Intro to Prob and Stat II
GSIA, Carnegie Mellon University
45-734, Spring 2002 (mini 4)
Midterm, Thursday, April 11, 2002

1. **(25 points)** You are negotiating rates for hospital care for an HMO. In your market, it is typical for hospitals to be paid a per diem (to be paid some number of dollars per day that your insureds spend in the hospital). Thus, you are interested in getting an idea of how much it costs a hospital to treat a patient for a day.

   You have a dataset on California hospital which contains quarterly costs in millions of dollars and inpatient days in thousands of days (inpatient days are the total number of days spent in the hospital by all patients over a quarter). There are 7535 observations.

   Here are some statistics.

Variable	Mean	Std Dev
costs	16.27	19.37
days	20.79	24.17

   In addition, the covariance between costs and days is 365.57.

   (a) **(5 points)** Calculate and interpret the correlation between costs and days.

   Let X denote days, and Y denote costs. The sample correlation coefficient r can be computed in terms of the sample standard deviations s_x, s_y given above as:

   $$r = \texttt{corr}(x,y) = \frac{\texttt{cov}(x,y)}{s_x s_y} = \frac{365.57}{24.17 \times 19.37} = 0.781$$

   This indicates that the two variables `days` and `costs` have a reasonably strong, though not perfect, positive linear relationship.

(b) **(20 points)** Calculate and interpret the intercept and slope of a regression of costs on days.

We can use a formula for b_{OLS} derived in terms of the data that we have, as follows:

$$b_{OLS} = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sum(x_i - \overline{x})^2} = \frac{\texttt{cov}(x, y)}{s_x^2} = \frac{365.57}{24.17^2} = 0.626$$

After this, we can compute a_{OLS}:

$$a_{OLS} = \overline{y} - b_{OLS}\overline{x} = 3.255$$

Scaling the units, we get our estimated linear regression equation:

$$\texttt{costs} = \$3,255 + \$626 \times \texttt{days}$$

The intercept (in this scaled line) is \$3,255. The mathematical meaning is that our best estimate is that it costs the hospital \$3,255 when a patient is admitted for zero days. A more meaningful interpretation is that \$3,255 is the base or fixed cost (on average) for treating a patient who is admitted to the hospital.

The slope is \$626. This is interpreted as our best estimate for the additional cost (on average) for a patient staying one additional day in the hospital.

2. **(15 points)** For this and the following questions, refer to the mpg dataset available on the website.

   The federal government requires that auto makers maintain a minimum level of fuel economy (measured by miles per gallon) for the vehicles they sell. It is therefore important for them to understand how design decisions affect the fuel efficiency of cars.

   Please build a model to estimate the effects of weight, horsepower, displacement, and number of cylinders on fuel efficiency. Write down your model and your coefficient estimates. Interpret one of your estimates.

   We use a Classical Linear Regression Model for estimating the effects of the four parameters on fuel efficiency.

$$\begin{aligned}\texttt{mpg}_i = \alpha \quad & +\beta_{wt}\texttt{weight}_i + \beta_{hp}\texttt{horsepower}_i \\ & +\beta_{disp}\texttt{displacement}_i + \beta_{cyl}\texttt{cylinders}_i + \epsilon_i\end{aligned}$$

   We estimate the coefficients in Eviews by running the linear regression estimation command:

`ls mpg c weight horsepower displacement cylinders`.

We obtain the following estimates:

$$\begin{aligned}\texttt{mpg}_i = 45.757 \quad & -0.00528\ \texttt{weight}_i - 0.0428\ \texttt{horsepower}_i \\ & +0.000139\ \texttt{displacement}_i - 0.393\ \texttt{cylinders}_i\end{aligned}$$

The coefficient $b_{wt} = -0.00528$ means that keeping all other factors constant, our best estimate of the effect of a 1 lb. increase in the weight of the car is that it causes its fuel efficiency to decrease by 0.00528 miles per gallon.

3. **(15 points)** Please give me your best guess as to the effect of a 1 lb. increase in weight on mpg. What is a 90% confidence interval for this quantity and what does it mean?

Our best guess for the effect of a 1 lb. increase in weight on mpg is precisely b_{wt}, which is the OLS regression estimate for β_{wt}. In other words, our best guess is that a 1 lb. increase in weight causes mpg to decrease by 0.00528 miles per gallon (keeping all other factors constant).

We now compute a 90% confidence interval for β_{wt}. Note that Eviews discarded two observations, so we have $n = 392$ and $K = 4$. Since the number of data points is so high, we will use the normal distribution to compute the confidence interval. However, using the t-distribution gives the same value due to the high number of observations.

$$\begin{aligned}90\%\ CI &= b_{wt} \pm z_{\alpha/2} s_{b_{wt}} \\ &= b_{wt} \pm z_{0.05} s_{b_{wt}} \\ &= -0.00528 \pm 1.645 \times 0.000717 \\ &= [-0.00646, -0.00410]\end{aligned}$$

This means that we are 90% confident that the effect of a 1 lb. increase in weight is that the fuel efficiency decreases by between 0.00410 and 0.00646 miles per gallon.

4. **(15 points)** Please test, at the 5% level the hypothesis that more powerful engines (as measured by horsepower) have the same fuel efficiency as do less powerful engines.

We test the null hypothesis that the coefficient β_{wt} is zero against the two-sided alternative, though it is justifiable to do a one-sided test too.

$$\begin{aligned}H_O &: \beta_{wt} = 0 \\ H_A &: \beta_{wt} \neq 0\end{aligned}$$

In the Eviews output, we observe that the p-value of the coefficient `weight`is 0.0010. Since this is much less than 5%, we can reject the null hypothesis.

Alternatively, we can do the statistical test. We compute the t-statistic:

$$t = \frac{b_{hp}}{s_{b_{hp}}} = \frac{-0.0428}{0.0129} = -3.318$$

We also find that $t_{387,0.025} = 1.960$.

Our rule is that we reject H_O if $|t| > t_{387,0.025}$. Since this is true, we reject the null hypothesis in favour of the two-sided alternative.

We therefore have sufficient statistical evidence to conclude that horsepower does have an impact on fuel efficiency.

5. **(15 points)** Please test, at the 5% level the hypothesis that neither displacement nor cylinders affect fuel efficiency.

$$\begin{array}{ll} H_O: & \beta_{disp} = \beta_{cyl} = 0 \\ H_A: & \text{at least one of } \beta_{disp}, \beta_{cyl} \text{ is not zero.} \end{array}$$

In Eviews, we can easily do this test by running the original regression model, and then typing `drop displacement cylinders`. Another way to do it in Eviews is the Wald test: `wald c(4)=0, c(5)=0`.

In both cases, we get an F-statistic of 0.845 and a p-value of 0.43. Hence we accept the null hypothesis.

Alternatively, we could do the F-test. The original regression already gives us $SSE_{UR} = 6963.43$. We run a restricted model, by regressing `mpg`only on `weight`and `horsepower`. We find $SSE_R = 6993.85$. Our rule is to reject the null hypothesis if $F > F_{2,387,0.05} = 3.00$.

$$F = \frac{(SSE_R - SSE_{UR})/Q}{SSE_{UR}/(n-K-1)} = \frac{(6993.85 - 6963.43)/2}{6963.43/387} = 0.845$$

Since this is less than 3, we accept the null.

That is, we find sufficient statistical evidence at the 5% level that neither displacement nor cylinders affect fuel efficiency.

6. **(15 points)** How does your answer to problem 5 change if you estimate a model without horsepower (ie with an unrestricted model containing all the same variables but not horsepower).

Interpret the meaning of any difference.

In Eviews, we do this by first estimating a new unrestricted model: `ls mpg c weight displacement cylinders`. We then do `drop displacement`

`cylinders` (or the wald test), and get an F-statistic of 4.023 with a p-value of 0.0187. Since this is less than 5%, we now reject the null hypothesis.

Alternatively, we could compute the sums of squared errors, and get $SSE_{UR} = 7247.34$ and $SSE_R = 7396.85$. The computed F-stat is now 4.00, which is greater than $F_{2,387,0.05}$, so we reject the null.

That is, we now conclude that at least one of displacement and cylinders affects fuel efficiency.

Fuel efficiency is impacted by how "powerful" the engine is. While horsepower is a good indicater of engine power, so are displacement / cylinders if horsepower is absent from the data. Hence when we drop horsepower from our regression model, the effect of displacement / cylinders on mpg shows up. Another observation is that this seems to indicate that there is some correlation between displacement / cylinders and horsepower, which is certainly a very plausible hypothesis.