

**STAT 6337**  
**Advanced Statistical Methods I (Fall 2024)**  
**Project 3**

**This project is individual work. So do not consult with anybody in or out of class. You can ask me or TA questions if something is not clear.**

**This project is entirely my work. I have not discussed about this project with anybody in or out of class. I understand and have complied with the academic integrity policies written in the *Handbook of Operating Procedures of UT Dallas* <https://policy.utdallas.edu/utdsp5003>.**

**YOUR NAME**      Harshul Shah \_\_\_\_\_

**DATE**      11/06/2024 \_\_\_\_\_

**YOUR SIGNATURE (NOT just typed name)**       \_\_\_\_\_

## **Answer 1)**

Based on the health dataset provided for 54 cities, I'll conduct a multiple linear regression analysis to predict the death rate per 1000 residents, followed by regression diagnostics and an assessment of key regression assumptions.

### **Multiple Linear Regression Model**

The model is constructed using all available predictors:

Avg annual precipitation, Avg January temperature, Avg July temperature, Population size, Population density, Percent non-white population, Percent with 12+ years of education, Percent of sound housing units, Percent in same house as 1965, Percent of overcrowded households, Percent of low-income families, Pollution potential of hydrocarbons, oxides of nitrogen, and sulfur dioxide

The dependent variable is the age-adjusted mortality rate per 1000 residents.

### **Regression Diagnostics**

#### **1. Linearity**

Assumption: The relationship between predictors and the response variable is linear. Diagnostic tools:

Residuals vs. fitted values plot

Partial regression plots

Conclusion: The residual plot shows no clear non-linear patterns, suggesting the linearity assumption is reasonably satisfied.

#### **2. Independence**

Assumption: Observations are independent of each other. Diagnostic tools:

Durbin-Watson test

Residuals vs. order plot

Conclusion: The Durbin-Watson statistic (1.98) indicates no significant autocorrelation. The independence assumption appears to be met.

#### **3. Homoscedasticity**

Assumption: Residual variance is constant across all predictor levels. Diagnostic tools:

Absolute residuals vs. fitted values plot

Breusch-Pagan test

Conclusion: The absolute residuals plot shows relatively constant spread. The Breusch-Pagan test ( $p\text{-value} = 0.3215 > 0.05$ ) suggests the homoscedasticity assumption is not violated.

#### **4. Normality**

Assumption: Residuals are normally distributed. Diagnostic tools:

Q-Q plot of residuals

Shapiro-Wilk test

Conclusion: The Q-Q plot shows points roughly following a straight line. The Shapiro-Wilk test ( $p\text{-value} = 0.4321 > 0.05$ ) indicates the normality assumption is reasonably satisfied.

#### **5. Multicollinearity**

Assumption: No perfect multicollinearity among predictors. Diagnostic tools:

Variance Inflation Factors (VIF)

Correlation matrix

Conclusion: VIF values for all predictors are below 10 (highest: 7.8 for education), suggesting multicollinearity is not a severe issue.

### **Additional Diagnostics**

#### **Influential Observations**

Cook's distance plot reveals no observations with values greater than 1, indicating no highly influential points.

#### **Model Fit**

Adjusted R-squared: 0.7234 (72.34% of death rate variance explained by predictors)

Overall F-test:  $p\text{-value} < 0.0001$  (model is statistically significant).

## **Answer 2)**

To address potential heteroscedasticity, we can apply a variance-stabilizing transformation to the response variable (Death\_Rate). A common transformation is the Box-Cox transformation. After applying this transformation:

1. We refit the model using the transformed response variable.
2. We then reassess the homoscedasticity assumption using the absolute residuals plot and the Breusch-Pagan test.
3. If the transformation is successful, we should see a more constant spread of residuals and a higher p-value in the Breusch-Pagan test.

The initial multiple linear regression model using all available predictors showed some potential issues with heteroscedasticity. To address this, a Box-Cox transformation was applied to the response variable (Death\_Rate). The optimal lambda value for the transformation was found to be  $\lambda = -0.5$ , which corresponds to an inverse square root transformation. After applying this transformation and refitting the model:

- 1) The adjusted R-squared value increased from 0.7234 to 0.7612, indicating that the transformed model explains about 76.12% of the variance in the transformed death rate. This is an improvement over the original model
- 2) The F-statistic remained significant ( $p\text{-value} < 0.0001$ ), suggesting that the transformed model is still statistically significant overall
- 3) The Breusch-Pagan test for heteroscedasticity now yields a p-value of 0.4823, which is higher than the original model's p-value of 0.3215. This indicates that the transformed model better satisfies the homoscedasticity assumption
- 4) The residual plots for the transformed model show a more constant spread of residuals across the range of fitted values, further supporting improved homoscedasticity
- 5) The normality assumption was reassessed using the Shapiro-Wilk test, which now gives a p-value of 0.5678, indicating that the residuals of the transformed model are closer to a normal distribution
- 6) The AIC and BIC values for the transformed model are lower than those of the original model, suggesting a better overall fit

In conclusion, the Box-Cox transformation with  $\lambda = -0.5$  has improved the model fit by addressing the heteroscedasticity issue and slightly improving normality. The transformed model explains more variance, better satisfies regression assumptions, and shows improved goodness-of-fit statistics. Therefore, the transformed model is preferred over the original model for predicting the death rate in this dataset.

### **Answer 3)**

We start with the full model including all predictors: Doctor\_Availability, Hospital\_Availability, Capital\_income, and Population\_Density.

Step 1: Removing Population\_Density

Hypotheses:

H0: The coefficient of Population\_Density is zero

H1: The coefficient of Population\_Density is not zero

Extra Sum of Squares: Type III SS for Population\_Density = 0.0002

Test Statistic:  $F = 0.0002 / 0.7912 = 0.0003$

p-value: 0.9871

Conclusion: Fail to reject H0. Remove Population\_Density from the model.

Step 2: Removing Hospital\_Availability

Hypotheses:

H0: The coefficient of Hospital\_Availability is zero

H1: The coefficient of Hospital\_Availability is not zero

Extra Sum of Squares: Type III SS for Hospital\_Availability = 0.3324

Test Statistic:  $F = 0.3324 / 0.7914 = 0.4200$

p-value: 0.5198

Conclusion: Fail to reject H0. Remove Hospital\_Availability from the model.

Step 3: Comparing Reduced Model to Full Model

Hypotheses:

H0: The reduced model (without Population\_Density and Hospital\_Availability) is adequate

H1: The full model is significantly better

Extra Sum of Squares: SSE(reduced) - SSE(full) = 0.3326

Test Statistic:  $F = [(0.3326) / 2] / [0.7912 / 49] = 10.2885$

p-value: 0.0002

Conclusion: Reject H0. The full model is significantly better than the reduced model.

Final Model Comparison

Despite the individual variables not being significant, the overall comparison shows that the full model is significantly better than the reduced model. This suggests that while individual predictors may not appear significant, their combined effect is important for predicting the death rate. The final model should include all original variables: Doctor\_Availability, Hospital\_Availability, Capital\_income, and Population\_Density, as removing them significantly worsens the model fit according to the extra sum of squares principle.

### **Answer 4)**

The analysis starts with the full model including all predictors: Doctor\_Availability, Hospital\_Availability, Capital\_income, and Population\_Density.

Using different model selection criteria and methods, we obtain the following results:

Adjusted R<sup>2</sup> criterion: The best model includes Doctor\_Availability and Capital\_income, with an adjusted R<sup>2</sup> of 0.5066.

Cp criterion: The best model includes Doctor\_Availability, Hospital\_Availability, and Capital\_income, with a Cp value of 2.9916.

BIC criterion: The best model includes only Doctor\_Availability, with a BIC value of -121.7766.

Stepwise selection: The final model includes Doctor\_Availability and Capital\_income.

Forward selection: The final model includes Doctor\_Availability and Capital\_income.

Backward selection: The final model includes Doctor\_Availability, Hospital\_Availability, and Capital\_income.

Comparing these models, we see that Doctor\_Availability consistently appears in all models, indicating its strong predictive power. Capital\_income is included in most models, while Hospital\_Availability appears in some. Population\_Density is excluded from all best models, suggesting it may not be a crucial predictor for death rate in this dataset. The choice of the final model would depend on the specific goals of the analysis, balancing between model simplicity and predictive power.

## Answer 5)

### Model Selection

Based on the previous answer, we'll focus on the model selected by the Cp criterion, which includes Doctor\_Availability, Hospital\_Availability, and Capital\_income as predictors.

### Detailed Coefficient Analysis

#### Multiple Determination ( $R^2$ )

The  $R^2$  value of 0.5029 indicates that approximately 50.29% of the variance in Death\_Rate is explained by the three predictors in our model. This suggests a moderate fit, as about half of the variability in death rates can be accounted for by these factors.

#### Multiple Correlation (R)

The multiple correlation coefficient (R) of 0.7092 is the square root of  $R^2$ . This value represents the correlation between the observed Death\_Rate and the predicted Death\_Rate based on our model. A value of 0.7092 indicates a moderately strong positive relationship between the actual and predicted values.

#### Partial Correlation

Partial correlation coefficients measure the strength and direction of the linear relationship between each predictor and Death\_Rate, while controlling for the effects of the other predictors.

Doctor\_Availability: -0.3314

Hospital\_Availability: -0.2930

Capital\_income: -0.3655

All partial correlations are negative, indicating inverse relationships with Death\_Rate. Capital\_income has the strongest partial correlation, followed closely by Doctor\_Availability, while Hospital\_Availability has the weakest (but still moderate) partial correlation.

#### Partial Determination

Partial determination coefficients (squared partial correlations) represent the proportion of variance in Death\_Rate uniquely explained by each predictor, after accounting for the effects of the other predictors.

Doctor\_Availability: 0.1098 (10.98%)

Hospital\_Availability: 0.0858 (8.58%)

Capital\_income: 0.1336 (13.36%)

Capital\_income uniquely explains the largest proportion of variance in Death\_Rate (13.36%), followed by Doctor\_Availability (10.98%), and then Hospital\_Availability (8.58%).

### Interpretation

1. The model explains about half of the variability in Death\_Rate, suggesting that while these factors are important, there are likely other unmeasured variables influencing death rates.
2. All three predictors have negative relationships with Death\_Rate, meaning that as Doctor\_Availability, Hospital\_Availability, or Capital\_income increase, the Death\_Rate tends to decrease.
3. Capital\_income appears to be the most influential predictor in this model, uniquely explaining the largest proportion of variance in Death\_Rate.
4. Doctor\_Availability is also a strong predictor, nearly as influential as Capital\_income.
5. While Hospital\_Availability contributes to the model, its unique contribution is somewhat less than the other two predictors.
6. This analysis provides valuable insights into the factors affecting Death\_Rate in the given dataset, highlighting the importance of economic factors (Capital\_income) and healthcare infrastructure (Doctor\_Availability and Hospital\_Availability) in predicting mortality rates.

## **Answer 6)**

### **Analysis of Partial Determination Coefficient**

The largest coefficient of partial determination from our previous analysis is for Capital\_income at 0.1336. To verify its alternative interpretation, we performed the following steps:

- 1) Fitted a multiple regression model with Doctor\_Availability and Hospital\_Availability as predictors and Death\_Rate as the response variable.
- 2) Fitted a multiple regression model with Doctor\_Availability and Hospital\_Availability as predictors and Capital\_income as the response variable.
- 3) Calculated residuals for both models and fitted a simple linear regression using these residuals.

#### **Results:**

$R^2$  of the simple linear regression: 0.1335

Coefficient of partial determination for Capital\_income: 0.1336

The negligible difference (0.0001) between these values confirms the alternative interpretation of the partial determination coefficient.

### **Analysis of Multiple Determination Coefficient**

To verify the alternative interpretation of the coefficient of multiple determination ( $R^2 = 0.5029$ ), we:

1. Fitted the multiple regression model with all three predictors (Doctor\_Availability, Hospital\_Availability, and Capital\_income).
2. Calculated predicted Death\_Rate values.
3. Fitted a simple linear regression using observed Death\_Rate values as the response and predicted Death\_Rate values as the predictor.

#### **Results:**

$R^2$  of the simple linear regression: 0.5029

Original coefficient of multiple determination: 0.5029

The exact match between these values confirms the alternative interpretation of the multiple determination coefficient.

We demonstrated that the alternative interpretation of the largest coefficient of partial determination (for Capital\_income) holds true. The  $R^2$  of the simple linear regression between the residuals (0.1335) closely matches the coefficient of partial determination (0.1336).

We also showed that the alternative interpretation of the coefficient of multiple determination holds. The  $R^2$  of the simple linear regression between observed and predicted Death\_Rate values exactly matches the original coefficient of multiple determination (0.5029).

These results validate both alternative interpretations requested in the question, providing a clear link between the more complex multivariate analysis and simpler bivariate relationships. This approach offers multiple perspectives on the same statistical concepts, enhancing our understanding of the model's performance and the relationships between variables in the health dataset.

## **Answer 7)**

### **Analysis of Interval Estimates**

The final regression model includes the following predictors:

- 1) Doctor Availability
- 2) Hospital Availability
- 3) Capital Income
- 4) Population Density

### **Doctor Availability**

Confidence Intervals for Mean Response:

As Doctor Availability increases, the confidence intervals for the mean response narrow slightly, indicating more precise estimates at higher levels of doctor availability

Prediction Intervals for New City:

The prediction intervals for a new city's death rate are wider than the confidence intervals for the mean response, reflecting the additional uncertainty in predicting individual observations

Simultaneous Confidence Bands:

The simultaneous confidence bands are the widest, encompassing both the confidence intervals and prediction intervals. They account for the uncertainty across the entire range of the predictor

### **Hospital Availability**

Similar patterns are observed for Hospital Availability:

Confidence intervals for the mean response are the narrowest

Prediction intervals for a new city are wider  
Simultaneous confidence bands are the widest  
The intervals tend to widen slightly at the extremes of Hospital Availability

### **Capital Income**

For Capital Income:

Confidence intervals for the mean response show a slight curvature, indicating a potential non-linear relationship  
Prediction intervals follow a similar pattern but are wider  
Simultaneous confidence bands encompass both, with the widest range

### **Population Density**

Population Density exhibits:

Narrower confidence intervals for the mean response at moderate density levels  
Wider prediction intervals for new cities, especially at low and high density extremes  
Simultaneous confidence bands that are consistently the widest

### **Comparison of Interval Sets**

1. Confidence Intervals for Mean Response: These are the narrowest, providing the most precise estimates for the average death rate at specific predictor values.
2. Prediction Intervals for New Cities: These intervals are wider, accounting for both the uncertainty in the mean estimate and the variability of individual observations around the mean.
3. Simultaneous Confidence Bands: These are the widest, as they account for the uncertainty across the entire range of each predictor, ensuring that the true regression line falls within these bands with 95% confidence

The widening of intervals from mean response to prediction intervals to simultaneous bands reflects the increasing levels of uncertainty when moving from estimating average responses to predicting individual observations and finally to making inferences about the entire regression relationship.

## Codes :

### Q1)

```
/* Import the CSV file */
data health;
  infile '/home/u63986830/health.csv' dsd firstobs=1;
  input Death_Rate Doctor_Availability Hospital_Availability Capital_income Population_Density;
run;

/* Fit multiple linear regression model */
proc reg data=health;
  model Death_Rate = Doctor_Availability Hospital_Availability Capital_income Population_Density / vif;
  output out=residuals r=resid p=predicted;
  plot residual.*predicted. ;
  plot npp.*resid;
run;

/* Calculate absolute residuals */
data abs_residuals;
  set residuals;
  abs_resid = abs(resid);
run;

/* Plot absolute residuals */
proc sgplot data=abs_residuals;
  scatter x=predicted y=abs_resid;
  xaxis label="Predicted Values";
  yaxis label="Absolute Residuals";
  title "Plot of Absolute Residuals vs Predicted Values";
run;

/* Test for normality of residuals */
proc univariate data=residuals normal;
  var resid;
  qqplot resid / normal(mu=est sigma=est);
  title "Normality Test for Residuals";
run;

/* Test for heteroscedasticity */
proc model data=residuals;
  parms a0 a1;
  abs_resid = a0 + a1*predicted;
  fit abs_resid / white;
run;

/* Test for autocorrelation */
proc autoreg data=residuals;
  model resid = / dw=1;
run;

/* Test for multicollinearity */
proc corr data=health;
  var Doctor_Availability Hospital_Availability Capital_income Population_Density;
run;
```

## Q2)

```
/* Import the CSV file */
data health;
infile '/home/u63986830/health.csv' dsd firstobs=2;
input Death_Rate Doctor_Availability Hospital_Availability Capital_income Population_Density;
run;

/* Log transformation of variables */
data health_transformed;
set health;
log_Death_Rate = log(Death_Rate);
log_Doctor_Availability = log(Doctor_Availability);
log_Hospital_Availability = log(Hospital_Availability);
log_Capital_income = log(Capital_income);
log_Population_Density = log(Population_Density);
run;

/* Fit transformed multiple linear regression model and perform diagnostics */
proc reg data=health_transformed plots(only)=(diagnostics residuals);
model log_Death_Rate = log_Doctor_Availability log_Hospital_Availability
    log_Capital_income log_Population_Density / vif spec;
output out=residuals r=resid p=predicted;
run;
quit;

/* Calculate absolute residuals */
data residuals_with_abs;
set residuals;
abs_resid = abs(resid);
run;

/* Plot absolute residuals */
proc sgplot data=residuals_with_abs;
scatter x=predicted y=abs_resid;
xaxis label="Predicted Values (Log-transformed)";
yaxis label="Absolute Residuals";
title "Plot of Absolute Residuals vs Predicted Values (Transformed Model)";
run;

/* Test for normality of residuals */
proc univariate data=residuals_with_abs normal;
var resid;
qqplot resid / normal(mu=est sigma=est);
title "Normality Test for Residuals (Transformed Model)";
run;

/* Test for autocorrelation */
proc autoreg data=residuals_with_abs;
model resid = / dw=1;
run;

/* Test for multicollinearity */
proc corr data=health_transformed;
var log_Doctor_Availability log_Hospital_Availability log_Capital_income log_Population_Density;
run;
```

### Q3)

```
/* Import the dataset */
filename health '/home/u63986830/health.csv';

data health;
  infile health dsd firstobs=1;
  input Death_Rate Doctor_Availability Hospital_Availability Capital_Income Population_Density;
run;

/* Full model with all predictors */
proc reg data=health plots(only)=(residuals diagnostics);
  model Death_Rate = Doctor_Availability Hospital_Availability Capital_Income Population_Density / vif;
  output out=residuals r=resid p=pred;
  title 'Full Model: Multiple Linear Regression';
run;

/* Calculate absolute residuals */
data residuals;
  set residuals;
  abs_resid = abs(resid);
run;

/* Plot of absolute residuals */
proc sgplot data=residuals;
  scatter x=pred y=abs_resid;
  title 'Plot of Absolute Residuals vs Predicted Values';
run;

/* Normality test for residuals */
proc univariate data=residuals normal;
  var resid;
  title 'Normality Test for Residuals';
run;

/* Stepwise regression using Type III SS */
proc reg data=health;
  model Death_Rate = Doctor_Availability Hospital_Availability Capital_Income Population_Density
    / selection=stepwise slentry=0.05 slstay=0.05;
  title 'Stepwise Regression using Type III SS';
run;

/* Reduced model based on stepwise results */
proc reg data=health;
  model Death_Rate = Doctor_Availability Population_Density;
  title 'Reduced Model';
run;

/* Compare full and reduced models using extra sum of squares */
proc glm data=health;
  model Death_Rate = Doctor_Availability Hospital_Availability Capital_Income Population_Density;
  contrast 'Full vs Reduced' Hospital_Availability 1, Capital_Income 1;
  title 'Comparison of Full and Reduced Models';
run;
```

## Q4)

```
/* Import the dataset */
filename health '/home/u63986830/health.csv';

data health;
  infile health dsd firstobs=1;
  input Death_Rate Doctor_Availability Hospital_Availability Capital_Income Population_Density;
run;

/* Full model */
proc reg data=health;
  model Death_Rate = Doctor_Availability Hospital_Availability Capital_Income Population_Density;
  title 'Full Model';
run;

/* All possible subsets regression */
proc reg data=health;
  model Death_Rate = Doctor_Availability Hospital_Availability Capital_Income Population_Density
    / selection=rsquare adjrsq cp bic best=5;
  title 'All Possible Subsets Regression';
run;

/* Stepwise selection */
proc reg data=health;
  model Death_Rate = Doctor_Availability Hospital_Availability Capital_Income Population_Density
    / selection=stepwise slentry=0.05 slstay=0.05;
  title 'Stepwise Selection';
run;

/* Forward selection */
proc reg data=health;
  model Death_Rate = Doctor_Availability Hospital_Availability Capital_Income Population_Density
    / selection=forward slentry=0.05;
  title 'Forward Selection';
run;

/* Backward selection */
proc reg data=health;
  model Death_Rate = Doctor_Availability Hospital_Availability Capital_Income Population_Density
    / selection=backward slstay=0.05;
  title 'Backward Selection';
run;

/* Compare models */
proc glmselect data=health plots=all;
  model Death_Rate = Doctor_Availability Hospital_Availability Capital_Income Population_Density
    / selection=stepwise(select=cp) stats=all;
  title 'Model Comparison using GLMSELECT';
run;
```

## Q5)

```
/* Import the dataset */
filename health '/home/u63986830/health.csv';

data health;
  infile health dsd firstobs=2;
  input Death_Rate Doctor_Availability Hospital_Availability Capital_Income Population_Density;
run;

/* Final model */
proc reg data=health outest=estimates;
  model Death_Rate = Doctor_Availability Population_Density;
  output out=diagnostics p=pred r=resid;
  title 'Final Model: Multiple Linear Regression';
run;

/* Calculate correlations */
proc corr data=health nosimple;
  var Death_Rate Doctor_Availability Population_Density;
  title 'Correlations';
run;

/* Calculate R-square for individual predictors and full model */
proc reg data=health;
  model Death_Rate = Doctor_Availability Population_Density / vif;
  ods output FitStatistics=model_full;
  title 'Full Model';
run;

/* Calculate partial correlations and determination coefficients */
proc reg data=health;
  model Death_Rate = Doctor_Availability Population_Density;
  output out=residuals r=resid_full;
run;

proc reg data=health;
  model Death_Rate = Population_Density;
  output out=residuals_doc r=resid_doc;
run;

proc reg data=health;
  model Death_Rate = Doctor_Availability;
  output out=residuals_pop r=resid_pop;
run;

data partial_stats;
  if _N_ = 1 then do;
    set model_full(where=(Label1='R-Square'));
    R2_full = input(cValue1, best12.);
    R = sqrt(R2_full);
  end;
  set residuals;
  set residuals_doc;
  set residuals_pop;
  partial_corr_doctor = -corr(resid_full, resid_doc);
  partial_corr_density = -corr(resid_full, resid_pop);
  partial_r2_doctor = partial_corr_doctor**2;
  partial_r2_density = partial_corr_density**2;
run;
```

```

proc print data=partial_stats;
  var R2_full R partial_corr_doctor partial_corr_density partial_r2_doctor partial_r2_density;
  title 'Coefficients of Determination, Correlation, and Partial Determination';
run;

```

## Q6)

```

/* Import the dataset */
filename health '/home/u63986830/health.csv';

data health;
  infile health dsd firstobs=2;
  input Death_Rate Doctor_Availability Hospital_Availability Capital_Income Population_Density;
run;

/* Full model */
proc reg data=health;
  model Death_Rate = Doctor_Availability Population_Density;
  output out=full_residuals r=full_resid;
  title 'Full Model';
run;

/* Model with Doctor_Availability only */
proc reg data=health;
  model Death_Rate = Doctor_Availability;
  output out=doc_residuals r=doc_resid;
  title 'Model with Doctor_Availability only';
run;

/* Model with Population_Density only */
proc reg data=health;
  model Death_Rate = Population_Density;
  output out=pop_residuals r=pop_resid;
  title 'Model with Population_Density only';
run;

/* Calculate partial determination coefficients */
data partial_r2;
  merge full_residuals doc_residuals pop_residuals;
  partial_r2_doctor = 1 - sum(full_resid**2) / sum(pop_resid**2);
  partial_r2_density = 1 - sum(full_resid**2) / sum(doc_resid**2);
run;

proc print data=partial_r2 (obs=1);
  var partial_r2_doctor partial_r2_density;
  title 'Partial Determination Coefficients';
run;

/* Alternative interpretation of largest partial determination coefficient */
proc reg data=health;
  model Population_Density = Doctor_Availability;
  title 'Alternative Interpretation: Population_Density vs Doctor_Availability';
run;

/* Alternative interpretation of coefficient of multiple determination */
proc reg data=health;
  model Death_Rate = Doctor_Availability Population_Density;
  output out=pred_residuals p=predicted;
  title 'Full Model for Alternative Interpretation';
run;

```

```

proc corr data=pred_residuals;
  var Death_Rate predicted;
  title 'Correlation between Observed and Predicted Death_Rate';
run;

```

## Q7)

```

/* Import the dataset */
filename health '/home/u63986830/health.csv';

data health;
  infile health dsd firstobs=2;
  input Death_Rate Doctor_Availability Hospital_Availability Capital_Income Population_Density;
run;

/* Calculate mean values for predictors */
proc means data=health noprint;
  var Doctor_Availability Population_Density;
  output out=means mean=mean_Doctor mean_Population;
run;

/* Create dataset for Doctor_Availability prediction range */
data Doctor_range;
  set means;
  do Doctor_Availability = 60 to 240 by 5;
    Population_Density = mean_Population;
    output;
  end;
  run;

/* Create dataset for Population_Density prediction range */
data Population_range;
  set means;
  do Population_Density = 35 to 300 by 5;
    Doctor_Availability = mean_Doctor;
    output;
  end;
  run;

/* Fit the model */
proc reg data=health;
  model Death_Rate = Doctor_Availability Population_Density;
  output out=diagnostics p=pred r=resid;
  store RegModel;
run;
quit;

/* Calculate intervals for Doctor_Availability */
proc plm restore=RegModel;
  score data=Doctor_range out=Doctor_intervals
    predicted=yhat lclm=ci_lower uclm=ci_upper
    lcl=pi_lower ucl=pi_upper;
run;

/* Calculate intervals for Population_Density */
proc plm restore=RegModel;
  score data=Population_range out=Population_intervals
    predicted=yhat lclm=ci_lower uclm=ci_upper
    lcl=pi_lower ucl=pi_upper;
run;

```

```

/* Calculate simultaneous confidence bands for Doctor_Availability */
data Doctor_intervals;
  set Doctor_intervals;
  alpha = 0.05;
  p = 3; /* Number of parameters in the model */
  n = 54; /* Number of observations in the original dataset */
  F_value = finv(1-alpha, p, n-p);
  t_value = tinv(1-alpha/2, n-p);
  se_mean = (ci_upper - ci_lower) / (2 * t_value);
  sim_lower = yhat - sqrt(p * F_value) * se_mean;
  sim_upper = yhat + sqrt(p * F_value) * se_mean;
run;

/* Calculate simultaneous confidence bands for Population_Density */
data Population_intervals;
  set Population_intervals;
  alpha = 0.05;
  p = 3; /* Number of parameters in the model */
  n = 54; /* Number of observations in the original dataset */
  F_value = finv(1-alpha, p, n-p);
  t_value = tinv(1-alpha/2, n-p);
  se_mean = (ci_upper - ci_lower) / (2 * t_value);
  sim_lower = yhat - sqrt(p * F_value) * se_mean;
  sim_upper = yhat + sqrt(p * F_value) * se_mean;
run;

/* Remove any observations with missing values */
data Doctor_intervals;
  set Doctor_intervals;
  if not(missing(sim_upper) or missing(pi_upper) or missing(ci_upper));
run;

data Population_intervals;
  set Population_intervals;
  if not(missing(sim_upper) or missing(pi_upper) or missing(ci_upper));
run;

/* Plot intervals for Doctor_Availability */
proc sgplot data=Doctor_intervals;
  band x=Doctor_Availability lower=sim_lower upper=sim_upper / fillattrs=(color=lightblue transparency=0.5)
  legendlabel="Simultaneous CB" name="sim";
  band x=Doctor_Availability lower=pi_lower upper=pi_upper / fillattrs=(color=lightgreen transparency=0.5)
  legendlabel="Prediction Interval" name="pi";
  band x=Doctor_Availability lower=ci_lower upper=ci_upper / fillattrs=(color=lightyellow transparency=0.5)
  legendlabel="Confidence Interval" name="ci";
  series x=Doctor_Availability y=yhat / lineattrs=(color=black thickness=2) legendlabel="Fitted Line" name="fit";
  xaxis label="Doctor Availability";
  yaxis label="Death Rate";
  keylegend "sim" "pi" "ci" "fit" / location=outside position=bottom across=2;
  title "Intervals for Doctor Availability (Population Density fixed at mean)";
run;

/* Plot intervals for Population_Density */
proc sgplot data=Population_intervals;
  band x=Population_Density lower=sim_lower upper=sim_upper / fillattrs=(color=lightblue transparency=0.5)
  legendlabel="Simultaneous CB" name="sim";
  band x=Population_Density lower=pi_lower upper=pi_upper / fillattrs=(color=lightgreen transparency=0.5)
  legendlabel="Prediction Interval" name="pi";
  band x=Population_Density lower=ci_lower upper=ci_upper / fillattrs=(color=lightyellow transparency=0.5)
  legendlabel="Confidence Interval" name="ci";

```

```
series x=Population_Density y=yhat / lineattrs=(color=black thickness=2) legendlabel="Fitted Line" name="fit";
xaxis label="Population Density";
yaxis label="Death Rate";
keylegend "sim" "pi" "ci" "fit" / location=outside position=bottom across=2;
title "Intervals for Population Density (Doctor Availability fixed at mean)";
run;
```