***(20 points)* Question 1**

In one of the class sessions on regression, we discussed the use of simple linear regression to estimate the relationship between the return on a stock and the market return. Think CAPM.

The slope, called the *beta coefficient*, is used to measure how responsive a stock’s price is to movements in the market. In financial investment parlance, the beta coefficient is used as a measure of a firm’s systematic risk.

In the file named *Question 1*, the return on the stock of three companies is provided: Dell, Sabre, and Walmart. Also provided is the return on the market (This is the value-weighted return computed by CRSP, the Center for Research on Security Prices).

Five years of monthly returns are used so that there are a total of 60 observations for each company.

Run the regression and answer the following questions:

1. *(5 points)*What are the beta coefficients for each of the three companies? Comment on the riskiness of each stock based on the beta coefficients.
2. *(5 points)*Is there a linear relationship between the firm return and the market return for each of these three companies? Please state your hypotheses, the -values, and your decision. Use a 5% level of significance.
3. *(5 points)*Test to see if Dell’s beta coefficient is greater than 1. Again, state your hypotheses, the -values, and your decision. Use a 5% significance level. *Reminder*: This is a nonstandard case.
4. *(5 points)*Test to see if Walmart’s beta coefficient is less than 1. Again, state your hypotheses, the -values, and your decision. Use a 5% significance level. *Reminder*: This is a nonstandard case.

***(20 points)* Question 2**

In a large public agency in Sacramento, management is concerned that a supervisor has been reporting unusual costs for supplies and equipment within her cost center. The average value seems reasonable, but individual amounts swing wildly and just don’t “look right.” The management would like to have a systematic method to audit purchase amounts that signal when the amounts may be due to fraud. No one wants to accuse staff of fraud falsely, but no one wants to miss the problem either. In pursuit of this objective, the managers collected a sample of invoices approved by the supervisor. The data are available in the file *Question 2*. It is mathematically known that when numerical values range over several orders of magnitude (which our invoice data indicates), the first digit should not be uniformly distributed, but the leading digit should follow Benford’s law. You are hired as a consultant to solve this case using a systematic, data-driven approach. Identify and run an appropriate statistical test to infer if the purchase activities constitute evidence of fraud. The management has asked for a report from your team. While you don’t have to create a report for answering this final exam question, but I would love to hear your thoughts on any strengths and shortcomings of your study that would go on the findings report, if you were to create one. In addition, what follow-up action items would you recommend on your report to the management of this public agency?

***(5 points)* Question 3**

Let us say that you run a regression with two explanatory variables and notice that the -value in the ANOVA table is extremely small but the -values of both explanatory variables are larger than 0.05. What is the probable reason for this? Can you conclude that neither explanatory variable does a good job in predicting the dependent variable?

***(5 points)* Question 4**

A biostatistician who designed a study for investigating the efficacy of a new medication at a leading U.S. pharmaceutical company was fired after the study. The tested null hypothesis stated that the drug is no better than a placebo. The -test yielded a value of 20 in a study of 400 subjects. The null hypothesis was rejected showing that the drug has a beneficial effect. Why did the management fire the biostatistician?

***(10 points)* Question 5**

The file named *Question 5* contains the following data for each of the 50 states:

total expenditures on a state’s criminal justice system (in millions of dollars) (EXPEND)

total number of police employed in the state (POLICE)

State governments must try to project spending in many areas. Expenditure on the criminal justice system is one area of continually rising cost. Your job is to build a model that can be used to forecast spending on a state’s criminal justice system. Use one of the functional forms discussed in the class to build a model. Once your model is complete, predict expenditures for a state that plans to hire 10,000 police officers. Find a point prediction and a 95% prediction interval. The functional forms notes and solved examples should help. (*Source*: These data are a bit dated and were obtained from the U.S. Department of Criminal Justice website.)

***(10 points)* Question 6**

The project manager at a construction company is evaluating how crew size affects the productivity of framing jobs. He has experimented with varying crew size (the number of workers) on a weekly basis over the past 27 weeks and has recorded productivity (jobs/week). The CSV file *Question6* contains the data.

1. *(2 points)*Create a scatterplot of the data. Based on the scatterplot alone, what crew size(s) seems optimal?
2. *(5 points)*Estimate the linear and the quadratic regression models. Evaluate the two models in terms of variable significance and adjusted . Which model provides the best fit? Provide an intuitive justification for the chosen model.
3. *(1 point)*Use the best-fitting model to predict how many jobs a crew of 5 could be expected to complete in a week.
4. *(2 points)*Estimate the cubic regression model. Does it improve the fit as compared to the quadratic regression model?

***(15 points)* Question 7**

Numerous attempts have been made to relate happiness to various factors. Since there is no unique way to quantify happiness, researchers generally rely on surveys to capture a subjective assessment of well-being. One study relates happiness with age and finds that holding everything else constant, people seem to be least happy when they are in their mid- to upper-40s (*The Economist*, December 16, 2010). Perhaps with greater age comes maturity that contributes to a better sense of overall well-being. With regard to the influence of money, a study from Princeton University’s Woodrow Wilson School suggests that money does buy happiness, but its effect diminishes as incomes rise above $75,000 a year (*Time Magazine*, September 6, 2010). Perhaps people do not need more than $75,000 to do what matters most to their emotional well-being, such as spending time with friends and family and meeting their basic food, health, and leisure needs. Nick Fisher is a young business school graduate who is fascinated by these reports. He decides to collect his own data to better comprehend and also verify the results of these studies. He surveys working adults in his hometown and inputs information on the respondent’s self-assessed happiness on a scale of 0 to 100, along with age and family income. The data is shown on the CSV file *Question7*.

1. *(13 points)*Run four models – linear, linear-log, log-linear, and log-log. Summarize the four models like my solved example shows. Compare the four functional forms and choose the one that fits the data the best.
2. *(1 point)*Using the model chosen in part a, predict happiness associated with varying levels of age for a family with income of $80,000.
3. *(1 point)*Using the model chosen in part a, predict happiness associated with varying levels of family income for a 60-year-old working adult.

***(10 points)* Question 8**

Traffic congestion on roads and highways costs industry billions of dollars annually as workers struggle to get to and from work. Several suggestions have been made about how to improve this situation, one of which is called *flextime* – workers are allowed to determine their own schedules (provided they work a full shift). Such workers will likely choose an arrival and departure time to avoid rush-hour traffic. In a preliminary experiment designed to investigate such a program, the general manager of a large company wanted to compare the times it took workers to travel from their homes to work at 8 a.m. with travel time under the flextime program. A random sample of 32 workers was selected. The employees recorded the time (in minutes) it took to arrive at work at 8 a.m. on Wednesday of one week. The following week, the same employees arrived at work at times of their own choosing. The travel time on Wednesday of that week was recorded. These results are listed in the CSV file *Question8*.

Run an appropriate statistical test to infer at the 5% significance level if travel times under the flextime program are different from travel times to arrive at work at 8 a.m.? Which conditions did you assess to select the statistical method?

***(5 points)* Question 9**

An auditor counted the number of filing errors associated with random samples of three types of mortgages at four banks. A filing error might, for instance, be a missing signature or incomplete paperwork. The mortgage types are traditional, refinancing, and home equity loan. What would it mean to find an interaction between the type of loan and the bank?

***(2 points)* Question 10**

What is the difference between a standard deviation and a standard error?

***(5 points)* Question 11**

Suppose that you have generated three alternative multiple regression models to explain the variation in a particular regressand. The regression output for each model can be summarized as follows:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Model 1 | Model 2 | Model 3 |
| No. of regressors | 4 | 6 | 9 |
|  | 0.76 | 0.77 | 0.79 |
| Adjusted | 0.75 | 0.74 | 0.73 |

All three models have the same regressand. Which of the models would you select and why? Be precise yet succinct.

***(10 points)* Question 12**

The *WinterFun* Company sells winter sports merchandise including skis, ice skates, sleds, and so on. Quarterly sales (in thousands of dollars) for the *WinterFun* company are shown on the CSV file *Question12*. The time period represented starts in the first quarter of 2008 and ends in the fourth quarter of 2017.

A linear regression containing only the time variable like shown below is called a **linear trend model**.

One of the common objectives of such analysis is if the linear trend model is sufficient for predicting sales or are the sales influenced by seasonality. For this question, follow the steps provided in the parts below to construct a final model:

1. *(3 points)*Create a linear trend model. Write the model. Examine the residual plot (versus fits) for this model. Does it show any patterns?
2. *(2 points)*Now, you want to see if there is any seasonality. For this create a time plot of sales (Sales versus Time) and check if there are seasonal patterns. Examine evidence of seasonality visually.
3. Create indicator variables for quarters. No need to show your R code.
4. *(5 points)*Conduct a Partial F-test to assess if the seasonal indicator variables are necessary in the model. Write the full model, the reduced models, and the hypotheses clearly. The reduced model in this case is the linear trend model and the full model is the one with indicator variables. Write your interpretation.

***(5 points)* Question 13**

A bank encouraged customers who have its debit cards to use them more frequently. It collected a sample of customers and found the number of times that each used the debit card during the four-month promotion March – June. To learn whether the average number of charges over this period is different in any month, the bank plans to use a one-way ANOVA. Would this be the right analysis? If not, what should the bank do?

***(25 points)* Question 14**

A technology company in the bay area is interested in conducting a study of the factors that affect absenteeism among its employees. After reviewing the literature on absenteeism and interviewing a number of employees, the analyst in charge of the project defined the variables as shown below:

Absenteeism (ABSENT): The number of distinct occasions that the employee was absent

during a year. Each occasion consists of one or more

consecutive days of absence.

Job Complexity (COMPLX): An index ranging from 0 to 100.

Seniority (SENIOR): Number of complete years with the company on December 31st of the year

in which the study was completed.

The CSV file *Question14* contains data on 77 employees. The dependent variable is absenteeism (ABSENT). The possible explanatory variables are

COMPLX = measure of job complexity

SENIOR = seniority

SATIS = categorical (qualitative) variable – response to “How satisfied are you with your manager?”

In this question, use SENINV = 1/SENIOR and COMPLX as two of the explanatory variables. The SENINV is the reciprocal of the seniority variable and the variable SATIS should be transformed into indicator variables as follows:

FS1 = 1 if SATIS = 1 (very dissatisfied)

= 0 otherwise

FS2 = 1 if SATIS = 2 (somewhat dissatisfied)

= 0 otherwise

FS3 = 1 if SATIS = 3 (neither satisfied nor dissatisfied)

= 0 otherwise

FS4 = 1 if SATIS = 4 (somewhat satisfied)

= 0 otherwise

FS5 = 1 if SATIS = 5 (very satisfied)

= 0 otherwise

Five indicator variables are created to represent all five supervisor satisfaction categories. Recall that only four need to be used in the regression.

1. *(3 points)* Run the regression containing all the independent variables (continuous + indicator). This is your full model.
2. *(3 points)* Run the regression after removing the indicator variables. This is your reduced model.
3. *(5 points)* Using Partial F-test, identify if the indicator variables (degree of satisfaction) are needed to determine the behavior of absenteeism in the employees. Use 5% level of significance.
4. *(3 points)* Use the appropriate model (after determining the final model from the results of the partial F-test) and calculate the average absenteeism rate for all employees with COMPLX = 60 and SENIOR = 30 who were very dissatisfied with their managers.
5. *(3 points)* Use the appropriate model (after determining the final model from the results of the partial F-test) and calculate the average absenteeism rate for all employees with COMPLX = 60 and SENIOR = 30 who were very satisfied with their managers.
6. *(3 points)* Use the appropriate model (after determining the final model from the results of the partial F-test) and calculate the average absenteeism rate for all employees with COMPLX = 10 and SENIOR = 3 who were very dissatisfied with their managers.
7. *(3 points)* Use the appropriate model (after determining the final model from the results of the partial F-test) and calculate the average absenteeism rate for all employees with COMPLX = 10 and SENIOR = 3 who were very satisfied with their managers.

Please note that the variable used for seniority in the regression model is the reciprocal and NOT the seniority value. Keep this in mind while answering parts d, e, f, and g above.

1. *(2 points)* How could this study be used by the management to help identify employees who might be prone to absenteeism?

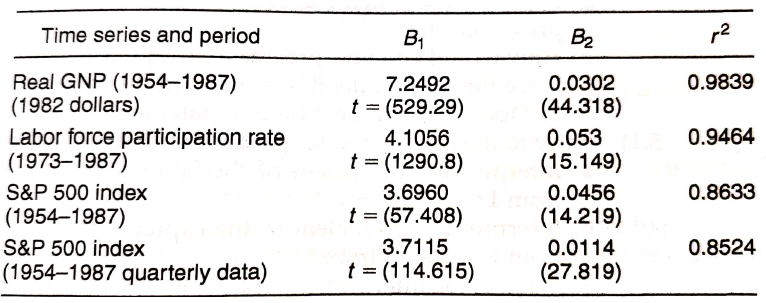
***(25 points)* Question 15**

The dataset on *Question15* file presents data from the case of *United States Department of the Treasury v. Harris Trust and Savings Bank* (1981). The data includes the salary of 93 employees of the bank, their education level, and their gender.

1. *(3 points)* Create a multiple regression model using education level and the male/female dummies.
2. *(5 points)* Interpret the differential intercept coefficient and the parameter estimate of the education level. Is there evidence of employment discrimination?
3. *(5 points)* Create a plot with the two regressions – one for male and another for female. Are these two regressions parallel, coincident, dissimilar, or concurrent regressions? Refer to my Week 9 note under the topic of Comparing Two Regressions.
4. *(5 points)* Now using the model from the above parts, determine if there is evidence of employment discrimination at Harris Bank even **after taking into account** education (i.e. interaction). Hint: For this, you will have to create an interaction variable EDUC\*MALES.
5. *(3 points)* Is the interaction significant? Is the dummy variable significant?
6. *(4 points)* What is the difference between the model with interaction variable and the model with only EDUCATION and the dummy variable?

***(8 points)* Question 16**

The growth model that we covered in the note on functional forms was fitted to several U.S. economic time series and the following results were obtained:



The B1 and B2 are the intercept and slope .

1. *(4 points)*In each case find out the instantaneous rate of growth.
2. *(4 points)*What is the compound rate of growth in each case?

***(5 points)* Question 17**

XYZ Realty sells homes along the East Coast of the United States. One of the questions most frequently asked by prospective buyers is: If we purchase this home, how much can we expect to pay to heat it during the winter? The research department at XYZ has been asked to develop some guidelines regarding heating costs for single-family homes.

Four variables are thought to relate to the heating costs:

the mean daily outside temperature,

the number of inches of insulation in the attic,

the age in years of the furnace, and

whether the home has a garage or not

To investigate, XYZ’s research department selected a random sample of 20 recently sold homes. It determined the cost to heat each home last January, as well as the January outside temperature in the region, the number of inches of insulation in the attic, the age of the furnace, and whether the home has a garage or not. The sample information is reported in CSV file *Question17*. Build a multiple regression model using the stepwise variable selection technique.

***(5 points)* Question 18**

In evaluating the performance of new hires, the HR division found that candidates with higher scores on its qualifying exam performed better. In a multiple regression that also used the education of the new hire as a regressor variable, the slope for the test score was zero. How can you explain this paradox to the HR manager?

***(5 points)* Question 19**

Managers in the HR department suspect that sick-day absentee rates are higher on some weekdays than others. What test cam they use to investigate this claim?

***(5 points)* Question 20**

A manager in the previous question thinks that the absentee rate is the same on Monday and Friday, but different from the rate on Tuesday and Thursday. What method should she use to test her suspicion?