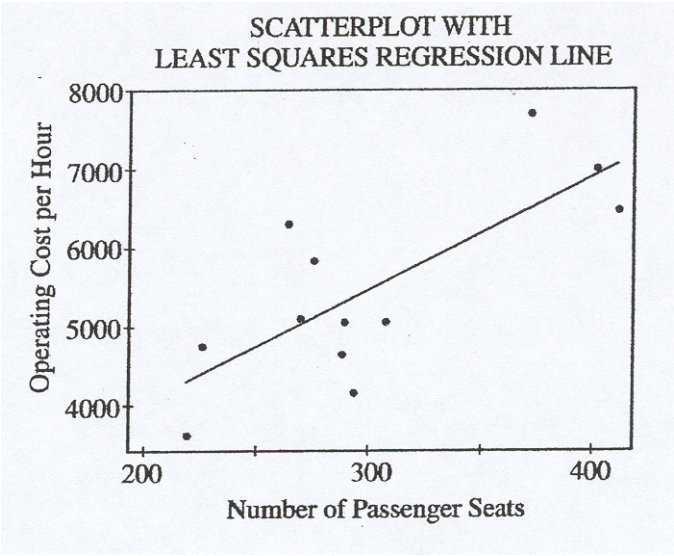
AP EXAM Review **Name** _____
Free Response Ch 7-10 **Per** ____

Linear Regression & Linear Transformations

1. Commercial airlines need to know the operating cost per hour of flight for each plane in their fleet. In a study of the relationship between operating cost per hour and number of passenger seats, investigators computed the regression of operating cost per hour on the number of passenger seats. The 12 sample aircraft used in the study included planes with as few as 216 passenger seats and planes with as many as 410 passenger seats. Operating cost per hour ranged between $3,600 and $7,800. Some computer output from a regression analysis of these data is shown below.



Predictor	Coef	SitDev	T	P
Constant	1136	1226	0.93	0.376
Seats	14.673	4.027	3.64	0.005
S = 845.3	R-Sq = 57.0%		R-Sq(adj) = 52.7%	

a. What is the equation of the least squares regression line that describes the relationship between operating cost per hour and number of passenger seats in the plane? Define any variables used in this equation.

operating cost = 14.673(Seats) + 1136 Seats = # of seats

b. What is the value of the correlation coefficient for operating cost per hour and number of passenger seats in the plane? Interpret this correlation.

$r = \sqrt{R^2} = \sqrt{.57} = .755$ There is a moderately strong linear positive correlation between # of seats and operating cost per hour.

c. Suppose that you want to describe the relationship between operating cost per hour and number of passenger seats in the plane for planes only in the range of 250 to 350 seats. Does the line shown in the scatterplot still provide the best description of the relationship for data in this range? Why or why not?

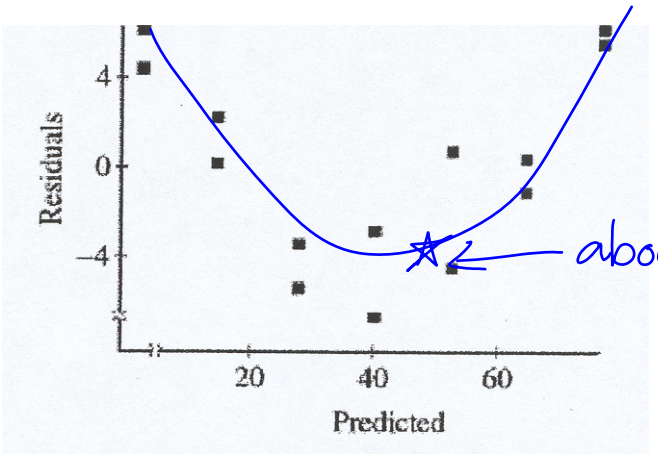
No, if we only looked at 250-350 seats, the data has a strong negative correlation. The best fit line operating cost = 14.673(Seats) + 1136 does not represent the data. The points around 400 are outliers and change the slope of the regression line.

2. In a study of the application of a certain type of weed killer, 14 fields containing large numbers of weeds were treated. The weed killer was prepared at seven different strengths by adding 1, 1.5, 2, 2.5, 3, 3.5, or 4 teaspoons to a gallon of water. Two randomly selected fields were treated with each strength of weed killer. After a few days, the percentage of weeds killed on each filed was measured. The computer output obtained from fitting a least squares regression line to the data is shown below. A plot of the residuals is provided as well.

Dependent variable is: percent killed
R squared = 97.2% R squared (adjusted) =96.9%
s = 4.505 with 14 — 2 = 12 degrees of freedom

Source	Sum of Squares	df	Mean Square	F-ratio
Regression	8330.16	1	8330.16	410
Residual	243.589	12	20.2920	

Variable	Coefficient	s.e. of Coeff	t-ratio	Prob
Constant	-20.5893	3.242	-6.35	≤0.0001
No. Teaspoons	24.3929	1.204	20.3	≤0.0001



about here

a. What is the equation of the least squares regression line given by this analysis? Define any variables used in this equation.

$\hat{y} = 24.3929x - 20.5893$

y = %of weeds killed x = # of teaspoons applied

b. If someone uses this equation to predict the percentage of weeds killed when 2.6 teaspoons of weed killer are used, which of the following would you expect?

- ○ The prediction will be too large.
- ○ The prediction will be too small.
- ○ A prediction cannot be made based on the information given on the computer output.

Explain your reasoning.

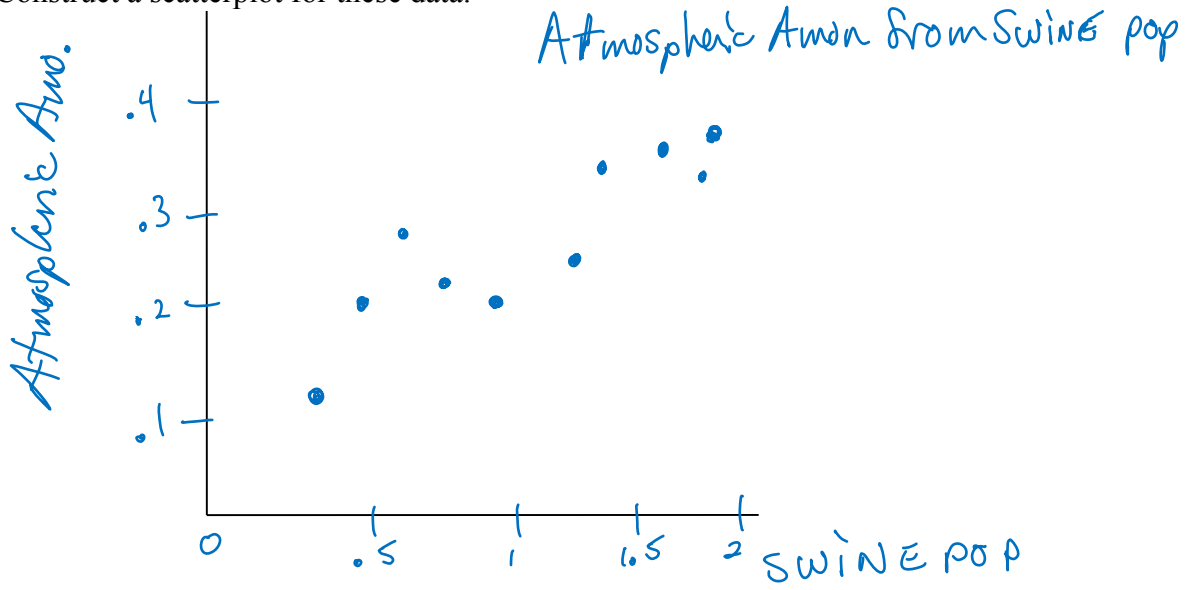
$\hat{y} = 24.3929(2.6) - 20.5893 = 42.83$ ← predicted.

Looking at the residual plot with a predicted value of 42.83 gives a residual of about −4. That means my actual is about 4 points below the predicted, so my actual % weeds killed is about 38.83%. The prediction would be too large.

3. Animal-waste lagoons and spray fields near aquatic environments may significantly degrade water quality and endanger health. The National Atmospheric Deposition Program has monitored the atmospheric ammonia at swine farms since 1978. The data on the swine population size (in thousands) and atmospheric ammonia (in parts per million) for one decade are given below.

Year	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997
Swine Population	0.38	0.50	0.60	0.75	0.95	1.20	1.40	1.65	1.80	1.85
Atmospheric Ammonia	0.13	0.21	0.29	0.22	0.19	0.26	0.36	0.37	0.33	0.38

a. Construct a scatterplot for these data.



b. The value for the correlation coefficient for these data is 0.85. Interpret this value.
There is a strong positive linear relationship between swine population and atmospheric ammonia.

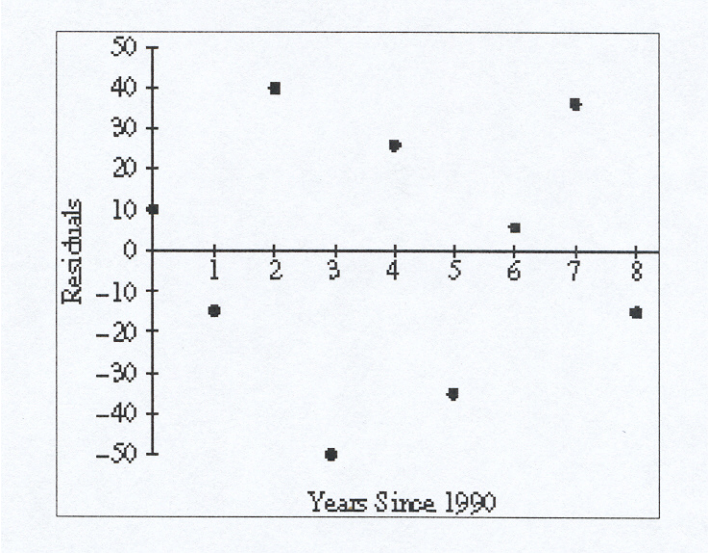
c. Based on the scatterplot in part **a** and the value of the correlation coefficient in part **b**, does it appear that the amount of atmosphere ammonia is linearly related to the swine population size? Explain.

Both the pattern of the scatterplot and the value of the correlation coefficient indicate atomospheric ammonia is linearly related to population size.

d. What percent of the variability in atmospheric ammonia can be explained by swine population size?

72% of the variability in ammonia can be explained by swine population Size.

. Lydia and Bob were searching the Internet to find information on air travel in the United States. They found data on the number of commercial aircraft flying in the United States during the years 1990-1998. The dates were recorded as years since 1990. Thus, the year 1990 was recorded as year 0. They fit a least squares regression line to the data. The graph of the residuals and part of the computer output for their regression are given below.



.

Predictor	Coef	Stdev	t-ratio	p
Constant	2939.93	20.55	143.09	0.000
Years	233.517	4.316	54.11	0.000
S =33.43				

a. Is a line an appropriate model to use for these data? What information tells you this?
Yes the model is appropriate because there is NO PATTERN in the RESIDUALS, and the t-ratio is enormous. If I tested whether the slope is positive or zero, I would overwhelmingly reject a zero slope in favor of a very positive slope.

b. What is the value of the slope of the least squares regression line?
Interpret the slope in the context of this situation.
B = slope = 233.517 aircraft/year. For every one additional year since 1990, about 233 more aircraft fly per year.

c. What is the value of the intercept of the least squares regression line?
Interpret the intercept in the context of this situation.
A = y int = 2939.93. It means in 1990 there were approximately 2939.93 commercial aircraft flying.

d. What is the predicted number of commercial aircraft flying in 1992?
Number of aircraft-hat = 233.517(years since 1990) + 2939.93 = 233.517(2) + 2939.93 = 3407 aircraft

e. What was the actual number of commercial aircraft flying in 1992?
Residual = actual – predicted
+40 = actual – 3407
Actual = 3447 aircraft in 1992.

5. Agricultural experts are trying to develop a bird deterrent to reduce costly damage to crops in the United States. An experiment is to be conducted using garlic oil to study its effectiveness as a nontoxic, environmentally safe bird repellant. The experiment will use European starlings, a bird species that causes considerable damage annually to the corn crop in the USA. Food granules made from corn are to be infused with garlic oil in each of five concentrations: 0%, 2%, 10%, 25%, 50%. The researchers will determine the adverse reaction of the birds to the repellant by measuring the number of food granules consumed during a two-our period following overnight food deprivation. There are 40 birds available for the experiment, and 8 will be used for each garlic concentration level. Each bird will be kept in a separate cage and provided with the same number of food granules.

A) For the experiment, identify the treatments.

Treatment is five different concentrations of garlic (0%, 2%, 10%, 25%, 50%0

B) For the experiment, identify the experimental units.

Experimental Units are the Birds eating the food pellets.

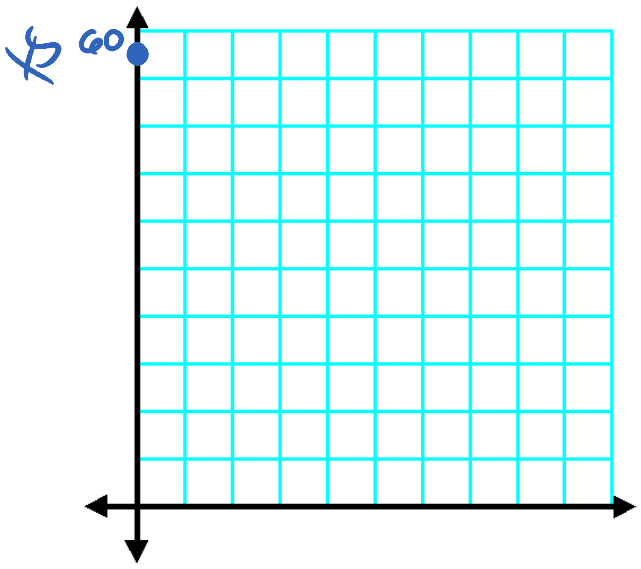
C) For the experiment, identify the response that will be measured.

We will measure the number of food granules consumed during a two-hour period following overnight food depreivation.

D) After performing the experiment, the researchers recorded the data shown in the table below.

Garlic Oil Concentration	0%	2%	10%	25%	50%
Mean # of food granules consumed	58	48	29	24	20
Number of Birds	8	8	8	8	8

E) Construct a graph of the data that could be used to investigate the appropriateness of a linear regression model for analyzing the results of the experiment.



F) Based on your graph, do you think a linear regression model is appropriate? Explain

6. Windmills generate electricity by transferring energy from wind to a turbine. A study was conducted to examine the relationship between wind velocity in miles per hour (mph) and electricity production in amperes for one particular windmill. For the windmill, measurements were taken on 25 randomly selected days, and the computer output for the regression analysis for predicting electricity production based on wind velocity is given below. The regression model assumptions were checked and determined to be reasonable over the interval of wind speeds represented in the data, which were from 10 mph to 40 mph.

Predictor	Coef	SE Coef	T	P
Constant	0.137	0.126	1.09	0.289
Wind Velocity	0.240	0.019	12.63	0.000
S = 0.237	R-sq = 0.873		R-sq adj= 0.868	

A) Use the computer output above to determine the equation of the least squares regression line. Identify all variables used in the equation.
Electricity production hat = 0.24 (wind velocity) – 0.137

Elect production is measured in amperes, and wind velocity in mph.

B) ow much more electricity would the windmill be expected to produce on a day when the wind velocity is 25 mph rather than 15 mph? Show how you arrived at your answer.

Elect prod hat = 0.24(25) -0.137 = 5.863
Elect prod hat = 0.24(15) – 0.137 = 3.463
5.863 – 3.463 = 2.4 Amperes more

C) What proportion of the variation in electricity production is explained by its linear relationship with wind velocity?
87.3% of the variation in electricity production is explained by its linear relationship with wind velocity.

D) Is there statistically convincing evidence that electricity production by the windmill is related to wind velocity? Explain.
Yes there is statistically convincing evidence b/c the t-test value is 12.63 and the probability associated with such a large t-value is about zero. That means the slope is NOT zero.
Also, the R-squared is high, the correlation coefficient (r) is high.

7. The statistics department at a large university is trying to determine if it is possible to predict whether an applicant will successfully complete the Ph.D. program or will leave before completing the program. The department is considering whether GPA (grade point average) in undergraduate statistics and mathematics courses (a measure of performance) and mean number of credit hours per semester (a measure of workload) would be helpful measures. A random sample of 20 entering students is taken.

Successfully completed PhD

Student	A	B	C	D	E	F	G	H	I	J	K	L	M
GPA	3.8	3.5	4.0	3.9	2.9	3.5	3.5	4.0	3.9	3.0	3.4	3.7	3.6
Credit Hours	12.7	13.1	12.5	13	15	14.7	14.5	12.0	13.1	15.3	14.6	12.5	14.0

Did NOT complete PhD

Student	N	O	P	Q	R	S	T
GPA	3.6	2.9	3.1	3.5	3.9	3.6	3.3
Credit Hours	11.1	14.5	14.0	10.9	11.5	12.1	12.0

The regression output below resulted from fitting a line to the data in each group. The residuals plots (not shown) indicated no unusual patterns, and the assumptions for inference were judged to be met.

Successfully completed PhD

Predictor	Coef	StDev	T	P
Constant	23.514	1.684	13.95	0.000
GPA	-2.7555	0.4668	-5.90	0.000
S=0.5658	Rsq = 76%			

Did NOT complete PhD

Predictor	Coef	StDev	T	P
Constant	24.200	3.474	6.97	0.001
GPA	-3.485	1.013	-3.44	0.018
S=0.8408	Rsq = 70.3%			

a) Use an appropriate graphical display to compare the GPAs for the two groups. Write a few sentences commenting on your display.

Range about the same. Iqr is the same. Phd is left skewed, NO Phd is more symmetric. 75% of NO Phd are below 50% of with a Phd.

b) For the students who successfully completed the PhD program, is there a significant relationship between GPA and mean number of credit hours per semester? Give a statistical justification.

High r = -.87. High r-squared = .76, low standard error. T-test value of -5.9 and corresponding p-value of zero is strong justification to reject Ho of slope = zero and say slope is negative. Checked residual plot and no pattern in the residuals. So all of these support a significant relationship between GPA and mean # of credit hours per semester.

c) In a new applicant has a GPA of 3.5 and mean number of credit hours per semester of 14, do you think this applicant will successfully complete the PhD program? Give a statistical justification.

Phd y-hat = -2.755(3.50 – 23.514 = 13.87 Resid = Act – Pred = 14 – 13.87 = 0.13

NO Phd y-hat = -3.485(3.5) – 24.2 = 12 Resid = Act – Pred = 14 – 12 = 2

Yes, she WILL complete the Phd program because the Phd line is a better predictor with a smaller residual.

8.

A) Given this scatterplot, is the linear model appropriate? Why or why not?
Does NOT look appropriate because the graphs curves and is either exponential decay or inverse.

B) Circle which best fit line is the correct equation?

$\widehat{HWY\ Mileage} = -13.5Ln(Engine\ Size) + 43$

$\widehat{Engine\ Size} = -13.5Ln(HWY\ Mileage) + 43$

$\widehat{Engine\ Size} = 43Ln(HWY\ Mileage) - 13.5$

$\widehat{HWY\ Mileage} = 43Ln(Engine\ Size) - 13.5$

C) With the correct model, estimate the HWY mileage for a truck with an 8 liter engine.
HWY mileage hat = -13.5(Ln(8)) + 43 = 14.928 mpg

D) With the correct model, calculate the residual for a 2 liter engine.
HWY mileage hat = -13.5 (Ln(2)) + 43 = 33.643 predicted
Actual from graph is 32
Residual = actual – predicted = 32 – 33.643 = -1.643