

**STAT:2100/4200, Lab 9**  
Inference in Regression

**You could treat the questions in this lab as practice exam questions.  
Solutions will be posted in the content section in ICON soon.**

## 1 Datasets

We will need the following dataset

OECD.dat

## 2 International health economics

Reference: <http://www.oecd.org/publications/figures/>

The OECD dataset is collated from the above web page of the Organization for Economic Cooperation and Development (OECD). It provides summary statistics for the 29 member nations. The variables are as follows:

name: name of country  
pcgdp: per capita gross domestic product (1998)  
reported in US dollars converted using Purchasing Power Parities to  
adjust for differences in price levels between countries  
pch: per capita health care expenditures (1996)  
reported in US dollars converted using Purchasing Power Parities  
beds: in-patient hospital beds per 1000 population (1996)  
los: average length of stay in days for hospital patients (1996)  
docs: doctors per 1000 population (1996)  
infmort: infant mortality (1996)  
number of deaths of infants < 1 yr of age per 1000 live births  
region: region of the world

Suppose we want to get predicted values of pch if we know pcgdp. We want predicted values for the countries that are in the dataset, as well as for a hypothetical new country with pcgdp = \$20,000.

We will put a dummy record (“country imaginary”) in the dataset with a missing value for pch and the desired value of the explanatory variable. This record will not be included in SAS’s calculation of the regression coefficients, but SAS will give us predicted values, as well as a confidence interval and prediction interval, for it. SAS’s symbol for a missing value is a period.

```
*****
* Reading in *
* the dataset *
***** ;
```

```
data OECD ;
input name $13. pcgdp pch beds los doc infmort ;
datalines;
Australia    22689    1775     8.7    15.5     2.5     0.6 16.4
...
UnitedStates 30514    3898     4.0    7.8     2.6     0.8   Amer
predict      20000 . . . . .
;
run ;
```

```
* Note the extra record for an imaginary country above, called "predict",
* with missing value for pch and other variables represented by one dot each.
```

```
proc print data = OECD ;
run ;
```

```
*****
* Regression *
***** ;
```

```
proc reg data = OECD ;
model pch = pcgdp / clb ; /* clb gets conf limits for
                          the parameter estimates */
run ;
```

```
*****
* Predicted values *
* and residuals *
***** ;
```

```
proc reg data = OECD ;
model pch = pcgdp / p ; /* p calculates predicted values */
id name ;
run ;
```

```
*****
* Confidence limits *
* for means of sub- *
* populations *
***** ;
```

```

proc reg data = OECD ;
model pch = pcgdp / clm ; /* clm gets conf limits for the mean */
id name ;
run ;

*****
* Prediction limits *
* for individual    *
* predictions       *
***** ;

proc reg data = OECD ;
model pch = pcgdp / cli ; /* cli gets prediction interval for
                           new individual */
id name ;
run ;

*****
* Scatterplots and *
* Residual plots   *
***** ;

proc reg data = OECD ;
model pch = pcgdp / p ;
plot pch * pcgdp / symbol = '.' ;
run ;
plot residual. * predicted. / symbol = '.' ;
run ;

```

Scatterplots and residual plots may also be obtained easily using the automated feature in SAS called “Insight”.

1. In the scatterplot, which variable, pch or pcgdp, is treated as the response variable? **pch**
2. Do you see any potentially influential points on either the scatter plot or the residual plot? Briefly explain. **One point (Luxenberg) has a pcgdp much higher than the rest of the points (it is an outlier on the x axis, not so much in the y axis.) Another point (USA) falls a bit far away from the linear pattern of all other points. Both points are influential in the sense that in case we did not observe either one or both of these points, the regression analysis would differ quite a bit. Anyway, they are not extreme outliers, we will go ahead and do regression for the entire dataset. (We can also repeat the regression analysis after removing one point or two and see if the results agree with when the outliers were included.)**

3. Check the assumptions needed for linear regression by examining the scatter plot. 1) The relationship between pcgdp and pch is close to being linear. 2) From this dataset, it is not easy to see if the subpopulations at each different  $x$  is normal since there are limited number of observations around each  $x$  values. But at least there is no sign of violation of the normality assumption 3) The variability in pch seems to be larger for larger pcgdp than for smaller pcgdps. So the equal variance assumption is not perfectly satisfied, yet not too seriously violated.
4. How well does the linear regression line fit the data?  $R^2 = .76$  suggests that the linear regression line based on the explanatory variable pcgdp accounts for 76% of the variability in pch, which is pretty high. Linear regression provided a pretty good fit to this data.
5. The null hypothesis is that there is no linear relationship between pch and pcgdp. Write this null hypothesis as a statement about a population parameter. Use conventional symbols.  $H_0 : \beta = 0$
6. We are interested in a linear relationship between pch and gdp. Write the alternative hypothesis as a statement about a population parameter.  $H_a : \beta \neq 0$
7. Give a point estimate and a 95% confidence interval for the parameter of interest in the hypotheses (numeric answers taken from the SAS output). Point estimate for  $\beta$  is  $b = 0.09682$ ; A 95% CI for  $\beta$  is  $(.07558, .11805)$ .
8. Based on your answer to the preceding question, would you reject the null hypothesis at significance level  $\alpha = .05$ ? (yes/no) Briefly explain. Although we got a confidence interval instead of obtaining a p-value above, we can still draw conclusions about the hypotheses: the 95% CI does not include 0, so the data suggests that  $\beta$  is significantly different from 0.
9. On the other hand, what are the numeric values of the test statistic and the p-value for the two-sided test of no linear relationship between pch and pcgdp (numeric answers from SAS output)? The test statistic is  $t = 9.36$  which corresponds to a p-value that is less than .0001.
10. Based on your answer to the preceding question, would you reject the null hypothesis at significance level  $\alpha = .05$ ? Yes, because p-value < .05.
11. What is the predicted pch in an individual country with pcgdp = 20000. Show your calculations or obtain it from SAS. SAS output reported 1471 as the predicted value for an imaginary country that has pcgdp=20000. If done by hand:  $\hat{y} = a + b * 20000 = -466.6 + 0.09682 * 20000 = 1469.8$ .
12. Give the interval in which you are 95% confident that the pch for **an individual country** with pcgdp=20000 would lie. Since the question is concerned with predicting the pch for **an individual country**, we look at the “95% CL Predict” columns in SAS, which gives a prediction interval: (658.4, 2256) .

13. Give the interval in which you are 95% confident that **the mean pch for all countries with pcgdp=20000** would lie. We look at the “95% CL Mean” columns in SAS, which gives a confidence interval: (1327,1614). (Note that this interval is a lot narrower than the interval from the previous question. Think of why.)