

Problem Set 2
ECON 30331
(Due Thursday, September 10th)

Dan Hungerman
Fall 2009

1. Suppose a researcher is interested in estimating the linear regression model $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ and in a sample of 48 points, the following descriptive statistics are generated:

$$\bar{x} = 30, \bar{y} = 63, \sum_{i=1}^n (x_i - \bar{x})^2 = 6900, \sum_{i=1}^n (y_i - \bar{y})^2 = 29,000$$

$$\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = 13,800$$

What are the OLS estimates of $\hat{\beta}_0$ and $\hat{\beta}_1$? The OLS estimates generate a value of $\hat{\sigma}_e^2 = 36$. What is the standard error on the estimate of $\hat{\beta}_1$? What is the R^2 for this model?

2. Using data from the 2004 baseball season, a researcher collects data on the number of wins a team had during the year and payroll in millions of dollars. The researcher wants to estimate a model to examine whether the size of the payroll alters wins, so they want to consider an OLS model of the form $wins_i = \beta_0 + \beta_1 payroll_i + \varepsilon_i$. The author gets as far as getting descriptive statistics and the correlation coefficient between wins and payroll (presented below), then their computer crashes. Using the data below, calculate just the estimates for $\hat{\beta}_0$ and $\hat{\beta}_1$. Interpret the results for $\hat{\beta}_1$. According to the model estimates, by how much will wins increase if a team spend \$15 million more on salary?

. sum wins payroll

Variable	Obs	Mean	Std. Dev.	Min	Max
wins	30	80.96667	13.36615	43	101
payroll	30	70.13708	27.26755	19.63	149.711

. corr wins payroll
 (obs=30)

	wins	payroll
wins	1.0000	
payroll	0.4176	1.0000

3. Suppose a researcher is interested in estimating the impact of gasoline taxes (X_i) on per capital gallons of gasoline consumed per year (Y_i). Assume tax is measured in cents per gallon. The researcher has data from 51 states for a 10 year period for a total of 510 observations. The researcher estimates the linear OLS model $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ and calculates $\hat{\beta}_1 = -0.90$. Suppose instead of measuring taxes in cents per gallon, the researcher measures taxes in dollars per gallon where the new model is $Y_i = \gamma_0 + \gamma_1 X_i + \varepsilon_i$

and $X^*=X/100$. What will be the estimate on the coefficient on γ_1 ? Suppose taxes are measured in cents as in the first case, but consumption is measured as gallons consumed per month, where $Y^*_i=Y_i/12$. The model now is of the form $Y_i^*=\alpha_1 + \alpha_2 X_i+\varepsilon_i$. What will be the estimate on α_2 ?

- On the Stata page for this class is a copy of a file called `state_cig_data.dta`. Download the data set then load that data set up into Stata and construct two new variables: The natural log of per capita consumption of cigarette packs, $\ln(Q)$ and the natural log of the real retail price ($\ln(P)$). For this second variable, use the statement, `gen ln_r_price=ln(retail_price/cpi)`. Next, run a regression of $\ln(Q)$ on $\ln(P)$, or the model $\ln(Q_i)=\beta_0 +\beta_1 \ln(P_i)+\varepsilon_i$. What are the estimates for $\hat{\beta}_1$ and $\hat{\beta}_0$ and what is the R^2 for the model? Next, interpret the estimate for $\hat{\beta}_1$. Be precise, explain the units of measure on the variable and give a numeric example.
- Below are STATA results from a SUMmary statement and a REGression, but some of the results have been whited-out. Please provide estimates for a, b and c.

. sum y x

Vari able	Obs	Mean	Std. Dev.	Mi n	Max
y	19906	2. 38902	. 5345824	. 2464662	3. 505295
x	19906	a	2. 795234	0	18

. reg y x

Source	SS	df	MS	Number of obs =	19906
Model	853. 046763	1	853. 046763	F(1, 19904) =	3511. 42
Resi dual	4835. 37212	19904	. 242934693	Prob > F =	0. 0000
Total	C	19905	. 285778392	R-squared =	b
				Adj R-squared =	0. 1499
				Root MSE =	. 49288

	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
y					
x	. 0740606	. 0012498	59. 26	0. 000	. 0716109 . 0765103
_cons	1. 414289	. 016816	84. 10	0. 000	1. 381328 1. 44725

- A researcher is interested in examining the impact of illegal music downloads on commercial music sales. The author collects data on commercial sales of the top 500 singles from 2007 (Y) and the number of downloads from a web site that allows ‘file sharing’ (X). The author estimates an OLS model of the form $Y_i=\beta_0 +\beta_1 X_i+\varepsilon_i$ and the author gets a large positive coefficient on the estimated parameter $\hat{\beta}_1$. The author concludes these results demonstrate that downloads actually spur on music sales. Is this an unbiased estimate of the impact of illegal music on sales? Why or why not? Do you expect the estimate to overstate or understate the true relationship between Y and X.
- A pharmaceutical company is interested in estimating the impact of a new drug on cholesterol levels. They enroll 200 people in a clinical trial. People are randomly assigned dosage levels or they are randomly assigned into the control group. Half of the

people are given dosages of the new drug and half the people are given a sugar pill with no active ingredient. To examine the impact of dosage on reductions in cholesterol levels, the authors of the study regress change in cholesterol levels (Y) on dosage level (X). For people in the control group, $x=0$ and for people in the treatment group, x measures milligrams of active ingredient and the model they estimate is of the form $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$. In this case, the authors find a statistically significant and negative coefficient on the estimate for β – larger dosages reduce cholesterol levels. Is this an unbiased estimate of the impact of dosage on change in cholesterol level? Why or why not? Do you expect the estimate to be too large or too small?

8. A researcher estimates the linear model $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ via OLS and assume in this case that $\bar{y} = \bar{x} = 0$. What is the estimate for $\hat{\beta}_0$ and provide an equation for the estimate of $\hat{\beta}_1$.
9. An author wants to estimate a model of the form $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$. Unfortunately, the author is unable to find exactly the variable for y but instead, they find a variable that has some ‘measurement error’, that is, the variable they use in the regression is y_i^* and $y_i^* = y_i + v_i$ where v_i is a random error with $E[v_i] = 0$, $V(v_i) = \sigma_v^2$ and $\text{Cov}(v_i, \varepsilon_i) = \text{Cov}(v_i, x_i) = 0$. Think of the problem this way. A survey asks people for their usual weekly earnings and instead of responding with their exact earnings, (y_i) they give an approximation y_i^* that varies randomly where the error in their response is given by the random variable (v_i). Suppose the researcher estimates the model with y_i^* instead of y , gets an estimate for $\hat{\beta}_1$ that equals the following

$$\hat{\beta}_1^* = \frac{\sum_{i=1}^n (y_i^* - \bar{y}^*)(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

In this example, will the use of y_i^* instead of the true y_i generate biased estimates?

HINT: Write the numerator as $\sum_{i=1}^n y_i^*(x_i - \bar{x})$ and recall that $y_i^* = y_i + v_i$. Substitute in the true value for y_i^* into the model and then take expectations.

10. (Bonus Problem, pretty hard) Consider a bivariate regression model of the form $y_i = \beta_0 + x_i \beta_1 + \varepsilon_i$. Show that the square of the correlation coefficient between y_i and \hat{y}_i is equal to the R^2 .

HINT: Write the definition of the squared correlation coefficient and there should be $\sum_{i=1}^n (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})$ in the numerator. Write this as $\sum_{i=1}^n y_i(\hat{y}_i - \bar{\hat{y}})$ and remember that $y_i = \hat{y}_i + \hat{\varepsilon}_i$