Overview

Stephanie Ranft

January 16, 2021

Statistics 2
PSBE2-07

Exercises

First partial exam - second sample

1. The strength (degree) of the correlation between a set of independent variables and a dependent variable is measured by

   (a) Coefficient of Correlation

   (b) Coefficient of Determination

   (c) Standard error of estimate

   (d) All of the above

2. The percent of total variation of the dependent variable explained by the set of independent variables is measured by

   (a) Coefficient of Correlation

   (b) Coefficient of Skewness

   (c) Coefficient of Determination

   (d) Standard Error of Estimate

   (e) Multicollinearity

3. A coefficient of correlation is computed to be -0.95 means that

   (a) The relationship between two variables is weak

   (b) The relationship between two variables is strong and positive

   (c) The relationship between two variables is strong and but negative

   (d) Correlation coefficient cannot have this value

4. Let the coefficient of determination computed to be 0.39 in a problem involving one independent variable and one dependent variable. This result means that

   (a) The relationship between two variables is negative

   (b) The correlation coefficient is 0.39 also

   (c) 39% of the total variation is explained by the independent variable

   (d) 39% of the total variation is explained by the dependent variable

5. Relationship between correlation coefficient and coefficient of determination is that

   (a) both are unrelated

   (b) The coefficient of determination is the coefficient of correlation squared

   (c) The coefficient of determination is the square root of the coefficient of correlation

   (d) both are equal

6. Multicollinearity exists when

   (a) Independent variables are correlated less than -0.70 or more than 0.70
   (b) An independent variables is strongly correlated with a dependent variable
   (c) There is only one independent variable
   (d) The relationship between dependent and independent variable is non-linear

7. If "time" is used as the independent variable in a simple linear regression analysis, then which of the following assumption could be violated

   (a) There is a linear relationship between the independent and dependent variables
   (b) The residual variation is the same for all fitted values of the dependent variable
   (c) The residuals are normally distributed
   (d) Successive observations of the dependent variable are uncorrelated

8. In multiple regression, when the global test of significance is rejected, we can conclude that

   (a) All of the net sample regression coefficients are equal to zero
   (b) All of the sample regression coefficients are not equal to zero
   (c) At least one sample regression coefficient is not equal to zero
   (d) The regression equation intersects the Y-axis at zero.

9. A residual is defined as

   (a) $y_i - \hat{y}_i$
   (b) Error sum of square
   (c) Regression sum of squares
   (d) Type I Error

10. What test statistic is used for a global test of significance?

    (a) Z test
    (b) t test
    (c) Chi-square test
    (d) F test

11. In multiple regression analysis, the correlation among the independent variables is termed

    (a) homoscedasticity
    (b) linearity
    (c) multicollinearity
    (d) adjusted coefficient of determination

12. In a multiple regression model, the error term e is assumed to

    (a) have a mean of 1
    (b) have a variance of zero
    (c) have a standard deviation of 1
    (d) be normally distributed

13. In order to test for the significance of a regression model involving 14 independent variables and 50 observations, the numerator and denominator degrees of freedom (respectively) for the critical value of F are

    (a) 13 and 48
    (b) 13 and 49
    (c) 14 and 48
    (d) 14 and 35
    (e) none of the above

14. A multiple regression analysis includes 4 independent variables results in sum of squares for regression of 1400 and sum of squares for error of 600. The VAF will be:

(a) 0.300

(b) 0.700

(c) 0.429

(d) 0.084

(e) none of the above

15. There are situations where a set of explanatory variables forms a logical group. The test to determine whether the extra variables provide enough extra explanatory power to warrant inclusion in the equation is the:

(a) complete F-test

(b) reduced F-test

(c) partial F-test

(d) reduced t-test

(e) none of the above

16. In the example of explaining a person's height by means of his/her right and left foot length, how would you treat for multicollinearity?

(a) Eliminate the right foot variable

(b) Eliminate the left foot variable

(c) Eliminate either foot variable

(d) Eliminate both feet variables

(e) None of the above

17. Determining which variables to include in regression analysis by estimating a series of regression equations by successively adding or deleting variables according to prescribed rules is referred to as:

(a) elimination regression

(b) logical regression

(c) forward regression

(d) backward regression

(e) stepwise regression

18. In Regression Analysis $\sum \hat{Y}$ is equal to

(a) 0

(b) $\sum Y$

(c) b_0

(d) $b_1 \sum X$

(e) None

19. In the Least Square Regression Line, $\sum(Y - \hat{Y})^2$ is always

(a) Negative

(b) Zero

(c) Non-Negative

(d) Fractional

(e) None

20. Which one is equal to explained variation divided by total variation?

(a) Sum of squares due to regression
(b) Coefficient of Determination
(c) Standard Error of Estimate
(d) Coefficient of Correlation

21. The best fitting trend is one for which the sum of squares of error is

(a) Zero
(b) Minimum (Least)
(c) Maximum
(d) None

22. If a straight line is fitted to data, then

(a) $\sum Y = \sum \hat{Y}$
(b) $\sum Y > \sum \hat{Y}$
(c) $\sum Y < \sum \hat{Y}$
(d) $\sum (Y - \hat{Y})^2 = 0$

23. In Regression Analysis two regression lines intersect at the point

(a) (0, 0)
(b) (b_0, b_0)
(c) (X, Y)
(d) $(\overline{X}, \overline{Y})$
(e) None

24. In the Least Square Regression line the quantity $\sum (Y - \hat{Y})$ is always

(a) Negative
(b) Zero
(c) Positive
(d) Fractional
(e) None

Solution. 1d, 2c, 3c, 4c, 5b, 6a, 7d, 8c, 9a, 10d, 11c, 12d, 13e, 14b, 15c, 16c, 17e, 18b, 19c, 20b, 21a, 22d, 23d, 24b, 25

Statistics 2 Exam 16th April 2020

1. What statement about the adjusted R^2 is **not** correct (n = sample size; p = number of predictors)?

   (a) When n/p is very large, R^2_{adj} and R^2 are very similar.

   (b) When p increases then R^2_{adj} will also increase.

   (c) R^2_{adj} estimate the proportion of variance accounted for in the population.

   (d) R^2_{adj} is always smaller than R^2.

2. In regression analysis, a categorical predictor with three categories is included through dummy variables. No other predictor variables included. Let β_i denote the regression coefficient associated to the i-th dummy variable. What null hypothesis is tested by the omnibus F test?

   (a) $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$

   (b) $H_0 : \beta_1 = \beta_2 = \beta_3$

   (c) $H_0 : \beta_1 = \beta_2 = 0$

   (d) $H_0 : \beta_1 = \beta_2$

3. A study as been performed to find out which of three teaching methods yielded the best results. To this end, 20 students were randomly allocated to each method. At the end of the teach period a multiple choice exam has been taken. The mean and standard deviation of the number of questions correct is as follows:

Method	Mean	SD
Method A	31.69	2.29
Method B	27.72	2.76
Method C	24.74	4.82
Total	28.05	4.46

   Compute s_p.

   (a) $s_p = 4.94$

   (b) $s_p = 3.47$

   (c) $s_p = 4.46$

   (d) $s_p = 3.29$

4. A oneway ANOVA has been carried out on a data set containing three groups and 20 measurements per group. All three sample means are exactly equal. Which claim below is **not** true?

   (a) $df_{\text{between}} < df_{\text{error}}$

   (b) $SS_{\text{between}} = 0$

   (c) $MSE = 0$

   (d) $df_{\text{total}} = 59$

5. What is not a questionable research practice?

   (a) Doing more observations as the result is not yet significant.

   (b) In reporting the results, focusing entirely on the significant results.

   (c) Silently removing values as outliers because they do not fit the model.

   (d) Combing two variables into one because of multicollinearity.

6. Consider the following relation between the variable A, B, and C:

$$\boxed{A} \longrightarrow \boxed{B} \longrightarrow \boxed{C}$$

How can this relation be best described?

(a) Variable B moderates the relation between A and C.
(b) Variable B mediates the relation between A and C.
(c) Variable A mediates the relation between B and C.
(d) Variable A moderates the relation between B and C.

7. Which of the alternatives below corresponds to the concept of 'power'?

(a) $\mathbb{P}\,(\text{not rejecting } H_0 \,|H_0 \text{ is true})$
(b) $\mathbb{P}\,(\text{rejecting } H_0 \,|H_0 \text{ is true})$
(c) $\mathbb{P}\,(\text{not rejecting } H_0 \,|H_0 \text{ is false})$
(d) $\mathbb{P}\,(\text{rejecting } H_0 \,|H_0 \text{ is false})$

8. A multiple regression with one dependent variable, Y, and two predictors, X_1 and X_2, has been carried out on a sample size $n = 31$. This yields $R^2 = 0.71$, $r_{Y,X_1} = -0.11$, and $r_{Y,X_2} = 0.47$. Compute the partial correlation coefficient for X_1.

(a) 0.84
(b) 0.70
(c) 0.79
(d) 0.68

9. What is an advantage of centering predictors when performing a regression with interaction between continuous variables?

(a) It improves the interpretability of B_3
(b) It improves the interpretability of B_1 and B_2
(c) The interaction will get a lower p-value and will be significant quicker
(d) All the other alternatives are correct

10. A study has collected data on 50 participants. Predictor variable A has a mean value of 12.0 and $SD = 2.0$. Predictor variable B has a mean of 7.5 and $SD = 1.5$. Dependent variable Y is regressed onto the standardised predictors, denoted a and b, yielding the regression equation:

$$\mathbb{E}(Y) = 12.4 + 2.6a + 3.4b - 1.3ab.$$

Compute the simple regression equation for Y on a for b on SD below the mean.

(a) $17.5 + 0.65a$
(b) $15.8 + 1.3a$
(c) $9 + 3.9a$
(d) $7.3 + 4.55a$

11. A sample of size $n = 56$ yields the following bivariate correlations:

	y	x_1	x_2
y	1.00	0.42	0.62
x_1		1.00	0.28
x_2			1.00

Which of the following four relations is true?

(a) $sr_1 < r_{1,2} < pr_1$

(b) $sr_1 < pr_1 < r_{1,2}$

(c) $r_{1,2} < sr_1 < pr_1$

(d) $pr_1 < r_{1,2} < sr_1$

12. In order to find the relation between exam grade (on the scale from 1 to 10) and time spent preparing for the exam (measured in hours), the following regression model is set up:

$$\text{grade}_i = \beta_0 + \beta_1 \text{time}_i + \varepsilon_i.$$

Consider the following two claims:

A: 'Since both grade and time must be positive, the intercept must be positive as well.'

B: 'If preparation time is measured in days rather than hours, the slope will become a factor 24 larger.'

(a) Claim A is correct, claim B is correct

(b) Claim A is incorrect, claim B is incorrect

(c) Claim A is incorrect, claim B is correct

(d) Claim A is correct, claim B is incorrect

13. In the context of simple linear regression, tests for three out of the following four null hypotheses always yield the same p-value. Which one does not?

(a) $H_0 : \beta_1 = 0$; $H_A : \beta_1 \neq 0$

(b) $H_0 : R^2 = 0$; $H_A : R^2 \neq 0$

(c) $H_0 : \rho = 0$; $H_A : \rho \neq 0$

(d) $H_0 : r = 0$; $H_A : r \neq 0$

14. Which claim is true?

(a) Which alternative is correct depends on the situation.

(b) The median of the likelihood always lies between the median of the prior distribution and the median of the posterior distribution.

(c) The median of the prior distribution always lies between the median of the posterior distribution and the median of the likelihood.

(d) The median of the posterior distribution always lies between the median of the prior distribution and the median of the likelihood.

15. A one-way ANOVA model was used to compare a number of groups with each other. The corresponding ANOVA table is as follows:

	Sum Sq	**df**	**Mean Sq**	**F**	**Sig.**
Between Groups	51.356		25.678	2.488	0.095
Within Groups	433.500		10.321		
Total	484.856	44			

How many groups were included in the analysis?

(a) 5
(b) 4
(c) 2
(d) 3

16. Consider the following two claims about simple linear regression.

A: "Prediction intervals for y at x-values close to $\bar{x}$ are smaller than those at x-values far from $\bar{x}$."

B: "The homogeneity assumption states that the variance of X is fixed."

(a) Claim A is incorrect; claim B is incorrect.
(b) Claim A is correct; claim B is correct.
(c) Claim A is incorrect; claim B is correct.
(d) Claim A is correct; claim B is incorrect.

17. A researcher collects data on 100 men and women aged 18 to 65 on their attitudes towards the environment. No predefined research hypotheses were stated. After studying the data, the researcher decides to write a manuscript entitled 'Grumpy old men: Why men aged 60+ are negative towards the environment'.

What has happened here?

(a) The researcher is HARKing.
(b) The result suggested by the title needs to be validated in future research.
(c) All the other alternatives are correct.
(d) The study is likely to have low power.

18. Consider the multiple regression model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$. Testing the null hypothesis $H_0 : R^2 = 0$ is equivalent to testing which hypothesis?

(a) $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$.
(b) H_0 : All regression coefficients are different from 0.
(c) H_0 : At least one regression coefficient is different from 0.
(d) $H_0 : \beta_1 = \beta_2 = \beta_3$.

19. In a certain town, the probability that it is raining at any given moment is 30%. The probability that there is a traffic jam at the town's main junction is 25%. The probability of a traffic jam during rain is 50%. What is the probability that it is raining, given that there is a traffic jam?

(a) 42%
(b) 50%
(c) 60%
(d) 25%

20. A regression has been performed on two continuous centred predictors x and z. It is known that $n = 81$, $s_x = 2.27$ and $s_z = 1.85$. The estimated regression line is

$$\hat{y}_i = 25.55 + 6.81x_i + 3.15z_i - 1.83x_iz_i.$$

Compute the simple slope for the regression of y on x when z is one standard deviation below mean.

(a) 35.7
(b) 10.2
(c) 3.4
(d) 1.0

21. The values of the dependent variable in a study are denoted by Y. A prior distribution is set up such that $\mathbb{P}(Y > 10) = 0.25$. Subsequently, a sample of size $n = 25$ is drawn, with all values being below 10. What can you say about the posterior probability of $Y > 10$?

(a) The posterior probability is zero.
(b) The posterior probability is larger than 0.25.
(c) The posterior probability is between 0 and 0.25.
(d) The posterior probability is 0.25.

22. A study has been performed on the effectiveness of three types of mindfulness training. In total 60 participants participated in the experiment. The results are summarised in the following table:

Type	$\bar{y}$	sd	n
A	10.00	3.20	15
B	14.00	3.60	30
C	8.00	2.80	15

Compute the upper bound of the 95% confidence interval for group C, based on that group's sd.

(a) 8.78
(b) 11.57
(c) 9.27
(d) 9.55

23. For a sample of size $n = 52$ the correlation coefficient between two variables has been computed as $r = 0.71$. Compute the test statistic for $H_0 : \rho = 0$ versus the two-sided alternative.

(a) $t = 10.12$
(b) $t = 7.13$
(c) $t = 6.99$
(d) $t = 9.32$

24. Consider the following claims about the Kruskal-Wallis test. Which alternative is correct?

Claim A: 'The Kruskal-Wallis test is a non-parametric alternative to two-way ANOVA.'
Claim B: 'Under H_0, the test statistic follows an F distribution with $I - 1$ and $N - I$ degrees of freedom.'

(a) Claim A is incorrect, claim B is incorrect
(b) Claim A is incorrect, claim B is correct
(c) Claim A is correct, claim B is incorrect
(d) Claim A is correct, claim B is correct

25. The population multiple regression model with p predictors

(a) is built on the principle DATA = FIT + RESIDUAL.
(b) is a straight line through the value of $x_1, \ldots, x_p$ and y.
(c) returns the value of the average of the dependent variable, for given values of the predictors.
(d) provides p predictions for the mean dependent variable.

26. Which of the following correctly describes an assumption in linear regression?

(a) There is a perfect linear relation between the scores of the predictor and the dependent variables in the population.
(b) The set of all scores from the dependent variable is normally distributed.
(c) All observations are independent of each other.
(d) There is a perfect linear relation between the scores of the predictor and the dependent variables in the sample.

27. In a sample, two variables A and B are positively correlated. What is **not** a possible explanation for this?

(a) Coincidence.
(b) A and B are common causes of C.
(c) A causes B through C.
(d) B causes A.

28. A one-way ANOVA experiment has been carried out. There were $I = 4$ groups with $n_i = 20$ measurements per group. Consider the contrast that compares group 1 to the mean of the three other groups. How many degrees of freedom does the corresponding t-test have?

(a) 3
(b) 75
(c) 76
(d) 77

29. A two-factor fixed-effecs ANOVA has been carried out, with two levels for both Factor A and Factor B. Based on four measurements per cell, the following ANOVA table was obtained:

Source	SS	df	MS	F
A	24.06	...	...	...
B	19.06	...	...	...
AB	8.56	...	...	...
Within	...	...	5.73	
Total	...	...		

Compute the F-value for Factor B.

(a) $F < 0.5$
(b) $0.5 < F < 1.0$
(c) $F > 2$
(d) $1 < F < 2$

30. The file drawer problem has to do with ...

(a) Mediator analysis
(b) Influential points
(c) Hidden moderators
(d) Publication bias

31. A simple linear regression is carried out in order to predict y from x. This provided estimates $b_0 = 20.00$, $b_1 = 84.00$ and $r_{xy} = 0.36$. Originally x was distance measured in metres. For compatibility with a similar American study, this variable is converted into yards (1 yard = 0.9144 metres). How many of the values b_0, b_1 and r_{xy} change because of this rescaling?

(a) 0

(b) 2

(c) 1

(d) 3

32. Scores of a response variable are collected for two independent groups (Control, Experiment). Some summary statistics are displayed below:

Source	n	Mean	SD
Control (C)	10	9.91	3.54
Experiment (E)	18	11.95	1.85

Consider code variable d such that $d_i = 0$ for subjects in the Control group and $d_i = 1$ for subjects in the Experiment group. What is the estimated regression model $\mu_Y = \beta_0 + \beta_1 d$?

(a) $\hat{y} = 9.91 + 2.04d$

(b) $\hat{y} = 11.95 - 9.91d$

(c) $\hat{y} = 11.95 + 2.04d$

(d) $\hat{y} = 9.91 + 11.95d$

33. For a bivariate sample of size $n = 67$ the correlation is $r = 0.52$. What is the distribution of the test statistic for $H_0 : \rho = 0.5$ versus alternative $H_A : \rho > 0.5$ if the null hypothesis is true?

(a) Standard normal distribution.

(b) t-distribution with 65 degrees of freedom.

(c) t-distribution with 64 degrees of freedom.

(d) t-distribution with 66 degrees of freedom.

34. In a two-way ANOVA setting Factor A has three levels and Factor B has 6 levels. What is $df_{A\times B}$, the degrees of freedom for the interaction term?

(a) 17

(b) 28

(c) 18

(d) 10

35. What is a Type I error?

(a) It is the probability of incorrectly not rejecting the null hypothesis.

(b) It is the probability of correctly rejecting the null hypothesis.

(c) It is the probability of incorrectly rejecting the null hypothesis.

(d) It is the probability of correctly not rejecting the null hypothesis.

36. For a given sample, the Fisher Z correlation is computed as $r_z = -0.44$. Compute the regular correlation coefficient r.

(a) $r = -0.47$

(b) $r = -0.41$

(c) $r = -0.17$

(d) $r = -0.14$

Solutions

1. Using the formula,
$$R^2_{\text{adj}} = 1 - \frac{p}{n-p-1}\left(1 - R^2\right) \leq R^2 \implies 1 - R^2_{\text{adj}} = \frac{p}{n-p-1}\left(1 - R^2\right)$$
we can see that R^2_{adj} and R^2 are very similar when $p/(n-p-1) \approx 1 \implies n \approx 2p+1$, i.e. when n/p is very large. This adjustment provides a better estimate of the VAF for the population than R^2, which overestimates, and therefore R^2_{adj} is a decreasing function of p.
B

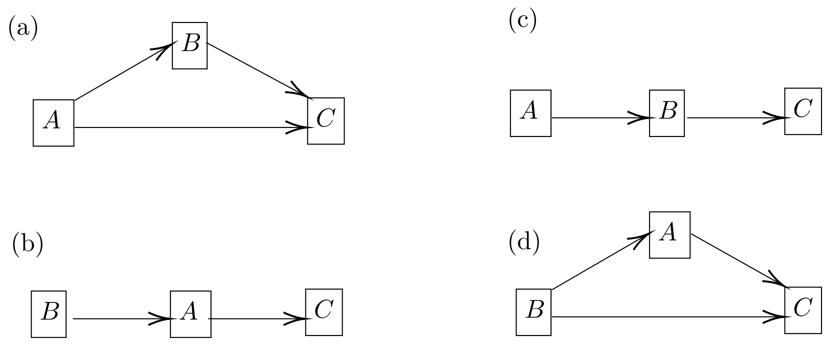
2. As there are three categories (A, B, and C), we select one as the reference group, e.g group A, and by creating two coded variables, we test the hypothesis that all groups are the same for the DV Y, i.e. that the beta coefficients in the following model equal zero:
$$\mu_Y = \mu_A + \underbrace{\left(\mu_B - \mu_A\right)}_{\beta_1} C_1 + \underbrace{\left(\mu_C - \mu_A\right)}_{\beta_2} C_2 + \varepsilon.$$
C

3. The wording of the question is misleading, as I think the professor wants you think that each group has 20 students, however it reads to me that $n = 20$ so each group has 20/3 people.
$$s_p = \sqrt{\frac{2.29^2 + 2.76^2 + 4.82^2}{3}} = 3.4686\cdots \approx 3.47.$$
B

4. As all three group means are equal, they must be equal to the overall mean of the DV and hence
$$SS_{\text{between}} = \sum_j \left(\bar{y}_j - \bar{y}\right)^2 = \underbrace{\left(\bar{y}_1 - \bar{y}\right)^2}_{=0} + \underbrace{\left(\bar{y}_2 - \bar{y}\right)^2}_{=0} + \underbrace{\left(\bar{y}_3 - \bar{y}\right)^2}_{=0} = 0.$$
Furthermore,
$$df_{\text{total}} = n - 1 = 3 \cdot 20 - 1 = 59 = \underbrace{df_{\text{between}}}_{=3-1} + df_{\text{error}} = 2 + df_{\text{error}} \implies df_{\text{error}} = 59 - 2 = 57 > df_{\text{between}} = 2.$$
As $\bar{y}_1 = \bar{y}_2 = \bar{y}_3 = \bar{y}$, if $MSE = 0$ then
$$SS_{\text{error}} = \sum_i \sum_j \left(y_{ij} - \bar{y}_j\right)^2 = \sum_i \left(y_i - \bar{y}\right)^2 = SS_{\text{total}} = 0,$$
i.e. the variance of the DV is zero and every observation is the same.
C

5. This is obvious. **D**

6. **B**



7. (a) and (d) are desirable outcomes; (b) is a Type I error α; (c) is a Type II error β.
D

8.
$$pr_1 = \sqrt{\frac{R^2 - r^2_{Y,X_2}}{1 - r^2_{Y,X_2}}} = \sqrt{\frac{0.71 - 0.47^2}{1 - 0.47^2}} = 0.7923\cdots \approx 0.79$$
C

9. Centring predictors ensures a meaningful zero interpretation of the predictors. Consider Y as explained by predictors X and Z, and by centring these predictors (X_c and Z_c) we interpret B_1 and B_2 as the change in Y for changes in X or Z. The same interpretation cannot be said for the interaction $X_c Z_c$ coefficient B_3, as it involves both centred predictors (biased).

B

10. If b is one SD below the mean, then we need to input $b = -1$ into the regression equal, as $b = (B - \bar{B})/SD_B = (B - 7.5)/1.5$ is standardised.

$$\implies \mathbb{E}(Y) = 12.4 + 2.6a + 3.4 \cdot (-1) - 1.3a \cdot (-1) = 9 + 3.9a.$$

C

11. Using the formula page,

$$pr_1 = \frac{r_{y,1} - r_{1,2}r_{y,2}}{\sqrt{1 - r_{1,2}^2}\sqrt{1 - r_{y,2}^2}} = \frac{0.42 - 0.28 \cdot 0.62}{\sqrt{1 - 0.28^2}\sqrt{1 - 0.62^2}} = 0.3272; \quad sr_1 = \frac{r_{y,1} - r_{1,2}r_{y,2}}{\sqrt{1 - r_{1,2}^2}} = \frac{0.42 - 0.28 \cdot 0.62}{\sqrt{1 - 0.28^2}} = 0.2567.$$

Therefore $sr_1 < r_{1,2} < pr_1$.

A

12. Depending on the data set, time_i may not have a meaningful zero and therefore b_0 may be less than or equal to zero. By transforming time_i from hours to days, we have

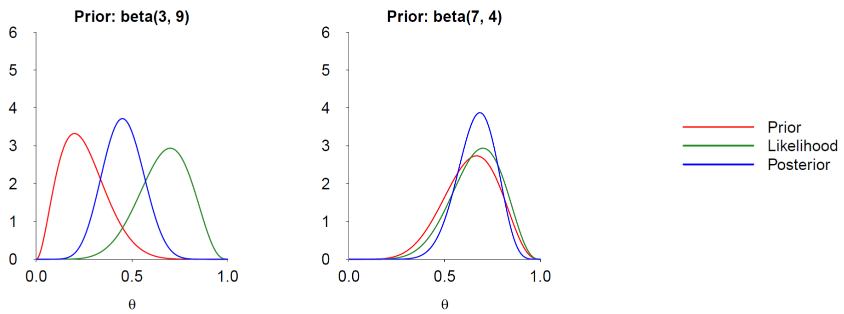
$$\text{days}_i = \frac{\text{time}_i}{24} \quad \implies b_1^* = \frac{\text{Cov}\,(\text{days}_i, \text{grade}_i)}{\text{Var}\,(\text{days}_i)} = \frac{\text{Cov}\,(\text{time}_i, \text{grade}_i)/24}{\text{Var}\,(\text{time}_i)/24^2} = 24 \cdot \frac{\text{Cov}\,(\text{time}_i, \text{grade}_i)}{\text{Var}\,(\text{time}_i)} = 24 \cdot b_1.$$

C

13. (a), (b) and (c) say the same thing, except (d) involves a sample rather than a population statistic and is obviously incorrect.

D

14. Recall this image from the slides:



The posterior distribution is somewhere between prior distribution and likelihood.

D

15. As $MS_G = SS_G/df_G$ we have $df_G = SS_G/MS_G = 51.356/25.678 = 2$, and the number of groups equals $df_G + 1$.

D

16. The prediction interval at some x-value, say x_h, is given by

$$\hat{y}_h \pm t^* \sqrt{MSE\left(1 + \frac{1}{n} + \frac{x_h - \bar{x}}{\sum_i x_i - \bar{x}}\right)},$$

where $\hat{y}_h$ is the predicted value of y at x_h and T^* is the critical value, for some $\alpha/2$ and df. The width of the interval is $2t^* \sqrt{MSE\left(1 + \frac{1}{n} + \frac{x_h - \bar{x}}{\sum_i x_i - \bar{x}}\right)}$, and if x_h is near $\bar{x}$ then the interval is narrower than if x_h is far away from $\bar{x}$.

The homogeneity assumption is that the variance of the error terms e_i is the same (or fixed) for all participants (for $i = 1, \ldots, n$), and that the error terms are uncorrelated. Nothing to do with X.

D

17. Obvious.

C

18. If $R^2 = 0$, then y is independent of x_1, x_2 and x_3, so this is equivalent to $\beta_1 = \beta_2 = \beta_3 = 0$.

A

19. Let R be the event that it is raining, and J be the event of a traffic jam. We are given $\mathbb{P}(R) = 0.30$, $\mathbb{P}(J) = 0.25$ and $\mathbb{P}(J|R) = 0.50$, and so

$$\mathbb{P}(R|J) = \frac{\mathbb{P}(R \cap J)}{\mathbb{P}(J)} = \frac{\mathbb{P}(J|R)\,\mathbb{P}(R)}{\mathbb{P}(J)} = \frac{0.50 \cdot 0.30}{0.25} = 0.60.$$

C

20. The simple slope for the regression of y on x is $6.81 - 1.83z_i$, and as z is centred we have that $z_i = -1.85$ when z is one sd below the mean. Therefore the simple slope for the regression of y on x is $6.81 - 1.83 \cdot (-1.85) = 10.1955$ when z is one sd below the mean.

B

21. As the sample contains no Y-values larger than 10 (likelihood of $Y > 10$ is near zero), the updated belief (posterior probability) of $Y > 10$ is less than 0.25.

C

22. The upper bound is given by

$$\bar{y}_C + t^*_{\substack{0.025\\ n_C-1}} \frac{s_C}{\sqrt{n_C}} = 8 + 2.145 \cdot \frac{2.8}{\sqrt{15}} = 9.55.$$

D

23. The test statistic is $t = r/SE_r$, where

$$SE_r = \sqrt{\frac{1-r^2}{n-2}} \implies t = \frac{0.71\sqrt{52-2}}{\sqrt{1-0.71^2}} = 7.13.$$

B

24. The Kruskal-Wallis test is a non-parametric alternative to one-way ANOVA, and does not require the normality assumption. Therefore the statistic cannot follow an F distribution as this is a part of the 'normal' family (exponential family of distributions).

A

25. The population model is $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon$; (a) refers to OLS, (b) implies that $R^2 = 1$, and (d) makes no sense.

C

26. Assumptions are linearity, normality, independence and homoskedasticity (in the population); (a) implies $\rho = \pm 1$, (b) implies that the sample is normally distributed, and (c) implies $r = \pm 1$.

C

27. (a) implies that A and B do not cause each other (theoretically independent), (c) implies $A \to C \to B$, and (d) implies $B \to A$, which are all plausible given r_{AB} is near +1. However (b) implies $A \to C \leftarrow B$.

B

28. There are $n = 4 \cdot 20 = 80$ participants and so the $df = n - I = 80 - 4 = 76$.

C

29. There are four measurements per cell, i.e. $I = 2 = J$ and $n = 4(I + J) = 16$, therefore $df_A = df_B = 1$, $df_{AB} = 1$, $df_E = n - IJ = 16 - 4 = 12$, and $df_T = n - 1 = 15$. As $MS = SS/df$, we have that $MS_A = 24.06$, $MS_B = 19.06$ and $MS_{AB} = 8.56$, and also $SS_E = 5.73 \cdot 12 = 68.76$ which gives $SS_T = 24.06 + 19.06 + 8.56 + 68.76 = 120.44$. This yields,

$$F_A = \frac{MS_A}{MS_E} = \frac{24.06}{5.73} = 4.20, \qquad F_B = \frac{MS_B}{MS_E} = \frac{19.06}{5.73} = 3.33, \qquad \text{and} \quad F_{AB} = \frac{MS_{AB}}{MS_E} = \frac{8.56}{5.73} = 1.49.$$

C

30. The ‘file drawer problem’ refers to the fact that many results remain unpublished - especially negative ones. This is a problem because it produces publication bias.

D

31. As 1 yard = 0.9144 metres, z yards = $0.9144x$ where x is measured in metres. Then,

$$r_{zy} = \frac{\mathrm{Cov}\,(0.9144x, y)}{\sqrt{\mathrm{Var}\,(0.9144x)\,\mathrm{Var}\,(y)}} = \frac{\cancel{0.9144}\,\mathrm{Cov}\,(x, y)}{\cancel{0.9144}\sqrt{\mathrm{Var}\,(x)\,\mathrm{Var}\,(y)}} = r_{xy} \qquad \text{and} \qquad \bar{z} = \frac{\sum_i z_i}{n} = \frac{0.9144\sum_i x_i}{n} = 0.9144\,\bar{x}.$$

$$\implies b_1^* = r_{zy}\frac{s_y}{s_z} = r_{xy}\frac{s_y}{0.9144\,s_x} = \frac{b_1}{0.9144} \qquad \text{and} \qquad b_0^* = \bar{y} - b_1^*\bar{z} = \bar{y} - \frac{b_1}{\cancel{0.9144}}\cdot\cancel{0.9144}\,\bar{x} = \bar{y} - b_1\bar{x} = b_0.$$

Only the regression slope changes when x is scaled.

C

32. b_0 is the mean of the reference group, therefore $b_0 = \bar{y}_C = 9.91$ and $b_1 = \bar{y}_E - \bar{y}_C = 11.95 - 9.91 = 2.04$ is the difference in the means.

A

33. Must use Fisher Z transformation for null hypotheses of the kind $\rho = \delta$ when $\delta \neq 0$, which is a z-test.

A

34. $df_{A\times B} = (I-1)(J-1) = (3-1)(6-1) = 10.$

D

35. (b) and (d) are desired outcomes, and (a) is a Type II error.

C

36. Using the Fisher Z inverse formula,

$$r = \frac{e^{2r_z} - 1}{e^{2r_z} + 1} = \frac{\exp\{2\cdot(-0.44)\} - 1}{\exp\{2\cdot(-0.44)\} + 1} = -0.41.$$

B