FINTECH545; Spring 2024 Project #02
Brandon Kaplan; (01/27)

Problem 1. (2*pts*)
Given the dataset in problem1.csv:

a) Calculate the first four moments values by using normalized formula in the "Week1 - Univariate Stats".

b) Calculate the first four moments values again by using your chosen statistical package.

c) Is your statistical package functions biased? Prove or disprove your hypothesis respectively. Explain your conclusion.

Solution.

a)

$$\hat{\mu} = 1.0489703904839585$$
$$\hat{\sigma^2} = 5.427220681881727$$
$$\hat{Skew}(X) = (0.8819320922598403 + 0j)$$
$$\hat{Kurtosis}(X) = (23.244253469616176 + 0j)$$

b) For this subproblem I calculated the first four moments using NumPy mean and variance and the Stats subpackage from SciPy. I calculated these using default parameters. They were as follows:

$$\hat{\mu} = 1.0489703904839585$$
$$\hat{\sigma^2} = 5.4217934611998455$$
$$\hat{Skew}(X) = 0.8806086425277364$$
$$\hat{Kurtosis}(X) = 23.122200789989723$$

c) I believed my statistical package(s) to be biased. This is because the sample variance, skewness, and kurtosis I calculated using the normalized formula in the Week1 PDF differed slightly (but significantly) from the values I computed for the same moments using the statistical packages. The differences were as follows:

Difference Between Formula and Biased Variance: 0.005427220681881728

Difference Between Formula and Biased Skewness: (0.001323449732103854+0j)

Difference Between Formula and Biased Kurtoses: (0.12205267962645294+0j)

I then adjusted the default parameters (setting ddof=1 for variance and bias=False for skewness and kurtosis) and calculated the three moments again. This time they were nearly (practically) identical to my moments calculated using the normalized formulas:

SciPy Unbiased Variance: 5.427220681881727

SciPy Unbiased Skewness: 0.8819320922598395

SciPy Unbiased Kurtosis: 23.24425346961619

The differences between the moments calculated using the normalized formulas and the the unbiased moments using the statistical packages were essentially zero:

Difference Between Formula and Unbiased Variances: 0.0

Difference Between Formula and Unbiased Skewness: (7.771561172376096e-16+0j)

Difference Between Formula and Unbiased Kurtoses: (-1.4210854715202004e-14+0j)

The differences between the unbiased moments and the biased moments calculated using the statistical packages was likewise nearly identical to the differences calculated between the moments calculated using the normalized formulas and the first set of moment I calculated using the statistical packages (which I hypothesized were biased):

Difference Between Scipy Calculated Variances: 0.005427220681881728

Difference Between SciPy Calculated Skewnesses: 0.001323449732103077

Difference Between SciPy Calculated Kurtoses: 0.12205267962646715

This was enough evidence to reassure me the default calculations for the second, third, and fourth moments using NumPy and SciPy are indeed biased.

□

Problem 2. (5*pts*)
Assume the multiple linear regression model $Y = X\beta + \epsilon where Y \in \mathbf{R^n}, X \in \mathbf{R^{n \times p}}, \beta \in \mathbf{R^p}, and\, \epsilon \in \mathbf{R^N}$:

a) *Fit the data in problem2.csv using OLS. Then, fit the data using MLE given the assumption of normality. Compare their beta and standard deviation of the OLS error to the fitted MLE* σ. *What's your finding? Explain any differences.*

b) *Fit the data in problem2.csv using MLE given the assumption of a T distribution of errors. Show the fitted parameters. Compare the fitted parameters among MLE under normality assumption and T distribution assumption. Which is the best of fit?*

c) *Fit the data in problem2_x.csv using MLE given* $X = [X_1, X_2]$, *follows the multivariate normal distribution. Assume* X *as a random variable, follows the fitted Gaussian distribution,* X_1 *(problem2_x1.csv) are a part of observed value of* X. *What's the distribution of* X_2 *given each observed value? Plot the expected value along with the 95% confidence interval.*

Solution.

a) I calculated the OLS betas (coefficients) to be -0.087384 and 0.075814 for the constant and x respectively. I calculated the standard error of these coefficients to be 0.071496 and 0.075814 for the constant and x respectively. I calculated the residual sum of squares (RSS) for the OLS solution to be 201.50534975420635 and the standard deviation of the OLS error, $\sigma_{OLS} = 1.008813058320225$. Because under the assumption of normally distributed errors the OLS estimator of β is the MLE of β I suspected that the MLE would return identical coefficients and RSS. The only difference should be in the standard deviation of the errors because the OLS and MLE compute standard deviation of errors slightly differently:

$$\sigma_{OLS} = \sqrt{\frac{RSS}{n-p}}$$

Where n is the number of observations and p is the number of parameters fitted. On the other hand:

$$\sigma_{MLE} = \sqrt{\frac{RSS}{n}}$$

Thus, for a sufficiently large n the $\sigma_{OLS} \approx \sigma_{MLE}$. The MLE would be preferred (even if no OLS assumptions are violated) given the cheaper cost of compute. To be sure, I fit the data using MLE given the assumption of normality by computing the negative log likelihood and then minimizing the function and found the following:

MLE (normal) Betas: [-0.08738446 0.7752741]

MLE (normal) RSS: 201.50534975420635

MLE (normal) Sigma: 1.003756319417732

MLS Sigma if Calculated Using OLS RSS: 1.003756319417732

Evidently, the estimated coefficients are identical as well as the RSS. The standard deviation is slightly smaller as given our sample size is not giant, we were dividing by a significantly bigger denominator to yield a smaller quotient.

b) I fitted the data using MLE assuming the Student's t-distribution of errors analogous to how I fitted the MLE assuming normality in part a). The fitted parameters I returned were:

MLE (t-dist) Betas: [-0.09726583 0.67497755]

MLE (t-dist) Sigma: 0.8550705386331746

MLE (t-dist) Degrees of Freedom: 7.158495188461883

The MLE under the t-distribution assumption returned smaller coefficients as well as a smaller standard deviation. I preformed AIC and BIC tests to get a better sense of goodness of fit for the data. For this I just calculated (positive) log likelihood because I would not be doing any optimization. I found:

AIC MLE (Normal): 575.0751261088556

BIC MLE (Normal): 584.9700782084997

AIC MLE (T): 570.5868066356559

BIC MLE (T): 583.7800761018481

As you can see the AIC and BIC for the data fitted using MLE under the t-distribution assumption returned a smaller AIC and BIC. Since the BIC for MLE (T) is lower than the BIC for MLE (Normal) and BIC penalizes greater for model complexity, this might reassure us that AIC is not simply picking the model with the higher number of parameters and overfitting. Thus the MLE (T) model seems to balance goodness of fit and model complexity better than the MLE (Normal) model when describing our data set. This might be on account of the Student's t-distribution's relative robustness towards outliers/fat tails compared to the normal distribution.

c) The distribution of X_2 given each observed values of X_1 is normal. In a multivariate normal distribution, the conditional distribution of one variable given the other will always be normal as well. I derived the conditional distribution of X_2 by calculating the $\hat{\mu_{X_1}}$, the $\hat{\mu_{X_2}}$, the covariance between X_1 and X_1, the covariance between X_2 and X_2, and the covariance between X_1 and X_2. I used this to calculate the mean and variance of the conditional distribution X_2 for each observation of X_1. To create the 95% confidence interval. I multiplied the standard deviation of the conditional distribution by 1.96 as approximately 95% of observations lie within $\pm 1.96\sigma$ of the mean. See Figure 1 on Page #5 to view the plot.

□

Problem 3. (3*pts*)
Fit the data in problem3.csv using AR (1) through AR (3) and MA (1) through MA (3), respectively. Which is the best of fit?

Solution. I fit the data using the ARIMA model from the Statsmodels package. For each AR(1) through AR(3) and MA(1) through (3) I found the AIC to get a sense of which model was the best of fit. The AIC I found were the following:

AR(1) AIC: 1644.6555047688475
AR(2) AIC: 1581.079265904978
AR(3) AIC: 1436.6598066945867
MA(1) AIC: 1567.4036263707874
MA(2) AIC: 1537.9412063807388
MA(3) AIC: 1536.8677087350309

As you can see the AR(3) model minimizes the AIC relative to the other five models. Given lower AIC generally indicates a better fit, I would argue that AR(3) is the best of fit.
□

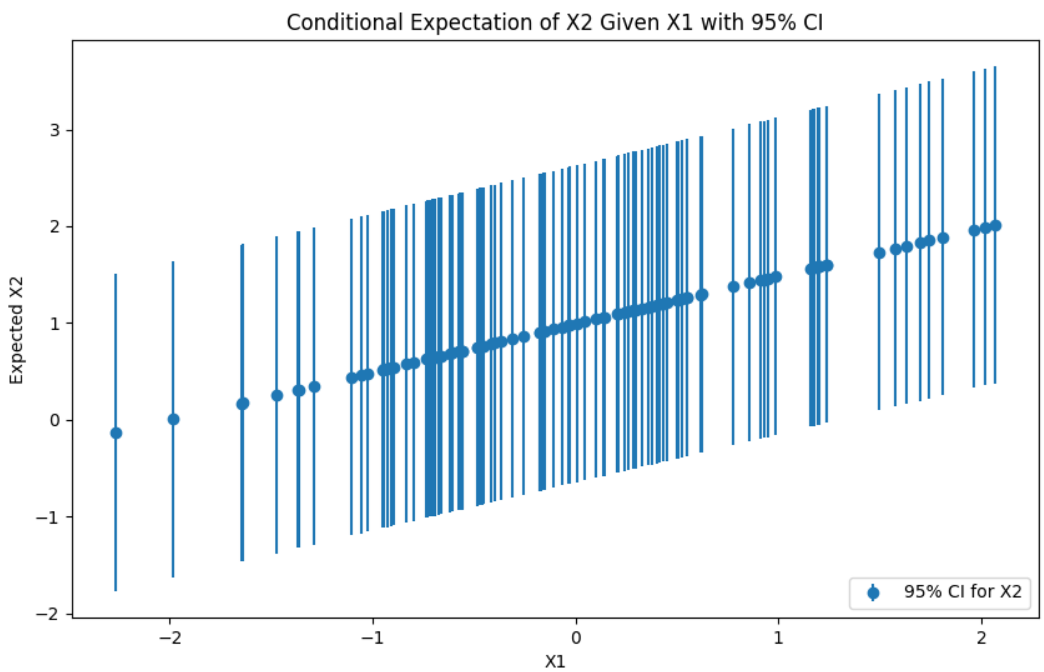

Figure 1: This figure demonstrates the expected values of X_2 given X_1 (blue dots) alongside a 95% confidence interval (blue lines).