Review of Regression Analysis

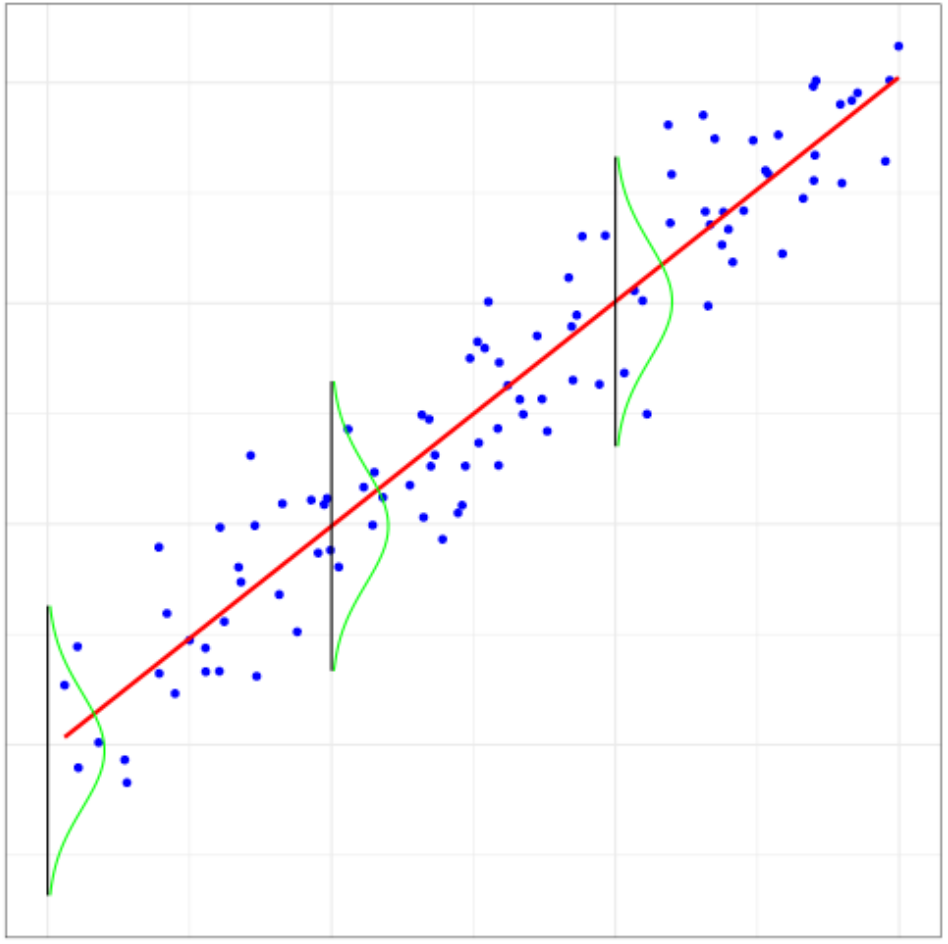
PSYC 575

Mark Lai

University of Southern California

2020/08/04 (updated: 2021-08-29)

Statistical Model



A set of statistical assumptions describing how data are generated

- Deterministic/fixed component

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots$$

- Stochastic/random component

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + e_i$$

$$e_i \sim N(0, \sigma)$$

Why Regression?

MLM is an extension of multiple regression to deal with data from multiple levels

Learning Objectives

Refresh your memory on regression

- Describe the statistical model
- Write out the model equations
- Simulate data based on a regression model
- Plot interactions

R Demonstration

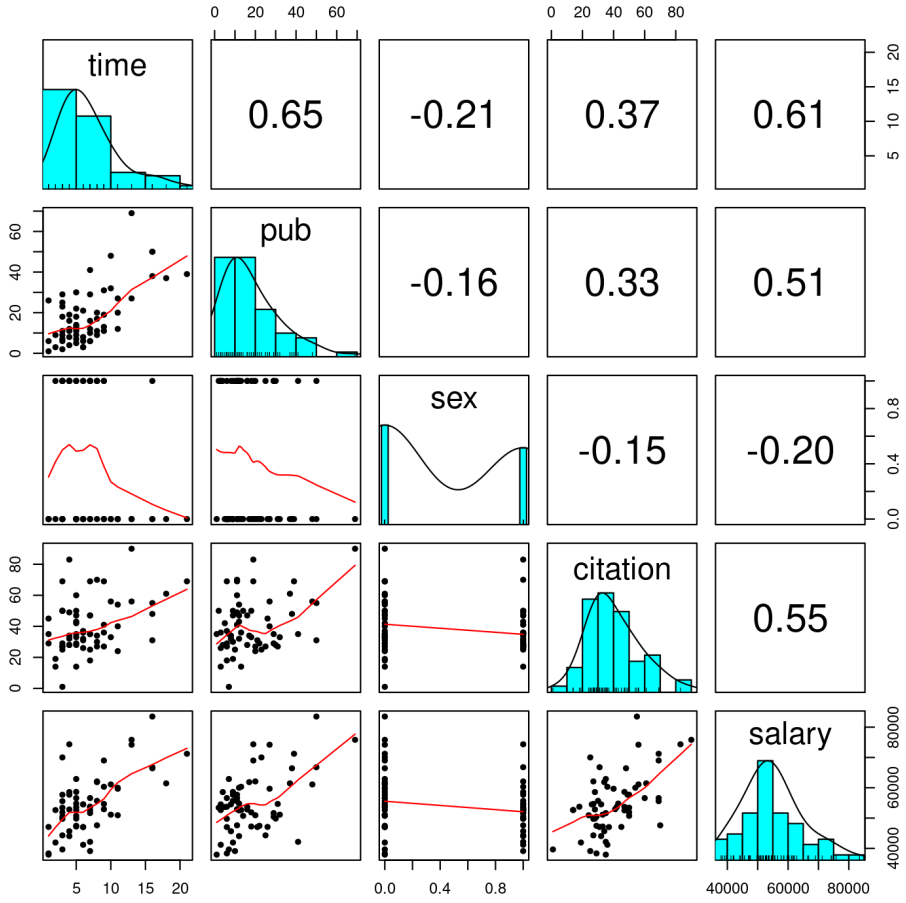
Salary Data

From Cohen, Cohen, West & Aiken (2003)

Examine factors related to annual salary of faculty in a university department

- time = years after receiving degree
- pub = # of publications
- sex = gender (0 = male, 1 = female)
- citation = # of citations
- salary = annual salary

Data Exploration



- How does the distribution of `salary` look?
- Are there more males or females in the data?
- How would you describe the relationship between number of publications and salary?

Simple Linear Regression

Sample regression line

Confidence intervals

Centering

Simulation

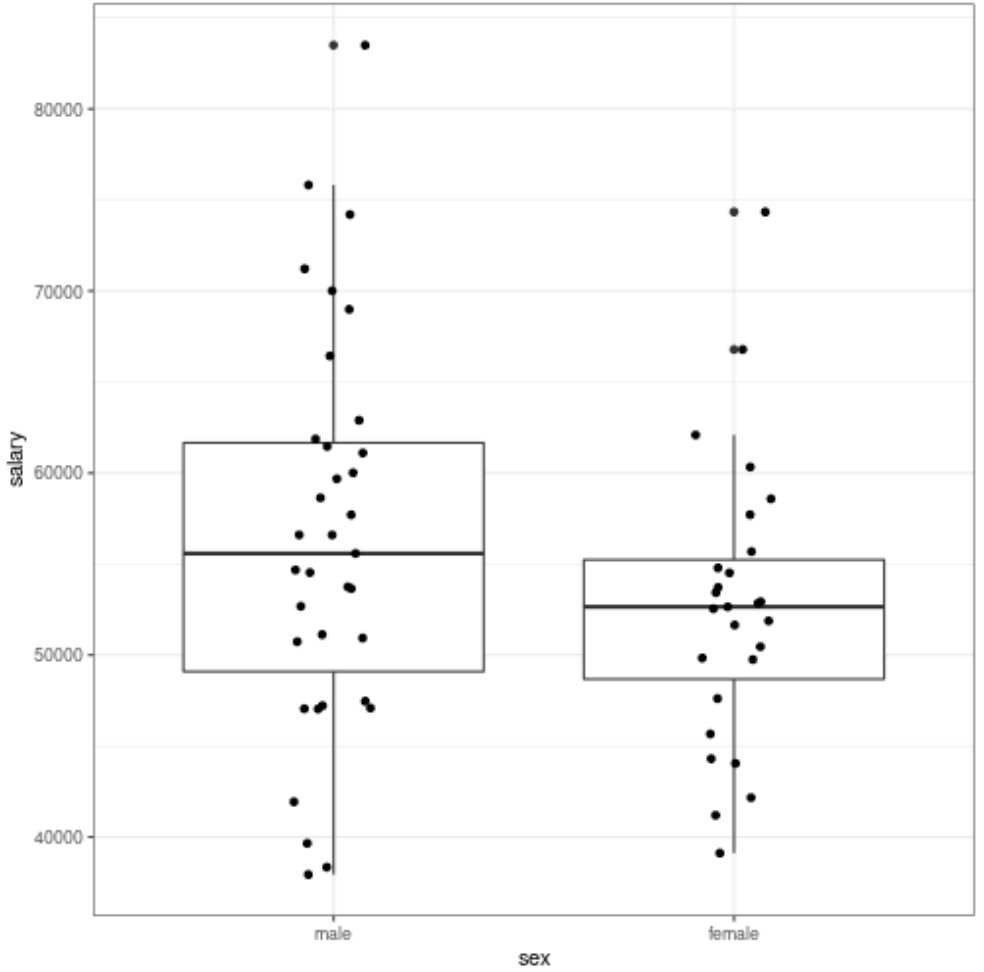
See lecture and R code

Categorical Predictors

Dummy Coding

With k categories, one needs $k-1$ dummy variables

The coefficients are differences relative to the reference group



Categorical Predictors

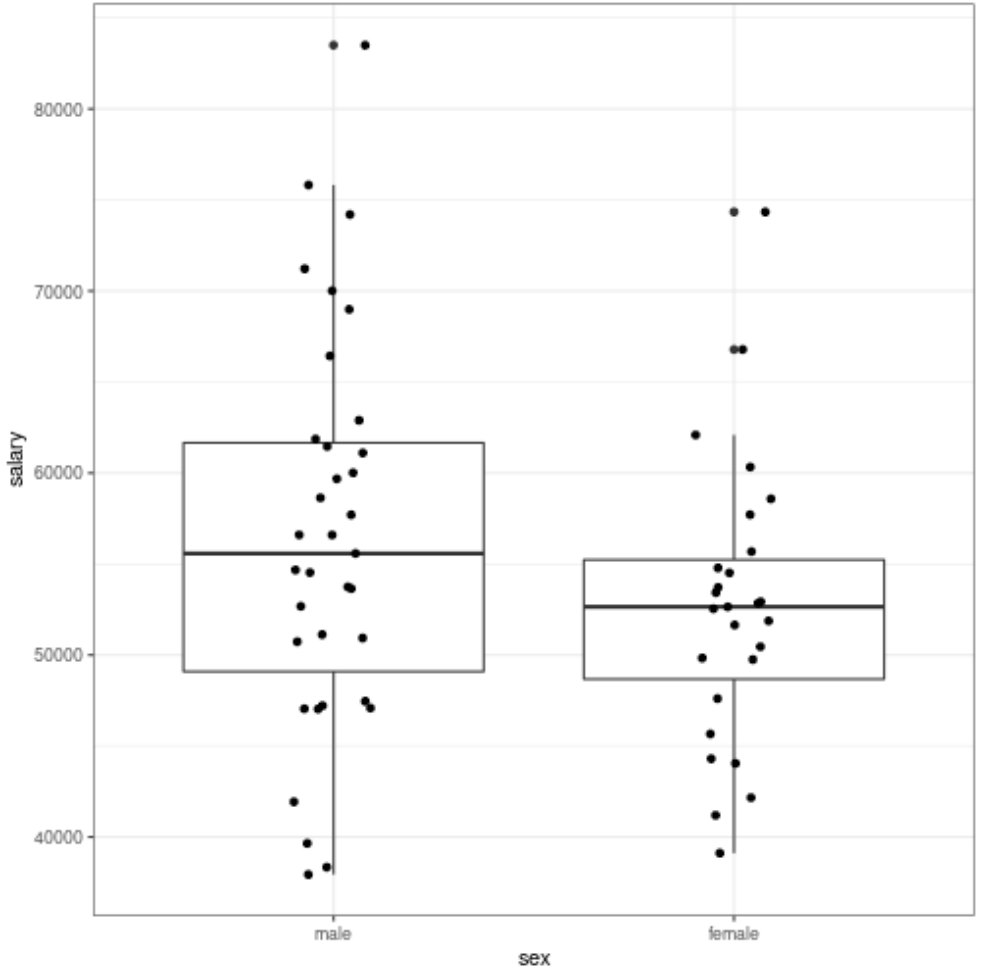
Dummy Coding

With k categories, one needs $k-1$ dummy variables

The coefficients are differences relative to the reference group

Male = 0

$$y = \beta_0 + \beta_1(0) = \beta_0$$



Categorical Predictors

Dummy Coding

With k categories, one needs $k-1$ dummy variables

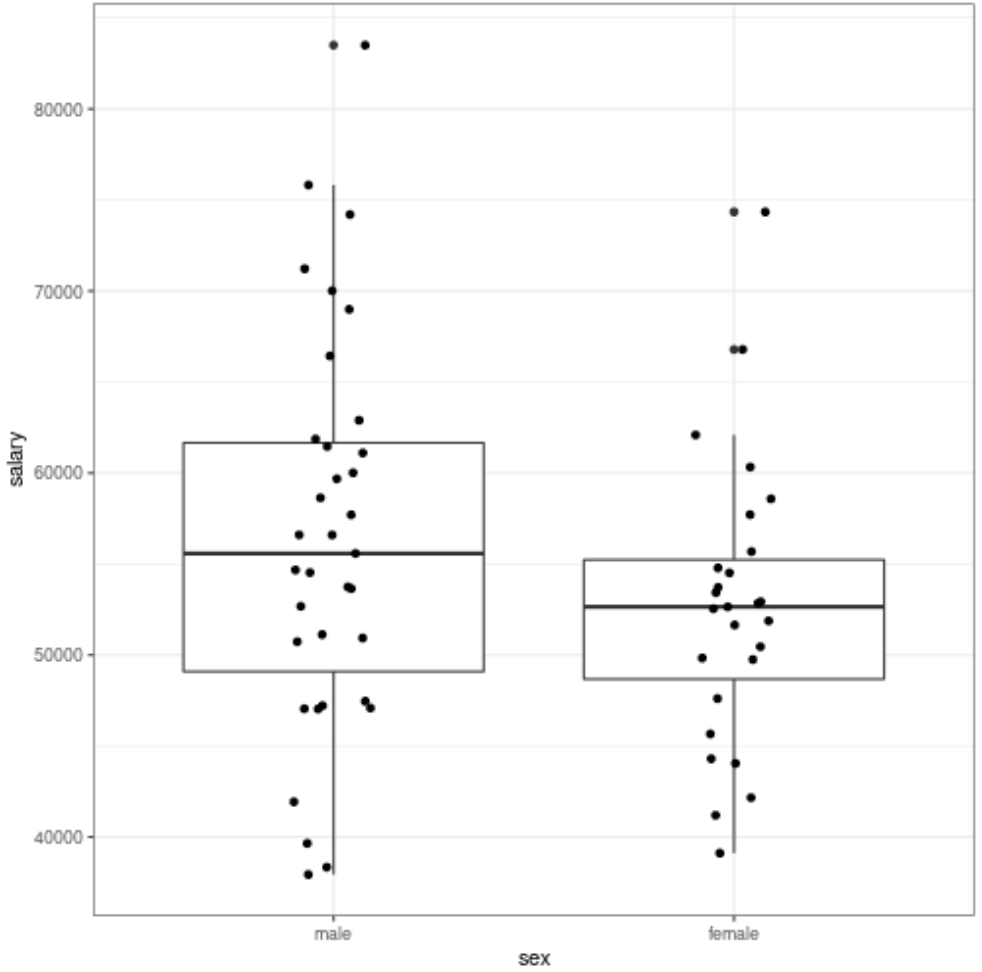
The coefficients are differences relative to the reference group

Male = 0

$$y = \beta_0 + \beta_1(0) = \beta_0$$

Female = 1

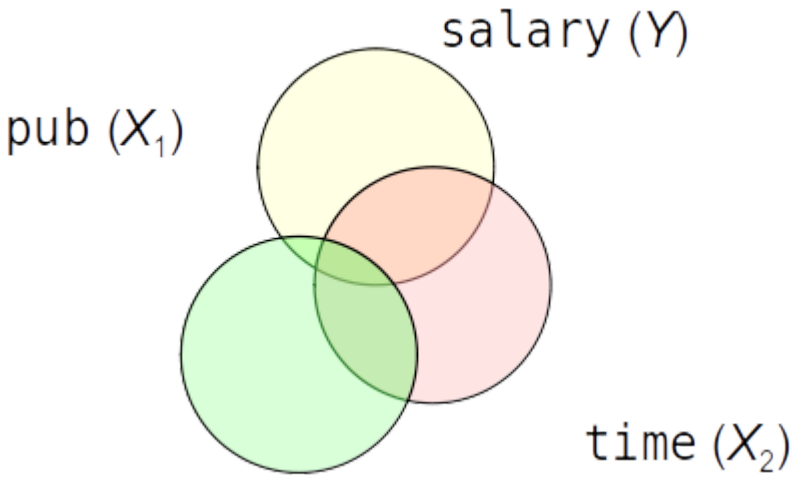
$$y = \beta_0 + \beta_1(1) = \beta_0 + \beta_1$$



Multiple Regression

Partial Effects

$$\text{salary}_i = \beta_0 + \beta_1 \text{pub}_i^c + \beta_2 \text{time}_i + e_i$$



Interpretations

Every unit increase in X is associated with β_1 unit increase in Y, **when all other predictors are constant**

Interactions

Regression slope of a predictor depends on another predictor

$$\begin{aligned}\widehat{\text{salary}} &= 54238 + 105 \times \text{pub}^c + 964 \times \text{time}^c \\ &\quad + 15(\text{pub}^c)(\text{time}^c)\end{aligned}$$

`time` = 7 $\Rightarrow$ `time_c` = 0.21

$$\begin{aligned}\widehat{\text{salary}} &= 54238 + 105 \times \text{pub}^c + 964(0.21) \\ &\quad + 15(\text{pub}^c)(0.21) \\ &= 54440 + 120 \times \text{pub}^c\end{aligned}$$

`time` = 15 $\Rightarrow$ `time_c` = 8.21

$$\begin{aligned}\widehat{\text{salary}} &= 54238 + 105 \times \text{pub}^c + 964(8.21) \\ &\quad + 15(\text{pub}^c)(8.21) \\ &= 62152 + 228 \times \text{pub}^c\end{aligned}$$

Interactions

Regression slope of a predictor depends on another predictor

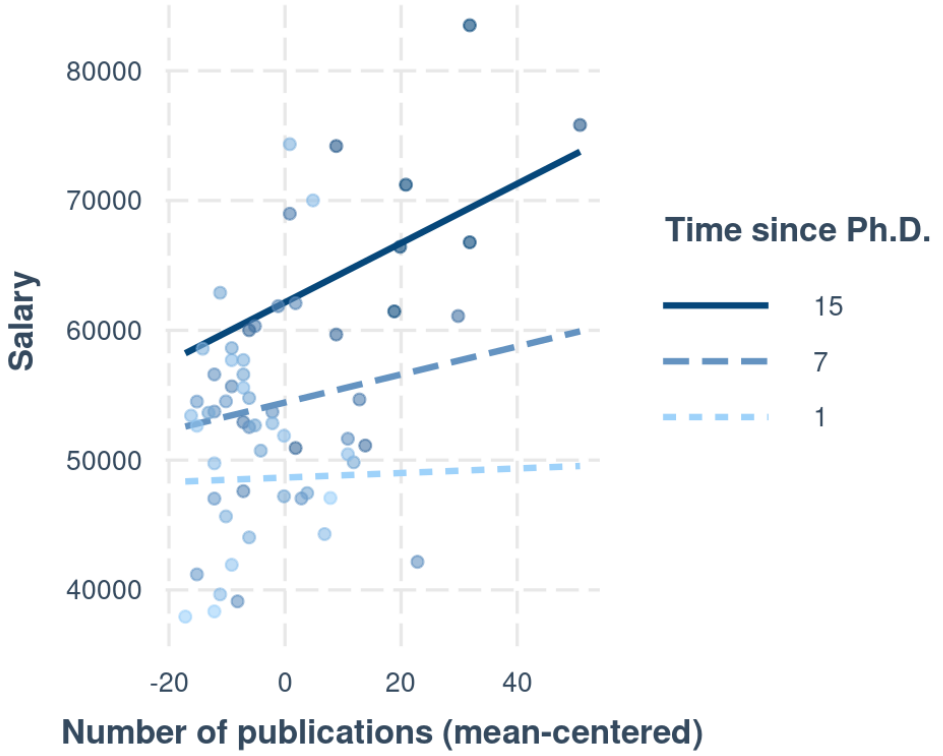
$$\widehat{\text{salary}} = 54238 + 105 \times \text{pub}^c + 964 \times \text{time}^c + 15(\text{pub}^c)(\text{time}^c)$$

`time` = 7 $\Rightarrow$ `time_c` = 0.21

$$\begin{aligned}\widehat{\text{salary}} &= 54238 + 105 \times \text{pub}^c + 964(0.21) \\ &\quad + 15(\text{pub}^c)(0.21) \\ &= 54440 + 120 \times \text{pub}^c\end{aligned}$$

`time` = 15 $\Rightarrow$ `time_c` = 8.21

$$\begin{aligned}\widehat{\text{salary}} &= 54238 + 105 \times \text{pub}^c + 964(8.21) \\ &\quad + 15(\text{pub}^c)(8.21) \\ &= 62152 + 228 \times \text{pub}^c\end{aligned}$$



modelsummary::msummary()

```r
library(modelsummary)
msummary(list("M3 + Interaction" = m4),
         fmt = "%.1f")  # keep one digit
```

	M3 + Interaction
(Intercept)	54238.1
	(1183.0)
pub_c	104.7
	(98.4)
time_c	964.2
	(339.7)
pub_c × time_c	15.1
	(17.3)
Num.Obs.	62
R2	0.399
R2 Adj.	0.368
AIC	1291.8
BIC	1302.4

Summary

Concepts

- What is a statistical model
- Linear/Multiple Regression
  - Centering
  - Categorical predictor
  - Interpretations
  - Interactions

HW 2

Try replicating the examples in the Rmd file