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 $\beta_0 = .54$)

$$logit(p_{callback}) = log(\frac{p_{callback}}{1-p_{callback}}) = log(\frac{p_{callback}}{p_{nocallback}}) = .54 - .99 * \text{Crim} - 1.25 * \text{Black} - .29 * \text{Crim} * \text{Black} + \epsilon$$

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

The coefficients represent how much a change from the reference category affects the log odds of getting a callback. So for example, ignoring the interaction effect, having a criminal record decreases the log odds of getting a callback by .99. Note that this is an *additive* effect, just like linear regression. Having a criminal record *subtracts* .99 from your log odds.

When we exponeniate the coefficients (show how inverse function works), we get an odds ratio, which is the probability of getting a callback over the probability of not getting a callback. This represents how many times the odds of success increase by when the X variable increases by one unit. Note that this transformation of the coefficient represents a *multiplicative* effect. So having a criminal record (ignoring interaction) *multiplies* your odds by exp(-.99).

$$\beta_{crim} = \frac{\text{crim odds}}{\text{no crim odds}}$$

So there are two ratios going on:

- Odds, or the expoentiated outcome variable (Y) = $\frac{p_{success}}{p_{failure}}$
- Odds ratios, or the exponentiated coefficients (betas) = $\frac{odds_{alt-group}}{odds_{reference-group}}$

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

Holding all other factors constant, having a criminal record means a job seeker's odds of receiving a callback are .37 times that of those without a criminal record.

Perhaps a more intuitive way to report these answers is through percentages.

$$\text{percent change} = \frac{\text{crim odds - no crim odds}}{\text{no crim odds}} * 100$$

But our coefficient is already a ratio of new odds to old odds:

$$\beta_{crim} = \frac{\text{crim odds}}{\text{no crim odds}}$$

So we can rearrange our formula:

$$\text{percent change} = \frac{\text{crim odds - no crim odds}}{\text{no crim odds}} * 100 = \frac{\text{no crim odds}(\frac{\text{crim odds}}{\text{no crim odds}} - 1)}{\text{no crim odds}} * 100 = \frac{\text{no crim odds}(\beta_{crim} - 1)}{\text{no crim odds}} * 100$$

$$\text{percent change} = (\beta_{crim} - 1) * 100$$

```
(.37-1)*100
```

```
## [1] -63
```

Holding all other factors constant, having a criminal record decreases a job seeker's odds of receiving a callback by 63%. Notice again, this is a multiplicative effect (we are not linearly adding the same value each time X increases by one unit).

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.

Remember that the odds are a ratio of probabilities, not a probability itself.

$$logit(p_{callback}) = log(\frac{p_{callback}}{1 - p_{callback}})$$

First, we plug our variable values into the regression equation.

$$\text{logit}(p_{\text{callback}}) = \log\left(\frac{p_{\text{callback}}}{1 - p_{\text{callback}}}\right) = .54 - .99 \times \text{Crim} - 1.25 \times \text{Black} - .29 \times \text{Crim} \times \text{Black} + \epsilon$$

```r
exp(.54-.99)
```

```
## [1] 0.6376282
```

Now we have our odds. To get probability from this, we plug it into the odds equation.

$$\frac{p_{callback}}{1-p_{callback}} = 0.6376$$

$$p_{callback} = 0.6376(1-p_{callback}) = 0.6376 - 0.6376 * p_{callback}$$

$$p_{callback} + 0.6376 * p_{callback} = 0.6376$$

$$1.6376 * p_{callback} = 0.6376$$
$$p_{callback} = \frac{0.6376}{1.6376} = .39$$

The probability of a White job seeker with a record receiving a callback is 39%.

$$log(\frac{p_{callback}}{1-p_{callback}}) = .54 - .99 * \text{Crim} - 1.25 * \text{Black} - .29 * \text{Crim} * \text{Black} = -1.25$$

```r
exp(.54-1.25)
```

```
## [1] 0.4916442
```

$$\frac{p_{callback}}{1-p_{callback}} = .49164$$

$$p_{callback} = 0.33$$

The probability of a Black job seeker without a record receiving a callback is 33%.

The probability of receiving a callback for White job seekers with a criminal record is 6 percentage *points* higher than it is for Black applicants with no 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 ($n - 1 -$ 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 in a previous class.

```
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 a previous class:

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

```r
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 - 14:13:12

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`:

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

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

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

```r
# 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
```

```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?

The null hypothesis is that all the variables included in the "full" model that are not in the simpler model have coefficients of 0. In this case, that means that the null hypothesis is that the coefficients for race = black and race = other are 0. Since the p-value of the F-statistic we calculate is $< .05$, we can reject the null hypothesis that these coefficients are 0.

- 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?

ANOVA vs. t-test: A two sample t-test compares the means of two groups, while ANOVA allows you to compare the means of three or more groups. However the theory behind these tests is slightly different. A t-test checks whether the difference in means between two groups is significantly different than one would expect under a null distribution. ANOVA tests whether there is more variance between two groups than within either of the groups, and if there is significantly more variation between the groups than within them, then the group means are declared different.

ANOVA vs. linear regression: An ANOVA is essentially a simple Y ~ X linear regression with a categorical IV, where X is dummies of the categorical variable.

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")
```

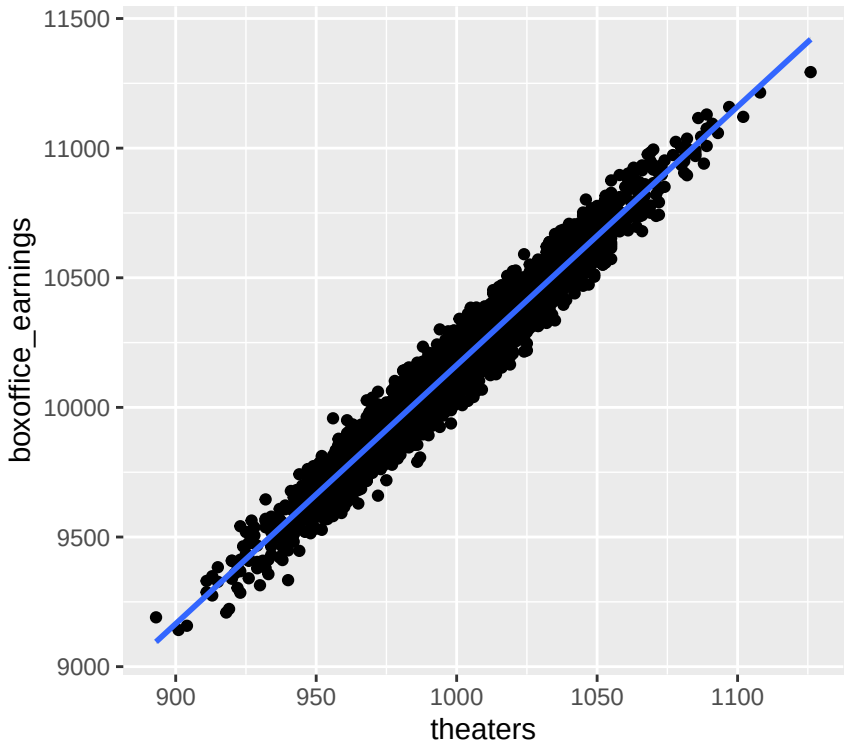
```
describe(boxoffice)
```

```
##                    vars    n     mean     sd   median  trimmed     mad     min
## X                     1 3000  1500.50 866.17  1500.50  1500.50 1111.95    1.00
## theaters              2 3000  1000.09  31.36  1000.00   999.95   31.13  893.00
## reviews               3 3000     5.80   3.22     6.00     5.98    4.45    0.00
## series                4 3000     0.11   0.32     0.00     0.02    0.00    0.00
## budget                5 3000    49.54  19.73    49.29    49.61   19.67  -19.62
## studio*               6 3000     2.00   0.82     2.00     2.00    1.48    1.00
## boxoffice_earnings    7 3000 10163.76 320.55 10163.61 10161.53  319.86 9141.51
##                       max   range  skew kurtosis    se
## X                  3000.00 2999.00  0.00    -1.20 15.81
## theaters           1126.00  233.00  0.04     0.00  0.57
## reviews              10.00   10.00 -0.26    -1.10  0.06
## series                1.00    1.00  2.44     3.97  0.01
## budget              111.13  130.75 -0.05    -0.15  0.36
## studio*               3.00    2.00  0.00    -1.50  0.01
## boxoffice_earnings 11293.32 2151.81  0.06    -0.01  5.85
```

- Check regression assumptions are met

```
# plotting theaters vs box office
boxoffice %>%
  ggplot(aes(x = theaters, y = boxoffice_earnings)) +
  geom_point() +
  geom_smooth(method = "lm")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



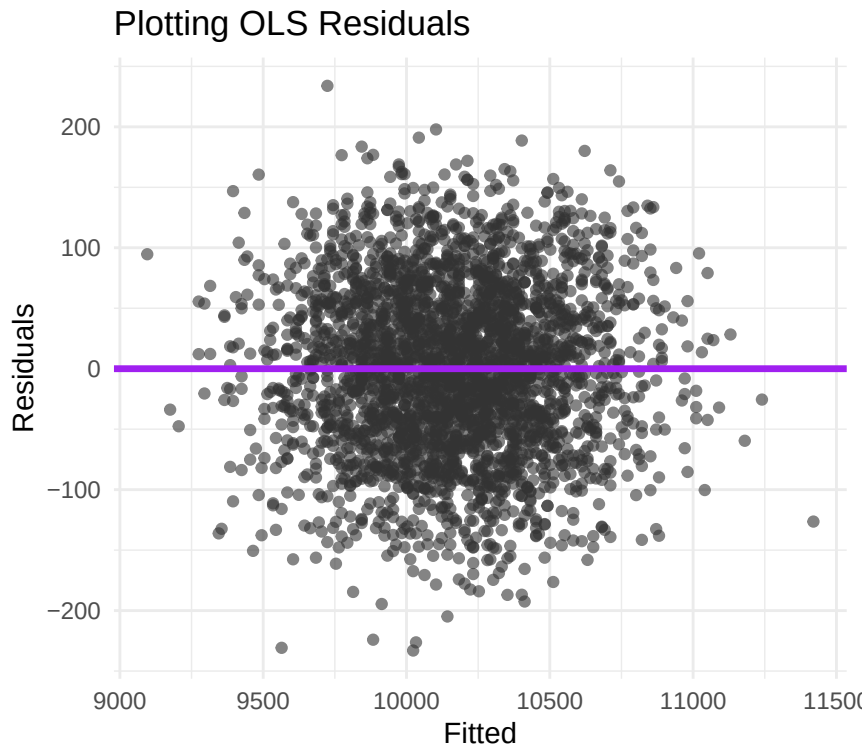
- Set up several candidates for regression models

```
# simplest regression model
mod1 <- boxoffice %>%
  lm(boxoffice_earnings ~ theaters,
     data = .)

# slightly more complicated model
mod2 <- boxoffice %>%
  lm(boxoffice_earnings ~ theaters + reviews,
     data = .)

# even more complicated model
mod3 <- boxoffice %>%
  lm(boxoffice_earnings ~ theaters + reviews*series,
     data = .)

# even more complicated!
mod4 <- boxoffice %>%
  lm(boxoffice_earnings ~ theaters + reviews*series + budget + studio,
     data = .)
```

```
# checking residuals are evenly spread accross levels of X and have a conditional mean of 0
ggplot(data = data.frame(
  fitted = fitted(mod1), # predicted values of Y for each value of X
  resid  = resid(mod1)
), aes(x = fitted, y = resid)) +
  geom_point(color = "grey20", alpha = 0.6) +        # scatterplot
  geom_hline(yintercept = 0, color = "purple", size = 1.2) + # reference line
  labs(
    title = "Plotting OLS Residuals",
```

```
    x = "Fitted",
    y = "Residuals"
  ) +
  theme_minimal()
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```




```
stargazer(mod1, mod2,
          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 - 14:13:13

```
stargazer(mod3, mod4,
          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 - 14:13:13

- 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).

```
anova(mod1, mod2, mod3, mod4)
```

```
## Analysis of Variance Table
##
## Model 1: boxoffice_earnings ~ theaters
## Model 2: boxoffice_earnings ~ theaters + reviews
## Model 3: boxoffice_earnings ~ theaters + reviews * series
```

Table 2:

	Dependent variable:	
	boxoffice_earnings	
	(1)	(2)
theaters	9.975***	9.988***
	(0.041)	(0.030)
reviews		14.817***
		(0.291)
Constant	187.990***	88.959***
	(40.754)	(29.888)
Observations	3,000	3,000
R^2	0.952	0.975
Adjusted R^2	0.952	0.974
Residual Std. Error	69.952 (df = 2998)	51.192 (df = 2997)
F Statistic	59,975.080*** (df = 1; 2998)	57,294.630*** (df = 2; 2997)
Note:		*p<0.1; **p<0.05; ***p<0.01

```
## Model 4: boxoffice_earnings ~ theaters + reviews * series + budget + studio
##   Res.Df      RSS Df Sum of Sq         F Pr(>F)
## 1   2998 14670222
## 2   2997  7853986  1   6816236 2721.4429 <2e-16 ***
## 3   2995  7495689  2    358297   71.5266 <2e-16 ***
## 4   2992  7493884  3      1805    0.2402 0.8683
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model 3 seems best!

- Display final model

```
stargazer(mod3,
          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 - 14:13:13

True data generating process

```
# create dataset
boxoffice <- data.frame(
  theaters = rpois(3000, lambda = 1000),
  reviews = rnorm(3000, mean = 6, sd = 4),
  series = rbinom(3000, size = 1, prob = .1),
  budget = rnorm(3000, mean = 50, sd = 20),
  studio = c(
    rep("warner_bros", 1000),
    rep("paramount", 1000),
    rep("disney", 1000)
```

Table 3:

	Dependent variable:	
	boxoffice_earnings	
	(1)	(2)
theaters	9.992***	9.993***
	(0.029)	(0.029)
reviews	15.266***	15.271***
	(0.301)	(0.301)
series	−6.531	−6.550
	(5.896)	(5.898)
budget		−0.007
		(0.046)
studioparamount		1.849
		(2.242)
studiowarner_bros		0.685
		(2.239)
reviews:series	−4.388***	−4.398***
	(0.902)	(0.902)
Constant	85.544***	83.904***
	(29.210)	(29.399)
Observations	3,000	3,000
R^2	0.976	0.976
Adjusted R^2	0.976	0.976
Residual Std. Error	50.027 (df = 2995)	50.046 (df = 2992)
F Statistic	30,032.430*** (df = 4; 2995)	17,148.430*** (df = 7; 2992)
Note:		*p<0.1; **p<0.05; ***p<0.01

Table 4:

	Dependent variable:
	boxoffice_earnings
theaters	9.992***
	(0.029)
reviews	15.266***
	(0.301)
series	−6.531
	(5.896)
reviews:series	−4.388***
	(0.902)
Constant	85.544***
	(29.210)
Observations	3,000
R^2	0.976
Adjusted R^2	0.976
Residual Std. Error	50.027 (df = 2995)
F Statistic	30,032.430*** (df = 4; 2995)
Note:	*p<0.1; **p<0.05; ***p<0.01

```
  )
) %>%
  mutate(
    reviews = ifelse(reviews > 10, 10, reviews),
    reviews = ifelse(reviews < 0, 0, reviews),
    reviews = round(reviews, digits = 0),

    boxoffice_earnings = 60 + 10 * theaters + 15 * reviews + (-5) * reviews *
      series + rnorm(3000, mean = 20, sd = 50)
  )

# saving to csv
write.csv(boxoffice, "data/boxoffice.csv")
```

$$boxoffice = 60 + 10 * theaters + 15 * reviews + -5 * reviews * series + \epsilon$$