SOC-GA 2332 Intro to Stats Lab 10

Risa Gelles-Watnick

11/07/2025

Part 0: Logistics

- Lab next week (11/14) will end at 12:15pm
- November 26th is a legislative Friday, so lab will meet Wednesday that week instead of Friday

Part 1: Quiz #4 Review & Extensions

APPENDIX B

TABLE B1

LOGISTIC REGRESSION OF THE EFFECTS OF CRIMINAL RECORD AND RACE ON APPLICANTS' LIKELIHOOD OF RECEIVING A CALLBACK

	Coefficient	Robust SE
Criminal record	−.99	.24***
Black	−1.25	.28***
Criminal record × black ...	−.29	.38

NOTE.—SEs are corrected for clustering on employer ID in order to account for the fact that these data contain two records per employer (i.e., criminal record versus no criminal record). This model also controls for location (city vs. suburb) and contact with the employer, variables that mediate the relationship between race, crime, and employer responses.

*** $P < .001$.

1. What is the regression formula for this model? (Assume an intercept of $\beta_0 = .54$)

2. What do these coefficients represent? If you exponentiate the value of the coefficients, what do they represent now?

3. Ignoring the interaction term and holding other factors constant, what is the ratio between the odds that a callback will happen when the job seeker has a criminal record and the odds that a callback will happen when the job seeker does not have a criminal record?

4. Ignoring the interaction term and holding other factors constant, what is the ratio between the odds that a callback will happen when the job seeker is White and the odds that a callback will happen when the job seeker is Black?

5. Ignoring the interaction effect, write a sentence causally interpreting the effect of having a criminal record on prospects of receiving a callback.

6. Now using the interaction term, among black job seekers, what is the ratio between the odds that a callback will happen when the job seeker has a criminal record and the odds that a callback will happen when the job seeker does not have a criminal record?

7. What is the probability of a White job seeker with a criminal record receiving a callback? What is the probability of a Black job seeker without a criminal record receiving a callback? Write a sentence comparing the chances of receiving a callback for a White job seeker with a criminal record and a Black job seeker without a criminal record.

Part 2: 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 $(n - 1 - \text{number of IVs in the complete model})$. The *df* of the F test statistic is (df_1, df_2).

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

```
knitr::opts_chunk$set(echo = TRUE,
                      cache = FALSE,
                      fig.align = "center",
                      fig.width = 4.5,
                      fig.height = 4,
                      letina = TRUE)


pacman::p_load(
  tidyverse,
  stargazer,
  psych
)
```

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 previous labs:

(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: Fri, Nov 07, 2025 - 10:01:37

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

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

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

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

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

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

- *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.

```
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

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