Stat 200 Exam 3 April 16, 2018

**PRINT NAME**_____ _____ net ID_____

(Last name) (First name) (University email)

Key

**Circle Section: L1 or Online**

**READ so you don't lose points:**

- **Filling out your name and net ID clearly and circling your section is worth 1 pt.**
- **Round all answers to 2 decimal places unless otherwise indicated.**
- **Show work when requested. No work, no credit.**

Write answers in appropriate blanks. When no blanks are provided **CIRCLE** your answers.

No notes or books are allowed. No phone calculators are allowed.
Do not use your own scrap paper. If you need some, ask a proctor.

**Make sure you have all 5 pages (8 problems).**

**DO NOT WRITE BELOW THIS LINE**

The numbers written in each blank below indicate how many points you missed on each page. The numbers printed to the right of each blank indicate how many points each page is worth.

Page 1 _____ 18

Page 2 _____ 28

Page 3 _____ 16

Page 4 _____ 20

Page 5 _____ 18

**Cover Page** _____**1**

**Score** _____

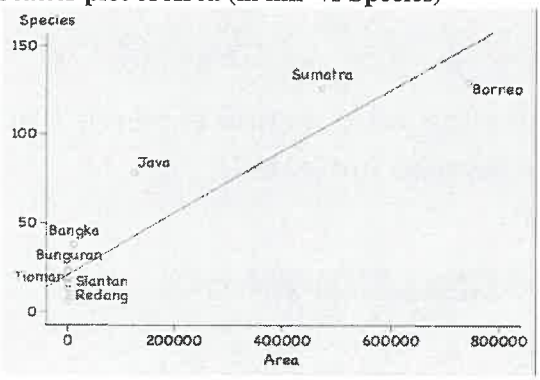
**NO CLASS on Tuesday!**

**Scores will be posted on Compass Wednesday night and exams will be returned in class on Thursday.**

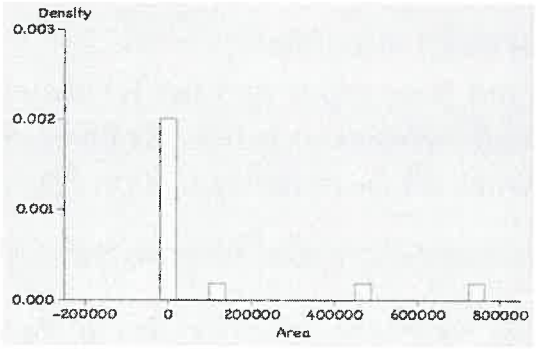
Stat 200 Exam 3 April 16, 2018

**Question 1** *(18 pts.)* pertains to the **Area** (in km$^2$) and the **number of mammal species** for 13 islands in Southeast Asia. How does the size of the island predict the number of species on the island?

**Scatter plot of Area (in km$^2$ vs Species)**



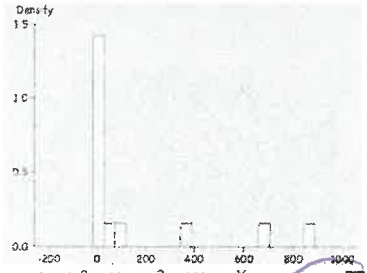

**Histogram of Area**




**a)** *(2 pts.)* Notice how most of the islands are all squished together in the corner. Also look how skewed the Area histogram is. I want to transform the X variable (Area) to make the histogram more normal. Which transformations should I try? Circle ALL that might work. i) $X^2$ ii) $X^3$ iii) $e^X$ iv) $\sqrt{X}$ v) ln(X)

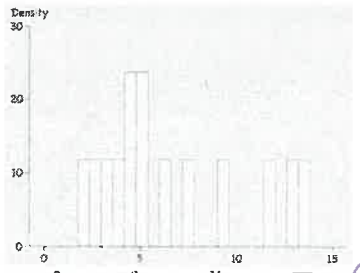
1pt 1pt

*b) (4 pts.)* You tried one of the transformations and it was a step in the right direction but it didn't go far enough. You tried another and it worked much better well. Below each histogram circle the transformation it depicts.




i) $X^2$ ii) $X^3$ iii) $e^X$ iv) $\sqrt{X}$ v) ln(X)

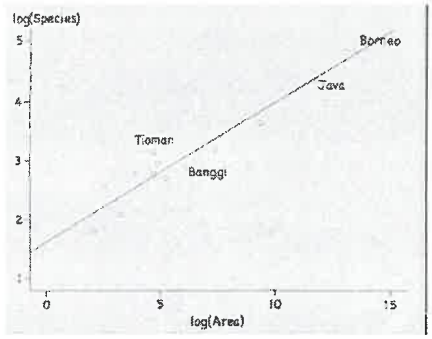
2 pt




i) $X^2$ ii) $X^3$ iii) $e^X$ iv) $\sqrt{X}$ v) ln(X)

2 pts

**c)** (12 pts.) Below is the scatter plot of ln(Species) vs ln(Area) where Species= the number of mammal species on each island and Area= area of each island in km$^2$ The regression equation is: **Predicted ln(Species)= 1.6 + 0.23 ln(Area)** **$SD_{errors}$= 0.2**




**i)** *(6 pts.)* Bangii has an area= 450 km$^2$

Use the regression equation to predict the **ln(Species)** and **Species** number for Bangii

**a)** (2 pts.) ln(Species) = 3.01 **b)** (2 pts.) Number of species= $e^{3.01}$ = 20.19

$-1.6 + 0.23 \ln(450)$

**c)** (2 pts.) 95% Confidence Interval for part(b) above= (13.6, 30.27)

(Use Z=2 for 95% CI) $e^{2.61}$ ; $e^{3.41}$

c.e from b

**ii)** (2 pts.) Another island has a 95% confidence interval = **(11.23, 25)** for the predicted number of species. What is the predicted number of species? 16.76 *Show work.*

$\sqrt{11.23 \times 25}$ or $e^{\left[\frac{\ln(11.23) + \ln(25)}{2}\right]}$

**iii)** (2 pts.) Change the regression equation **ln(Species) = 1.6 + 0.23 ln(Area)** to an equation in terms of **species** and **Area, not ln(Area)**.

Species = $e^{1.6}\ Area^{0.23}$ $e^{1.6} \cdot e^{\ln Area \times 0.23} = e^{1.6} \cdot Area^{0.23}$

**iv)** (2 pts.) One island has twice the area of another island. The regression estimate for the number of species on the smaller island is 9. What is the regression estimate for the number of species on the larger island? 10.56 ***Show work.***

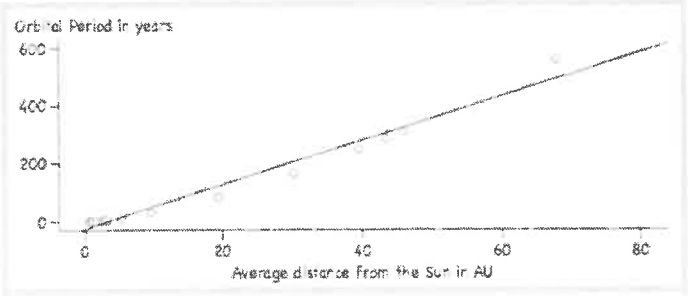
$2^{0.23} \times 9 = 1.173 \times 9$

No PC.

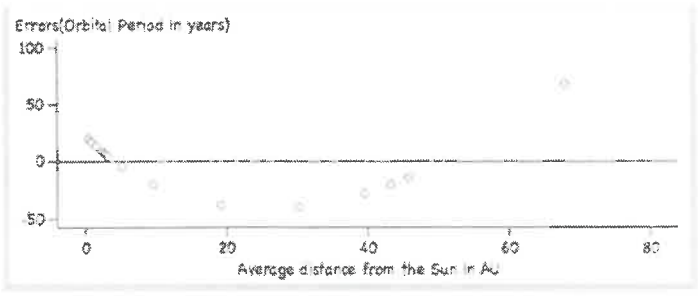
Stat 200 Exam 3 April 16, 2018

**Question 2** *(6 pts.)* The scatter plot below shows the average distance from the Sun in AU (astronomical units) on the X axis and the Orbital period in years (length of time to orbit sun) on the Y axis of 16 solar systems objects. (Imagine these 16 objects were randomly chosen from a large collection of objects orbiting the sun.)

**Scatter Plot**




**Residual Plot**




Here's the regression equation: **Predicted Orbital Period = -23.12 + 7.57(Distance from Sun) r= 0.9864 and $SD_{errors}$ = 26.04**

**a)** (3 pts.) Why do the 16 points closely follow a line in the scatter plot but follow a curve in the residual plot?

- **i.** Residual plots always transform linear plots into curves that either point up or down depending on the whether the correlation is positive or negative.
- **ii.** It's because the correlation is so high, the higher the correlation the stronger the curvature.
- **iii.** It's because the scale on the Y axis for the residual plot has been changed, making it easier to see the curvature.

**b)** (3 pts.) Is it appropriate to use the regression equation above to describe the relation between distance from the sun and orbital period for all the objects ?

- **i.** Yes, because the scatter plot follows a line very closely.
- **ii.** No, because the residual plot shows a clear pattern violating the assumptions needed to use a linear model.
- **iii.** Yes, because the 16 objects were randomly selected so there is no need to check whether assumptions were violated.

**Question 3** *(10 pts.)*

For each of the following is it appropriate to use logistic regression? **Circle Yes or No.**

- **a)** Predicting income based on years of college. YES NO
- **b)** Predicting ln(income) based on years of college YES NO
- **c)** Predicting graduating college based on family income. YES NO
- **d)** Predicting getting a scholarship based on gender and ethnicity. YES NO
- **e)** Predicting favorite color based on gender YES NO

**Question 4** (6 pts.) Circle True or False for each statement below.

- **a)** The logistic regression model only handles X values that can be coded as 1's and 0's. i)True ii)False
- **b)** Transforming non-linear scatter plots into linear ones by converting Y to ln(Y) is called logistic regression. i)True ii)False
- **c)** The assumptions needed to make inferences for linear and logistic regression are the same i)True ii)False

**Question 5** (4 pts.)

How are the parameters chosen in logistic regression and linear regression?

*Fill in the first blank below with "logistic" or "linear" and the second blank with "minimize" or "maximize".*

a) In Linear regression, the parameters are chosen to minimize the sum of the squared errors

b) In Logistic regression, the parameters are chosen to maximize the likelihood of getting our sample data.

**Question 6** (2 pts.) Are F and t tests ever appropriate to test significance in Logistic regression models?

***Choose one:***

- **a)** Yes, when the sample size is small the F and t tests give more accurate results.
- **b)** No, because F and t tests can never be done on variables that have undergone log transformations.
- **c)** No, because F and t tests are never done when we are predicting counts (when Y is binary), since the SD can be estimated directly from the count.

-½ for careless errors

Stat 200 Exam 3 April 16, 2018

**Question 7 Part I** *(16 pts.)*
On our survey, **178** students anonymously answered these 2 questions:
"Would you volunteer to be randomly assigned to either the online or in person section?"(No = 0, Yes =1)
"Which section are you in?" (L1=0, online=1)

To predict the probability of volunteering from section, we fit a logistic regression model. Here's the ln(odds) form of the

regression equation: $\ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = -0.5261 + -0.7267(\text{Section})$

**a)** *(2 pts.)* Are online students more or less likely to volunteer?

***Choose one:*** **i)** More **ii)** Less **iii)** Same **iv)** Not enough info

**b)** *(2 pts.)* What is the probability that an L1 student would volunteer? p= 0.37 *Show work.*

Check with chart below

$\ln(\text{odds}) = -0.5261$

$\text{odds} = e^{-0.5261} = 0.59$

$p = .59/1.59 = 0.37$

$\frac{26}{70} = 0.37$ ✓

**c)** *(2 pts.)* What is the probability that an online student would volunteer? p= 0.22 *Show work.*

$\ln(\text{odds}) = -0.5261 - 0.7267 = -1.2528$

$\text{odds} = e^{-1.2528} = 0.29$

$p = .29/1.29 = 0.22$

$\frac{24}{108} = 0.22$ ✓

**d)** (2 pts.) The Odds Ratio = _____ *Show work.*

$e^{-0.7267} = 0.48$

**e)** (2 pts.) If we switched the coding for section to online = 0 and L1 =1 what would change?
***Choose one:*** **i)** Odds **ii)** Probabilities **iii)** Odds Ratio **iv)** All **v)** None

**f)** (6 pts.) Look at the table below showing the 178 responses to the 2 questions.

| | No | Yes | Total |
|---|---|---|---|
| L1 | 44 | 26 | 70 |
| Online | 84 | 24 | 108 |
| Total | 128 | 50 | 178 |

OR is the change in odds as you go from X=0 to X=1
so if we switch coding it will change ONL odds to L1 odds / L1 odds to ONL odds. Probabilities & odds don't change bc we're not changing the volunteer coding (Y)

Use the table to compute the odds for an L1 and online student volunteering. ***Please leave your answers in fraction form.***

i) (2pts.) Odds for L1 = $\frac{26}{44} = \frac{13}{22}$ = # success / # failures = # Yes / # No for L1

ii) (2pts.) Odds for Online = $\frac{24}{84} = \frac{2}{7}$ " " for ONL

iii) (2pts.) Should you get the same OR as in (d) above? (Assuming you compute the ratio of Online odds to L1 odds.)

**a)** Yes, within rounding error **b)** No

$OR = \frac{24/84}{26/44} = e^{-0.7267}$

Stat 200 Exam 3 April 16, 2018

**Question 7 Part II** (10 pts.)

A third question on the same survey was: "How many people have you been in a serious relationship with?" Adding relationships to the model gives us: $\ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = -1.33 + -1.03(\text{Section}) + 0.64(\text{Relationships})$

a) (2 pts.) The $\chi^2$ test for the overall regression effect: $H_0$: All $\beta$'s $=0$ yielded a $\chi^2$ **stat = 26.**
How many **degrees of freedom?** = 2 (p − 1 → 3 − 1)

b) (2 pts.) The p value < 0.1%. This means that the probability that…***Choose only one:***
i) the null is true < 0.1% ii) the null is false > 99.9% iii) we'd get a $\chi^2$ stat ≥ 26 if the null was true < 0.1%
We assume null is true, P-value is probability of getting test stat

c) (2pts.)The relationship slope has a SE = 0.14. To test $H_0$: $\beta_{relationship} = 0$ against $H_A$: $\beta_{relationship} \neq 0$ compute the Z stat. ***Show work.*** $\frac{0.64}{0.14} =$ Z = 4.57 (2 pts; one more extra)

d) (3 pts.) Since p ≤ 5%, a 95% Confidence interval for the Relationship slope does not include 0. (1 pt)
*Fill in the first blank with > or < , the second with "does" or "does not", and the third blank with a number.*

e) (2 pts.) The OR for Relationship = 1.90 ($e^{0.64}$ =) and the OR for Section= 0.36 ($e^{-1.03}$) does not, 1 (−1)

f) (2 pts) Comparing two people in the same section, the person with 2 more relationships has 3.6 (accept 3.61) times the odds of volunteering. ***Fill in the blank with a number. Show work.***
$e^{0.64 \times 2} = 3.6$ or $1.9^2 = 3.61$

g) (2 pts.) Comparing an L1 student with 4 relationships to an online student with 2 relationships, the L1 student has 10.07 times the odds of volunteering. ***Fill in the blank with a number. Show work.***
$e^{0.64} \cdot e^{0.64} \cdot e^{1.03} = 10.07$ or $\frac{3.6}{0.36} = 10$

h) (2 pts.) What's the probability that an L1 student with 10 relationships will volunteer? 0.99. ***Show work.***
ln odds $(-1.33 + 6.4) = 5.07 \Rightarrow e^{5.07} = 159.17$
$p = \frac{159.17}{160.17} = 0.99$

i) (2 pts.) Would the ln(odds) equation at the top of the page change if we reversed the coding for Section so that L1=1 and online=0 and kept everything else the same? If so, write the new equation in the blank provided.

a) No, it would not change. b) Yes, it would change to $\ln\left(\frac{\hat{p}}{1-\hat{p}}\right) =$ −2.36 + 1.03(S) + 0.64(R (1 pt; +1 extra credit)

For ONL=1, equation at top gives ln(odds) = −1.33 + −1.03(1) + 0.64(R)
= −2.36 + 0.64(R)
Now ONL=0 so to get same ln(odds)
new intercept must be −2.36 ln(odds) = −2.36 + 1.03(S) + 0.64(
Sign for Section changes bc L1 is more likely to volunteer.

**Question 8** *(18 pts.)*

A predictor of whether esophageal cancer has not metastasized to the lymph nodes is the diameter of the tumor. Below is the log odds regression equation predicting the probability of no metastasis from the diameter of the tumor (measured in cm) from a hypothetical study of 200 patients.

**ln (p/(1-p)) = 2 - 0.5 (Diameter)**

**a)** *(4 pts.)* Use the equation to estimate the **odds** and **probability** of no metastasis for a tumor of diameter = 8 cm. ***Show work.***

*i)* Odds= 0.14 *ii)* Probability = 0.12

$\ln(\text{odds}) = 2 - 4 = -2$

$\text{Odds} = e^{-2}$

$\frac{e^{-2}}{1+e^{-2}} = 0.12$

**b)** *(2 pts.)* How do the estimated *odds* of no metastasis change if the tumor increases in diameter by 1 cm ?

**i)** odds are multiplied by 0.61 **ii)** the odds decrease by 0.5 **iii)** not enough info

$\times e^{-0.5} = 0.61$

**c)** *(2 pts.)* How does the estimated *probability* of no metastasis change if the tumor increases in diameter by 1 cm?

**i)** the probability is multiplied by 0.61 **ii)** the probability decreases by 0.5 **iii)** not enough info

**d)** *(3 pts.)* How big a tumor would give a 50% probability of metastasis? 4 cm

***Show work.*** $0 = 2 - 0.5D$

$D = 2/.5 = 4$

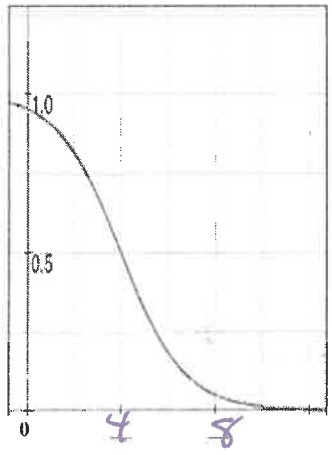
**e)** *(3 pts.)* How big a tumor would give a 40% probability of no metastasis? 4.81 cm

***Show work.*** $p = .4 \Rightarrow \text{odds} = \frac{.4}{.6} \Rightarrow \ln\left(\frac{.4}{.6}\right) = -.405$

$-.405 = 2 - 0.5D$ $\quad D = \frac{2.405}{.5} = 4.81$

f) ( 4 pts) Below is a graph of the probability form of the model.

Write its equation: $p = \frac{e^{2-0.5(D)}}{1+e^{2-0.5(D)}}$ and fill in the 2 blanks on the X-axis with the correct diameter values (in cm).




**Fill in the 2 blanks above with the correct numbers.** 1pt 1pt

c.e. from d

↑ 2pts for equation

$\ln(\text{odds}) = 2 - 0.5(D)$

$\text{odds} = e^{2-0.5(D)}$

$p = \frac{e^{2-0.5(D)}}{1+e^{2-0.5(D)}}$