Question 1 of 16
A researcher estimates the relationship between height and weight. He chooses height (in cm) as the explanatory variable and weight (in kg) as the response variable. He finds that in the linear relationship \( y = a + bx \), \( a = -85.45 \) and \( b = 0.958 \). This implies that an additional 10 cm of height is associated with an additional ...

A) 8.545 kg
B) 85.45 kg
C) 95.8 kg
D) 958 g
E) 9.58 kg

Points: 0 out of 1

\( b = 0.958 \) hence a 1 cm increase in height is associated with a 0.958 kg increase in height. Thus a 10 cm increase in height is associated with a 9.58 kg increase in height.

Question 2 of 16
A researcher estimates the relationship between height and weight. He chooses height (in cm) as the explanatory variable and weight (in kg) as the response variable. He finds that in the linear relationship \( y = a + bx \), \( a = -85.45 \) and \( b = 0.958 \). The standard deviation of height is 22.3893 cm and the standard deviation of weight is 29.8945 kg. Therefore the correlation between height and weight is:

A) 0.717
B) -0.717
C) 0.516
D) -0.516
E) 0.233
F) -0.233
G) 0.987

Points: 1 out of 1

The correlation between height and weight is \((s_x/s_y)b\) which is \((22.3893/29.8945) \times 0.958 = 0.717\)
Question 3 of 16
A researcher estimates the relationship between height and weight. He chooses height (in cm) as the explanatory variable and weight (in kg) as the response variable. The standard deviation of the predicted weight is 21.45 kg and the standard deviation of weight is 29.89 kg. Therefore the r squared is

A) 51%
B) 45%
C) 23%
D) 98%

Points: 1 out of 1

Two approaches: the r squared is the ratio of the variance of the predicted weight to the actual weight, hence the square of the ratio of the standard deviation of the predicted weight to the actual weight. That gives 0.51, or equivalently 51%. The second approach is to use the result from the previous question: the square of the correlation between y and x gives 0.51 or 51%.

Question 4 of 16
For the following variables in a regression analysis, which variable more naturally plays the role of the explanatory variable and which "more naturally" plays the role of the response variable?

College GPA or High school GPA

A) College GPA as y, High school GPA as x
B) High school GPA as y, College GPA as x

Points: 1 out of 1

The high school GPA is measured before the college GPA is, hence it is hard to imagine that the college GPA would cause a rise in the high school GPA. The College GPA is thus the natural response variable.

Question 5 of 16
The OECD (Organization for Economic Cooperation and Development) consists of 20 advanced, industrialized countries. For these nations, the prediction equation relating child poverty rate in 2000 (in percentage points) to social expenditure as a percent of gross domestic product (in percentage points) is \( y = 22 - 1.3x \). The y values ranged from 2.8% (Finland) to 21.9% (U.S.). The x-values ranged from 2% (U.S.) to 16% (Denmark).

Find the predicted poverty rates for the U.S. and for Denmark.

A) U.S: 19.4% ; Denmark : 1.2%
B)
U.S: 14.9% ; Denmark: 2.1%

C)
U.S: 14.9% ; Denmark: 1.2%

D)
U.S: 14.9% ; Denmark: 1.0%

Points: 1 out of 1

Plug in the x values for the US and for Denmark in the linear relationship \( y = 22 - 1.3 \times x \) to find that for the US the predicted value is \( 22 - 1.3 \times 2 = 19.4 \) percent, and for Denmark the predicted value is \( y = 22 - 1.3 \times 16 = 2.1 \) percent. Note the importance of the assumption that x is measured in percentage points.

Question 6 of 16
A college admissions officer uses regression to approximate the relationship between \( y = \) college GPA and \( x = \) high school GPA (both measured on a four-point scale) for students at that college. Which equation is more realistic? Which equation is more realistic:

A) \( y = 0.5 + 7.0 \times x \)

✓ B) \( y = 0.5 + 0.7 \times x \)

C) \( y = 5 + 7 \times x \)

D) \( y = 5 + 0.7 \times x \)

Points: 1 out of 1

The GPAs are measured on a four point scale. Hence it is likely that the predicted values y should be in that four point scale. Since \( 0.5 + 0.7 = 1.2 \) and \( 0.5 + 0.7 \times 4 = 3.3 \), the predicted values of y range between 1.2 and 3.3. Notice that neither A, C, D would give predicted values in this 1-4 range.

Question 7 of 16
The figure below is a scatterplot relating \( y = \) percent of people using cell phones and \( x = \) per capita gross domestic product (GDP) for nations listed in the Human Development Report.

The least squares prediction equation is \( y = -0.13 + 2.62 \times x \).

Is the correlation positive, or negative?
A) Impossible to know

B) Positive

C) Negative

Points: 1 out of 1

From the formula that $r(x,y) = \frac{s_x}{s_y} b$, and given that standard deviations are always positive, the correlation $r(x,y)$ is of the same sign as $b$. $b$ is positive, hence $r$ is positive.

**Question 8 of 16**

The figure below is a scatterplot relating $y =$ percent of people using cell phones and $x =$ per capita gross domestic product (GDP) for nations listed in the Human Development Report.

The least squares prediction equation is $y = -0.13 + 2.62 x$. 
For the U.S., \( x = 34.3 \) and \( y = 45.1 \). Find the predicted cell-phone use and the residual.

A) Predicted cell phone use : 89.7 Residual : 44.6

B) Predicted cell phone use : 89.7 Residual : 44.6

C) Predicted cell phone use : -89.7 Residual : -44.6

D) Predicted cell phone use : 98.1 Residual : 41.6

E) Predicted cell phone use : 41.6 Residual : 89.7

Points: 1 out of 1

For the US \( x=34.3 \). Plugging in \( x \) in this linear relationship, we find that the predicted value \( y = -0.13 + 2.62 \times 34.3 = 89.736 \). The residual (also known as error) is the difference between \( y \) and the predicted value of \( y = -44.6 \).

**Question 9 of 16**

For two given variables, one the explanatory variable (\( x \)) the other one the response variable (\( y \)). The slope of the regression line and the correlation:

A) always have the same sign
The correlation \( r(x,y) = \left( \frac{s_x}{s_y} \right) \times b \) and standard deviations are positive hence \( r \) has the same sign as \( b \).

**Question 10 of 16**

A high school student analyzes whether a relationship exists between \( x = \) number of books read for pleasure in the previous year and \( y = \) daily average number of hours spent watching television. For her three best friends, the table below shows the observations.

<table>
<thead>
<tr>
<th>x</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
</tr>
</tbody>
</table>

Choose the right correlation in the following list:

- **A)** 0
- **B)** 1 ✔
- **C)** -1
- **D)** 0.5
- **E)** -0.5
- **F)** 2
- **G)** 0.2
- **H)** -0.2

**Points: 1 out of 1**

Notice that \( y \) is exactly equal to \( 5 - 2/5 \times x \). There is an exact linear relationship tying \( y \) and \( x \), and the sign of the slope is negative, hence \( r(x,y) = -1 \) (see slides defining the correlation).

**Question 11 of 16**
The prediction equation relating \( x = \) years of education and \( y = \) annual income (in dollars) is \( y = -20,000 + 4,000x \), and the correlation equals 0.50. The standard deviations were 2.0 for \( x \) and 16,000 for \( y \).

Results were translated to units of euros, at a time when the exchange rate was $1.5 per euro. Find the prediction equation.

Correct answer: A)
Selected answer: A)

A) \( y = -13,333 + 2,666 x \)

B) \( y = -30,000 + 6,000 x \)

C) \( y = -20,000 + 4,000 x \), unchanged

D) \( y = -20,000 + 6,000 x \)

Points: 1 out of 1

Notice that \( y = -20,000 + 4,000 x \) gives the annual income in dollars given \( x \) in years of education. To convert any quantity from dollars to euros at the exchange rate of $1.5 per euro, divide a quantity in dollars by 1.5. Hence the annual income in euros given \( x \) is \( ( -20,000 + 4,000 x )/1.5 = -13,333 + 2,666 x \). (I have ignored digits after the decimal).

**Question 12 of 16**

The variables \( y = \) annual income (thousands of dollars), \( x_1 = \) number of years of education, and \( x_2 = \) number of years experience in job are measured for all the employees having city-funded jobs, in Knoxville, Tennessee. The following prediction equations and correlations apply.

(i.) \( y = 10 + 1.0 \times x_1, r = 0.30 \)

(ii.) \( y = 14 + 0.4 \times x_2, r = 0.60 \)

The correlation is -0.40 between \( x_1 \) and \( x_2 \). Which of the following statements are true?

Tick all that apply. (points added for right answers, deducted for wrong answers)

A) The strongest sample association is between \( y \) and \( x_2 \).

B) The weakest sample association is between \( x_1 \) and \( x_2 \).

C) The prediction equation using \( x_2 \) to predict \( x_1 \) has negative slope.

D) A standard deviation increase in education corresponds to a predicted increase of 0.3 standard deviations in income.

E) The predicted mean income for employees having 20 years of experience is $4000 higher than the predicted mean income for employees having 10 years of experience.
If $s_x = 8$ for the model using $x_1$ to predict $y$ (only for this answer), then it is not unusual to observe an income of $70,000 for an employee who has 10 years of education. Assume a bell shaped distribution for $y$.

If $s_x = 8$ for the model using $x_1$ to predict $y$ (only for this answer), then it is unusual to observe an income of $50,000 for an employee who has 10 years of education. Assume a bell shaped distribution for $y$.

**A**: yes, the strongest sample association is between $y$ and $x_2$ as the correlation between $y$ and $x_2$ is the highest.

**B**: the weakest sample association is between $y$ and $x_1$, as it is the absolute value of the correlation that measures how strongly two variables are associated.

**C**: regressing $x_1$ on $x_2$ would lead to a slope with the same sign as the correlation between the two variables.

**D**: a standard deviation increase in education corresponds to a $s \times b / s_y$ predicted increase in standard deviations of income, hence the correlation $0.3$.

**E**: an increase in experience by $20 - 10 = 10$ years raises income by $0.4 \times 10 = 4$ thousand dollars.

**F**: see slides for session 8 (coming up). The standard deviation of the predicted values of $y$ is equal to the standard deviation of $x$ (the slope of the linear relationship is 1), but the distribution of the actual income $y$ has a larger standard deviation than $x$, since the $R^2$ squared of the regression is not 1 $(r < 1)$. We find that $s_y = (s_x / r) b$. Assuming the distribution of $y$ is bell shaped, then almost all observations fall between the mean of $y$ given $x$ ($10 + 1.0 \times 10 = 20$) and $-3 s_y$.

**Question 13 of 16**

The correlation is **inappropriate** as a measure of association between two quantitative variables.

Tick all that apply (points deducted for wrong answers).

**A)** When different people measure the variables using different units

**B)** When the data points fall exactly on a straight line

**C)** When $y$ tends to decrease as $x$ increases

**D)** When we have data for the entire population rather than a sample

**E)** When the slope of the prediction equation is 0 using nearly all the data, but a couple of outliers are extremely high on $y$ at the high end of the $x$ scale

**Points**: 1 out of 1

Outliers can severely impact the calculated $b$, while the true $\beta$ is zero. A computation of $b$ excluding outlier observations – judged on a case by case basis – will lead to a $b$ that is closer to the true slope $\beta$.

I have given the point regardless of answer A’s tick or not given the ambiguity here. However, the point made here is important: if say $y$ is measured in dollars for some observations, and $y$ is
measured in euros for other observations, the correlation and the slope are inadequate measures of association.

B is wrong: when the data points fall exactly on a straight line, the correlation is exactly 1 or -1.

C is wrong: of course the correlation can be negative, corresponding to a negative slope b.

D is wrong: of course statistical analysis (including the calculation of the correlation) is performed mostly on samples rather than the whole population. This is why we write b as the calculated slope, which is not equal to $\beta$, the slope of the entire population.

**Question 14 of 16**

A study in 2000 by the National Highway Traffic Safety Administration estimated that 73% of people wear seat belts, that failure to wear seat belts led to 9200 deaths in the previous year, and that that value would decrease by 270 for every 1 percentage point gain in seat belt usage. Let $y =$ predicted number of deaths in a year and $x =$ percentage of people who wear seat belts (from 0 to 100). Find the prediction equation that yields these results.

**A)**
$y = 28,910 - 270x$

**B)**
$y = 29,810 - 260x$

**C)**
$y = 29,810 - 250x$

**D)**
$y = 28,190 - 270x$

**E)**
$y = 82,910 - 270x$

The exercise says that when the $x$ increases by 1 percentage point ($x$ is measured between 0 and 100) the $y$ decreases by 270. Hence the slope is $b = -270$. Now, to find $a$, realize that when $x = 73$, then $y = 9200$. Hence $a - 270$ times 73 = 9200 and $a = 28,910$.

**Question 15 of 16**

A recent study, after pointing out that diets high in fats and sugars (bad for our health) are more affordable than diets high in fruit and vegetables (good for our health), reported, "Every extra 100 g of fats and sweets eaten decreased diet costs by 0.2 Euros, whereas every extra 100 g of fruit and vegetables eaten increased diet costs by 0.3 Euros." Indicate the parameters to which these interpretations refer.

**A)**
The response variable is diet costs ($y$), the explanatory variables are fats and sweets ($x$, in hundreds of grams), and fruits and vegetables ($z$ in hundreds of grams). The linear relationship is $y = 200 - 0.2x + 0.3z$.

**B)**
The explanatory variables are diet costs ($y$), fats and sweets ($x$, in hundreds of grams), and fruits and vegetables ($z$ in hundreds of grams). The linear relationship is $x = 200 + 0.2y + 0.3z$.

**C)**
the response variables are diet costs (y), fats and sweets (x, in hundreds of grams), and the explanatory variable is fruits and vegetables (z in hundreds of grams). The linear relationship is \( x + 0.2\ y = 0.3\ z \).

**Points:** 1 out of 1

The first relationship has a negative slope for x (fats and sweets), since an increase in x decreases y (diet cost), while it has a positive slope for z (fruits and vegetables) since an increase in x increases y.

**Question 16 of 16**
The coefficient of the regression of weight on height is 0.9580879. The standard deviation of weight is 29.89457 and the standard deviation of height is 22.38937. Therefore the coefficient of the regression of height on weight is...

A) 0.5374
B) 0.7899
C) 0.2323
D) -0.1292
E) 0.9878

**Points:** 1 out of 1

See the slides of session 8 for the explanation that the solution is \((s_x/s_y)^2 b\).

Note that the correct answer is not \(1/b\) (to convince yourself of this, take the formula for \(b\) and switch x and y in that formula, and it will not be equal to \(1/b\)).