Homework #4

answers

Problem 1 Fire damage in the United States amounts to billions of dollars each year, much of it insured. The time taken for firefighters to arrive at the fire is critical. This raises the question, should insurance companies lower premiums if the home to be insured is close to a fire station? To help make a decision, a study was undertaken wherein a number of fires were investigated. The distance to the nearest fire station (in miles) and the percentage of fire damage were recorded (dataset Xr16-11). Determine the least squares regression line and interpret the coefficients.

The estimated regression line is $Damage = 23.11 + 5.35 * Distance + \epsilon$. This means that being 1 mile farther away from a fire station increases the percent of fire damage by 5.35. The coefficient estimate is highly statistically significant.

Problem 2 Each year in the US, there are about 12 suicides per 100,000 people. In a December 12, 2013 article on slate.com, two economists describe their research studying the link between gun ownership and number of suicides. Their research relies on a regression of the gun suicides per capita on the percentage of individuals living in households with at least one gun, across states.

$$Suicide\_rate_i = \beta_0 + \beta_1 * Guns_i + \epsilon \tag{1}$$

where $Suicide\_rate_i$ is state $i's$ number of suicides per 100,000 people, and $Guns_i$ is state i's percentage of individuals living in households with at least one gun. For example, if 45.6% of Kentuckians live in a household with at least one gun, $Guns_i = 45.6$ for Kentucky.

The authors obtain the following estimates:

Regression	Statistics
F statistic	134.479
R Square	.3952
Observations	50

	coefficients	Standard error	t stat	P-value 95%
Intercept	11.456	.5879	19.4863	5.46E-89
Guns	.06	.028571	.3524	.0357

a. Do the authors find a significant effect of gun ownership on the number of suicides? Explain.

Yes. The p-value is .0357, so Guns are significant at the 5% level.

b. Interpret the estimate of β_1 in plain English. (hint: be very careful with the units.)

A one percentage point increase in the percentage of individuals living in households with guns is correlated with an increase of .06 in the number of suicides per 100,000 people.

c. Suggest one "control variable" which could be added to equation r̊egress to account for factors which affect both guns and suicides. Explain.

There are many good answers. For example, the unemployment rate may be correlated with both gun sales and the suicide rate, so it would be an ideal control variable.

Problem 3 Consider the following regression equation:

$$y = \beta_0 + \beta_1 * X + \epsilon \tag{2}$$

a. Explain in plain English how to interpret the coefficient β_1.

β_1 is the amount by which y increases given a one-unit increase in x.

b. Explain in plain English why the ϵ term is necessary.

The ϵ term represents noise, or luck. When we model the relationship between two variables using a linear regression model, we do not expect that there is an exact linear relationship between y and x, but instead expect that the average value of y has a linear relationship with x. The ϵ term is sometimes positive, sometimes negative, meaning that data will sometimes be above and sometimes below a regression line.

c. Suppose that x is mileage and y is the selling price of a certain type of used car. You estimate r̊egress using data provided by a local dealership. Do you expect the regression results to indicate that $\hat{\beta}_1 > 0$ or that $\hat{\beta}_1 < 0$? Why?

We would expect $\hat{\beta}_1 < 0$, as cars with greater mileage will, on average, sell for a lower dollar amount than cars with lower mileage.

d. Suppose that x is the number of McDonald's restaurants in a country and y is the GDP of a country. You estimate r̊egress using a dataset containing information on 150 countries, and find a large, positive, and statistically significant estimate $\hat{\beta}_1$. Does this mean that McDonald's restaurants cause a high GDP? Explain.

No, it means that there is a correlation between the number of McDonald's in a country and GDP, but it does not mean that the relationship is causal. For example, it may be that a high GDP is something McDonald's looks for when entering a country, and so the causality may even go from y to x, and not the other way around.

Problem 4 Consider the following regression model:

$$WAGE = \beta_0 + \beta_1 * EXPERIENCE + \epsilon \tag{3}$$

where $WAGE$ is hourly wage and $EXPERIENCE$ is years of full-time work experience. Suppose you estimate the regression parameters using a large Census dataset (a random sample of all U.S. residents). Your results are as follows:

Regression	**Statistics**
Multiple R	.22358
R Square	.24985
Adjusted R Square	.24985
Standard error	5.2198
Observations	189,253

	coefficients	Standard error	t stat	P-value	Lower 95%	Upper 95%
Intercept	7.11	2.00	3.56	.000186	4.0924	14.1276
EXPERIENCE	.96	.31	3.097	.000978	.3524	1.5676

a. Consider the following hypothesis test:

$$H_0 : \beta_1 = 0$$
$$H_A : \beta_1 \neq 0$$

Is there sufficient evidence to support rejecting H_0, with $\alpha = .05$? How do you know?

Yes, we would reject H_0 in favor of H_A here, as the p-value associated with the test is $.000978 < .05$.

b. Interpret the estimate of β_1 precisely.

The estimate of .96 means that one year of additional full-time work experience is associated with a 96 cent hourly pay increase, on average.

Differing levels of full-time experience is often suggested as one reason why women earn a lower wage than men. Suppose that, among 45-year old workers, women have an average of 17.7 years of full-time work experience, while men have an average of 22.6 years of full time work experience.

c. Give a point estimate for the wage of a worker with 17.7 years of experience. Give a point estimate for the wage of a worker with 22.6 years of experience.

Using the estimated regression line $WAGE = 7.11 + .96 * EXPERIENCE$, we get point estimates of \$24.10 and \$28.81, respectively.

d. Suppose that the average 45-year old man earns a wage which is \$6.76 higher than the average wage of a 45-year old woman. What fraction of this difference is explained by differing levels of full-time work experience, according to the results?

The results of part c predict that differing levels of full-time work experience account for \$4.70 of the difference between men and women, or fraction .696 of the difference.

Problem 5 Answer the following questions related to the 2008 General Social Survey (GSS2008.xls).

a. Conduct an analysis of the relationship between income (INCOME) and age (AGE). Estimate with 95% confidence the average increase in income for each additional year of age. (hint: you may need to drop observations from the dataset for which you are missing information on either income or age)

Dropping all observations with missing data for either age or income, the estimated regression line is $Income = 22921.57 + 436.82 * Age + \epsilon$.

b. Is there sufficient evidence to conclude that more educated people (EDUC) watch less television (TVHOURS)?

After dropping observations for which there is missing data on EDUC or TVHOURS, we are left with 1322 observations. The estimate regression line is $TVHours = 5.79 - .21 * Educ + \epsilon$. The education coefficient is highly significant.

c. What does the dataset say about the relationship between education and income?

For the 1189 individuals with data on both, the estimated regression line is $Income = -28926 + 5110 * Educ + \epsilon$. The education variable is a highly significant correlate of income.

Problem 6 A computer dating service typically asks for various pieces of information such as height, weight, and income. One such service requests the length of index fingers. The only plausible reason for this request is to act as a proxy on height. Women in particular have often complained that men lie about their heights. If there is a strong relationship between heights and index fingers, the information can be used to "correct the false claims about heights. To test the relationship between the two variables, researchers gathered the heights and lengths of index fingers (in centimeters) of 121 students (dataset Xr16-107).

a. Using a computer, draw a scatter plot depicting the relationship between the two variables.

b. Is there sufficient evidence to infer that height and length of index fingers are linearly related?

c. Predict with 95% confidence the marginal increase in height associated with a 1cm increase in index finger length.

The estimated regression line is $Height = 119.07 + 7.29 * Index + \epsilon$. The index finger variable is a highly significant correlate of height. The 95% confidence interval for β_1 is $[5.39, 9.19]$.

Problem 7 A researcher is interested in the effect of the number of police officers working in a city on that city's violent crime rate. She obtains data on the 500 largest cities in the US, with the 2012 violent crime rate and the number of police officers per capita employed in 2012 for each city. She then uses a computer to regress the violent crime rate on police per capita, and finds the following results:

$$R^2 = .64$$

	Coefficients	Standard error	t stat	P-value
Intercept	604.32	158.65	3.81	$1.39 * 10^{-4}$
Police per capita	22.13	10.43	2.12	.034

a. Interpret the regression results. What do the results predict is the effect of hiring one more police officer? Is this effect significant?

The results suggest that an increase of 1 in police per capita is correlated with an increase in crime of 22.13. That is, more police lead to more crime.

b. A colleague points out that the positive coefficient estimated for the police variable suggests that police cause crime. Can you suggest an alternative explanation?

The most likely alternative explanation is that the cities with higher crime rates have hired more police officers per capita to combat the crime. That is, the data exhibit reverse causality. Not only do police (presumably) deter crime, but, across cities, crime causes police, and the latter effect is evidently larger than the former.

Problem 8 An economist for the federal government is attempting to produce a better measure of poverty than is currently in use. To help acquire information, she recorded the annual household income (in thousands of dollars) and the amount of money spent on food during one week for a random sample of households (dataset Xr16-15).

a. Determine the regression line and interpret the coefficients.

There are no obvious clues here that one variable should be the y-variable and one the x-variable here (though probably food should be the dependent variable, as it is a function of income, but if you did it the other way, this is fine too). Regressing food on income, we get:

$$Food = 153.899 + 1.958 * Income \tag{4}$$

This means that an increase in annual income of \$1,000 leads to an increase in weekly food spending of \$1.96, suggesting that about 10% of annual income is spent on food (as there are 52 weeks in a year, and 52*\$1.96=\$101.92, which is 10% of \$1,000.

b. Determine the coefficient of determination and describe what it tells you.

The coefficient of determination, or R^2 is .245871. This will not depend on which variable was the dependent variable. This means that about 25% of the variation in one variable is explained using variation in the other variable. To put it another way, knowing the value of income reduces the uncertainty associated with food by 25%.

c. Conduct a test to determine whether there is evidence of a linear relationship between household income and food budget.

The P-value for income is $1.1 * 10^{-10}$ (again, if you regressed income on food, you will get the same P-value). This means that income is a highly statistically significant variable, and so there is strong evidence of a linear relationship between income and food.