Faculty of Engineering

University of Moratuwa

In20-Semester 07

MA4014 - Linear Models and Multivariate Statistics

Assignment 1

KUMARASINGHE H.K.N.S.

200322R

Due Date of Submission

[22 / 10 / 2024]

Question 1

(a) Test the hypothesis that the variable Female is not needed in the regression equation relating Sales to the six predictor variables.

This can be evaluated using a t-test to determine if the beta coefficient for the *Female* variable is statistically significant. Alternatively, an F-test can be conducted to compare the full model with a reduced model that excludes the *Female* variable.

The results of fitting the model are as follows.

```
> full_model <- lm(Sales ~ Age + HS + Income + Black + Female + Price , data = cigarette.data)
> summary(full_model)

Call:
lm(formula = Sales ~ Age + HS + Income + Black + Female + Price,
    data = cigarette.data)

Residuals:
    Min      1Q  Median      3Q     Max 
-48.398 -12.388  -5.367   6.270 133.213 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)   
(Intercept) 103.34485  245.60719   0.421  0.67597   
Age           4.52045    3.21977   1.404  0.16735   
HS           -0.06159    0.81468  -0.076  0.94008   
Income        0.01895    0.01022   1.855  0.07036 . 
Black         0.35754    0.48722   0.734  0.46695   
Female       -1.05286    5.56101  -0.189  0.85071   
Price        -3.25492    1.03141  -3.156  0.00289 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 28.17 on 44 degrees of freedom
Multiple R-squared:  0.3208,	Adjusted R-squared:  0.2282 
F-statistic: 3.464 on 6 and 44 DF,  p-value: 0.006857
```

```
> reduced_model <- lm(Sales ~ Age + HS + Income + Black + Price , data = cigarette.data)
> summary(reduced_model)

Call:
lm(formula = Sales ~ Age + HS + Income + Black + Price, data = cigarette.data)

Residuals:
    Min      1Q  Median      3Q     Max 
-47.089 -11.767  -5.525   5.650 132.873 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)   
(Intercept) 59.463336  80.387585   0.740  0.46332   
Age          4.117758   2.391184   1.722  0.09193 . 
HS          -0.066817   0.805446  -0.083  0.93425   
Income       0.019458   0.009746   1.997  0.05194 . 
Black        0.311472   0.417577   0.746  0.45961   
Price       -3.252022   1.020186  -3.188  0.00261 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 27.87 on 45 degrees of freedom
Multiple R-squared:  0.3203,	Adjusted R-squared:  0.2448 
F-statistic: 4.241 on 5 and 45 DF,  p-value: 0.003039
```

```
> anova_result = anova(reduced_model, full_model)
> anova_result
Analysis of Variance Table

Model 1: Sales ~ Age + HS + Income + Black + Price
Model 2: Sales ~ Age + HS + Income + Black + Female + Price
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1     45 34954
2     44 34926  1    28.453 0.0358 0.8507
> df1 <- anova_result[2, "Df"]
> df2 <- anova_result[2, "Res.Df"]
> p_value_from_f_dist <- 1 - pf(0.95, df1, df2)
> p_value_from_f_dist
[1] 0.3350476
> F_value_from_f_dist <- qf(0.95, df1, df2)
> F_value_from_f_dist
[1] 4.061706
> pt(0.025, 44)
[1] 0.509916
> pt(0.975, 44)
[1] 0.8325548
> qt(0.025, 44)
[1] -2.015368
> qt(0.975, 44)
[1] 2.015368
```

t-test

- **Null Hypothesis (H0):** The coefficient (β) for the variable "Female" is 0.
- **Alternative Hypothesis (H1):** The coefficient (β) for "Female" is not 0.

With a test statistic range of $-2.0153 < -0.189 < 2.0153$, this falls within the interval $t_{44,0.025} < t < t_{44,0.975}$. Therefore, we do not reject the null hypothesis, and the β\betaβ value for the "Female" variable is not statistically significant.

F-test

- **Null Hypothesis (H0):** The reduced model is sufficient.
- **Alternative Hypothesis (H1):** The reduced model is not sufficient.

With F=0.0358, which is less than $F_{2,44,0.95}$ =3.209, we do not reject the null hypothesis. This indicates that the reduced model is sufficient.

(b) Test the hypothesis that the variables Female and HS are not needed in the above regression equation.

We can perform an F-test to assess whether removing the variables "Female" and "HS" significantly impacts the model, by comparing the variances of the full and reduced models.

```
> reduced_model <- lm(Sales ~ Age + Income + Black + Price , data = cigarette.data)
> summary(reduced_model)

Call:
lm(formula = Sales ~ Age + Income + Black + Price, data = cigarette.data)

Residuals:
    Min      1Q  Median      3Q     Max 
-46.784 -11.810  -5.380   5.758 132.789 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)   
(Intercept) 55.329580  62.395293   0.887   0.3798   
Age          4.191538   2.195535   1.909   0.0625 . 
Income       0.018892   0.006882   2.745   0.0086 **
Black        0.334162   0.312098   1.071   0.2899   
Price       -3.239941   0.998778  -3.244   0.0022 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 27.57 on 46 degrees of freedom
Multiple R-squared:  0.3202,    Adjusted R-squared:  0.2611 
F-statistic: 5.416 on 4 and 46 DF,  p-value: 0.001168

> anova_result = anova(reduced_model, full_model)
> anova_result
Analysis of Variance Table

Model 1: Sales ~ Age + Income + Black + Price
Model 2: Sales ~ Age + HS + Income + Black + Female + Price
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1     46 34960                           
2     44 34926  2    33.799 0.0213 0.9789
> df1 <- anova_result[2, "Df"]
> df2 <- anova_result[2, "Res.Df"]
> p_value_from_f_dist <- 1 - pf(0.95, df1, df2)
> p_value_from_f_dist
[1] 0.3945298
> F_value_from_f_dist <- qf(0.95, df1, df2)
> F_value_from_f_dist
[1] 3.209278
```

F-test

- **Null Hypothesis (H0):** The reduced model is sufficient.
- **Alternative Hypothesis (H1):** The reduced model is not sufficient.

With F=0.0213, which is less than the critical value $F_{44,2,0.95}$ = 3.209278, we do not reject the null hypothesis. This indicates that the reduced model is sufficient without the "Female" and "HS" variables.

(c) Compute the 95% confidence interval for the true regression coefficient of the variable Income.

```
> full_model <- lm(Sales ~ Age + HS + Income + Black + Female + Price , data = cigarette.data)
> summary = summary(full_model)
> income_estimate <- summary$coefficients ["Income", "Estimate"]
> income_estimate
[1] 0.01894645
```

```
> conf_int <- confint (full_model, level = 0.95)
> conf_int
                    2.5 %       97.5 %
(Intercept) -3.916439e+02 598.33360254
Age         -1.968565e+00  11.00946945
HS          -1.703475e+00   1.58030249
Income      -1.642517e-03   0.03953542
Black       -6.243909e-01   1.33946122
Female      -1.226033e+01  10.15461632
Price       -5.333583e+00  -1.17625412
> conf_int["Income", ]
       2.5 %       97.5 %
-0.001642517  0.039535423
```

Therefore, the 95% confidence interval for the coefficient is [-0.001642517, 0.039535423]. This interval includes zero, supporting the conclusion that the coefficient is not statistically significant at the 5% significance level.

(d) What percentage of the variation in Sales can be accounted for when Income is removed from the above regression equation? Explain.

```
> reduced_model <- lm(Sales ~ Age + HS + Black + Female + Price , data = cigarette.data)
> summary(reduced_model)

Call:
lm(formula = Sales ~ Age + HS + Black + Female + Price, data = cigarette.data)

Residuals:
    Min      1Q  Median      3Q     Max 
-37.414 -16.454  -5.746   8.541 133.319 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept) 162.3245   250.0537   0.649  0.51954   
Age           7.3073     2.9238   2.499  0.01616 * 
HS            0.9717     0.6103   1.592  0.11836   
Black         0.8447     0.4213   2.005  0.05101 . 
Female       -3.7815     5.5063  -0.687  0.49576   
Price        -2.8603     1.0362  -2.760  0.00832 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 28.93 on 45 degrees of freedom
Multiple R-squared:  0.2678,	Adjusted R-squared:  0.1864 
F-statistic: 3.291 on 5 and 45 DF,  p-value: 0.01287

> # Calculate R-squared values for both models
> r_squared_full_model <- summary(full_model)$r.squared
> r_squared_full_model
[1] 0.3208426
> r_squared_reduced_model <- summary(reduced_model)$r.squared
> r_squared_reduced_model
[1] 0.2677526
> # Calculate adjusted R-squared values for both models
> adjusted_r_sq_full_model <- summary(full_model)$adj.r.squared
> adjusted_r_sq_full_model
[1] 0.2282303
> adjusted_r_sq_reduced_model <- summary(reduced_model)$adj.r.squared
> adjusted_r_sq_reduced_model
[1] 0.1863918
> # Calculate the variation explained by adding the Income variable
> variation_explained_by_income <- (adjusted_r_sq_full_model - adjusted_r_sq_reduced_model) / adjusted_r_sq_full_model * 100
> variation_explained_by_income
[1] 18.33169
```

When comparing the adjusted R^2 values of the two models:

- Adjusted R^2 (Full Model) = 0.2282
- Adjusted R^2 (Reduced Model) = 0.1864

Since 0.2282>0.18640.2282 > 0.18640.2282>0.1864, the adjusted R^2 for the full model is higher than that of the reduced model, indicating that the full model provides a better fit to the data than the reduced model.

The difference in variation explained by adding the "Income" variable is:

0.2282−0.1864=0.04180.

This accounts for an increase of approximately 18.33% in the explained variation.

(e) What percentage of the variation in Sales can be accounted for by the three variables: Price, Age, and Income? Explain.

```
> age_price_income_model <- lm(Sales ~ Price + Age + Income, data = cigarette.data)
> summary(age_price_income_model)

Call:
lm(formula = Sales ~ Price + Age + Income, data = cigarette.data)

Residuals:
    Min      1Q  Median      3Q     Max 
-50.430 -13.853  -4.962   6.691 128.947 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)   
(Intercept) 64.248227  61.933008   1.037  0.30487   
Price       -3.399234   0.989172  -3.436  0.00124 **
Age          4.155909   2.198699   1.890  0.06491 . 
Income       0.019281   0.006883   2.801  0.00737 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 27.61 on 47 degrees of freedom
Multiple R-squared:  0.3032,	Adjusted R-squared:  0.2588 
F-statistic: 6.818 on 3 and 47 DF,  p-value: 0.0006565

> r_squared_sales_age_price = summary(age_price_income_model)$r.squared
> r_squared_sales_age_price
[1] 0.3032434
> adj_r_squared_sales_age_price = summary(age_price_income_model)$adj.r.squared
> adj_r_squared_sales_age_price
[1] 0.2587696
> ratio = (adj_r_squared_sales_age_price / adjusted_r_sq_full_model) * 100
> ratio
[1] 113.3809
> _
```

Therefore, the three variables—price, age, and income—can explain a variance of 0.3032434 in sales.

The difference in variance captured by the full model and the new model is:

$$0.2587696 - 0.2282303 = 0.03060$$

This indicates that the new model outperforms the full model by capturing additional variance in the sales data.

(f) What percentage of the variation in Sales that can be accounted for by the variable Income, when Sales are regressed on only Income? Explain.

```
> income_model <- lm(Sales ~ Income, data = cigarette.data)
> summary(income_model)

Call:
lm(formula = Sales ~ Income, data = cigarette.data)

Residuals:
    Min      1Q  Median      3Q     Max
-54.550 -15.772  -6.517   4.491 144.628

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 55.362454  27.743082   1.996   0.0516 .
Income       0.017583   0.007283   2.414   0.0195 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 30.63 on 49 degrees of freedom
Multiple R-squared:  0.1063,    Adjusted R-squared:  0.08808
F-statistic: 5.829 on 1 and 49 DF,  p-value: 0.01954

> r_squared_sales_age_price = summary(income_model)$r.squared
> r_squared_sales_age_price
[1] 0.1063203
> adj_r_squared_sales_age_price = summary(income_model)$adj.r.squared
> adj_r_squared_sales_age_price
[1] 0.08808191
> ratio = (adj_r_squared_sales_age_price / adjusted_r_sq_full_model) * 100
> ratio
[1] 38.59344
> 
```

Therefore, the variable "income" can capture 0.10632030 variance in sales.

The difference in variance captured by the full model and the new model is:

$$0.2282303 - 0.08808191 = 0.140148390$$

This indicates that the full model is better than the new model, as it captures more variance in the sales data.

Question 2

Initial multiple regression model

```
> # Fit initial multiple regression model
> model1 <- lm(Y ~ X1 + X2 + X3, data = edu_expenditure.data)
> summary(model1)

Call:
lm(formula = Y ~ X1 + X2 + X3, data = edu_expenditure.data)

Residuals:
    Min      1Q  Median      3Q     Max 
-30.787  -9.202  -2.578  10.590  48.548 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -11.404627  28.976165  -0.394    0.696    
X1            0.044933   0.007667   5.860 4.69e-07 ***
X2            0.066223   0.048834   1.356    0.182    
X3           -0.028954   0.019293  -1.501    0.140    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.7 on 46 degrees of freedom
Multiple R-squared:  0.4721,	Adjusted R-squared:  0.4376 
F-statistic: 13.71 on 3 and 46 DF,  p-value: 1.608e-06
```

Model with regions

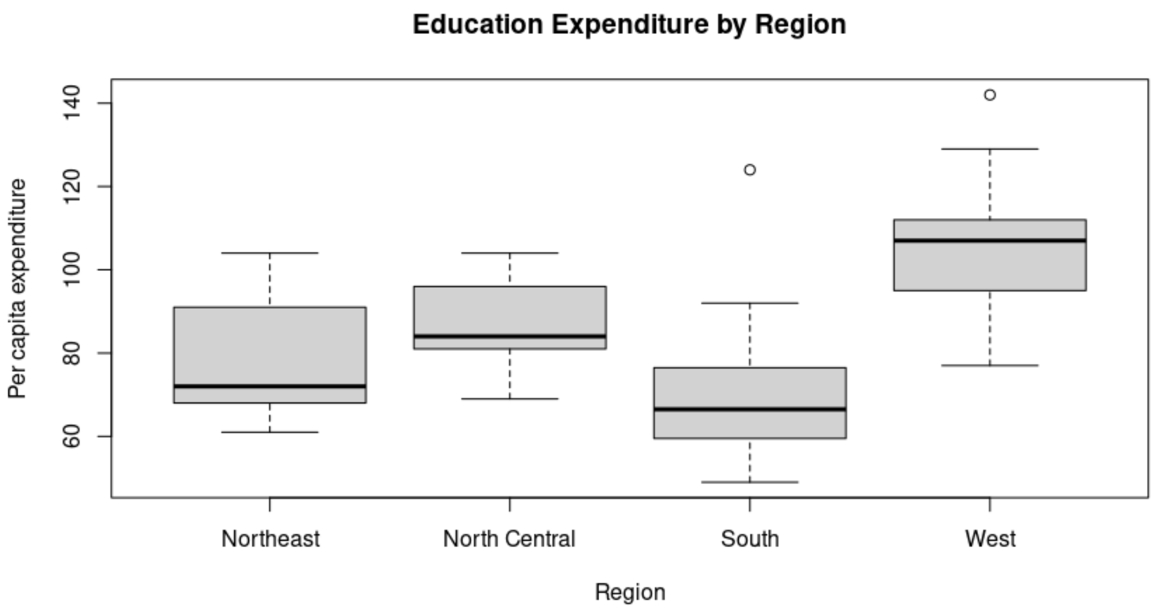
```
> # Fit model with region
> model2 <- lm(Y ~ X1 + X2 + X3 + Region, data = edu_expenditure.data)
> summary(model2)

Call:
lm(formula = Y ~ X1 + X2 + X3 + Region, data = edu_expenditure.data)

Residuals:
    Min      1Q  Median      3Q     Max 
-19.906  -5.216  -0.845   6.035  33.162 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  8.338181  21.835344   0.382  0.70444    
X1           0.038208   0.005787   6.602 4.86e-08 ***
X2           0.002590   0.036030   0.072  0.94302    
X3          -0.023655   0.013956  -1.695  0.09732 .  
Region2     15.938968   5.087174   3.133  0.00311 ** 
Region3     10.826957   5.640893   1.919  0.06159 .  
Region4     33.083081   5.123367   6.457 7.90e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.07 on 43 degrees of freedom
Multiple R-squared:  0.7546,	Adjusted R-squared:  0.7204 
F-statistic: 22.04 on 6 and 43 DF,  p-value: 1.18e-11
```

Adding regional effects significantly improved the model (R^2 increased from 0.472 to 0.755)

Basic statistics by region

```
> # Basic statistics by region
> aggregate(Y ~ Region, data = edu_expenditure.data, mean)
         Region         Y
1     Northeast  77.66667
2 North Central  87.25000
3         South  70.50000
4          West 106.00000
> aggregate(Y ~ Region, data = edu_expenditure.data, sd)
         Region        Y
1     Northeast 16.07016
2 North Central 10.46314
3         South 17.84750
4          West 17.73885
```

1. The West region spends the most on education and significantly higher than other regions
2. The South spends the least on education
3. North Central has the most consistent spending (lowest standard deviation)
4. South and West show the most variation in spending (highest standard deviations)
5. There's about a $35.50 difference between the highest spending region (West) and lowest spending region (South)



- West region has the largest box, showing wide spread in spending
- North Central has the most compact box, indicating more consistent spending
- West region has two outliers (dots above the box) showing some states with unusually high expenditure
- Clear regional differences in spending
- West consistently spends more on education

- South generally spends less
- More variation in spending in Western states
- North Central states have more uniform spending patterns

This suggests there are substantial regional differences in education expenditure, both in terms of average spending and spending consistency.

Model with interaction effects

```
> # Test for interaction effects
> model_interaction <- lm(Y ~ X1*Region + X2*Region + X3*Region, data = edu_expenditure.data)
> summary(model_interaction)

Call:
lm(formula = Y ~ X1 * Region + X2 * Region + X3 * Region, data = edu_expenditure.data)

Residuals:
     Min       1Q   Median       3Q      Max 
-17.2713  -6.4982  -0.2577   5.4192  30.9564 

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)  
(Intercept)             9.979e+01  1.949e+02   0.512   0.6119  
X1                      3.599e-02  1.339e-02   2.688   0.0111 *
RegionNorth Central    -2.458e+02  2.307e+02  -1.066   0.2941  
RegionSouth            -7.312e+01  1.965e+02  -0.372   0.7122  
RegionWest             -2.307e+02  2.157e+02  -1.070   0.2922  
X2                     -2.105e-01  4.408e-01  -0.477   0.6361  
X3                     -3.694e-02  3.508e-02  -1.053   0.2997  
X1:RegionNorth Central -3.542e-02  4.441e-02  -0.798   0.4307  
X1:RegionSouth          2.318e-03  1.515e-02   0.153   0.8793  
X1:RegionWest           3.597e-02  2.540e-02   1.416   0.1659  
RegionNorth Central:X2  7.219e-01  5.005e-01   1.442   0.1584  
RegionSouth:X2          1.884e-01  4.424e-01   0.426   0.6729  
RegionWest:X2           4.757e-01  4.664e-01   1.020   0.3150  
RegionNorth Central:X3  9.350e-02  8.511e-02   1.099   0.2797  
RegionSouth:X3          1.944e-02  4.232e-02   0.459   0.6489  
RegionWest:X3           4.679e-03  4.940e-02   0.095   0.9251  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.58 on 34 degrees of freedom
Multiple R-squared:  0.8227,	Adjusted R-squared:  0.7445 
F-statistic: 10.52 on 15 and 34 DF,  p-value: 9.764e-09
```

Adding regional effects significantly improved the model (R^2 increased from 0.755 to 0.8227)

But None of the interactions between regions and variables (X1, X2, X3) are statistically significant:

- Income (X1) interactions with regions: all $p > 0.16$
- Youth population (X2) interactions with regions: all $p > 0.15$
- Urban population (X3) interactions with regions: all $p > 0.27$

The gain in R^2 (from 0.7546 to 0.8227) isn't substantial enough to justify the added complexity

Anova test

The ANOVA test provides statistical evidence to justify why we prefer the simpler model despite the higher R^2 in the interaction model.

```
> # ANOVA comparison of models
> anova(model_region, model_interaction)
Analysis of Variance Table

Model 1: Y ~ X1 + X2 + X3 + Region
Model 2: Y ~ X1 * Region + X2 * Region + X3 * Region
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     43 5267.5
2     34 3806.3  9    1461.2 1.4502 0.2064
```

- F-statistic = 1.4502
- p-value = 0.2064 (> 0.05)

The high p-value (0.2064) indicates that the improvement from adding interactions is not statistically significant

- Residual degrees of freedom: decreased from 43 to 34 (lost 9 df)
- RSS (Residual Sum of Squares) decreased from 5267.5 to 3806.3

The decrease in RSS isn't large enough relative to the loss in degrees of freedom to justify the more complex model

Therefore, we can formally conclude that while the interaction model has a higher R^2 (0.8227 vs 0.7546), the improvement is not statistically significant ($p = 0.2064$). This statistical test supports our decision to prefer the simpler model without interactions.