Statistical Inference

Lecturer: Abdol-Hossein Vahabie

Spring Semester 1401-1402

Writing Assignment IV Deadline 1402/03/04

1 Short Answers

Answer each question in less than 2 lines.

1. What is the difference between simple linear regression and multiple linear regression?
2. How do you interpret the slope coefficient in a linear regression model?
3. What is the purpose of residual analysis in linear regression?
4. What are link functions and how does the choice of link function affect the interpretation of coefficients in a Generalized Linear Regression model?
5. Can you explain how Ridge Regression works and how it differs from Ordinary Least Squares Regression?
6. In what situations might a Poisson regression model be more appropriate than a linear regression model?
7. What is multicollinearity and how do you handle multicollinearity in a linear regression model?
8. Explain the concept of heteroscedasticity and how it can impact the results of linear regression analysis. How do you test for heteroscedasticity in a linear regression model?
9. What is the difference between logistic regression and linear regression?
10. What is overdispersion in generalized linear models and how do you address it?
11. How do you interpret odds ratios in logistic regression models?

12. Can you use categorical variables in a linear or generalized linear regression model? If so, how?

13. What is regularization and why might it be used in a linear or generalized linear regression model?

14. Can outliers affect the results of a linear or generalized linear regression model? If so, how can they be addressed? What will you do to encounter outliers in your data?

15. How do you assess goodness of fit for a logistic regression model?

16. Can interaction terms be included in a multiple linear or generalized linear regression model? If so, how are they interpreted?

17. What is the difference between parametric and nonparametric methods for modeling relationships between variables?

18. Imagine you have some time series data and some missing values. What choices will you have for filling in these missing values? If your data belongs to the stock market, which one of the methods you proposed should be applied?

19. Can nonlinear relationships be modeled using either multiple or generalized linear regressions? If so, how can this be done?

20. Explain what ACF (auto-correlation function) and PACF (Partial auto-correlation function) plots are and when we use them.

21. What statistical test do we apply to find out if a time series is a white noise or not? Explain two of them by your choice.

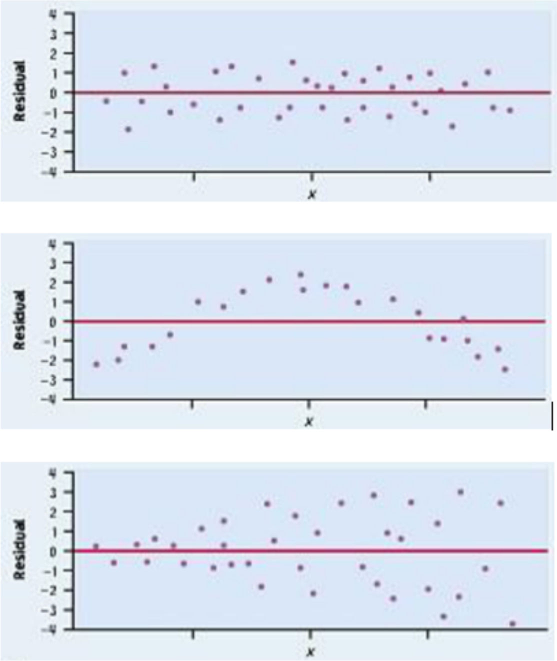
22. How do we measure the accuracy and performance of a regression model?

23. What is stationarity in time series data? What should we do if our data is non-stationary?

24. Can we use ANOVA in a linear Regression context? If so, how? What will the meaning of each variable in this setting be?

25. What are the conditions for inference in the problem of linear regression?

26. What is the meaning of each of these residual plots below? What will you do if you encounter such a residual plot?



27. Why do we rely on n-2 degree of freedom t-distribution when estimating regression parameters β_0 and β_1?

28. You get a confidence interval of (-0.34,4) for β_1 in a problem you are trying to solve. What does this confidence interval mean?

29. What are AIC and BIC metrics and when do we use them?

30. The table below shows the percent of obese people in the United States. What will this percentage be in 2050? Is this extrapolation logical?

Year	**1990**	**1995**	**2000**	**2005**	**2010**	**2015**	**2020**
Percent	30%	32%	35%	39%	46%	50%	55%

2 True/False

1. Linear regression assumes that the relationship between the dependent variable and independent variables is linear.

2. Generalized linear regression can handle non-normal response variables.

3. White noise is always stationary.

4. Autocorrelation in time series data means that there is no relationship between consecutive observations.

5. In linear regression, multicollinearity occurs when two or more independent variables are highly correlated with each other.

6. Generalized linear regression can handle both continuous and categorical response variables.

7. A stationary time series has a constant mean and variance over time.

8. In linear regression, the residuals should be normally distributed with a mean of zero and constant variance across all values of the independent variable(s).

9. In generalized linear regression, the link function relates the expected value of the response variable to a linear combination of the predictor variables.

10. In linear regression, outliers can have a significant impact on the model's coefficients and predictions.

11. In generalized linear regression, deviance measures how well the model fits the data.

12. In linear regression, adding more independent variables will always improve the model's fit to the data.

13. In generalized linear regression, the dispersion parameter measures the degree of overdispersion in the response variable.

14. A stationary time series has a constant autocorrelation over time.

15. In linear regression, the coefficient of determination (R-squared) measures the proportion of variance in the dependent variable that is explained by the independent variables.

16. In generalized linear regression, the likelihood function is used to estimate the model's parameters.

17. Autocorrelation in time series data means that there is a relationship between consecutive observations.

18. The least-squares regression line $(y = \beta_0 + \beta_1 x)$ is an estimate of the true population regression line $(\mu_y = \beta_0 + \beta_1 x)$. The fitted values β_0 and β_1 are unbiased estimators of the intercept and slope of the population regression line.

19. After fitting the regression line, it is important to investigate the residuals to determine whether or not they appear to fit the assumption of a normal distribution.

20. In a regression problem, the null hypothesis states that the slope coefficient, β_1, is equal to 0. If this is true, then there is no linear relationship between the explanatory and dependent variables, and the linear regression equation, $(y = \beta_0 + \beta_1 x + \varepsilon)$, simply becomes $(y = \beta_0 + \varepsilon)$.

21. The alternative hypothesis in a regression problem may be one-sided or two-sided, stating that β_1 is either less than 0, greater than 0, or simply not equal to 0.

22. Correlation is a measure of the association between any two variables.

23. Survival from a disease has been calculated and the coefficient for sex is x and 2x for age. Does it mean that age is twice as important as sex?

24. If the variable is not a meaningful predictor, R2 will be very close to 1 and adjusted R2 will decrease.

25. If the variable is a meaningful predictor, adjusted R2 will increase and be higher than R2.

3 Linear Regression

A nutritionist has collected the number of minutes swam per day and the corresponding daily calorie consumption for a sample of athletes. The data is shown in the table below:

Minutes Swam	Daily Calorie Consumption
35	450
40	500
45	550
50	600
55	650

(a) Use linear regression to estimate the relationship between minutes swam and daily calorie consumption. (You can create a scatter plot to visualize the relationship between two variables)

(b) Predict the daily calorie consumption for an athlete who swims 48 minutes per day.

(c) Find the value of the coefficient of determination (R^2) of the regression equation and interpret it.

4 Multiple Linear Regression

Suppose we want to investigate the relationship between a company's sales revenue (y) and their advertising spending on TV (x1), radio (x2), and newspaper (x3). We collect data from a random sample of 50 weeks and fit the following multiple linear regression model using least squares regression :
$y = beta_0 + beta_1 X_1 + beta_2 X_2 + beta_3 X_3$

Source	df	SS	MS	F
Regression	2	1200000	600000	9.85
Error	47	2800000	59574.47	
Total	49	400000		

(a) Perform the overall significance test at a significance level of 0.05.

(b) Test the significance of each individual regression coefficient at the same significance level.

5 EEG Data (R)

Electroencephalography (EEG) is a method to record an electrogram of the spontaneous electrical activity of the brain. It is typically non-invasive, with the EEG electrodes placed along the scalp. The Device that Extracted this data had 126 channels.

In this data, we have 7 second of recording with 500 Hz brain signals of 90 people. The first 45 people are shown a picture of a face and the other 45 people are shown a piano picture. We want to find out which signals are important in detecting what picture (face or piano) the person saw. Nothing is shown from 0 to 1 second. At the beginning of the first second the picture is shown.

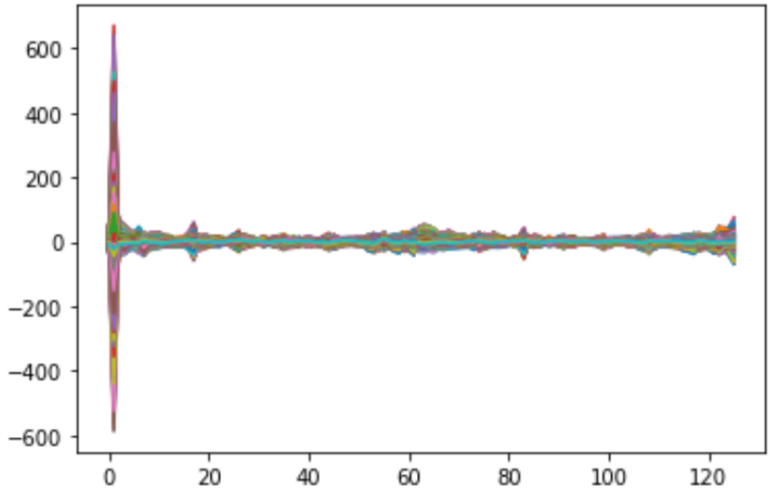
EEG is a time series data and time series are not good features by themselves. We usually extract some features from time series that help us classify them. These features are domain-specific and require domain knowledge. But there are some features that are statistical and don't require domain knowledge. We are going to use those features to solve this classification problem.

Load the X and Y into R and explore the data for yourself.

The data has the shape of $numberOfTrials \times numberOfChannels \times Time$.

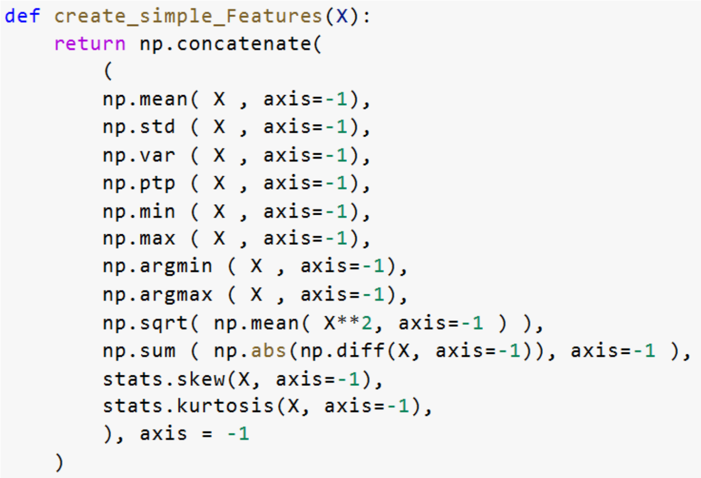
We have 3500 time points which are made out of 7 second of recording 500 Hz.

1. Plot one sample from each group randomly and compare them. (Figure 1 is a sample of the output you should get that has been drawn with matplotlib in Python)



2. Get the first dimension of the sample you plotted (first sensor value) and plot its fast Fourier transform (FFT) and explain what you see. What is the use of these different frequency bands?

3. With the function below (which is written in Python) we created a dataset of simple features from the raw EEG. Run a logistic regression on this data.



```
def create_simple_Features(X):
    return np.concatenate(
        (
        np.mean( X , axis=-1),
        np.std ( X , axis=-1),
        np.var ( X , axis=-1),
        np.ptp ( X , axis=-1),
        np.min ( X , axis=-1),
        np.max ( X , axis=-1),
        np.argmin ( X , axis=-1),
        np.argmax ( X , axis=-1),
        np.sqrt( np.mean( X**2, axis=-1 ) ),
        np.sum ( np.abs(np.diff(X, axis=-1)), axis=-1 ),
        stats.skew(X, axis=-1),
        stats.kurtosis(X, axis=-1),
        ), axis = -1
    )
```

4. Using the backward selection and p-value as the criterion, find a logistic regression model which is the best for predicting.

5. Use the p-value to see which feature is useful in classifying out EEG data. Can you rank the features and tell which one is the most important?

6. Explain what are wavelets and how do they help us deal with time series data.

6 Bitcoin Data (R)

Bitcoin Data has been given to you. The columns are low, high, open, close, high, and volume of a daily candle of bitcoin. Scientists usually try to forecast close prices using these five features but market data is so noisy (Random walk) that makes it almost impossible to forecast. Forecasting cryptocurrencies and stocks can be both a classification or a regression problem, here you experience the latter.

1. Plot the histogram of price return and explain what you see. You can use QQ-plot to back up your theory.

2. Draw ACF and PACF plot for Bitcoin and find out which lags of the price has a significant correlation with the price.

3. Search about what are a stationary time series and its relation with the I component of models like ARIMA.

4. What is the right way of doing cross-validation in time series data? Try to validate the result of the previous part with cross-validation.

5. Do a linear regression and plot the histogram of residuals. Are you seeing white noise?

Required Document

Please upload a file in ZIP format (not RAR) to the elearn platform(`https://elearn5.ut.ac.ir/`).

General Rules

- Please upload a file in ZIP format (~~not RAR~~) to the elearn platform(`https://elearn5.ut.ac.ir/course/view.php?id=14838`).
- You are allowed a total grace period of 3 days to submit late assignments for all of your exercises.
- It is prohibited to use ~~handwritten material~~ and only material produced through typing in the HW template is permissible.
- Utilizing a LaTeXto compose the report will grant an additional 5 points.
- If you submit your exercise in LaTeXformat, please include your complete LaTeXproject as a Zip file along with the pdf and r codes.
- You must make a decision to select and address either one of the two R inquiries (answering both questions will not earn any points).

Deadline

Thursday 23:59. 1402/03/04.

Contact Information

Please direct your questions regarding Homework 3 only to the teaching assistants, Hamid Nemati and Hossein Valipour, through the course mail ((hosseinvalipour96@gmail.com) and (hamid.nemati@ut.ac.ir)). Use ”HW4” as the subject line.

Good Luck