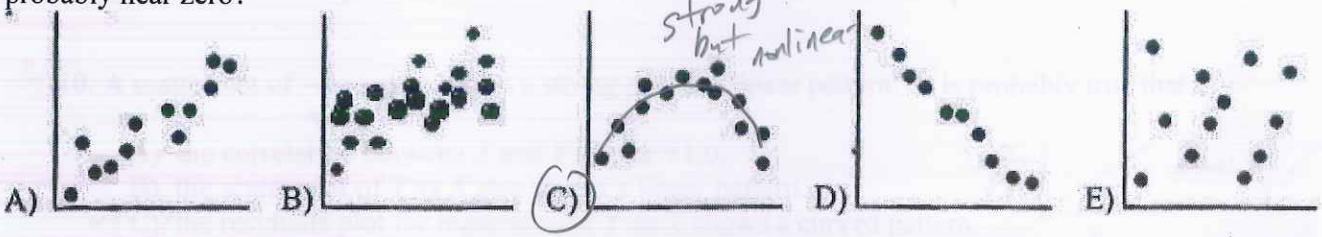
AP Statistics – Unit 2 (combined) Practice Test SOLUTIONS

C 1. Researchers studying growth patterns of children collect data on the heights of fathers and sons. The correlation between the fathers' heights and the heights of their 16 year-old sons is most likely to be…
A) near -1.0 B) near 0 C) near +0.7 D) exactly +1.0 E) somewhat greater than 1.0

D 2. The auto insurance industry crashed some test vehicles into a cement barrier at speeds of 5 to 25 mph to investigate the amount of damage to the cars. They found a correlation of $r = 0.60$ between speed (MPH) and damage ($). If the speed at which a car hit the barrier is 1.5 standard deviations above the mean speed, we expect the damage to be ___?___ the mean damage.
A) equal to B) 0.36 SD above C) 0.60 SD above D) 0.90 SD above E) 1.5 SD above

$r = \frac{s_y}{s_x}$, so $s_y = s_x = 1$ $\quad s_y = r s_x = 0.6(1.5) = 0.90$

C 3. Which scatterplot shows a strong association between two variables even though the correlation is probably near zero?



A 4. The correlation between X and Y is $r = 0.35$. If we double each X value, decrease each Y by 0.20, and interchange the variables (put X on the Y-axis and vice versa), the new correlation:
A) is 0.35 B) is 0.50 C) is 0.70 D) is 0.90 E) cannot be determined

B 5. The correlation between a family's weekly income and the amount they spend on restaurant meals is found to be $r = 0.30$. Which must be true?
✗ I. Families tend to spend about 30% of their incomes in restaurants. (r is not a proportion)
✓ II. In general, the higher the income, the more the family spends in restaurants. r+, so b+
✗ III. The line of best fit passes through 30% of the (*income, restaurant$*) data points. (LSRL may pass through no points)
A) I only B) II only C) III only D) II and III only E) I, II, and III

D 6. A medical researcher finds that the more overweight a person is, the higher his pulse rate tends to be. In fact, the model suggests that 12-pound differences in weight are associated with differences in pulse rate of 4 beats per minute. Which is true?
$\frac{4 \text{ beats}}{12 \text{ min}}$ is a slope, not r
✗ I. The correlation between pulse rate and weight is 0.33.
✗ II. If you lose 6 pounds, your pulse rate will slow down 2 beats per minute. too strongly worded
✓ III. A positive residual means a person's pulse rate is higher than the model predicts. always true
A) none B) I only C) II only D) III only E) II and III only

A 7. Education research consistently shows that students from wealthier families tend to have higher SAT scores. The slope of the line that predicts *SAT score* from *family income* is 6.25 points per $1000, and the correlation between the variables is 0.48. Then the slope of the line that predicts *family income* from *SAT score* (in $1000 per point)…
A) is 0.037 B) is 0.16 C) is 3.00 D) is 6.25 E) is 13.02

$b_1 = r\frac{s_y}{s_x}$ $\quad 6.25 = .48\left(\frac{s_y}{s_x}\right)$ $\quad \left(\frac{s_y}{s_x}\right) = \frac{6.25}{.48}$

$b_2 = r\frac{s_x}{s_y}$ $\quad b_2 = .48\left(\frac{.48}{6.25}\right) = 0.0368$

A 8. A regression analysis of company profits and the amount of money the company spent on advertising found $r^2 = 0.72$. Which of these is true?

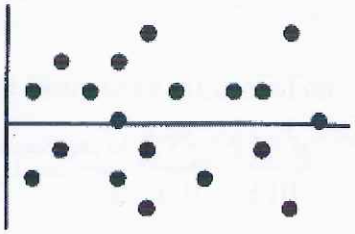
About 72% of the variation in profits is explained by the LSRL model which predicts profit from advertising.

X I. This model can correctly predict the profit for 72% of companies.
X II. On average, about 72% of a company's profit results from advertising.
X III. On average, companies spend about 72% of their profits on advertising.

A) none B) I only C) II only D) III only E) I and II only

A 9. A least squares line of regression has been fitted to a scatterplot; the model's residuals plot is shown. Which is true?

A) The linear model is appropriate. (no pattern in residuals)
B) The linear model is poor because some residuals are large.
C) The linear model is poor because the correlation is near 0.
D) A curved model would be better.
E) None of the above.



C 10. A scatterplot of $\frac{1}{\sqrt{y}}$ vs. x shows a strong positive linear pattern. It is probably true that…

A) the correlation between X and Y is near +1.0.
B) the scatterplot of Y vs X also shows a linear pattern.
C) the residuals plot for regression of Y on X shows a curved pattern.
D) large values of X are associated with large values of Y.
E) accurate predications can be made for Y even if extrapolation is involved.

$\frac{1}{\sqrt{y}}$ vs x | linear

means y vs x must have been curved

A 11. A company's sales increase by the same amount each year. This growth is…

A) linear B) exponential C) logarithmic D) power E) quadratic

sales | ... | yr

B 12. A company's sales increase by the same percent each year. This growth is…

A) linear B) exponential C) logarithmic D) power E) quadratic

sales | ... | year

E 13. It's easy to measure the circumference of a tree's trunk, but not so easy to measure its height. Foresters developed a model for ponderosa pines that they use to predict the tree's height (in feet) from the circumference of its trunk (in inches): $\ln \hat{h} = -1.2 + 1.4(\ln C)$. A lumberjack finds a tree with a circumference of 60"; how tall does this model estimate the tree to be?

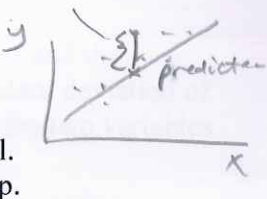
A) 5' B) 11' C) 19' D) 83' E) 93'

($\ln = \log_e$) $\ln(\hat{h}) = -1.2 + 1.4\ln(60)$, $\ln(\hat{h}) = 4.532$, $\log_e(\hat{h}) = 4.532$, $\hat{h} = e^{4.532} = 92.95$

B 14. All but one of these statements contains an error. Which statement could be true?

X A) The correlation between a football player's weight and the position he plays is 0.54. (position is categorical)
✓ B) The correlation between the amount of fertilizer used and yield of beans is 0.42.
X C) The correlation between a car's length and its fuel efficiency is 0.71 miles per gallon. r has no units
X D) There is a high correlation (1.09) between height of a corn stalk and its age in weeks. $-1 \le r \le 1$
X E) There is a correlation of 0.63 between gender and political party. both are categorical

resid = $y_{actual} - y_{pred.}$



C 15. Residuals are…

A) possible models not explored by the researcher.
B) variation in the data that is explained by the model.
C) the difference between observed responses and values predicted by the model.
D) data collected from individuals that is not consistent with the rest of the group.
E) none of these.

B 16. Which statement about influential points is true?

✓ I. Removal of an influential point changes the regression line.
✓ II. Data points that are outliers in the horizontal direction are more likely to be influential on slope than points that are outliers in the vertical direction.
× III. Points that are influential on slope have large residuals. (if leverage is high enough they pull line to themselves)

A) I only B) I and II C) I and III D) II and III E) I, II, and III

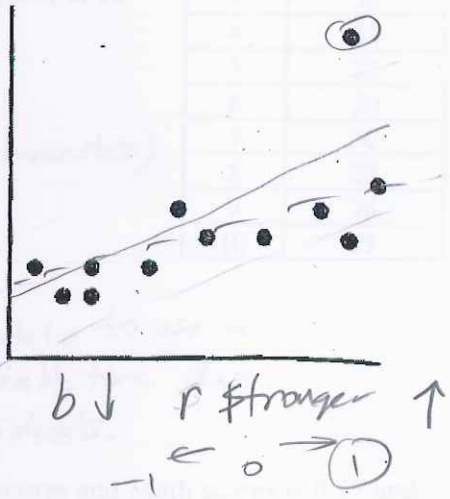
A 17. Which is true?

× I. Random scatter in the residuals indicates a model with high predictive power. can have good linear model w/ high or low r^2
× II. If two variables are very strongly associated, then the correlation between them will be near +1.0 or -1.0. only if linearly associated
× III. The higher the correlation between two variables the more likely the association is based in cause and effect. no, this requires experiment design

A) none B) I only C) II only D) I and II only E) I, II, and III

D 18. If the point in the upper right corner of this scatterplot is removed from the data set, then what will happen to the slope of the line of best fit (*b*) and to the correlation (*r*)?

A) both will increase.
B) both will decrease.
C) *b* will increase, and *r* will decrease.
D) *b* will decrease, and *r* will increase.
E) both will remain the same.



19. **Earning power** – A college's job placement office collected data about students' GPAs and the salaries they earned in their first jobs after graduation. The mean GPA was 2.9 with a standard deviation of 0.4. Starting salaries had a mean of $47200 with a SD of $8500. The correlation between the two variables was $r = 0.72$. The association appeared to be linear in the scatterplot. *(Show all work)*

a. Write an equation of the model that can predict salary based on GPA: $\hat{y} = a + bx$ need the y x

$b = r\frac{s_y}{s_x} = .72\frac{(8500)}{(0.4)} = 15300 \rightarrow \hat{y} = a + 15300x$ still need

$(\bar{x}, \bar{y})$ is on LSRL: $(47200) = a + 15300(2.9)$

$a = 2830$ so: $\hat{y} = 2830 + 15300x$

(2.9, 47200)

answer: $\hat{y} = 2830 + 15300x$ where x = GPA, y = Salary ($)

b. Do you think these predictions will be reliable? Explain.

$r = .72$

$r^2 = (.72)^2 = .52$

52% of the variation in salary is explained by the LSRL relating salary to GPA. This is a medium-reliable model.

c. Your brother just graduated from that college with a GPA of 3.30. He tells you that based on this model the residual for his pay is -$1880. What salary is he earning?

$\hat{\text{Salary}} = 2830 + 15300(3.3) = 53320$

resid = salary − $\hat{\text{salary}}$

$-1880 = \text{salary} - 53320$

Salary = $51440

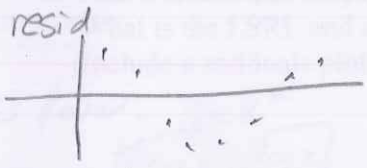
20. **Assembly line** – Your new job at *Panasony* is to do the final assembly of an electronic product. As you learn how, you get faster. The company tells you that you will qualify for a raise if after 13 weeks your assembly time averages under 20 minutes. The data shows your average assembly time during each of your first 10 weeks.

Week	Time(min)
1	43
2	39
3	35
4	33
5	32
6	30
7	30
8	28
9	26
10	25

a. Which is the explanatory variable? Week

b. What is the correlation between these variables $r = -.97$ (1-var stats)

c. You want to predict whether or not you will qualify for that raise. Would it be appropriate to use a linear model? Explain.



It would not be appropriate to use a linear model to predict assembly time due to the pattern in the residuals.

21. **Math and Verbal** – Suppose the correlation between SAT Verbal scores and Math scores is 0.57 and that these scores are normally distributed. If a student's Verbal score places her at the 90th percentile, at what percentile would you predict her Math score to be? *(Show work)*

(1) $r = 0.57$

$b = r\frac{s_y}{s_x}$

$s_y = s_x = 1$ for z-scores

$b = r\frac{(1)}{(1)} = r$

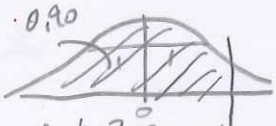
and LSRL goes through (0,0)

so $z_y = rz_x$

$(y = a + bx)$

or r

(2) 0.90

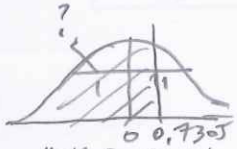


$z_x = \text{invNorm}(0.90, 0, 1)$ area to left, mean, SD

$z_x = 1.2816$

so $z_y = (0.57)1.2816 = 0.7305$

(3)



normalcdf(−999, 0.7305, 0, 1) lower, upper, mean, SD

$= .767$

$\approx 77^{th}$ percentile

22. **Gas mileage** – An important factor in the amount of gasoline a car uses is the size of the engine. Called "displacement", engine size measures the volume of the cylinders in cubic inches. A regression analysis on data collected for a representative sample of cars is shown.

Dependent variable is: MPG
89 total cases of which 0 are missing
R squared = 60.9% *R squared (adjusted) = 60.0%*
s = 3.056 with 89 − 2 = 87 degrees of freedom

Source	Sum of Squares	df	Mean Square	F-ratio
Regression	696.744	1	696.744	74.6
Residual	448.236	48	9.33826	

Variable	Coefficient	s.e. of Coeff	t-ratio	prob
Constant	a = 34.9799	1.231	28.4	≤ 0.0001
Eng. Displcmt	b = -0.066196	0.0077	-8.64	≤ 0.0001

a. How many cars were included in this analysis? 89

b. What is the correlation between engine size and fuel economy? −.780 $r^2 = .609$
$r = \pm\sqrt{.609} = \pm .780$
$= -.780$ (bc negative)

c. Write the LSRL found by this analysis:
$\hat{y} = 34.9799 - 0.066196x$ where x: engine displacement (in^3)
y: fuel economy (MPG)

d. A car you are thinking of buying is available with two different size engines, 190 cubic inches or 240 cubic inches. How much difference might this make in your gas mileage? *(Show work)*
$\hat{mpg} = 34.9799 - 0.066196(190) = 22.40266$ mpg
$\hat{mpg} = 34.9799 - 0.066196(240) = 19.09286$ mpg
3.3 mpg

23. **Breaking strength** – A company manufactures polypropylene rope in six different sizes. To assess the strength of the ropes they test two samples of each size to see how much force (in kilograms) the ropes will hold without breaking. The table shows the results of the tests. We want to create a model for predicting the breaking strength from the diameter of the rope.

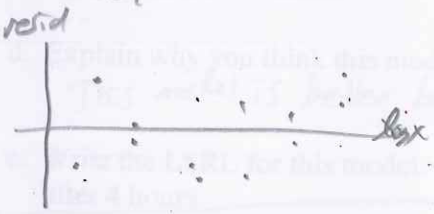
Diameter (mm)	Strength (kg)	
4	60	76
7	157	153
10	254	262
12	334	388
15	551	529
20	938	893

x	y
4	60
4	76
7	157
7	153
10	254
10	262

a. Find a model that uses re-expressed data to straighten the scatterplot. What is the LSRL and coefficient of determination for your model? (Include a residuals plot as evidence that your model straightens the data)

best is Power: $y = x^k$
$\log y = k \log x$
L4 L3

$\log(\hat{strength}) = 0.8367 + 1.609 \log(diam)$
$r^2 = .9913$
(coefficient of determination)



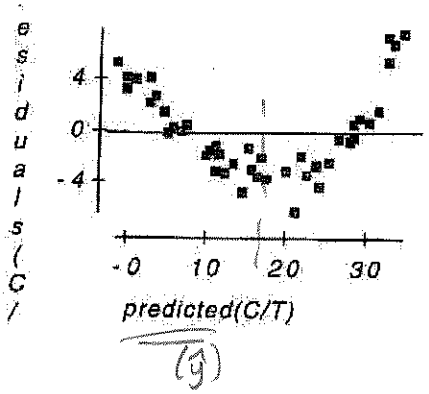
b. The company is thinking of introducing a new 25mm diameter rope. How strong should it be?
$\log(\hat{strength}) = 0.8367 + 1.609 \log(25)$
$\log(\hat{strength}) = 3.0864$
$\hat{strength} = 10^{3.0864} =$ 1220 kg

24. **Penicillin** – Doctors studying how the human body assimilates medication inject some patients with penicillin, and then monitor the concentration of the drug (in units/cc) in the patients' blood for seven hours. The data are shown in the scatterplot. First they tried to fit a linear model to the original data. The regression analysis and residuals plots are shown.

Dependent variable is: **Concentration**
No Selector
R squared = 90.8% *R squared (adjusted) = 90.6%*
s = 3.472 with 43 - 2 = 41 degrees of freedom

Source	Sum of Squares	df	Mean Square	F-ratio
Regression	4900.55	1	4900.55	407
Residual	494.199	41	12.0536	

Variable	Coefficient	s.e. of Coeff	t-ratio	prob
Constant	a=40.3266	1.295	31.1	≤ 0.0001
Time	b=-5.95956	0.2956	-20.2	≤ 0.0001



a. Find the correlation between time and concentration.

$r^2 = .908 \Rightarrow r = \pm\sqrt{.908} = \pm .95289$, b is negative, so $r = -.9529$

b. Write the LSRL for this model. Then, using this model, estimate what the concentration of pencillin will be after 4 hours.

$\hat{y} = 40.3266 - 5.95956x$
x: time (hrs)
y: Concentration (units/cc)

$\hat{y} = 40.3266 - 5.95956(4)$
$\hat{y} =$ 16.488 units/cc

c. Is that estimate likely to be accurate, too low, or too high? Explain.

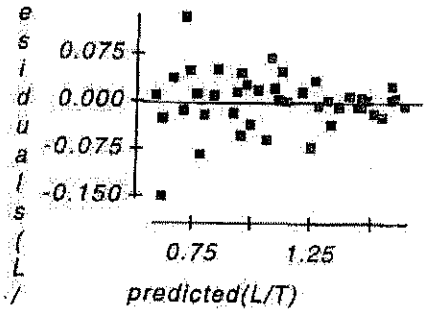
In residuals plot, for $\hat{y} \approx 17$, the residuals are negative, meaning actual is lower than predicted, so this estimate is likely to be too high.

Now the researchers try a new model, using the re-expression log(*Concentration*). Examine the regression analysis and the residuals plot below.

Dependent variable is: **LogCnn**
No Selector
R squared = 98.0% *R squared (adjusted) = 98.0%*
s = 0.0451 with 43 - 2 = 41 degrees of freedom

Source	Sum of Squares	df	Mean Square	F-ratio
Regression	4.11395	1	4.11395	2022
Residual	0.083412	41	0.002034	

Variable	Coefficient	s.e. of Coeff	t-ratio	prob
Constant	a=1.80184	0.0168	107	≤ 0.0001
Time	b=-0.172672	0.0038	-45.0	≤ 0.0001



d. Explain why you think this model is better than the original linear model.

This model is better because there is less pattern in the residuals.

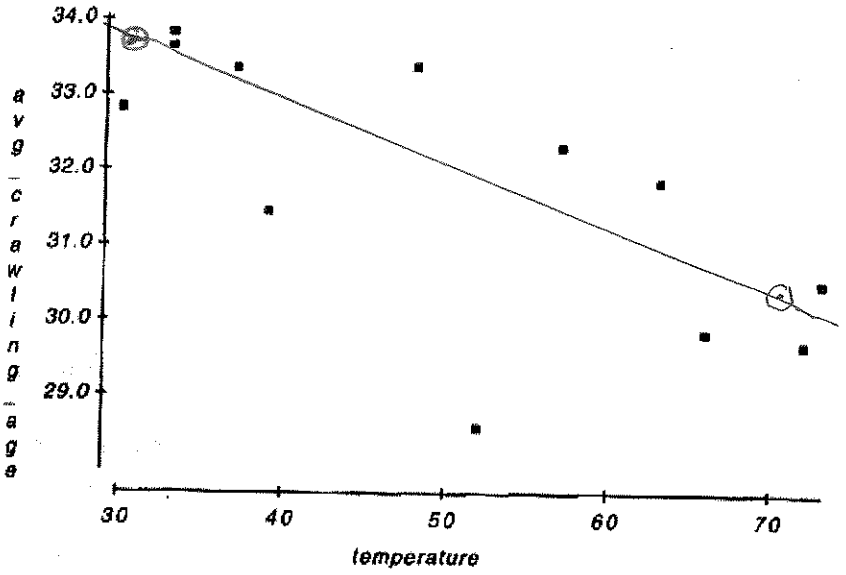
e. Write the LSRL for this model. Then, using this new model, estimate the concentration of penicillin after 4 hours.

$\log(\hat{y}) = 1.80184 - 0.172672x$
x: time (hrs)
y: concentration (units/cc)

$\log(\hat{y}) = 1.80184 - 0.172672(4)$
$\log(\hat{y}) = 1.111152$
$\hat{y} = 10^{1.111152} =$ 12.9 units/cc

25. **Crawling** – Researchers at the University of Denver Infant Study Center investigated whether babies take longer to learn to crawl in cold months (when they are often bundled in clothes that restrict their movement) than in warmer months. The study sought an association between babies' first crawling age (in weeks) and the average temperature during the month they first try to crawl (about 6 months after birth). Between 1988 and 1991 parents reported the birth month and age at which their child was first able to creep or crawl a distance of four feet in one minute. Data were collected on 208 boys and 206 girls. The graph below plots average crawling ages (in weeks) against the mean temperatures when the babies were 6 months old. The researchers found a correlation of $r = -0.70$ and their line of best fit was

$$\hat{AvAge} = 36 - 0.08 AvTemp.$$



a. Draw the line of best fit on the graph (show your method clearly). find two predicted pairs:
$\hat{age} = 36 - .08(30) = 33.6$ (30, 33.6) $\hat{age} = 36 - .08(70) = 30.4$ (70, 30.4)

b. Describe the association in contest.

There is a medium-strong, negative, fairly linear association between average crawling age and temperature.

c. Explain (in context) what the slope of the line means.

$b = -.08 \frac{\text{weeks}}{°F}$ For every 1 additional °F in temperature, the average crawling age for babies decreases by 0.08 weeks, on average.

d. Explain (in context) what the y-intercept of the line means.

At a temperature of 0°F, average crawling age for babies is predicted to be 36 weeks.

e. Explain (in context) what r^2 means.

$r^2 = (-0.70)^2 = .49$ About 49% of the variation in avg crawling age of babies is explained by the LSRL model which relates crawling age to temperature.

f. In this context, what does a negative residual indicate?

This particular baby is crawling earlier than the LSRL model predicts for that temperature.

26. **Music and grades** – A couple of years ago a local newspaper published research results claiming a positive association between the number of years high school children had taken instrumental music lessons and their performances in school (GPA).

a. What does "positive association" mean in this context?

The more years students take music lessons, the higher their GPAs, on average.

b. A group of parents then went to the School Board demanding more funding for music programs as a way to improve student chances for academic success in high school. As a statistician, do you agree or disagree with their reasoning? Explain.

I disagree. Association does not imply causation.

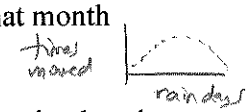
27. **Associations** – for each pair of variables, indicate whether you expect the association to be positive, negative, curved, or no association:

a. Power level setting of a microwave vs. Number of minutes it takes to boil water

negative

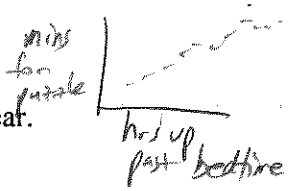
b. Number of days it rained in a month vs. Number of times you mowed your lawn that month

curved



c. Number of hours a person has been up pass a normal bedtime vs. Number of minutes it takes the person to do a crossword puzzle.

postive



d. Length of a student's hair vs. Number of credits the student earned last year.

No association