Introduction to Statistics: Homework 4

Review Homework

DUE THURSDAY, December 9th

Please type your responses. When you are asked to "interpret the coefficient" you should evaluate the statistical significance of the coefficient as well as the magnitude of the relationship. ***Be sure to interpret coefficients in the context of the full regression!***

1. Let's say we are interested in how exercise affects weight loss. Our dependent variable is weight loss (in pounds) over the past year (negative values mean that the individual *gained* weight). We estimate a model predicting change in weight with a variable that measures the number of days per week the individual exercises (0 to 7). We get the following regression estimates:

	Coefficient	Standard Error	T
# of days per week exercise	6.6	1.1	6.0
Constant	-8.0	.9	8.9

   a. Interpret the coefficient on # of days per week exercised. [4 points]
   b. Interpret the coefficient on the constant. [4 points]
   c. Based on this model, what would be the expected change in weight for someone who exercises 3 days per week? [4 points]
   d. The estimate of the relationship between exercise and weight loss is probably biased. Describe one confounding variable that you could control for that would reduce this bias and explain why controlling for this variable would lead to a less biased estimate of the relationship between exercise and weight loss. [10 points]

2. Apart from a failure to control for confounding factors, we might also think that the relationship between exercise and weight loss is different for men and women. For example, there is some evidence that men tend to lose weight faster than women. We can test this using an interaction term.

	Coefficient	Standard Error	T
# of days per week exercise	9.6	1.1	8.7
Gender (1=female, 0=male)	-2.3	1.0	-2.3
Days exercise x Gender	-2.4	1.2	-2.0
Constant	-8.0	.9	8.9

   a. Interpret the coefficient on each of the two component terms (# of days per week exercise and Gender) [8 points]

b. What does the statistical significance of the coefficient on the interaction term tell us? [6 points]
c. What is the estimated (slope of the) relationship between days per week exercise and weight loss among men? Among women? [8 points]

3. Most Americans celebrate the winter holiday season in some way, regardless of their religious affiliation. Let's say we are interested in estimating the relationship between religious affiliation and spending on gifts during the winter holiday season. We predict the number of dollars an individual spends based on indicators for religious affiliation (leaving "atheist or agnostic" as the omitted category) and a measure of household income (in tens of thousands of dollars; for example, $40,000 per year = 4).

	Coefficient	Standard Error	T
Protestant	150.2	40.2	3.7
Roman Catholic	115.5	38.7	3.0
Jewish	95.2	35.4	2.7
Other Religion	45.6	34.4	1.3
Income ($10,000s)	256.4	30.8	8.3
Constant	85.0	34.2	2.5

a. Interpret the coefficient on Other Religion [4 points]
b. Interpret the coefficient on Income [4 points]
c. What is the predicted amount of holiday season spending for a Roman Catholic with a household income of $200,000 per year? [4 points]
d. What is the predicted amount of holiday season spending for a Protestant with a household income of $2,500,000 per year? [4 points]
e. In class we talked about transforming independent variables as a way of estimating non-linear relationships. In this case we might want to use a logged (natural logarithm) measure of income, rather than the linear measure that is included. Draw a quick sketch of the type of relationship a logged measure would allow us to estimate (household income in $10,000s on the x-axis and holiday spending on the y-axis). Why might this relationship fit the data better than a linear measure of income? [10 points]

4. We want to estimate the relationship between an individual's age and the amount of money their health care costs per year (in dollars). We estimate the following model:

	Coefficient	Standard Error	T
Age	-170	40	4.3
Age-squared	2.5	1.0	2.5
Gender (1=female, 0=male)	220	44	5.0
Constant	5000	250	20.0

a. What does the coefficient on age-squared tell us? [4 points]

b. Fill in this table with predicted values *for females* [10 points]

Age	Predicted value (health care cost)
10	
30	
50	
70	
90	

c. Sketch a graph of this relationship and describe the estimated relationship between age and health spending. [10 points]

5. How to improve the education system in the US is a matter of much debate. Some people have argued that more students should be encouraged to attend private schools, because they think private schools do a better job of educating students.

   a. Suppose we examine the test scores of the elementary school students in New Haven and find that, on average, students enrolled in private schools score 20 percent higher on achievement tests than those enrolled in public schools. Should we conclude that private schools in New Haven provide better education than public schools? Why or why not? [10 points]

   b. Imagine that a pilot program was instituted where some New Haven students entering grade school were randomly assigned to attend private school and the city would pay the tuition. Four years later we compare the test scores of those who were randomly selected to attend private school (funded by city money) and those who were not. We find that the test scores across these two groups are the same. What are the strengths of this approach to assessing the benefits of private schooling compared with the comparison described in a) above? If we estimated a regression model predicting test scores with an indicator variable (1=selected to attend private school; 0=not selected), would we need to control for other variables? Why or why not? [10 points]