Stat 200 Exam 2 Study Guide Updated covering Chapters 21-41

The only 2 formulas that will be given to you on Exam 2 are:

$$SD_{errors} = \sqrt{1-r^2} * SD_y \quad \text{and} \quad SE_{slope} = \frac{SD_{errors}}{\sqrt{n} * SD_x} = \frac{\sqrt{1-r^2} * SD_y}{\sqrt{n} * SD_x}$$

Formulas not given to you that you need to know:

- Slope of the regression line = $r\frac{SD_y}{SD}$
- **Correlation Coefficient,** $r = \frac{\sum_{i=1}^{n} Z_x Z_y}{n}$
- **Z and t test stats for testing H_0 : slope=0 in simple regression (1 slope):**

$$Z = \frac{r}{\sqrt{1-r^2}} * \sqrt{n} \qquad t_{(n-2)} = \frac{r}{\sqrt{1-r^2}} * \sqrt{n-2}$$

NOTE: The Z and t formulas are the same as the square root of the χ^2 and F formulas below when p=2.

- **Chi square and F stats for testing H_0 : All slopes = 0 in multiple regression**

$$\chi^2_{(p-1)} = \frac{R^2}{1-R^2} * n \qquad F_{(p-1,n-p)} = \frac{R^2}{1-R^2} * \frac{n-p}{p-1}$$

- **ANOVA for regression: SST= SSM + SSE and ANOVA for group means SST= SSB + SSW (see summary on p.175)**

- **Formulas on page 186 for testing group means using:**

$SE^+_{diff} = SD^+_{errors}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ and Bonferoni corrected p-values= p-value * g(g-1)/2

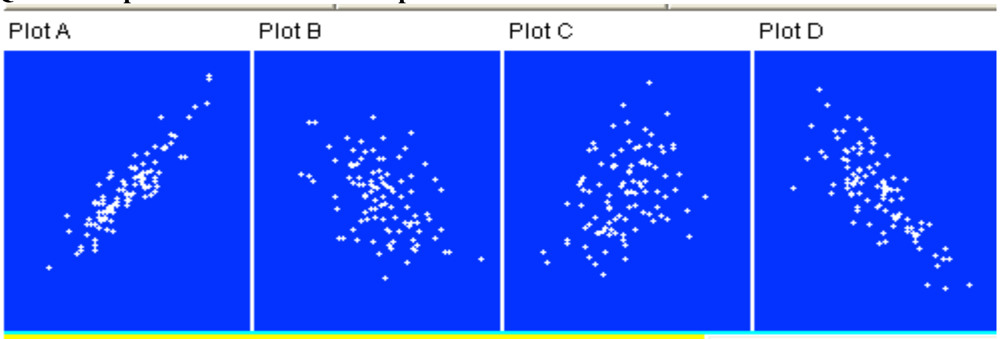
For **regression**: #parameters (p) = # of β's in regression equation, for **means: #** parameters (p) = # of groups (g)

Source	SS (Sum of Squares)	df
Model	$R^2 SST$ SSM (reg) SSB (means)	 p-1 g-1
Error	$(1-R^2)SST$ SSE (reg) SSW (means)	 n-p n-g
Total	SST	n-1

Stat 200 Exam 2 Study Guide Updated covering Chapters 21-41

Part VIII Simple Regression: Chapters 21-23

Question 1 pertains to the 4 scatter plots below:



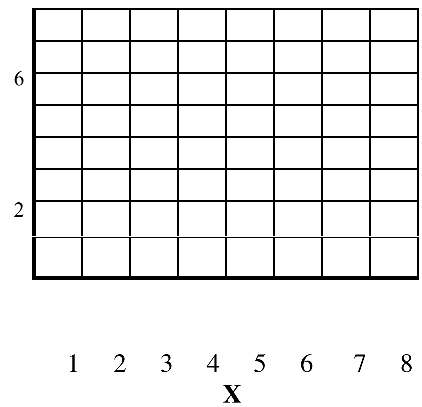
Write the letter of the plot next to the correlation coefficient that is closest to it.

a) r = 0.36______ **b)** r = 0.9__________ **c)** r = -0.79_________ **d)** r = -0.46________

Question 2

Compute the correlation coefficient (**r**) between X and Y by filling in the table below. Plot the points on the graph and check that the plot and r agree.

X	Y	X in Standard Units	Y in Standard Units	Products
2	4			
4	6			
5	5	0		
6	2	0.5		
8	8			



a) The correlation coefficient, r = ____________

b) Using the result of part (a), determine the correlation coefficient for each of the following data sets. No computation is necessary. *Write your answers in the blanks provided*. ***Your answer should be a number.***

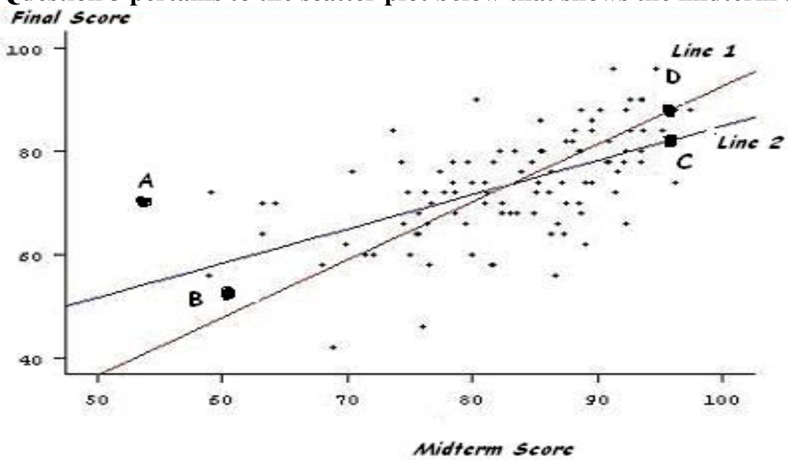
x	*y*
2	-8
4	-12
5	-10
6	-4
8	-16
r = __________	

x	*y*
8	8
5	5
2	4
4	6
6	2
r = ________	

x	*y*
4	4
6	6
7	5
8	2
10	8
r = _____	

x	*y*
4	2
6	4
5	5
2	6
8	8
r = _____	

Question 3 pertains to the scatter plot below that shows the midterm and final exam scores for 107 students.



	Average	*SD*
Midterm	*83*	*9*
Final	*74*	*10*

Correlation: r = 0.6

a) Which is the regression line? ***Choose one:*** i) Line 1 ii) Line 2

b) Look at students A, B, C and D on the graph. How did their actual scores on the final compare to their predicted scores? For each student circle whether their actual final exam scores were better than, worse than, or the same as the regression line predicted from their midterm scores.

	Actual Final Scores Compared to Predicted Ones			
Student A	*Choose One:*	*Better*	*Worse*	*The Same*
Student B	*Choose One:*	*Better*	*Worse*	*The Same*
Student C	*Choose One:*	*Better*	*Worse*	*The Same*
Student D	*Choose One:*	*Better*	*Worse*	*The Same*

c) Without any information about a particular student's midterm score, what would you expect him to score on the Final? ______________

d) About 68% of the time, your prediction in part (c) will be correct to within _________ points.

e) Suppose you are told that the student has a midterm score of 74. Now what would you predict for his score on the final exam? Use the 3 step process (not the regression equation) ***Show your work!*** ***Circle answer.***

f) About 68% of the time, your prediction in part (e) will be correct to within __________ points. ***Show your work!***

g) If a student was exactly average on both the midterm and the final which line would he fall on?
Choose one: Only the SD Line Only the Regression Line Both Neither

h) If a student was exactly1 SD above average on both the midterm and the final which line would he fall on?
Choose one: Only the SD Line Only the Regression Line Both Neither

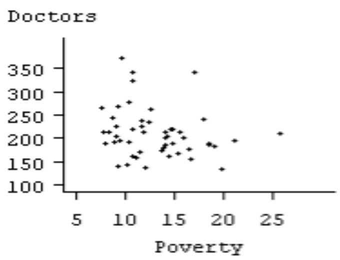
i) If a new scatter plot was drawn with 10 pts. added to everyone's final score then the correlation between midterm and final scores would**….** ***Choose one:*** i) increase ii) decrease iii) stay the same
(For (i) and (j) assume that final scores are allowed to exceed 100)

j) If a new scatter plot was drawn with 10 % added to everyone's final score then the correlation between midterm and final scores would___________ ***Choose one:*** i) stay the same ii) decrease iii) increase

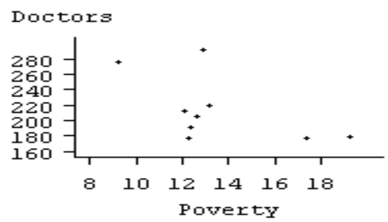
h) If point A was removed the, r would … i) Decrease ii) Increase iii) Stay the Same

Question 4
The following scatter plots show the relation between poverty level (percentage of people living below the poverty line) and number of doctors (per 100,000 people) by state and by geographical region. The graph on the left has 50 points, one for each **individual** state's poverty and doctor level. The graph on the right has the same information condensed into 9 points, one for each of the 9 geographical regions. (In other words, the 50 states were divided into 9 geographical regions. The average poverty and doctor level was computed for each region.)

By State



By Division



a) The correlation coefficient for the graph on the left is -0.2. The correlation for the graph on the right is closest to
i) -0.2 ii) -0.6 iii) 0 iv) 0.2 v) 0.6

b) The scatter plots above are an illustration of
i) The Regression Effect ii) Simpson's Paradox iii) Ecological Correlation iv)Negative Correlation

Question 5 For each of the following pairs of variables, check the box under the column heading that best describes its correlation among typical STAT 100 students:

	Correlation		Exactly -1	Between -1 and 0	About 0	Between 0 and 1	Exactly +1
a)	Weight in lbs.	Weight in kilograms (There are 2.2 lbs./kg)	☐	☐	☐	☐	☐
b)	Weight in lbs.	GPA	☐	☐	☐	☐	☐
c)	Freshman GPA	Sophomore GPA	☐	☐	☐	☐	☐
d)	How much you fall asleep in class	How much sleep you got the night before	☐	☐	☐	☐	☐
e)	Number of Points scored on Exam 1	Number of points missed on Exam 1	☐	☐	☐	☐	☐

Question 6

Here are the (rounded) summary statistics for height and weight of the 325 men in our class who completed Survey 1.

	Average	***SD***
Height	*71"*	*3"*
Weight	*175 lbs.*	*30 lbs.*

Correlation: r = 0.5

a) One student is exactly one SD above average in height and falls on the regression line. How many lbs. does he weigh?

b) Another student is 65" tall, predict how many lbs he weighs. ***Show work. Circle answer.***

c) What is the RMSE when predicting weight from height? ***Show work. Circle answer. Round your answer to the nearest lb.***

d) If a student is 71" and weighs 175 lbs. he would fall on the …..

Choose one:

i) SD line only ii) regression line only iii) Neither iv) Both

e) What is the slope for ***predicting weight*** from height?

Show work, circle answer.

f) The men in our class who are 68" weigh 160 lbs. on the average. Can you conclude that the men in our class who weigh 160 lbs. are 68" tall on the average?

Choose one:

i) Yes
ii) No, they'd be taller than 68" on the average.
iii) No, they'd be shorter than 68" on the average.

g) The regression equation for predicting height from weight is : **Height = .05 inch/lb * (Weight) + _______**

Find the y-intercept. ***Show work, write answer in blank below. Give your answer to 2 decimal places.***

h) If all the heights of the men were converted to centimeters (by multiplying each height by 2.54 cm/inch) the correlation coefficient would …

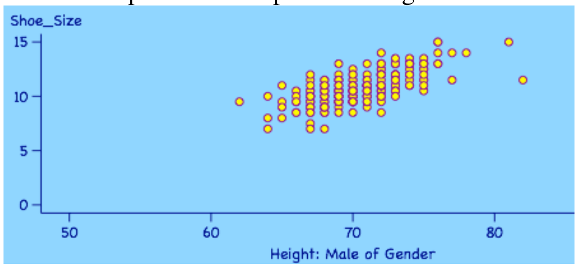
Choose one: i) increase ii) decrease iii) stay the same iv) not enough information given

Stat 200 Exam 2 Study Guide Updated covering Chapters 21-41

Part X: Inference for Regression: Chapters 24-27

Question 7 Part I

The scatter plot below depicts the height and shoe size of 100 UI male undergrads



	Avg	*SD*
Height	71”	3”
Shoe Size	11	1.5

r = 0.7

a) Find the slope and y-intercept of the regression equation for predicting shoe size from height.

Shoe Size = ___________ Height + ___________ (Round to 2 decimal places.)

b) What is the SD_{errors} for predicting shoe size from height?

i) 3 ii) 1.5 iii) 0.51 iv) 0.71 v) 1.07 vi) 2.14

Question 7 part II deals with *inference*—using the sample slope to make inferences about the population slope.

Now suppose the 100 students from Question 7 were randomly chosen from all male UI undergrads.

a) This corresponds to drawing _______points, at random ___________replacement from a scatter plot depicting (write a number in the first blank and “with” or “without” in the second blank)

i) the heights and shoe sizes of all male UI undergrads
ii) the heights and shoe sizes of the 100 randomly drawn students
iii) the heights and shoe sizes of all UI undergrads

b) Our best estimate of the **slope** for the whole population = _____ with a SE = _____.
Show work for SE. Round to 3 decimal places. You don’t need to re-calculate the sample slope.

c) Find the following confidence intervals for the slope of ***all*** UI undergrads when predicting shoe size from height. (Round answers to 3 decimal places.) Use the Normal Curve.

90% Confidence Interval = ______ +/- _____SE_{slope} = (_____________to ____________)

95% Confidence Interval = _______ +/- _____SE_{slope} = (_____________to ____________)

d) In part (c) above we saw that a 90% confidence interval for slope did not include 0.
Based only on that information, you could conclude that a Z test for slope would _________ the null hypothesis that $slope_{pop} = 0$ against the alternative that slope ____0 at $\alpha = 10\%$.

Fill in the 1st blank with “reject” or “not reject” and the 2nd with “ >” or “≠”.
(Hint: 90% CI interval has 5% area in each tail.)

Question 7 Part III: Z and t tests for Slope in Simple Regression
Formulas you'll need to know. (Or derive them from the 2 formulas you're given.)

$$Z_{\text{slope}} = \frac{\text{obs slope - exp slope}}{SE_{\text{slope}}} = \sqrt{n}\frac{r}{\sqrt{1-r^2}} \qquad t_{\text{slope}} = \frac{\text{obs slope - exp slope}}{SE^{+}_{\text{slope}}} = \sqrt{n-2}\frac{r}{\sqrt{1-r^2}}$$

a) Compute the Z statistic to test **H_0: $\text{slope}_{\text{pop}}$=0** **$H_a$: $\text{slope}_{\text{pop}}$>0**

b) To change the Z-stat above to a t-statistic you would multiply by ________.

i) $\sqrt{\frac{98}{100}}$ ii) $\sqrt{\frac{100}{98}}$ iii) $\frac{100}{98}$ iv) $\frac{98}{100}$ v) $\sqrt{\frac{99}{100}}$ vi) $\sqrt{\frac{100}{99}}$

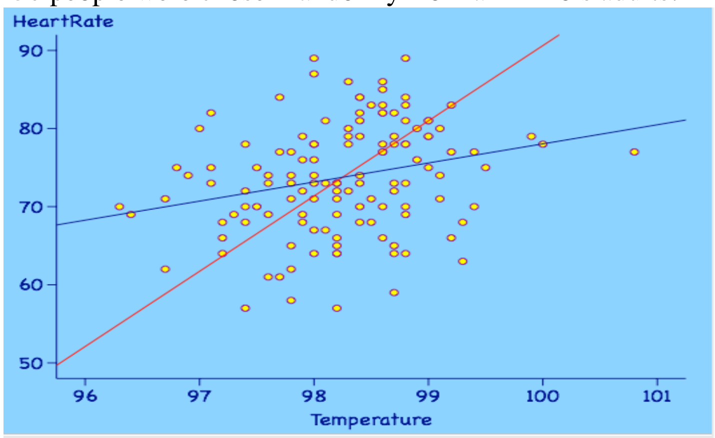
c) How many degrees of freedom does the t-test have? _____

d) How do p-values for Z and t tests compare when performed on the same data sets with the same null and alternative hypotheses?

i) Z tests will always yield smaller p-values
ii) Z tests will always yield larger p-values
iii) Both tests will yield exactly the same p-values
iv) Depending on the sample size the p-values from the z test could be larger, smaller or the same as the corresponding p-values from the t-test.

Question 8

The scatter plot below depicts the body temperatures and heart rates (beats per minute) of 130 adults. Pretend the 130 people were chosen randomly from all Illinois adults.



	Avg	*SD*
Temp	98	0.7
HR	74	7

r = 0.25

Sample Regression Equation

*Heart Rate= -171 +**2.5**(Temperature)*

a)What is the SE of the sample slope? Show work and round your answer to 2 decimal places.

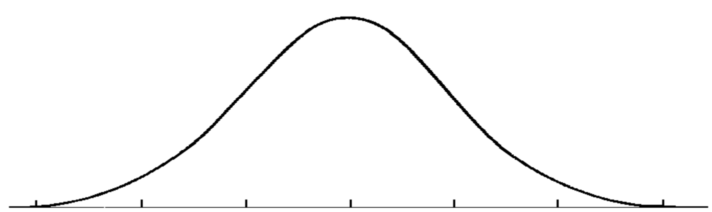
b) A 95% confidence interval for the population slope using the Normal Curve is (__________ to _____________). Round your answers to 2 decimal places.

c) The confidence interval above didn't include 0, so if we did a 2 sided Z test, testing the null hypothesis that the slope = 0 for the whole population we should _____________the null. Reject? or Not Reject? Circle one.

d) Do the hypothesis test by calculating Z and the p-value. The null and alternative are:

> **H_0:** Slope of the regression equation for the *whole* population is 0. We just happened to get a small slope of **2.5** in our sample of n=130 due to the luck of the draw.
>
> **H_a:** Slope of the regression equation for the whole population ≠ 0. Our sample slope of **2.5** is too big to be due to chance variation.

i) Calculate the test statistic Z for the slope.



ii) Mark Z on the Normal Curve and find p-value.

iii) Conclusion? Reject null?

Question 9

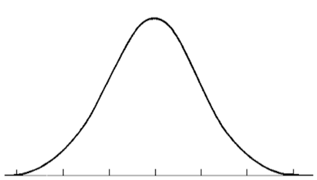
We're trying to fit a simple linear regression model for the whole population: $Y=\beta_0 + \beta_1X + \varepsilon$. (Assume ε are independent and normally distributed with constant variance). We draw a random sample of **n=7** from the population and get a sample correlation **r = 0.6**. Compute the 4 test statistics for testing the **null H_0: β_1=0**. (same as testing **H_0: $r_{population}$=0.**) ***(Round your final answers to 4 decimal places, but don't round during intermediate steps.)***

a) R^2 = _________ 1-R^2 = ____

b) Now compute the 4 statistics below.

	Z	χ^2	t	F
Compute the values of the 4 test statistics. *Show work below your answers.*	Z= ______ *(1 pt.)*	χ^2 = ______ *(1 pt.)*	t = _____ *(1 pt.)*	F=_____ *(1 pt.)*

c) Compute the p-values for each statistic. Assume the alternative for the Z and t test is 1-sided: **H_A: β_1 > 0**, and assume the alternative for the χ^2 and F is 2-sided: **H_A: $\beta_1 \neq 0$**.

Z	χ^2	t	F
p-value= ______% Label Z on the normal curve below and shade the area representing the p-value.  *	p-value=______% How many degrees of freedom? _____	Choose one: i) 1% ii) 2% ii) 7.7% How many degrees of freedom? _____	p-value=_______% Choose one: i) 2% ii) 4% iii) 15.4% How many degrees of freedom in numerator?___ in denominator?___

*If Z is between 2 lines on the Normal Table you may approximate middle area.

d) Suppose our sample y values are: 1, 2, 3, 4, 5, 6, 7. Compute the SST. (Show work).

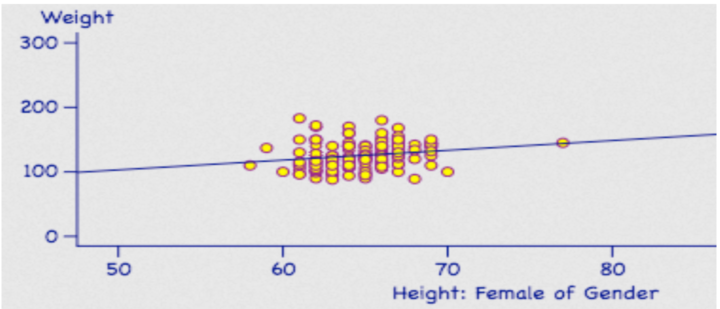
e) Compute SSM. _______ Hint: Use part (a)

Part X: Binary Variables in a Regression Model (Chapter 28--30)

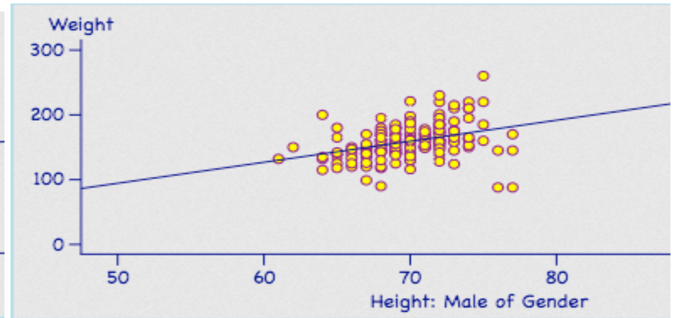
Question 10

The scatter plots below show the Height (in inches) on the X axis and the Weight (in lbs.) on the Y axis of the 123 females and 165 males in this class who responded to Survey 1.

Females



Males



Female: Weight = 27.18 + 1.517 × Height **Male: Weight = -67.68 + 3.242 × Height**

a) Translate the 2 simple regression equations into the multiple regression equation below. Assume Gender is a 0-1 variable coded with Males=0 and Females=1.

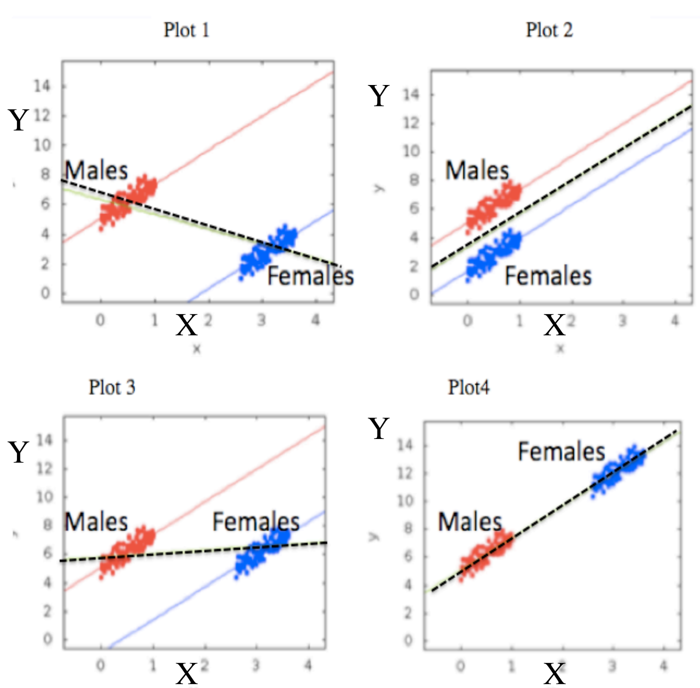
Weight = __________ + ___________***Height +** __________ **Gender** + ____________ **Gender*Height**

b) If you switched the code so that Males=1 and Females=0, what would the multiple regression equation be?

Weight = __________ + ___________***Height +** __________ **Gender** + ____________ **Gender*Height**

Question 11 Let's say the 4 plots below depict data from 4 populations and we're trying to figure if X causes Y in these 4 populations. Each plot consists of 2 groups (males and females as marked).

.



a) First let's focus on the relation between X and Ywithin each group.
Is there the same strong positive relation between X and Y for both males and females in each population?

i) No because males and females have different X values in some of the populations.

ii) Yes because they all have the same slope

iii) No, because males and females have different Y levels in some of the populations.

b) Now, let's focus on the overall regression effect (indicated by the dashed line) in the 4 plots.
For which plots does the overall regression effect agree with the group regression effects?

i) Plots 2 and 4 only, since the overall slope is the same as the group slopes.

ii) Only Plot 4 since the overall slope and the overall intercept is the same as the group slopes and intercepts.

iii) None of them because men and women are clearly separate groups in all 4 plots.

c) In which plots is the overall influence of X on Y confounded because of gender?

Circle all that apply: **i)** Plot 1 ii) Plot 2 iii) Plot 3 iv) Plot 4 v) None

d) In which plot is there an interaction effect between Gender and X?

Circle all that apply: i) Plot 1 ii) Plot 2 iii) Plot 3 iv) Plot 4 v) None

Question 12 (Also watch this [video](https://www.youtube.com/watch?v=Tw_M1GHZVhg) (https://www.youtube.com/watch?v=Tw_M1GHZVhg) if you need help with this type of problem)
Suppose A and B are 2 drugs designed to help improve test scores. The numbers in each table indicate the average number of points gained in 4 groups—those who received neither drug, those who received only Drug A, those who received only Drug B, and those who received both drugs. Each table describes a differenthypothetical study.
Fill in the missing blanks so that the equation and the tables match. Fill in ALL 12 blanks.

No Interaction
Points= _____ + _____A + ____B

	A=0	A=1
B=0	1	3
B=1	5	_____

Only Interaction
Points= _____ + ____ AB

	A=0	A=1
B=0	1	____
B=1	_____	6

Points= 3A + 4B – 2 AB

	A=0	A=1
B=0	____	____
B=1	_____	_____

Part XI: Multiple Regression (Quantitative X's) (Chaps 31-36)

Question 13 When the null hypothesis is true in a regression model with 6 parameters and large n, you'd expect your F stat to be about ______ and your χ^2 stat to be about __________ when the null is true. *Write a number in each blank*.

Question 14
To find out how education affects household income in Illinois, researcher collected data from **177** randomly selected Illinois Husband-Wife households on the following 3 variables: Years of Education of Wife (EducationW), Years of Education of Husband (EducationH), Total household Income. (The data is from 1989). Here's the multiple regression equation for predicting Household Income from Husband's and Wife's Education Years:

Predicted Household Income = -\$7, 580 + \$1,500/year * EducationW + \$3,000/year*EducationH

a) What does the \$3000/year slope mean in the multiple regression equation above?

i. For those husbands with wives at the **same** educational level, each extra year of husband's education increases household income \$3000 on the average.

ii. For all husbands, regardless of how educated their wives are, each extra year of husband's education increases household income \$3000 on the average.

b) Calculate the predicted Household Income for a married couple who each have only a 10th grade education (10 years of education each)?

\$_________________

c) Based on the correlation matrix at right, do you think the slopes in the 2 simple regression equations:

$$\text{Predicted Household Income} = \hat{\beta}_0 + \hat{\beta}_1(\text{EducationH})$$

$$\text{Predicted Household Income} = \hat{\beta}_0 + \hat{\beta}_1(\text{EducationW})$$

are the same as the slopes in the multiple regression above?

	EducationW	EducationH	Income
EducationW	1.000	0.5943	0.3280
EducationH	0.5943	1.000	0.3973
Income	0.3280	0.3973	1.000

i) Yes, they would still be \$1500 and \$3000 in the simple regression equations.
ii) No, they'll both be larger in the simple regressions since all variables are positively correlated.
iii) No, they'll both be smaller in the simple regressions because they're fewer variables.
iv) It's impossible to know because they're all correlated with each other.

d) The multiple correlation is **R= 0.4** (rounded). How was that calculated?

i. All 3 variables were converted to Z scores. Then R is the correlation between those 3 sets of numbers.

ii. Each of the 177 husband-wife pairs has a predicted income from the regression equation and an actual income. R is the correlation between those 2 sets of numbers, calculated by converting both sets to Z scores then taking the average of the product of their Z scores.

iii. R is the absolute value of the correlation coefficient between income, years of education of husbands, and years of education of wives.

e) Use **R= 0.41** and **n= 177** to compute the **Chi Square statistic** for testing the **overall** regression effect H_0: Both slopes=0 in the population. (Round to 2 decimal places.)

Chi Square =________ *Show work.*

Look at the Chi Square table. You need a χ^{2*} = _______ to reject the null at P= 0.05 (5%) and a χ^{2*} = _______ to reject the null at P=0.01 (1%).

f) Now compute F- statistic. F= ________ (Round to 2 decimal places.) *Show work.*

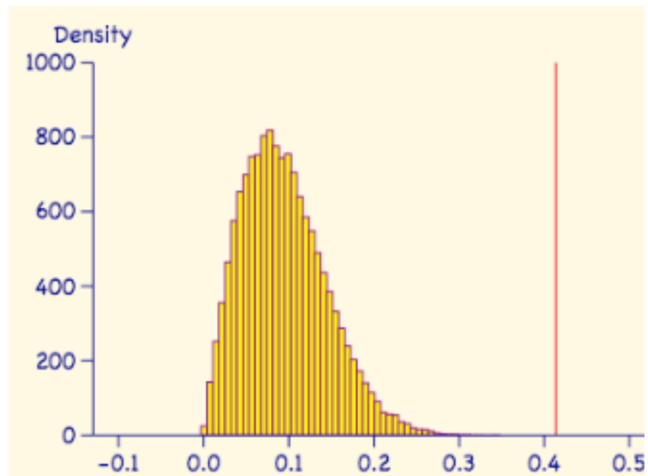
Look at the F table, you need a F^*=_____ to reject the null for P=0.05 (5%)
(177 isn't on the table so use 120 instead).

g) Conclusion from both the F and the Chi Square:

i) Reject null, both slopes must be significant
ii) Reject null, neither slope is significant
iii) Reject null, at least one of the slopes is significant.
iv) Cannot reject null, at least one of the slopes is significant

h) Another way to compute the p-value is by the re-randomization test. The histogram on the right shows the randomization test results of 50,000 randomizations showing the distribution of R's. What does the vertical line mark?

i. the specified significance level α
ii. the randomized R's that land at p-value = 0.1 %
iii. the value of our sample R.



i) The p-value given by the randomization test is closest to ….

i) 0
ii) 1%
iii) 5%
iv) not enough info

j) I did a t-test and a Z-test for the wife's education slope and got a p-value just under 5% by one test and just over 5% by the other test, which p-value belongs with which test?

i) The t-test must have given the bigger p-value since the t-curve has fatter tails.
ii) The z-test must have given the bigger p-value since the Z statistic is bigger.
iii) If done correctly the tests should have given exactly the same p-value.

k) If I delete Wife's Education from the model will R^2 go up or down?

i) It has to go down.
ii) It has to go up or stay the same.
iii) It could go up, down or stay the same depending on whether it is significant.

l) I decide to add a 3rd variable, either $\mathbf{X_{3a}}$ or $\mathbf{X_{3b}}$ to the full model since both look like good predictors of income on their own. I check the correlation matrix and see that X_{3a} has almost no correlation with either X variable already in the model, while $\mathbf{X_{3b}}$ has a correlation of 0.95 with Husband's education.

Which variable should I add to the full model?

i. It's a toss up**--** the higher the correlation the better the fit will be so $\mathbf{X_{3b}}$ is a good candidate, but X_{3a} adds a completely new element to the mix.
ii. Choose $\mathbf{X_{3b}}$**,** there's no point in adding something that does not fit well with the other X's. The X's need to work together. No correlation is equivalent to no communication. Predictive power is lost.
iii. Choose $\mathbf{X_{3a}}$, putting 2 variables that are highly correlated in the same model causes problems.

Question 15 There are 3 sections to the MCAT: Physical Science (PS); Biological Science (BS); and Verbal Reasoning (VR). Each is scored on a scale of 1-15. Suppose we randomly selected 55 UI pre-meds from all UI pre-meds who took the MCAT last year and got the following sample multiple regression equation for predicting PS from both VR and BS: $\hat{PS} = 1.6 + 0.2 \times VR + 0.6 \times BS$

Summary Stats

	Average	Median	SD	Min	Max	n
VR	9.764	10.00	1.768	6.000	13.00	55.00
BS	9.782	10.00	1.522	6.000	14.00	55.00
PS	9.709	10.00	1.659	5.000	14.00	55.00

a) Here's the ANOVA table to test the overall regression effect. Fill in them missing values. You'll need to use some info from the summary stats above to calculate SST.

Source	SS (Round to nearest whole number)	df	MS (Round to 1 decimal place)	(Round to 2 decimal places)
Model	SSM= 63			F=
Error	SSE=			SD^{+}_{errors}=
Total	SST=			R^2 =

b) Our F is **> or <** F^* =______ so our p-value **> or <** ______**%** so we **can or cannot** reject the null.
Circle the correct **">" or "<"** signs, fill in the 2 blanks and circle **"can" or "cannot"**

F Distribution critical values for P=0.001

Denominator DF	Numerator DF 1	2	3	4	5	7	10	15	20	30	60	120	500	1000
1	405284	499999	540379	562500	576405	592873	605621	615764	620908	626099	631337	633972	635983	636301
2	998.50	999.00	999.17	999.25	999.30	999.36	999.40	999.43	999.45	999.47	999.48	999.49	999.50	999.50
3	167.03	148.50	141.11	137.10	134.58	131.58	129.25	127.37	126.42	125.45	124.47	123.97	123.59	123.53
4	74.137	61.245	56.177	53.436	51.712	49.658	48.053	46.761	46.100	45.429	44.746	44.400	44.135	44.093
5	47.181	37.122	33.202	31.085	29.752	28.163	26.917	25.911	25.395	24.869	24.333	24.061	23.852	23.819
7	29.245	21.689	18.772	17.198	16.206	15.019	14.083	13.324	12.932	12.530	12.119	11.909	11.747	11.722
10	21.040	14.905	12.553	11.283	10.481	9.5174	8.7539	8.1288	7.8038	7.4688	7.1224	6.9443	6.8065	6.7846
15	16.587	11.339	9.3352	8.2526	7.5673	6.7408	6.0808	5.5351	5.2484	4.9502	4.6378	4.4749	4.3478	4.3275
20	14.819	9.9526	8.0984	7.0960	6.4606	5.6920	5.0753	4.5618	4.2900	4.0051	3.7030	3.5439	3.4184	3.3981
30	13.293	8.7734	7.0544	6.1245	5.5339	4.8173	4.2389	3.7528	3.4928	3.2171	2.9197	2.7595	2.6310	2.6100
60	11.973	7.7678	6.1712	5.3067	4.7565	4.0864	3.5415	3.0781	2.8265	2.5549	2.2522	2.0821	1.9390	1.9150
120	11.380	7.3212	5.7814	4.9471	4.4157	3.7669	3.2372	2.7833	2.5345	2.2621	1.9502	1.7668	1.6027	1.5736

c) When the null is true we'd expect our F to be about _____. Given how your F compares to that you'd expect the p-value to be about ____%.

d) Suppose you decided to reject the null, you'd conclude that

i. Both slopes must be significant
ii. The VR slope must be significant
iii. The BS slope must be significant
iv. The intercept must be significant
v. Either the VR or the BS slope or both must be significant

Question 16

On a Stat 100 Survey, 764 students reported how many drinks they typically consumed per week, how many hours they typically exercised per week and their GPA. The multiple regression equation predicting GPA from drinks and exercise yielded R=0.04. Assume these students were randomly sampled from a larger population of possible Stat 100 students.

a) Do a χ^2 test for the overall regression effect. How many degrees of freedom? _____

b) Compute the F stat. How many df in the numerator____? The denominator ____?

c) Here are the p-values for the 2 tests. Which one is for the χ^2 and which is for the F?
55.74% and 55.58% Label each as either χ^2 or F.

Question 17

In the overall regression test, the null hypothesis is that the population slopes all = 0. That's equivalent to the null hypothesis that in the population ….

i) R = 0
ii) R^2= 0
iii) $Y = \bar{Y}$
iv) all of the above
v) none of the above

Question 18

If the χ^2 test doesn't yield significant results, is it possible the F test still would?

i) Yes, since the F test yields slightly more precise tests.
ii) Yes, if the sample size is relatively small, the F test results could yield significantly different results.
iii) No, the p-value for the F test will always be greater so it could never yield more significant results.
iv) It's impossible to know since F is centered at 1 when the null is true and the χ^2 is centered at its degrees of freedom making comparisons of results statistically meaningless.

Part XIII Chapters 38- 41

Question 19
9 numbers are divided into 3 groups as shown below.

Group 1	Group 2	Group 3	
0	4	5	
2	6	7	
4	8	9	
Mean = 2	Mean = 6	Mean = 7	Overall Mean = 5

SST= 66

a) Compute SSB

b) Compute SSW (same as SSE)

c) The SST = 66. Use the SST to compute the SD. (Hint: The SST is the sum of the squared deviations.)

Question 20: 717 Stat 100 students rated their belief in the existence of ghosts on a scale of 1-10 (1 is certain ghosts don't exist and 10 is certain they do exist). They also classified their hometowns into 4 types: Small Town, Medium City, Suburb, and Big City.
Here are the results:

	Level of hometown	Average	SD	n
Ghosts	small_town	4.769	3.096	121
Ghosts	medium_city	4.115	2.833	104
Ghosts	suburb	5.140	3.106	356
Ghosts	big_city	5.309	3.064	136

a) What's the appropriate significance test the null that all 4 group means are the SAME in the "population". We just happen to see small differences in our sample due to chance?
  i) Z test only ii) t test only iii) F-test only iv) either z, t or F v) either t or F

b) What's the alternative hypothesis
  i) All 4 group means are different than each other in the population.
  ii) Some group means are different than each other in the population.
  iii) One of the group means is different than the others in the population.
  iv) That either i, ii, or iii is true.

Fill out the ANOVA table below to test the null hypothesis that all the group means are the same in the population we just happen to see differences in the sample due to sampling variation. ***Show work inside each box (except for the df column). Write your answers in the blanks provided.***

Source	SS (Sum of Squares)	df	Mean Square	F Statistic	P-value
Model	SSB=107	df=______	MSB=_________ (round to 1 decimal place)	F=______	You would have to look at the F curve with $df_{between}$ = ____, df_{within} = __ F^* at a=0.05 is = ___ Reject null at at a=0.05? Yes or No
Error	SSW=___________	df=______	MSW=_________ (round to 2 decimal places)	SD^+_{errors} = ____________ (Round to 2 decimal places)	
Total	SST=6813	df=______		R^2 = _________ (Round to 3 decimal places)	

Stat 200 Exam 2 Study Guide Updated covering Chapters 21-41

Question 21 On the Fall 2015 survey **707** Stat100 students rated whether to legalize marijuana on a scale of 0-10 (with 0 meaning strongly for legalization and 10 meaning strongly against). They also classified themselves into **6** ethnics groups. Imagine the students were randomly sampled from a much larger population of all Stat 100 students.

	Ethnicity	Average (rounded)	SD (rounded	n
Legalize Marijuana?	White	4.05	3.16	354
Legalize Marijuana?	Black	3.99	3.68	65
Legalize Marijuana?	Hispanic	4.46	3.34	100
Legalize Marijuana?	Asian	5.33	3.57	145
Legalize Marijuana?	Mixed	4.40	3.55	30
Legalize Marijuana?	Other	4.08	4.16	13

R= 0. 15

a) Compute the Chi Square Statistic to test the null that all group means are the same in the population. Show work. Round to 2 decimal places. **Circle answer.**

b) How many degrees of freedom for the χ^2 ? _________

c) Compute the **F** Statistic to test the null that all group means are the same in the population. Show work. Round to 2 decimal places. **Circle answer.**

d) How many degrees of freedom for the numerator_____? the denominator? _________

e) Below are the p-values for the two tests but I can't remember which is which? Identify the correct test by filling in the blanks with χ^2 **or F** **i) 0.6645%** is for the ____ **ii) 0.7472%** is for the ____

f) What do you conclude? **Choose one.**

- **i)** That all the group averages are significantly different from each other .
- **ii)** That at least one of the group averages is significantly different than the others.
- **iii)** That none of the group averages are significantly different from each other.

g) Compute the t-statistic to test whether the difference between Asians and Whites is significant.

i) What is the $SE^{+}_{difference}$? Use SD^{+}_{errors} = 3.375. Round your answer to 2 decimals.

ii) What is the t statistic?

iii) How many degrees of freedom? _______

iv) The uncorrected p-value is 0.013%. The Bonferroni correction would ____________the p-value by _____

Fill in the first blank with "multiply" or "divide" and the second with a number.

F Distribution critical values for P=0.05

Denominator DF	Numerator DF 1	2	3	4	5	7	10
1	161.45	199.50	215.71	224.58	230.16	236.77	241.88
2	18.513	19.000	19.164	19.247	19.296	19.353	19.396
3	10.128	9.5522	9.2766	9.1172	9.0135	8.8867	8.7855
4	7.7086	6.9443	6.5915	6.3882	6.2560	6.0942	5.9644
5	6.6078	5.7862	5.4095	5.1922	5.0504	4.8759	4.7351
7	5.5914	4.7375	4.3469	4.1202	3.9715	3.7871	3.6366
10	4.9645	4.1028	3.7082	3.4780	3.3259	3.1354	2.9782
15	4.5431	3.6823	3.2874	3.0556	2.9013	2.7066	2.5437
20	4.3512	3.4928	3.0983	2.8660	2.7109	2.5140	2.3479
30	4.1709	3.3159	2.9223	2.6896	2.5336	2.3343	2.1646
60	4.0012	3.1505	2.7581	2.5252	2.3683	2.1666	1.9927
120	3.9201	3.0718	2.6802	2.4473	2.2898	2.0868	1.9104
500	3.8601	3.0137	2.6227	2.3898	2.2320	2.0278	1.8496
1000	3.8508	3.0047	2.6137	2.3808	2.2230	2.0187	1.8402

F Distribution critical values for P=0.01

Denominator DF	Numerator DF 1	2	3	4	5	7	10
1	4052.2	4999.5	5403.4	5624.6	5763.6	5928.4	6055.8
2	98.503	99.000	99.166	99.249	99.299	99.356	99.399
3	34.116	30.817	29.457	28.710	28.237	27.672	27.229
4	21.198	18.000	16.694	15.977	15.522	14.976	14.546
5	16.258	13.274	12.060	11.392	10.967	10.455	10.051
7	12.246	9.5467	8.4513	7.8466	7.4605	6.9929	6.6201
10	10.044	7.5594	6.5523	5.9944	5.6363	5.2001	4.8492
15	8.6831	6.3588	5.4169	4.8932	4.5557	4.1416	3.8049
20	8.0960	5.8489	4.9382	4.4306	4.1027	3.6987	3.3682
30	7.5624	5.3903	4.5098	4.0179	3.6990	3.3046	2.9791
60	7.0771	4.9774	4.1259	3.6491	3.3388	2.9530	2.6318
120	6.8509	4.7865	3.9490	3.4795	3.1736	2.7918	2.4720
500	6.6858	4.6479	3.8210	3.3569	3.0539	2.6751	2.3564
1000	6.6603	4.6264	3.8012	3.3379	3.0356	2.6571	2.3387