ST 312: Final Review Problems

Chapters 10-13

Part I: Multiple Choice, True/False Questions

1. The residual is defined as the difference between the:
   A) actual value of y and the estimated value of y
   B) actual value of x and the estimated value of x
   C) actual value of y and the estimated value of x
   D) actual value of x and the estimated value of y
   E) least squares estimated slope $\hat{\beta}$ and the actual population slope β

2. In regression models, which of the following are **true** regarding the error term ε?
   A) $\mu_e = 1$
   B) $\sigma_e = 0$
   C) $\sigma_e = 1$
   D) ε must have negative values
   E) ε values are assumed to have normal distribution.

3. In a one-way ANOVA, the rejection region …
   A) Is always to the left of the t critical value.
   B) Is always to the right of the t critical value.
   C) Is always to the left of the F critical value.
   D) Is always to the right of the F critical value.

4. A large F statistic means…
   A) The between group variation is larger than the within group variation
   B) The between group variation is smaller than the within group variation
   C) The sample means are equal.
   D) The sample variances are significantly different.

5. Which of the following is **true**?
   A) s_ϵ has the same unit as the Y variable
   B) s_ϵ has the same unit as the X variable
   C) s_ϵ has no unit

6. Which of the following is **true** about the regression line $Y = \beta_0 + \beta_1 X$?
   A) X is the response variable and Y is the explanatory variable
   B) β_0 is the slope and β_1 is the intercept
   C) The slope is the amount that X increases when Y increases by 1 unit
   D) None of the above are true

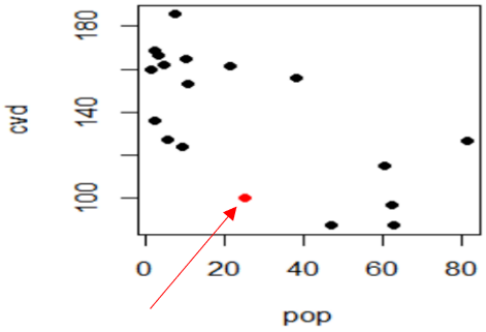
7. Let σ^2 be the common variance of X for each population. In an ANOVA, an estimate for σ^2 that would be valid only under the null hypothesis is given by the
   A) Within group sum of squares
   B) Between group sum of squares
   C) Within group mean squares
   D) Between group mean squares
   E) F test statistic

8. **True / False** : The larger R^2 is, the better the fitted regression line is.

9. **True / False** : The larger s_ϵ is, the better the fitted regression line is.

10. **True / False** : In two-way ANOVA, the interaction test must be completed before the tests for the main effects.

11. A regression model is built to study the relationship between population size (X) and cardiovascular disease incidence rate (Y) throughout the 18 European nations, i.e., $Y = b_0 + b_1 X$. It is known that $\bar{X}$ = 25.3 and $\bar{Y}$ = 137.5.

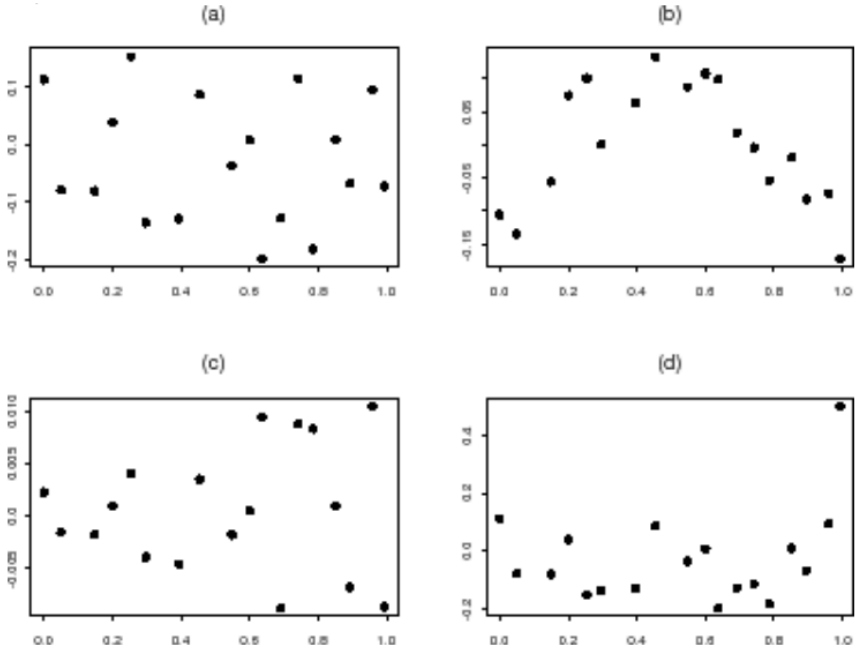


Is Country A (noted by the arrow) an influential point that greatly affects the slope?
   A) Yes
   B) No
   C) Cannot tell

12. In an ANOVA, which of the following is also known as the mean square error (MSE)?
    A) Within group sum of squares
    B) Between group sum of squares
    C) Within group mean squares
    D) Between group mean squares

13. To check for equal variance in ANOVA, we need to know if
    A) The smallest standard deviation is less than 4 times the largest.
    B) The largest standard deviation is less than 4 times the smallest.
    C) The smallest variance is less than 4 times the largest.
    D) The largest variance is twice the smallest.
    E) The largest standard deviation is less than 2 times the smallest.

14. The residual plots of four datasets are shown below. Connect each plot to the appropriate description.



- Figure _____ suggests that the linearity assumption between X and Y may be broken.
- Figure _____ suggests potential outliers (i.e., most of the residuals are randomly scattered around 0, but one observation has produced a residual which is much larger than any of the other residuals.)
- Figure _____ suggests that the variance of residuals is not constant.
- Figure _____ suggests regression assumptions are satisfied since the residual plot has no systematic pattern.

15. In a one-way ANOVA, the total number of observations/units is equal to…
    A) The number of treatments/groups – 1.
    B) The sum of squares total divided by the total degrees of freedom (i.e. SST/df_T).
    C) The total degrees of freedom + 1.
    D) The group degrees of freedom + the error degrees of freedom (df_G + df_E).

16. An F test for a regression model with 3 explanatory variables has which of the following null hypotheses?
    A) $H_0: \mu_1 = \mu_2 = \mu_3$
    B) $H_0: \mu_1 = \mu_2 = \mu_3 = 0$
    C) $H_0: \beta_1 = \beta_2 = \beta_3$
    D) $H_0: \beta_1 = \beta_2 = \beta_3 = 0$
    E) $H_0: \beta_0 = \beta_1 = \beta_2 = \beta_3$
    F) $H_0: \beta_0 = \beta_1 = \beta_2 = \beta_3 = 0$

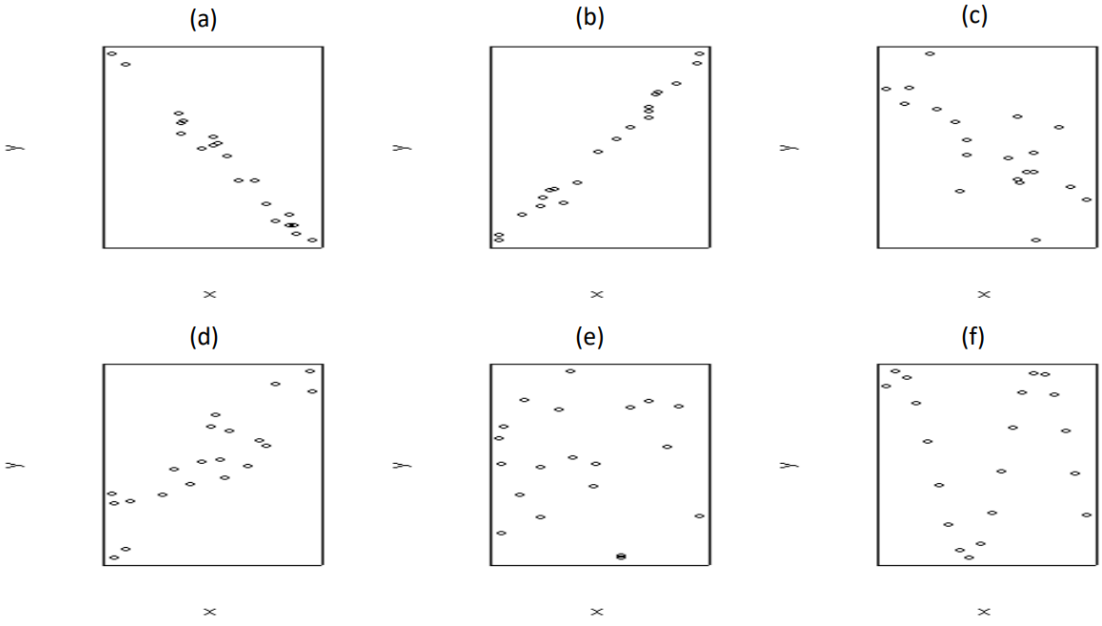
Questions 17-18: A study plans to determine how detergent type and water temperature affect the dirt removal of laundry. There are two detergent types (A and B), and three temperatures (cold, warm, hot). The ANOVA table is provided.

Source	Df	SS	MS	F
Detergent	1	12.03	12.03	2.53
Water Temp	2	312.46	156.23	32.89
Interaction	2	60.46	30.23	6.36
Error	24	114	4.75	
Total	29	498.96		

17. At an α level of 0.05, what is the critical value for the interaction effect?
    A) 3.40
    B) 3.33
    C) 19.00
    D) 4.32

18. Select all of the following statements that are **true**:
    A) We cannot conclude that there is a significant effect from detergent type because the F statistic is not significant at the $\alpha = 0.05$ level.
    B) We can conclude that the water temperature significantly effects the mean amount of dirt removed because the F statistic is in the RR for $\alpha = 0.05$.
    C) We can conclude that detergent type affects mean dirt removal differently across water temperatures.
    D) We can conclude that the change in mean amount of dirt removed between water temperatures varies by detergent type.
    E) We cannot conclude there is a significant effect from detergent type because it's effect cannot be separated from the effect of water temperature.
    F) We cannot conclude there is a significant effect from water temperature because it's effect cannot be separated from the effect of detergent type.

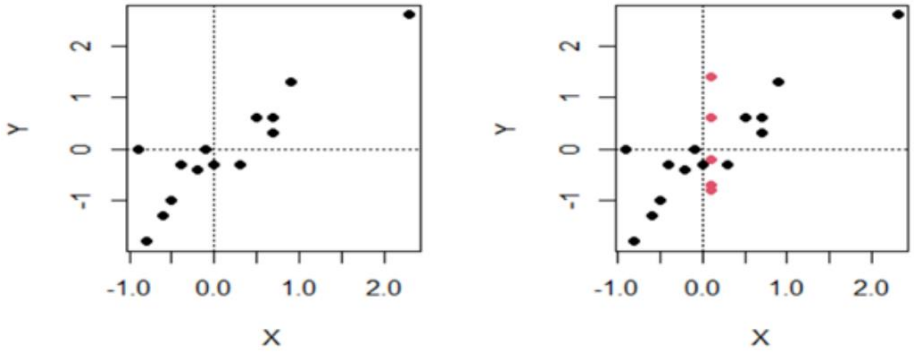
Questions 19-20: Use the plots below to answer the following questions.



19. Suppose that the *correlation* coefficient of the regression line between X (explanatory variable) and Y (response variable) was found to be 0. Which of the above plots may be obtained from X and Y? List **all** that apply.

20. Suppose that the *slope* coefficient between X (explanatory variable) and Y (response variable) was found to be negative. Which of the above plots may be obtained from X and Y? List **all** that apply.

21. In both datasets of the left plot and right plot, the mean of X and Y are both 0. Which plot has the larger correlation coefficient?



A) Left
B) Right
C) Cannot Tell

Questions 22-28. A random sample of 79 companies from the Forbes 500 list (which actually consists of nearly 800 companies) was selected, and the relationship between sales (in hundreds of thousands of dollars) and profits (in hundreds of thousands of dollars) was investigated by regression. The following (partial) results were obtained from statistical software.

Variable	Parameter estimate	Standard error
Constant	–176.644	61.16
Sales	0.092498	0.0075

R^2 = 0.662

(partial) ANOVA table:

Source	DF	Sum of squares	F statistic
Model	1	32,809,212	**
Error	**	**	
Total	**	**	

The researchers also wish to estimate the profits (in hundreds of thousands of dollars) for companies that had sales (in hundreds of thousands of dollars) of 500. The following results were obtained from statistical software.

Sales	Predicted profit	Standard error	95.0% C.I.	95.0% P.I.
500	–130.4	59.3	(–248.5, –12.3)	(–1066.4, 805.6)

22. What is the correlation coefficient between sales and profits?
    A) 0.092
    B) 0.81
    C) −0.81
    D) 0.662
    E) −0.662 11

23. What is an approximate 90% confidence interval for the slope β_1?
    A) –0.09 ± 0.0075
    B) 0.09 ± 0.0075
    C) –0.09 ± 0.0125
    D) 0.09 ± 0.0125

24. If the researchers wish to estimate the mean profits for all companies that had sales of 500, what would be an interval estimate at a 95% confidence level for the mean profits?
    A) (–1066.4, 805.6)
    B) –130.4 ± 59.3
    C) (–248.5, –12.3)
    D) 500 ± 59.3

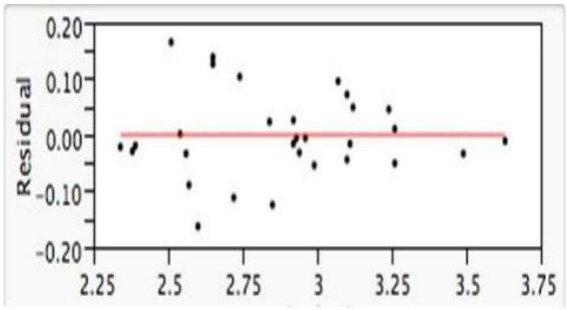
25. If the researchers wish to estimate the profits for a particular company that had sales of 500, what would be an interval estimate at a 95% confidence level for the profits?
    A) (–1066.4, 805.6)
    B) –130.4 ± 59.3
    C) (–248.5, –12.3)
    D) 500 ± 59.3

26. What are the degrees of freedom for SSE, the error sum of squares?
    A) 2
    B) 77
    C) 78
    D) 46

27. What is the value of the total sum of squares?
    A) 21,719,698
    B) 32,809,212
    C) 16,751,531
    D) 49,560,743

28. What is the value of the F statistic for testing the hypotheses H_0: $\beta_1 = 0$ versus H_A: $\beta_1 \neq 0$?
    A) 1.96
    B) 152.1
    C) 77
    D) 217,328

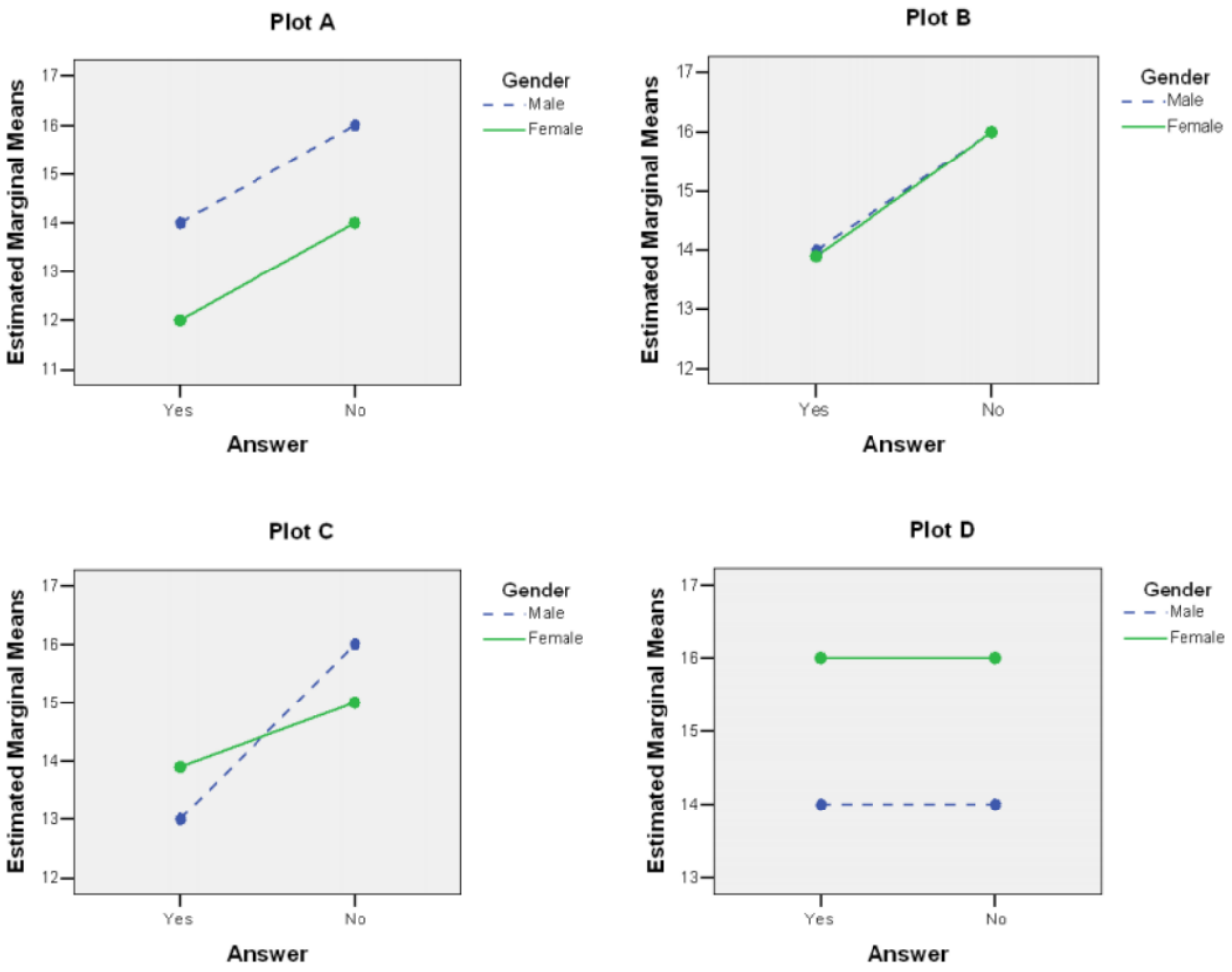
Questions 29-30. In a multiple regression with five explanatory variables, data are collected on 30 observations with the residual plot (residual vs. $\hat{Y}$) shown here.



29. What are the degrees of freedom for the ANOVA F test?
    A) 4 and 24
    B) 5 and 25
    C) 5 and 24
    D) 5 and 29
    E) 6 and 25

30. What statements about residuals and/or about *this* residual plot are **FALSE**? Select all false statements.
    A) The residual plot indicates that the variability is not constant.
    B) The residual plot shows that the residuals do approximately follow a Normal distribution, as the statistical model requires.
    C) Residuals from a least-squares fit in linear regression always sum to zero.
    D) None of the residuals look as though they would be considered to be outliers.

Questions 31-33. The following four graphs are the interaction plots where the response was compared between gender (male/female) and answer (yes/no).



31. Plot _____ shows a pattern of interaction between gender and answer.

32. Plot _____ shows no (main) gender effect.

33. Plot _____ shows no (main) answer effect.

34. Select the procedures that would be appropriate to answer each of the following questions. You only need to give the letter of the answer. (CI = confidence interval)

    A) One sample t test/CI for a single mean
    B) Two-sample t test/CI for a difference in two means
    C) One sample z test/CI for a single proportion
    D) Two-sample z test/CI for a difference in two proportions
    E) Paired t test/CI for a difference in two means
    F) One-Way ANOVA
    G) Two-way ANOVA
    H) Simple Linear regression (with one explanatory variable)
    I) Multiple Linear regression (with two or more explanatory variables)
    J) χ 2 test of independence
    K) χ 2 test for goodness of fit

    (a) _____ What is the average cost to send a child to private school for one year?

    (b) _____ Is the proportion of road accidents the same at intersections involving traffic lights and stop signs?

    (c) _____ How much will my electric bill decrease if I reduce my temperature setting and limit light usage hours?

    (d) _____ Does the distance of a house from the nearest high school impact its selling price?

    (e) _____ How do region (northeast, southeast, northwest, and southwest) and season (spring, summer, fall and winter) affect a home's electric bill?

    (f) _____ How much faster are African swallows than European swallows?

    (g) _____ Do freshmen, sophomores, juniors, and seniors work the same amount of time (in hours per week) at part-time jobs?

    (h) _____ Do Congressional committees with more members propose more laws?

    (i) _____ Does the number of drug side effects decrease after a patient starts an exercise program?

    (j) _____ Is a die fair?

Part II: Computational Problems

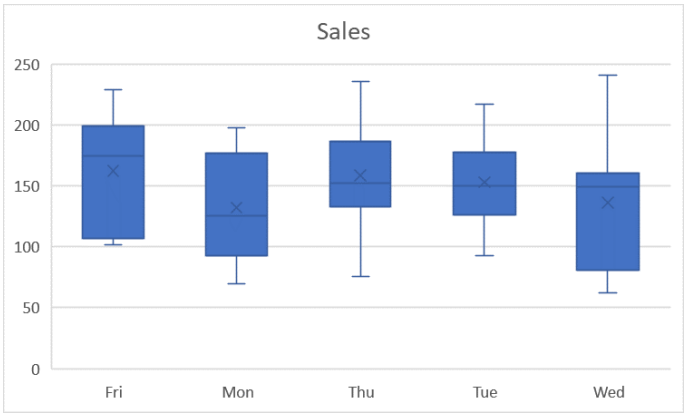
Questions 35-36. In a multiple regression with two explanatory variables, the total sum of squares SST = 1000 and the mean square error MSE = 40. There are 13 observations.

35. What is the value of R^2?

36. What is the value of adjusted R^2?

Questions 37-43. Over a two-month period, data is randomly collected from a café to determine whether overall sales (in dollars) differ significantly based on the day of the week. Summary statistics and ANOVA output are provided below.

Groups	Variance
Fri	2365.216
Mon	2221.271
Thu	1628.181
Tue	1576.208
Wed	3514.81



Source of Variation	SS	df	MS	F
Between Groups	6546.285	4	1636.571	
Within Groups	79139.8	35	2261.137	
Total	85686.09	39		

37. What are the explanatory & response variables?

38. How many total units were measured/observed?

39. Write the hypotheses tested in the ANOVA output provided above.

40. Based on the output, are the assumptions met for the results of this ANOVA test to be valid? Explain.

41. Calculate the F statistic.

42. State the rejection region for a significance level of 5%.

43. State your decision and write an interpretation.

Questions 44-46. In a post-hoc ANOVA analysis, a total of 10 comparisons are performed with α_I being the p-value threshold for each of the comparisons.

44. What is the experiment-wise error rate α_E of this post-hoc ANOVA analysis if $\alpha_I = 0.05$?

45. What is α_I if the Bonferroni procedure is used to control for $\alpha_E = 0.05$?

46. What is the experiment-wise error rate α_E if α_I is set at the value in Q45?

Questions 47-51. Parents with young children are convinced that supermarkets place high-sugar cereals on shelves that are at eye-level with children. Below is the summary statistics of a cereal's sugar content (in grams of sugar per gram of cereal) and a cereal's shelf location with values 1, 2, or 3 (where 1 is the lowest shelf; shelf 2 is at the child's eye-level; 3 is the highest shelf.)

Shelf	Sample size	Sample mean	Sample SD
1	6	0.18	0.173
2	6	0.46	0.022
3	4	0.21	0.091

Analysis of Variance results:

Source	**SS**	**df**	**MS**	**F-Stat**
Groups	0.2715			
Error				
Total	0.4485			

47. Complete the ANOVA table above.

48. Conduct a 5-step ANOVA F Test. Let $\alpha = 0.05$.

49. Specify the contrasts for the following comparisons (Report your contrasts with NO fractions):

ψ_1: The mean difference in sugar content between the cereal on Shelf 2 and the average of cereals on Shelf 1 and Shelf 3.

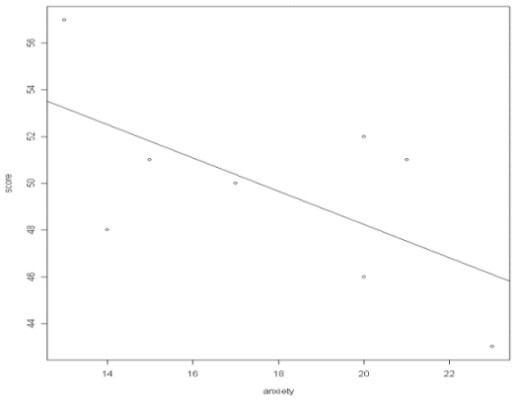
ψ_2: The mean difference in sugar content between the cereal on Shelf 1 and Shelf 3.

50. Test if these comparisons equal 0 by completing the table below. Use Bonferroni's method to control the experiment-wise error rate at 0.05.

Contrast	H_0	Estimate (c_ℓ)	SE of c_ℓ	t*	df	p-value	Reject H_0?
ψ_1							
ψ_2							

51. Draw the practical conclusions obtained in Q51.

Questions 52-61. Is there a significant relationship between test anxiety and exam performance? Data on anxiety score and exam score were obtained from a sample of n = 8 students. Let X=anxiety score and Y=exam score. Below are some summary statistics of this dataset.



$\sum_{i=1}^{8} X_i = 143,\ \sum_{i=1}^{8}(X_i - \bar{X})^2 = 92.875,$

$\sum_{i=1}^{8} Y_i = 398,\ \sum_{i=1}^{8}(Y_i - \bar{Y})^2 = 123.5,$

$\sum_{i=1}^{8}(X_i - \bar{X})(Y_i - \bar{Y}) = -66.25$, and $\hat{\sigma}_{b_1} = 0.37$

52. Judging from the scatter plot, the relation between anxiety level and exam score appear to be (circle any of the following which apply):

    A) positive

    B) negative

    C) linear

    D) nonlinear

53. Find the equation of the least squares regression line for predicting the exam score given an anxiety score. Use 4 decimal places for the slope.

54. Calculate the coefficient of determination and interpret its meaning (Hint: correlation). Round the answer to 3 digits after decimal point and use this value for future calculations.

55. Complete the ANOVA table below and perform a F test to examine if the relationship between the anxiety level and exam score is significant at α=0.05. (Hint: Use R^2)

Source of Variation	SS	df	MS	F
Model				
Error				
Total				

56. Calculate and interpret the sample standard deviation around the regression line.

57. Is there significant evidence that anxiety level is negatively associated with exam score? Use rejection region method with α = 0.05.

58. Provide an estimate of the mean exam score for students with a low anxiety score at 15. Comment on how trustworthy your predictions is and why.

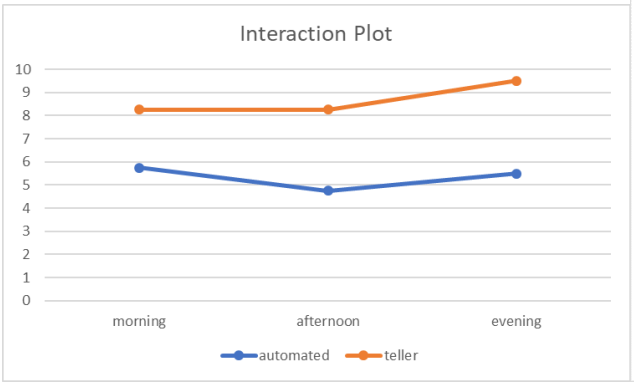
59. **True / False** : The SE in Q58 for anxiety score 15 is the larger than that for anxiety score 17.

60. Estimate the exam score for an individual with a low anxiety score at 15.

61. (Fill in the blank.) A 95% confidence interval for the mean exam score for an anxiety level of 15 is _____ than the 95% prediction interval for the exam score of a student whose anxiety level is 15.

Questions 62-65. A bank is interested in average customer satisfaction regarding the issue of obtaining bank balances and a summary of recent transactions. The bank uses two systems – automated voicemail system, and bank tellers or representatives. The bank takes a random sample of 24 customers and notes both the type of contact (automated or teller) and the time of day (morning, afternoon, evening). The data is summarized below and a plot is provided. Note: this is a balanced design.

Means (SDs)	automated	teller
morning	5.75 (1.708)	8.25 (0.957)
afternoon	4.75 (1.258)	8.25 (1.708)
evening	5.5 (1.0)	9.5 (0.577)



62. Discuss the conditions for conducting a two-way ANOVA in scenario.

63. Complete the ANOVA table below.

Source of Variation	SS	df	MS	F
Time	4		2	
Type	66.67			
Interaction	2.33			
Error		18		
Total	102			

64. Conduct the appropriate test(s) using $\alpha = 0.05$.

65. Write a statement describing your conclusions in the context of the problem.

Questions 66-79. Nutrition facts labels provide consumers with information about the nutritional value of food products that they buy. A study of these labels collected data from 152 consumers who were sent information about a frozen chicken dinner. Each subject was asked to give an overall product nutrition score and also evaluated each of 10 nutrients on a 9-point scale, with higher values indicating that the product has a healthy value for the given nutrient. Composite scores for favorable nutrients (such as protein and fiber) and unfavorable nutrients (such as fat and sodium) were used in a multiple regression to predict the overall product nutrition score. The following was reported:

Variable	**Estimate**	**SE**	**t**
Constant	3.33	0.13	26.1
Unfavorable_nutrients (X_1)	0.82	0.12	6.8
Favorable_nutrients (X_2)	0.57	0.1	**(XXX)**

Model $F^* = 33.4$, $R^2 = 0.31$

66. What is the equation of the multiple regression?

67. What is the missing value (XXX) in the output?

68. What are the degrees of freedom associated with the t statistics that you explained in Q67?

69. Give the null and alternative hypotheses associated with the t test for Favorable Nutrients, report the p-value and interpret this result.

70. Give the null and alternative hypotheses associated with the value labeled ``Model F'' in the output above. Also complete the ANOVA table below, and interpret the result.

Source of Variation	SS	df	MS	F
Model				
Error				
Total	558.3			

71. What proportion of the variation in the overall product nutrition score is explained by the explanatory variables X_1 and X_2?

72. What is the residual SD?

73. Construct the 95% confidence interval for the slope of unfavorable nutrient scores.

74. If the unfavorable nutrient score is 10 and the favorable nutrient score is 5, what is the predicted overall nutrition score?

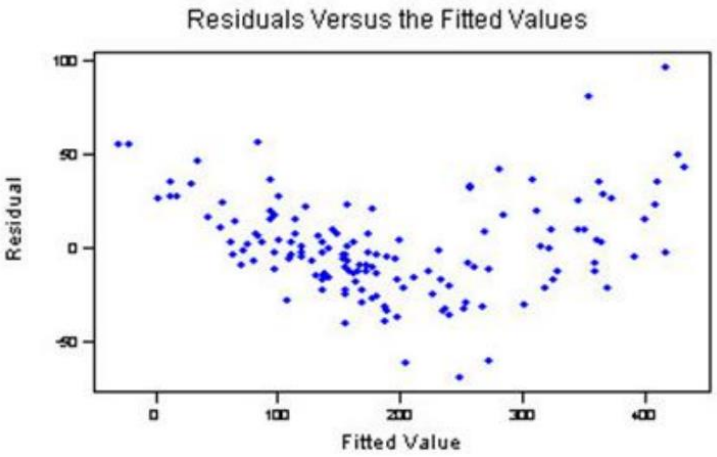
75. If the observed overall nutrition value for a product with favorable and unfavorable scores as given in Q74 is 12.9, what is the residual value for that product?

76. What is the adjusted R^2?

77. Which of the following statement is **true** about the adjusted R^2? Select all that apply.

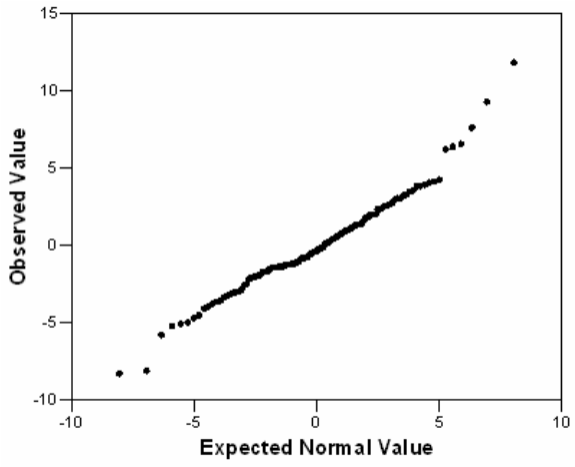
- O Used when a numerical predictor has a curvilinear relationship with the response.
- O Used to check the assumptions of the regression model.
- O Proportion of the variability in y explained by the regression model.
- O A point that lies far away from the rest.
- O Used in a regression model to represent categorical variables.
- O Used when trying to decide between two models with different numbers of predictors

78. Assume that the residual plot is as shown here. Does it indicate any problems with the regression model as fit?



- O There are no problems.
- O Yes, the plot indicates there is nonconstant variance around this line.
- O Yes, the plot indicates curvature is present.

79. The normal quantile plot of residuals is as given below. What assumption do we check with this graph?



- O The Normality of the error terms
- O The independence of the residuals
- O The constant variance assumption of the predicted values
- O None of the above