INFO251 – Applied Machine Learning

Lab 6

Suraj R. Nair

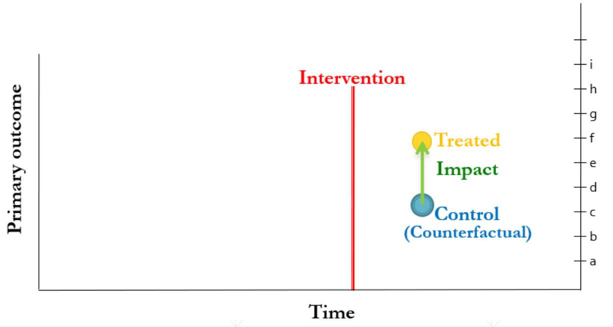
Announcements

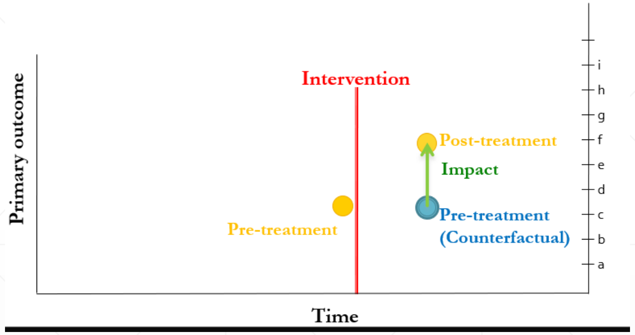
- Quiz 1 on March 4
- PS2 Grades released, PS3 Grades next week
- PS 4 due on March 13
- Today:
  - Quiz review: code demo + quiz questions discussion
  - For derivations / discussions related to mathematical intuition: office hours

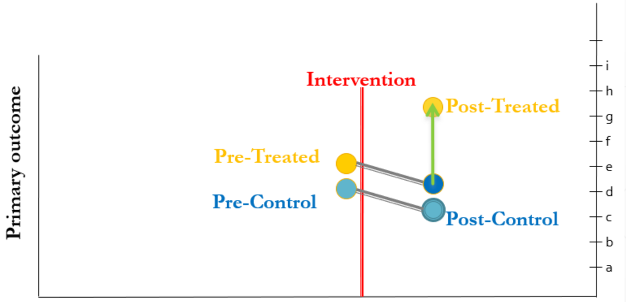
Quiz 1 Review

Quick Review: Research Designs

Design	Key Identifying Assumption	Confounds / Threats to identification
Randomized experiment (T v/s C)	?	?
Pre v/s Post	?	?
Double Difference	?	?







1. Diff-in-diff

- Suppose you are evaluating the impact of a minimum wage program on employment rates. In the treatment group, the employment rate changed from 74% (pre) to 82% (post). In the control group, during the same time, the employment rate changed from 71% (pre) to 68% (post).
- Estimate the true impact of the minimum wage program.

2. Linear Regression

We run a linear regression of the form

$$\text{GPA} = \alpha + \beta \text{ StudyingHours} + \gamma \text{ ChatGPT}$$

StudyingHours is continuous (time spent reviewing lecture notes); ChatGPT is binary (indicator for whether student uses ChatGPT to write assignments)

$\alpha = 0.5$

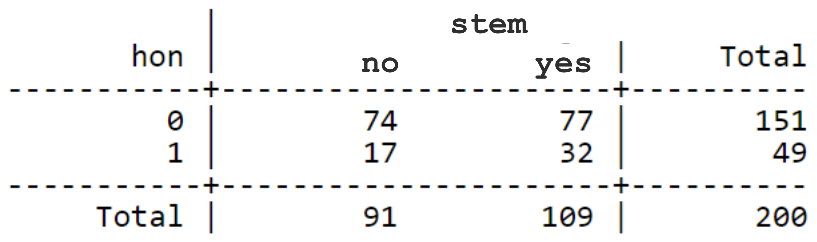
$\beta = 0.12$

$\gamma = -0.05$

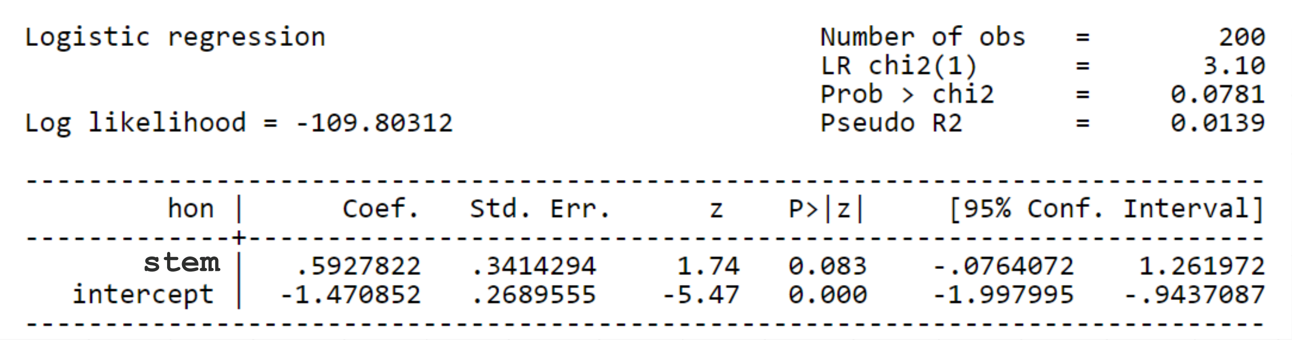
What is the difference between the GPA of a student who spends 20 hours studying + uses ChatGPT, and the GPA of a student who spends 40 hours studying + does not use ChatGPT?

3. Logistic Regression

- $\text{logit}(honor_i) = \alpha + \beta STEM_i + \epsilon_i$
- Calculate (from the regression results below):
  - odds of a non-STEM student pursing an honors degree?
  - odds of a STEM student pursuing an honors degree?
  - the odds ratio (STEM vs Non-STEM)
  - probability that a STEM student is an honors student?



```
           |          stem
       hon |        no        yes |     Total
-----------+----------------------+----------
         0 |        74         77 |       151
         1 |        17         32 |        49
-----------+----------------------+----------
     Total |        91        109 |       200
```



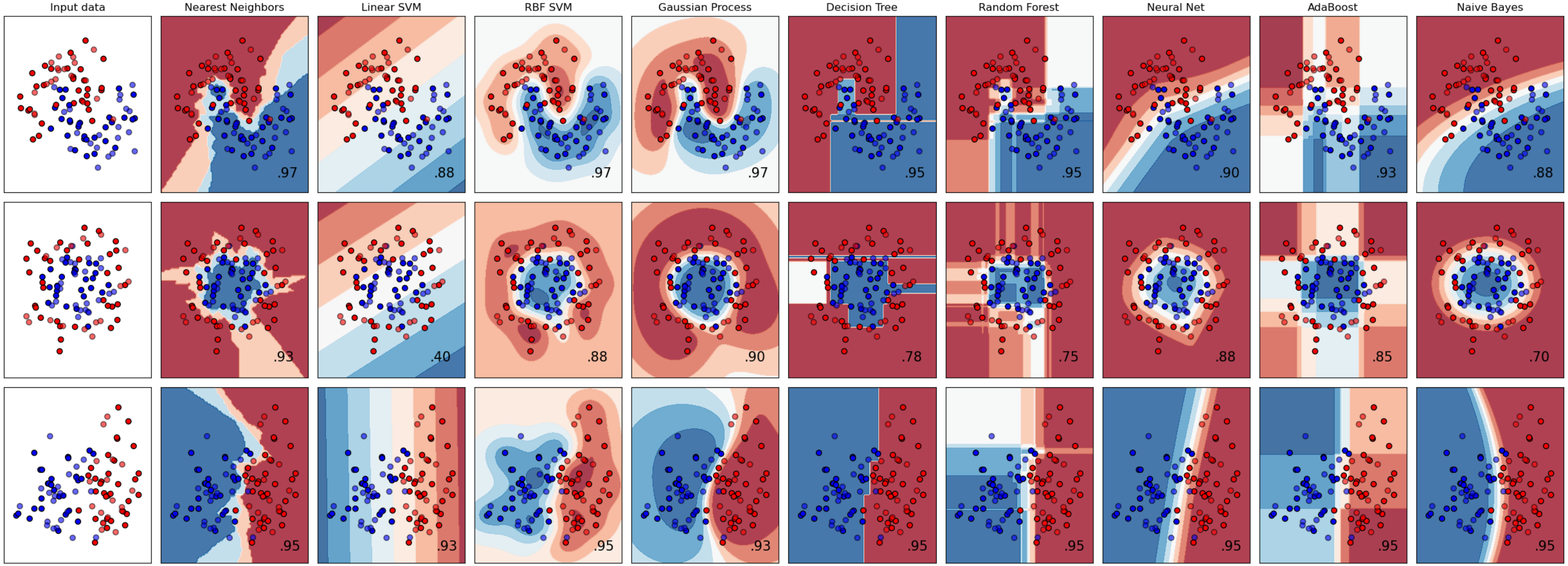
```
Logistic regression                             Number of obs     =        200
                                                LR chi2(1)        =       3.10
                                                Prob > chi2       =     0.0781
Log likelihood = -109.80312                     Pseudo R2         =     0.0139

------------------------------------------------------------------------------
         hon |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        stem |   .5927822   .3414294     1.74   0.083    -.0764072    1.261972
   intercept |  -1.470852   .2689555    -5.47   0.000    -1.997995   -.9437087
------------------------------------------------------------------------------
```

4. Ridge regression

- Statement A: As the regularization penalty becomes larger, ridge regression coefficients approach infinity
- Statement B: Ridge regression forces some coefficients to zero

1. A is True, B is True
2. A is True, B is False
3. A is False, B is True
4. A is False, B is False

Quick Review: Decision Boundaries



Source

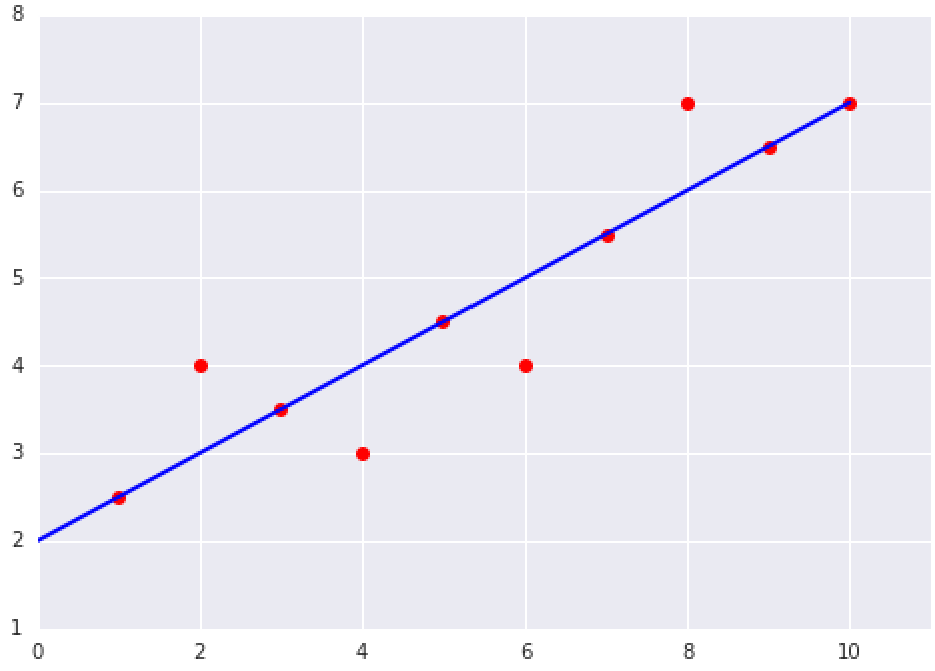
5. Decision Boundaries

- Which of the following algorithms recovers non-linear decision boundaries:
  - K-nearest neighbors (K = 5)
  - SVM
  - Logistic Regression
  - Logistic Regression with lasso regularization

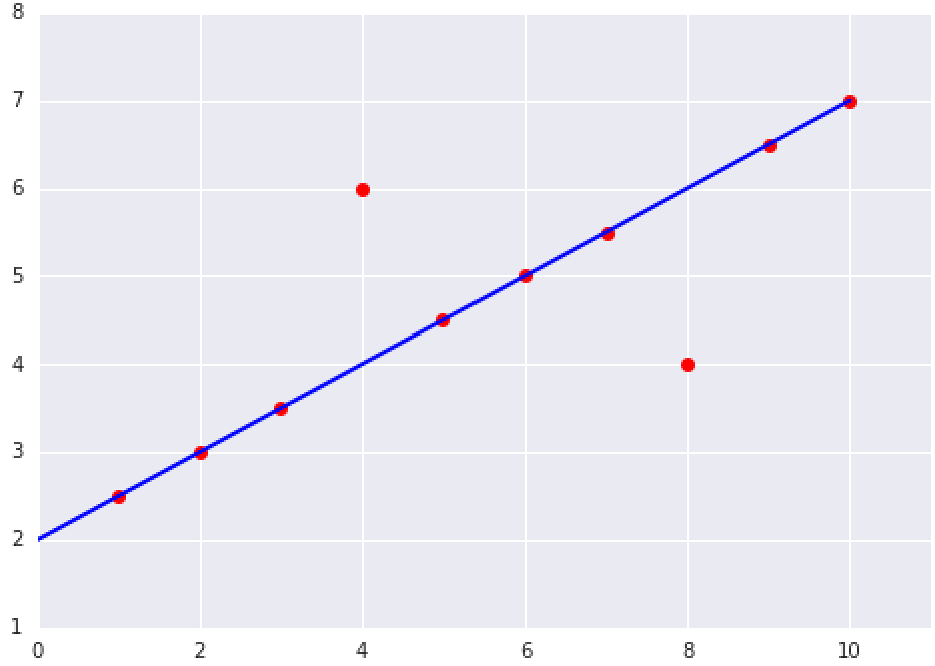
6. Mean Squared Error

- Suppose you build a linear regression model which predicts y = f(x). Which of these two cases has a higher MSE?
  - A
  - B

A



B



7. Bayes Theorem

- A doctor knows that having a cold causes you to sneeze 50% of the time.
- Prior probability of any patient having a cold is 1/10,000
- Prior probability of any patient sneezing is 1/15
- If a patient is sneezing, what is the probability they have a cold?

8. Cross-validation

- Suppose you want to estimate the out of sample performance of a K-nearest neighbors algorithm using nested cross-validation. If you have 5 outer loops, 10 inner loops and 20 different values for K in the hyperparameter grid, how many times will the learning algorithm nearest_neighbor(K) be called?
- *Hint: Don't forget the refit step!*

9. Classification

- Calculate accuracy, TPR, FPR and Precision for the “green” class.

		Predicted	
		Green	**Orange**
Actual	**Green**	9	3
	Orange	2	1