22401 HW2

Bin Yu

January 27, 2025

Question 1

(a)

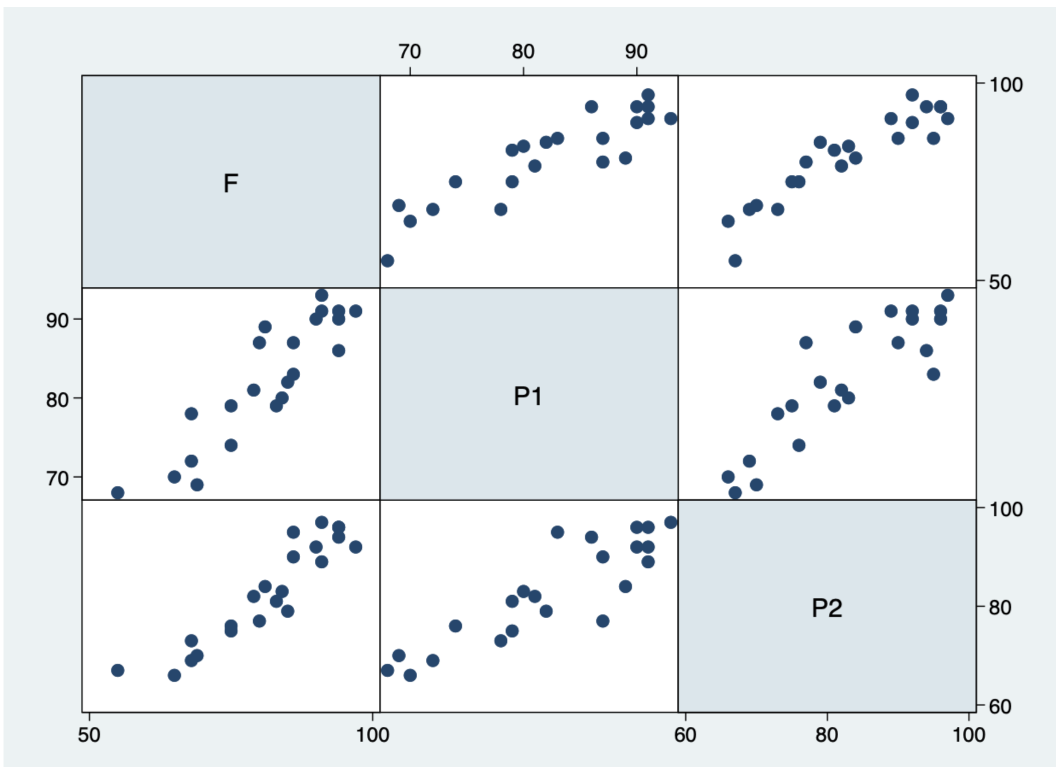

Figure 1: relationships among the three variables

Comment:

- **Strong Linear Relationships:** There appears to be a strong positive linear relationship between *F* (final assessment) and *P1* (preliminary assessment 1). Similarly, *F* and *P2* (preliminary assessment 2) also exhibit a positive linear relationship, but it seems slightly weaker compared to the relationship between *F* and *P1*.
- **Correlation Between P1 and P2:** The scatterplot between *P1* and *P2* shows a positive linear relationship, indicating that individuals who perform well in one preliminary assessment tend to perform well in the other.

(b)

```
. * F = β0 + β1 * P1 + ε
. regress F P1

      Source |       SS           df       MS      Number of obs   =        22
-------------+----------------------------------   F(1, 20)        =     81.14
       Model |  2094.74806         1  2094.74806   Prob > F        =    0.0000
    Residual |  516.342849        20  25.8171425   R-squared       =    0.8023
-------------+----------------------------------   Adj R-squared   =    0.7924
       Total |  2611.09091        21  124.337662   Root MSE        =    5.0811

------------------------------------------------------------------------------
           F |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          P1 |   1.260516   .1399383     9.01   0.000     .9686097    1.552422
       _cons |  -22.34244   11.56395    -1.93   0.068    -46.46442     1.77955
------------------------------------------------------------------------------
```

Figure 2: Model 1: $F = \beta_0 + \beta_1 P_1 + \epsilon$

```
. * F = β0 + β2 * P2 + ε
. regress F P2

      Source |       SS           df       MS      Number of obs   =        22
-------------+----------------------------------   F(1, 20)        =    122.89
       Model |  2245.63144         1  2245.63144   Prob > F        =    0.0000
    Residual |  365.459467        20  18.2729734   R-squared       =    0.8600
-------------+----------------------------------   Adj R-squared   =    0.8530
       Total |  2611.09091        21  124.337662   Root MSE        =    4.2747

------------------------------------------------------------------------------
           F |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          P2 |   1.004267   .0905909    11.09   0.000     .8152974    1.193236
       _cons |  -1.853547   7.561809    -0.25   0.809     -17.6272    13.92011
------------------------------------------------------------------------------
```

Figure 3: Model 2: $F = \beta_0 + \beta_2 P_2 + \epsilon$

```
. * F = β0 + β1 * P1 + β2 * P2 + ε
. regress F P1 P2

      Source |       SS           df       MS      Number of obs   =        22
-------------+----------------------------------   F(2, 19)        =     74.07
       Model |  2314.26087         2  1157.13043   Prob > F        =    0.0000
    Residual |  296.830042        19  15.6226338   R-squared       =    0.8863
-------------+----------------------------------   Adj R-squared   =    0.8744
       Total |  2611.09091        21  124.337662   Root MSE        =    3.9525

------------------------------------------------------------------------------
           F |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          P1 |   .4883376   .2329926     2.10   0.050     .0006785    .9759966
          P2 |   .6720356   .1792831     3.75   0.001     .2967916     1.04728
       _cons |  -14.50054   9.235645    -1.57   0.133    -33.83097     4.82989
------------------------------------------------------------------------------
```

Figure 4: Model 3: $F = \beta_0 + \beta_1 P_1 + \beta_2 P_2 + \epsilon$

Figure 5: Regression results

(c)

- **Model 1:** $F = \beta_0 + \beta_1 P_1 + \epsilon$
  From the regression results:
  - $\beta_1 = 1.2605$ (p-value < 0.001), indicating that there is a significant positive linear relationship between P_1 and F. More specifically, a one-unit increase in P_1 is associated with an average increase of 1.26 in F, holding all else constant.
  - The R-squared value is 0.8023, meaning that 80.23% of the variation in F is explained by P_1.
- **Model 2:** $F = \beta_0 + \beta_2 P_2 + \epsilon$
  From the regression results:
  - $\beta_2 = 1.0043$ (p-value < 0.001), indicating that there is a significant positive linear relationship between P_2 and F. More specifically, a one-unit increase in P_2 is associated with an average increase of 1.004 in F, holding all else constant.
  - The R-squared value is 0.8600, meaning that 86.00% of the variation in F is explained by P_2.

(d)

Based on the R-squared values:

- P_2 (R-squared = 0.8600) explains more variation in F than P_1 (R-squared = 0.8023), making P_2 a better individual predictor of F.

(e)

From Model 3 ($F = \beta_0 + \beta_1 P_1 + \beta_2 P_2 + \epsilon$):

- $\beta_2 = 0.6720$ indicates that for a one-unit increase in P_2, F increases by 0.672 on average, holding P_1 unchanged. This accounts for the effect of P_1 when interpreting the effect of P_2.

(f)

Using the coefficients from Model 3:

$$\hat{F} = \beta_0 + \beta_1 P_1 + \beta_2 P_2$$

$$\hat{F} = -14.5005 + 0.4883(78) + 0.6720(85)$$

$$\hat{F} = -14.5005 + 38.0874 + 57.1200 = 80.7069$$

The predicted final assessment score is approximately 80.7069.

Question 2

Given a simple linear regression model:

$$Y = \beta_0 + \beta_1 X_1 + \epsilon$$

with 20 observations ($n = 20$).

Calculations

1. Degrees of Freedom

$$df_{\text{total}} = n - 1 = 20 - 1 = 19$$
$$df_{\text{model}} = 1$$
$$df_{\text{residual}} = df_{\text{total}} - df_{\text{model}} = 19 - 1 = 18$$

2. Calculating the Coefficient for X_1

Given:

$$\text{t-test for } X_1 = 8.32$$
$$\text{Std. Error for } X_1 = 0.1528$$
$$\text{Coefficient for } X_1 = t \times \text{Std. Error} = 8.32 \times 0.1528 \approx 1.271$$

3. F-Statistic Calculation

In simple linear regression, the F-statistic is related to the t-statistic of the predictor by:

$$F = t^2$$

Given the t-test for X_1 is 8.32:

$$F = (8.32)^2 \approx 69.2224$$

4. Mean Square Residual (MSE)

The F-statistic is also the ratio of the Mean Square Regression (MSR) to the Mean Square Error (MSE):

$$F = \frac{MSR}{MSE}$$

Given:

$$MSR = \frac{SSR}{1} = 1848.76$$
$$69.2224 = \frac{1848.76}{MSE}$$
$$MSE = \frac{1848.76}{69.2224} \approx 26.7075$$

$$\hat{\sigma} = \sqrt{MSE} = \sqrt{26.7075} \approx 5.1679$$

5. Sum of Squares for Residuals (SSE)

$$SSE = MSE \times df_{\text{residual}} = 26.7075 \times 18 = 480.735$$

6. Total Sum of Squares (SST)

$$SST = SSR + SSE = 1848.76 + 480.735 = 2329.495$$

$$MST = SST/n - 1 = \frac{2329.495}{19} \approx 122.605$$

7. Coefficient of Determination(R^2)

$$R^2 = \frac{SSR}{SST} = \frac{1848.76}{2329.495} \approx 0.7936$$

8. Adjusted Coefficient of Determination ($R^2_{\textbf{adjusted}}$)

$$R^2_{\text{adjusted}} = 1 - \left(\frac{(1-R^2)(n-1)}{n-2}\right) = 1 - \left(\frac{(1-0.7936)(19)}{18}\right) = 1 - (0.2179) = 0.7821$$

9. Variance of Y (Var(Y))

The total variance of Y is calculated as:

$$\text{Var}(Y) = \frac{SST}{n-1} = \frac{2329.495}{19} \approx 122.605$$

This represents the overall variability in the response variable Y.

10. Standard Error and t-test for the Constant Term

Given:

$$\text{Coefficient for Constant} = -23.4325$$
$$\text{std} = 12.74$$
$$t = \frac{\text{Coefficient} - 0}{std} = \frac{-23.4325}{12.74} \approx -1.8392$$

Completed ANOVA Table

ANOVA Table				
Source	**Sum of Squares**	**df**	**Mean Square**	**F-Test**
Model	1848.76	1	1848.76	69.2224
Residuals	480.735	18	26.7075	
Total	2329.495	19	122.605	

Completed Coefficients Table

Coefficients Table				
	Coefficient	**Std. Error**	**t-test**	**p-value**
Constant	-23.4325	12.74	-1.8392	0.0824
X_1	1.271	0.1528	8.32	< 0.0001
$n = 20$	$R^2 = 0.7936$	$R^2_{adjusted} = 0.7821$	$\hat{\sigma} = 5.1679$(root of MSE)	

$$\text{Var}(Y) = 122.605$$

Question 3

(a)

Stata Output:

```
. correlate BMI age cholest glucose
(obs=58)

             |      BMI      age  cholest  glucose
-------------+------------------------------------
         BMI |   1.0000
         age |   0.1863   1.0000
     cholest |   0.2814   0.1697   1.0000
     glucose |   0.2211   0.2671   0.0827   1.0000
```

Figure 6: Correlation Matrix of BMI Predictors

Comment:

- **BMI and Age:** The correlation coefficient is 0.1863, indicating a weak positive correlation. This suggests that as age increases, BMI tends to increase slightly, but the relationship is not strong.
- **BMI and Cholesterol:** The correlation coefficient is 0.2814, showing a positive correlation. Higher cholesterol levels are slightly associated with higher BMI.
- **BMI and Glucose:** The correlation coefficient is 0.2211, indicating a positive correlation. Higher glucose levels have a slight association with higher BMI.
- **Age and Cholesterol:** Correlation of 0.1697 suggests apositive relationship between age and cholesterol, but not very strong.
- **Age and Glucose:** Correlation of 0.2671 indicates a positive relationship between the age and glucose.
- **Cholesterol and Glucose:** Correlation of 0.0827 shows a very weak positive relationship between the cholesterol and glucose.

Overall, cholesterol shows the strongest correlation with BMI among the predictors, followed by glucose and age.

(b)

Regression of BMI on Age

Stata Output:

```
. regress BMI age

      Source |       SS           df       MS      Number of obs   =        58
-------------+----------------------------------   F(1, 56)        =      2.01
       Model |  86.3186659         1  86.3186659   Prob > F        =    0.1615
    Residual |  2401.20521        56  42.8786644   R-squared       =    0.0347
-------------+----------------------------------   Adj R-squared   =    0.0175
       Total |  2487.52387        57  43.6407697   Root MSE        =    6.5482

------------------------------------------------------------------------------
         BMI |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0785774   .0553817     1.42   0.161    -.0323654    .1895202
       _cons |   26.30496   2.815726     9.34   0.000     20.66438    31.94554
------------------------------------------------------------------------------
```

Figure 7: Simple Linear Regression of BMI on Age

Model Summary:

$$R^2 = 0.0347, \quad \text{Adjusted } R^2 = 0.0175$$
$$\text{F}(1, 56) = 2.01, \quad \text{Prob} > \text{F} = 0.1615$$
$$\text{Root MSE} = 6.5482$$

Interpretation:

- **Coefficient for Age**: 0.0786 suggests that for each additional year of age, BMI increases by approximately 0.0786 units on average. However, this effect is not statistically significant ($p = 0.161$).
- **Model Fit**: The model explains only 3.47% of the variance in BMI, indicating a poor fit.

2. Regression of BMI on Cholesterol

Stata Output:

```
. regress BMI cholest

      Source |       SS           df       MS      Number of obs   =        58
-------------+----------------------------------   F(1, 56)        =      4.81
       Model |  196.932516         1  196.932516   Prob > F        =    0.0324
    Residual |  2290.59136        56  40.9034171   R-squared       =    0.0792
-------------+----------------------------------   Adj R-squared   =    0.0627
       Total |  2487.52387        57  43.6407697   Root MSE        =    6.3956

------------------------------------------------------------------------------
         BMI |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     cholest |   .0542134   .0247075     2.19   0.032     .0047185    .1037084
       _cons |   20.00025   4.683001     4.27   0.000     10.61907    29.38143
------------------------------------------------------------------------------
```

Figure 8: Simple Linear Regression of BMI on Cholesterol

Model Summary:

$$R^2 = 0.0792, \quad \text{Adjusted } R^2 = 0.0627$$
$$\text{F}(1,\, 56) = 4.81, \quad \text{Prob} > \text{F} = 0.0324$$
$$\text{Root MSE} = 6.3956$$

Interpretation:

- **Coefficient for Cholesterol**: 0.0542 indicates that for each unit increase in cholesterol, BMI increases by 0.0542 units on average. This effect is statistically significant ($p = 0.032$).
- **Model Fit**: The model explains 7.92% of the variance in BMI, which is still relatively low.

3. Regression of BMI on Glucose

Stata Output:

```
. regress BMI glucose

      Source |       SS           df       MS      Number of obs   =        58
-------------+----------------------------------   F(1, 56)        =      2.88
       Model |  121.563209         1  121.563209   Prob > F        =    0.0954
    Residual |  2365.96066        56  42.2492976   R-squared       =    0.0489
-------------+----------------------------------   Adj R-squared   =    0.0319
       Total |  2487.52387        57  43.6407697   Root MSE        =    6.4999

------------------------------------------------------------------------------
         BMI |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     glucose |   .0250111   .0147449     1.70   0.095    -.0045265    .0545488
       _cons |   27.16779   1.932712    14.06   0.000     23.29611    31.03948
------------------------------------------------------------------------------
```

Figure 9: Simple Linear Regression of BMI on Glucose

Model Summary:

$$R^2 = 0.0489, \quad \text{Adjusted } R^2 = 0.0319$$
$$\text{F}(1,\, 56) = 2.88, \quad \text{Prob} > \text{F} = 0.0954$$
$$\text{Root MSE} = 6.4999$$

Interpretation:

- **Coefficient for Glucose**: 0.0250 suggests that for each unit increase in glucose, BMI increases by approximately 0.0250 units on average. This effect is not significant ($p = 0.095$).
- **Model Fit**: The model explains 4.89% of the variance in BMI, indicating a poor fit.

Overall Summary

- **Cholesterol** is the only predictor with a statistically significant relationship with BMI at the 0.05 level.
- **Age** and **Glucose** show non-significant relationships with BMI.
- All models have low R^2 values, indicating that each predictor alone explains only a small portion of the variability in BMI.

(c)

Stata Output:

```
. regress BMI age cholest glucose

      Source |       SS           df       MS      Number of obs   =        58
-------------+----------------------------------   F(3, 54)        =      2.62
       Model |  316.246324         3  105.415441   Prob > F        =    0.0599
    Residual |  2171.27755        54  40.2088435   R-squared       =    0.1271
-------------+----------------------------------   Adj R-squared   =    0.0786
       Total |  2487.52387        57  43.6407697   Root MSE        =     6.341

------------------------------------------------------------------------------
         BMI |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0410063     .05632     0.73   0.470    -.0719085    .1539212
     cholest |   .0482577   .0248764     1.94   0.058    -.0016165    .0981319
     glucose |   .0197315   .0149381     1.32   0.192    -.0102177    .0496806
       _cons |   16.80503   5.072879     3.31   0.002     6.634519    26.97554
------------------------------------------------------------------------------
```

Figure 10: Multiple Linear Regression of BMI on Age, Cholesterol, and Glucose

Model Summary:

$$R^2 = 0.1271, \quad \text{Adjusted } R^2 = 0.0786$$
$$\text{F(3, 54)} = 2.62, \quad \text{Prob} > \text{F} = 0.0599$$
$$\text{Root MSE} = 6.341$$

Comment

- **Age**: The coefficient is 0.0410 but is not statistically significant ($p = 0.470$).
- **Cholesterol**: The coefficient is 0.0483, marginally significant ($p = 0.058$).
- **Glucose**: The coefficient is 0.0197 and not statistically significant ($p = 0.192$).
- **Model Fit**: The model explains 12.71% of the variance in BMI, with an adjusted R^2 of 7.86%, it is increased comparing to the simple linear regression model, but still a poor fit.

Overall Summary

- **Cholesterol** remains the most significant predictor in the multiple regression model, albeit marginally significant.
- **Age** and **Glucose** do not significantly predict BMI when controlling for the other variables.
- The combined model has a slightly better fit than individual models but still explains a limited portion of BMI variability.

(d)

Stepwise Selection Based on $p < 0.15$

Stata Output:

```
. stepwise, pr(.15): regress BMI age cholest glucose
                      begin with full model
p = 0.4697 >= 0.1500  removing age

      Source |       SS           df       MS      Number of obs   =        58
-------------+----------------------------------   F(2, 55)        =      3.70
       Model |  294.930722         2  147.465361   Prob > F        =    0.0311
    Residual |  2192.59315        55    39.86533   R-squared       =    0.1186
-------------+----------------------------------   Adj R-squared   =    0.0865
       Total |  2487.52387        57  43.6407697   Root MSE        =    6.3139

------------------------------------------------------------------------------
         BMI |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     glucose |   .0225336   .0143721     1.57   0.123    -.0062687    .0513358
     cholest |   .0510412   .0244757     2.09   0.042     .0019908    .1000916
       _cons |   17.94173   4.806009     3.73   0.000     8.310277    27.57319
------------------------------------------------------------------------------
```

Figure 11: Stepwise Selection Regression of BMI on Glucose and Cholesterol

Model Summary:

$$R^2 = 0.1186, \quad \text{Adjusted } R^2 = 0.0865$$

$$\text{F(2, 55)} = 3.70, \quad \text{Prob} > \text{F} = 0.0311$$

$$\text{Root MSE} = 6.3139$$

Interpretation:

- **Cholesterol**: Remains a significant predictor ($p = 0.042$).
- **Glucose**: Still a non-significant predictor ($p = 0.123$) but near significant (as $p < 0.15$.
- **Age**: Removed from the model as it was not significant.
- **Model Fit**: The reduced model explains 11.86% of the variance in BMI, with an adjusted R^2 of 8.65%.

Based on the stepwise selection and manual evaluation, the most suitable reduced model includes **Cholesterol** and **Glucose**, as both have p-values below or near the 0.15 threshold. And we can see this model increased the adjusted R^2 comparing to the model with all 3 predictors.

(e)

Stata Output:

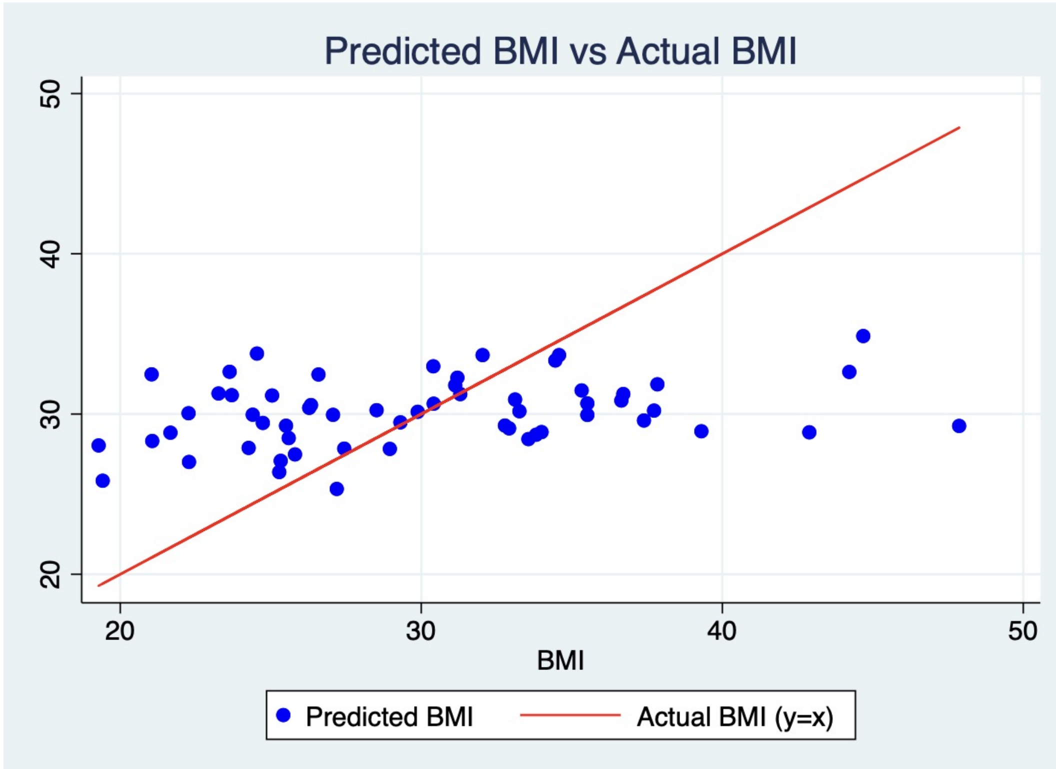

Figure 12: Predicted BMI vs Actual BMI

Based on Figure 12, the scatterplot of predicted BMI ($\hat{y}$) against actual BMI values does not appear to be fairly random. Specifically, the points show a systematic deviation rather than random scatter around the line $y = x$ (where predicted value = actual value):

- For lower actual BMI values, the model tends to overestimate, giving higher predicted BMI values than the actual ones.
- Conversely, for higher actual BMI values, the model underestimates, providing lower predicted values compared to the actual BMI.

Question 4

(a)

Hypotheses:

$$H_0 : \beta_{\text{Age}} = \beta_{\text{HS}} = \beta_{\text{Income}} = \beta_{\text{AA}} = \beta_{\text{Female}} = \beta_{\text{Price}} = 0$$
$$H_A : \text{At least one } \beta_i \neq 0 \quad \text{for } i \in \{\text{Age}, \text{HS}, \text{Income}, \text{AA}, \text{Female}, \text{Price}\}$$

Test Used:

- **F-test**: To assess the joint significance of all six predictor variables.

$$F = \frac{\text{MSR}}{\text{MSE}} = \frac{\left(\frac{\text{SSR}}{k}\right)}{\left(\frac{\text{SSE}}{n-k-1}\right)}$$

Where:

- SSR = Sum of Squares of Regression
- SSE = Sum of Squares of Error
- k = Number of predictors (here, $k = 6$)
- n = Sample size = 51

Stata Output:

```
. test Age HS Income AA Female Price

 ( 1)  Age = 0
 ( 2)  HS = 0
 ( 3)  Income = 0
 ( 4)  AA = 0
 ( 5)  Female = 0
 ( 6)  Price = 0

       F(  6,    44) =    3.46
            Prob > F =    0.0069
```

Figure 13: Stata Output for Part (a): Overall Significance Test

Conclusion:

Since the p-value (0.0069) is less than the significance level of 0.05, we reject the null hypothesis. This indicates that there are at least some useful predictors among the six candidate variables in explaining the variation in Sales.

(b)

Hypotheses:

$$H_0 : \beta_{\text{Female}} = 0$$
$$H_A : \beta_{\text{Female}} \neq 0$$

Test Used:

- **t-test**: To assess the significance of the *Female* coefficient.
- Alternatively, an **F-test** can be used for testing a single coefficient.

$$t = \frac{\hat{\beta}_{\text{Female}}}{\text{SE}(\hat{\beta}_{\text{Female}})}$$

Where:

- $\hat{\beta}_{\text{Female}}$ = Estimated coefficient for *Female*
- $\text{SE}(\hat{\beta}_{\text{Female}})$ = Standard error of the *Female* coefficient

For a single coefficient, the F-statistic is:

$$F = t^2$$

Stata Output:

```
. test Female

 ( 1)  Female = 0

       F(  1,    44) =    0.04
            Prob > F =    0.8507
```

Figure 14: Stata Output for Part (b): Significance of *Female* Variable

Conclusion:

Since the p-value (0.8507) is much greater than 0.05, we fail to reject the null hypothesis. This suggests that the *Female* variable is not a significant predictor of Sales and may not be needed in the regression model.

(c)

Hypotheses:

$$H_0 : \beta_{\text{Female}} = 0 \quad \text{and} \quad \beta_{\text{HS}} = 0$$
$$H_A : \text{At least one of } \beta_i \neq 0 \quad \text{for } i \in \{\text{Female}, \text{HS}\}$$

Test Used:

- **F-test**: To assess the joint significance of the *Female* and *HS* coefficients. We use the reduced model that doesn't include female and hs, and the full model that include all 6 predictors.

$$F = \frac{(\text{SSR}_{\text{reduced}} - \text{SSR}_{\text{full}})/q}{\text{SSR}_{\text{full}}/(n-k-1)}$$

Where:

- $\text{SSR}_{\text{reduced}}$ = Regression Sum of Squares for the reduced model (H_0 true)

- SSR_{full} = Regression Sum of Squares for the full model
- $q = df_{reduced} - df_{full} = 2$
- n = Sample size = 51
- k = Number of predictors in the full model

Stata Output:

```
. test Female HS

 ( 1)  Female = 0
 ( 2)  HS = 0

       F(  2,    44) =    0.02
            Prob > F =    0.9789
```

Figure 15: Stata Output for Part (c): Joint Significance of *Female* and *HS* Variables

Conclusion:

Since the joint p-value (0.9789) is significantly higher than 0.05, we fail to reject the null hypothesis. This indicates that neither *Female* nor *HS* are significant predictors of Sales when considered together. Therefore, what we can do to these predictors is that both variables can be excluded from the regression model.

(d)

Model:

$$\text{Sales} = \beta_0 + \beta_1 \cdot \text{Age} + \beta_2 \cdot \text{Income} + \beta_3 \cdot \text{AA} + \beta_4 \cdot \text{Price} + \epsilon$$

Stata Output:

```
. regress Sales Age Income AA Price

      Source |       SS           df       MS      Number of obs   =        51
-------------+----------------------------------   F(4, 46)        =      5.42
       Model |  16465.6761         4  4116.41902   Prob > F        =    0.0012
    Residual |  34959.7693        46  759.994986   R-squared       =    0.3202
-------------+----------------------------------   Adj R-squared   =    0.2611
       Total |  51425.4454        50  1028.50891   Root MSE        =    27.568

------------------------------------------------------------------------------
       Sales |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         Age |   4.191537   2.195535     1.91   0.062    -.2278452     8.61092
      Income |   .0188921   .0068822     2.75   0.009      .005039    .0327452
          AA |   .3341623   .3120983     1.07   0.290    -.2940589    .9623836
       Price |  -3.239941   .9987778    -3.24   0.002    -5.250376   -1.229506
       _cons |   55.32961   62.39529     0.89   0.380     -70.2656    180.9248
------------------------------------------------------------------------------
```

Figure 16: Stata Output for Part (d): 95% Confidence Interval for 'Income' Coefficient

Conclusion:

The 95% confidence interval for the *Income* coefficient is [0.0050, 0.0327]. Since this interval does not include zero, we conclude that *Income* is a statistically significant predictor of Sales at the 5% significance level.

(e)

Model:

$$\text{Sales} = \beta_0 + \beta_1 \cdot \text{Age} + \beta_2 \cdot \text{AA} + \beta_3 \cdot \text{Price} + \epsilon$$

Stata Output:

```
. regress Sales Age AA Price
```

Source	SS	df	MS
Model	10738.7868	3	3579.5956
Residual	40686.6586	47	865.673588
Total	51425.4454	50	1028.50891

Number of obs	= 51
F(3, 47)	= 4.14
Prob > F	= 0.0111
R-squared	= 0.2088
Adj R-squared	= 0.1583
Root MSE	= 29.422

Sales	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
Age	5.490043	2.288183	2.40	0.020	.8868126	10.09327
AA	.3793874	.3326267	1.14	0.260	-.2897713	1.048546
Price	-2.781837	1.050974	-2.65	0.011	-4.896124	-.6675495
_cons	72.87416	66.24195	1.10	0.277	-60.38746	206.1358

Figure 17: Stata Output for Part (e): R-squared After Removing Income, Female, and HS

Conclusion:

The R-squared value after removing *Income*, *Female*, and *HS* is 0.2088. This means that approximately 20.88% of the variability in Sales is explained by the remaining predictors: Age, AA, and Price.

(f)

Model:

$$\text{Sales} = \beta_0 + \beta_1 \cdot \text{Price} + \beta_2 \cdot \text{Age} + \beta_3 \cdot \text{Income} + \epsilon$$

Stata Output:

```
. regress Sales Price Age Income
```

Source	SS	df	MS
Model	15594.4257	3	5198.1419
Residual	35831.0197	47	762.362122
Total	51425.4454	50	1028.50891

Number of obs	= 51
F(3, 47)	= 6.82
Prob > F	= 0.0007
R-squared	= 0.3032
Adj R-squared	= 0.2588
Root MSE	= 27.611

Sales	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
Price	-3.399234	.9891719	-3.44	0.001	-5.389191	-1.409277
Age	4.155908	2.198699	1.89	0.065	-.2673039	8.579119
Income	.019281	.0068833	2.80	0.007	.0054337	.0331284
_cons	64.24826	61.93301	1.04	0.305	-60.34488	188.8414

Figure 18: Stata Output for Part (f): R-squared with Price, Age, and Income

Conclusion:

The R-squared value with *Price*, *Age*, and *Income* is 0.3032, indicating that approximately 30.32% of the variability in Sales is explained by these three predictors.

(g)

Model:

$$\text{Sales} = \beta_0 + \beta_1 \cdot \text{Income} + \epsilon$$

Stata Output:

```
. regress Sales Income

      Source |       SS           df       MS      Number of obs   =        51
-------------+----------------------------------   F(1, 49)        =      5.83
       Model |  5467.56673         1  5467.56673   Prob > F        =    0.0195
    Residual |  45957.8787        49  937.915892   R-squared       =    0.1063
-------------+----------------------------------   Adj R-squared   =    0.0881
       Total |  51425.4454        50  1028.50891   Root MSE        =    30.625

------------------------------------------------------------------------------
       Sales |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      Income |   .0175834   .0072826     2.41   0.020     .0029484    .0322184
       _cons |   55.36246   27.74308     2.00   0.052    -.3893547    111.1143
------------------------------------------------------------------------------
```

Figure 19: Stata Output for Part (g): R-squared with Income Alone

Conclusion:

The R-squared value with *Income* alone is 0.1063, meaning that approximately 10.63% of the variability in Sales is explained by *Income* alone.

(h)

I use stepwise regression for model selection. When use $p < 0.05$, the final model will not include age as a predictor. When use $p < 0.1$, the final model will include age as a predictor.

Model with Age:

$$\text{Sales} = \beta_0 + \beta_1 \cdot \text{Age} + \beta_2 \cdot \text{Income} + \beta_3 \cdot \text{Price} + \epsilon$$

Stata Output:

```
. stepwise, pr(0.10): regress Sales Age HS Income AA Female Price
                      begin with full model
p = 0.9401 >= 0.1000  removing HS
p = 0.8469 >= 0.1000  removing Female
p = 0.2899 >= 0.1000  removing AA

      Source |       SS           df       MS      Number of obs   =        51
-------------+----------------------------------   F(3, 47)        =      6.82
       Model |  15594.4257         3  5198.1419   Prob > F        =    0.0007
    Residual |  35831.0197        47  762.362122   R-squared       =    0.3032
-------------+----------------------------------   Adj R-squared   =    0.2588
       Total |  51425.4454        50  1028.50891   Root MSE        =    27.611

------------------------------------------------------------------------------
       Sales |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         Age |   4.155908   2.198699     1.89   0.065    -.2673039    8.579119
       Price |  -3.399234   .9891719    -3.44   0.001    -5.389191   -1.409277
      Income |    .019281   .0068833     2.80   0.007     .0054337    .0331284
       _cons |   64.24826   61.93301     1.04   0.305    -60.34488    188.8414
------------------------------------------------------------------------------
```

Figure 20: Stata Output for Part (h): Final Model with Age

Model without Age:

$$\text{Sales} = \beta_0 + \beta_1 \cdot \text{Income} + \beta_2 \cdot \text{Price} + \epsilon$$

Stata Output:

```
. stepwise, pr(.05): regress Sales Age HS Income AA Female Price
                      begin with full model
p = 0.9401 >= 0.0500  removing HS
p = 0.8469 >= 0.0500  removing Female
p = 0.2899 >= 0.0500  removing AA
p = 0.0649 >= 0.0500  removing Age

      Source |       SS           df       MS      Number of obs   =        51
-------------+----------------------------------   F(2, 48)        =      8.01
       Model |  12870.7112         2  6435.3556   Prob > F        =    0.0010
    Residual |  38554.7342        48   803.22363   R-squared       =    0.2503
-------------+----------------------------------   Adj R-squared   =    0.2190
       Total |  51425.4454        50  1028.50891   Root MSE        =    28.341

------------------------------------------------------------------------------
       Sales |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      Income |    .022078   .0069001     3.20   0.002     .0082043    .0359516
       Price |  -3.017565    .993955    -3.04   0.004    -5.016045   -1.019084
       _cons |   153.3384    41.2389     3.72   0.001     70.42206    236.2548
------------------------------------------------------------------------------
```

Figure 21: Stata Output for Part (h): Final Model without Age

Comparison

Table 1: Comparison of Models With and Without *Age*

Metric	**With Age**	**Without Age**
R-squared	0.3032	0.2503
Adj R-squared	0.2588	0.2190
Root MSE	27.611	28.341
Coefficients		
Income	0.019281 (Std. Err: 0.0068833)	0.022078 (Std. Err: 0.0069001)
Price	-3.399234 (Std. Err: 0.9891719)	-3.017565 (Std. Err: 0.993955)
Age	4.155908 (Std. Err: 2.198699)	–

- The model with *Age* has a higher R-squared (0.3032) compared to the model without *Age* (0.2503), indicating that the model with *Age* explains a greater proportion of the variability in Sales.
- The Root MSE is slightly lower in the model with *Age* (27.611) compared to the model without *Age* (28.341), suggesting a better fit of model with age.
- The standard errors of the coefficients for *Income* and *Price* are slightly lower in the model with *Age* (*Income*: 0.0068833 vs. 0.0069001; *Price*: 0.9891719 vs. 0.993955), indicating more precise estimates.

Final Model Selection:

Based on the comparison, the final model with *Age* is preferred due to its higher R-squared, lower Root MSE, and more precise coefficient estimates. Therefore, the final regression model is:

$$\text{Sales} = \beta_0 + \beta_1 \cdot \text{Income} + \beta_2 \cdot \text{Price} + \beta_3 \cdot \text{Age} + \epsilon$$

Final Model Interpretation:

- **Income**: For each additional unit increase in *Income*, Sales increase by approximately 0.022 packs per capita on average, holding *Price* and *Age* constant. This relationship is statistically significant (p = 0.019).
- **Price**: For each unit increase in *Price*, Sales decrease by approximately 3.02 packs per capita on average, holding *Income* and *Age* constant. This relationship is statistically significant (p = 0.004).
- **Age**: For each additional year increase in *Age*, Sales increase by approximately 4.16 packs per capita on average, holding *Income* and *Price* constant. This relationship is marginally significant (p = 0.065).

Stata Output for Final Model:

```
. stepwise, pr(0.10): regress Sales Age HS Income AA Female Price
                      begin with full model
p = 0.9401 >= 0.1000  removing HS
p = 0.8469 >= 0.1000  removing Female
p = 0.2899 >= 0.1000  removing AA
```

Source	SS	df	MS
Model	15594.4257	3	5198.1419
Residual	35831.0197	47	762.362122
Total	51425.4454	50	1028.50891

Number of obs	=	51
F(3, 47)	=	6.82
Prob > F	=	0.0007
R-squared	=	0.3032
Adj R-squared	=	0.2588
Root MSE	=	27.611

Sales	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
Age	4.155908	2.198699	1.89	0.065	-.2673039	8.579119
Price	-3.399234	.9891719	-3.44	0.001	-5.389191	-1.409277
Income	.019281	.0068833	2.80	0.007	.0054337	.0331284
_cons	64.24826	61.93301	1.04	0.305	-60.34488	188.8414

Figure 22: Stata Output for Part (h): Final Regression Model with *Age*