ISyE 6414 Practice Midterm Solutions

Summer Semester 2022

June 7, 2022

Part 1 T/F

Question 1 If the confidence interval for a regression coefficient contains the value zero, we interpret that the regression coefficient is definitely equal to zero.

False. The coefficient is plausibly zero, but we cannot be certain that it is. See Topic 1.2 Lesson 4 and Lesson 5

Question 2 The larger the coefficient of determination or R-squared, the higher the variability explained by the simple linear regression model.

True. R-squared represents the proportion of total variability in Y (response) that can be explained by the regression model (that uses X). R-squared is the proportion of variability explained by the model.See Topic 1.3 Lesson 9.

Question 3 The estimators of the error term variance and of the regression coefficients are random variables. True. The estimators are $\widehat{\beta} = \left(X^T X\right)^{-1} X^T Y$ and $\widehat{\sigma^2} = \widehat{\epsilon}^T \widehat{\epsilon}/(n-p-1)$, where $\widehat{\epsilon} = (I-H)Y$. These estimators are functions of the response, which is a random variable. Therefore they are also random. See Topicl.2 Lesson4.

Question 4 The one-way ANOVA is a linear regression model with one qualitative predicting variable. True. One-way ANOVA is a linear regression model with one predicting factor/ categorical variable. See Topic 2.2 Lesson7

Question 5 We can assess the assumption of constant-variance in simple linear regression by plotting residuals against fitted values.residuals against fitted values.

True. In a residuals Vs fitted plot, if the residuals are scattered around the 0 line, it indicates that the constant variance assumption of errors hold. See Topic1.3 Lesson8

Question 6 If one confidence interval in the pairwise comparison includes zero under ANOVA, we conclude that the two corresponding means are plausibly equal.

True. If the confidence interval includes zero, it is plausible that the corresponding means are equal.See Topic2.2 Lesson5.

Question 7 In Anova, the pooled variance estimator or MSE is the variance estimator assuming equal means.

False. The pooled variance estimator is the variance estimator assuming equal variances. We assume that the variance of the response variable is the same across all populations and equal to sigma square. Module 2 Topic 2.1 Lessons $1-3$

Question 8 Assuming the model is a good fit, the residuals in simple linear regression have constant variance.

True. Goodness of fit refers to whether the model assumptions hold, one of which is constant variance. See Topic 1.8, Lesson 8.

Question 9 You are interested in understanding the relationship between education level and IQ, with IQ as the response variable. In your model, you also include age. Age would be considered a controlling variable while the education level would be an explanatory variable.

True. Controlling variables can be used to control for bias selection in a sample. They're used as default variables to capture more meaningful relationships with respect to other explanatory or predicting factors. Explanatory variables can be used to explain variability in the response variable, in this case the education level.See Topic3.1 Lesson4.

Question 10 If a predicting variable is categorical with 5 categories in a linear regression model without intercept, we will include 5 dummy variables in the model.

True. When we have qualitative variables with k levels, we only include $k-1$ dummy variables if the regression model has an intercept. If not, we will include k dummy variables. See Topic3.1 Lesson2

Question 11 In ANOVA, the number of degrees of freedom of the chi-squared distribution for the variance estimator (not pooled variance estimator) is $N-k-1$ where k is the number of groups.

False. This variance estimator has N-1 degrees of freedom. We lose one DF because we calculate

one mean and hence its N-1. See Topic2.1 Lesson4

Question 12 The only assumptions for a simple linear regression model are linearity, constant variance, and normality. False The assumptions of simple Linear Regression are Linearity, Constant Variance assumption, Independence and normality. See Topic1.1 Lesson2.

Question 13 In simple linear regression, the confidence interval of the response increases as the distance between the predictor value and the mean value of the predictors decreases.

False: The confidence interval bands increase as a predictor increases in distance from the mean of the predictors. See Topic1.2 Lesson6.

Question 14 If the constant variance assumption does not hold in multiple linear regression, we apply a Box-Cox transformation to the predicting variables.

False. (3.11. Assumptions and Diagnostics) If constant variance or normality assumptions do not hold, we apply a Box-Cox transformation to the response variable.

Question 15 Multicollinearity in multiple linear regression means that the columns in the design matrix are (nearly) linearly dependent.

True. (3.13. Model Evaluation and Multicollinearity) Problems arise when the columns of $X^T X$ are not linearly independent, or the value of one predictor can be closely estimated from the other predictors. We call this condition multicollinearity.

Part 2 Multiple Choice

Problem 1. You are thinking about starting a new business. However, your initial capital is limited. To start, you are thinking about a pizza business but your are open to explore options with lower initial investments. For this, you collected the following data on initial investment for several types of industries:

Pizza	Bakery	Shoes	Gifts	Pets
80	150	48	100	25
125	40	35	96	80
35	120	95	35	30
58	75	45	99	35
110	160	75	75	30
140	60	115	150	28
97	45	42	45	20
50	100	78	100	75
65	86	65	120	48
79	87	125	50	20

Consider the following (incomplete) ANOVA table.

Source	Df	Sum of Squares	Mean Squares	F-statistics	p-value
Treatments	A	18186	B	C	0.00662
Error	D	E	1114		
Total	F	68336			

Question 16 What is the value for A in the ANOVA table? 4

If k represents the number of levels of the qualitative variable (here k = 5), this is k − 1.

Question 17 What is the value for B in the ANOVA table? 4546

This is Sum of Squares Treatments / Df Treatments (i.e. 18186/4).

Question 18 What is the value for C in the ANOVA table? 4.08

This is Mean Squares Treatment / Mean Squares Error = 4546/1114

Question 19 What is the value for D in the ANOVA table? 45

If k represents the number of levels of the qualitative variable (here k = 5) and N the number of observations (here $N = 5^*10 = 50$), this is $N - k$.

Question 20 What is the value for E in the ANOVA table? 50150

This is Sum of Squares Total - Sum of Squares Treatments = $68336 - 18186$. Alternatively, this can be solved similar to B in Q15 using 1114D = 111445 (though it doesn't exactly match the solution value due to rounding).

Question 21 What is the value for F in the ANOVA table? 49

If N the number of observations (here N = 5*10 = 50), this is N − 1.

Question 22 What are the null and alternative hypotheses? Null: the mean initial capital is the same for all industries; Alternative: at least two industries have unequal mean initial capital

See Topic2.1 Lesson4

Question 23 Should we reject the null hypothesis? What are the implications in terms of the business problem? Yes, we should reject the null hypothesis. Further analysis is needed to choose the best business.

We reject the null, meaning at least two industries have unequal means. Determining which is the lowest requires further analysis.

Problem 2.

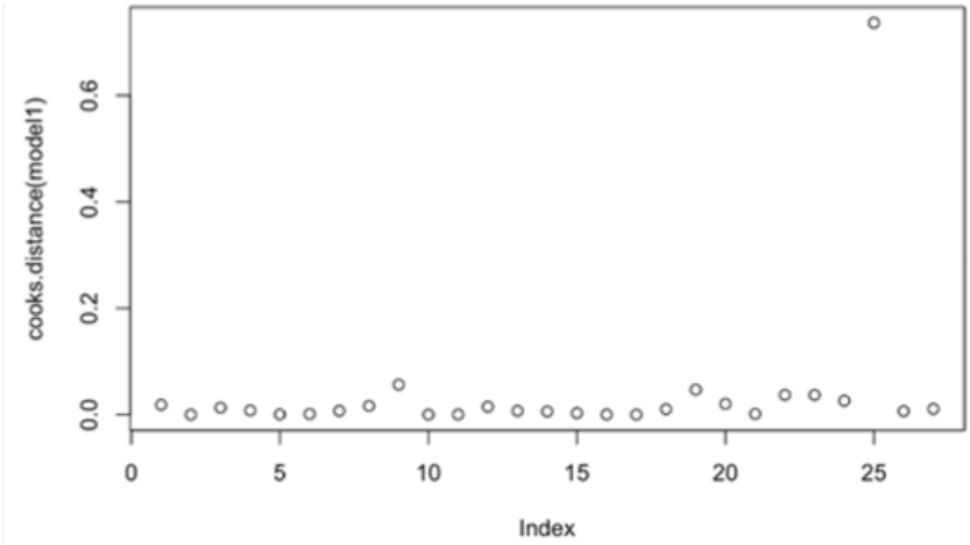
An experiment was conducted to determine the effect of gamma radiation on the numbers of chromosomal abnormalities observed in cells. A multiple linear regression model was fitted to estimate the effect of the number of cells, amount of the radiation dose (Grays), and the rate of the radiation dose (Grays/hour) on the number of chromosomal abnormalities observed. The data frame has 27 observations.
Here is the model summary and Cook's Distance plot.

Coefficient	Estimate	SE	t-value	Pr(>|t|)
(Intercept)	-74.15392	42.24544	-1.755	0.092518
cells	0.06871	0.02196	3.129	0.004709**
doseamt	41.33160	9.13907	4.523	0.000153***
doserate	20.28402	8.29071	2.447	0.022482*

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 54.05 on X degrees of freedom
Multiple R-squared: 0.5213, Adjusted R-squared: 0.4588

F-statistic: 8.348 on Y and X DF, p-value: 0.0006183



Question 24 How does an increase in 1 unit in doserate affect the expected number of chromosome abnormalities, given that the other predictors in the model are held constant? (A) Increase of 8.291 (B) Decrease of 41.331 (C) Increase of 20.284 (D) Decrease of 9.134

Answer (C). (3.2 Lesson 6: Inference for Regression Parameters) The estimated coefficient for doserate is 20.284.

Question 25 Given the value of the coefficient of determination for this model, then (A) Sum of squared errors (SSE) = Sum of squared total (SST) (B) Sum of squared errors (SSE) = 1 (C) Sum of squares for regression (SSR) = Sum of squared total (SST) (D) Sum of squares for regression (SSR) < Sum of squared total (SST)

Answer (D). (3.3 Lesson 13: Model Evaluation and Multicollinearity)

$$R^2 = 0.523 < 1, R^2 = \frac{SSR}{SST} \Longrightarrow SSR < SST$$

Question 26 For an F-test of overall significance of the regression model, what degrees of freedom would be used? (A) 3,24 (B) 2,27 (C) 3,23 (D) 1,23

Answer (C). (3.3 Lesson 13: Model Evaluation and Multicollinearity) $\mathrm{p} = 3; \mathrm{n} - \mathrm{p} - 1 = 27 - 3 - 1 = 23$

Question 27 Calculate the Sum of Squared Regression (SSR) from the model summary. (A) $17,484.25$ (B) $73,163.60$ (C) $67,181.18$ (D) $55,284.40$

Answer (B). (3.3 Lesson 13: Model Evaluation and Multicollinearity)

$$F_{stat} = \frac{MSReg}{MSE} = \frac{\mathrm{SSReg}}{MSE * p}$$
$$SSReg = F_{stat} * MSE * p = 8.348 * 54.05^2 * 3 = 73,163.60$$

Question 28 Based on the cook's distance plot and the rule of thumb $4/n$, how many data points may be outliers potentially influencing the model? (A) None (B) One (C) Two (D) Three

Answer (B). (3.3 Lesson 11: Assumptions and Diagnostics) There are 27 observations hence the rule of thumb is $4/\mathrm{n} = 4/27 = 0.15$. From the graph there is only one observation with a Cook's distance measurement > 0.15.