CS 59970 - Intro to Data Science
Midterm Exam
October 22, 2018

Answer Key

This is a 50 minute exam. Please use your time wisely.

Academic dishonesty is prohibited in The City University of New York. Penalties for academic dishonesty include academic sanctions, such as failing or otherwise reduced grades, and/or disciplinary sanctions, including suspension or expulsion.

Exam 3079

Part I: Multiple Choice. (2 points per question)

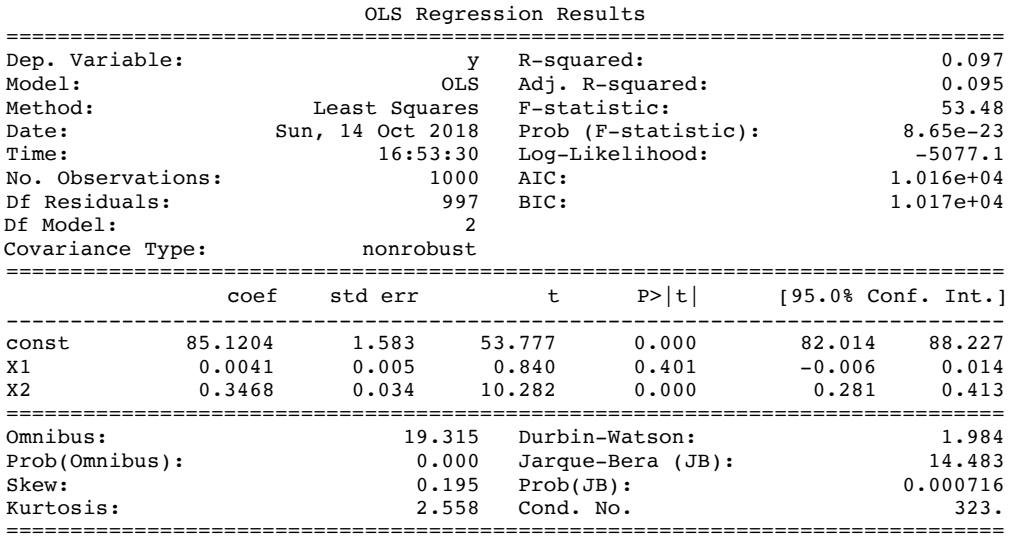
Circle the best answer(s).

1. Which of the following is ***NOT*** an assumption of linear regression? *Select all that apply.*
    a. The data are linear in form.
    b. The sample data extend beyond the bounds of [0,1].
    c. The dependent variable is correlated with all independent variables in the model.
    d. The errors are normally distributed.

2. Which of the following packages are helpful with the ETL process? *Select all that apply.*
    a. luigi
    b. pipemaster
    c. airflow
    d. bokeh

3. Which of the following packages can be helpful in web-scraping? *Select all that apply.*
    a. seaborn
    b. BeautifulSoup
    c. requests
    d. sklearn

Figure



OLS Regression Results

Dep. Variable:	y	R-squared:	0.097
Model:	OLS	Adj. R-squared:	0.095
Method:	Least Squares	F-statistic:	53.48
Date:	Sun, 14 Oct 2018	Prob (F-statistic):	8.65e-23
Time:	16:53:30	Log-Likelihood:	-5077.1
No. Observations:	1000	AIC:	1.016e+04
Df Residuals:	997	BIC:	1.017e+04
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf.	Int.]
const	85.1204	1.583	53.777	0.000	82.014	88.227
X1	0.0041	0.005	0.840	0.401	-0.006	0.014
X2	0.3468	0.034	10.282	0.000	0.281	0.413

Omnibus:	19.315	Durbin-Watson:	1.984
Prob(Omnibus):	0.000	Jarque-Bera (JB):	14.483
Skew:	0.195	Prob(JB):	0.000716
Kurtosis:	2.558	Cond. No.	323.

4. Based on the figure above, which of the following numbers below best reflects the overall goodness of fit of the model?
    a. 19.315
    b. 0.097
    c. 997
    d. 8.65e-23

5. Based on the figure above, what is the percentage likelihood that the coefficient for the independent variable X1 is equal to zero?
    a. 0.0041
    b. 0.005
    c. 0.401
    d. 0.840

6. Which of the following python packages can be used to run a linear regression? *Select all that apply.*
    a. **statsmodels**
    b. **sklearn (sci-kit learn)**
    c. **numpy**
    d. seaborn

7. Which of the following is ***NOT*** a package that is commonly used for data visualization?
    a. matplotlib
    b. **luigi**
    c. bokeh
    d. seaborn

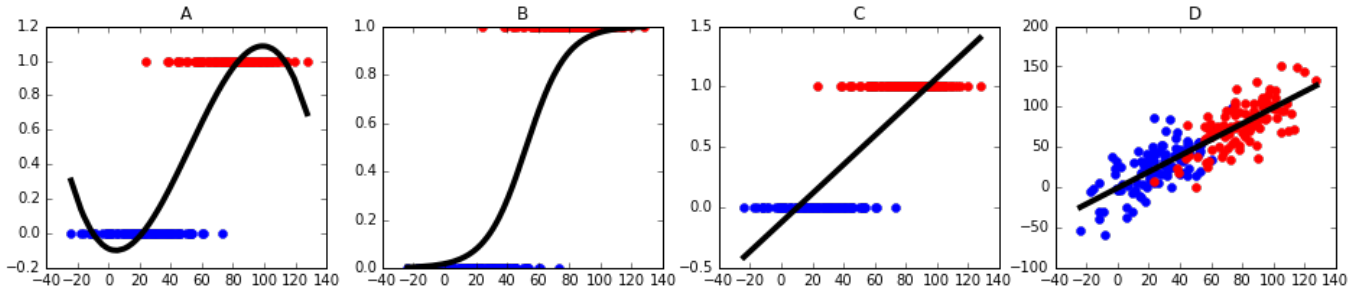
8. Which of the follow are stakeholders in the ETL process? *Select all that apply.*
    a. **software engineers**
    b. **data engineers**
    c. **data scientists**
    d. **product managers**

9. In statistics, a sequence or a vector of random variables is homoskedastic if:
    a. all variables in the sequence or vector have increasing variance.
    b. **all variables in the sequence or vector have the same finite variance.**
    c. one predictor variable can be linearly predicted from the others with a substantial degree of accuracy.
    d. all variables in the sequence or vector are bounded to the range [-1,1].

10. In statistics, a sequence or a vector of random variables exhibits multicollinearity if:
    a. all variables in the sequence or vector have increasing variance.
    b. all variables in the sequence or vector have the same finite variance.
    c. one predictor variable can be linearly predicted from the others with a substantial degree of accuracy.
    d. all variables in the sequence or vector are bounded to the range [-1,1].

11. What were data scientists at Target famously able to do?
    a. Identify and send advertisements to a consumer who was pregnant before she told her family she was expecting.
    b. Build an artificial intelligence system to defeat two reigning champions on the quiz show Jeopardy.
    c. Discover that customers shopping for greeting cards were five times more likely to also buy groceries than those shopping for clothing.
    d. Recommend products to online customers based upon what similar customers had purchased in stores.

12. Which of the follow figures below depicts the output of logistic regression?



    a. A
    b. B
    c. C
    d. D

13. Which of the following is the best example of a regression problem?
    a. Predicting which sports team will win an upcoming match.
    b. Estimating the likelihood that an Instagram user will like a given photo.
    c. Predicting how much a customer will spend at an online retailer in the next year based on their past purchase history.
    d. Predicting whether a customer will purchase an item at an online retailer based on their browsing history.

14. Which of the following is the best example of a classification problem?
    a. Predicting how much a customer will spend on groceries in the next three weeks based on their credit card transaction history.
    b. Predicting whether a customer will purchase an item at an online retailer based on their browsing history.
    c. Predicting the number of online views an article on the New York Times website.
    d. Estimating the total amount of time the average user will spend using an application on a mobile device.

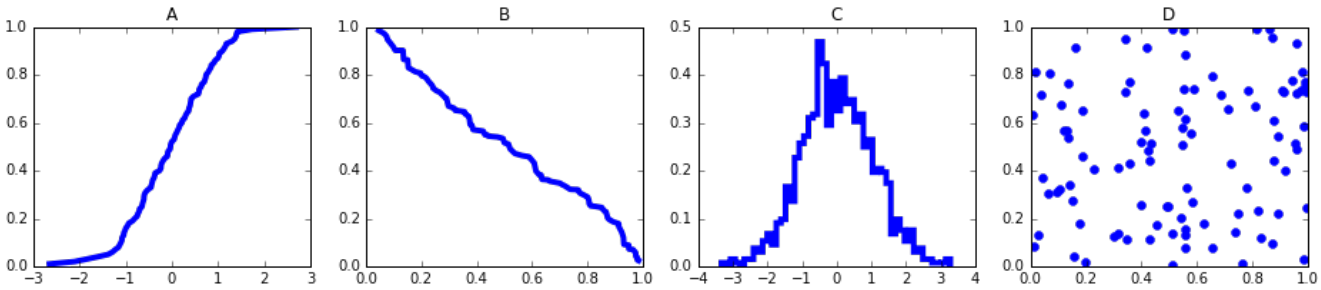
The *law of large numbers* tells us that the average of the results obtained from a large number of trials should be [ANSWER TO QUESTION 15], and will tend to [ANSWER TO QUESTION 16].

15.
    a. homoskedastic
    b. identical among most trials
    c. increasing with the number of trials
    d. close to the expected value

16.
    a. tending toward zero as more trials are performed
    b. be uniformly distributed as more trials are performed
    c. become closer to the expected value as more trials are performed
    d. be positively skewed for distributions with a mean greater than zero

17. According to the central limit theorem, when independent random variables are added:
    a. their properly normalized sum tends toward a uniform distribution
    b. their properly normalized sum tends toward a normal distribution
    c. their properly normalized sum tends toward zero
    d. their variance tends to increase

18. Which of the following charts could be an empirical cumulative distribution function?



    a. A
    b. B
    c. C
    d. D

19. Which of the following statements about logistic regression is ***FALSE***? *Select all that apply.*
    a. Logistic regression always outputs 1 or 0.
    b. Logistic regression is an example of a linear model.
    c. Stochastic gradient descent is commonly used to fit logistic regression.
    d. Logistic regression is commonly used in machine learning.

20. Suppose we're interested in determining whether there is a relationship exists between a company's stock performance and the number of tweets about that company in a given day. Which of the following statistics might be helpful? *Select all that apply*.
    a. **The coefficient for the number tweets regressed again the stock performance.**
    b. **The correlation between the two distributions.**
    c. **The covariance between the two distributions.**
    d. The median of each distribution.

Part II: Short Answer. (6 points per question)

Answer each of the following in a few short sentences.

1. Explain the difference between linear and logistic regression. (Explain at least two differences and one similarity in detail).

    - *Linear regression predicts a continuous target, returns an unbounded continuous output, and is useful in regression problems.*
    - *Logistic regression predicts a binary target, outputs a value between 0,1 and is a classification algorithm.*
    - *Both models are predictive models that are linear in their parameters.*

2. Explain the difference between structured and unstructured data, give an example of each, and a justification for why each example fits into that category.

    - *Structured data is well organized and has a clearly defined model. Stock prices are an example of structured data: they are meticulously maintained records upon which large amounts wealth is allocated.*
    - *Unstructured data has little underlying order and requires processing to be useful in analytical exercises. Call center transcripts are an example of unstructured data: while there may be some metadata surrounding each call, the transcripts themselves must processed to extract keywords and sentiment, among other attributes.*

3. In the data science lifecycle, does business understanding need to come before data understanding? Why or why not?

   Both business understanding and data understanding are at the beginning of the data science lifecycle. While business understanding typically precedes data understanding, the two go hand in hand given the availability of data will drive the business questions that can be answered.

4. Say we want to create a model of rents in New York City. What model would be helpful in predicting rents? Give examples of three features you think might be useful.

   In class we used linear regression to predict rents. Three features that might be helpful are size of the apartment, distance from the subway, and floor within the building.

5. Which of the following charts depicts the best fitting model? For each chart give a reason why it is or why is not the best fitting model.

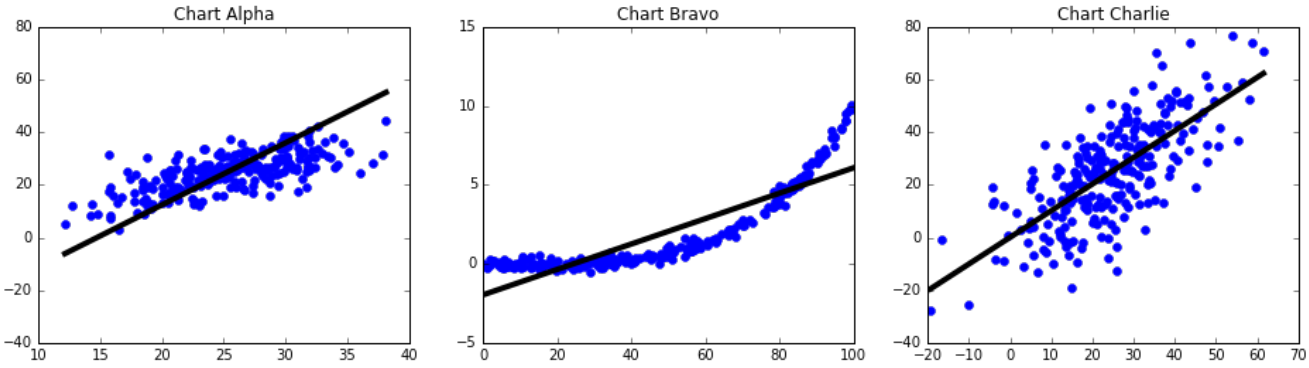


Chart Charlie demonstrates the best fit data: the errors are normally distributed around the regression line, which cuts evenly through the distribution of points. The regression line in Chart Alpha is not centered on the data. The data in Chart Bravo is not linear.

6. A mysterious disease causing uncontrollable, maniacal laughter has broken out among City College students, infecting 1,000 people. Public health authorities have identified a pill that will stop the maniacal laughter without any side effects, but the manufacturer can only make 1,000 pills. CCNY administration has asked for your help to identify which students are likely infected, and which are just thinking about the hilarious gifs their CS professor put in this semester's slides.

   You've trained a classifier, and the output is as follows:

		Predicted	
		Infected	Not Infected
Actual	Infected	900	100
	Not Infected	100	8,900

   Based on the scenario presented and the information below:

   1. What metric is best suited to evaluate the performance of your classifier (please give the proper name of the metric)?
   2. Why is this metric the best-suited for this scenario?
   3. What is the value of this metric?

Given that we want to find all of the cases of the disease to stop its spread, we'll want to focus on recall (true positives / (true positives + false negatives). 900 / (900 + 100) = 90%.