Quantitative Methods

共享题干题

【题干】Aaliyah Schultz is a fixed-income portfolio manager at Aries Investments. Schultz supervises Ameris Steele, a junior analyst.A few years ago, Schultz developed a proprietary machine learning (ML) model that aims to predict downgrades of publicly-traded firms by bond rating agencies. The model currently relies only on structured financial data collected from different sources. Schultz thinks the model's predictive power may be improved by incorporating sentiment data derived from textual analysis of news articles and Twitter content relating to the subject companies.Schultz and Steele meet to discuss plans for incorporating the sentiment data into the model. They discuss the differences in the steps between building ML models that use traditional structured data and building ML models that use textual big data. Steele tells Schultz:Statement 1 The second step in building text-based ML models is text preparation and wrangling, whereas the second step in building ML models using structured data is data collection.Statement 2 The fourth step in building both types of models encompasses data/text exploration.Steele expresses concern about using Twitter content in the model, noting that research suggests that as much as 10%–15% of social media content is from fake accounts. Schultz tells Steele that she understands her concern but thinks the potential for model improvement outweighs the concern.Steele begins building a model that combines the structured financial data and the sentiment data. She starts with cleansing and wrangling the raw structured financial data. Exhibit 1 presents a small sample of the raw dataset before cleansing: Each row represents data for a particular firm.

Exhibit 1 Sample of Raw Structured Data Before Cleansing

ID	Ticker	IPO Date	Industry (NAICS)
1	ABC	4/6/17	44
2	BCD	November 15, 2004	52
3	HIJ	26-Jun-74	54
4	KLM	14-Mar-15	72

After cleansing the data, Steele then preprocesses the dataset. She creates two new variables: an "Age" variable based on the firm's IPO date and an "Interest Coverage Ratio" variable equal to EBIT divided by interest expense. She also deletes the "IPO Date" variable from the dataset. After applying these transformations, Steele scales the financial data using normalization. She notes that over the full sample dataset, the "Interest Expense" variable ranges from a minimum of 0.2 and a maximum of 12.2, with a mean of 1.1 and a standard deviation of 0.4.Steele and Schultz then discuss how to preprocess the raw text data. Steele tells Schultz that the process can be completed in the following three steps:Step 1 Cleanse the raw text data.Step 2 Split the cleansed data into a collection of words for them to be normalized.Step 3 Normalize the collection of words from Step 2 and create a distinct set of tokens from the normalized words.With respect to Step 1, Steele tells Schultz:"I believe I

should remove all html tags, punctuations, numbers, and extra white spaces from the data before normalizing them.” After properly cleansing the raw text data, Steele completes Steps 2 and 3. She then performs exploratory data analysis. To assist in feature selection, she wants to create a visualization that shows the most informative words in the dataset based on their term frequency (TF) values. After creating and analyzing the visualization, Steele is concerned that some tokens are likely to be noise features for ML model training; therefore, she wants to remove them.Steele and Schultz discuss the importance of feature selection and feature engineering in ML model training. Steele tells Schultz:“Appropriate feature selection is a key factor in minimizing model overfitting, whereas feature engineering tends to prevent model underfitting.” Once satisfied with the final set of features, Steele selects and runs a model on the training set that classifies the text as having positive sentiment (Class “1” or negative sentiment (Class “0”). She then evaluates its performance using error analysis. The resulting confusion matrix is presented in Exhibit 2.

Exhibit 2 Confusion Matrix

		Actual Training Results	
		Class “1”	Class “0”
Predicted Results	Class “1”	TP = 182	FP = 52
	Class “0”	FN = 31	TN = 96

1.【单项选择题】Which of Steele’s statements relating to the steps in building structured data-based and text-based ML models is correct?

A.Only Statement 1 is correct.

B.Only Statement 2 is correct.

C.Statement 1 and Statement 2 are correct.

参考答案：B

【莽学解析】The five steps in building structured data-based ML models are: 1) conceptualization of the modeling task, 2) data collection, 3) data preparation and wrangling, 4) data exploration, and 5) model training. The five steps in building text-based ML models are: 1) text problem formulation, 2) data (text) curation, 3) text preparation and wrangling, 4) text exploration, and 5) model training. Statement 1 is incorrect: Text preparation and wrangling is the third step in building text ML models and occurs after the second data (text) curation step. Statement 2 is correct: The fourth step in building both types of models encompasses data/text exploration.

2.【单项选择题】Steele’s concern about using Twitter data in the model best relates to:

A.volume.

B.velocity.

C.veracity.

参考答案：C

【莽学解析】Veracity relates to the credibility and reliability of different data sources.

Steele is concerned about the credibility and reliability of Twitter content, noting that research suggests that as much as 10%–15% of social media content is from fake accounts.

3.【单项选择题】What type of error appears to be present in the IPO Date column of Exhibit 1?

A. invalidity error.

B. inconsistency error.

C. non-uniformity error.

参考答案：C

【莽学解析】A non-uniformity error occurs when the data are not presented in an identical format. The data in the "IPO Date" column represent the IPO date of each firm. While all rows are populated with valid dates in the IPO Date column, the dates are presented in different formats (e.g., mm/dd/yyyy, dd/mm/yyyy).

4.【单项选择题】What type of error is most likely present in the last row of data (ID #4) in Exhibit 1?

A. Inconsistency error

B. Incompleteness error

C. Non-uniformity error

参考答案：A

【莽学解析】There appears to be an inconsistency error in the last row (ID #4). An inconsistency error occurs when a data point conflicts with corresponding data points or reality. In the last row, the interest expense data item has a value of 1.5, and the total debt item has a value of 0.0. This appears to be an error: Firms that have interest expense are likely to have debt in their capital structure, so either the interest expense is incorrect or the total debt value is incorrect. Steele should investigate this issue by using alternative data sources to confirm the correct values for these variables.

5.【单项选择题】During the preprocessing of the data in Exhibit 1, what type of data transformation did Steele perform during the data preprocessing step?

A. Extraction

B. Conversion

C. Aggregation

参考答案：A

【莽学解析】During the data preprocessing step, Steele created a new "Age" variable based on the firm's IPO date and then deleted the "IPO Date" variable from the dataset. She also created a new "Interest Coverage Ratio" variable equal to EBIT divided by interest expense. Extraction refers to a data transformation where a new variable is extracted from a current variable for ease of analyzing and using for training an ML model, such as creating an age variable from a date variable or a ratio variable. Steele also performed a selection transformation by deleting the IPO Date variable, which refers to deleting the data columns that are not needed for the project.

6.【单项选择题】Based on Exhibit 1, for the firm with ID #3, Steele should compute the scaled value for the "Interest Expense" variable as:

A. 0.008.

B. 0.083.

C. 0.250.

参考答案：B

【莽学解析】Steele uses normalization to scale the financial data. Normalization is the process of rescaling numeric variables in the range of [0, 1]. To normalize variable X, the minimum value (Xmin) is subtracted from each observation (Xi), and then this value is divided by the difference between the maximum and minimum values of X (Xmax - Xmin):

$$X_{i\,(\text{normalized})} = \frac{X_i - X_{\min}}{X_{\max} - X_{\min}}$$

The firm with ID #3 has an interest expense of 1.2. So, its normalized value is calculated as:

$$X_{i\,(\text{normalized})} = \frac{1.2 - 0.2}{12.2 - 0.2} = 0.083$$

7.【单项选择题】Is Steele's statement regarding Step 1 of the preprocessing of raw text data correct?

A. Yes.

B. No, because her suggested treatment of punctuation is incorrect.

C. No, because her suggested treatment of extra white spaces is incorrect.

参考答案：B

【莽学解析】Although most punctuations are not necessary for text analysis and should be removed, some punctuations (e.g., percentage signs, currency sym-bols, and question marks) may be useful for ML model training. Such punctuations should be substituted with annotations (e.g., /percentSign/, /dollarSign/, and /questionMark/) to preserve their grammatical meaning in the text. Such annotations preserve the semantic meaning of important characters in the text for further text processing and analysis stages.

8.【单项选择题】Steele's Step 2 can be best described as:

A. tokenization.

B. lemmatization.

C. lemmatization.

参考答案：A

【莽学解析】Tokenization is the process of splitting a given text into separate tokens. This step takes place after cleansing the raw text data (removing html tags, numbers, extra white spaces, etc.). The tokens are then normalized to create the bag-of-words (BOW).

9.【单项选择题】The output created in Steele's Step 3 can be best described as a:

A. bag-of-words.

B. set of n-grams.

C. document term matrix.

参考答案：A

【莽学解析】After the cleansed text is normalized, a bag-of-words is created. A bag-of-words

(BOW) is a collection of a distinct set of tokens from all the texts in a sample dataset.

10.【单项选择题】Given her objective, the visualization that Steele should create in the exploratory data analysis step is a:
A. scatter plot.
B. word cloud.
C. document term matrix.
参考答案：B
【莽学解析】Steele wants to create a visualization for Schultz that shows the most informative words in the dataset based on their term frequency (TF, the ratio of the number of times a given token occurs in the dataset to the total number of tokens in the dataset) values. A word cloud is a common visual-ization when working with text data as it can be made to visualize the most informative words and their TF values. The most commonly occurring words in the dataset can be shown by varying font size, and color is used to add more dimensions, such as frequency and length of words.

11.【单项选择题】To address her concern in her exploratory data analysis, Steele should focus on those tokens that have:
A. low chi-square statistics.
B. low mutual information (ML) values.
C. very low and very high term frequency (TF) values.
参考答案：C
【莽学解析】Frequency measures can be used for vocabulary pruning to remove noise features by filtering the tokens with very high and low TF values across all the texts. Noise features are both the most frequent and most sparse (or rare) tokens in the dataset. On one end, noise features can be stop words that are typically present frequently in all the texts across the dataset. On the other end, noise features can be sparse terms that are present in only a few text files. Text classification involves dividing text documents into assigned classes. The frequent tokens strain the ML model to choose a decision boundary among the texts as the terms are present across all the texts (an example of underfitting). The rare tokens mislead the ML model into classifying texts containing the rare terms into a specific class (an example of overfitting). Thus, identifying and removing noise features are critical steps for text classification applications.

12.【单项选择题】Is Steele’s statement regarding the relationship between feature selection/feature engineering and model fit correct?
A. Yes.
B. No, because she is incorrect with respect to feature selection.
C. No, because she is incorrect with respect to feature engineering.
参考答案：A
【莽学解析】A dataset with a small number of features may not carry all the characteristics that explain relationships between the target variable and the features. Conversely, a large number of features can complicate the model and potentially distort patterns in the data due to low degrees of freedom, causing overfitting. Therefore, appropriate feature selection is a key factor in minimizing such model overfitting. Feature engineering tends to prevent underfitting in the training of the model. New features, when engineered properly, can elevate the

underlying data points that better explain the interactions of features. Thus, feature engineering can be critical to overcome underfitting.

13.【单项选择题】Based on Exhibit 2, the model's precision metric is closest to:
A.78%.
B.81%.
C.81%.
参考答案：A
【莽学解析】Precision, the ratio of correctly predicted positive classes (true positives) to all predicted positive classes, is calculated as: Precision (P) = TP/(TP＋FP) = 182/(182＋52) = 0.7778 (78%).

14.【单项选择题】Based on Exhibit 2, the model's F1 score is closest to:
A.77%.
B.81%.
C.85%.
参考答案：B
【莽学解析】The model's F1 score, which is the harmonic mean of precision and recall, is calculated as:F1 score = (2 × P × R)/(P＋R).F1 score = (2 × 0.7778 × 0.8545)/(0.7778＋0.8545) = 0.8143 (81%).

15.【单项选择题】Based on Exhibit 2, the model's accuracy metric is closest to:
A.77%.
B.81%.
C.85%.
参考答案：A
【莽学解析】The model's accuracy, which is the percentage of correctly predicted classes out of total predictions, is calculated as:Accuracy = (TP＋TN)/(TP＋FP＋TN＋FN).Accuracy = (182＋96)/(182＋52＋96＋31) = 0.7701 (77%).

【题干】Espey Jones is examining the relation between the net profit margin (NPM) of companies, in percent, and their fixed asset turnover (FATO). He collected a sample of 35 companies for the most recent fiscal year and fit several different functional forms, settling on the following model:
lnNPM_i = b₀+ b₁FATO_i.
The results of this estimation are provided in Exhibit 1.

16.【单项选择题】The coefficient of determination is closest to:
A.0.0211.
B.0.9789.
C.0.9894.
参考答案：B
【莽学解析】The coefficient of determination is 102.9152 ÷ 105.1303 = 0.9789.

17.【单项选择题】The standard error of the estimate is closest to:

Exhibit 1 Results of Regressing NPM on FATO

Source	df	Sum of Squares	Mean Square	*F*	*p*-Value
Regression	1	102.9152	102.9152	1,486.7079	0.0000
Residual	32	2.2152	0.0692		
Total	33	105.1303			

	Coefficients	Standard Error	*t*- Statistic	*p*-Value
Intercept	0.5987	0.0561	10.6749	0.0000
FATO	0.2951	0.0077	38.5579	0.0000

A. 0.2631.

B. 1.7849.

C. 38.5579

参考答案：A

【莽学解析】The standard error is the square root of the mean square error, or √0.0692=0.2631.

18.【单项选择题】At a 0.01 level of significance, Jones should conclude that:

A. the mean net profit margin is 0.5987%.

B. the variation of the fixed asset turnover explains the variation of the natural log of the net profit margin.

C. a change in the fixed asset turnover from 3 to 4 times is likely to result in a change in the net profit margin of 0.5987%.

参考答案：B

【莽学解析】The p-value corresponding to the slope is less than 0.01, so we reject the null hypothesis of a zero slope, concluding that the fixed asset turnover explains the natural log of the net profit margin.

19.【单项选择题】The predicted net profit margin for a company with a fixed asset turnover of 2 times is closest to:

A. 1.1889%.

B. 1.8043%.

C. 3.2835%

参考答案：C

【莽学解析】The predicted natural log of the net profit margin is 0.5987＋(2 × 0.2951) = 1.1889. The predicted net profit margin is $e^{1.1889}$ = 3.2835%.

【题干】Angela Martinez, an energy sector analyst at an investment bank, is concerned about the future level of oil prices and how it might affect portfolio values. She is considering whether to recommend a hedge for the bank portfolio's exposure to changes in oil prices. Martinez examines West Texas Intermediate (WTI) monthly crude oil price data, expressed in US dollars per barrel, for the 181-month period from August 2000 through August 2015. The end-of-month WTI

oil price was $51.16 in July 2015 and $42.86 in August 2015 (Month 181).

After reviewing the time-series data, Martinez determines that the mean and variance of the time series of oil prices are not constant over time. She then runs the following four regressions using the WTI time-series data.

Linear trend model: Oil pricet = b0 + b1t + et

Log-linear trend model: ln Oil pricet = b0 + b1t + et

AR(1) model: Oil pricet = b0 + b1Oil pricet-1 + et

AR(2) model: Oil pricet = b0 + b1Oil pricet-1 + b2Oil pricet-2 + et

Exhibit 1 presents selected data from all four regressions, and Exhibit 2 presents selected autocorrelation data from the AR(1) models.

Exhibit 1. Crude Oil Price per Barrel, August 2000–August 2015

	Regression Statistics (t-statistics for coefficients are reported in parentheses)			
	Linear	**Log-Linear**	**AR(1)**	**AR(2)**
R^2	0.5703	0.6255	0.9583	0.9656
Standard error	18.6327	0.3034	5.7977	5.2799
Observations	181	181	180	179
Durbin-Watson	0.10	0.08	1.16	2.08
RMSE			2.0787	2.0530
Coefficients:				
Intercept	28.3278	3.3929	1.5948	2.0017
	(10.1846)	(74.9091)	(1.4610)	(1.9957)
t (Trend)	0.4086	0.0075		
	(15.4148)	(17.2898)		
Oil Price$_{t-1}$			0.9767	1.3946
			(63.9535)	(20.2999)
Oil Price$_{t-2}$				-0.4249
				(-6.2064)

In Exhibit 1, at the 5% significance level, the lower critical value for the Durbin-Watson test statistic is 1.75 for both the linear and log-linear regressions.

Exhibit 2. Autocorrelations of the Residual from AR(1) Model

Lag	Autocorrelation	t-Statistic
1	0.4157	5.5768
2	0.2388	3.2045
3	0.0336	0.4512
4	-0.0426	-0.5712

Note: At the 5% significance level, the critical value for at-statistic is 1.97.

After reviewing the data and regression results, Martinez draws the following conclusions.Conclusion 1: The time series for WTI oil prices is covariance stationary.Conclusion 2: Out-of-sample forecasting using the AR(1) model appears to be more accurate than that of the AR(2) model.

20.【单项选择题】Based on Exhibit 1, the predicted WTI oil price for October 2015 using the linear trend model is closest to:

A.$29.15.

B.$74.77.

C.$103.10.

参考答案：C

【莽学解析】The predicted value for period t from a linear trend is calculated as

$$\hat{y}_t = \hat{b}_0 + \hat{b}_1(t).$$

October 2015 is the second month out of sample, or t = 183. So, the predicted value for October 2015 is calculated as

$$\hat{y}_t = 28.3278 + 0.4086(183) = \$103.10.$$

Therefore, the predicted WTI oil price for October 2015 based on the linear trend model is $103.10.

21.【单项选择题】Based on Exhibit 1, the predicted WTI oil price for September 2015 using the log-linear trend model is closest to:

A.$29.75.

B.$29.98.

C.$116.50.

参考答案：C

【莽学解析】The predicted value for period t from a log-linear trend is calculated as

$$\ln\hat{y}_t = \hat{b}_0 + \hat{b}_1(t).$$

September 2015 is the first month out of sample, or t = 182. So, the predicted value for September 2015 is calculated as follows:

$$\ln\hat{y}_t = 3.3929 + 0.0075(182)$$

$$\ln\hat{y}_t = 4.7579$$

$$\hat{y}_t = e^{4.7579} = \$116.50$$

Therefore, the predicted WTI oil price for September 2015, based on the log-linear trend model, is $116.50.

22.【单项选择题】Based on the regression output in Exhibit 1, there is evidence of positive serial correlation in the errors in:

A. the linear trend model but not the log-linear trend model.
B. both the linear trend model and the log-linear trend model.
C. neither the linear trend model nor the log-linear trend model.
参考答案：B
【莽学解析】The Durbin-Watson statistic for the linear trend model is 0.10 and for the log-linear trend model is 0.08. Both of these values are below the critical value of 1.75. Therefore, we can reject the hypothesis of no positive serial correlation in the regression errors in both the linear trend model and the log-linear trend model.

23.【单项选择题】Martinez's Conclusion 1 is:
A. correct.
B. incorrect because the mean and variance of WTI oil prices are not constant over time.
C. incorrect because the Durbin-Watson statistic of the AR(2) model is greater than 1.75.
参考答案：B
【莽学解析】There are three requirements for a time series to be covariance stationary. First, the expected value of the time series must be constant and finite in all periods. Second, the variance of the time series must be constant and finite in all periods. Third, the covariance of the time series with itself for a fixed number of periods in the past or future must be constant and finite in all periods. Martinez concludes that the mean and variance of the time series of WTI oil prices are not constant over time. Therefore, the time series is not covariance stationary.

24.【单项选择题】Based on Exhibit 1, the forecasted oil price in September 2015 based on the AR(2) model is closest to:
A. $38.03.
B. $40.04.
C. $61.77.
参考答案：B
【莽学解析】The last two observations in the WTI time series are July and August 2015, when the WTI oil price was $51.16 and $42.86, respectively. Therefore, September 2015 represents a one-period-ahead forecast. The one-period-ahead forecast from an AR(2) model is calculated as

$$\hat{x}_{t+1} = \hat{b}_0 + \hat{b}_1 x_t + \hat{b}_2 x_{t-1}$$

So, the one-period-ahead (September 2015) forecast is calculated as

$$\hat{x}_{t+1} = 2.0017 + 1.3946(\$42.86) - 0.4249(\$51.16) = \$40.04$$

Therefore, the September 2015 forecast based on the AR(2) model is $40.04.

25.【单项选择题】Based on the data for the AR(1) model in Exhibits 1 and 2, Martinez can conclude that the:
A. residuals are not serially correlated.
B. autocorrelations do not differ significantly from zero.
C. standard error for each of the autocorrelations is 0.0745.
参考答案：C

【莽学解析】

The standard error of the autocorrelations is calculated as $\frac{1}{\sqrt{T}}$, where T represents the number of observations used in the regression. Therefore, the standard error for each of the autocorrelations is $\frac{1}{\sqrt{180}} = 0.0745$. Martinez can conclude that the residuals are serially correlated and are significantly different from zero because two of the four autocorrelations in Exhibit 2 have a t-statistic in absolute value that is greater than the critical value of 1.97. Choices A and B are incorrect because two of the four autocorrelations have a t-statistic in absolute value that is greater than the critical value of the t-statistic of 1.97.

26.【单项选择题】Based on the mean-reverting level implied by the AR(1) model regression output in Exhibit 1, the forecasted oil price for September 2015 is most likely to be:

A. less than $42.86.

B. equal to $42.86.

C. greater than $42.86.

参考答案：C

【莽学解析】The mean-reverting level from the AR(1) model is calculated as

$$\hat{x}_t = \frac{b_0}{1-b_1} = \frac{1.5948}{1-0.9767} = \$68.45$$

Therefore, the mean-reverting WTI oil price from the AR(1) model is $68.45. The forecasted oil price in September 2015 will likely be greater than $42.86 because the model predicts that the price will rise in the next period from the August 2015 price of $42.86.

【题干】Doris Honoré is a securities analyst with a large wealth management firm. She and her colleague Bill Smith are addressing three research topics: how investment fund characteristics affect fund total returns, whether a fund rating system helps predict fund returns, and whether stock and bond market returns explain the returns of a portfolio of utility shares run by the firm.

To explore the first topic, Honoré decides to study US mutual funds using a sample of 555 large-cap US equity funds. The sample includes funds in style classes of value, growth, and blend (i.e., combining value and growth characteristics). The dependent variable is the average annualized rate of return (in percent) over the past five years. The independent variables are fund expense ratio, portfolio turnover, the natural logarithm of fund size, fund age, and three dummy variables. The multiple manager dummy variable has a value of 1 if the fund has multiple managers (and a value of 0 if it has a single manager). The fund style is indicated by a growth dummy (value of 1 for growth funds and 0 otherwise) and a blend dummy (value of 1 for blend funds and 0 otherwise). If the growth and blend dummies are both zero, the fund is a value fund. The regression output is given in Exhibit 1.

Exhibit 1 Multiple Regression Output for Large-Cap Mutual Fund Sam

	Coefficient	Standard Error	*t*-Stati
Intercept	10.9375	1.3578	8.05
Expense ratio (%)	−1.4839	0.2282	−6.50
Portfolio turnover (%)	0.0017	0.0016	1.07
ln (fund size in $)	0.1467	0.0612	2.39
Manager tenure (years)	−0.0098	0.0102	−0.95
Multiple manager dummy	0.0628	0.1533	0.41
Fund age (years)	−0.0123	0.0047	−2.62
Growth dummy	2.4368	0.1886	12.91
Blend dummy	0.5757	0.1881	3.06

ANOVA	df	SS	MSS
Regression	8	714.169	89.27
Residual	546	1583.113	2.89
Total	554	2297.282	
Multiple R	0.5576		
R^2	0.3109		
Adjusted R^2	0.3008		
Standard error (%)	1.7028		
Observations	555		

Based on the results shown in Exhibit 1, Honoré wants to test the hypothesis that all of the regression coefficients are equal to zero. For the 555 fund sample, she also wants to compare the performance of growth funds with the value funds.

Honoré is concerned about the possible presence of multicollinearity in the regression. She states that adding a new independent variable that is highly correlated with one or more independent variables already in the regression model, has three potential consequences:

1. The R2 is expected to decline.
2. The regression coefficient estimates can become imprecise and unreliable.
3. The standard errors for some or all of the regression coefficients will become inflated.

Another concern for the regression model (in Exhibit 1) is conditional heteroskedasticity.

Honoré is concerned that the presence of heteroskedasticity can cause both the F-test for the overall significance of the regression and the t-tests for significance of individual regression coefficients to be unreliable. She runs a regression of the squared residuals from the model in Exhibit 1 on the eight independent variables and finds the R2 is 0.0669.

As a second research project, Honoré wants to test whether including Morningstar's rating system, which assigns a one- through five-star rating to a fund, as an independent variable will improve the predictive power of the regression model. To do this, she needs to examine whether values of the independent variables in a given period predict fund return in the next period. Smith suggests three different methods of adding the Morningstar ratings to the model:

Method 1: Add an independent variable that has a value equal to the number of stars in the rating of each fund.

Method 2: Add five dummy variables, one for each rating.

Method 3: Add dummy variables for four of the five ratings.

As a third research project, Honoré wants to establish whether bond market returns (proxied by returns of long-term US Treasuries) and stock market returns (proxied by returns of the S&P 500 Index) explain the returns of a portfolio of utility stocks being recommended to clients. Exhibit 2 presents the results of a regression of 10 years of monthly percentage total returns for the utility portfolio on monthly total returns for US Treasuries and the S&P 500.

Exhibit 2. Regression Analysis of Utility Portfolio Returns

	Coefficient	Standard Error	t-Statistic	p-Value
Intercept	-0.0851	0.2829	-0.3008	0.7641
US Treasury	0.4194	0.0848	4.9474	<0.0001
S&P 500	0.6198	0.0666	9.3126	<0.0001

ANOVA	df	SS	MSS	*F*	Significance *F*
Regression	2	827.48	413.74	46.28	<0.0001
Residual	117	1045.93	8.94		
Total	119	1873.41			

Multiple R	0.6646
R^2	0.4417
Adjusted R^2	0.4322
Standard error (%)	2.99
Observations	120

For the time-series model in Exhibit 2, Honoré says that positive serial correlation would not require that the estimated coefficients be adjusted, but that the standard errors of the regression coefficients would be underestimated. This issue would cause the t-statistics of the regression coefficients to be inflated. Honoré tests the null hypothesis that the there is no serial correlation in the regression residuals and finds that the Durbin–Watson statistic is

equal to 1.81. The critical values at the 0.05 significance level for the Durbin－Watson statistic are dl = 1.63 and du = 1.72.

Smith asks whether Honoré should have estimated the models in Exhibit 1 and Exhibit 2 using a probit or logit model instead of using a traditional regression analysis.

27.【单项选择题】Considering Exhibit 1, the F-statistic is closest to:

A. 3.22.

B. 8.06.

C. 30.79.

参考答案：C

【莽学解析】

$$\text{The F-statistic is } F=\frac{RSS/k}{SSE/[n-(k+1)]}=\frac{714.169/8}{1583.113/546}=\frac{89.2712}{2.8995}=30.79$$

Because F = 30.79 exceeds the critical F of 1.96, the null hypothesis that the regression coefficients are all 0 is rejected at the 0.05 significance level.

28.【单项选择题】Based on Exhibit 1, the difference between the predicted annualized returns of a growth fund and an otherwise similar value fund is closest to:

A. 1.86%.

B. 2.44%.

C. 3.01%.

参考答案：B

【莽学解析】The estimated coefficients for the dummy variables show the estimated difference between the returns on different types of funds. The growth dummy takes the value of 1 for growth funds and 0 for the value fund. Exhibit 1 shows a growth dummy coefficient of 2.4368. The estimated difference between the return of growth funds and value funds is thus 2.4368.

29.【单项选择题】Honoré describes three potential consequences of multicollinearity. Are all three consequences correct?

A. Yes

B. No, 1 is incorrect

C. No, 2 is incorrect

参考答案：B

【莽学解析】The R^2 is expected to increase, not decline, with a new independent variable. The other two potential consequences Honore describes are correct.

30.【单项选择题】Which of the three methods suggested by Smith would best capture the ability of the Morningstar rating system to predict mutual fund performance?

A. Method 1

B. Method 2

C. Method 3

参考答案：C

【莽学解析】Using dummy variables to distinguish among n categories would best capture the ability of the Morningstar rating system to predict mutual fund performance. We need n - 1 dummy variables to distinguish among n categories. In this case, there are five possible ratings, and we need four dummy variables. Adding an independent variable that has a value equal to the number of stars in the rating of each fund is not appropriate because if the coefficient for this variable is positive, this method assumes that the extra return for a two-star fund is twice that of a one-star fund, the extra return for a three-star fund is three times that of a one-star fund, and so forth, which is not a reasonable assumption.

31.【单项选择题】Honoré is concerned about the consequences of heteroskedasticity. Is she correct regarding the effect of heteroskedasticity on the reliability of the F-test and t-tests?

A. Yes

B. No, she is incorrect with regard to the F-test

C. No, she is incorrect with regard to the t-tests

参考答案：A

【莽学解析】Heteroskedasticity causes the F-test for the overall significance of the regression to be unreliable. It also causes the t-tests for the significance of individual regression coefficients to be unreliable because heteroskedasticity introduces bias into estimators of the standard error of regression coefficients.

32.【单项选择题】Is Honoré's description of the effects of positive serial correlation (in Exhibit 2) correct regarding the estimated coefficients and the standard errors?

A. Yes

B. No, she is incorrect about only the estimated coefficients

C. No, she is incorrect about only the standard errors of the regression coefficients.

参考答案：A

【莽学解析】The model in Exhibit 2 does not have a lagged dependent variable. Positive serial correlation will, for such a model, not affect the consistency of the estimated coefficients. Thus, the coefficients will not need to be corrected for serial correlation. Positive serial correlation will, however, cause the standard errors of the regression coefficients to be understated; thus, the corresponding t-statistics will be inflated.

33.【单项选择题】Based on her estimated Durbin-Watson statistic, Honoré should:

A. fail to reject the null hypothesis.

B. reject the null hypothesis because there is significant positive serial correlation.

C. reject the null hypothesis because there is significant negative serial correlation.

参考答案：A

【莽学解析】The critical Durbin-Watson (D-W) values are d_l = 1.63 and d_u = 1.72. Because the estimated D-W value of 1.81 is greater than d_u = 1.73 (and less than 2), she fails to reject the null hypothesis of no serial correlation.

34.【单项选择题】Should Honoré have estimated the models in Exhibit 1 and Exhibit 2 using probit or logit models instead of traditional regression analysis?

A. Both should be estimated with probit or logit models.

B. Neither should be estimated with probit or logit models.

C. Only the analysis in Exhibit 1 should be done with probit or logit models.

参考答案：B

【莽学解析】Probit and logit models are used for models with qualitative dependent variables, such as models in which the dependent variable can have one of two discrete outcomes (i.e., 0 or 1). The analysis in the two exhibits are explaining security returns, which are continuous (not 0 or 1) variables.

【题干】Adele Chiesa is a money manager for the Bianco Fund. She is interested in recent findings showing that certain business condition variables predict excess US stock market returns (one-month market return minus one-month T-bill return). She is also familiar with evidence showing how US stock market returns differ by the political party affiliation of the US President. Chiesa estimates a multiple regression model to predict monthly excess stock market returns accounting for business conditions and the political party affiliation of the US President:Excess stock market returnt= a₀+ a₁Default spread_{t－1}+ a₂Term spread_{t－1}+ a₃Pres party dummy_{t－1}+ e_tDefault spread is equal to the yield on Baa bonds minus the yield on Aaa bonds. Term spread is equal to the yield on a 10-year constant-maturity US Treasury index minus the yield on a 1-year constant-maturity US Treasury index. Pres party dummy is equal to 1 if the US President is a member of the Democratic Party and 0 if a member of the Republican Party.Chiesa collects 432 months of data (all data are in percent form, i.e., 0.01 = 1 percent). The regression is estimated with 431 observations because the independent variables are lagged one month. The regression output is in Exhibit 1. Exhibits 2 through 5 contain critical values for selected test

Exhibit 1. Multiple Regression Output (the Dependent Variable Is the One-Month Market Return in Excess of the One-Month T-Bill Return)

	Coefficient	*t*-Statistic	*p*-Value
Intercept	−4.60	−4.36	<0.01
Default spread$_{t-1}$	3.04	4.52	<0.01
Term spread$_{t-1}$	0.84	3.41	<0.01
Pres party dummy$_{t-1}$	3.17	4.97	<0.01
Number of observations		431	
Test statistic from Breusch–Pagan (BP) test		7.35	
R^2		0.053	
Adjusted R^2		0.046	
Durbin–Watson (DW)		1.65	
Sum of squared errors (SSE)		19,048	
Regression sum of squares (SSR)		1,071	

An intern working for Chiesa has a number of questions about the results in Exhibit 1:Question 1: How do you test to determine whether the overall regression model is significant?Question 2: Does the estimated model conform to standard regression assumptions? For instance, is the error term serially correlated, or is there conditional heteroskedasticity?Question 3: How do you interpret the coefficient for the Pres party dummy variable?Question 4: Default spread appears to be quite important. Is there some way to assess the precision of its estimated coefficient? What is the economic interpretation of this variable?After responding to her intern's questions, Chiesa concludes with the following statement: "Predictions from Exhibit 1 are subject to parameter estimate uncertainty, but not regression model uncertainty."

Exhibit 2. Critical Values for the Durbin–Watson Statistic ($\alpha = 0.05$)

	$K = 3$	
N	d_l	d_u
420	1.825	1.854
430	1.827	1.855
440	1.829	1.857

Exhibit 3. Table of the Student's *t*-Distribution (One-Tailed Probabilities for df = ∞)

P	*t*
0.10	1.282
0.05	1.645
0.025	1.960
0.01	2.326
0.005	2.576

Exhibit 4. Values of χ^2

	Probability in Right Tail			
df	0.975	0.95	0.05	0.025
1	0.0001	0.0039	3.841	5.024
2	0.0506	0.1026	5.991	7.378
3	0.2158	0.3518	7.815	9.348
4	0.4840	0.7110	9.488	11.14

Exhibit 5. Table of the *F*-Distribution (Critical Values for Right-Hand Tail Area Equal to 0.05) Numerator: df1 and Denominator: df2

	df1				
df2	1	2	3	4	427
1	161	200	216	225	254
2	18.51	19.00	19.16	19.25	19.49
3	10.13	9.55	9.28	9.12	8.53
4	7.71	6.94	6.59	6.39	5.64
427	3.86	3.02	2.63	2.39	1.17

35.【单项选择题】Regarding the intern's Question 1, is the regression model as a whole significant at the 0.05 level?

A. No, because the calculated F-statistic is less than the critical value for F.

B. Yes, because the calculated F-statistic is greater than the critical value for F.

C. Yes, because the calculated x^2 statistic is greater than the critical value for x^2.

参考答案：B

【莽学解析】The F-test is used to determine if the regression model as a whole is significant. F = Mean square regression (MSR) ÷ Mean squared error (MSE) MSE = SSE/[n - (k + 1)] = 19,048 ÷ 427 = 44.60 MSR = RSS/k = 1071 ÷ 3 = 357 F = 357 ÷ 44.60 = 8.004

The critical value for degrees of freedom of 3 and 427 with α = 0.05 (one-tail) is F = 2.63 from Exhibit 5. The calculated F is greater than the critical value, and Chiesa should reject the null hypothesis that all regression coefficients are equal to zero.

36.【单项选择题】Which of the following is Chiesa's best response to Question 2 regarding

serial correlation in the error term? At a 0.05 level of significance, the test for serial correlation indicates that there is:c

A. no serial correlation in the error term.

B. positive serial correlation in the error term.

C. negative serial correlation in the error term.

参考答案：B

【莽学解析】The Durbin-Watson test used to test for serial correlation in the error term, and its value reported in Exhibit 1 is 1.65. For no serial correlation, DW is approximately equal to 2. If DW $<$ d_1, the error terms are positively serially correlated. Because the DW = 1.65 is less than d_1 = 1.827 for n = 431 (see Exhibit 2), Chiesa should reject the null hypothesis of no serial correlation and conclude that there is evidence of positive serial correlation among the error terms.

37.【单项选择题】Regarding Question 3, the Pres party dummy variable in the model indicates that the mean monthly value for the excess stock market return is:

A. 1.43 percent larger during Democratic presidencies than Republican presidencies.

B. 3.17 percent larger during Democratic presidencies than Republican presidencies.

C. 3.17 percent larger during Republican presidencies than Democratic presidencies.

参考答案：B

【莽学解析】The coefficient for the Pres party dummy variable (3.17) represents the increment in the mean value of the dependent variable related to the Democratic Party holding the presidency. In this case, the excess stock market return is 3.17 percent greater in Democratic presidencies than in Republican presidencies.

38.【单项选择题】In response to Question 4, the 95 percent confidence interval for the regression coefficient for the default spread is closest to:

A. 0.13 to 5.95.

B. 1.72 to 4.36.

C. 1.93 to 4.15.

参考答案：B

【莽学解析】The confidence interval is computed as a1 $\pm$ $s(a_1)$ $\times$ $t(95\%, \infty)$. From Exhibit 1, a_1 = 3.04 and $t(a_1)$ = 4.52, resulting in a standard error of a_1 = $s(a_1)$ = 3.04/4.52 = 0.673. The critical value for t from Exhibit 3 is 1.96 for p = 0.025. The confidence interval for a_1 is 3.04 $\pm$ 0.673 $\times$ 1.96 = 3.04 $\pm$ 1.31908 or from 1.72092 to 4.35908.

39.【单项选择题】With respect to the default spread, the estimated model indicates that when business conditions are:

A. strong, expected excess returns will be higher.

B. weak, expected excess returns will be lower.

C. weak, expected excess returns will be higher.

参考答案：C

【莽学解析】The default spread is typically larger when business conditions are poor, i.e., a greater probability of default by the borrower. The positive sign for default spread (see Exhibit 1) indicates that expected returns are positively related to default spreads, meaning that excess returns are greater when business conditions are poor.

40.【单项选择题】Is Chiesa's concluding statement correct regarding parameter estimate uncertainty and regression model uncertainty?

A. Yes.

B. No, predictions are not subject to parameter estimate uncertainty.

C. No, predictions are subject to regression model uncertainty and parameter estimate uncertainty.

参考答案：C

【莽学解析】Predictions in a multiple regression model are subject to both parameter estimate uncertainty and regression model uncertainty.

【题干】Doug Abitbol is a portfolio manager for Polyi Investments, a hedge fund that trades in the United States. Abitbol manages the hedge fund with the help of Robert Olabudo, a junior portfolio manager.

Abitbol looks at economists' inflation forecasts and would like to examine the relationship between the US Consumer Price Index (US CPI) consensus forecast and the actual US CPI using regression analysis. Olabudo estimates regression coefficients to test whether the consensus forecast is unbiased. If the consensus forecasts are unbiased, the intercept should be 0.0 and the slope will be equal to 1.0. Regression results are presented in Exhibit 1. Additionally, Olabudo calculates the 95% prediction interval of the actual CPI using a US CPI consensus forecast of 2.8.

Exhibit 1 Regression Output: Estimating US CPI

Regression Statistics

R^2	0.9859
Standard error of estimate	0.0009
Observations	60

	Coefficients	Standard Error	*t*-Statistic
Intercept	0.0001	0.0002	0.5000
US CPI consensus forecast	0.9830	0.0155	63.4194

Notes:

1 The absolute value of the critical value for the *t*-statistic is 2.002 at the 5% level of significance.

2 The standard deviation of the US CPI consensus forecast is $s_x = 0.7539$.

3 The mean of the US CPI consensus forecast is $\bar{X} = 1.3350$.

Finally, Abitbol and Olabudo discuss the forecast and forecast interval:

Observation 1: For a given confidence level, the forecast interval is the same no matter the US CPI consensus forecast.

Observation 2: A larger standard error of the estimate will result in a wider confidence interval.

41.【单项选择题】Based on Exhibit 1, Olabudo should:

A. conclude that the inflation predictions are unbiased.
B. reject the null hypothesis that the slope coefficient equals one.
C. reject the null hypothesis that the intercept coefficient equals zero.
参考答案：A
【莽学解析】We fail to reject the null hypothesis of a slope equal to one, and we fail to reject the null hypothesis of an intercept equal to zero. The test of the slope equal to 1.0 is t=(0.9830-1.0000)/0.0155=-1.09766. The test of the intercept equal to 0.0 is t=(0.0001-0.0000)/0.0002=0.5000. Therefore, we conclude that the forecasts are unbiased.

42.【单项选择题】Based on Exhibit 1, Olabudo should calculate a prediction interval for the actual US CPI closest to:
A. 2.7506 to 2.7544.
B. 2.7521 to 2.7529.
C. 2.7981 to 2.8019.
参考答案：A
【莽学解析】The forecast interval for inflation is calculated in three steps:Step 1 - Make the prediction given the US CPI forecast of 2.8:

Step 2 - Compute the variance of the prediction error:

$$s_f^2 = s_e^2\left\{1 + (1/n) + \left[(X_f - \bar{X})^2\right] / \left[(n-1) \times s_x^2\right]\right\}.$$

$$s_f^2 = 0.0009^2\left\{1 + (1/60) + \left[(2.8 - 1.3350)^2\right] / \left[(60-1) \times 0.75\right.\right.$$

$$s_f^2 = 0.00000088.$$

$$s_f = 0.0009.$$

Step 3 - Compute the prediction interval:

$$\hat{Y} \pm t_c \times s_f$$

2.7525 ± (2.0 x 0.0009)2.7525 - (2.0 x 0.0009) = 2.7506; lower bound2.7525＋(2.0 x 0.0009) = 2.7544; upper boundSo, given the US CPI forecast of 2.8, the 95% prediction interval is 2.7506 to 2.7544.

43.【单项选择题】Which of Olabudo’s observations of forecasting is correct?
A. Only Observation 1.
B. Only Observation 2

C. Both Observation 1 and Observations 2.

参考答案：B

【莽学解析】The confidence level influences the width of the forecast interval through the critical t-value that is used to calculate the distance from the forecasted value: The larger the confidence level, the wider the interval.

Therefore, Observation 1 is not correct.

Observation 2 is correct. The greater the standard error of the estimate, the greater the standard error of the forecast.

【题干】Kenneth McCoin, CFA, is a challenging interviewer. Last year, he handed each job applicant a sheet of paper with the information in the following table, and he then asked several questions about regression analysis. Some of McCoin's questions, along with a sample of the answers he received to each, are given below. McCoin told the applicants that the independent variable is the ratio of net income to sales for restaurants with a market cap of more than $100 million and the dependent variable is the ratio of cash flow from operations to sales for those restaurants. Which of the choices provided is the best answer to each of McCoin's questions?

44.【单项选择题】The coefficient of determination is closest to:

A. 0.7436.

B. 0.8261.

C. 0.8623.

参考答案：A

【莽学解析】The coefficient of determination is the same as R^2, which is 0.7436 in the table.

45.【单项选择题】The correlation between X and Y is closest to:

A. -0.7436.

B. 0.7436.

C. 0.8623.

参考答案：C

【莽学解析】Because the slope is positive, the correlation between X and Y is simply the square root of the coefficient of determination: $\sqrt{0.7436}=0.8623$.

46.【单项选择题】If the ratio of net income to sales for a restaurant is 5%, the predicted ratio of cash flow from operations (CFO) to sales is closest to:

A. -4.054.

B. 0.524.

C. 4.207.

参考答案：C

【莽学解析】To make a prediction using the regression model, multiply the slope coefficient by the forecast of the independent variable and add the result to the intercept. Expected value of CFO to sales = 0.077 + (0.826 × 5) = 4.207.

47.【单项选择题】Is the relationship between the ratio of cash flow to operations and the ratio of net income to sales significant at the 0.05 level?

Regression Statistics	
R^2	0.7436
Standard error	0.0213
Observations	24

Source	df	Sum of Squares	Mean Square	F	p-V
Regression	1	0.029	0.029000	63.81	
Residual	22	0.010	0.000455		
Total	23	0.040			

	Coefficients	Standard Error	t-Statistic
Intercept	0.077	0.007	11.328
Net income to sales (%)	0.826	0.103	7.988

A. No, because the R^2 is greater than 0.05.
B. No, because the p-values of the intercept and slope are less than 0.05.
C. Yes, because the p-values for F and t for the slope coefficient are less than 0.05.
参考答案：C
【莽学解析】The p-value is the smallest level of significance at which the null hypotheses concerning the slope coefficient can be rejected. In this case, the p-value is less than 0.05, and thus the regression of the ratio of cash flow from operations to sales on the ratio of net income to sales is significant at the 5% level.

【题干】Brad Varden, a junior analyst at an actively managed mutual fund, is responsible for research on a subset of the 500 large-cap equities the fund follows. Recently, the fund has been paying close attention to management turnover and to publicly available environmental, social, and governance (ESG) ratings. Varden is given the task of investigating whether any significant relationship exists between a company's profitability and either of these two characteristics. Colleen Quinni, a senior analyst at the fund, suggests that as an initial step

in his investigation, Varden should perform a multiple regression analysis on the variables and report back to her.Varden knows that Quinni is an expert at quantitative research, and she once told Varden that after you get an idea, you should formulate a hypothesis, test the hypothesis, and analyze the results. Varden expects to find that ESG rating is negatively related to ROE and CEO tenure is positively related to ROE. He considers a relationship meaningful when it is statistically significant at the 0.05 level. To begin, Varden collects values for ROE, CEO tenure, and ESG rating for a sample of 40 companies from the large-cap security universe. He performs a multiple regression with ROE (in percent) as the dependent variable and ESG rating and CEO tenure (in years) as the independent variables: $Y_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + \varepsilon_i$.Exhibit 1 shows the regression results.

DF Associates is one of the companies Varden follows. He wants to predict its ROE using his regression model. DF Associates' corporate ESG rating is 55, and the company's CEO has been in that position for 10.5 years.Varden also wants to check on the relationship between these variables and the dividend growth rate (divgr), so he completes the correlation matrix shown in Exhibit 2.

Exhibit 2. Correlation Matrix

	ROE	ESG	Tenure	Divgr
ROE	1.0			
ESG	0.446	1.0		
Tenure	0.369	0.091	1.0	
Divgr	0.117	0.046	0.028	1.0

Investigating further, Varden determines that dividend growth is not a linear combination of CEO tenure and ESG rating. He is unclear about how additional independent variables would affect the significance of the regression, so he asks Quinni, "Given this correlation matrix, will both R^2 and adjusted R^2 automatically increase if I add dividend growth as a third independent variable?"The

Exhibit 1. Regression Statistics					
$\hat{Y}_i = 9.442 + 0.069X_{1i} + 0.681X_{2i}$					
	Coefficient	**Standard Error**		**t-Statistic**	**p-Value**
Intercept	9.442	3.343		2.824	0.008
b_1 (ESG variable)	0.069	0.058		1.201	0.238
b_2 (Tenure variable)	0.681	0.295		2.308	0.027
ANOVA	**df**	**SS**	**MSS**	**F**	**Significance F**
Regression	2	240.410	120.205	4.161	0.023
Residual	37	1069.000	28.892		
Total	39	1309.410			
Multiple R	0.428				
R^2	0.183				
Adjusted R^2	0.139				
Standard error (%)	5.375				
Observations	40				

discussion continues, and Quinni asks two questions.1. What does your F-statistic of 4.161 tell you about the regression?2. In interpreting the overall significance of your regression model, which statistic do you believe is most relevant: R
2
, adjusted R
2
, or the F-statistic?Varden answers both questions correctly and says he wants to check two more ideas. He believes the following:1. ROE is less correlated with the dividend growth rate in firms whose CEO has been in office more than 15 years, and2. CEO tenure is a normally distributed random variable.Later, Varden includes the dividend growth rate as a third independent variable and runs the regression on the fund's entire group of 500 large-cap equities. He finds that the adjusted R
2
 is much higher than the results in Exhibit 1. He reports this to Quinni and says, “Adding the dividend growth rate gives a model with a higher adjusted R
2
. The three-variable model is clearly better.” Quinni cautions, "I don't think you can conclude that yet.”

48.【单项选择题】Based on Exhibit 1 and given Varden's expectations, which is the best null hypothesis and conclusion regarding CEO tenure?

A. $b_2 \leq 0$; reject the null hypothesis
B. $b_2 = 0$; cannot reject the null hypothesis
C. $b_2 \geq 0$; reject the null hypothesis
参考答案：A
【莽学解析】Varden expects to find that CEO tenure is positively related to the firm's ROE. If he is correct, the regression coefficient for tenure, b_2, will be greater than zero ($b_2 > 0$) and statistically significant. The null hypothesis supposes that the "suspected" condition is not true, so the null hypothesis should state the variable is less than or equal to zero. The t-statistic for tenure is 2.308, significant at the 0.027 level, meeting Varden's 0.05 significance requirement. Varden should reject the null hypothesis.

49.【单项选择题】At a significance level of 1%, which of the following is the best interpretation of the regression coefficients with regard to explaining ROE?
A. ESG is significant, but tenure is not.
B. Tenure is significant, but ESG is not.
C. Neither ESG nor tenure is significant.
参考答案：C
【莽学解析】The t-statistic for tenure is 2.308, indicating significance at the 0.027 level but not the 0.01 level. The t-statistic for ESG is 1.201, with a p-value of 0.238, which means we fail to reject the null hypothesis for ESG at the 0.01 significance level.

50.【单项选择题】Based on Exhibit 1, which independent variables in Varden's model are significant at the 0.05 level?
A. ESG only
B. Tenure only
C. Neither ESG nor tenure
参考答案：B
【莽学解析】The t-statistic for tenure is 2.308, which is significant at the 0.027 level. The t-statistic for ESG is 1.201, with a p-value of 0.238. This result is not significant at the 0.05 level.

51.【单项选择题】Based on Exhibit 1, the predicted ROE for DF Associates is closest to:
A. 10.957%.
B. 16.593%.
C. 20.388%.
参考答案：C
【莽学解析】

The regression equation is as follows:

$$\hat{Y}_i = 9.442 + 0.069X_{1i} + 0.681X_{2i}$$

$$\text{ROE} = 9.442 + 0.069(\text{ESG}) + 0.681(\text{Tenure})$$

$$= 9.442 + 0.069(55) + 0.681(10.5)$$

$$= 9.442 + 3.795 + 7.151$$

$$= 20.388.$$

52.【单项选择题】Based on Exhibit 2, Quinni's best answer to Varden's question about the effect of adding a third independent variable is:

A. no for R^2 and no for adjusted R^2.

B. yes for R^2 and no for adjusted R^2.

C. yes for R^2 and yes for adjusted R^2.

参考答案：B

【莽学解析】When you add an additional independent variable to the regression model, the amount of unexplained variance will decrease, provided the new variable explains any of the previously unexplained variation. This result occurs as long as the new variable is even slightly correlated with the dependent variable. Exhibit 2 indicates the dividend growth rate is correlated with the dependent variable, ROE. Therefore, R² will increase. Adjusted R², however, may not increase and may even decrease if the relationship is weak. This result occurs because in the formula for adjusted R², the new variable increases k (the number of independent variables) in the denominator, and the increase in R² may be insufficient to increase the value of the formula.

$$\text{adjusted } R^2 = 1 - \left(\frac{n-1}{n-k-1}\right)(1-R^2)$$

53.【单项选择题】Based on Exhibit 1, Varden's best answer to Quinni's question about the F-statistic is:

A. both independent variables are significant at the 0.05 level.

B. neither independent variable is significant at the 0.05 level.

C. at least one independent variable is significant at the 0.05 level.

参考答案：C

【莽学解析】Exhibit 1 indicates that the F-statistic of 4.161 is significant at the 0.05 level. A significant F-statistic means at least one of the independent variables is significant.

54.【单项选择题】Varden's best answer to Quinni's question about overall significance is:

A. R^2.

B. adjusted R^2.

C. the F-statistic.

参考答案：C

【莽学解析】In a multiple linear regression (as compared with simple regression), R^2 is less appropriate as a measure of whether a regression model fits the data well. A high adjusted R^2 does not necessarily indicate that the regression is well specified in the sense of including the correct set of variables. The F-test is an appropriate test of a regression's overall significance in either simple or multiple regressions.

55.【单项选择题】If Varden's beliefs about ROE and CEO tenure are true, which of the following would violate the assumptions of multiple regression analysis?

A. The assumption about CEO tenure distribution only

B. The assumption about the ROE/dividend growth correlation only

C. The assumptions about both the ROE/dividend growth correlation and CEO tenure distribution

参考答案：C

【莽学解析】Multiple linear regression assumes that the relationship between the dependent variable and each of the independent variables is linear. Varden believes that this is not true for dividend growth because he believes the relationship may be different in firms with a long-standing CEO. Multiple linear regression also assumes that the independent variables are not random. Varden states that he believes CEO tenure is a random variable.

56.【单项选择题】The best rationale for Quinni's caution about the three-variable model is that the:

A. dependent variable is defined differently.

B. sample sizes are different in the two models.

C. dividend growth rate is positively correlated with the other independent variables.

参考答案：B

【莽学解析】If we use adjusted R^2 to compare regression models, it is important that the dependent variable be defined the same way in both models and that the sample sizes used to estimate the models are the same. Varden's first model was based on 40 observations, whereas the second model was based on 500.

【题干】Elena Vasileva recently joined EnergyInvest as a junior portfolio analyst. Vasileva's supervisor asks her to evaluate a potential investment opportunity in Amtex, a multinational oil and gas corporation based in the United States. Vasileva's supervisor suggests using regression analysis to examine the relation between Amtex shares and returns on crude oil.

Vasileva notes the following assumptions of regression analysis:

Assumption 1: The error term is uncorrelated across observations.

Assumption 2: The variance of the error term is the same for all observations.

Assumption 3: The dependent variable is normally distributed.

Vasileva runs a regression of Amtex share returns on crude oil returns using the monthly data she collected. Selected data used in the regression are presented in Exhibit 1, and selected regression output is presented in Exhibit 2. She uses a 1% level of significance in all her tests.

Critical t-values for a 1% level of significance:

One-sided, left side: -2.441

One-sided, right side: +2.441

Two-sided: ±2.728

Vasileva expects the crude oil return next month, Month 37, to be -0.01. She computes the standard error of the forecast to be 0.0469.

57.【单项选择题】Which of Vasileva's assumptions regarding regression analysis is incorrect?

A. Assumption 1

B. Assumption 2

C. Assumption 3

参考答案：C

【莽学解析】The assumptions of the linear regression model are that (1) the relationship between the dependent variable and the independent variable is linear in the parameters b0 and b1, (2) the residuals are independent of one another, (3) the variance of the error term is the

Exhibit 1 Selected Data for Crude Oil Returns and Amtex Share Returns

	Oil Return (X_i)	Amtex Return (Y_i)	Cross-Product $(X_i - \bar{X})(Y_i - \bar{Y})$	Predicted Amtex Return $\hat{Y}_i$	Regression Residual $Y_i - \hat{Y}_i$
Month 1	−0.032000	0.033145	−0.000388	0.002011	−0.031134
⋮	⋮	⋮	⋮	⋮	⋮
Month 36	0.028636	0.062334	0.002663	0.016282	−0.046053
Sum			0.085598		
Average	−0.018056	0.005293			

Exhibit 2 Selected Regression Output, Dependent Variable: Amtex Share Return

	Coefficient	Standard Error
Intercept	0.0095	0.0078
Oil return	0.2354	0.0760

same for all observations, and (4) the error term is normally distributed. Assumption 3 is incorrect because the dependent variable need not be normally distributed.

58.【单项选择题】Based on Exhibit 1, the standard error of the estimate is closest to:

A. 0.04456.

B. 0.04585.

C. 0.05018.

参考答案：B

【莽学解析】The standard error of the estimate for a linear regression model with one independent variable is calculated as the square root of the mean square error:

$$s_e = \sqrt{\frac{0.071475}{34}} = 0.04585.$$

59.【单项选择题】Based on Exhibit 2, Vasileva should reject the null hypothesis that:

A. the slope is less than or equal to 0.15.

B. the intercept is less than or equal to zero.

C. crude oil returns do not explain Amtex share returns.

参考答案：C

【莽学解析】Crude oil returns explain the Amtex share returns if the slope coefficient is statistically different from zero. The slope coefficient is 0.2354, and the calculated t-statistic is

$$t\text{-statistic} = \frac{\hat{b}_1 - b_1}{s_{\hat{b}_1}} = \frac{0.2354 - 0.0000}{0.0760} = 3.0974$$

which is outside the bounds of the critical values of ±2.728.

Therefore, Vasileva should reject the null hypothesis that crude oil returns do not explain Amtex share returns, because the slope coefficient is statistically different from zero.

60.【单项选择题】Based on Exhibit 2 and Vasileva's prediction of the crude oil return for Month 37, the estimate of Amtex share return for Month 37 is closest to:

A. -0.0024.

B. 0.0071.

C. 0.0119.

参考答案：B

【莽学解析】The predicted value of the dependent variable, Amtex share return, given the value of the independent variable, crude oil return, -0.01, is calculated as

$$\hat{Y} = \hat{b}_0 + \hat{b}_1 X_i = 0.0095 + [0.2354 \times (-0.01)] = 0.0071$$

61.【单项选择题】Using information from Exhibit 2, the 99% prediction interval for Amtex share return for Month 37 is best described as:

A.

$\hat{Y}_f$ ± 0.0053.

B.

$\hat{Y}_f$ ± 0.0469.

C.

$\hat{Y}_f$ ± 0.1279

参考答案：C

【莽学解析】The predicted share return is 0.0095 + [0.2354 × (-0.01)] = 0.0071. The lower

limit for the prediction interval is 0.0071 - (2.728 × 0.0469) = -0.1208, and the upper limit for the prediction interval is 0.0071 + (2.728 × 0.0469) = 0.1350.

【题干】An analyst is examining the annual growth of the money supply for a country over the past 30 years. This country experienced a central bank policy shift 15 years ago, which altered the approach to the management of the money supply. The analyst estimated a model using the annual growth rate in the money supply regressed on the variable (SHIFT) that takes on a value of 0 before the policy shift and 1 after. She estimated the following:

	Coefficients	Standard Error	*t*-Stat.
Intercept	5.767264	0.445229	12.953
SHIFT	−5.13912	0.629649	−8.161

Critical t-values, level of significance of 0.05:One-sided, left side: -1.701One-sided, right side: +1.701Two-sided: ±2.048

62.【单项选择题】The variable SHIFT is best described as:
A. an indicator variable.
B. a dependent variable.
C. a continuous variable.
参考答案：A
【莽学解析】SHIFT is an indicator or dummy variable because it takes on only the values 0 and 1.

63.【单项选择题】The interpretation of the intercept is the mean of the annual growth rate of the money supply:
A. over the enter entire period.
B. after the shift in policy.
C. before the shift in policy.
参考答案：C
【莽学解析】In a simple regression with a single indicator variable, the intercept is the mean of the dependent variable when the indicator variable takes on a value of zero, which is before the shift in policy in this case.

64.【单项选择题】The interpretation of the slope is the:
A. change in the annual growth rate of the money supply per year.
B. average annual growth rate of the money supply after the shift in policy.
C. difference in the average annual growth rate of the money supply from before to after the shift in policy.
参考答案：C
【莽学解析】Whereas the intercept is the average of the dependent variable when the indicator variable is zero (that is, before the shift in policy), the slope is the difference in the mean of the dependent variable from before to after the change in policy.

65.【单项选择题】Testing whether there is a change in the money supply growth after the shift in policy, using a 0.05 level of significance, we conclude that there is:

A. sufficient evidence that the money supply growth changed.

B. not enough evidence that the money supply growth is different from zero.

C. not enough evidence to indicate that the money supply growth changed

参考答案：A

【莽学解析】The null hypothesis of no difference in the annual growth rate is rejected at the 0.05 level: The calculated test statistic of -8.16188 is outside the bounds of ±2.048.

【题干】Howard Golub, CFA, is preparing to write a research report on Stellar Energy Corp. common stock. One of the world's largest companies, Stellar is in the business of refining and marketing oil. As part of his analysis, Golub wants to evaluate the sensitivity of the stock's returns to various economic factors. For example, a client recently asked Golub whether the price of Stellar Energy Corp. stock has tended to rise following increases in retail energy prices. Golub believes the association between the two variables is negative, but he does not know the strength of the association.Golub directs his assistant, Jill Batten, to study the relationships between (1) Stellar monthly common stock returns and the previous month's percentage change in the US Consumer Price Index for Energy (CPIENG) and (2) Stellar monthly common stock returns and the previous month's percentage change in the US Producer Price Index for Crude Energy Materials (PPICEM). Golub wants Batten to run both a correlation and a linear regression analysis. In response, Batten compiles the summary statistics shown in Exhibit 1 for 248 months. All the data are in decimal form, where 0.01 indicates a 1% return. Batten also runs a regression analysis using Stellar monthly returns as the dependent variable and the monthly change in CPIENG as the independent variable. Exhibit 2 displays the results of this regression model.

Exhibit 2 Regression Analysis with CPIENG

Regression Statistics

R^2	0.0211
Standard error of the estimate	0.0710
Observations	248

Regression Statistics

	Coefficients	Standard Error	*t*-St
Intercept	0.0138	0.0046	3
CPIENG (%)	−0.6486	0.2818	−2

Exhibit 1 Descriptive Statistics

	Stellar Common Stock Monthly Return	Lagged Cha
		CPIENG
Mean	0.0123	0.0023
Standard deviation	0.0717	0.0160
Covariance, Stellar vs. CPIENG	−0.00017	
Covariance, Stellar vs. PPICEM	−0.00048	
Covariance, CPIENG vs. PPICEM	0.00044	
Correlation, Stellar vs. CPIENG	−0.1452	

Critical t-values
One-sided, left side: -1.651
One-sided, right side: +1.651
Two-sided: ±1.967

66.【单项选择题】Which of the following best describes Batten's regression?
A. Time-series regression.
B. Cross-sectional regression.
C. Time-series and cross-sectional regression.
参考答案：A
【莽学解析】The data are observations over time.

67.【单项选择题】Based on the regression, if the CPIENG decreases by 1.0%, the expected return on Stellar common stock during the next period is closest to:
A. 0.0073 (0.73%).
B. 0.0138 (1.38%).
C. 0.0203 (2.03%).
参考答案：C
【莽学解析】From the regression equation, Expected return = 0.0138＋ 0.006486 = 0.0203, or 2.03%.

68.【单项选择题】Based on Batten's regression model, the coefficient of determination indicates that:

A. Stellar's returns explain 2.11% of the variability in CPIENG.
B. Stellar's returns explain 14.52% of the variability in CPIENG.
C. changes in CPIENG explain 2.11% of the variability in Stellar's returns.
参考答案：C
【莽学解析】R^2 is the coefficient of determination. In this case, it shows that 2.11% of the variability in Stellar's returns is explained by changes in CPIENG.

69.【单项选择题】For Batten's regression model, 0.0710 is the standard deviation of:
A. the dependent variable.
B. the residuals from the regression.
C. the predicted dependent variable from the regression.
参考答案：C
【莽学解析】The standard error of the estimate is the standard deviation of the regression residuals.

70.【单项选择题】For the analysis run by Batten, which of the following is an incorrect conclusion from the regression output?
A. The estimated intercept from Batten's regression is statistically different from zero at the 0.05 level of significance.
B. In the month after the CPIENG declines, Stellar's common stock is expected to exhibit a positive return.
C. Viewed in combination, the slope and intercept coefficients from Batten's regression are not statistically different from zero at the 0.05 level of significance.
参考答案：C
【莽学解析】The slope and intercept are both statistically different from zero at the 0.05 level of significance.

【题干】Iesha Azarov is a senior analyst at Ganymede Moon Partners (Ganymede), where he works with junior analyst Pàola Bector. Azarov would like to incorporate machine learning (ML) models into the company's analytical process. Azarov asks Bector to develop ML models for two unstructured stock sentiment datasets, Dataset ABC and Dataset XYZ. Both datasets have been cleaned and preprocessed in preparation for text exploration and model training.
Following an exploratory data analysis that revealed Dataset ABC's most frequent tokens, Bector conducts a collection frequency analysis. Bector then computes TF－IDF (term frequency－inverse document frequency) for several words in the collection and tells Azarov the following:
Statement 1: IDF is equal to the inverse of the document frequency measure.
Statement 2: TF at the collection level is multiplied by IDF to calculate TF－IDF.
Statement 3: TF－IDF values vary by the number of documents in the dataset, and therefore, model performance can vary when applied to a dataset with just a few documents.
Bector notes that Dataset ABC is characterized by the absence of ground truth.
Bector turns his attention to Dataset XYZ, containing 84,000 tokens and 10,000 sentences. Bector chooses an appropriate feature selection method to identify and remove unnecessary tokens from the dataset and then focuses on model training. For performance evaluation purposes, Dataset XYZ is split into a training set, cross-validation (CV) set, and test set. Each of the sentences has already been labeled as either a positive sentiment (Class "1") or

a negative sentiment (Class “0”) sentence. There is an unequal class distribution between the positive sentiment and negative sentiment sentences in Dataset XYZ. Simple random sampling is applied within levels of the sentiment class labels to balance the class distributions within the splits. Bector’s view is that the false positive and false negative evaluation metrics should be given equal weight. Select performance data from the cross-validation set confusion matrices is presented in Exhibit 1:

Exhibit 1 Performance Metrics for Dataset XYZ

Confusion Matrix	CV Data (threshold *p*-value)	Performance M		
		Precision	Recall	F
A	0.50	0.95	0.87	
B	0.35	0.93	0.90	
C	0.65	0.86	0.97	

Azarov and Bector evaluate the Dataset XYZ performance metrics for Confusion Matrices A, B, and C in Exhibit 1. Azarov says, “For Ganymede’s purposes, we should be most concerned with the cost of Type I errors.”

Azarov requests that Bector apply the ML model to the test dataset for Dataset XYZ, assuming a threshold p-value of 0.65. Exhibit 2 contains a sample of results from the test dataset corpus.

Bector makes the following remarks regarding model training:

Remark 1: Method selection is governed by such factors as the type of data and the size of data.

Remark 2: In the performance evaluation stage, model fitting errors, such as bias error and variance error, are used to measure goodness of fit.

71.【单项选择题】Based on the text exploration method used for Dataset ABC, tokens that potentially carry important information useful for differentiating the sentiment embedded in the text are most likely to have values that are:

A. low.

B. intermediate.

C. high.

参考答案：B

【莽学解析】When analyzing term frequency at the corpus level, also known as collection frequency, tokens with intermediate term frequency (TF) values potentially carry important information useful for differentiating the sentiment embedded in the text. Tokens with the highest TF values are mostly stop words that do not contribute to differentiating the sentiment embedded in the text, and tokens with the lowest TF values are mostly proper nouns or sparse

Exhibit 2 10 Sample Results of Test Data for Dataset

Sentence #	Actual Sentiment
1	1
2	0
3	1
4	1
5	0
6	1
7	0
8	1
9	0
10	0

terms that are also not important to the meaning of the text.
A is incorrect because tokens with the lowest TF values are mostly proper nouns or sparse terms (noisy terms) that are not important to the meaning of the text.
C is incorrect because tokens with the highest TF values are mostly stop words (noisy terms) that do not contribute to differentiating the sentiment embedded in the text.

72.【单项选择题】Which of Bector's statements regarding TF, IDF, and TF－IDF is correct?
A. Statement 1
B. Statement 2
C. Statement 3
参考答案：C
【莽学解析】Statement 3 is correct. TF－IDF values vary by the number of documents in the dataset, and therefore, the model performance can vary when applied to a dataset with just a few documents. Statement 1 is incorrect because IDF is calculated as the log of the inverse, or reciprocal, of the document frequency measure. Statement 2 is incorrect because TF at the sentence (not collection) level is multiplied by IDF to calculate TF－IDF.
A is incorrect because Statement 1 is incorrect. IDF is calculated as the log of the inverse, or reciprocal, of the document frequency (DF) measure.

B is incorrect because Statement 2 is incorrect. TF at the sentence (not collection) level is multiplied by IDF to calculate TF－IDF.

73.【单项选择题】What percentage of Dataset ABC should be allocated to a training subset?

A. 0%.

B. 20%.

C. 60%.

参考答案：A

【莽学解析】0% of the master dataset of Dataset ABC should be allocated to a training subset. Dataset ABC is characterized by the absence of ground truth (i.e., no known outcome or target variable) and is therefore an unsupervised ML model. For unsupervised learning models, no splitting of the master dataset is needed, because of the absence of labeled training data. Supervised ML datasets (with labeled training data) contain ground truth, the known outcome (target variable) of each observation in the dataset.

B is incorrect because 20% is the commonly recommended split for the cross-validation set and test set in supervised training ML datasets. Dataset ABC is an unsupervised ML dataset, for which no splitting (0%) of the master dataset is needed, because of the absence of labeled training data. In supervised ML models (which contain labeled training data), the master dataset is split into three subsets (a training set, cross-validation set, and test set), which are used for model training and testing purposes.

C is incorrect because 60% is the commonly recommended split for the training set in supervised training ML datasets. Dataset ABC is an unsupervised ML dataset, for which no splitting (0%) of the master dataset is needed, because of the absence of labeled training data. In supervised ML models (which contain labeled training data), the master dataset is split into three subsets (a training set, cross-validation set, and test set), which are used for model training and testing purposes.

74.【单项选择题】Based only on Dataset XYZ's composition and Bector's view regarding false positive and false negative evaluation metrics, which performance measure is most appropriate?

A. Recall.

B. F1 score.

C. Precision.

参考答案：B

【莽学解析】F1 score is the most appropriate performance measure for Dataset XYZ. Bector gives equal weight to false positives and false negatives. Accuracy and F1 score are overall performance measures that give equal weight to false positives and false negatives. Accuracy is considered an appropriate performance measure for balanced datasets, where the number of "1" and "0" classes are equal. F1 score is considered more appropriate than accuracy when there is unequal class distribution in the dataset and it is necessary to measure the equilibrium of precision and recall. Since Dataset XYZ contains an unequal class distribution between positive and negative sentiment sentences, F1 score is the most appropriate performance measure.

Precision is the ratio of correctly predicted positive classes to all predicted positive classes and is useful in situations where the cost of false positives or Type I errors is high.

Recall is the ratio of correctly predicted positive classes to all actual positive classes and is useful in situations where the cost of false negatives or Type II errors is high.

A is incorrect because Bector gives equal weight to false positives and false negatives.

Accuracy and F1 score are overall performance measures that give equal weight to false positives and false negatives. Recall is the ratio of correctly predicted positive classes to all actual positive classes and is useful in situations where the cost of false negatives or Type II errors is high.
C is incorrect because Bector gives equal weight to false positive and false negatives. Accuracy and F1 score are overall performance measures that give equal weight to false positives and false negatives. Precision is the ratio of correctly predicted positive classes to all predicted positive classes and is useful in situations where the cost of false positives or Type-I error is high.

75.【单项选择题】Based on Exhibit 1, which confusion matrix demonstrates the most favorable value of the performance metric that best addresses Azarov's concern?
A. Confusion Matrix A.
B. Confusion Matrix B.
C. Confusion Matrix C.
参考答案：A
【莽学解析】Precision is the ratio of correctly predicted positive classes to all predicted positive classes and is useful in situations where the cost of false positives or Type I errors is high. Confusion Matrix A has the highest precision and therefore demonstrates the most favorable value of the performance metric that best addresses Azarov's concern about the cost of Type I errors. Confusion Matrix A has a precision score of 0.95, which is higher than the precision scores of Confusion Matrix B (0.93) and Confusion Matrix C (0.86).
B is incorrect because precision, not accuracy, is the performance measure that best addresses Azarov's concern about the cost of Type I errors. Confusion Matrix B demonstrates the most favorable value for the accuracy score (0.92), which is higher than the accuracy scores of Confusion Matrix A (0.91) and Confusion Matrix C (0.91). Accuracy is a performance measure that gives equal weight to false positives and false negatives and is considered an appropriate performance measure when the class distribution in the dataset is equal (a balanced dataset). However, Azarov is most concerned with the cost of false positives, or Type I errors, and not with finding the equilibrium between precision and recall. Furthermore, Dataset XYZ has an unequal (unbalanced) class distribution between positive sentiment and negative sentiment sentences.
C is incorrect because precision, not recall or F1 score, is the performance measure that best addresses Azarov's concern about the cost of Type I errors. Confusion Matrix C demonstrates the most favorable value for the recall score (0.97), which is higher than the recall scores of Confusion Matrix A (0.87) and Confusion Matrix B (0.90). Recall is the ratio of correctly predicted positive classes to all actual positive classes and is useful in situations where the cost of false negatives, or Type II errors, is high. However, Azarov is most concerned with the cost of Type I errors, not Type II errors.
F1 score is more appropriate (than accuracy) when there is unequal class distribution in the dataset and it is necessary to measure the equilibrium of precision and recall. Confusion Matrix C demonstrates the most favorable value for the F1 score (0.92), which is higher than the F1 scores of Confusion Matrix A (0.91) and Confusion Matrix B (0.91). Although Dataset XYZ has an unequal class distribution between positive sentiment and negative sentiment sentences, Azarov is most concerned with the cost of false positives, or Type I errors, and not with finding the equilibrium between precision and recall.

76.【单项选择题】Based on Exhibit 2, the accuracy metric for Dataset XYZ's test set sample is closest to:

A. 0.67.

B. 0.70.

C. 0.75.

参考答案：B

【莽学解析】Accuracy is the percentage of correctly predicted classes out of total predictions and is calculated as (TP + TN)/(TP + FP + TN + FN). In order to obtain the values for true positive (TP), true negative (TN), false positive (FP), and false negative (FN), predicted sentiment for the positive (Class "1") and the negative (Class "0") classes are determined based on whether each individual target p-value is greater than or less than the threshold p-value of 0.65. If an individual target p-value is greater than the threshold p-value of 0.65, the predicted sentiment for that instance is positive (Class "1"). If an individual target p-value is less than the threshold p-value of 0.65, the predicted sentiment for that instance is negative (Class "0"). Actual sentiment and predicted sentiment are then classified as follows:

Actual Sentiment	Predicted Sentiment
1	1
0	1
1	0
0	0

Exhibit 2, with added "Predicted Sentiment" and "Classification" columns, is presented below:

Based on the classification data obtained from Exhibit 2, a confusion matrix can be generated:

Exhibit 2 10 Sample Results of Test Data for Data

Sentence #	Actual Sentiment	Target *p*-Value	Predi Sentir
1	1	0.75	1
2	0	0.45	0
3	1	0.64	0
4	1	0.81	1
5	0	0.43	0
6	1	0.78	1
7	0	0.59	0
8	1	0.60	0
9	0	0.67	1
10	0	0.54	0

Using the data in the confusion matrix above, the accuracy metric is computed as follows:
Accuracy = (TP + TN)/(TP + FP + TN + FN).
Accuracy = (3 + 4)/(3 + 1 + 4 + 2) = 0.70.
A is incorrect because 0.67 is the F1 score, not accuracy metric, for the sample of the test set for Dataset XYZ, based on Exhibit 2. To calculate the F1 score, the precision (P) and the recall (R) ratios must first be calculated. Precision and recall for the sample of the test set for Dataset XYZ, based on Exhibit 2, are calculated as follows:
Precision (P) = TP/(TP + FP) = 3/(3 + 1) = 0.75.
Recall (R) = TP/(TP + FN) = 3/(3 + 2) = 0.60.
The F1 score is calculated as follows:
F1 score = (2 × P × R)/(P + R) = (2 × 0.75 × 0.60)/(0.75 + 0.60) = 0.667, or 0.67.
C is incorrect because 0.75 is the precision ratio, not the accuracy metric, for the sample of the test set for Dataset XYZ, based on Exhibit 2. The precision score is calculated as follows:
Precision (P) = TP/(TP + FP) = 3/(3 + 1) = 0.75.

Confusion Matrix for Dataset XYZ Sample Test Data
0.65

	Actual Train
Predicted Results	Class "1"
Class "1"	TP = 3
Class "0"	FN = 2

77.【单项选择题】Which of Bector's remarks related to model training is correct?
A. Only Remark 1.
B. Only Remark 2.
C. Both Remark 1 and Remark 2.
参考答案：A
【莽学解析】Only Remark 1 is correct. Method selection is the first task of ML model training and is governed by the following factors: (1) supervised or unsupervised learning, (2) the type of data, and (3) the size of data. The second and third tasks of model training, respectively, are performance evaluation and tuning.
Remark 2 is incorrect because model fitting errors (bias error and variance error) are used in tuning, not performance evaluation. The techniques used in performance evaluation, which measure the goodness of fit for validation of the model, include (1) error analysis, (2) receiver operating characteristic (ROC) plots, and (3) root mean squared error (RMSE) calculations.
B and C are incorrect because Remark 2 is incorrect. Model fitting errors (bias error and variance error) are used in tuning, not performance evaluation. The techniques used in performance evaluation, which measure the goodness of fit for validation of the model, include (1) error analysis, (2) receiver operating characteristic plots, and (3) root mean squared error calculations.

【题干】Gary Hansen is a securities analyst for a mutual fund specializing in small-capitalization growth stocks. The fund regularly invests in initial public offerings (IPOs). If the fund subscribes to an offer, it is allocated shares at the offer price. Hansen notes that IPOs frequently are underpriced, and the price rises when open market trading begins. The initial return for an IPO is calculated as the change in price on the first day of trading divided by the offer price. Hansen is developing a regression model to predict the initial return for IPOs. Based on past research, he selects the following independent variables to predict IPO initial returns:

Underwriter rank	=	1–10, where 10 is highest rank
Pre-offer price adjustment[a]	=	(Offer price - Initial filing price)/Initial filing price
Offer size ($ millions)	=	Shares sold × Offer price
Fraction retained[a]	=	Fraction of total company shares retained by insiders

[a]Expressed as a decimal

Hansen collects a sample of 1,725 recent IPOs for his regression model. Regression results appear in Exhibit 1, and ANOVA results appear in Exhibit 2.

Exhibit 1. Hansen's Regression Results Dependent Variable: IPO Initial Retur
(Expressed in Decimal Form, i.e., 1% = 0.01)

Variable	Coefficient (b_j)	Standard Error	*t*-Statistic
Intercept	0.0477	0.0019	25.11
Underwriter rank	0.0150	0.0049	3.06
Pre-offer price adjustment	0.4350	0.0202	21.53
Offer size	-0.0009	0.0011	−0.82
Fraction retained	0.0500	0.0260	1.92

Exhibit 2. Selected ANOVA Results for Hansen's Regression

	Degrees of Freedom (df)	Sum of Squares (SS)
Regression	4	51.433
Residual	1,720	91.436
Total	1,724	142.869

Multiple R-squared = 0.36

Hansen wants to use the regression results to predict the initial return for an upcoming IPO. The upcoming IPO has the following characteristics:■ underwriter rank = 6;■ pre-offer price adjustment = 0.04;■ offer size = $40 million;■ fraction retained = 0.70.Because he notes that

the pre-offer price adjustment appears to have an important effect on initial return, Hansen wants to construct a 95 percent confidence interval for the coefficient on this variable. He also believes that for each 1 percent increase in pre-offer price adjustment, the initial return will increase by less than 0.5 percent, holding other variables constant. Hansen wishes to test this hypothesis at the 0.05 level of significance.Before applying his model, Hansen asks a colleague, Phil Chang, to review its specification and results. After examining the model, Chang concludes that the model suffers from two problems: 1) conditional heteroskedasticity, and 2) omitted variable bias. Chang makes the following statements:Statement 1: "Conditional heteroskedasticity will result in consistent coefficient estimates, but both the t-statistics and F-statistic will be biased, resulting in false inferences." Statement 2: "If an omitted variable is correlated with variables already included in the model, coefficient estimates will be biased and inconsistent and standard errors will also be inconsistent." Selected values for the t-distribution and F-distribution appear in Exhibits 3 and 4, respectively.

Exhibit 3. Selected Values for the *t*-Distribution (df = ∞)

Area in Right Tail	*t*-Value
0.050	1.645
0.025	1.960
0.010	2.326
0.005	2.576

Exhibit 4. Selected Values for the *F*-Distribution (α=0.01) (df1/df2: Numerator/Denominator Degrees of Freedom)

		df1	
		4	∞
df2	4	16.00	13.50
	∞	3.32	1.00

78.【单项选择题】Based on Hansen's regression, the predicted initial return for the upcoming IPO is closest to:

A.0.0943.

B.0.1064.

C. 0.1541.

参考答案：C

【莽学解析】The predicted initial return (IR) is:IR = 0.0477 + (0.0150 × 6) + (0.435 × 0.04) − (0.0009 × 40) + (0.05 × 0.70) = 0.1541

79.【单项选择题】The 95 percent confidence interval for the regression coefficient for the pre-offer price adjustment is closest to:

A. 0.156 to 0.714.

B. 0.395 to 0.475.

C. 0.402 to 0.468.

参考答案：B

【莽学解析】The 95% confidence interval is 0.435 ± (0.0202 × 1.96) = (0.395, 0.475).

80.【单项选择题】The most appropriate null hypothesis and the most appropriate conclusion regarding Hansen's belief about the magnitude of the initial return relative to that of the pre-offer price adjustment (reflected by the coefficient b_j) are:

A. Null Hypothesis: H_0: $b_j = 0.5$; Conclusion about b_j (0.05 Level of Significance): Reject H_0

B. Null Hypothesis: H_0: $b_j \geq 0.5$; Conclusion about b_j (0.05 Level of Significance): Fail to reject H_0

C. Null Hypothesis: H_0: $b_j \geq 0.5$; Conclusion about b_j (0.05 Level of Significance): Reject H_0

参考答案：C

【莽学解析】To test Hansen's belief about the direction and magnitude of the initial return, the test should be a one-tailed test. The alternative hypothesis is H_a: $b_j < 0.5$, and the null hypothesis is H_0: $b_j \geq 0.5$. The correct test statistic is: t = (0.435 − 0.50)/0.0202 = −3.22, and the critical value of the t-statistic for a one-tailed test at the 0.05 level is −1.645. The test statistic is significant, and the null hypothesis can be rejected at the 0.05 level of significance.

81.【单项选择题】The most appropriate interpretation of the multiple R-squared for Hansen's model is that:

A. unexplained variation in the dependent variable is 36 percent of total variation.

B. correlation between predicted and actual values of the dependent variable is 0.36.

C. correlation between predicted and actual values of the dependent variable is 0.60.

参考答案：C

【莽学解析】The multiple R-squared for the regression is 0.36; thus, the model explains 36 percent of the variation in the dependent variable. The correlation between the predicted and actual values of the dependent variable is the square root of the R-squared or

$$\sqrt{0.36} = 0.60.$$

82.【单项选择题】Is Chang's Statement 1 correct?

A. Yes.

B. No, because the model's F-statistic will not be biased.

C. No, because the model's t-statistics will not be biased.

参考答案：A

【莽学解析】Chang is correct because the presence of conditional heteroskedasticity results in consistent parameter estimates, but biased (up or down) standard errors, t-statistics, and F-statistics.

83.【单项选择题】Is Chang's Statement 2 correct?

A. Yes.

B. No, because the model's coefficient estimates will be unbiased.

C. No, because the model's coefficient estimates will be consistent.

参考答案：A

【莽学解析】Chang is correct because a correlated omitted variable will result in biased and inconsistent parameter estimates and inconsistent standard errors.

【题干】Anh Liu is an analyst researching whether a company's debt burden affects investors' decision to short the company's stock. She calculates the short interest ratio (the ratio of short interest to average daily share volume, expressed in days) for 50 companies as of the end of 2016 and compares this ratio with the companies' debt ratio (the ratio of total liabilities to total assets, expressed in decimal form).

Liu provides a number of statistics in Exhibit 1. She also estimates a simple regression to investigate the effect of the debt ratio on a company's short interest ratio. The results of this simple regression, including the analysis of variance (ANOVA), are shown in Exhibit 2.

In addition to estimating a regression equation, Liu graphs the 50 observations using a scatter plot, with the short interest ratio on the vertical axis and the debt ratio on the horizontal axis.

Exhibit 1 Summary Statistics

Statistic	Debt Ratio X_I
Sum	19.8550
Sum of squared deviations from the mean	$\sum_{i=1}^{n}(X_i - \bar{X})^2 = 2.2225.$
Sum of cross-products of deviations from the mean	$\sum_{i=1}^{n}(X_i - \bar{X})(Y_i$

Critical t-values for a 0.05 level of significance:
One-sided, left side: -1.677
One-sided, right side: +1.677
Two-sided:±2.011
Liu is considering three interpretations of these results for her report on the relationship between debt ratios and short interest ratios:
Interpretation 1: Companies' higher debt ratios cause lower short interest ratios.
Interpretation 2: Companies' higher short interest ratios cause higher debt ratios.
Interpretation 3: Companies with higher debt ratios tend to have lower short interest ratios.She is especially interested in using her estimation results to predict the short interest ratio for MQD Corporation, which has a debt ratio of 0.40.

84.【单项选择题】Based on Exhibits 1 and 2, if Liu were to graph the 50 observations, the scatter plot summarizing this relation would be best described as:
A.horizontal.
B.upward sloping.
C.downward sloping.
参考答案：C
【莽学解析】The slope coefficient (shown in Exhibit 2) is negative. We could also determine this by looking at the cross-product (Exhibit 1), which is negative.

Exhibit 2 Regression of the Short Interest Ratio o

ANOVA	Degrees of Freedom (df)	Sum of Squares
Regression	1	38.4404
Residual	48	373.7638
Total	49	412.2042

Regression Statistics	
R^2	0.0933
Standard error of estimate	2.7905
Observations	50

	Coefficients	Standard
Intercept	5.4975	0.841
Debt ratio (%)	−4.1589	1.871

85.【单项选择题】Based on Exhibit 1, the sample covariance is closest to:

A. -9.2430.

B. -0.1886.

C. 8.4123.

参考答案：B

【莽学解析】The sample covariance is calculated as

$$\frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{n-1}=-9.2430\div 49=-0.1886$$

86.【单项选择题】Based on Exhibits 1 and 2, the correlation between the debt ratio and the short interest ratio is closest to:
A. -0.3054.
B. 0.0933.
C. 0.3054.
参考答案：A
【莽学解析】In simple regression, the R^2 is the square of the pairwise correlation. Because the slope coefficient is negative, the correlation is the negative of the square root of 0.0933, or -0.3054.

87.【单项选择题】Which of the interpretations best describes Liu's findings?
A. Interpretation 1
B. Interpretation 2
C. Interpretation 3
参考答案：C
【莽学解析】Conclusions cannot be drawn regarding causation; they can be drawn only about association; therefore, Interpretations 1 and 2 are incorrect.

88.【单项选择题】The dependent variable in Liu's regression analysis is the:
A. intercept.
B. debt ratio.
C. short interest ratio.
参考答案：C
【莽学解析】Liu explains the variation of the short interest ratio using the variation of the debt ratio.

89.【单项选择题】Based on Exhibit 2, the degrees of freedom for the t-test of the slope coefficient in this regression are:
A. 48.
B. 49.
C. 50.
参考答案：A
【莽学解析】The degrees of freedom are the number of observations minus the number of parameters estimated, which equals 2 in this case (the intercept and the slope coefficient). The number of degrees of freedom is 50 - 2 = 48.

90.【单项选择题】Which of the following should Liu conclude from the results shown in Exhibit 2?
A. The average short interest ratio is 5.4975.
B. The estimated slope coefficient is different from zero at the 0.05 level of significance.
C. The debt ratio explains 30.54% of the variation in the short interest ratio.
参考答案：B
【莽学解析】The t-statistic is -2.2219, which is outside the bounds created by the critical t-

values of ±2.011 for a two-tailed test with a 5% significance level. The value of 2.011 is the critical t-value for the 5% level of significance (2.5% in one tail) for 48 degrees of freedom. A is incorrect because the mean of the short interest ratio is 192.3 ÷ 50 = 3.846. C is incorrect because the debt ratio explains 9.33% of the variation of the short interest ratio.

91.【单项选择题】Based on Exhibit 2, the short interest ratio expected for MQD Corporation is closest to:

A.3.8339.

B.5.4975.

C.6.2462.

参考答案：A

【莽学解析】The predicted value of the short interest ratio = 5.4975 + (-4.1589 × 0.40) = 5.4975 - 1.6636 = 3.8339.

92.【单项选择题】Based on Liu's regression results in Exhibit 2, the F-statistic for testing whether the slope coefficient is equal to zero is closest to:

A.-2.2219.

B.3.5036.

C.4.9367.

参考答案：C

【莽学解析】

$$F = \frac{\text{Mean regression sum of squares}}{\text{Mean squared error}} = \frac{38.4404}{7.7867} = 4.9367$$

【题干】Alef Associates manages a long-only fund specializing in global smallcap equities. Since its founding a decade ago, Alef maintains a portfolio of 100 stocks (out of an eligible universe of about 10,000 stocks). Some of these holdings are the result of screening the universe for attractive stocks based on several ratios that use readily available market and accounting data; others are the result of investment ideas generated by Alef's professional staff of five securities analysts and two portfolio managers.Although Alef's investment performance has been good, its Chief Investment Officer, Paul Moresanu, is contemplating a change in the investment process aimed at achieving even better returns. After attending multiple workshops and being approached by data vendors, Moresanu feels that data science should play a role in the way Alef selects its investments. He has also noticed that much of Alef's past outperformance is due to stocks that became takeover targets. After some research and reflection, Moresanu writes the following email to the Alef's CEO.Subject: Investment Process ReorganizationI have been thinking about modernizing the way we select stock investments. Given that our past success has put Alef Associates in an excellent financial position, now seems to be a good time to invest in our future. What I propose is that we continue managing a portfolio of 100 global small-cap stocks but restructure our process to benefit from machine learning (ML). Importantly, the new process will still allow a role for human insight, for example, in providing domain knowledge. In addition, I think we should make a special effort to identify companies that are likely to be acquired. Specifically, I suggest following the four steps which would be repeated every quarter.Step 1 We apply ML techniques to

a model including fundamental and technical variables (features) to predict next quarter's return for each of the 100 stocks currently in our portfolio. Then, the 20 stocks with the lowest estimated return are identified for replacement.Step 2 We utilize ML techniques to divide our investable universe of about 10,000 stocks into 20 different groups, based on a wide variety of the most relevant financial and non-financial characteristics. The idea is to prevent unintended portfolio concentration by selecting stocks from each of these distinct groups.Step 3 For each of the 20 different groups, we use labeled data to train a model that will predict the five stocks (in any given group) that are most likely to become acquisition targets in the next one year.Step 4 Our five experienced securities analysts are each assigned four of the groups, and then each analyst selects their one best stock pick from each of their assigned groups. These 20 "high-conviction" stocks will be added to our portfolio (in replacement of the 20 relatively underperforming stocks to be sold in Step 1).A couple of additional comments related to the above:
Comment 1 The ML algorithms will require large amounts of data. We would first need to explore using free or inexpensive historical datasets and then evaluate their usefulness for the ML-based stock selection processes before deciding on using data that requires subscription.
Comment 2 As time passes, we expect to find additional ways to apply ML techniques to refine Alef's investment processes.
What do you think?
Paul Moresanu

93.【单项选择题】The machine learning techniques appropriate for executing Step 1 are most likely to be based on:
A.regression
B.classification
C.clustering
参考答案：A
【莽学解析】The target variable (quarterly return) is continuous, hence this calls for a supervised machine learning based regression model.B is incorrect, since classification uses categorical or ordinal target variables, while in Step 1 the target variable (quarterly return) is continuous.C is incorrect, since clustering involves unsupervised machine learning so does not have a target variable.

94.【单项选择题】Assuming regularization is utilized in the machine learning technique used for executing Step 1, which of the following ML models would be least appropriate:
A.Regression tree with pruning.
B.LASSO with lambda (λ) equal to 0.
C.LASSO with lambda (λ) between 0.5 and 1.
参考答案：B
【莽学解析】It is least appropriate because with LASSO, when λ = 0 the penalty (i.e., regularization) term reduces to zero, so there is no regularization and the regression is equivalent to an ordinary least squares (OLS) regression.A is incorrect. With Classification and Regression Trees (CART), one way that regularization can be implemented is via pruning which will reduce the size of the regression tree—sections that provide little explanatory power are pruned (i.e., removed).C is incorrect. With LASSO, when λ is between 0.5 and 1 the

relatively large penalty (i.e., regularization) term requires that a feature makes a sufficient contribution to model fit to offset the penalty from including it in the model.

95.【单项选择题】Which of the following machine learning techniques is most appropriate for executing Step 2:

A.K-Means Clustering

B.Principal Components Analysis (PCA)

C.Classification and Regression Trees (CART)

参考答案：A

【莽学解析】K-Means clustering is an unsupervised machine learning algorithm which repeatedly partitions observations into a fixed number, k, of non-overlapping clusters (i.e., groups).B is incorrect. Principal Components Analysis is a long-established statistical method for dimension reduction, not clustering. PCA aims to summarize or reduce highly correlated features of data into a few main, uncorrelated composite variables.C is incorrect. CART is a supervised machine learning technique that is most commonly applied to binary classification or regression.

96.【单项选择题】The hyperparameter in the ML model to be used for accomplishing Step 2 is?

A.100, the number of small-cap stocks in Alef's portfolio.

B.10,000, the eligible universe of small-cap stocks in which Alef can potentially invest.

C.20, the number of different groups (i.e. clusters) into which the eligible universe of small-cap stocks will be divided.

参考答案：C

【莽学解析】Here, 20 is a hyperparameter (in the K-Means algorithm), which is a parameter whose value must be set by the researcher before learning begins.A is incorrect, because it is not a hyperparameter. It is just the size (number of stocks) of Alef's portfolio.B is incorrect, because it is not a hyperparameter. It is just the size (number of stocks) of Alef's eligible universe.

97.【单项选择题】The target variable for the labelled training data to be used in Step 3 is most likely which one of the following?

A.A continuous target variable.

B.A categorical target variable.

C.An ordinal target variable.

参考答案：B

【莽学解析】To predict which stocks are likely to become acquisition targets, the ML model would need to be trained on categorical labelled data having the following two categories: "0" for "not acquisition target", and "1" for "acquisition target".A is incorrect, because the target variable is categorical, not continuous.C is incorrect, because the target variable is categorical, not ordinal (i.e., 1st, 2nd, 3rd, etc.).

98.【单项选择题】Comparing two ML models that could be used to accomplish Step 3, which statement（ ） best describe（ ） the advantages of using Classification and Regression Trees

(CART) instead of K-Nearest Neighbor (KNN)?Statement I For CART there is no requirement to specify an initial hyperparameter (like K).Statement II For CART there is no requirement to specify a similarity (or distance) measure.Statement III For CART the output provides a visual explanation for the prediction.
A.Statement I only.
B.Statement III only.
C.Statements I, II and III.
参考答案：C
【莽学解析】The advantages of using CART over KNN to classify companies into two categories ("not acquisition target" and "acquisition target"), include all of the following: For CART there are no requirements to specify an initial hyperparameter (like K) or a similarity (or distance) measure as with KNN, and CART provides a visual explanation for the prediction (i.e., the feature variables and their cut-off values at each node).A is incorrect, because CART provides all of the advantages indicated in Statements I, II and III.B is incorrect, because CART provides all of the advantages indicated in Statements I, II and III.

99.【单项选择题】Assuming a Classification and Regression Tree (CART) model is used to accomplish Step 3, which of the following is most likely to result in model overfitting?
A.Using the k-fold cross validation method
B.Including an overfitting penalty (i.e., regularization term).
C.Using a fitting curve to select a model with low bias error and high variance error.
参考答案：C
【莽学解析】A fitting curve shows the trade-off between bias error and variance error for various potential models. A model with low bias error and high variance error is, by definition, overfitted.A is incorrect, because there are two common methods to reduce overfitting, one of which is proper data sampling and cross-validation. K-fold cross validation is such a method for estimating out-of-sample error directly by determining the error in validation samples.B is incorrect, because there are two common methods to reduce overfitting, one of which is preventing the algorithm from getting too complex during selection and training, which requires estimating an overfitting penalty.

100.【单项选择题】Assuming a Classification and Regression Tree (CART) model is initially used to accomplish Step 3, as a further step which of the following techniques is most likely to result in more accurate predictions?
A.Discarding CART and using the predictions of a Support Vector Machine (SVM) model instead.
B.Discarding CART and using the predictions of a K-Nearest Neighbor (KNN) model instead.
C.Combining the predictions of the CART model with the predictions of other models - such as logistic regression, SVM, and KNN - via ensemble learning.
参考答案：C
【莽学解析】Ensemble learning is the technique of combining the predictions from a collection of models, and it typically produces more accurate and more stable predictions than the best single model.A is incorrect, because a single model will have a certain error rate and will make noisy predictions. By taking the average result of many predictions from many models (i.e., ensemble learning) one can expect to achieve a reduction in noise as the average result

converges towards a more accurate prediction.B is incorrect, because a single model will have a certain error rate and will make noisy predictions. By taking the average result of many predictions from many models (i.e., ensemble learning) one can expect to achieve a reduction in noise as the average result converges towards a more accurate prediction.

101.【单项选择题】Regarding Comment #2, Moresanu has been thinking about the applications of neural networks (NNs) and deep learning (DL) to investment management. Which statement ( ) best describe ( ) the tasks for which NNs and DL are well-suited?Statement I NNs and DL are well-suited for image and speech recognition, and natural language processing.Statement II NNs and DL are well-suited for developing single variable ordinary least squares regression models.Statement III NNs and DL are well-suited for modelling non-linearities and complex interactions among many features.

A.Statement II only.

B.Statements I and III.

C.Statements I, II and III.

参考答案：B

【莽学解析】NNs and DL are well-suited for addressing highly complex machine learning tasks, such as image classification, face recognition, speech recognition and natural language processing. These complicated tasks are characterized by non-linearities and complex interactions between large numbers of feature inputs.A is incorrect, because NNs and DL are well-suited for addressing highly complex machine learning tasks, not simple single variable OLS regression models.C is incorrect, because NNs and DL are well-suited for addressing highly complex machine learning tasks, not simple single variable OLS regression models.

102.【单项选择题】Regarding neural networks (NNs) that Alef might potentially implement, which of the following statements is least accurate?

A.NNs must have at least 10 hidden layers to be considered deep learning nets.

B.The activation function in a node operates like a light dimmer switch since it decreases or increases the strength of the total net input.

C.The summation operator receives input values, multiplies each by a weight, sums up the weighted values into the total net input, and passes it to the activation function.

参考答案：A

【莽学解析】It is the least accurate answer because neural networks with many hidden layers—at least 3, but often more than 20 hidden layers—are known as deep learning nets.B is incorrect, because the node’s activation function operates like a light dimmer switch which decreases or increases the strength of the (total net) input.C is incorrect, because the node’s summation operator multiplies each (input) value by a weight and sums up the weighted values to form the total net input. The total net input is then passed to the activation function.