Name: _____ Student ID: _____

Cogs 109: Modeling and Data Analysis

Midterm exam

Tuesday 10/30/2018

- Calculators allowed, but not smartphones. Calculators should not be necessary.
- There will not be any coding exercises.
- 1 page (single sided) of handwritten notes are allowed.

1. **Prediction and inference.** Consider a data set containing the number of people who voted in San Diego county in the last election in each age group (18-30 year olds, 30-40, 40-50, 50-60, 60-70, 70+). You could model these data to address different types of questions.
    a. (1 point) In 1-3 sentences, describe one question that you could address using prediction. (Be sure to phrase your answer as a question!)

How many people aged 55 years will vote?

    b. (1 point) In 1-3 sentences, describe one question that you could address using inference.

Is there a relationship between age group and voting? Are people aged 30-40 more likely to vote?

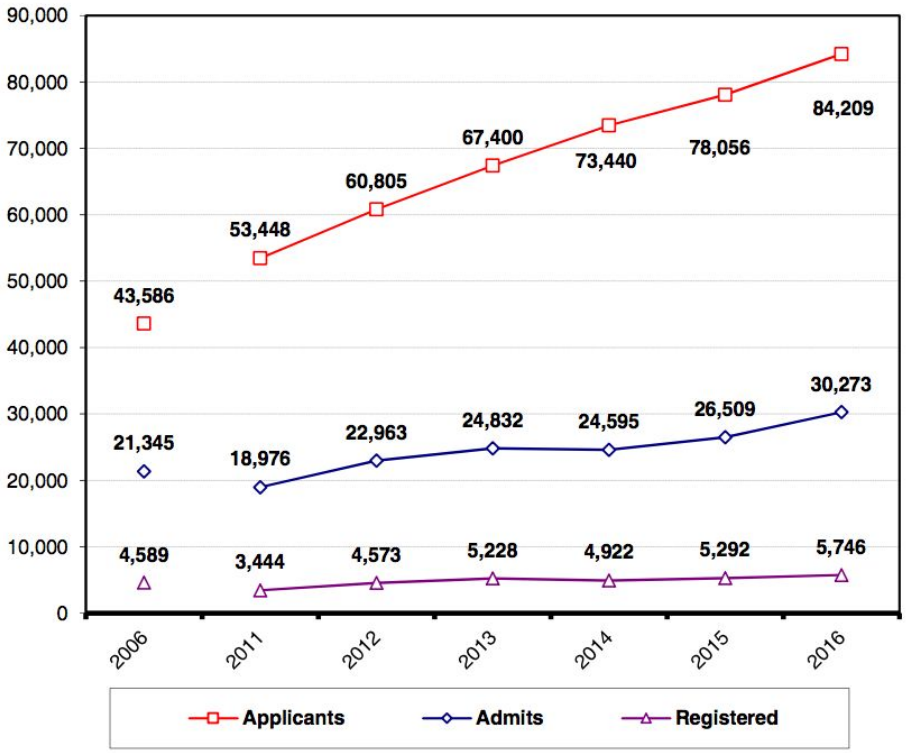
2. **Model flexibility.**
    a. (1 point) Name one <u>advantage</u> of a flexible model.

A flexible model can capture complicated relationships in data sets. Because of this, they have a lower bias. It might therefore make more accurate predictions and enable more reliable inferences -- if the data set is sufficiently large.

    b. (1 point) Name one <u>disadvantage</u> of a flexible model.

Flexible models are more at risk of overfitting. They generally have higher variance than simple models.

3. The plot below shows the number of applicants, admitted, and registered students at UCSD over time.



a. (1 point) A data scientist in the Admission Office would like to predict how many applicants will apply next year (2019). Would a regression or a classification model be most appropriate, and why?

A regression is appropriate because the outcome (number of applicants) is a quantitative variable, not a categorical variable.

b. (1 point) The data scientist decides to try two different models. In these models, y is the number of applicants and x is the year. Circle the more flexible model:

Model 1: $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \varepsilon$ Model 2: $y = \beta_0 + \beta_1 x + \varepsilon$

Model 1 is more flexible since it has more parameters (4) compared to model 2 (2 parameters).

A summary of the model fit for a linear regression (model 2) is shown below. (Recall that the notation `5e-02` means $5 \times 10^{-2} = 0.05$).

```
Linear regression model:
    Applicants ~ 1 + Year
Estimated Coefficients:
                   Estimate          SE          tStat       pValue
```

(Intercept)	-1.2103e+07	4.2491e+05	-28.485	9.0393e-06
Year	6000	200	28.649	8.8353e-06

c. (2 points) Is there a statistically significant relationship between Applicants and Year? Justify your answer by referring to at least one of the entries in the table above.

Yes - the p-value for the slope is $8.8e-6 = 8.8 \times 10^{-6} < 0.05$ **, and the t-statistic is 28.6 > 2.**

d. (1 point) What is the null hypothesis that corresponds to the "Year" term in the model (with p-value p=8.8353e-06)?
There is no relationship between year and the number of applicants.

e. (2 point) According to this model, what is the expected increase in the number of applicants to UCSD per year?
6000 applicants

f. (1 point) What is the 95% confidence interval for the expected increase in the number of applicants per year?

The 95% confidence interval is 6000 ± 2*200 = 6000±400 = [5600 - 6400]

f. (2 points) After fitting the two models to the data for 2011 through 2016, the data scientist determines that the mean squared error for the training data is $MSE_1 = 2.7 \times 10^5$ for Model 1, while it is higher for Model 2: $MSE_2 = 8.1 \times 10^6$. Does this warrant choosing Model 1 over Model 2? Why or why not?

No - Model 1 is more flexible, and therefore is expected to have lower training error. It may or may not have lower test error.

g. (2 points) In a short paragraph (3-5 sentences), describe how you could use leave-one-out cross-validation (LOOCV) to compare the performance of the two models on test data.

There are 6 data points. First, choose one sample (e.g. 2011) to set aside for testing, and use the remaining 5 samples to find the best fit parameters for model 1 and model 2. Then measure the squared residual error for the test data point for each model. Repeat this procedure 6 times, setting aside a different data point for testing each time. Finally, average the squared residual error to estimate the test-set mean squared error (MSE_{test}) for each model. Choose the model with the lower test error.

4. Two researchers (Alice and Bob) are studying the effect of age (x, measured in years) on the speed of a motor reflex (y, measured in cm/s) in older adults. Each researcher selects N=20 subjects at random from a nursing home with a large population of residents.

a. (2 points) The researchers propose to model the data using a linear regression: $y = \beta_0 + \beta_1 x_1 + \varepsilon$. In this formula, what does ε represent? What does β_1 represent?

ε represents irreducible noise.

β_1 represents the slope, i.e. the change in expected value of y for each additional year of age.

b. (2 points) Alice determines that the best fitting model for her sample has $\hat{\beta}_0 = 180\ cm/s$, $\hat{\beta}_1 = -1\ cm/s/year$. What motor reflex speed should Alice predict for a 60 year old subject?

$\hat{y} = 180 - 1 * 60 = 120\ cm/s$.

c. (1 point) If Alice recruits a new subject who is 60 years old and has a reflex speed of 110 cm/s, what is the squared residual error for that subject? (Please include units)

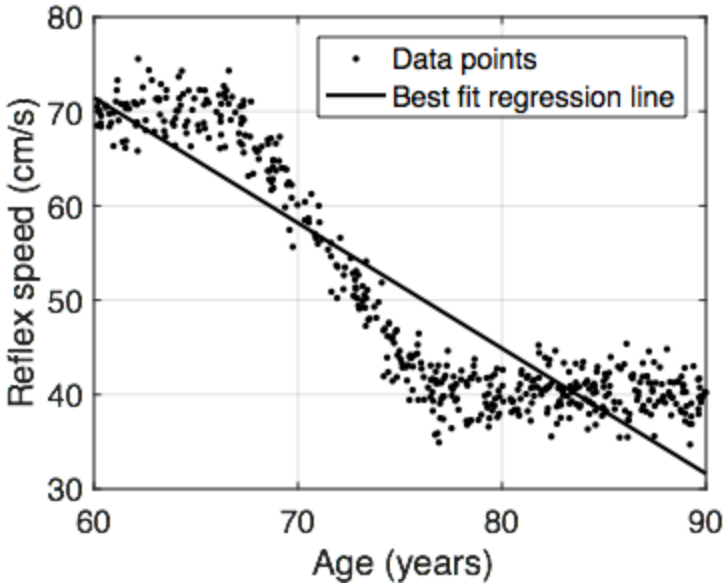
$Err = (y - \hat{y})^2 = (110 - 120)^2 = (-10)^2 = 100\ (cm/s)^2$.

d. (1 point) Bob finds that the best fit model for his sample has $\hat{\beta}_0 = 150\ cm/s$, $\hat{\beta}_1 = -1.5\ cm/s/year$. These values are similar, but not identical, to Alice's parameter estimates. Does the difference between the two regression lines correspond to model variance, or model bias?

Variance

e. (1 point) Alice and Bob team up to conduct a larger study, including N=500 subjects. The data points and best fit regression line are shown below. They notice that the model predictions are systematically too low for 60-70 year old subjects, and too high for 70-80 year old subjects. Does this problem correspond to model variance or model bias?

Bias



f. (2 points) In addition to age, reflex speed could be influenced by a subject's physical activities. Assume each subject is in one of 3 categories: inactive, plays shuffleboard, or swims. Which of the following would be a valid model for the data? (Circle all acceptable models)

i. $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$, where

$x_1 =$ Age and

$x_2 = 0$ (inactive), 1 (shuffleboard) or 2 (swims)

ii. $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$, where

$x_1 =$ Age and

$x_2 = 0$ (inactive), 1 (shuffleboard) or 0 (swims) and

$x_3 = 0$ (inactive), 0 (shuffleboard) or 1 (swims)

iii. $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$, where

$x_1 =$ Age and

$x_2 = 1$ (inactive), 0 (shuffleboard) or 0 (swims) and

$x_3 = 0$ (inactive), 1 (shuffleboard) or 0 (swims)

Both ii and iii are valid, but i is not. This is because model 1 treats golf and basketball as quantitative, not categorical, variables.

5. Consider a logistic regression model that predicts whether a customer will buy a new car (y=yes or no) based on the customer's employment status (Unemployed, Part-time, or Full-time). The model is: $Prob(y = yes\ |x) = f(\beta_0 + \beta_1 x_1 + \beta_2 x_2)$ where $f(z) = e^z/(1 + e^z)$ is the logistic function. We have defined dummy variables:

$x_1 = 1$ if the customer is employed part-time, 0 otherwise

$x_2 = 1$ if the customer is employed full-time, 0 otherwise

a. (1 point) Give one reason why a logistic regression is more appropriate for this data set than a linear regression.

Since the outcome (yes or no) is a binary variable, logistic regression is appropriate as it models the probability of a binary outcome. Linear regression would give positive and negative values that don't have a clear interpretation in terms of the binary outcomes.

b. (2 points) What is the predicted probability of buying a new car for an unemployed customer? (Your answer should be written in terms of the parameters, $\beta_0, \beta_1 \beta_2$).

For an unemployed customer, $x_1 = 0,\ x_2 = 0$ **, so** $Prob(y = 1|x) = f(\beta_0 + \beta_1 x_1 + \beta_2 x_2) = f(\beta_0) = e^{\beta_0}/(1 + e^{\beta_0})$

b. (2 points) Write the formula for the odds-ratio in terms of the posterior probability, $Prob(y = yes \mid x)$.

The odds ratio is $Prob(y = yes|x)/(1 - Prob(y = yes|x))$

c. (2 points) Refer to the regression summary table, below.

Generalized linear regression model:

logit(y) ~ 1 + x1 + x2

Distribution = Binomial

Estimated Coefficients:

	Estimate	SE	tStat	pValue
(Intercept)	-0.25131	0.35635	-0.70525	0.48066
x1	0.76214	0.51021	1.4938	0.13524
x2	-1.8281	0.63893	-2.8612	0.0042201

Based on this table, is there a statistically significant difference between unemployed customers (x1=0, x2=0) and part-time employed customers (x1=1, x2=0)? Justify your answer.

No - the p-value is 0.135 > 0.05, and the t-statistic is 1.494 < 2.

6. Extra credit (1 point): Do you believe the sun will rise tomorrow? Why or why not? Is your belief based on a statistical model, or something else? (Answer in 3-5 sentences).