COMP615 – Foundations of Data Science

Lab 5 – Interpretation of results and Multiple Regressions

Introduction

This lab will cover interpretation of simple linear regression results from lab 04. In addition, multiple regression analysis will be applied to predict MPI urban using the selected predictors from lab 04. All the methods discussed in the lectures will form part of the lab. Use the sklearn documentation online to configure the methods. Above all, make sure you understand what you are doing; simply configuring the methods is not enough without an understanding of how they work. The basic code appears below.

Submit: To Do 1:3

1. Simple Linear Regression: Interpretation of the Results

To start with run a simple linear regression model using 'child_mort' from the 'combined' dataset (lab 04) as Independent Variable (also known as covariable or predictor) to predict the Dependent Variable (DV) 'mpi_urban'.

```
# create linear regression class object
reg = linear_model.LinearRegression()

# libraries for plotting of residual plots
import statsmodels.api as sm
from statsmodels.formula.api import ols
```

```
#fit simple linear regression model
model = ols('mpi_urban ~ child_mort', data=combined).fit()

#view model summary
print(model.summary())

#define figure size
fig = plt.figure(figsize=(12,8))

#produce regression plots
fig = sm.graphics.plot_regress_exog(model, 'child_mort', fig=fig)
```

(A)

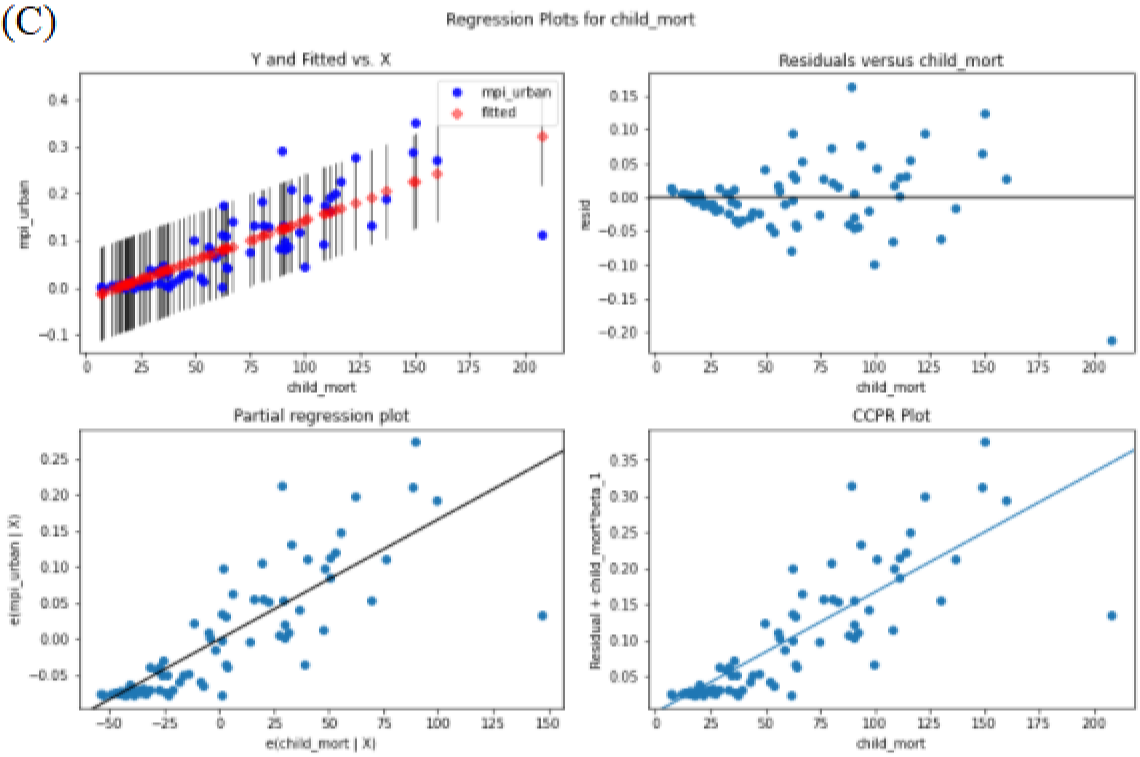
```
                            OLS Regression Results
==============================================================================
Dep. Variable:              mpi_urban   R-squared:                       0.680
Model:                            OLS   Adj. R-squared:                  0.676
Method:                 Least Squares   F-statistic:                     163.5
Date:                Mon, 28 Mar 2022   Prob (F-statistic):           9.91e-21
Time:                        09:39:01   Log-Likelihood:                 127.06
No. Observations:                  79   AIC:                            -250.1
Df Residuals:                      77   BIC:                            -245.4
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
```

(B)

```
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -0.0230      0.010     -2.388      0.019      -0.042      -0.004
child_mort     0.0017      0.000     12.786      0.000       0.001       0.002
==============================================================================
Omnibus:                       17.128   Durbin-Watson:                   2.167
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               78.736
Skew:                          -0.262   Prob(JB):                     7.99e-18
Kurtosis:                       7.863   Cond. No.                         130.
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

(C)



Looking at the generated regression results, Figure1(A), we can see that 68% (R^2) of the changes in 'mpi_urban' are explained by changes in our independent variable ('child_mort'). Adjusted R-squared are used for analysing multiple dependent variables' effectiveness on the model. In some cases, adding more independent variable might increase R^2 but lower the adjusted score. This can be an indication that some variables are not contributing to your model's R^2 properly. Prob (F-Statistic) indicates the accuracy of the null hypothesis (H_0: Effect of all independent variables in regression model is zero). In our model, given the low value of 9.91e-21<0.05 we reject the H_0 in favour of the alternative hypothesis that our model fits the data better than the intercept-only model.

P>|t| is one of the most important statistics in the summary. The ***p*-value** of 0.000 <0.05 rejects the H_0 (no statistically significant association between the variables child_mort' and 'mpi_urban'). Durbin-Watson tests for H_0 of no autocorrelation among the residuals. In our case, the value (2.167) is slightly higher than ideal (not equal to 2) therefore we reject the H_0. In addition, the 'Residuals vs Feature" plot in Figure 1 (C) also indicates heteroscedasticity (cone shape pattern) . When the residuals centre on zero, then we can say the residuals are random and they indicate that the model's predictions are correct on average.
The Partial regression plot and Component and Component Plus Residual (CCPR) plots of Figure 1 (C) will be discussed later when we add more variables to our model. 'Condition Number' measures the sensitivity of model's output as compared to the input. Multicollinearity is strongly indicated by a high condition number.

2. Multiple Independent Variables

Let's use all features of the 'combined' dataset to predict 'mpi_urban':

```
reg.fit(combined[['child_mort','exports','health','imports','income','
inflation','life_expec','total_fer','gdpp']],combined.mpi_urban)
```

Perform accuracy assessment by calculating the R-squared (R^2): R^2 indicates the proportion of variance in y (mpi_urban), explained by x (other features selected). I used this value to complete Table 1 (see page 5).

```
reg.score(combined[['child_mort','exports','health','imports','income',
'inflation','life_expec','total_fer','gdpp']],combined.mpi_urban)
```

Calculate Adjusted R-squared: The adjusted R-squared is a modified version of R^2 that adjusts for the number of predictors in a regression model. I used this value to complete Table 1 (see page 5).

```
1- (1-
reg.score(combined[['child_mort','exports','health','imports','income',
'inflation','life_expec','total_fer','gdpp']],combined.mpi_urban))*(len
(combined.mpi_urban)-1)/(len(combined.mpi_urban)-
combined[['child_mort','exports','health','imports','income','inflation
','life_expec','total_fer','gdpp']].shape[1]-1)
```

OR just simply fit a multiple regression model and print the summary.

```
Model1 = ols('mpi_urban ~
child_mort+exports+health+imports+income+inflation+life_expec+total_fer+gd
pp', data=combined).fit()
print(Model1.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:              mpi_urban   R-squared:                       0.835
Model:                            OLS   Adj. R-squared:                  0.813
Method:                 Least Squares   F-statistic:                     38.76
Date:                Mon, 28 Mar 2022   Prob (F-statistic):           1.45e-23
Time:                        22:31:26   Log-Likelihood:                 153.21
No. Observations:                  79   AIC:                            -286.4
Df Residuals:                      69   BIC:                            -262.7
Df Model:                           9
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -0.2707      0.082     -3.300      0.002      -0.434      -0.107
child_mort     0.0013      0.000      5.061      0.000       0.001       0.002
exports       -0.0002      0.000     -0.471      0.639      -0.001       0.001
health         0.0038      0.002      1.685      0.097      -0.001       0.008
imports    -7.742e-05      0.000     -0.216      0.830      -0.001       0.001
income     -1.822e-06    2.5e-06     -0.730      0.468    -6.8e-06    3.16e-06
inflation     -0.0007      0.000     -2.070      0.042      -0.001   -2.68e-05
life_expec     0.0026      0.001      2.598      0.011       0.001       0.005
total_fer      0.0273      0.005      5.600      0.000       0.018       0.037
gdpp         1.51e-06   3.29e-06      0.460      0.647   -5.04e-06    8.06e-06
==============================================================================
Omnibus:                       15.533   Durbin-Watson:                   2.027
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               23.219
Skew:                           0.796   Prob(JB):                     9.08e-06
Kurtosis:                       5.126   Cond. No.                     1.70e+05
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.7e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Figure 1: Result of Model1: fitting all available independent variables to predict ‘mpi_urban’.

To Do 1: Create a second model without features with multicollinearity and heteroscedasticity. Provide the code and complete Table1.

To Do 2: Create a third model with features with highest R-squared value found on linear regression. Provide the code and complete Table1.

Table 1: R^2 and Adjusted R^2 obtained for three different models.

	R^2	Adjusted R^2
Model 1	**0.835**	**0.813**
Model 2		
Model 3		

To Do 3: Use Table 1 to compare the R^2 and Adjusted R^2 obtained for each model and discuss your findings. Which model do you think is the best for predicting ‘mpi_urban’? Justify your answer.