Priyasri Sankaran
Project 3
Consider the health dataset, which consists of data from 54 cities. We want to build a multiple linear regression for predicting death rate per 1000 residents.
1)Fit a multiple linear regression model using all available predictors. Carry out regression diagnostics (including plot of absolute residuals). The analysis should include an assessment of the degree to which the key regression assumptions are satis ed. Clearly state each assumption, the diagnostic tools used to check it, and the conclusion. Use all tests and plots that were discussed in class.

Fit a multiple linear regression model using all available predictors. Health dataset has 53 observations and 4 predictors:-Doctor_Availability, Hospital_Availability,Capital_income and Population_Density.
The relationship between the Death_Rate according to the fitted regession model.
Death_rate = 12.26626+(0.00739* Doctor_Availability)+(0.00058372*Hospital_Availability)-(0.33023*Capital_income)-(0.00946*Population_Density)
Doctor_Availability: β_1=0.00739, P value =0.2917
Hospital_Availability: β_2=0.00058, P value =0.4228
Capital_income: β_3=−0.33023, P value=0.1656
Population_Density: β_4=−0.00946, P value =0.0587 *
None of the P values are significant at 0.05 level except Population Density very close to the 0.05 level.
Model Performance: R^2 = 0.1437 and adjusted R^2=0.0723 gives us the variation in Death rates. The F value=2.01 and the P value=0.1075, the P value is not significant at 0.05. Model is not statistically significant at 0.05 level.
Assumptions and Diagnostic: Linearity: We assume that the relationship between each predictor and the response variable is linear. The scatter plot shows non linear trend.
Shapiro-Wilk test W=0.9088 with a P Value-0.007< 0.05, shows non-normality.
Skewness= -1.2351339 and kurtosis= 2.7387. The negative skew values shows that the data is not normally distributed and sked to the left. We can observe that on the Q-Q plot.
The residuals deviate from the normality, evidence that our assumption of linearity is not satisfied. No significance if independent of each other variable.The model may not explain the variability on the variable death rate. If we take a look at the scatter plot predicted values and Death rate we can see non linearity. Q-Q Plot for residual, normality test shows that some data points are not on the line, indicating that the is no pattern.
Lack of fit test: We performed lack of fit test to check linearity on the full model. No significant value is obtained from the result. That's because there were no replication in the data.

Breucsh Pagan test P Value=0.4255 > 0.05, suggests that the assumption is not met. This indicates that no significant heteroscedasticity.
The variance of residuals is constant across different levels of predictor variables. Here is no significant evidence of heteroscedasticity in the model, the variance of the residuals appears to be constant, the assumption of homoscedasticity is satisfied.
The Brown-Forsythe test results were not significant. The p Value=0.284<0.05. this indicates the variance homogeneity.

2) If an assumption is not met, attempt to remedy the situation. Explain the steps used to obtain the transformed model. Comment on the fit the transformed model using appropriate tests and statistics.
Note: For the remaining parts, continue in transformed scale, if a transformation was applied earlier.

Since the assumption is not met, remedy of the situation is to perform transformation on the model. By looking at the Q-Q plots some data points are off from the line. Looks like left skewed, also the skewed value=-1.2351339 is a negative skewed data. Also the Death_Rate indicates non-normality in residuals.
So performed a square transformation on the model. Observed the Q-Q Plot after the transformation, now the data points are on the fitted line.

We did lack of test on transformed data, no significant value is obtained from the result. That's because there were no replication in the data.
The scatter plot after transformation looks same as the full model. The R^2=0.1786. Adjust R^2=0.1102. Indicates the variance is around 17.86% in square_Death_Rate is explained by the model.
F-Statistics=2.61, very low with the P-Value=0.0470 is now statistically significant at 0.05 level, different from original model.
Normality of the residuals:- Shapiro-Wilks test W=0.99215, P-Value=0.9790.
Breucsh pagan test result indicates no significant heteroscedasticity >0.05. Capital_income shows Marginal Pvlaue=0.0204.
The square transformation improved the model, over all variance remains moderate.

3) Use the principle of extra sum of squares (type I and III SS) to determine which variables can be removed from the model (try removing one variable at a time | the least significant one). Once a tentative model is obtained, compare it with the initial model (with all variables but in transformed scale obtained earlier) using appropriate extra sum of squares. Clearly state at each step the hypotheses being tested, the appropriate extra sum of squares, test statistic, p-value, and conclusion.

Model	Sum of Square	Error Sum Squares	DF(model)	DF (Error)
Full-transformed Model	7857.09	36135	4	48
Reduced Model(Hospital_Availability)	7555.067	36437.186	3	49
Reduced Model(Population_Density Doctor_Avalilability)	5675.34	38316	2	50

Null Hypothesis(H_0): coefficient of Hospital_Availability and is not a significant variable
Alternative Hypothesis(H_a):coefficient of Hospital _Availability is a significant variable.
The transformed full model includes all predictors. R^2=0.1786, F= 2.61, P value=0.047. Please see the below results for both Final model and the transformed full model. Based on the high P value We are dropping Doctor_Availability and Hospital_Availability.
We have a final model with Capital_income and Population_Density.
Doctor_Availability: β_1=0.15131, P value =0.2089
Hospital_Availability: β_2=0.00784, P value =0.5295
Capital_income: β_3=−07.45199, P value=0.0699
Population_Density: β_4=−17533, P value =0.0416 **
Model without Doctor_Availability and Hospital_Availability

H_0: The coefficient of Doctor_Availability is 0.

H_a: The coefficient of Doctor_Availability is not 0.

R2=0.1290R2=0.1290, F=3.70, p=0.0316. Type I and Type III SS for remaining predictors shows:
Capital_income**:** p = 0.1643 and Population_Density**:** p = 0.0385 (significant).

The Extra sum of the square please see the below image and the values:

Type 1 SS for capital_income=3880.96, F=5.16, P value=0.0277<0.05, reject H_0

Type III SS =2588.165 F value=3.44, P value=0.0699>0.05

Type I SS For the Population_Density=3300.651845, F-value=4.38, p value=0.0416<0.05, reject H_0

Type III SS=3300.65, F value=4.38, P value=0.0416<0.05. reject H_0

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Doctor_Availability	1	531.042251	531.042251	0.71	0.4051
Hospital_Availabilit	1	144.412739	144.412739	0.19	0.6634
Capital_income	1	3880.986458	3880.986458	5.16	0.0277
Population_Density	1	3300.651845	3300.651845	4.38	0.0416

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Doctor_Availability	1	1221.117277	1221.117277	1.62	0.2089
Hospital_Availabilit	1	302.026915	302.026915	0.40	0.5295
Capital_income	1	2588.165960	2588.165960	3.44	0.0699
Population_Density	1	3300.651845	3300.651845	4.38	0.0416

Therefore we conclude that we will keep Capital income and the Population_Density in the model.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	7857.09329	1964.27332	2.61	0.0470
Error	48	36135	752.81581		
Corrected Total	52	43992			

Root MSE	27.43749	R-Square	0.1786
Dependent Mean	89.30717	Adj R-Sq	0.1102
Coeff Var	30.72260		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	156.83437	34.61504	4.53	<.0001
Doctor_Availability	1	0.15131	0.11881	1.27	0.2089
Hospital_Availability	1	0.00784	0.01237	0.63	0.5295
Capital_income	1	-7.45199	4.01902	-1.85	0.0699
Population_Density	1	-0.17533	0.08374	-2.09	0.0416

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Capital_income	1	2212.952724	2212.952724	2.89	0.0955
Population_Density	1	3462.392088	3462.392088	4.52	0.0385

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Capital_income	1	1526.855179	1526.855179	1.99	0.1643
Population_Density	1	3462.392088	3462.392088	4.52	0.0385

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	156.5461757	34.22137620	4.57	<.0001
Capital_income	-5.0808897	3.59957368	-1.41	0.1643
Population_Density	-0.1744056	0.08205074	-2.13	0.0385

Also from the partial F test we can see that Pvalue=0.1921>0.05, we do not reject null hypothesis indicating that the Doctor_Availability and Hospital_Availability and capital income do not improve the model.

Partial F test removing Doctor_Availability Hospital_Availability Capital_income

The REG Procedure
Model: MODEL1

Test x1x2x3 Results for Dependent Variable square_Death_Rate

Source	DF	Mean Square	F Value	Pr > F
Numerator	3	1236.20122	1.64	0.1921
Denominator	48	752.81581		

4) Starting with the full model, find the best model(s) using adjusted R_2, C_p, and BIC criterion. Also, nd models using stepwise, forward, and backward selection methods. Compare all these models.

Adjusted R-Square Selection Method

Number of Observations Read	53
Number of Observations Used	53

Number in Model	Adjusted R-Square	R-Square	C(p)	BIC	Variables in Model
3	0.1210	0.1717	3.4012	356.9939	Doctor_Availability Capital_income Population_Density
4	0.1102	0.1786	5.0000	358.8306	Doctor_Availability Hospital_Availability Capital_income Population_Density
3	0.0989	0.1508	4.6221	358.1074	Hospital_Availability Capital_income Population_Density
2	0.0942	0.1290	3.8981	357.1622	Capital_income Population_Density
2	0.0806	0.1159	4.6617	357.8611	Hospital_Availability Population_Density
1	0.0765	0.0943	3.9263	356.9943	Population_Density
2	0.0693	0.1051	5.2971	358.4352	Doctor_Availability Population_Density
2	0.0672	0.1031	5.4132	358.5394	Doctor_Availability Capital_income
3	0.0659	0.1198	6.4380	359.7181	Doctor_Availability Hospital_Availability Population_Density
3	0.0487	0.1036	7.3844	360.5370	Doctor_Availability Hospital_Availability Capital_income
1	0.0317	0.0503	6.4974	359.3188	Capital_income
2	0.0214	0.0590	7.9874	360.7941	Hospital_Availability Capital_income
1	-.0073	0.0121	8.7315	361.2533	Doctor_Availability
1	-.0119	0.0076	8.9924	361.4744	Hospital_Availability
2	-.0240	0.0154	10.5397	362.9324	Doctor_Availability Hospital_Availability

Adjusted R square: By looking at the model $R_{adj}^2 = 0.1210$ is high for the model with these variables Doctor_Availability,Capital_income, Population_Density.

C_p: Lowest C_p= 3.4012~3value for the model with Doctor_Availability ,Capital_income, Population_Density.

BIC: Lowest BIC value=356.9943, model with variable Doctor_Availability ,Capital_income, Population_Density.

Comparison table:

Table :

Selection	Variable	R^2	C_p	F Value	P Value
Stepwise	Population _Density	0.0943	3.9263	5.31	0.0253
Forward	Population _Density	0.0943	3.9263	124.85	0.0253

Table :

Backward- Selection	R^2	C_p	P Value
removed (Hospital_avalilability	0.1717	3.4012	0.0494
Removed- Doctor_Availability	0.1290	3.8981	0.0385
Removed- Capital_income	0.0943	3.9263	0.0253

Analysis:

R^2 (highest)and C_p(smallest) both has best values with these variables Doctor_Availability,Capital_income, Population_Density. Please refer above.

Stepwise selection small p value with the variable as Population_Density.

Forward Selection gives a Population_Density with P value=0.025<0.05

BIC Criterion has the lowest value with the Doctor_Availability ,Capital_income, Population_Density variables.

If we observe Doctor_Availability ,Capital_income, Population_Density variables selected as a significant variable on various evaluation. These various selection methods giving us the best model to fit our given health data.

No other variable met the 0.0500 significance level for entry into the model.

Summary of Forward Selection							
Step	Variable Entered	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	Population_Density	1	0.0943	0.0943	3.9263	5.31	0.0253

Bounds on condition number: 1, 1

All variables left in the model are significant at the 0.0500 level.

Summary of Backward Elimination							
Step	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	Hospital_Availability	3	0.0069	0.1717	3.4012	0.40	0.5295
2	Doctor_Availability	2	0.0427	0.1290	3.8981	2.53	0.1183
3	Capital_income	1	0.0347	0.0943	3.9263	1.99	0.1643

5) For one of the “final" models obtained from the previous part, report coefficients of multiple determination, multiple correlation, partial correlation, and partial determination. Multiple determination (R^2)=0.1717. Proportion of variance is explained in transformed model by the variable population_Density. This is not a very high R^2 value. There might be other factors that are contributing to the variance.

Multiple Correlation R=√0.1717 ≈0.414. This explains the linear relationship between the transformed model and the variable Population_Density. The linear relationship is not very high. Ther is a moderate relationship is there.
Partial correlation(Population_Density)= - 0.27667.
Partial correlation (Doctor_Availability)=0.2214
Partial correlation(Capital_income=-0.2729

Coefficient of Partial Determination(Population_Density)=0.0765
Coefficient of Partial correlation (Doctor_Availability)=0.049
Coefficient of Partial correlation(Capital_income)=0.07447

This explains the linear relationship between the transformed model and the variable Population_ Density, Doctor_Availability and Capital_income. There is a negative correlation. The variance is not too high. Overall we see medium impact on the transformed model.

6) Consider the largest coefficient of partial determination. Show that its alternative interpretation in terms of coefficient of simple determination holds by fitting appropriate models and calculating R_2. Similarly, show that the alternative interpretation of the coefficient of multiple determination holds.

$R^2_{partial}$ =0.1936, R^2_{full}= 0.1786, $R^2_{reduced}$= 0.1210. The result shows that the relationship between the coefficient of partial determination and the coefficient of simple determination, confirming the interpretation holds.

Number of Observations Used	53

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	14448	7224.03725	6.00	0.0046
Error	50	60192	1203.84908		
Corrected Total	52	74641			

Root MSE	34.69653	R-Square	0.1936
Dependent Mean	116.09434	Adj R-Sq	0.1613
Coeff Var	29.88649		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-24.39094	42.89170	-0.57	0.5721
Capital_income	1	15.61347	4.51156	3.46	0.0011
Population_Density	1	-0.06183	0.10284	-0.60	0.5504

7) Use the final model to obtain 95% interval estimates for the mean response for the entire range of predictor and plot them against each predictor in the final model separately (by fixing other predictors' value to their averages). In the same plot, add two corresponding sets of intervals | (i) 95% interval estimates for death rate of a new city (not in the dataset) (ii) 95% simultaneous confidence bands for the entire regression line. Compare the three sets of intervals.

Please refer the below image of the confidence intervals. The confidence interval for mean response is narrow because it only counts the mean estimate’s uncertainty. The Predictions interval is the widest, because it accounts for the individual variability. The simulation confidence bands are slightly wider than the confidence interval for the mean. The confidence and the prediction intervals for the predicted values are below. We see that the Predicted values fall in the interval.

Confidence and Prediction Intervals for Predicted Values

Obs	Population_Density	Predicted	Lower_CI	Upper_CI	Lower_PI	Upper_PI
1	109	85.477	74.6847	96.269	29.6244	141.329
2	144	81.124	67.6942	94.553	24.7025	137.545
3	113	98.305	80.4648	116.145	40.6745	155.936
4	97	92.187	83.1400	101.235	36.6458	147.729
5	206	75.184	54.9027	95.466	16.7519	133.617
6	124	74.744	60.5630	88.925	18.1392	131.349
7	152	71.214	56.6000	85.829	14.4994	127.929
8	162	92.717	74.8654	110.568	35.0827	150.351
9	150	82.617	70.2092	95.025	26.4301	138.804
10	134	95.094	79.1120	111.075	38.0110	152.176