University of Tehran
School of Electrical and Computer Engineering

Statistical Inference

Instructor: Dr. Hossein Vahabie

Assignment 4

Statistical Inference Tests

Teaching Assistants:

Amirali Soltani [aa.soltanitehrani@gmail.com]

Kimia Afrazande [kimia.afrazande@ut.ac.ir]

Winter 2022

Homework 4

Statistical Inference, Winter 2022

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.

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

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

d) 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.

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

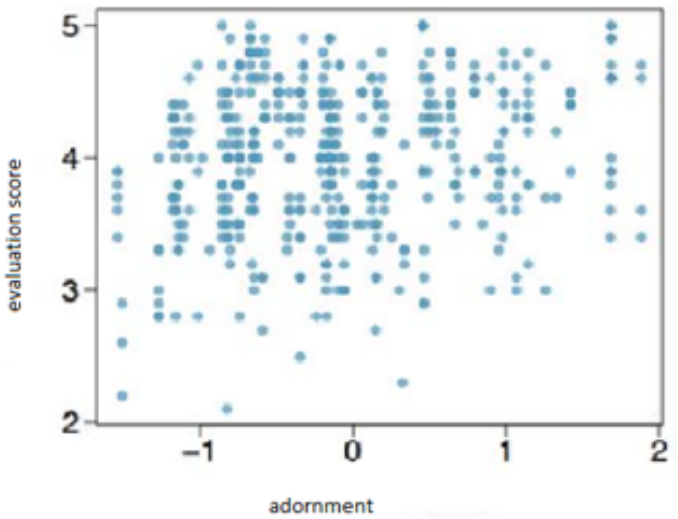
g) If R^2 = 0.791, there is a strong positive linear relationship with strength 0.88 between variables.

h) The residuals measure the distance between the observed value and the regression 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		0.032		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.

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

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.

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.

b) Interpret the slopes of smoke and water consumption

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.

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 .

b) Complete the below table.

	Df	Sum sq	Mean sq	F_value	Pr(>f)
Class			572		0.002326
Residual		39518			

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?

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

(c) (*R*) We are going to work with the NBA_stats.csv dataset which is given to you. You can read about this dataset here. At the cleaning data phase, remove players who have been traded this year. Then replace players to their primary position. You can also use any cleaning data method you have learned on this data.

(d) (*R*) We are going to study whether there is a relation between field goal attempts and field goal percentage. Use a proper plot to see field goal attempts and field goal percentage. Why did you use this plot and did you see any relation between these two variables?

(e) (*R*) Apply linear regression model to this data. Discuss statistics that *R* gives to you. Is this model good? (You can use residual and QQ plot for you discussion)

(f) (*R*) Jannice, who is a statistics enthusiast, was working with this dataset. At first she thought that there is a relation between these two variables but she figured that "what if there is any dependency between these two variables?", so she decided to use her knowledge. Plot boxplot of field goals percentage in all different positions. Do you see any difference in boxplots of each position? Explain it. Use the plot of part d by coloring the position of players. Does position affect these values? Explain it. Can this phenomenon be an example of random effect? Why?

(g) (*R*) Extend the linear model you used in part e by adding position variable to it. (you can use *lme4* library in *R* for this part) Discuss about the statistics that *R* gives to you. How much of the variance is not explained by fixed effect? Does this library report p-value as a statistic? Why?

(h) (*R*) Can you design a statistic to show that adding this extra variable was useful? Explain your test.

Homework 4

Statistical Inference, Winter 2022

Question 6- Generalized Linear Models

For this problem, use the hypothetical data below.

(a) Find the skew and kurtosis for Y_i. (*R*) Create a histogram of Y_i.

(b) A gamma distribution can be used to model skewed metric distributions.

(*R*) Find the gamma probabilities associated with the numbers x = 1 through 10. Set the shape parameter alpha = 2.0 and the scale parameter beta = 1.0. For graphic purposes, add a row into the dataset for x = 0.5, Prob(x) = 8%, to account for the probabilities of x < 1.0 and to make the total add up to 100%. Plot the probability distribution with changing the parameter values to see how the distribution's shape and scale changes.

(c) Use the "Generalized Linear Model" procedure in (*R*) to regress Y_i on X_i. Report the Pearson χ^2 for the following three models: linear, gamma with log link, gamma with identity link. Which model predictions best fit the data?

(d) (*R*) Rerun the linear model and save the residuals and the "Predicted value of linear predictor." Create a scatterplot of the former against the latter. Can you detect the skew in the error distributions from this plot? What is the skew value of the residuals?

(e) What is the skew value of the residuals from the best-fitting model in Part (c)?

Case	X	Y	Case	X	Y	Case	X	Y	Case	X	Y	Case	X	Y
1	1	15	5	5	9	9	9	22	13	13	25	17	17	18
2	2	1	6	6	9	10	10	33	14	14	26	18	18	19
3	3	17	7	7	40	11	11	13	15	15	57	19	19	40
4	4	28	8	8	6	12	12	14	16	16	17	20	20	21

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.

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

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

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

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

Question 8- Multiple Comparison

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

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

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

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

(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