استنباط آماری

تمرین چهارم

مینا فریدی

810100430

Question 1-

True or false? Explain your reason.

a) In linear regression if we increase the training set size, the mean training error will decrease.

False, the error term is estimated by the expression, Sum of Squares for Error divided by the Degrees of Freedom for Error. Assuming that the variability of the data set is somewhat stable, this means that the size of the error term is dependent on the degrees of freedom for error. The degrees of freedom for error is the sample size reduced by the number of parameters in the model. Thus, the larger the sample size, the larger the degrees of freedom for error (df(error)). Since the df(error) is in the denominator of the estimate, the larger it is, the smaller the error term will be. Thus, the best way to increase the sensitivity of your test is to reduce the error term by increasing the sample size.

b) In linear regression, the mean of residuals is always zero.

True, the mean of residuals in linear regression is always zero. In regression analysis, the difference between the observed value of the dependent variable (y) and the predicted value ($\hat{y}$) is called the residual (e). Each data point has one residual. Both the sum and the mean of the residuals are equal to zero.

c) Standardization of features is required before training a logistic regression model.

False, feature standardization makes the values of each feature in the data have zero-mean. Standardization isn't required for logistic regression. The main goal of standardizing features is to help convergence of the technique used for optimization.

d) Correlated variables can have zero correlation coefficient.

True, Correlation is for measuring linear association and correlated variables can have zero correlation coefficient.

e) In Ridge regression if you apply a very large penalty, some of the coefficients will become absolute zero.

False, by changing the values of alpha, we are controlling the penalty term. The higher the values of alpha, the bigger is the penalty and therefore the magnitude of coefficients is reduced. The consequence of imposing this penalty, is to reduce (i.e. shrink) the coefficient values towards zero. But since the lambda variable is in the denominator the value doen't get absolute zero.

f) If $R^2 = 0$ then there is no relation between x,y.

False, The coefficient of determination r^2 and the correlation coefficient r quantify the strength of a linear relationship. It is possible that $r^2 = 0\%$ and $r = 0$, suggesting there is no linear relation between x and y, and yet a perfect curved (or "curvilinear" relationship) exists.

g) If $R^2 = 0.791$, there is a strong positive linear relationship with strength 0.88 between variables.

False, R might be negative while R^2 is positive and the strength of the relationship depends on the field.

h) The residuals measure the distance between the observed value and the regression line.

True, Residual is the difference between the observed value and the predicted value. Observed value is the actual data point while predicted value is the value obtained from the regression equation. On a residual plot, it is the vertical difference between the observed point and the line.

Question 2-

The principal of a language school wants to evaluate the teachers of the school. He polls students. But using this survey alone will not necessarily be valid because other indicators such as teacher adornment will affect the survey. The table below shows the output of a regression for teacher evaluation among 463 votes.

	estimate	Std.error	T_value	Pr(>|t|)
intercept	4.010	0.025	157.21	0
adornment	0.13	0.032	4.14	0

a) Given that the average adornment score is -0.0883 and average evaluation score is 3.9983. Calculate the missing values of the predictor table.

Mean of the two dimensions

$$\begin{bmatrix} 0 \\ 4.01 \end{bmatrix}$$

$$\begin{bmatrix} -0.08 \\ 3.99 \end{bmatrix}$$

Now we calculate the slope:

$$\frac{3.99 - 4.01}{-0.08} = 0.13$$

SE=0.32 so the t-value is:

$$\frac{0.13}{0.32} = 4.14$$

b) Do these data provide convincing evidence that the slope of the relationship between the evaluation score and adornment is positive? Explain your reasoning.

Since p-value=0 so the hypothesis of H0:B=0 is rejected and there is a slight relationship between the evaluation score and adornment.

H0: B=0

Ha: B>0

c) Calculate a 95% confidence interval for the slope of evaluation score, and interpret it in the context of the data. Do the results from the hypothesis test agree with the confidence interval? Explain.

$t^* = -1.91$

$0.13 \pm (-1.91 * 0.032) = \begin{cases} 0.071 \\ 0.193 \end{cases}$

We are 95% sure that for each unit of increase in the amount of teachers' education, the average evaluation rate is expected to be between 0.07 and 0.19 higher. If we want to judge only on the basis of a safe interval, hypothesis 0 is also rejected because the slope = 0 is not in this interval. Therefore, we are 95% sure that the slope of a positive value is non-zero and in the mentioned range.

Question 3-

Researchers are interested in finding a relation between blood pressure and variables such as age, weight, gender, Water consumption per day, and so on. The results of a regression model for predicting blood pressure based on all of the variables included in the data set is shown in the following table.

	estimate	std.error	t-value	pr(>|t|)
(intercept)	-80.41	14.35	-5.60	0.0
smoke	0.44	0.03	15.26	0.0
exercise hours	-3.33	1.13	-2.95	0.0033
age	-0.01	0.09	-0.10	0.9170
height	1.15	0.21	5.63	0.0
weight	0.05	0.03	1.99	0.0471
water consumption	-8.40	0.95	-8.81	0.0

a) Write the equation of the regression line.

$$y = -80.41 + 0.44B1 - 3.33B2 - 0.01B3 + 1.15B4 + 0.05B5 - 8.40B6$$

b) Interpret the slopes of smoke and water consumption

For each unit increase in smoke, we expect blood pressure to increase by 0.44, and for each unit increase in water consumption, we expect blood pressure to decrease by 8.4. That is, smoking increases blood pressure and water consumption reduces blood pressure, but because the amount of slope for water consumption is higher,

it means that in general, water consumption has a greater effect on blood pressure than smoking.

c) The variance of the residuals is 249.28, and the variance of the blood pressure of people in the data set is 332.57. Calculate the R-squared and the adjusted R-squared.

$$R^2 = 1 - \frac{RSS}{TSS} = 0.25$$

$$adjustedR2 = 1 - \frac{(1 - R^2)(N - 1)}{N - p - 1} = 7.25$$

Question 4-

We want to check the effect of TOEFL score on the chance of admission to universities by using anova test and α=0.01. According to the language score of 428 students, there are three categories of good, medium and bad.

a) Write a hypothesis .

H0: mean of the all categories are the same

Ha: there is a category with a different mean

b) Complete the below table.

	Df	Sum sq	Mean sq	F_value	Pr(>f)
Class	2	1716	572	6.16	0.002326
Residual	426	39518	92983		

DF = N – P

Where:

N = sample size

P = the number of parameters or relationships

DF(class)=2

DF(residual)=426

DF(total) = 428

Sum sq (class)= 2*572=1144

Mean sq(residual)= $\frac{39518}{426} = 92.76$

$$f - value = \frac{mean\ sq(class)}{mean\ sq(residual)} = 6.16$$

c) Suppose the number of good and medium students is 200 and 150, respectively. If the average distance of language scores in these two groups is 8 points, can it be concluded that the chance of admit in these two groups is different?

Question 5- Random Effect

(a) What is the problem of assuming our model data is independent?

There are four assumptions associated with a linear regression model:

1) **Linearity:** The relationship between X and the mean of Y is linear.
2) **Homoscedasticity:** The variance of residual is the same for any value of X.
3) **Independence: Observations are independent of each other**.

   This assumption requires that the observed values of the independent variables be determined independently of the error term.

   Let's say that x is correlated with the error term. we will not be able to estimate the slope of x separately from movements in the error term. we will give credit to x for movement in Y that is due to the error term.

4) **Normality**: For any fixed value of X, Y is normally distributed.

(b) When should we use random effect on our linear model?

A random-effects model assumes that explanatory variables have fixed relationships with the response variable across all observations, but that these fixed effects may vary from one observation to another.

Using random effects is an efficient way to improve the estimates in the linear models. if we have some grouping structures in the investigating dataset that are not directly related to our keen question to answer, it's better to include them as random effects in our linear model.

Question 7- Bootstrap

Ali was so curious about the number of holes in Sangak bread. One day he decided to buy eight loaves of bread at eight different times of the day. Then he counted the number of holes in the bread and here are the results.

43, 59, 22, 25, 36, 47, 19, 21

(a) Find mean and standard deviation of this sample.

$$\text{Mean} = (43 + 59+22+25+36+47+19+21)/8 = 34$$

Standard Deviation, s: **14.628738838328**

Count, N: 8
Sum, Σx: 272
Mean, x̄: 34
Variance, s^2: 214

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2},$$

$$s^2 = \frac{\Sigma(x_i - \bar{x})^2}{N - 1}$$
$$= \frac{(43 - 34)^2 + \ldots + (21 - 34)^2}{8 - 1}$$
$$= \frac{1498}{7}$$
$$= 214$$
$$s = \sqrt{214}$$
$$= 14.628738838328$$

(b) Learn Ali to use bootstrap resampling duo to his samples. Can he use this method with his samples? Explain why.

For doing the bootstrap resampling we should do these steps:

- First, we choose the size of the sample. It can be N (size of the all samples)

- we must randomly choose the first observation from the dataset.
- This observation is returned to the dataset and we repeat this step N-1 more times.

He can do resampling but he should pay attention that resampling is very complex and time consuming.

And it might reduce the accuracy of the experiment since some samples may not be chosen. He should consider that the amount of data is not going to increase by resampling.

(c) What is your expectation about the shape, density, and center of this bootstrap density?

The center is around the lower numbers like 30. And it might be a gamma distribution. Since the numbers are around 20-40 and there are a few numbers lower or more than this range.

(d) What is your interesting parameter of population? Find an estimation for that parameter.

For choosing a parameter we can test the breads from multiple breadshops. And calculate the mean of them.

(e) A bootstrap density from bootstrap statistic has 4.85 SE. Use this SE to calculate the 95% confidence interval of the bootstrap parameter.

Standard Error = 4.85 → 4.85*1.95 = 9.45

$$\text{CI} = \bar{X} \pm margin\ of\ error = 34 \pm 9.45 = \begin{cases} 24.55 \\ 43.45 \end{cases}$$

Question 8- Multiple Comparison

(a) What is the purpose of multiple comparison methods? Which problems these methods are willing to solve?

In statistics, the multiple comparisons occurs when one considers a set of statistical inferences simultaneously or infers a subset of parameters selected based on the observed values.We are not always interested in comparison of two groups per experiment. Sometimes (in practice, very often), we may have to determine whether differences exist among the means of three or more groups. The most common analytical method used for such determinations is analysis of variance (ANOVA). [1] When the null hypothesis (H_0) is rejected after ANOVA, that is, in the case of three groups, H_0: $\mu_A = \mu_B = \mu_C$, we do not know how one group differs from a certain group. The result of ANOVA does not provide detailed information regarding the differences among various combinations of groups. Therefore, researchers usually perform additional analysis to clarify the differences between particular pairs of experimental groups. If the null hypothesis (H_0) is rejected in the ANOVA for the three groups, the following cases are considered:

$$\mu_A \neq \mu_B \neq \mu_C \text{ or } \mu_A \neq \mu_B = \mu_C \text{ or } \mu_A = \mu_B \neq \mu_C \text{ or } \mu_A \neq \mu_C = \mu_B$$

we use the 'multiple comparison test' (MCT) to find out in which of these cases the null hypothesis rejected.

(b) Describe the Bonferroni correction and the Benjamini-Hochberg procedure method in a few lines. What does each method is willing to control?

Bonferroni method:

Bonferroni adjustment is used to control the family-wise error rate (**FWER**). With an increase in the number of hypotheses tested, type I error increases. Therefore, the significance level is divided into numbers of hypotheses tests. In this manner, type I error can be lowered. In other words, the higher the number of hypotheses to be tested, the more stringent the criterion, the lesser the probability of production of type I errors, and the lower the power.

Benjamini-Hochberg:

This method is used to control the false discovery rate (**FDR**). Adjusting the rate helps to control for the fact that sometimes small p-values (less than 5%) happen by chance, which could lead you to incorrectly reject the true null hypotheses. In other words, the B-H Procedure helps us to avoid Type I errors (false positives).

(c) What are the assumptions of each method? Describe them.

Bonferroni designed his method of correcting for the increased error rates in hypothesis testing that had multiple comparisons. A threshold value of α less than 0.05, which is conventionally used, can be set. If the H0 is true for all tests, the probability of obtaining a significant result from this new, lower critical value is 0.05. In other words, if all the null hypotheses, H0, are true, the probability that the family of tests includes one or more false positives due to chance is 0.05.

The **Benjamini-Hochberg** adjustment is very popular due to its simplicity. Rearrange all the P values in order from the smallest to largest value. The smallest P value has a rank of $i = 1$, the next smallest has $i = 2$, and so on.

(d) In which situations we should not have correct for multiple comparisons?

Multiple comparisons can be accounted for with Bonferroni and other corrections, or by the approach of controlling the False Discover Rate. But these approaches are not always needed. Here are three situations were special calculations are not needed.

when interpreting the results rather than in the calculations:

Some statisticians recommend never correcting for multiple comparisons while analyzing data. Instead report all of the individual P values and confidence intervals, and make it clear that no mathematical correction was made for multiple comparisons. This approach requires that all comparisons be reported.

if you make only a few planned comparisons:

in these situations:

- we have chosen a a few scientifically sensible comparisons rather than every possible comparison
- The choice of which comparisons to make was part of the experimental design
- after looking at the data we see no need to do more comparisons

we should set the significance level for each individual comparison without correction for multiple comparisons. In this case, the 5% traditional significance level applies to each individual comparisons, rather than the whole family of comparisons.

when the comparisons are complementary:

In the tests which the samples are divided into subgroups and we are looking for the answer of a same question for all of them.

For example when we want to find the effect of a drug on different groups of people like men, women or different age ranges.

(e) For below p-values use Benjamini-Hochberg procedure to control FDR with $\delta = 0.05$ control level.

0.0008, 0.009, 0.205, 0.165, 0.450, 0.396, 0.641, 0.781, 0.993, 0.90

First we rank the p-values:

0.0008 < 0.009 < 0.165 < 0.396 < 0.450 < 0.641 < 0.781 < 0.90 < 0.993

Now we compare each individual P value to its Benjamini-Hochberg critical value achieved by this equation:

Benjamini-Hochberg critical value = $(i / m) \cdot Q$

(i, rank; m, total number of tests; Q, chosen FDR)

P-values	0.0008	0.009	0.165	0.396	0.450	0.641	0.781	0.90	0.993
Rank	1	2	3	4	5	6	7	8	9
Critical Value	0.0055	0.011	0.016	0.022	0.027	0.033	0.038	0.044	0.05

Now we choose the largest p-value which is smaller than its critical value: 0.009

And also we choose the p-values less than that: 0.0008