# Formulas

Stephanie Ranft S2459825

October 24, 2019

**Statistics 2**
**PSBE2-07**

## Exercises

### First partial exam

1. What is the purpose of a simple linear regression?
   (a) To predict scores on an independent variable from scores on a single dependent variable.
   (b) To predict scores on an independent variable from scores on multiple dependent variables.
   (c) To assess whether there is a significant difference between repeated measures.
   (d) To assess whether there is a significant difference between independent groups.
   (e) To predict scores on a dependent variable from scores on multiple independent variables.
   (f) To predict scores on a dependent variable from scores on a single independent variable.
2. What is the purpose of a multiple regression?
   (a) To assess whether there is a significant difference between repeated measures.
   (b) To assess whether there is a significant difference between independent groups.
   (c) To predict scores on an independent variable from scores on multiple dependent variables.
   (d) To predict scores on a dependent variable from scores on a single independent variable.
   (e) To predict scores on a dependent variable from scores on multiple independent variables.
   (f) To predict scores on an independent variable from scores on a single dependent variable
3. What does the Adjusted R squared value tell you?
   (a) The Adjusted R squared value tells you if there is a positive relationship.
   (b) The Adjusted R squared value tells you if there is a negative relationship.
   (c) The Adjusted R squared value tells you if there is a significant difference.
   (d) The Adjusted R squared value tells you how much of the variance in the dependent variable can be accounted for by the independent variable.
   (e) The Adjusted R squared value tells you if there is a significant relationship.
   (f) None of these.
4. Which of the following points are not true when conducting a multiple regression?
   (a) Data must be free from outliers for a multiple regression.
   (b) Data must be homogeneous for a multiple regression.
   (c) The assumption of multicollinearity must be met for a multiple regression.
   (d) Multiple regression can be used to assess linear relationships.
   (e) Data must be normally distributed for multiple regression.
   (f) Multiple regression can be used to assess quadratic relationships
5. Which of these points reflect the assumption of multicollinearity?
   (a) An independent variable cannot be a combination of other independent variables.

(b) Data must be normally distributed and not skewed.

(c) There must not be any extreme scores in the data set.

(d) The relationship between your independent variables must not be above $r = 0.7$.

(e) The variance across your variables must be equal.

(f) None of these.

6. What are residuals?

   (a) Residuals are the differences between the observed and expected dependent variable scores.

   (b) Extreme scores.

   (c) Confidence intervals.

   (d) Uncontrolled variables.

   (e) Serendipitous findings.

   (f) Left over scores

7. The assumption that the variance of the residuals about the predicted dependent variable scores should be the same for all predicted scores reflects which assumption?

   (a) Singularity.

   (b) Multicollinearity.

   (c) Normality.

   (d) Homoscedasticity.

   (e) Homogeneity.

   (f) All of these.

8. What do you report in a multiple regression to say whether your model was significant or not?

   (a) ANOVA.

   (b) Correlation.

   (c) R squared.

   (d) Chi-squared.

   (e) Beta.

   (f) Adjusted R squared.

9. What degrees of freedom do you report in a multiple regression?

   (a) Error and residual degree of freedom.

   (b) Regression and residual degrees of freedom.

   (c) Adjusted R squared and regression degrees of freedom.

   (d) Residual degree of freedom.

   (e) Regression degree of freedom.

   (f) None.

10. What does a beta of 0.478 mean?

    (a) That one model is a better predictor than another.

    (b) That the relationship between the independent and dependent variables is not linear.

    (c) This means that for every unit increase in your independent variable, your dependent variable increases by 0.478 units.

    (d) That there is no predictive power in your independent variable.

    (e) That the regression is not significant.

    (f) That the correlation is significant.

11. What is the correct format for reporting the ANOVA in a multiple regression?

    (a) N = 23, P = 0.000, F = 963.

(b) R (12) = -78.97, $p > 0.001$.

(c) R2 = 78%, F = 278, $p > 0.05$.

(d) T (18) = +8.90, $p < 0.05$.

(e) F (3, 89) = 789.34, $p < 0.001$.

(f) None of these.

12. In a multiple regression problem involving two independent variables, what can you say about their relationship if b1 = 2.0?

(a) The relationship between X1 and Y is significant.

(b) The estimated value of Y increases by an average of 2 units for each increase of 1 unit of X1, holding X2 constant.

(c) The estimated value of Y increases by an average of 2 units for each increase of 1 unit of X1, without regard to X2.

(d) The estimated average value of Y is 2 when X1 equals zero.

13. What does the coefficient of multiple determination measure?

(a) It measures the variation around the predicted regression equation.

(b) It measures the proportion of variation in Y explained by X1 and X2.

(c) It measures the proportion of variation in Y that is explained by X1 holding X2 constant.

(d) It will have the same sign as b1.

14. What formula would you use to calculate the coefficient of multiple determination?

(a) SSR/SST

(b) SSE/SST

(c) SSR/SSE

(d) (SSR+SSE)/SST

15. What is adjusted r2 "adjusted" for?

(a) The number of predictors only.

(b) The sample size only.

(c) The number of predictors and the sample size.

(d) None of the above.

16. Which of the following is not a plot of residuals typically used in multiple regression analysis?

(a) Residuals versus time.

(b) Residuals versus X1.

(c) Residuals versus X2.

(d) Residuals versus correlation coefficients.

17. What is the formula for the F statistic for testing the entire regression model?

(a) SSR/SSE.

(b) MSE/MSR.

(c) MSR/MSE.

(d) MSR/SST.

18. What test would you use to test for the significance of individual regression coefficients in a multiple regression model with more than two explanatory variables?

(a) The Z test.

(b) The t test.

(c) The F test.

(d) None of the above.

19. How are the degrees of freedom associated with the multiple regression model when running a t test for the individual coefficients determined?

    (a) n-p.
    (b) n-1.
    (c) n-p-1.
    (d) n-p+1.

20. Which of the following is correct regarding the value of the adjusted r2 in a multiple regression model?

    (a) It can be negative.
    (b) It has to be positive.
    (c) It has to be larger than the coefficient of multiple determination.
    (d) It can be larger than 1.

21. How are the degrees of freedom determined for SST?

    (a) k
    (b) n-k-1
    (c) n-1
    (d) None of the above.

22. Besides the estimated regression coefficient and appropriate t statistic, what else is needed to construct a confidence interval for a regression coefficient?

    (a) The standard error of the regression coefficient.
    (b) The F statistic.
    (c) The standard error of the estimate.
    (d) The coefficient of determination.

23. In least squares regression, which of the following is not a required assumption about the error term $\varepsilon$?

    (a) The expected value of the error term is one.
    (b) The variance of the error term is the same for all values of x.
    (c) The values of the error term are independent.
    (d) The error term is normally distributed.

24. Larger values of $r^2$ imply that the observations are more closely grouped about the

    (a) average value of the independent variables
    (b) average value of the dependent variable
    (c) least squares line
    (d) origin

25. In a regression analysis if $r^2 = 1$, then

    (a) SSE must also be equal to one
    (b) SSE must be equal to zero
    (c) SSE can be any positive value
    (d) SSE must be negative

26. If the correlation coefficient is 0.8, the percentage of variation in the response variable explained by the variation in the explanatory variable is

    (a) 0.80%
    (b) 80%
    (c) 0.64%
    (d) 64%

27. If the correlation coefficient is a positive value, then the slope of the regression line

    (a) must also be positive
    (b) can be either negative or positive
    (c) can be zero
    (d) can not be zero

28. When the error terms have a constant variance, a plot of the residuals versus the independent variable x has a pattern that

    (a) fans out
    (b) funnels in
    (c) fans out, but then funnels in
    (d) forms a horizontal band pattern
    (e) forms a linear pattern that can be positive or negative

29. Consider the following regression equation: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$. What does $\beta_1$ imply?

    (a) $\beta_1$ measures the marginal effect of $x_1$ on $x_2$.
    (b) $\beta_1$ measures the marginal effect of $y$ on $x_1$.
    (c) $\beta_1$ measures the marginal effect of $x_1$ on $y$.
    (d) $\beta_1$ measures the marginal effect of $x_1$ on $\varepsilon$.

30. If the explained sum of squares is 35 and the total sum of squares is 49, what is the residual sum of squares?

    (a) 10
    (b) 12
    (c) 18
    (d) 14

31. Which of the following is true of R2?

    (a) R2 is also called the standard error of regression.
    (b) A low R2 indicates that the Ordinary Least Squares line fits the data well.
    (c) R2 usually decreases with an increase in the number of independent variables in a regression.
    (d) R2 shows what percentage of the total variation in the dependent variable, Y, is explained by the explanatory variables.

32. If an independent variable in a multiple linear regression model is an exact linear combination of other independent variables, the model suffers from the problem of

    (a) perfect collinearity
    (b) homoskedasticity
    (c) heteroskedasticty
    (d) omitted variable bias

33. Exclusion of a relevant variable from a multiple linear regression model leads to the problem of

    (a) misspecification of the model
    (b) multicollinearity
    (c) perfect collinearity
    (d) homoskedasticity

34. High (but not perfect) correlation between two or more independent variables is called

    (a) heteroskedasticty
    (b) homoskedasticty
    (c) multicollinearity

(d) micronumerosity

35. Find the degrees of freedom in a regression model that has 10 observations and 7 independent variables.

    (a) 17
    (b) 2
    (c) 3
    (d) 4

36. True or False:

    (a) The term "linear" in a multiple linear regression model means that the equation is linear in parameters.
    (b) The key assumption for the general multiple regression model is that all factors in the unobserved error term be correlated with the explanatory variables.
    (c) The coefficient of determination (R2) decreases when an independent variable is added to a multiple regression model.
    (d) A larger error variance makes it difficult to estimate the partial effect of any of the independent variables on the dependent variable.

37. The normality assumption implies that:

    (a) the population error $\varepsilon$ is dependent on the explanatory variables and is normally distributed with mean equal to one and variance $\sigma^2$.
    (b) the population error $\varepsilon$ is independent of the explanatory variables and is normally distributed with mean equal to one and variance $\sigma$.
    (c) the population error $\varepsilon$ is dependent on the explanatory variables and is normally distributed with mean zero and variance $\sigma$.
    (d) the population error $\varepsilon$ is independent of the explanatory variables and is normally distributed with mean zero and variance $\sigma^2$.

38. A normal variable is standardized by:

    (a) subtracting off its mean from it and multiplying by its standard deviation.
    (b) adding its mean to it and multiplying by its standard deviation.
    (c) subtracting off its mean from it and dividing by its standard deviation.
    (d) adding its mean to it and dividing by its standard deviation.

39. Which of the following is a statistic that can be used to test hypotheses about a single population parameter?

    (a) F statistic
    (b) t statistic
    (c) $\chi^2$ statistic
    (d) Durbin Watson statistic

40. Consider the equation, $Y = \beta_1 + \beta_2 X_2 + \varepsilon$. A null hypothesis, $H_0 : \beta_2 = 0$ states that:

    (a) $X_2$ has no effect on the expected value of $\beta_2$.
    (b) $X_2$ has no effect on the expected value of Y.
    (c) $\beta_2$ has no effect on the expected value of Y.
    (d) Y has no effect on the expected value of $X_2$.

41. The significance level of a test is:

    (a) the probability of rejecting the null hypothesis when it is false.
    (b) one minus the probability of rejecting the null hypothesis when it is false.
    (c) the probability of rejecting the null hypothesis when it is true.
    (d) one minus the probability of rejecting the null hypothesis when it is true.

42. The general t statistic can be written as:

    (a) t = hypothesized value / standard error

    (b) t = estimate - hypothesized value

    (c) t = (estimate - hypothesised value) / variance

    (d) t = (estimate - hypothesised value) / standard error

43. Which of the following statements is true of hypothesis testing?

    (a) The t test can be used to test multiple linear restrictions.

    (b) A test of single restriction is also referred to as a joint hypotheses test.

    (c) A restricted model will always have fewer parameters than its unrestricted model.

    (d) OLS estimates maximize the sum of squared residuals.

44. Which of the following statements is true?

    (a) If the calculated value of F statistic is higher than the critical value, we reject the alternative hypothesis in favor of the null hypothesis.

    (b) The F statistic is always non-negative as $SSR_0$ is never smaller than $SSR_1$.

    (c) Degrees of freedom of a restricted model is always less than the degrees of freedom of an unrestricted model.

    (d) The F statistic is more flexible than the t statistic to test a hypothesis with a single restriction.

45. True or False:

    (a) If the calculated value of the t statistic is greater than the critical value, the null hypothesis, H0 is rejected in favor of the alternative hypothesis, H1.

    (b) $H_1 : \beta_j \neq 0$, where $\beta_j$ is a regression coefficient associated with an explanatory variable, represents a one-sided alternative hypothesis.

    (c) Standard errors must always be positive.

*Solution.* 1f, 2e, 3d, 4f, 5d, 6a, 7d, 8a, 9b, 10c, 11e, 12b, 13b, 14a, 15c, 16d, 17c, 18b, 19c, 20a (independence, too many $p$, or small $n$), 21c, 22a, 23a, 24c, 25b, 26d, 27a, 28d 29c, 30d, 31d, 32a, 33b, 34c, 35b, 36(T, F, F, T), 37d, 38c, 39b, 40b, 41c, 42d, 43c, 44b, 45(T, F, T)

**20a:** $\bar{R} \le 0 \implies \begin{cases} R^2 \le \dfrac{p}{n-1} \\ p \ge R^2(n-1) \\ n \le 1 + \dfrac{p}{R^2} \end{cases}$