NRES 776 Lecture 19

GLM - multiple variables and Simpson’s paradox

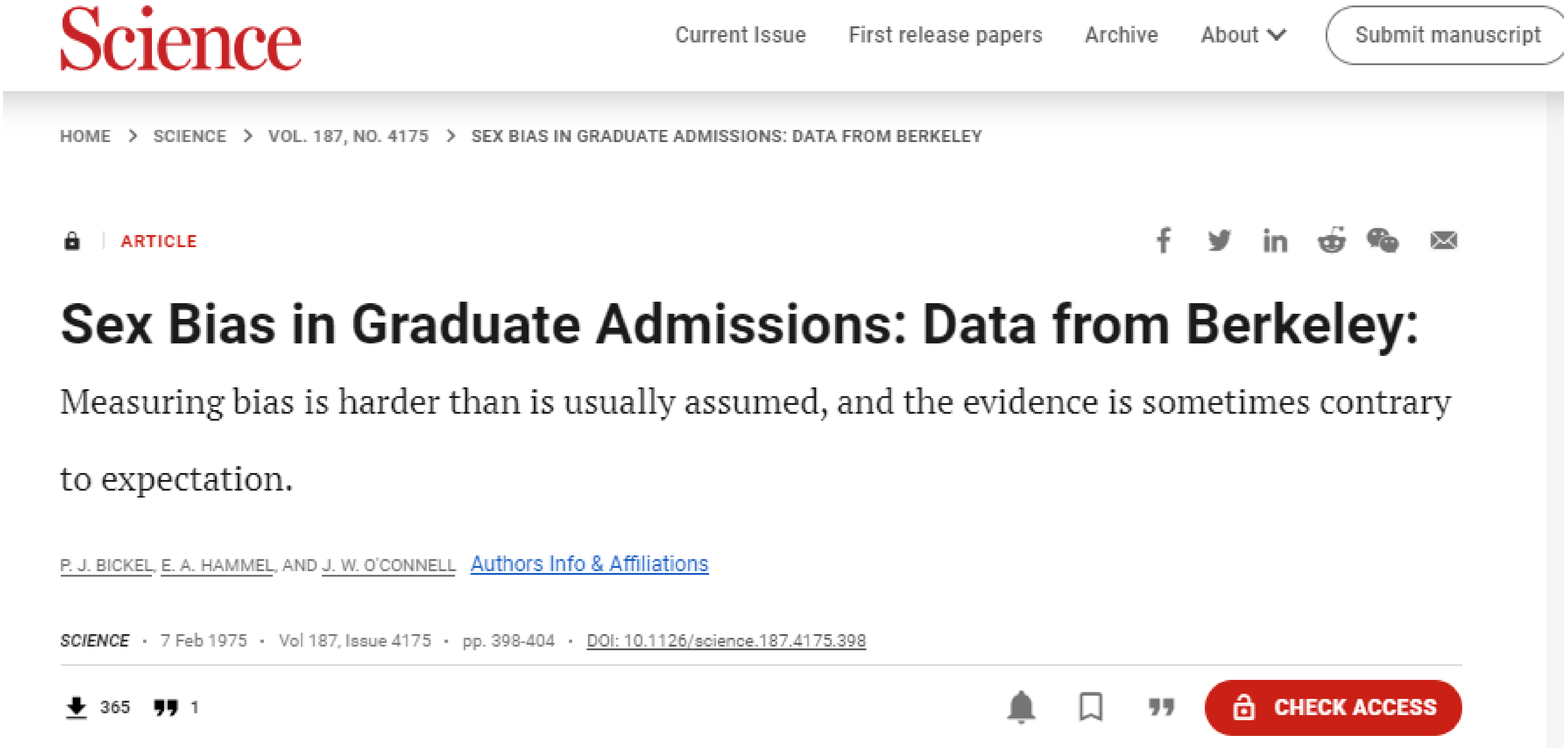
Sunny Tseng

Our schedule today

- Announcement (3 min)
  - zoom recording
  - topic for next week's lab
- Binomial regression with multiple variables (30 min)
- Wrap up (5 min)

UCB Admission data set

Aggregate data on applicants to graduate school at Berkeley for the six largest departments in 1973 classified by admission and sex.



Science

Current Issue First release papers Archive About Submit manuscript

HOME > SCIENCE > VOL. 187, NO. 4175 > SEX BIAS IN GRADUATE ADMISSIONS: DATA FROM BERKELEY

ARTICLE

Sex Bias in Graduate Admissions: Data from Berkeley:

Measuring bias is harder than is usually assumed, and the evidence is sometimes contrary to expectation.

P. J. BICKEL, E. A. HAMMEL, AND J. W. O'CONNELL Authors Info & Affiliations

SCIENCE • 7 Feb 1975 • Vol 187, Issue 4175 • pp. 398-404 • DOI: 10.1126/science.187.4175.398

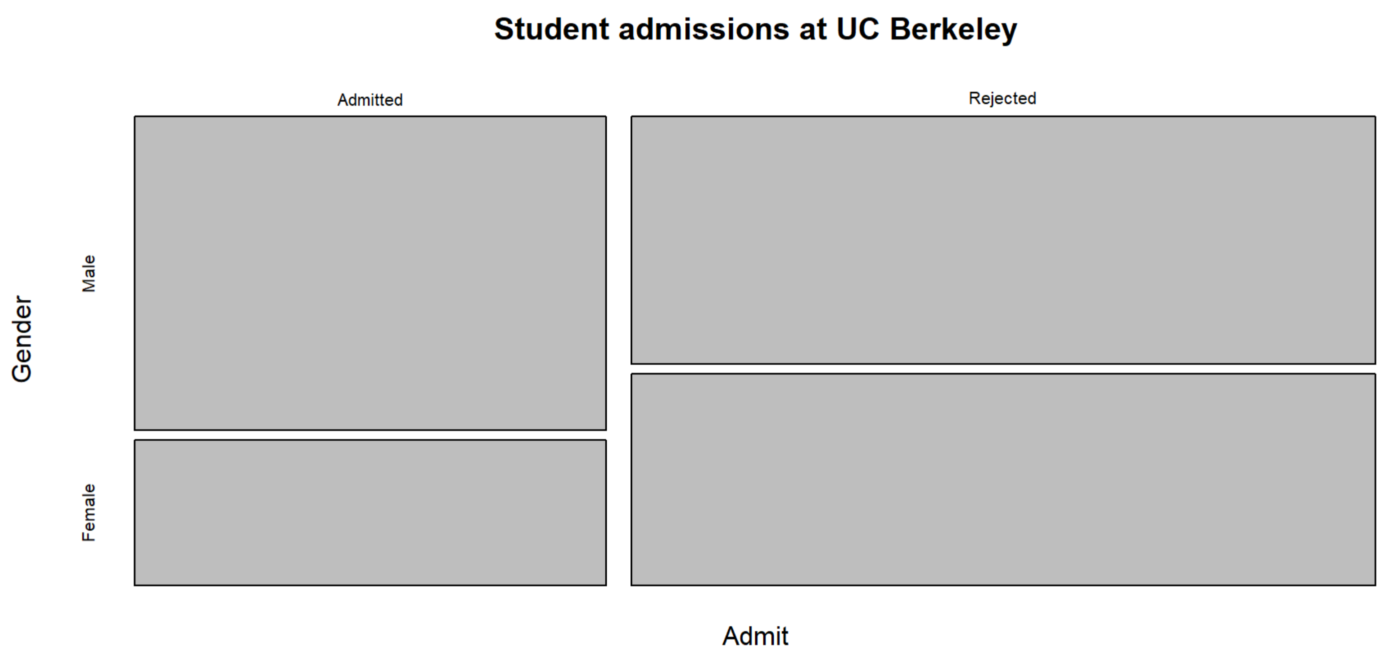
365 1 CHECK ACCESS

Exploratory data visualization

```r
apply(UCBAdmissions,
      c(1, 2),
      sum)
```

```
           Gender
Admit       Male Female
  Admitted 1198    557
  Rejected 1493   1278
```

```r
mosaicplot(apply(UCBAdmissions, c(1, 2), sum),
           main = "Student admissions at UC Berkeley")
```



Model formulation (glm_1)

$$logit(admission_i) = \beta_0 + \beta_1 genderM_i$$

```r
glm_1 <- glm(formula = cbind(Admitted, Rejected) ~ Gender,
             data = UCBAdmissions_clean,
             family = "binomial")

glm_1 %>% summary
```

```

Call:
glm(formula = cbind(Admitted, Rejected) ~ Gender, family = "binomial",
    data = UCBAdmissions_clean)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-16.7915   -4.7613   -0.4365    5.1025   11.2022

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.83049    0.05077 -16.357   <2e-16 ***
GenderMale   0.61035    0.06389   9.553   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 877.06  on 11  degrees of freedom
Residual deviance: 783.61  on 10  degrees of freedom
AIC: 856.55
```

```
Number of Fisher Scoring iterations: 4
```

Model goodness of fit

```r
glm_null <- glm(formula = cbind(Admitted, Rejected) ~ 1,
                data = UCBAdmissions_clean,
                family = "binomial")

lrtest(glm_1, glm_null)
```

```
Likelihood ratio test

Model 1: cbind(Admitted, Rejected) ~ Gender
Model 2: cbind(Admitted, Rejected) ~ 1
  #Df  LogLik Df  Chisq Pr(>Chisq)
1   2 -426.27
2   1 -473.00 -1 93.449  < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Coef. interpretation

$$logit(p_i) = \beta_0 + \beta_1 genderM_i$$

$$\beta_1 = -0.83$$

$$\beta_1 = 0.61$$

For female applicants:

$$logit(p_F) = log(\frac{p_F}{1-p_F}) = \beta_0$$

$$\frac{p_F}{1-p_F} = exp(\beta_0) = 0.43$$

Coef. interpretation (con'd)

For male applicants:

$$logit(p_M) = log(\frac{p_M}{1 - p_M}) = \beta_0 + \beta_1$$

$$\frac{p_M}{1 - p_M} = exp(\beta_0 + \beta_1)$$

$$\frac{Odd(p_M)}{Odd(p_F)} = exp(\beta_1) = 1.84$$

- Male students has 1.84 times higher odds in getting admitted in the university
- Evidence of sex bias in admission practices!

Model formulation (glm_2)

$$logit(admission_i) = \beta_0 + \beta_1 genderM_i + \beta_2 deptB_i + \beta_3 deptC_i + \beta_4 deptD_i + beta_5 deptE_i + beta_6 de$$

```r
glm_2 <- glm(formula = cbind(Admitted, Rejected) ~ Gender + Dept,
                data = UCBAdmissions_clean,
                family = "binomial")
glm_2 %>% summary()
```

```

Call:
glm(formula = cbind(Admitted, Rejected) ~ Gender + Dept, family = "binomial",
    data = UCBAdmissions_clean)

Deviance Residuals:
      1        2        3        4        5        6        7        8
-1.2487   3.7189  -0.0560   0.2706   1.2533  -0.9243   0.0826  -0.0858
      9       10       11       12
 1.2205  -0.8509  -0.2076   0.2052

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.68192    0.09911   6.880 5.97e-12 ***
GenderMale  -0.09987    0.08085  -1.235    0.217
DeptB       -0.04340    0.10984  -0.395    0.693
DeptC       -1.26260    0.10663 -11.841  < 2e-16 ***
DeptD       -1.29461    0.10582 -12.234  < 2e-16 ***
DeptE       -1.73931    0.12611 -13.792  < 2e-16 ***
DeptF       -3.30648    0.16998 -19.452  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 877.056  on 11  degrees of freedom
Residual deviance:  20.204  on  5  degrees of freedom
AIC: 103.14
```

Model goodness of fit

```r
glm_null <- glm(formula = cbind(Admitted, Rejected) ~ 1,
                data = UCBAdmissions_clean,
                family = "binomial")

lrtest(glm_1, glm_null)
```

```
Likelihood ratio test

Model 1: cbind(Admitted, Rejected) ~ Gender
Model 2: cbind(Admitted, Rejected) ~ 1
  #Df  LogLik Df  Chisq Pr(>Chisq)
1   2 -426.27
2   1 -473.00 -1 93.449  < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Coef. interpretation

$$logit(admission_i) = \beta_0 + \beta_1 genderM_i + \beta_2 deptB_i + \beta_3 deptC_i + \beta_4 deptD_i + beta_5 deptE_i + beta_6 de$$

$$\beta_0 = 0.6$$

$$\beta_1 = -0.09$$

$$\beta_2 = -0.04$$

For female student in dept B:

$$logit(p_{FB}) = log(\frac{p_{FB}}{1 - p_{FB}}) = \beta_0 + \beta_2$$

$$\frac{p_{FB}}{1 - p_{FB}} = exp(\beta_0 + \beta_2) = 1.89$$

Coef. interpretation (con'd)

For male student in dept B:

$$logit(p_{MB}) = log(\frac{p_{MB}}{1 - p_{MB}}) = \beta_0 + \beta_1 + \beta_2$$

$$\frac{p_{MB}}{1 - p_{MB}} = exp(\beta_0 + \beta_1 + \beta_2)$$

$$\frac{Odd(p_{MB})}{Odd(p_{FB})} = exp(\beta_1) = 0.9$$

- Male students has 0.9 times less odds in getting admitted in the department B
- In general, male students has 0.9 times less odds in getting admitted in the university
- Evidence of sex bias in admission practices! <- no!

Comparison between models

```
glm_3 <- glm(formula = cbind(Admitted, Rejected) ~ Dept,
             data = UCBAdmissions_clean,
             family = "binomial")
```

- `glm_null`: null model
- `glm_1`: include Gender
- `glm_2`: include Gender and Dept
- `glm_3`: include Dept

Compare AIC values

GLM model	AIC
`glm_null`	948
`glm_1`	856
`glm_2`	103
`glm_3`	102

Comparison between models

Use likelihood ratio test for nested models

```r
lrtest(glm_1, glm_2) # select glm_2
```

```
Likelihood ratio test

Model 1: cbind(Admitted, Rejected) ~ Gender
Model 2: cbind(Admitted, Rejected) ~ Gender + Dept
  #Df  LogLik Df Chisq Pr(>Chisq)
1   2 -426.27
2   7  -44.57  5 763.4  < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

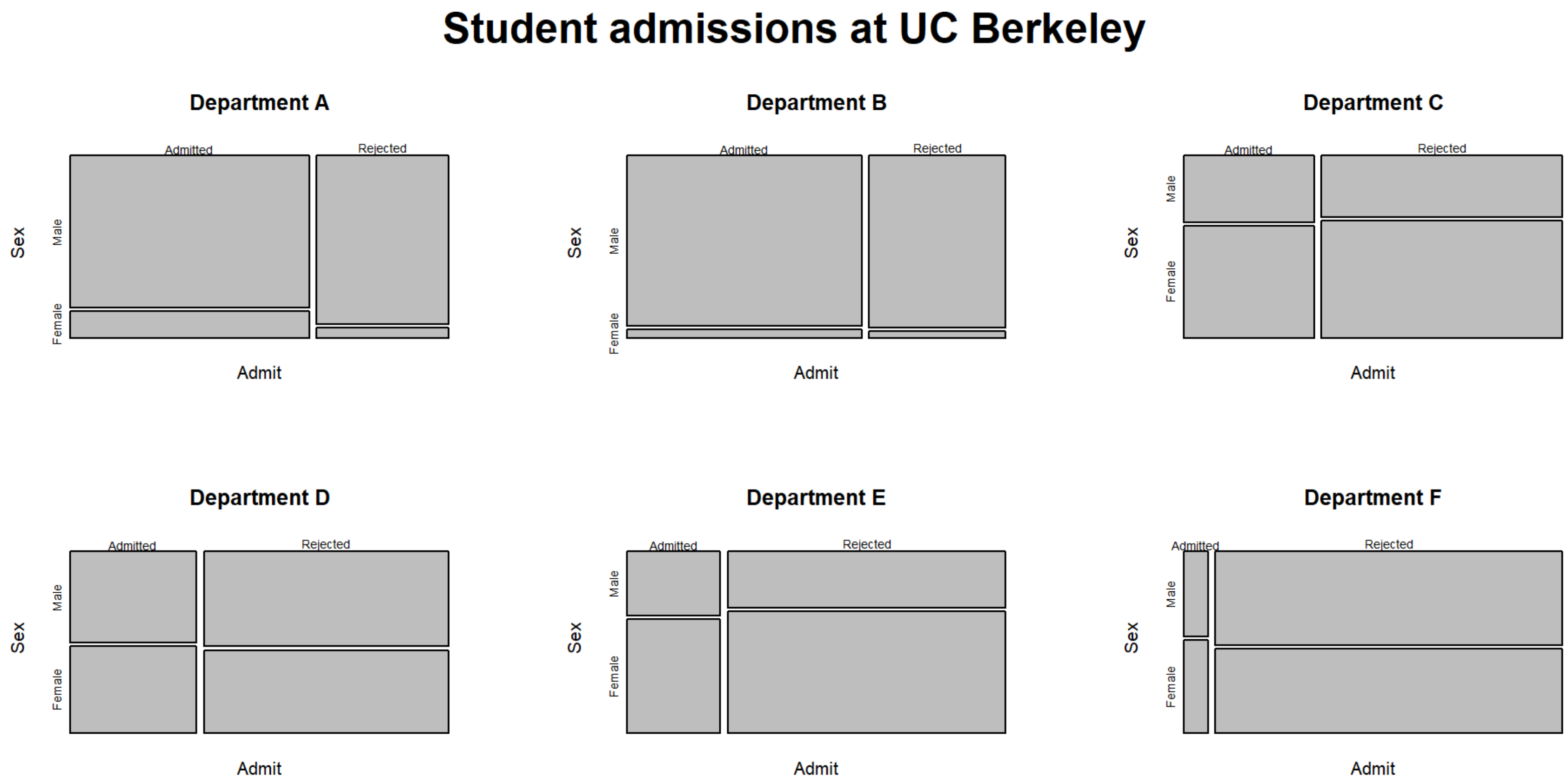
```r
lrtest(glm_2, glm_3) # select glm_3
```

```
Likelihood ratio test

Model 1: cbind(Admitted, Rejected) ~ Gender + Dept
Model 2: cbind(Admitted, Rejected) ~ Dept
  #Df  LogLik Df  Chisq Pr(>Chisq)
1   7 -44.572
2   6 -45.338 -1 1.5312     0.2159
```

What is actually going on

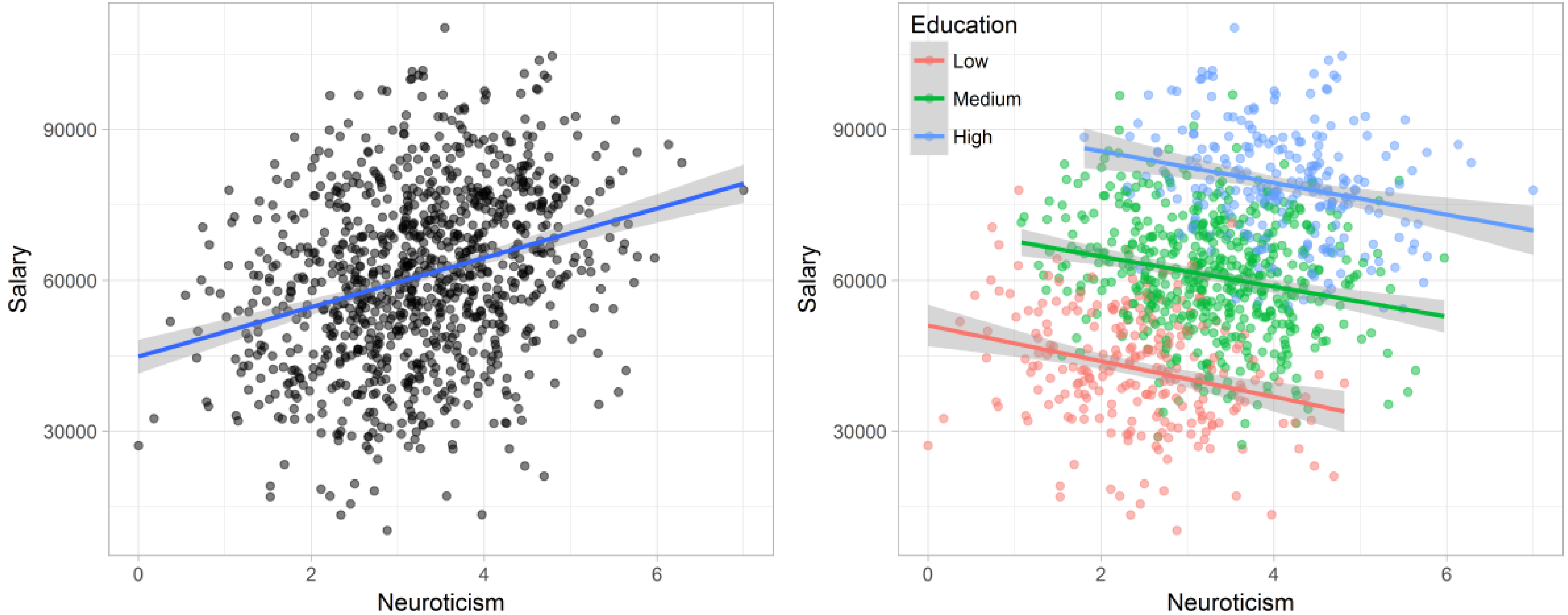
- Admission rate is actually influenced by department, not gender
- Just happen to be that more male students applied to the department with higher admission rate



Simpson’s Paradox

A phenomenon in probability and statistics in which a trend appears in several groups of data but disappears or reverses when the groups are combined.

Which means, in a easier language, missing of important variable in a model.



Wrap up

What we learned today

- Multiple GLM with Binomial regression
- Compare models using AIC and likelihood ratio test
- The importance of including critical variable in the model

Next time

- Next Tuesday in person lecture with Lisa
- Next Thursday lab 10, virtual on zoom