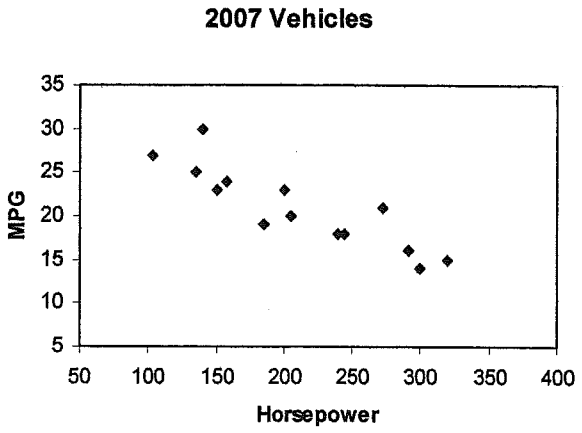
AP Statistics
Linear Regression - Review Problems

1) Here are advertised horsepower ratings and expected gas mileage for several 2007 vehicles.

Audi A4	200 hp	23 mpg
Buick LeSabre	205	20
Chevy TrailBlazer	291	16
Chevy Aveo	103	27
Ford Excursion	300	14
GMC Yukon	320	15
Honda Civic	140	30
Hyundai Elantra	135	25
Lexus 350	272	21
Lincoln Town Car	239	18
Mazda CX-7	244	18
Olds Silhouette	185	19
Toyota Camry	158	24
VW Beetle	150	23



(a) Describe the association in context.

(b) Find the equation of the least squares regression.

(c) Draw the line of best fit on the graph.

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

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

(f) Explain the meaning of R^2 in this context.

(g) Do you think a linear model is appropriate?

(h) In general, what would a positive residual mean in this context?

2) Tests were conducted on 11 fast food chicken sandwiches. It was found that the mean fat content was 20.6 grams with a standard deviation of 9.8 grams, and the mean number of calories was 472.7 with a standard deviation of 144.2. A scatterplot showed the association to be reasonably linear, and the correlation between fat and calories was 0.947.

(a) Write the equation a linear model that estimates the number of calories.

(b) Do you think predictions made with this model will be reliable?

(c) What does it mean if a certain sandwich has a negative residual?

(d) How many calories would you predict are in a chicken sandwich with 35 grams of fat?

3)

(a) What is the response variable?

_____

(b) How many cars is this analysis based on?

_____

(c) What is the correlation between weight and MPG?

_____

Dependent variable is: MPG
No Selector
R squared = 75.6% R squared (adjusted) = 75.1%
s = 2.413 with 50 - 2 = 48 degrees of freedom

Source	Sum of Squares	df	Mean Square	F-ratio
Regression	865.410	1	865.410	149
Residual	279.570	48	5.82436	

Variable	Coefficient	s.e. of Coeff	t-ratio	prob
Constant	48.7393	1.976	24.7	Š 0.0001
Weight	-8.21362e-3	0.0007	-12.2	Š 0.0001

(d) What is the model – the equation of the line of best fit? _____

(e) How strong is the model? Explain.

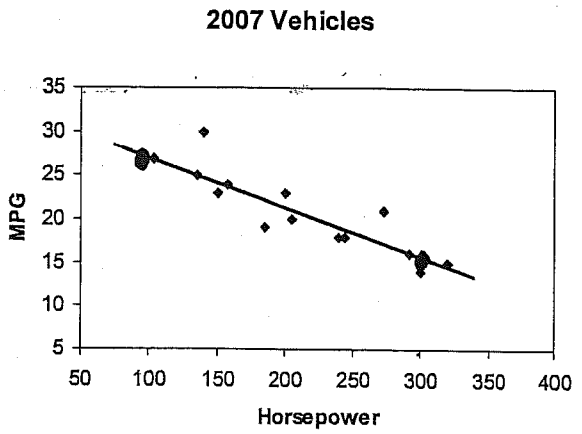
(f) A 1992 Geo Prizm weighed 2608 pounds. Use your model to estimate how many miles per gallon it should get.

(g) The owner actually averaged about 33.7 mpg. What is the residual?

AP Statistics
Linear Regression - Review Problems

1) Here are advertised horsepower ratings and expected gas mileage for several 2007 vehicles.

Audi A4	200 hp	23 mpg
Buick LeSabre	205	20
Chevy TrailBlazer	291	16
Chevy Aveo	103	27
Ford Excursion	300	14
GMC Yukon	320	15
Honda Civic	140	30
Hyundai Elantra	135	25
Lexus 350	272	21
Lincoln Town Car	239	18
Mazda CX-7	244	18
Olds Silhouette	185	19
Toyota Camry	158	24
VW Beetle	150	23



(a) Describe the association in context.
There is a negative, linear association between horsepower and MPG. The relationship is fairly strong, with a correlation of −0.885. In general, the higher the horsepower, the lower the MPG. The Honda Civic seems unusual with a high MPG for its horsepower.

(b) Find the equation of the least squares regression.

$\hat{MPG} = 33.534 - 0.05998(HP)$

(c) Draw the line of best fit on the graph.

$\hat{MPG} = 33.534 - 0.05998(100) = 27.536$

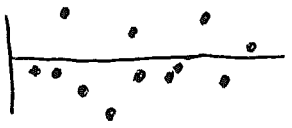
$\hat{MPG} = 33.534 - 0.05998(300) = 15.54$

(d) Explain what the slope of the line means in this context.
The model predicts that an increase of 100 in horsepower is associated with a decrease of about 6 mpg.

(e) Explain what the y-intercept of the line means in this context.
According to this model, a car with 0 horsepower would get about 33.5 mpg.

(f) Explain the meaning of R^2 in this context.
78.4% of the variability in MPG is explained by this model.

(g) Do you think a linear model is appropriate?
Yes, a linear model is appropriate because there is no pattern in the residuals plot.



(h) In general, what would a positive residual mean in this context?
A car gets a higher MPG than the model predicts.

2) Tests were conducted on 11 fast food chicken sandwiches. It was found that the mean fat content was 20.6 grams with a standard deviation of 9.8 grams, and the mean number of calories was 472.7 with a standard deviation of 144.2. A scatterplot showed the association to be reasonably linear, and the correlation between fat and calories was 0.947.

x = fat
y = calories

(a) Write the equation a linear model that estimates the number of calories.

$$b_1 = r\frac{S_y}{S_x} = \frac{0.947(144.2)}{9.8} = 13.934$$

$$b_0 = \bar{y} - b_1\bar{x} = 472.7 - 13.934(20.6) = 185.66$$

$$\hat{\text{calories}} = 185.66 + 13.934\,(\text{fat})$$

(b) Do you think predictions made with this model will be reliable?

Predictions will be relatively reliable because 89.7% of the variability in the # of calories is explained by this model.

(c) What does it mean if a certain sandwich has a negative residual?

The actual # of calories in the sandwich is less than the model predicts.

(d) How many calories would you predict are in a chicken sandwich with 35 grams of fat?

$$\hat{\text{calories}} = 185.66 + 13.934(35) = 673.35$$

The model estimates 673 calories in a chicken sandwich with 35 grams of fat.

3)

Dependent variable is: MPG
No Selector
R squared = 75.6% R squared (adjusted) = 75.1%
s = 2.413 with 50 - 2 = 48 degrees of freedom

Source	Sum of Squares	df	Mean Square	F-ratio
Regression	865.410	1	865.410	149
Residual	279.570	48	5.82436	

Variable	Coefficient	s.e. of Coeff	t-ratio	prob
Constant	48.7393	1.976	24.7	≤ 0.0001
Weight	-8.21362e-3	0.0007	-12.2	≤ 0.0001

(a) What is the response variable?

MPG

(b) How many cars is this analysis based on?

50

(c) What is the correlation between weight and MPG?

-0.869

(d) What is the model – the equation of the line of best fit?

$$\hat{MPG} = 48.7393 - 0.008214(wt)$$

(e) How strong is the model? Explain.

The model is moderately strong. 75.6% of the variability in MPG is explained by this model.

(f) A 1992 Geo Prizm weighed 2608 pounds. Use your model to estimate how many miles per gallon it should get.

$$\hat{MPG} = 48.7393 - 0.008214(2608) = 27.317$$

The model estimates about 27 mpg for the '92 Prizm.

(g) The owner actually averaged about 33.7 mpg. What is the residual?

residual = actual − predicted
= 33.7 − 27.3
= 6.4 mpg

AP Statistics
Linear Regression - Review Problems 2

1) The association between a family's weekly income and the amount they spend on restaurant meals is found to be linear with $r = 0.40$.

(a) **Answer true or false:**

___ Families tend to spend about 40% of their incomes on restaurant meals.

___ The line of best fit passes through 40% of the data points.

___ When income increases by $100 per month, families tend to spend an additional $40 on restaurant meals.

___ If a family's income increases, then they will spend more in restaurants.

___ In general, families with higher incomes spend more in restaurants.

___ 40% of the variability in amount spent in restaurants is explained by differences in family income.

(b) If a family's weekly income is 2 standard deviations above the average family weekly income, the predicted amount this family spends on restaurant meals is _____ standard deviations _____ the mean.

(c) If a family's weekly income is 1.5 standard deviations below the average family weekly income, the predicted amount this family spends on restaurant meals is _____ standard deviations _____ the mean.

(d) What are the units for the correlation coefficient?

(e) If we change "weekly income" to "yearly income" (by multiplying by 52 weeks/year), how will the correlation change?

2) A regression analysis of company profits and the amount of money the company spent on advertising found $R^2 = 0.85$.

Answer true or false:

___ This model can correctly predict the profit for 85% of companies.

___ On average, companies spend about 85% of their profits on advertising.

___ Advertising spending explains about 85% of profits.

___ Differences in advertising spending explain about 85% of the variability in profits.

___ The R^2 value of 85% shows that this linear model is appropriate.

___ The R^2 value of 85% shows that this linear model will be fairly reliable.

3) A consumer group creates a linear model to estimate the cost of a standard TV (in $) based on the screen size (in inches). Which is the most likely value of the slope of the line of best fit?

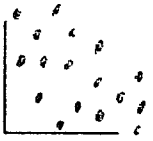
(A) 0.15 (B) 1.5 (C) 15 (D) 150 (E) 1500

4) 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 (in thousands of dollars) is 6.25, and the correlation between the variables is 0.48. Then the slope of the line that predicts family income (in thousands of dollars) from SAT score is

(A) 0.037 (B) 0.16 (C) 3.00 (D) 6.25 (E) 13.02

5) Studies have shown that nations with a higher number of television sets per person have higher life expectancies. Does this mean that we can lengthen the lives of people in Botswana by shipping them TV sets? Explain.

6) Match each of the following scatterplots with its correlation:

(A)  (B)  (C)  (D)  (E) 

___ $r = -0.3$

___ $r = 0.5$

___ $r = -0.7$

___ $r = 0.9$

___ $r = 0 - 0.99$

AP Statistics
Linear Regression - Review Problems 2

1) The association between a family's weekly income and the amount they spend on restaurant meals is found to be linear with $r = 0.40$.

(a) **Answer true or false:**

F Families tend to spend about 40% of their incomes on restaurant meals.

F The line of best fit passes through 40% of the data points.

F When income increases by \$100 per month, families tend to spend an additional \$40 on restaurant meals. (This would be true if slope was 0.40.)

F If a family's income increases, then they will spend more in restaurants. (not necessarily)

T In general, families with higher incomes spend more in restaurants.

F 40% of the variability in amount spent in restaurants is explained by differences in family income. (This would be true if R^2 was 0.40.)

(b) If a family's weekly income is 2 standard deviations above the average family weekly income, the predicted amount this family spends on restaurant meals is 0.8 standard deviations above the mean.

$\hat{z}_y = r \cdot z_x = (0.4)(2) = 0.8$

(c) If a family's weekly income is 1.5 standard deviations below the average family weekly income, the predicted amount this family spends on restaurant meals is 0.6 standard deviations below the mean.

$\hat{z}_y = r \cdot z_x = (0.4)(-1.5) = -0.6$

(d) What are the units for the correlation coefficient?

Correlation coefficients are unitless.

(e) If we change "weekly income" to "yearly income" (by multiplying by 52 weeks/year), how will the correlation change?

It won't change. Correlation is not affected by changes in the center or scale of either variable.

2) A regression analysis of company profits and the amount of money the company spent on advertising found $R^2 = 0.85$.

Answer true or false:

F This model can correctly predict the profit for 85% of companies.

F On average, companies spend about 85% of their profits on advertising.

F Advertising spending explains about 85% of profits.

T Differences in advertising spending explain about 85% of the variability in profits.

F The R^2 value of 85% shows that this linear model is appropriate. (use a residuals plot)

T The R^2 value of 85% shows that this linear model will be fairly reliable.

3) A consumer group creates a linear model to estimate the cost of a standard TV (in $) based on the screen size (in inches). Which is the most likely value of the slope of the line of best fit?

(A) 0.15 (B) 1.5 (C) 15 (D) 150 (E) 1500

If you increase the screen size by 10 inches, the cost is most likely to increase by $150. $\frac{150}{10} = 15$

4) 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 (in thousands of dollars) is 6.25, and the correlation between the variables is 0.48. Then the slope of the line that predicts family income (in thousands of dollars) from SAT score is

(A) 0.037 (B) 0.16 (C) 3.00 (D) 6.25 (E) 13.02

$b_1 = r\frac{S_y}{S_x}$

$6.25 = (0.48)\frac{S_{SAT}}{S_{inc.}}$

$\frac{S_{SAT}}{S_{inc.}} = 13.020833$

$\frac{S_{inc.}}{S_{SAT}} = \frac{1}{13.020833} = 0.0768$

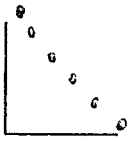
new slope: $b_1 = r\frac{S_{inc.}}{S_{SAT}}$

$= (0.48)(0.0768) = \boxed{0.0369}$

5) Studies have shown that nations with a higher number of television sets per person have higher life expectancies. Does this mean that we can lengthen the lives of people in Botswana by shipping them TV sets? Explain.

Shipping TV's to Botswana will probably not lengthen the people's lives. Just because 2 variables are associated does not mean that one causes the other to happen. There may be other factors, such as a lurking variable of national wealth.

6) Match each of the following scatterplots with its correlation:

(A)  (B)  (C)  (D)  (E) 

C $r = -0.3$

D $r = 0.5$

B $r = -0.7$

A $r = 0.9$

E $r = 0 - 0.99$