_candidate, data = blackturnout)
cand_fit
```

```
##
## Call:
## lm(formula = black_turnout ~ black_candidate, data = blackturnout)
##
## Coefficients:
##     (Intercept)  black_candidate
##          0.3939           0.0616
```

```
summary(cand_fit)$r.squared
```

```
## [1] 0.0135
```

Question 3

Run a multiple regression with Black turnout as the dependent variable and `black_share` and `year` as the independent variables. Interpret the coefficients. Evaluate both the R^2 and the adjusted R^2.

Does the relationship between `black_share` and `black_turnout` change from the previous regression?

Answer 3

```
cvapyear_fit <- lm(black_turnout ~ black_share + year, data = blackturnout)
cvapyear_fit
```

```
##
## Call:
## lm(formula = black_turnout ~ black_share + year, data = blackturnout)
##
## Coefficients:
## (Intercept)  black_share         year
##   -11.29365      0.19453      0.00581
```

```
summary(cvapyear_fit)$r.squared
```

```
## [1] 0.0315
```

```
summary(cvapyear_fit)$adj.r.squared
```

```
## [1] 0.0299
```

Question 4

Run a multiple regression with Black turnout as the dependent variable and `black_candidate` and `black_share` as the independent variables. Interpret the coefficients. Evaluate both the R^2 and the adjusted R^2.

Does the relationship between having a Black candidate and Black turnout change from the previous models?

Answer 4

```
candcvap_fit <- lm(black_turnout ~ black_candidate + black_share, data = bl
candcvap_fit
```

```
##
## Call:
## lm(formula = black_turnout ~ black_candidate + black_share, data = blackturn
##
## Coefficients:
##     (Intercept)  black_candidate      black_share
##         0.37528         -0.00736          0.20739
```

```
summary(candcvap_fit)$r.squared
```

```
## [1] 0.0285
```

```
summary(candcvap_fit)$adj.r.squared
```

```
## [1] 0.027
```

Question 5

Create a factor version of the `year` variable and run a regression with `black_turnout` as the dependent variable and this year factor as an independent variable. What does R do with these factors? How do we interpret the coefficients?

Run the same regression without the intercept. How do we interpret these coefficients?

Answer 5

```
table(blackturnout$year)
```

```
##
## 2006 2008 2010
##  398  416  423
```

```
blackturnout$year_fac <- as.factor(blackturnout$year)

year_fit <- lm(black_turnout ~ year_fac, data = blackturnout)
year_fit
```

```
##
## Call:
## lm(formula = black_turnout ~ year_fac, data = blackturnout)
##
## Coefficients:
##  (Intercept)  year_fac2008  year_fac2010
##       0.2934        0.2899        0.0301
```

Answer 5 (cont'd)

```
year_noint_fit <- lm(black_turnout ~ year_fac - 1, data = blackturnout)
year_noint_fit
```

```
##
## Call:
## lm(formula = black_turnout ~ year_fac - 1, data = blackturnout)
##
## Coefficients:
## year_fac2006  year_fac2008  year_fac2010
##        0.293         0.583         0.324
```

Question 6

Run a regression with `black_turnout` as the dependent variable and `state` as an independent variable. How do we interpret the coefficients? (Hint: use `table()` or `unique()` to find which state is omitted.)

Run the same regression without the intercept. How do we interpret these coefficients?

Answer 6

```
unique(blackturnout$state)
```

```
##  [1] "AK" "AL" "AR" "AZ" "CA" "CO" "CT" "DE" "FL" "GA" "IA" "IL" "IN"
## [14] "KS" "KY" "LA" "MA" "MD" "ME" "MI" "MN" "MO" "MS" "NC" "NE" "NH"
## [27] "NJ" "NM" "NV" "NY" "OH" "OK" "OR" "PA" "RI" "SC" "TN" "TX" "UT"
## [40] "WA" "WI" "WV"
```

```
state_fit <- lm(black_turnout ~ state, data = blackturnout)
## state_fit
```

Answer 6 (cont'd)

```
##
## Call:
## lm(formula = black_turnout ~ state, data = blackturnout)
##
## Coefficients:
## (Intercept)      stateAL      stateAR      stateAZ      stateCA
##      0.5512      -0.1188      -0.1885      -0.1954      -0.1549
##     stateCO      stateCT      stateDE      stateFL      stateGA
##     -0.0470      -0.1210      -0.0185      -0.1243      -0.1394
##     stateIA      stateIL      stateIN      stateKS      stateKY
##     -0.0301      -0.1829      -0.2029      -0.1652      -0.1145
##     stateLA      stateMA      stateMD      stateME      stateMI
##     -0.0904      -0.1934      -0.0623       0.3537      -0.0318
##     stateMN      stateMO      stateMS      stateNC      stateNE
##     -0.0864      -0.1622      -0.1494      -0.1020      -0.1604
##     stateNH      stateNJ      stateNM      stateNV      stateNY
##      0.0495      -0.1521      -0.1286      -0.1610      -0.1976
##     stateOH      stateOK      stateOR      statePA      stateRI
##     -0.1050      -0.0327       0.1368      -0.2161      -0.1205
##     stateSC      stateTN      stateTX      stateUT      stateWA
##     -0.1133      -0.1452      -0.2604      -0.1640      -0.2047
##     stateWI      stateWV
##     -0.1712      -0.1712
```

Answer 6 (cont'd)

```
state_noint_fit <- lm(black_turnout ~ state - 1, data = blackturnout)
state_noint_fit
```

```
##
## Call:
## lm(formula = black_turnout ~ state - 1, data = blackturnout)
##
## Coefficients:
## stateAK  stateAL  stateAR  stateAZ  stateCA  stateCO  stateCT
##   0.551    0.432    0.363    0.356    0.396    0.504    0.430
## stateDE  stateFL  stateGA  stateIA  stateIL  stateIN  stateKS
##   0.533    0.427    0.412    0.521    0.368    0.348    0.386
## stateKY  stateLA  stateMA  stateMD  stateME  stateMI  stateMN
##   0.437    0.461    0.358    0.489    0.905    0.519    0.465
## stateMO  stateMS  stateNC  stateNE  stateNH  stateNJ  stateNM
##   0.389    0.402    0.449    0.391    0.601    0.399    0.423
## stateNV  stateNY  stateOH  stateOK  stateOR  statePA  stateRI
##   0.390    0.354    0.446    0.519    0.688    0.335    0.431
## stateSC  stateTN  stateTX  stateUT  stateWA  stateWI  stateWV
##   0.438    0.406    0.291    0.387    0.346    0.380    0.380
```