Multiple Regression Analysis
Assignment 04

CST 316 -2
Statistical Methods - II

CST/20/094
F.F.Thaufeek

Computer Science and Technology
Department of Computer Science and Informatics
Faculty of Applied Sciences
Uva Wellassa University of Sri Lanka

1. This example examines the relationship between annual family Food Expenditure and family Income. The maintained hypothesis is that family food expenditure increases as family income increases and conversely, ceteris paribus (including family size).
   The following set of data is obtained from 20 families in a certain area in 2023.

Family	Annual Food Expenditure (Rs. 000')	Annual Income (Rs. 000')	Family Size (number in family)
1	5.2	28	3
2	5.1	26	3
3	5.6	32	2
4	4.6	24	1
5	11.3	54	4
6	8.1	59	2
7	7.8	44	3
8	5.8	30	2
9	5.1	40	1
10	18.0	82	6
11	4.9	42	3
12	11.8	58	4
13	5.2	28	1
14	4.8	20	5
15	7.9	42	3
16	6.4	47	1
17	20.0	112	6
18	13.7	85	5
19	5.1	31	2
20	29	26	2

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	2	219.992	109.996	3.25	0.064
Annual Income	1	64.760	64.760	1.91	0.185
Family Size (number in family)	1	14.067	14.067	0.42	0.528
Error	17	576.110	33.889		
Lack-of-Fit	16	571.610	35.726	7.94	0.273
Pure Error	1	4.500	4.500		
Total	19	796.102			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
5.82141	27.63%	19.12%	8.34%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	2.37	3.02	0.78	0.444	
Annual Income	0.1045	0.0756	1.38	0.185	1.84
Family Size (number in family)	0.73	1.13	0.64	0.528	1.84

Regression Equation

Annual Food Expenditure = 2.37 + 0.1045 Annual Income + 0.73 Family Size (number in family)

Fits and Diagnostics for Unusual Observations

Obs	Annual Food Expenditure	Fit	Resid	Std Resid	
14	4.80	8.10	-3.30	-0.80	X
17	20.00	18.44	1.56	0.36	X
20	29.00	6.54	22.46	4.03	R

R Large residual
X Unusual X

Analysis of Variance (ANOVA) Table

The ANOVA table helps to summarize the sources of variability in the data. It separates the total variability into variability due to the regression model and variability due to error. The table shows that the regression model has an F-value of 3.25 with a p-value of 0.064, indicating that the model is not statistically significant at the 5% level. This means that we do not have enough evidence to say that the regression model explains a significant portion of the variability in Annual Food Expenditure. The p-values for the individual predictors, Annual Income (0.185) and Family Size (0.528), are both greater than 0.05, indicating that neither predictor is significant at the 5% level.

Model Summary

The model summary provides several key statistics about the fit of the regression model. The standard error of the estimate (S) is 5.82141, indicating the average distance that the observed values fall from the regression line. The R^2 value of 27.63% suggests that approximately 27.63% of the variability in Annual Food Expenditure can be explained by the model. However, the adjusted R^2 is lower at 19.12%, which adjusts for the number of predictors in the model. The predicted R^2 value of 8.34% is even lower, indicating that the model may not perform well in predicting new data.

Coefficients Table

The coefficients table provides the estimated regression coefficients, their standard errors, t-values, p-values, and variance inflation factors (VIF). The constant term has a coefficient of 2.37 with a p-value of 0.444, indicating it is not significant. The coefficient for Annual Income is 0.1045 with a p-value of 0.185, and the coefficient for Family Size is 0.73 with a p-value of 0.528. Both predictors have VIF values of 1.84, indicating low multicollinearity. However, neither predictor is statistically significant at the 5% level.

Fits and Diagnostics for Unusual Observations

The fits and diagnostics table highlights unusual observations in the data. Observation 20 has a large residual of 22.46 with a standardized residual of 4.03, indicating it is an outlier. Observations 14 and 17 are marked with "X" for unusual X values, suggesting they have unusual predictor values. These unusual observations can significantly affect the model's performance and may warrant further investigation to understand their impact.

Interpretation

1. **ANOVA Table**: The p-value for the regression model is 0.064, which is slightly above the typical significance level of 0.05, indicating that the model is not statistically significant at the 5% level. The p-values for both predictors (Annual Income and Family Size) are greater than 0.05, indicating that neither is a significant predictor of Annual Food Expenditure in this model.
2. **Model Summary**: The R^2value of 27.63% indicates that approximately 27.63% of the variability in Annual Food Expenditure can be explained by the model. However, the adjusted R^2 is lower at 19.12%, suggesting that the model may not be very effective at predicting new data.
3. **Coefficients Table**: The coefficients for Annual Income and Family Size are 0.1045 and 0.73, respectively. However, their p-values (0.185 and 0.528) indicate that these predictors are not statistically significant at the 5% level.
4. **Fits and Diagnostics**: Observation 20 has a large residual, indicating it is an outlier. Observations 14 and 17 are marked as unusual X, suggesting they have unusual predictor values

In summary, relationship between annual food expenditure and the predictors (annual income and family size) is positive, the predictors are not statistically significant at the 5% level. The model explains a moderate portion (27.6%) of the variance in food expenditure, but further investigation with a larger dataset or additional predictors might be needed for more conclusive results.

2. Assume that during a three-hour period spent outside, a person recorded the outside temperature, the time spent mowing the grass, and their water consumption. Use Minitab to find the Multiple Correlation Coefficient, RRR, to determine if the amount of water consumed is dependent on the temperature and the time spent mowing the grass. Use the data given below to build the relationship among these variables. Write a summary report on your findings, which should include all relevant and sufficient hypotheses.

Water Consumption (ounces)	Time mowing the grass (hours)	Temperature (F)
16	1.85	75
20	1.25	83
25	1.50	85
27	1.75	85
32	1.15	92
48	1.75	97
48	1.60	99
24	1.45	84
31	1.36	91

Regression Analysis: Water Consumptio versus Time mowing the , Temperature (F)

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	2	1008.99	504.494	305.72	0.000
Time mowing the grass (hours)	1	77.01	77.011	46.67	0.000
Temperature (F)	1	994.44	994.435	602.61	0.000
Error	6	9.90	1.650		
Total	8	1018.89			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
1.28460	99.03%	98.70%	97.69%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-121.36	6.51	-18.65	0.000	
Time mowing the grass (hours)	13.08	1.91	6.83	0.000	1.03
Temperature (F)	1.4976	0.0610	24.55	0.000	1.03

Regression Equation

Water Consumption (ounces) = -121.36 + 13.08 Time mowing the grass (hours)
+ 1.4976 Temperature (F)

Analysis of Variance (ANOVA) Table

The ANOVA table breaks down the total variability in water consumption into components that can be attributed to the regression model and the error term. The regression model includes two predictors: time mowing the grass (hours) and temperature (F). The table shows that the regression model is highly significant, with an F-value of 305.72 and a p-value of less than 0.001. This indicates that the model explains a significant portion of the variability in water consumption.

- **Time mowing the grass (hours):** This predictor has an F-value of 46.67 and a p-value of less than 0.001, indicating it is highly significant.
- **Temperature (F):** This predictor has an F-value of 602.61 and a p-value of less than 0.001, indicating it is also highly significant.

Model Summary

The model summary provides statistics that describe the fit of the regression model:

- **S (Standard Error of the Estimate):** The standard error of the estimate is 1.28460, which indicates the average distance that the observed values fall from the regression line.
- **R^2 (Coefficient of Determination):** The R^2value is 99.03%, meaning that approximately 99.03% of the variability in water consumption can be explained by the regression model.
- **Adjusted** R^2**:** The adjusted R^2 is 98.70%, which adjusts for the number of predictors in the model. It indicates a very high level of explanatory power.
- **Predicted** R^2**:** The predicted R^2 is 97.69%, suggesting that the model is also very effective in predicting new data.

Coefficients Table

The coefficients table includes the estimated regression coefficients, their standard errors, t-values, p-values, and variance inflation factors (VIF):

- **Constant:** The constant term has a coefficient of -121.36 with a p-value of less than 0.001, indicating it is statistically significant.
- **Time mowing the grass (hours):** This predictor has a coefficient of 13.08 with a p-value of less than 0.001. This suggests that, holding temperature constant, each additional hour of mowing the grass increases water consumption by 13.08 ounces.
- **Temperature (F):** This predictor has a coefficient of 1.4976 with a p-value of less than 0.001. This suggests that, holding the time mowing the grass constant, each additional degree Fahrenheit increases water consumption by 1.4976 ounces.
- **VIF (Variance Inflation Factor):** Both predictors have VIF values of 1.03, indicating low multicollinearity.

Interpretation

In summary, the regression analysis indicates that both time spent mowing the grass and temperature are significant predictors of water consumption. The model explains a very high percentage of the variability in water consumption, as indicated by the R^2 and adjusted R^2 values. Both predictors have highly significant p-values, and the low VIF values suggest that there is no problematic multicollinearity. Therefore, this regression model can be considered reliable for predicting water consumption based on the given predictors.