Assignment 1

MA 4014 Linear Models and Multivariate Statistics

Index number: 200417M

All R codes used in this assignment can be found in the GitHub repository at https://github.com/Saeedha-N/linear-models-assignment.git.

Q1.

Data Analysis

This study investigates cigarette consumption across all 50 U.S. states and the District of Columbia using regression analysis. The original linear model considering the given six predictor variables for the response variable `Sales` is as follows:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6$$

Where, Y = `Sales`, X_1 = `Age`, X_2 = `HS`, X_3 = `Income`, X_4 = `Black`, X_5 = `Female`, and X_6 = `Price`.

The following classical assumptions must hold to validly interpret for linear regression and related tests:

- The relationship between predictors and the response is **linear**
- Observations are **independent**
- **Homoscedasticity**
- **Normality** of residuals
- **No multicollinearity**

Firstly, we will verify whether the above assumptions hold.

Model summary

```
> ex1 <- read.table(file = "ex1.txt", header = TRUE)
> full_model <- lm(Sales ~ Age + HS + Income + Black + Female + Price, data = ex1)
> summary(full_model)

Call:
lm(formula = Sales ~ Age + HS + Income + Black + Female + Price,
    data = ex1)

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
```

The model summary indicates a **small value for adjusted R²**, meaning the variation in `Sales` is not captured properly by the six predictors. Only the predictor `Price` is statistically significant. Existing outliers or influential points could be a reason for these issues.

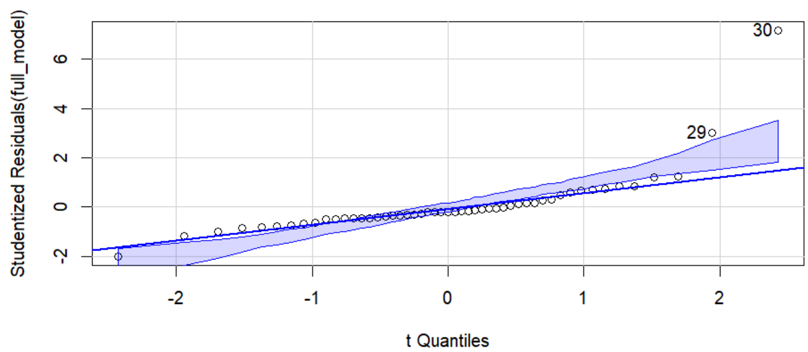
Verify Assumptions

Verify homoscedasticity assumption

```
> ncvTest(full_model) # check for heteroscedasticity
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 3.496498, Df = 1, p = 0.061499
>
```

This result indicates no strong evidence of heteroscedasticity, meaning **homoscedasticity is likely present** and the assumption of constant variance is reasonably satisfied.

Verify normal distribution of residuals



The Q-Q plot shows that most residuals **follow** the expected **normal distribution**, but observations 29 and especially 30, strongly violate the normality assumption.

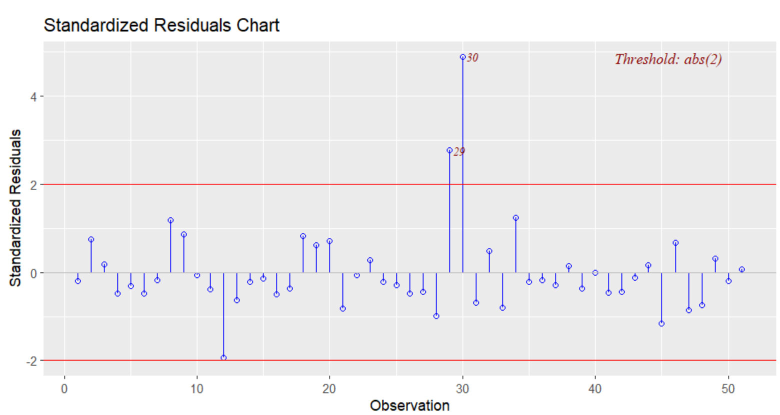
```
> shapiro.test(full_model$residuals) # check for normality

        Shapiro-Wilk normality test

data:  full_model$residuals
W = 0.73931, p-value = 3.629e-08
```

The Shapiro-Wilk test indicates a significant **violation of the normality assumption**. Observations 29 and 30 could strongly be a reason for this.

Verify independence of observations assumption



The standardized residuals chart shows no systematic pattern or trend across observations, suggesting that the observations likely **satisfy the independent assumption**. However, it also shows that observations 29 and 30 exceed the ±2 threshold, indicating they are outliers with unusually large prediction errors.

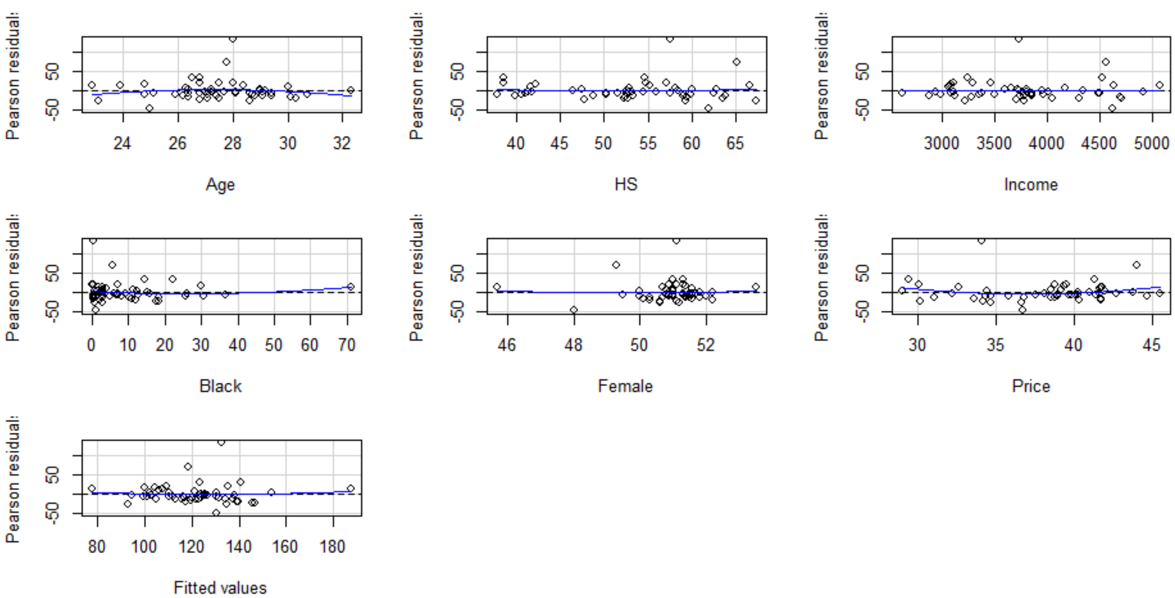
Verify no multicollinearity assumption

```
> vif(full_model) # check for multicollinearity
     Age       HS   Income    Black   Female    Price
2.300617 2.676465 2.325164 2.392152 2.406417 1.142181
> |
```

All VIF values are below 5, indicating that **multicollinearity is not a concern** in the model.

<u>Verify linearity assumption</u>

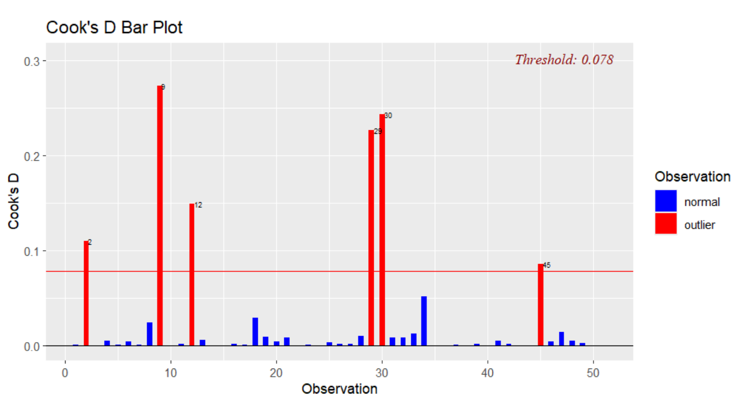
```
> residualPlots(full_model) # Pearson residuals
           Test stat Pr(>|Test stat|)
Age          -0.9865           0.3294
HS            0.3838           0.7030
Income       -0.1817           0.8567
Black         0.9245           0.3604
Female        0.2547           0.8001
Price         1.3529           0.1832
Tukey test    0.4710           0.6376
```



The residualPlots output shows that none of the predictors exhibit significant nonlinearity, as all p-values are well above 0.05. Hence, transformations of the predictor variables are not needed. This verifies our assumption that the relationship between the predictors and the response is linear, **satisfying the linearity assumption**. However, several outliers can be noticed across all plots. We identify the outliers through the following plots.

<u>Outlier Identification</u>

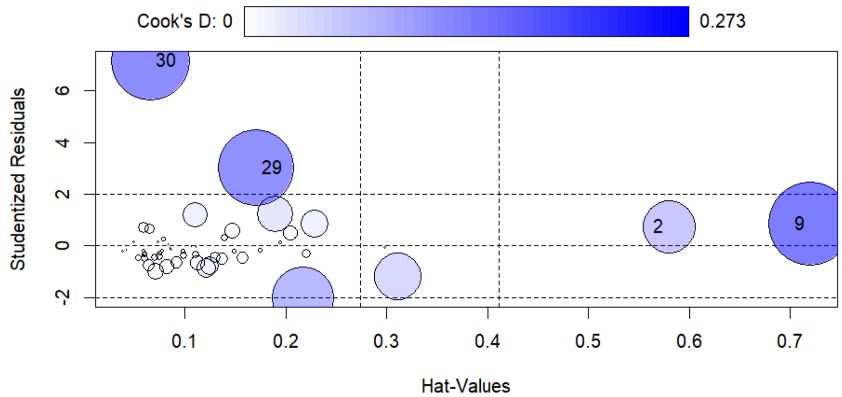
Cook's D bar plot:



This Cook's D bar plot shows that observations 2, 9, 12, 23, 30, and 45 exceed the influence threshold (0.078), indicating they are potentially influential outliers.

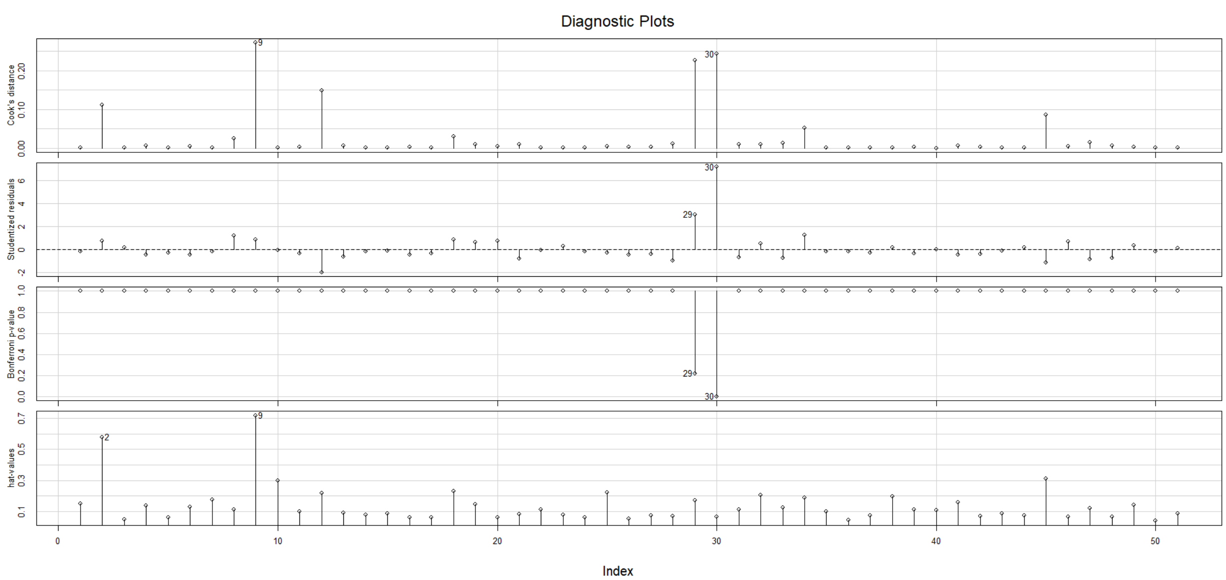
Influence plot:

```
> influencePlot(full_model)
      StudRes        Hat     CookD
2   0.7442479 0.58016035 0.1104657
9   0.8597660 0.71971284 0.2727724
29  3.0179945 0.17115641 0.2268847
30  7.1652223 0.06634506 0.2430739
```



This influence plot shows that observations 30, 29, 9, and 2 are influential, with high studentized residuals and/or hat values, and large Cook's D values, which can disproportionately affect the regression model's estimates.

Diagnostic plot:



This diagnostic plot confirms that observations 29 and 30 stand out as major outliers and influential points, with high studentized residuals, low Bonferroni-adjusted p-values, and notable Cook's distance and hat values.

Considering the above outputs, I proceeded to remove observations 2, 9, 29, and 30.

Outlier removal:

```
# remove outliers
ex1_clean <- ex1[-c(2, 9, 29, 30), ]
full_model_clean <- lm(Sales ~ Age + HS + Income + Black + Female + Price, data = ex1_clean)
```

Post-removal of outliers

Model summary:

```
> summary(full_model_clean)

Call:
lm(formula = Sales ~ Age + HS + Income + Black + Female + Price,
    data = ex1_clean)

Residuals:
    Min      1Q  Median      3Q     Max 
-23.358  -8.047  -2.164   6.683  38.204 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -3.102e+02  1.787e+02  -1.736 0.090340 .  
Age          1.021e+00  1.643e+00   0.621 0.537982    
HS          -1.243e+00  4.775e-01  -2.604 0.012870 *  
Income       1.974e-02  4.952e-03   3.986 0.000277 ***
Black       -6.324e-01  3.814e-01  -1.658 0.105102    
Female       1.024e+01  4.002e+00   2.560 0.014343 *  
Price       -3.349e+00  5.528e-01  -6.058 3.92e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.59 on 40 degrees of freedom
Multiple R-squared:  0.5896,    Adjusted R-squared:  0.528 
F-statistic: 9.576 on 6 and 40 DF,  p-value: 1.58e-06
```

This updated regression model, after removing influential observations, shows a significant improvement in fit with an **Adjusted R-squared of 0.528**, meaning ~52.8% of the variation in cigarette sales is explained by the predictors. Additionally, the predictor variables, `Income` ($p = 0.0003$), `HS` ($p = 0.0129$), and `Female` ($p = 0.0143$) are now statistically significant.

The model's overall p-value (1.58e-06) confirms it is statistically significant. Removing outliers improved both the model's explanatory power and the significance of key predictors.

Shapiro-Wilk test:

```
> shapiro.test(full_model_clean$residuals) # check for normality

	Shapiro-Wilk normality test

data:  full_model_clean$residuals
W = 0.95608, p-value = 0.0752
```

The Shapiro-Wilk test for the updated regression model indicates that the **normality assumption is reasonably satisfied**.

Thus, we can say that the updated regression model on the cleaned dataset successfully satisfies all aforementioned assumptions necessary for regression analysis.

(a). **Test the hypothesis that the variable `Female` is not needed**

From the above **model summary**, we can already see that the **variable `Female` is statistically significant,** hence we can easily **reject this hypothesis.**

However, to further assess this, we conducted an F-test by comparing a reduced model (excluding `Female`) and a full model (including `Female`), using the cleaned dataset `ex1_clean`.

```
# (a)
model_no_female_clean <- lm(Sales ~ Age + HS + Income + Black + Price, data = ex1_clean)
```

Hypothesis testing:

H_0: Reduced model is adequate
H_1: Full model is adequate

ANOVA output:

```
> anova(model_no_female_clean, full_model_clean)
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     41 8602.6
2     40 7391.4  1    1211.2 6.5546 0.01434 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> |
```

This shows a 1211.2 reduction in RSS (residual sum of squares), with an F-statistic of 6.5546 and a corresponding p-value of 0.01434 (<0.05).

Thus, we reject the null hypothesis and conclude that the full model is adequate, i.e., variable **Female is statistically significant and improves the explanatory power of the regression model**. This further aligns with the model summary (after removing outliers) above.

(b). **Test the hypothesis that both the variables Female and HS are not needed**

From the above **model summary**, we can already see that the variables **Female and HS are statistically significant,** hence we can easily **reject this hypothesis.** However, to further assess this, we conducted an F-test by comparing a reduced model (excluding both Female and HS) and a full model (including both), using the cleaned dataset ex1_clean.

```
# (b)
model_no_female_hs_clean <- lm(Sales ~ Age + Income + Black + Price, data = ex1_clean)
```

Hypothesis testing:

H_0: Reduced model is adequate
H_1: Full model is adequate

ANOVA output:

```
> anova(model_no_female_hs_clean, full_model_clean)
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     42 9689.4
2     40 7391.4  2      2298 6.218 0.004453 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> |
```

- **Reduction in RSS:** 2298
- **Diff. in DoF:** 2 (= 42 - 40)
- **F-statistic:** 6.218
- **p-value:** 0.004453 (< 0.01)

Since the p-value is well below 0.05, we reject the null hypothesis and conclude that the full model is significantly better. This means that at least one of the variables, **Female or HS, adds explanatory power to the model**. These findings are also consistent with the full model summary, where both predictors showed statistical significance after removing outliers.

(c). **<u>95% CI for the true regression coefficient of the `Income` variable</u>**

```
> # (c)
> confint(full_model_clean, "Income", level = 0.95)
             2.5 %     97.5 %
Income 0.009731336 0.02974993
```

The 95% confidence interval for the regression coefficient of `Income` is **[0.0097, 0.0297]**, which means we are 95% confident that the true effect of `Income` on `Sales` lies within this range.

Since the interval does not include 0, it indicates that `Income` is a statistically significant predictor of `Sales` in the model using the cleaned dataset. This further aligns with the model summary after removing outliers, where the `Income` variable shows statistical significance.

(d). **<u>Percentage of variation in `Sales` when `Income` is removed</u>**

According to the model summary (after removing outliers) above, the adjusted R-squared for the full model (including `Income`) is 0.528.

Model summary after removing `Income` variable:

```
> # (d)
> model_no_income_clean <- lm(Sales ~ Age + HS + Black + Female + Price, data = ex1_clean)
> summary(model_no_income_clean)

Call:
lm(formula = Sales ~ Age + HS + Black + Female + Price, data = ex1_clean)

Residuals:
    Min      1Q  Median      3Q     Max 
-32.221  -9.820  -2.100   8.632  50.925 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -282.0223   208.5018  -1.353   0.1836    
Age            3.9297     1.7184   2.287   0.0274 *  
HS            -0.1405     0.4543  -0.309   0.7587    
Black         -0.1189     0.4191  -0.284   0.7781    
Female         8.0503     4.6275   1.740   0.0894 .  
Price         -2.9740     0.6360  -4.676 3.16e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.87 on 41 degrees of freedom
Multiple R-squared:  0.4265,	Adjusted R-squared:  0.3566 
F-statistic: 6.099 on 5 and 41 DF,  p-value: 0.0002613
```

We can see that the adjusted R-squared value of the reduced model (without `Income`) has dropped to 0.3566.

Explanation

By comparing the full model (which includes `Income`) to the reduced model (without `Income`), we find that the adjusted R-squared drops from 0.528 to 0.3566. Thus, only **35.66% of the variation in `Sales` can be accounted for when `Income` is removed from the original regression equation**. The drop in adjusted R-squared after removing `Income` confirms that the variation in the response variable cannot be accurately explained using the remaining predictors, ultimately leading to a notably poorer model fit.

This suggests that `Income` is a meaningful predictor, as indicated by its statistical significance in the full model's summary.

(e). **Percentage of variation in `Sales` from `Price`, `Age`, and `Income` variables**

Model summary with variables `Price`, `Age`, and `Income`:

```
> # (e)
> model_price_age_income_clean <- lm(Sales ~ Price + Age + Income, data = ex1_clean)
> summary(model_price_age_income_clean)

Call:
lm(formula = Sales ~ Price + Age + Income, data = ex1_clean)

Residuals:
    Min      1Q  Median      3Q     Max 
-34.167  -7.879  -2.500   6.898  44.561 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 63.783365  35.992686   1.772  0.08346 .  
Price       -2.859677   0.606371  -4.716 2.55e-05 ***
Age          4.580830   1.400088   3.272  0.00211 ** 
Income       0.009283   0.004434   2.093  0.04224 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.51 on 43 degrees of freedom
Multiple R-squared:  0.4254,	Adjusted R-squared:  0.3854 
F-statistic: 10.61 on 3 and 43 DF,  p-value: 2.392e-05
```

We can see that the adjusted R-squared value of this reduced model has dropped to 0.3854.

Explanation

By comparing the full model (which includes all predictors) to the reduced model (which includes only `Price`, `Age`, and `Income`), we find that the adjusted R-squared decreases from 0.528 to 0.3854. This means that only **38.54% of the variation in `Sales` can be accounted for by the three predictor variables: Price, Age, and Income.**

This indicates a moderate explanatory power, showing that these variables together contribute meaningfully to predicting cigarette sales. As seen in the full model's summary, `Price` and `Income` are statistically significant (Pr(>|t|) < 0.05); however, a significant portion of variation remains unexplained because other statistically significant variables, such as `Female` and `HS`, are excluded.

(f). **<u>Percentage of variation in `Sales` from only `Income` variable</u>**

We need to fit a simple linear regression model with only `Income` as the predictor of `Sales`.

Model summary with `Income` variable alone:

```
> # (f)
> model_income_only_clean <- lm(Sales ~ Income, data = ex1_clean)
> summary(model_income_only_clean)

Call:
lm(formula = Sales ~ Income, data = ex1_clean)

Residuals:
    Min      1Q  Median      3Q     Max 
-45.282  -9.465  -2.218   7.162  61.377 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 79.616634  19.008482   4.188 0.000129 ***
Income       0.009658   0.005080   1.901 0.063719 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19.25 on 45 degrees of freedom
Multiple R-squared:  0.07434,	Adjusted R-squared:  0.05377 
F-statistic: 3.614 on 1 and 45 DF,  p-value: 0.06372
```

We can see that the adjusted R-squared value has significantly dropped to 0.05377 in this reduced simple linear regression model.

<u>Explanation</u>

By comparing the full model (which includes all predictors) to the reduced simple linear regression model (which includes only `Income`), we find that the adjusted R-squared decreases significantly from 0.528 to 0.05377. This means that when `Sales` is regressed on `Income` alone, **only 5.377% of the variation in `Sales` can be accounted for by `Income`**.

This low percentage indicates that although `Income` is statistically significant in the full model summary, it is a weak predictor of cigarette sales on its own. Excluding other statistically significant variables such as `Price`, `Female`, and `HS` substantially weakens the model's explanatory power, highlighting the importance of including multiple relevant predictors.

Q2.