Stat 200

STATISTICS 200 EXAM 3

Dec 4, 2017

PRINT NAME______________________________ ____________________ KEY

(Last name) (First name) Net ID (email, not UIN)

Circle Section: **L1** **Online**

Write answers in appropriate blanks. When no blanks are provided **CIRCLE** your answers. **SHOW WORK** whe requested.

No notes or books are allowed. Calculators (except for ones connected to the internet) are allowed.
Do not use your own scrap paper. If you need some, ask me.

Rounding Instructions: Please round all answers to 2 decimal places unless otherwis stated.

Make sure you have all 4 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. Th numbers printed to the right of each blank indicate how many points each page is worth.

Page 1_________10

Page 2_________35

Page 3_________30

| ~~10~~

Page 4_________25

| ~~13~~

Score _________

Scores will be posted on Compass Wed night and exams will be returned in class on Thursday. Online student can pick up their exams during office hours between 4-6 pm in 23 IH.

Stat 200 Dec 4, 2017

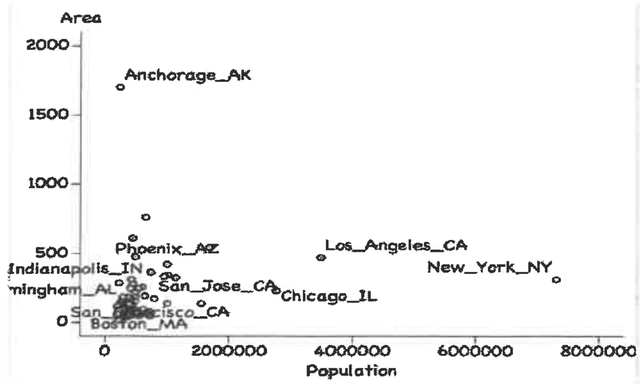
Question 1 *(2 pts)*

Suppose you'd like to do linear regression but your scatter plot is not close to linear. You see that the histogram of the Y variable is right skewed and you'd like to transform it to be more normal. Which transformation(s) would be possible candidates?

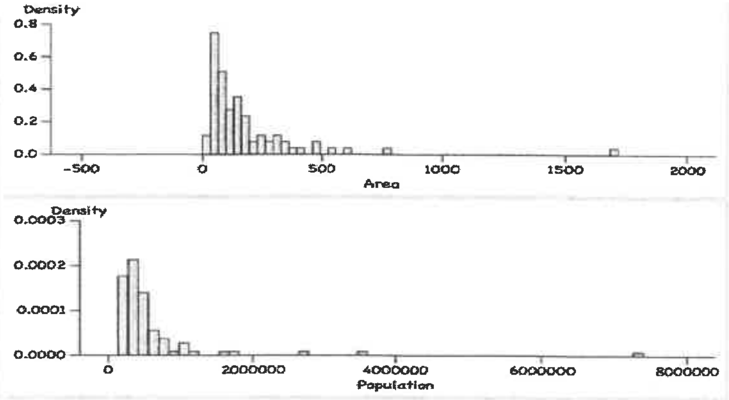
Circle all that could be. i)Y^2 ii) Y^3 iii) e^Y [✓ iv) $\sqrt{Y}$] [✓ v) ln(Y)]

Question 2 *(8 pts.)* pertains to the Area (in square miles) and the population of 77 US cities.

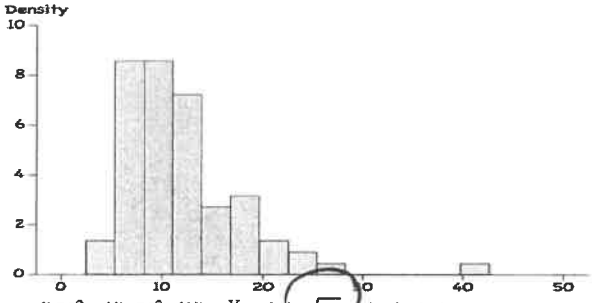
Scatter plot of Area vs Population



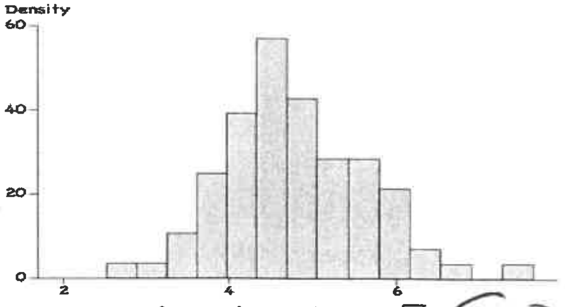
Histograms of Area (top) and Population (bottom)



a) *(2 pts.)* Below are histograms of the transformed Area. Which of the 5 transformations do the histograms depict? Below each histogram circle the transformation the histogram represents.

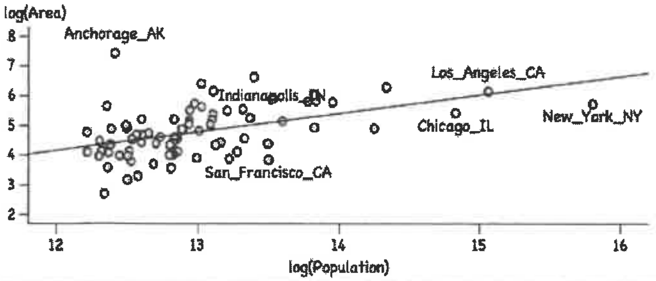


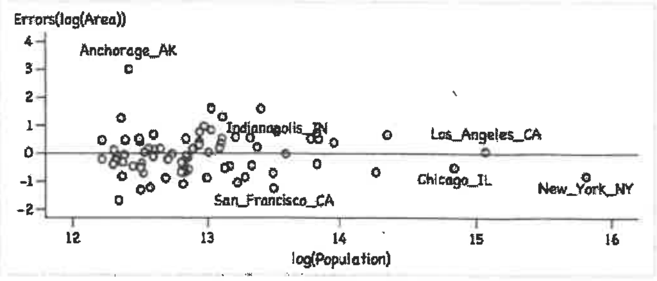
i)Y^2 ii) Y^3 iii) e^Y [✓ iv) $\sqrt{Y}$] v) ln(Y)



i)Y^2 ii) Y^3 iii) e^Y iv) $\sqrt{Y}$ [✓ v) ln(Y)]

Below is the scatter plot of ln(Area) vs ln (Population) on the left and the residual plot on the right.





b)*(2 pts.)* **The regression Equation is: ln(Area) = -3.3 + 0.62 *ln(Population) and SD$_{errors}$ = 0.75**

Use the regression equation above to predict the **ln(Area)** and **Area** of a city with a population of 3,000,000. 383.75 unrounded

Round final answers to 2 decimal places. *You may use your rounded answer for ln(Area) to compute Area.*

$-3.3 + 0.62\ln(3,000,000)$

ln(Area) = 5.95 Area = $e^{5.95}$ sq miles

382.51

c) *(2 pts.)* Build a **95%** Confidence Interval for your estimate of **Area** in part (b). Your answer should be a confidence interval for Area (NOT ln(Area)). ***Show work. Circle answer.*** Round to 2 decimal places. (Use Z = 2 as the critical value for 95%)

$e^{5.95 - 2(0.75)}$ to $e^{5.95 + 2(0.75)} \rightarrow e^{4.45}$ $e^{7.45}$

(85.63 sq miles to 1719.86 sq miles)

d) *(2 pts.)* A certain % Confidence Interval for the area of another city was computed to be (90 sq miles to 403 sq miles), but we don't kn the % CI. If possible calculate the estimated area of the city and show work. If not, write not enough info.

$\sqrt{90 \times 403} = 190.45$ 190.45 sq miles

OR $e^{(\ln(90) + \ln(403))/2}$

Stat 200 Dec 4, 2017

Question 3 *(10 pts.)*

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

i) Predicting eye color from hair color. YES [✓ NO]
ii) Predicting year in school from age. YES [✓ NO]
iii) Predicting passing the final from class attendance. [✓ YES] NO
iv) Predicting passing the final from gender. [✓ YES] NO
v) Predicting ln(Childrens Income) from ln(Parents Income). YES [✓ NO]

Question 4 *(15 pts.)*
For the following problems p is defined as the probability of "success" and 1-p is the probability of "failure".

Fill out the 15 missing blanks in the table below.

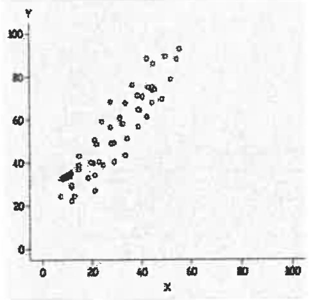
p	1/13	2/7	1/2	5/7	12/13	←Express p as a fraction.
1-p	12/13	5/7	1/2	2/7	1/13	← Express 1-p as a fraction.
odds	1/12	2/5	1	5/2	12	← Express odds as a fraction.
ln(odds)	-2.48	-0.92	0	0.92	2.48	← Round ln(odds) to 2 decimal places.

Question 5 (6 pts.)
a) Which plot violates linearity? *Circle one:* A B [✓ C]

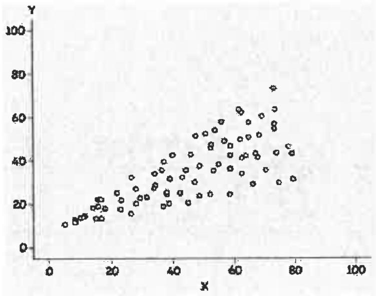
b) Which plot is linear but violates equal variability of the errors around the regression line? *Circle one:* A [✓ B] C

c) Which plot is well suited to linear regression analysis as is? *Circle one:* [✓ A] B C

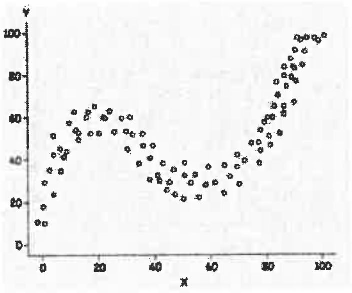
Plot A



Plot B



Plot C



Question 6 (4 pts.) True or False?

i) The logistic regression model only handles Y values that can be coded as 1's and 0's. *Circle one:* [✓ True] False

ii) A log transformation of any variable turns a linear regression model into a logistic regression model. *Circle one:* True [✓ False]

Stat 200 Dec 4, 2017

Question 7 *(30 pts.)* Below is the output from the logistic regression model predicting the probability of being Greek (a member of a fraternity or sorority) from gender (Males=0, Females=1) and # drinks per week, based on the 778 students who answered Survey 2. Let's treat them as if they were a random sample.

Y	R	n	# X's	Chi-square	df	p-value
greek	0.3352	778	2	113.8		0%

	Slopes	SE	Z	p-values
Intercept	-1.665	0.1856	-8.974	0%
gender	0.3956	0.1804		2.83%
drinks_per_week	0.09134	0.009694	9.422	0%

a) *(2 pts.)* A χ^2 test was done for the overall regression and a Z-test for the individual slopes. Could we have used F and t tests instead? *Circle one:*

[✓ i) No] ii) Yes, but it's not needed since the sample size is large.

b) *(1 pt.)* How many df for the χ^2 test? 2

c) *(2 pts.)* Calculate the Z stat to test: H_o: $\text{Slope}_{\text{gender}} = 0$. *Show work and round answer to 2 decimal places.*

$\frac{0.3956}{0.1804}$ Z= 2.19

d) *(1 pt.)* What's the **log (odds)** form of the logistic regression equation for the probability of being Greek?

$\ln\left(\frac{\hat{p}}{1-\hat{p}}\right)$ = -1.665 + 0.3956 **Gender** + 0.09134 **Drinks**

e) *(2 pts.)* Are females more or less likely than males to be Greek given the same level of drinking? **Circle one:**

[✓ i) More] ii) Less iii) Same iv) Not enough info

f) *(4 pts.)* Calculate the odds ratio for Gender and Drinks? ***Show work and round answer to 2 decimal places.***

i) Gender $\hat{OR}$ = $e^{0.3956}$ = 1.49 ii) Drinks $\hat{OR}$ = $e^{0.09134}$ = 1.1

g) *(6 pts.)* Use this rounded equation: **ln(odds) = -1.7 + 0.4 Gender + 0.1 Drinks** to predict the ln(odds), odds, and probability of being Greek for the individuals in the table below: ***Show work and Round answers to 2 decimal places.***

Gender: 0 =M 1=F	Drinks	ln(odds)	Odds	p
Male	20	$-1.7 + 0.1(20) = 0.3$	$e^{0.3} = 1.35$	$\frac{1.35}{2.35} = 0.57$
Female	$0 = -1.7 + 0.4 + 0.1\text{Drnk}$ $\text{Drinks} = \frac{1.7 - .4}{.1} = 13$	0	1	0.5

h) *(2 pts.)* Two males differ in their number of drinks per week by 5, compare their **odds** of being in a fraternity (given our logistic model). The heavier drinker has__________ times greater **odds** of being Greek.

i) 1.1 x 5 [✓ ii) 1.1^5] iii) 1.49 x 5 iv) 1.49^5 v) Not enough info

i) *(2 pts.)* Would your answer to (h) above change if you're comparing the odds of 2 females with a 5 drink difference?

i) Yes, it would be bigger ii) Yes, it would be smaller [✓ iii) No, it would be the same.]

j) *(2 pts.)* Two males differ in their number of drinks per week by 5, compare their **probability** of being Greek (given our logistic model). The heavier drinker has__________ times greater **probability** of being in a fraternity.

i) same answer as in (h) above ii) answer in h/(1 + answer in h) [✓ iii) Not enough info]

k) *(2 pts.)* Construct a 95% Confidence Interval for the **Gender** slope. (Use Gender slope = 0.4 with SE = 0.18)

a) 0.4 +/- 0.18 [✓ b) 0.4 +/- 0.36] c) 0.4 +/- 0.95(0.18)

l) *(2 pts.)* Construct a 95% Confidence Interval for the **Odds Ratio** for **Gender**.

$e^{0.4-0.36}$ to $e^{0.4+0.36}$

a) $e^{0.4} \pm e^{0.18}$ b) $e^{0.4} \pm e^{0.36}$ c) $e^{0.4} \pm e^{0.95(0.18)}$ [✓ d) $\left(\frac{e^{0.4}}{e^{0.36}} \text{ to } e^{0.4}e^{0.36}\right)$]

m) *(2 pts)* Since the 95% Confidence Interval for the Gender slope did not include 0, the p-value for a 2-sided Z test < 5 % and the p-value for a 1-sided test < 2.5 %. *Fill in the 2 blanks with numbers.*

Stat 200 Dec 4, 2017

Question 8 *(25 pts.)*A predictor of 5 year survival rate from breast cancer is the diameter of the tumor. Below is the log odds regression equation predicting the probability of survival after 5years from the diameter of the tumor measured in cm from a hypothetical study of a 100 patients.

$$\mathbf{ln}\left(\frac{\hat{p}}{1-\hat{p}}\right) = \mathbf{3 - 0.6\ Diameter}$$

a) *(2 pts.)*Use the above equation to estimate the ln(odds) and odds of 5 yr survival for a patient with a tumor of 3 cm.
Round answers to 2 decimals.
ln (odds)= $3 - 0.6(3) = \boxed{1.2}$ Odds= $e^{1.2} = 3.32$

b) *(2 pts.)* What is the ***probability*** of 5 yr survival for a patient with a tumor of 3 cm.
Round answer to 2 decimals.
Probability is $\frac{3.32}{4.32} \times 100 =$ % $\boxed{76.85\%}$

c) *(2 pts.)* How does the estimated *probability* of surviving 5 years change if the tumor increases in diameter by 1 cm? **Circle one:**

i) It changes by a fixed additive amount regardless of the tumor size. (i.e., there's a constant slope in the probability vs. size plot)

ii) It changes by the fixed multiplicative factor, $e^{-0.6}$

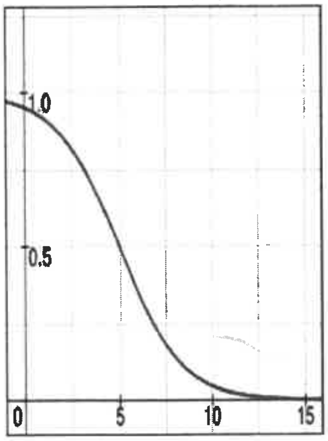
[✓ iii)] Neither of the above, you can't describe how the probability changes with either an additive or multiplicative constant since probability is bounded between 0 and 1.

d) *(6 pts)* What diameter does the tumor have to have for the estimated probability of 5-year survival to be 20% and 80%? Answer by filling out the table below. **ln (p/(1-p)) = 3 - 0.6 Diameter.** *Round answers to 2 decimal places. Show work for 1st column.*

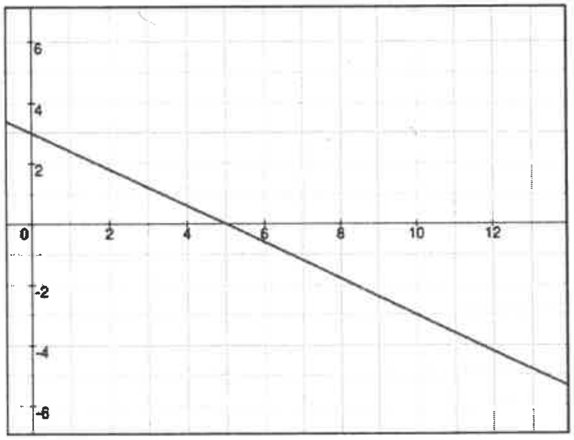
Tumor Diameter in cm	ln(Odds)	Odds	P
$-1.39 = 3 - 0.6D \rightarrow 0.6D = 4.39$ $D = \frac{4.39}{0.6} = \boxed{7.32}$	-1.39	$\frac{.2}{.8} = \frac{1}{4}$	0.2
$1.39 = 3 - 0.6D \rightarrow D = \frac{3-1.39}{0.6} = \boxed{2.68}$	1.39	$\frac{.8}{.2} = 4$	0.8

Below are plots depicting the probability, the odds or the ln(odds) of surviving 5 years based on the breast tumor size. The X axis is diameter of the tumor in cm and the Y axis is either probability, odd or ln(odds) of survival.

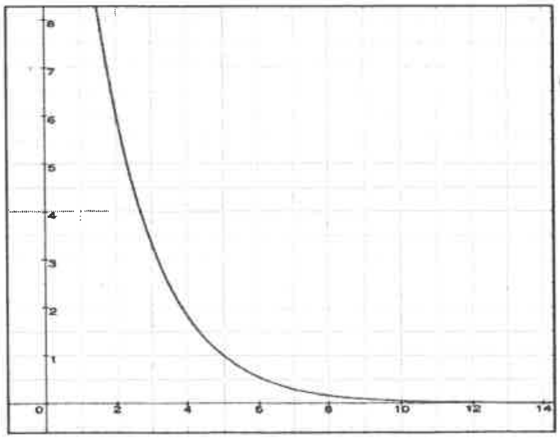
Plot A



Plot B



Plot C



e) *(3 pts)* Which plot is which? Plot A depicts probability, Plot C depicts odds, and Plot B depicts ln(odds).
Fill in the 3 blanks above with A, B, or C.

f) *(4 pts)* The ln(odds) equation is **ln (p/(1-p)) = 3 - 0.6 X**, where X= diameter of tumor. What are the odds and probability equations?

i) *(2 pts)* Odds equation: p/(1-p) = $e^{3-0.6X}$ **ii)** *(2 pts)* Probability equation: p = $\frac{e^{3-0.6X}}{1+e^{3-0.6X}}$

g) *(2 pts)* Judging from the plots what tumor diameter size gives a 50-50 chance of surviving 5 years? 5 cm

h) *(4 pts)* The diameter size that gives a 50-50 chance of surviving has a y-value = 0 in Plot B and a y-value = 1 in Plot C.