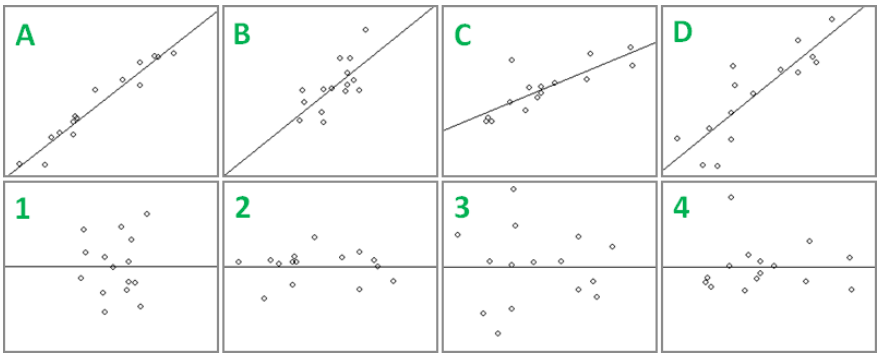
MET CS 555 - Data Analysis and Visualization
Quiz - 4

1. A-D show 4 separate regressions of x on y. 1-4 show the corresponding residual plots. However, the ordering was mixed. Which of the following matches are correct? (Select all that are true.)



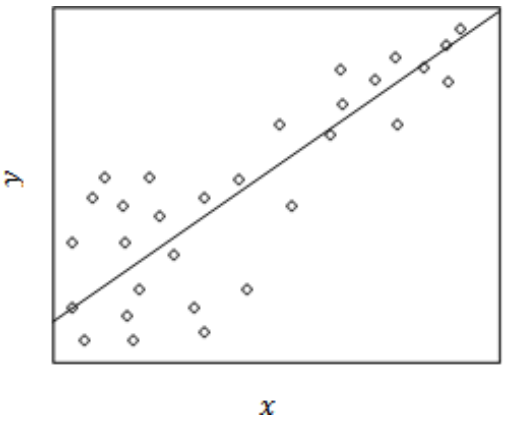
A. A & 1
B. A & 2
C. A & 3
D. A & 4
E. D & 1
F. D & 2
G. D & 3
H. D & 4

Answer (B) and (G)

Description. For this question you just need to understand what the residual plot is and how it is generated from the regression line. You just need to rotate the regression line to a horizontal line and match the data points on the scatterplot with the residual plot. For example for the plot A you can see that at the beginning there is a data point above and another data point below the line. You need to look at the residual plots and check where you can find this case.

---

2. In the figure below, the association between x and y is shown in a scatterplot with the least squares linear regression line. Which of the following assumptions of regression would you be most concerned about violating if you were the analyst working with this data from review of the scatterplot below?



A. Linearity
B. Independence
C. Constant variance
D. Normality

Answer (C)

Description. For checking the normality you need to check the histogram of the residuals and cannot judge the normality only from the scatterplot. You can not check the independence of data points from the visualization and plots. By loooking at the scatterplot you can see that there is a linear relationship betweent the variables and for example we do not see a curve pattern on the scatterplot. Variability of data is something that you can check by looking at the scatterplot or the residual plots. We can see there there is higher variability on lower values of x and the higher values of x have less variability. So, we would be more concerned about the *"Constant variance"*.

---

3. The association between resting heart rate (independent variable) and diastolic blood pressure (dependent variable) was investigated. A least-squares regression line was fit which gave the following equation: $\hat{y} = 1.2x - 20$. One subject had a resting heart rate of 70 and their diastolic blood pressure was 60. Calculate the residual associated with this subject.

A. -8
B. -4
C. -2
D. 2
E. 4
F. 8
G. 60
H. 72
I. 84

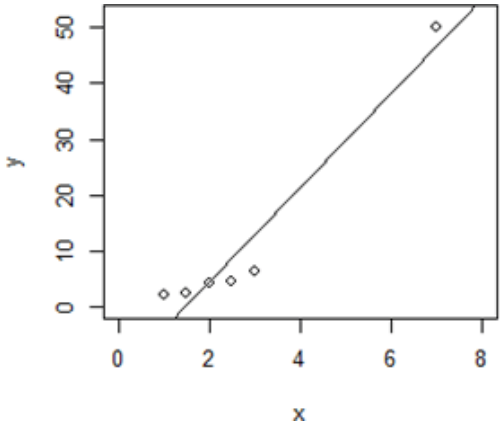
Answer (B)

Description. You need to understand what is the definition of residual value. It is difference between the observed value and the predicted value by the regression equation. You just need to plugin the values, for the predicted value $\hat{y} = 1.2x - 20 = 1.2 * 70 - 20 = 64$ and so the residual would be then between the observed value 60 and the predicted value 64, and it is $60 - 64 = -4$

---

4. A least squares regression line was fit to the data in the table below. Which of the following is true about the observation from ID 4? (Select all that apply)

ID	x	Y
1	2.5	4.5
2	1	2.1
3	1.5	2.5
4	7	50
5	2	4.3
6	3	6.4



A. It is an outlier in the x direction.

B. It is an outlier in the y direction.

C. It has a large residual.

D. It is likely an influence point.

E. It has little effect on the regression given the size of the residual, so no additional action is necessary.

F. It should be removed and the regression should be re-run to see what effect it has on the regression.

G. We should re-check the data to ensure there wasn't a data entry error.

Answer (A), (B), (D), (F) and (G)

Description. First of all, we should re-check the data to ensure there wasn't a data entry error. It has not a large residual, but it is an outlier in both direction of x and y. It should be removed and the regression should be re-run to see what effect it has on the regression.

The size of the residual does not tell you how much effect the data point has on the regression.

If we look at the other data points we can see that we could fit a flat line to those data point with a much smaller slope that the slope of the current line. We can say that it has an effect on the regression line and it is probably an influence point. (It is likely an influence point.)

If you check the data and you see that it is a real data and not a data entry error (or any other kind of data collection error), we should remove the data point and re-calculate the regression to check what are the difference with and without.

---

5. A summary of the output from a least-squares regression model is shown below. What is the value of k for this regression?

	Estimate	**SE**	t	$Pr(>\|t\|)$
Intercept	-4.4570	0.22	-20.0	< 0.0001
Age	0.065509	0.009489	6.90	< 0.0001
Weight	0.15710	0.03321	4.73	< 0.0001
Temp	-0.08725	0.05925	-1.47	0.141

A. 1

B. 2

C. 3

D. 4

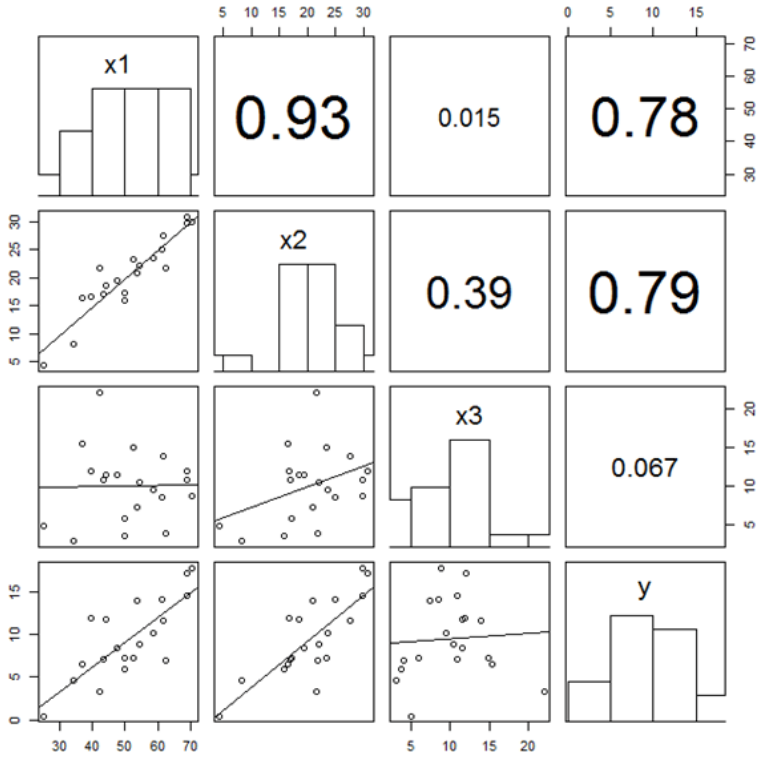
E. 5

Answer (C)

Description. It is number of variables in the table without the intercept so that we have the 3 independent variables of Age, Weight and Temperate.

---

6. A scatterplot matrix was created before a regression was fit. Which of the following might we be most concerned about if we are interested in fitting a multiple linear regression model predicting y from x1, x2, and x3?



A. x3 being a lurking variable

B. collinearity

C. curved or non-linear association between x1 and y

D. outliers and/or influence points

Answer (B)

Description. Here we have scatterplot matrix and we are interested in setting a multiple linear regression by setting y from x1, x2, and x3. We would be concerned about the collinearity because we see really strong correlation (0.93) between x1 and x2.

We see a good linear relation between x1 and y by looking at the last plot on left down side.

We do not see any clear outlier or influence points.

X3 is not lurking variables because we have measured that variable and we see the correlation of the variable x3 with the y.

---

7. A least-squares multiple linear regression model was fit on 49 observations. The resulting regression equation is given by $\hat{y} = 75 + 3x_1 + 5x_2 - 3x_3$. Calculate the F-statistic for the regression by filling in the ANOVA table.

	SS	df	MS	F-statistic
Regression				
Residual	180			
Total	200			

A. 0.19

B. 0.39

C. 0.60

D. 1.67

E. 2.56

F. 5.22

Answer (D)

Description. You need to know the relations in the ANOVA table.

	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 df for regression to be 3 and df for the residuals is $n - k - 1 = 49 - 3 - 1 = 45$

	SS	df	MS	F-statistic
Regression	20	3	6.66667	1.666667
Residual	180	45	4	
Total	200			

and so the **F-statistic** = 1.67

---

8. A least-squares multiple linear regression model was fit on 49 observations. The resulting regression equation is given by $\hat{y} = 75 + 3x_1 + 5x_2 - 3x_3$. Calculate the percent of variance in y explained by x_1 and x_2 and x_3

	SS	df	MS	F-statistic
Regression				
Residual	180			
Total	200			

A. 1%

B. 7%

C. 10%

D. 11%

E. 81%

F. 89%

G. 90%

H. 93%

I. 99%

Answer (C)

Description. Here we have the same ANOVA table as above

	SS	df	MS	F-statistic
Regression	20	3	6.66667	1.666667
Residual	180	45	4	
Total	200			

The percent of variance in y explained by y is the definition of R^2 value. And this is $Regression/Total = 20/200 = 0.10$ equals to 10%

---

9. A least-squares multiple linear regression model was fit for 20 observations taken from mothers that were between 32 and 40 weeks pregnant. The model predicted infant birth weight (in pounds) from gestational age (weeks) and estriol level (measured just before birth from the mother's urine). The resulting regression equation is given by $\hat{y} = 4.18 + 0.04x_{gestage} + 0.15x_{estriol}$. The F-statistic was 20.76 with 2 and 17 degrees of freedom and $p < 0.001$. Which of the following are accurate statements based on (only) these results? (Select all that apply.)

   A. Gestational age, after controlling for estriol level, is a significant predictor of infant birth weight.

   B. Estriol level, after controlling for gestational age, is a significant predictor of infant birth weight.

   C. Gestational age and estriol level, taken together, are significant predictors of infant birth weight.

   D. There is evidence of a linear relationship between infant birth weight and gestational age and estriol level.

   E. Estriol level is a more important predictor of infant birth weight than gestational age.

Answer (C) and (D)

Description. The conclusion of the global F-statistic is that the variables are significant predictors, so in this case it can saythat *"gestational age and estriol level, taken together, are significant predictors of infant birth weight"*.

And as a result of the siginificant F-statistic test, we have linear relation evidence between these birth weight and gestational age and estriol level so that we can say *"There is evidence of a linear relationship between infant birth weight and gestational age and estriol level"*.

We do not have the results of the t-test here so that we do not have results about the beta stimates and we can not say something about "after controlling of one of the variables".

Also the global F-test does not tell us which one of the variables is a more important predictor.

---

10. A least-squares multiple linear regression model was fit for 20 observations taken from mothers that were between 32 and 40 weeks pregnant. The model predicted infant birth weight (in pounds) from gestational age (weeks) and estriol level (measured just before birth from the mother's urine). The resulting regression equation is given by $\hat{y} = 4.18 + 0.04x_{gestage} + 0.15x_{estriol}$.

    The t-statistics for each parameter in the model are summarized in the table below. Which of the following are accurate statements based on (only) these results? Assume a significance level of 0.05. (Select all that apply)

	Estimate	**SE**	t	$Pr(>\|t\|)$
Intercept	4.18	0.88	4.75	< 0.000185
$Gestational_{Age}$	0.04	0.008	5	< 0.00011
$Estriol_{level}$	0.15	0.1	1.5	< 0.151957

   A. Gestational age, after controlling for estriol level, is a significant predictor of infant birth weight.

   B. Gestational age, after controlling for estriol level, is not a significant predictor of infant birth weight.

   C. For every additional week in gestation, infant birth weight increases by approximately 0.04 pounds, after accounting for estriol level.

   D. The average difference in infant birth weights between 32 and 34 week fetuses is predicted by the model to be the same as the average difference in infant birth weight between 36 and 38 week fetuses.

E. Advanced gestational age causes an increase in infant birth weight, after controlling for estriol level.

Answer (A), (C) and (D)

Description. We can see from the results that gestational age is a significant predictor (p-value of 0.00011). So we can say *"Gestational age, after controlling for estriol level, is a significant predictor of infant birth weight."*

By using linearity of the linear regression model, we can say that increase of amount of units can always cause the amount of change in the predicited y vaules. So that we can say *" The average difference in infant birth weights between 32 and 34 week fetuses is predicted by the model to be the same as the average difference in infant birth weight between 36 and 38 week fetuses."* .

And the beta estimate of gestational age 0.04, specifies in the multiple linear regression model the changes in y for each unit change after accounting for estriol level, so that we can say *" The average difference in infant birth weights between 32 and 34 week fetuses is predicted by the model to be the same as the average difference in infant birth weight between 36 and 38 week fetuses."*.

Also in regression model we do not say *"cause"*, we say *"associated with "* or we say *"significant predictors of "*.

---

11. A least-squares multiple linear regression model was fit for 20 observations taken from mothers that were between 32 and 40 weeks pregnant. The model predicted infant birth weight (in pounds) from gestational age (weeks) and estriol level (measured just before birth from the mother's urine). The resulting regression equation is given by $\hat{y} = 4.18 + 0.04x_{gestage} + 0.15x_{estriol}$.

    The t-statistics for each parameter in the model are summarized in the table below. Which of the following are accurate statements based on (only) these results? Assume a significance level of 0.05. (Select all that apply)

	Estimate	**SE**	t	$Pr(>\|t\|)$
Intercept	4.18	0.88	4.75	< 0.000185
$Gestational_{Age}$	0.04	0.008	5	< 0.00011
$Estriol_{level}$	0.15	0.1	1.5	< 0.151957

    A. Reject the null hypothesis that $H0 : \beta_{estriol} = 0$ (after controlling for gestational age)

    B. Accept the null hypothesis that $H0 : \beta_{estriol} = 0$ (after controlling for gestational age)

    C. Fail to reject the null hypothesis that $H0 : \beta_{estriol} = 0$ (after controlling for gestational age)

Answer (C)

Description. Here the p-value for $Estriol_{level}$ is 0.151957 which is not significant (not smaller than 0.05). So we can say that we *" Fail to reject the null hypothesis that* $H0 : \beta_{estriol} = 0$ *(after controlling for gestational age)"*

---

12. A least-squares multiple linear regression model was fit for 20 observations taken from mothers that were between 32 and 40 weeks pregnant. The model predicted infant birth weight (in pounds) from gestational age (weeks) and estriol level (measured just before birth from the mother's urine). The resulting regression equation is given by $\hat{y} = 4.18 + 0.04x_{gestage} + 0.15x_{estriol}$.

    The investigator wants to know what the predicted average infant birth weight for a fetus at 28 weeks of gestation with an estriol level of 25. Select the most appropriate response.

    A. 9.10 pounds

    B. 9.38 pounds

    C. This regression equation should not be used for this purpose as the observations used were all from women who were at 32 to 40 weeks gestation.

Answer (C)

Description. The value of 28 weeks of gestation is outside the values of 32 and 40 weeks that we used to develop our linear regression model, so that we can say *"This regression equation should not be used for this purpose as the observations used were all from women who were at 32 to 40 weeks gestation."*.

Because the value is outside and it causes an extrapolation. We can not be sure that relationship that our model is based on can be used for the values outside the input data for learning the model (If this relation exist the same way outside this range of values).

We can not say that our model can be accurate and predictive for ranges outside the given sample data.
