MET CS 555 - Data Analysis and Visualization
Quiz - 3

1. You are interested in understanding the relationship between childhood ear infections and adult weight. If you were to conduct a linear regression analysis to explore this relationship, which variable would be the explanatory variable?

   A. Number of childhood ear infections

   B. Whether or not the subject had a childhood ear infection (yes/no)

   C. Adult weight in pounds

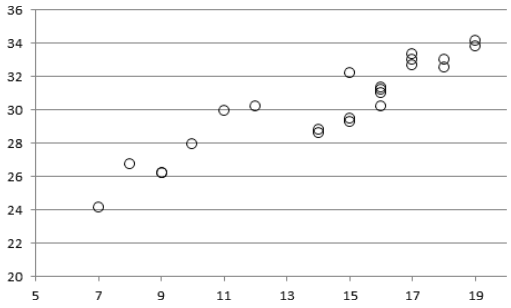
   D. Adult weight in categories: Average, Overweight, Obese

   Answer (A)

   Description. We choose the varible of *"Number of childhood ear infections"* because it is a continuous variables and we are interested to explain the adult weight by using the number of childhood ear infections.

---

2. The scatterplot below shows the association between two factors. The correlation was calculated as 0.93. If an additional datapoint (7, 30) was added, what effect would this have on the correlation coefficient?



   A. No effect

   B. The addition of the point would increase the correlation coefficient

   C. The addition of the point would decrease the correlation coefficient

   Answer (C)

   Description. If you take a look at scatterplot, an additional datapoint (7, 30) would make the correlation less stronger and so the correlation coefficient would decrease.

---

3. The news reports on the results from a new study investigating the relationship between number of childhood ear infections and adulthood weight. The reporter mentions that "the correlation between the number of childhood ear infections and weight in adulthood is +0.30." What does this mean about the relationship between these factors? (Select all that apply)

   A. Those with a higher number of childhood ear infections tend to weigh more as adults on average than those with a lower number of childhood ear infections.

   B. Those with a lower number of childhood ear infections tend to weight more as adults on average than those with a higher number of childhood ear infections.

   C. Those with a higher number of childhood ear infections tend to weigh less as adults on average than those with a lower number of childhood ear infections.

D. Those with a lower number of childhood ear infections tend to weight less as adults on average than those with a higher number of childhood ear infections.

E. Adults who weigh more tend to have had more childhood ear infections than those who weigh less.

F. Adults who weigh less tend to have had more childhood ear infections than those who weigh more.

Answer (A), (D), (E)

Description. The important factor is that as the correlation is a positive +0.3, then any increase in number of childhood ear infections would result to increase of adult weight because of the positive correlation between these two factors. And also any decrease in number would result in decrease of adult weight so that these two factors are always increase and decrease together. In short the two factors are going together (increase and decrease).

---

4. A study shows that the correlation between energy levels and sugar consumption is -0.80. If a least-squares regression was performed, which of the following would be true?

A. $\hat{\beta}_1 < 0$

B. $\hat{\beta}_1 = 0$

C. $\hat{\beta}_1 > 0$

Answer (A)

Description. Because our correlation is a negative correlation, the $\hat{\beta}_1$ should be a negative value so that a decrease in x would result in increase in y (our dependant variable) and also an increase in x would result in decrease in y.

---

5. A study shows that the correlation between head circumference and IQ score is 0.50. If a least-squares regression was performed, which of the following would be true? (Select all that apply)

A. 50% of the varability in IQ score is explained by head circumference

B. 50% of the varability in head circumference is explained by IQ score

C. 25% of the varability in IQ score is explained by head circumference

D. 25% of the varability in head circumference is explained by IQ score

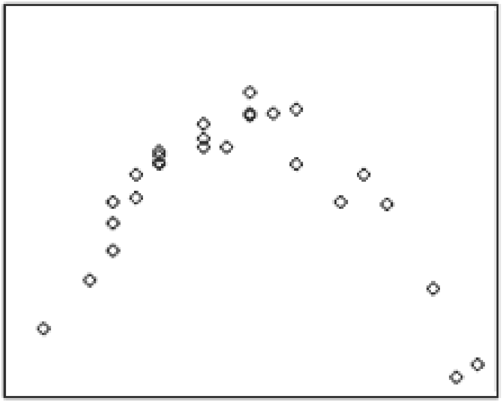
Answer (C) and (D)

Description. The coefficient of variation (or the coefficient of determination) is the square of the sample correlation coefficient (that is, $R^2 = r^2$) and represents the proportion (percentage) of the variation in the response variable explained by the regression model (equation). The percentage of variability is the R^2, so that we need to square the correlation to get the R^2.

In this case we have $0.5^2 = 0.25$ or 25%

The correlation and R^2 are the same regardless which variable is the explanatory variable and which one is the response variable.

6. The relationship between dose (in milligrams) of a new drug for depression and scores on a scale measuring ”happiness” is shown in the graph below. There appears to be a moderately strong relationship between these two variables. If we calculated the correlation coefficient, what would you guess the result would be?



A. $r \approx -0.80$

B. $r \approx 0$

C. $r \approx 0.80$

Answer (B) Description. The correlation coefficient measures the strength of the leaner associations between explanatory and response variable. Here we see a non-linear relationship between the variables, and if we fit a leaner line here we would get some horizontal line (or very close to a horizontal line). Hence, the r is approximately equal to zero $r \approx 0$ and we have no leaner associations between these two variables.

7. Match the correlation with graph.

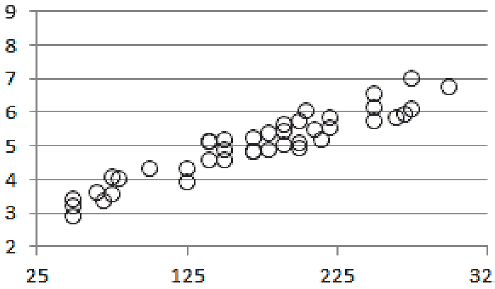


Figure 1: B

A. $r = 0.95$

B. $r = 0.75$

C. $r = 0.60$

Answers (A) - Fig.1, (B) - Fig.2 , and (C) - Fig.3

Description. If we look at the plots we can see that the plot A has the strongest correlation and we would then match that to be the plot for the $r = 0.95$

If we look at the second plot, we see a less strong scatterplot because some of the data points comparing

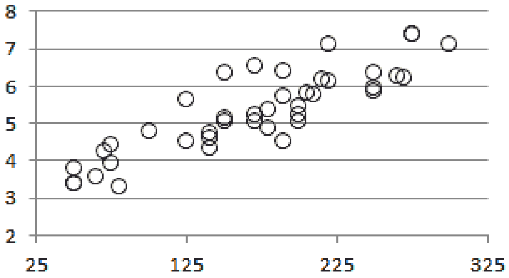


Figure 2: A

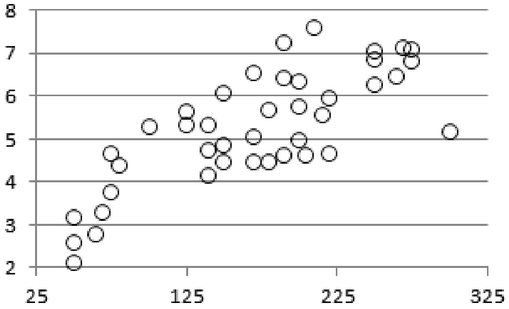


Figure 3: C

to the first plot are more distributed along a straight line and we would match that to be $r = 0.75$. and the last plot shows less strong correlation so that we can match that to the less strong r value here $r = 0.60$.

---

8. A least-squares simple linear regression model was fit predicting duration (in minutes) of a dive from depth of the dive (in meters) from a sample of 12 penguins diving depths and times. Calculate the F-statistic for the regression by filling in the ANOVA table.

	SS	**df**	**MS**	**F-statistic**
Regression				
Residual			2	
Total	30			

A. 2

B. 5

C. 6

D. 10

E. 20

Answer (B)

Description. We need to remember and understand the relations in the ANOVA table. If you take a look at the ANOVA table, you can calculate first the degree of freedom for Regression *Reg df* $= k$ (for simple linear regression $k = 1$) and Residual *Res df* $= n-k-1$ in our case it would be *df* $= 12-1-1 = 10$.

	SS	**df**	**MS**	**F-statistic**
Regression	$Reg\ SS$	$Reg\ df = k$	$Reg\ MS = Reg\ SS/Reg\ df$	$F = Reg\ MS/Res\ MS$
Residual	$Res\ SS$	$Res\ df = n - k - 1$	$Res\ MS = Res\ SS/Res\ df$	
Total	$Total\ SS = Reg\ SS + Res\ SS$			

We can calculate the following table

	SS	**df**	**MS**	**F-statistic**
Regression	10	1	10	5
Residual	20	10	2	
Total	30			

We have $Total\ SS = Reg\ SS + Res\ SS$ so in our case $30 = Reg\ SS + 20$, so that $Reg\ SS = 10$ And then $Reg\ MS = Reg\ SS/Reg\ df$ we can have $Reg\ MS = 10/1 = 10$ And so $F = Reg\ MS/Res\ MS$ is $F = 10/2 = 5$

---

9. A least-squares simple linear regression model was fit predicting duration (in minutes) of a dive from depth of the dive (in meters) from a sample of 12 penguins' diving depths and times. Calculate the R-squared value for the regression by filling in the ANOVA table.

	SS	**df**	**MS**	**F-statistic**
Regression				
Residual			2	
Total	30			

A. 7%

B. 17%

C. 33%

D. 50%

E. 66%

F. 75%

Answer (C)

Description. We have the same ANOVA table as above

	SS	**df**	**MS**	**F-statistic**
Regression	10	1	10	5
Residual	20	10	2	
Total	30			

We know that the R^2

$R^2 = \frac{Reg\ SS}{Total\ SS} = \frac{30}{10} = 33\%$

Also just for info, we know that $R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$

---

10. The shear strength of the bond between two types of propellant is important in the manufacturing of a rocket motor. An investigator is interested in whether the age of the propellant is related to the shear strength? Data from 20 paired observations were used to fit a a least-squares simple linear regression model predicting shear strength from propellant age (in weeks). The equation for the regression is given by $\hat{y} = 1000 - 38x$. What is the correct interpretation of the slope parameter?

A. Each additional week of age of the propellant is associated with a 38 unit increase in shear strength.

B. Each additional week of age of the propellant is associated with a 38 unit decrease in shear strength.

C. Each additional unit of shear strength is associated with a 38 week increase in the age of the propellant.

D. Each additional unit of shear strength is associated with a 38 week decrease in the age of the propellant.

Answer (B)

Description. We have two variables **"age of the propellant"** and **"shear strength"** so that the **"age of the propellant"** is the explanatory variable x and the **"shear strength"** is the response variable y. We have the equation $\hat{y} = 1000 - 38x$ with a negative slope of -38. Now you can see that answer B is correct, because each additional week of age of the propellant is associated with a 38 unit decrease in shear strength.

---

11. The shear strength of the bond between two types of propellant is important in the manufacturing of a rocket motor. An investigator is interested in whether the age of the propellant is related to the shear strength? Data from 20 paired observations were used to fit a a least-squares simple linear regression model predicting shear strength from propellant age (in weeks). The equation for the regression is given by $\hat{y} = 1000 - 38x$. Use this equation to predict the shear strength of a propellant that is 20 weeks old.

    A. 240

    B. 520

    C. 760

    D. 924

    E. 1000

    F. 19,240

Answer (A)

Description. You have the equation $\hat{y} = 1000 - 38x$. You would then plug in your value of 20 weeks $\hat{y} = 1000 - 38 * 20 = 1000 - 760 = 240$.

---

12. Calculate the correlation coefficient for the following data:

x	y
1	1
2	2
2	3
3	2
4	4

A. $r = -0.81$

B. $r = -0.65$

C. $r = 0.65$

D. $r = 0.81$

Answer (D)

Description. You can calculate it manually or use a software like R.

In R code you can write.

```
> x <- c(1, 2, 2, 3, 4)
> y <- c(1, 2, 3, 2, 4)
> cor(x,y)
[1] 0.8076923
```

13. A least-squares simple linear regression model was fit predicting number of headache days per month from daily amount of caffeine consumed. The equation for the regression is given by $\hat{y} = -5 + 0.55x$. What is true about the correlation coefficient between number of headaches and caffeine consumption?

    A. $r < 0$, there is a positive association between the variables

    B. $r > 0$, there is a positive association between the variables

    C. $r < 0$, there is a negative association between the variables

    D. $r > 0$, there is a negative association between the variables

    Answer (B)

    Description. β coefficient is a positive value, and slope of the equation is positive so that correlation coefficient is a positive value and there is a positive association between the variables.

14. A least-squares simple linear regression model was fit predicting variable y from x. See the output below. What is the equation for the least-squares regression line?

Estimate	**SE**	**t-statistic**	**p-value**	
Intercept	-45	5	-9	< 0.001
x	3	2	1.5	0.1441

    A. $\hat{y} = -45 + 3x$

    B. $\hat{y} = 3 - 45x$

    C. $\hat{y} = 5 + 2x$

    D. $\hat{y} = 2 + 5x$

    Answer (A)

    Description. Equation of the simple leaner regression is $\hat{y} = \beta_0 + \beta_1 x$ β_0 is the intercept and β_1 is the slope of the line . You just need to plugin the values from the table into the equation.

15. A least-squares simple linear regression model was fit predicting variable y from x. See the output below. What conclusions can we draw from this output? (Select all that apply)

Estimate	**SE**	**t-statistic**	**p-value**	
Intercept	-45	5	-9	< 0.001
x	3	2	1.5	0.1441

    A. As x increases, y increases.

    B. As x increases, y decreases.

    C. We have sufficient evidence at the alpha = 0.05 level that $\beta_1 \neq 0$ (there is a linear association).

    D. We do not have sufficient evidence at the alpha = 0.05 level that $\beta_1 \neq 0$ (there is not a linear association).

    Answer (A), and (D)

    Description. If you look at the $\beta_1 = +3$, you see a positive value of 3 which means that there is a positive correlation between variables so that as x increases, y increases (you choose the Answer A). If you look at the p-values = 0.1441 for the β_1 you can see that it is not smaller than 0.05 and we can say that we do not have sufficient evidence at the alpha = 0.05 level that $\beta_1 \neq 0$.