MANB 1123 ASSIGNMENT #2

RULES:

1. This assignment should be conducted in pair.
2. Answer ALL questions given.
3. You may use any statistical packages/tools as you prefer to get the result.
4. Submission should follow the date given and should be in softcopy (via e-learning) and hardcopy.

HYPOTHESIS TESTING

Question 1:

Reviewers from the Oregon Evidence-Based Practice Center at the Oregon Health and Science University investigated the effectiveness of prescription drugs in assisting people to fall asleep and stay asleep. The Oregon reviewers, led by Susan Carson, M.P.H., concluded that Sonata was better than Ambien at putting people to sleep quickly, whereas patients on Ambien slept longer and reported having a better quality sleep than those taking Sonata. Samples taken by Carson and her associates are contained in a file entitled **Shuteye**. The samples reflect an experiment in which individuals were randomly given the two brands of pills on separate evenings. Their time spent sleeping was recorded for each of the brands of sleeping pills.

a) Does the experiment seem to have dependent or independent samples? Explain your reasoning.
b) Do the data indicate that the researchers were correct? Conduct a statistical procedure to determine this.
c) Conduct a procedure to determine the plausible differences in the average number of hours slept by those taking Ambien and Sonata.

Question 2:

The Center on Budget and Policy Priorities reported that average out-of-pocket medical expenses for prescription drugs for privately insured adults with incomes over 200% of the poverty level was $173 in 2002. Suppose an investigation was conducted in 2009 to determine whether the increased availability of generic drugs, Internet prescription drug purchases, and cost controls have reduced out-of-pocket drug expenses. The investigation randomly sampled 196 privately insured adults with incomes over 200% of the poverty level, and the respondents' 2009 out-of-pocket medical expenses for prescription drugs were recorded. These data are in the file **Drug Expenses**. Based on the sample data, can it be concluded that 2009 out-of-pocket prescription drug expenses are lower than the 2002 average reported by the Center on Budget and Policy Priorities? Use a level of significance of 0.01 to conduct the hypothesis test.

Question 3:

A treadmill manufacturer has developed a new machine with softer tread and better fans than its current model. The manufacturer believes these new features will enable runners to run for longer times than they can on its current machines. To determine whether the desired result is achieved, the manufacturer randomly sampled 35 runners. Each runner was measured for one week on the current

machine and for one week on the new machine. The weekly total number of minutes for each runner on the two types of machines was collected. The results are contained in the file **Treadmill**. At α = 0.02 level of significance, can the treadmill manufacturer conclude that the new machine has the desired result?

Question 4:

Cell phones are becoming an integral part of our daily lives. Commissioned by Motorola, a new behavioral study took researchers to nine cities worldwide from New York to London. Using a combination of personal interviews, field studies, and observation, the study identified a variety of behaviors that demonstrate the dramatic impact cell phones are having on the way people interact. The study found cell phones give people a newfound personal power, enabling unprecedented mobility and allowing them to conduct their business on the go. Interesting enough, gender differences can be found in phone use. Women see their cell phone as a means of expression and social communication, whereas males tend to use it as an interactive toy. A cell phone industry spokesman stated that half of all cell phones in use are registered to females.

a. State the appropriate null and alternative hypotheses for testing the industry claim.
b. Based on a random sample of cell phone owners shown in the data file called **Cell Phone Survey**, test the null hypothesis. (Use α = 0.05.)

Question 5:

It is a commonly held belief that SUVs are safer than cars. If an SUV and car are in a collision, does the SUV sustain less damage (as suggested by the cost of repair)? The Insurance Institute for Highway Safety crashed SUVs into cars, with the SUV moving 10 miles per hour and the front of the SUV crashing into the rear of the car.

SUV into Car	SUV Damage	Car Damage
Honda CR-V into Honda Civic	1721	1274
Toyota RAV4 into Toyota Corolla	1434	2327
Hyundai Tucson into Kia Forte	850	3223
Volkswagen Tiguan into VW Golf	2329	2058
Jeep Patriot into Dodge Caliber	1415	3095
Ford Escape into Ford Focus	1470	3386
Nissan Rogue into Nissan Sentra	2884	4560

Source: Insurance Institute for Highway Safety

(a) Why are these matched-pairs data?
(b) Draw a boxplot of the differenced data. Does the visual evidence support the belief that SUVs have a lower repair cost?
(c) Do the data suggest the repair cost for the car is higher? Use α = 0.05 level of significance.

Note: A normal probability plot indicates the differenced data are approximately normal with no outliers.

Question 6:

Do people walk faster in the airport when they are departing (getting on a plane) or when they are arriving (getting off a plane)? Researcher Seth B. Young measured the walking speed of travelers in San Francisco International Airport and Cleveland Hopkins International Airport. His findings are summarized in the table below. Do individuals walk at different speeds depending on whether they are departing or arriving at $\alpha = 0.05$ level of significance?

Direction of Travel	Departure	Arrival
Mean speed (feet per minute)	260	269
Standard deviation (feet per minute)	53	34
Sample size	35	35

Source: Seth B. Young. "Evaluation of Pedestrian Walking Speeds in Airport Terminals." *Transportation Research Record*. Paper 99-0824.

In further of his study, Seth B. Young want to find "Do business travelers walk at a different pace than leisure travelers"? Thus, he measured the walking speed of business and leisure travelers in San Francisco International Airport and Cleveland Hopkins International Airport. His findings are summarized in the table below. Determine whether business travelers walk at a different speed from leisure travelers at $\alpha = 0.05$ level of significance?

Type of Traveler	Business	Leisure
Mean speed (feet per minute)	272	261
Standard deviation (feet per minute)	43	47
Sample size	20	20

Source: Seth B. Young. "Evaluation of Pedestrian Walking Speeds in Airport Terminals." *Transportation Research Record*, Paper 99-0824.

REGRESSION ANALYSIS

Question 1:

Alex Court, the cost accountant for A & A Industrial Products, was puzzled by the repair cost analysis report he had just reviewed. This was the third consecutive report where unscheduled plant repair costs were out of line with the repair cost budget allocated to each plant. A & A budgets for both scheduled maintenance and unscheduled repair costs for its plants' equipment, mostly large industrial machines. Budgets for scheduled maintenance activities are easy to estimate and are based on the equipment manufacturer's recommendations. The unscheduled repair costs, however, are harder to determine. Historically, A & A Industrial Products has estimated unscheduled maintenance using a formula based on the average number of hours of operation between major equipment failures at a plant. Specifically, plants were given a budget of $65.00 per hour of operation between major failures. Alex had arrived at this amount by dividing aggregate historical repair costs by the total number of hours between failures. Then plant averages would be used to estimate unscheduled repair cost. For example, if a plant averaged 450 hours of run time before a major repair occurred, the plant would be allocated a repair budget of 450 hours × $65 = $29,250 per repair. If the plant was expected to be in operation 3,150 hours per year, the company would anticipate seven unscheduled repairs (3,150/450) annually and budget $204,750 for annual unscheduled repair costs. Alex was becoming more and more convinced that this approach was not working. Not only was upper management upset about the variance between predicted and actual costs of repair, but plant managers believed that the model did not account for potential differences among the company's three plants when allocating dollars for unscheduled repairs. At the weekly management meeting, Alex was informed that he needed to analyze his cost projections further and produce a report that provided a more reliable method for predicting repair costs. On leaving the meeting, Alex had his assistant randomly pull 64 unscheduled repair reports. The data are in the file **A & A Costs**. The management team is anxiously waiting for Alex's analysis.

- **(a)** Identify the major issue(s) of the case.
- **(b)** Analyze the overall cost allocation issues by developing a scatterplot of Cost v. Hours of Operation. Which variable, cost or hours of operation, should be the dependent variable? Explain why.
- **(c)** Fit a linear regression equation to the data.
- **(d)** Explain how the results of the linear regression equation could be used to develop a cost allocation formula. State any adjustments or modification you have made to the regression output to develop a cost allocation formula that can be used to predict repair costs.
- **(e)** Sort the data by plant.
- **(f)** Fit a linear regression equation to each plant's data.
- **(g)** Explain how the results of the individual plant regression equations can help the manager determine whether a different linear regression equation could be used to develop a cost allocation formula for each plant. State any adjustments or modification you have made to the regression output to develop a cost allocation formula.
- **(h)** Based on the individual plant regression equations determine whether there is reason to believe there are differences among the repair costs of the company's three plants.
- **(i)** Summarize your analysis and findings in a report to the company's manager.

Question 2:

The athletic director of State University is interested in developing a multiple regression model that might be used to explain the variation in attendance at football games at his school. A sample of 16 games was selected from home games played during the past 10 seasons. Data for the following factors were determined:

y : Game attendance
x_1 : Team win/loss percentage to date
x_2 : Opponent win/loss percentage to date
x_3 : Games played this season
x_4 : Temperature at game time

The data collected are in the file called **Football**.

a) Produce scatter plots for each independent variable versus the dependent variable. Based on the scatter plots, produce a model that you believe represents the relationship between the dependent variable and the group of predictor variables represented in the scatter plots.
b) Based on the correlation matrix developed from these data, comment on whether you think a multiple regression model will be effectively developed from these data.
c) Use the sample data to estimate the multiple regression models that contains all four independent variables.
d) What percentage of the total variation in the dependent variable is explained by the four independent variables in the model?
e) Test to determine whether the overall model is statistically significant. Use $\alpha = 0.05$.
f) Which, if any, of the independent variables is statistically significant? Use a significance level of $\alpha = 0.08$ and the *p*-value approach to conduct these tests.
g) Estimate the standard deviation of the model error and discuss whether this regression model is acceptable as a means of predicting the football attendance at State University at any given game.
h) Define the term *multicollinearity* and indicate the potential problems that multicollinearity can cause for this model. Indicate what, if any, evidence there is of multicollinearity problems with this regression model.
i) Develop a 95% confidence interval estimate for each of the regression coefficients and interpret each estimate. Comment on whether the interpretation of the intercept is relevant in this situation.

Question 3:

A nutritionist wants to develop a model that describes the relation between the calories, total fat content, protein, sugar, and carbohydrates in cheeseburgers at fast-food restaurants. She obtains the following data from the websites of the companies. She will use calories as the response variable and the others as explanatory variables.

(a) Construct a correlation matrix. Is there any reason to be concerned about multicollinearity?
(b) Find the least-squares regression equation.
(c) Test the regression coefficient.

Restaurant	Fat (g)	Protein (g)	Sugar (g)	Carbs (g)	Calories
1/4-pound single with cheese (Wendy's)	20	25	9	39	430
Whataburger (Whataburger)	32	30	10	61	640
Cheeseburger (In-n-Out)	27	22	10	39	480
Big Mac (McDonald's)	29	25	9	45	540
Whopper with cheese (Burger King)	47	33	11	52	760
Jumbo Jack (Jack in the Box)	42	25	12	54	690
1/4 Pounder with Cheese (McDonald's)	26	29	9	40	510
Cheeseburger (Sonic)	31	29	15	59	630

Source: Each company's Web site

(d) Test the regression slope. Should any of the explanatory variables be removed from the model? If so, which one?

(e) Determine the regression model with the explanatory variable identified in part (d) removed. Are the remaining slope coefficients significantly different from zero? If not, remove the appropriate explanatory variable and compute the least-squares regression equation.

(f) Interpret the regression coefficients for the least-squares regression equation found in part (e).

(g) Determine and interpret R^2 and the adjusted R^2.

(h) Construct 95% confidence and prediction intervals for the calories in a fast-food cheeseburger that has 38 g of fat, 29 g of protein, 11 g of sugar, and 52 g of carbohydrates. Interpret the results.

Question 4:

Researchers developed a model to explain the age gap between husbands and wives at first marriage. The model is below:

$\hat{y} = 3.8483 + 0.0321x_1 + 0.9848x_2 + 0.5391x_3 - 0.000145x_4^2$

Where;

y: Age gap at first marriage (male - female)

x_1: Percent of children aged 10 to 14 involved in child labor

x_2: Indicator variable where 1 is an African country, 0 otherwise

x_3: Percent of the population that is Muslim

x_4^2: Percent of the population that is literate

Source: Xu Zhang and Solomon W. Polachek, State University of New York at Binghamton "The Husband Wife Age Gap at First Marriage: A Cross-Country Analysis"

(a) Use the model to predict the age gap at first marriage for an African country where the percent of children aged 10 to 14 who are involved in child labor is 12, the percent of the population that is Muslim is 30, and the percent of the population that is literate is 75.

(b) What would be the mean difference in age gap between an African country and a non-African country?

(c) Interpret the coefficient of "percent of children aged 10 to 14 involved in child labor."

(d) The coefficient of determination for this model is 0.593. Interpret this result.

(e) The P-value for the test H_0: b1= 0 versus H_1: b1≠ 0 is 0.008. What would you conclude about this test?