Azreen Haque

4/11/2025

Solutions

Problem 1

Solutions below.

Part A

```
data <- read.csv("hw07pr01.csv", header = TRUE, sep = ",")

# Fit linear model
model1 <- lm(steroid ~ age, data = data)

# Get summary and store it in 's'
s <- summary(model1)

# Extract regression components
intercept <- s$coefficients["(Intercept)", "Estimate"]
slope <- s$coefficients["age", "Estimate"]
r_squared <- s$r.squared
adj_r_squared <- s$adj.r.squared
p_value_slope <- s$coefficients["age", "Pr(>|t|)"]

# Print all extracted values
cat("Fitted equation: Ŷ =", round(intercept, 5), "+", round(slope, 5), "* age\n")
```

```
## Fitted equation: Ŷ = -15.88032 + 2.21459 * age
```

```
cat("R-squared:", round(r_squared, 4), "\n")
```

```
## R-squared: 0.9498
```

```
cat("Adjusted R-squared:", round(adj_r_squared, 4), "\n")
```

```
## Adjusted R-squared: 0.9492
```

```
cat("P-value for age coefficient:", format.pval(p_value_slope, digits = 3), "\n")
```

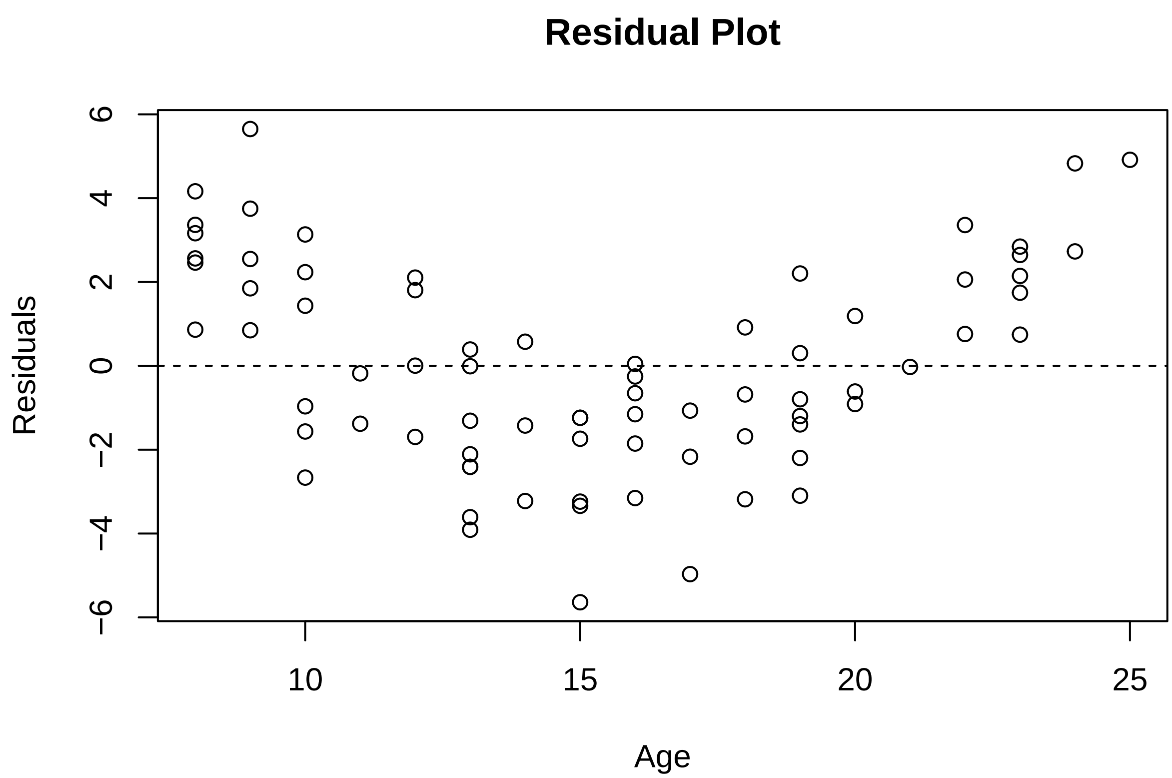
```
## P-value for age coefficient: <2e-16
```

```
cat(
  "The p-value for the slope coefficient ( 1) is less than 2e-16,\n",
  "which is extremely small. This provides strong evidence against\n",
  "the null hypothesis H0:  1 = 0.\n\n",
  "We conclude that age is a statistically significant predictor\n",
  "of steroid level in this sample of 75 individuals.\n"
)
```

```
## The p-value for the slope coefficient ( 1) is less than 2e-16,
##  which is extremely small. This provides strong evidence against
##  the null hypothesis H0:  1 = 0.
##
##  We conclude that age is a statistically significant predictor
##  of steroid level in this sample of 75 individuals.
```

Part B

```r
# Create residual plot
plot(data$age, resid(model1),
     main = "Residual Plot",
     xlab = "Age",
     ylab = "Residuals")
abline(h = 0, lty = 2)
```




```r
# Create standardized residual plot
sse <- sum(resid(model1)^2)
n <- nrow(data)
mse <- sse / (n - 2)
standardized_res <- resid(model1) / sqrt(mse)

plot(data$age, standardized_res,
     main = "Standardized Residual Plot",
```

```r
     xlab = "Age",
     ylab = "Standardized Residuals",
     ylim = c(-3.5, 3.5))
abline(h = 0, lty = 2)
abline(h = c(-3, 3), lty = 3)
```


```r
cat(
  "The residual plot for the linear model shows a U-shaped pattern,\n",
  "suggesting that the relationship between age and steroid level\n",
  "is not adequately captured by a simple linear model.\n\n",

  "This is a violation of the linearity assumption.\n\n",

  "The variance of residuals appears roughly constant,\n",
  "and no standardized residuals exceed ±3,\n",
  "so the constant variance and outlier assumptions\n",
  "are reasonably met.\n\n",

  "Conclusion: The linear model violates the linearity assumption,\n",
  "and a more appropriate model may be quadratic."
)
```

```
## The residual plot for the linear model shows a U-shaped pattern,
##  suggesting that the relationship between age and steroid level
##  is not adequately captured by a simple linear model.
```

```
##
##  This is a violation of the linearity assumption.
##
##  The variance of residuals appears roughly constant,
##  and no standardized residuals exceed ±3,
##  so the constant variance and outlier assumptions
##  are reasonably met.
##
##  Conclusion: The linear model violates the linearity assumption,
##  and a more appropriate model may be quadratic.
```

Part C

```r
# Create quadratic term
data$age2 <- data$age^2

# Fit quadratic regression model
model2 <- lm(steroid ~ age + age2, data = data)

# Summarize the model
s2 <- summary(model2)

# Extract coefficients
b0 <- s2$coefficients["(Intercept)", "Estimate"]
b1 <- s2$coefficients["age", "Estimate"]
b2 <- s2$coefficients["age2", "Estimate"]

# Extract R-squared and Adjusted R-squared
r2 <- s2$r.squared
adj_r2 <- s2$adj.r.squared

# Extract p-value for  1 (age)
p_b1 <- s2$coefficients["age", "Pr(>|t|)"]

# Print results
cat("Fitted equation:\n")
```

```
## Fitted equation:
```

```r
cat("Ŷ = ", round(b0, 5), " + ", round(b1, 5), "* age + ", round(b2, 5), "* age²\n\n")
```

```
## Ŷ =  2.89313  +  -0.41824 * age +  0.08365 * age²
```

```r
cat("R-squared: ", round(r2, 4), "\n")
```

```
## R-squared:  0.9796
```

```r
cat("Adjusted R-squared: ", round(adj_r2, 4), "\n")
```

```
## Adjusted R-squared:  0.979
```

```r
cat("P-value for  1 (age): ", format.pval(p_b1, digits = 3), "\n")
```

```
## P-value for  1 (age):  0.112
```

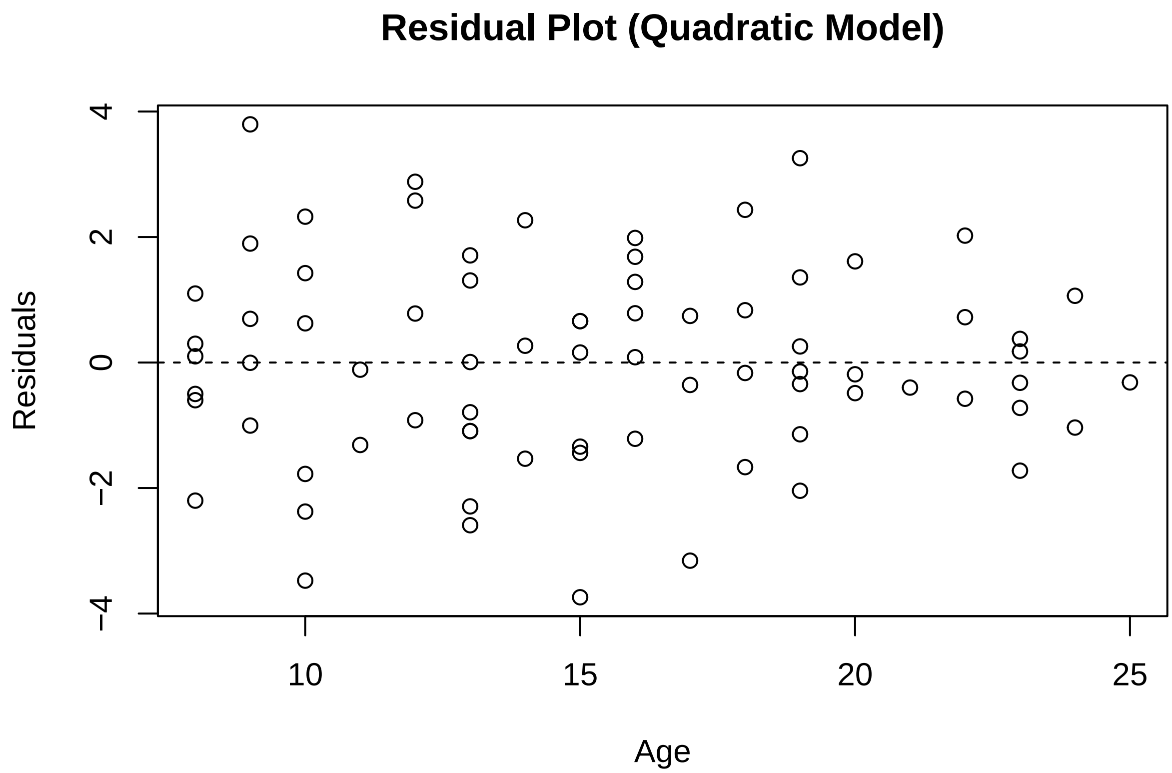
```r
cat(
  "Comparison to Part (a):\n",
  "The quadratic model has a higher R² (0.9796 vs. 0.9498)\n",
  "and higher adjusted R² (0.979 vs. 0.9492),\n",
  "indicating a better fit.\n\n",

  "However, the p-value for age is no longer significant (0.112),\n",
  "suggesting age may not be linearly associated with steroid\n",
  "after accounting for the quadratic term.\n"
)
```

```
## Comparison to Part (a):
##  The quadratic model has a higher R² (0.9796 vs. 0.9498)
##  and higher adjusted R² (0.979 vs. 0.9492),
##  indicating a better fit.
##
##  However, the p-value for age is no longer significant (0.112),
##  suggesting age may not be linearly associated with steroid
##  after accounting for the quadratic term.
```

Part D

```r
# Residual plot for quadratic model
plot(data$age, resid(model2),
     main = "Residual Plot (Quadratic Model)",
     xlab = "Age",
     ylab = "Residuals")
abline(h = 0, lty = 2)
```




```r
# Standardized residuals
sse2 <- sum(resid(model2)^2)
n <- nrow(data)
mse2 <- sse2 / (n - 3)  # 3 parameters: intercept, age, age^2
standardized_res2 <- resid(model2) / sqrt(mse2)

# Standardized residual plot
plot(data$age, standardized_res2,
     main = "Standardized Residual Plot (Quadratic)",
     xlab = "Age",
     ylab = "Standardized Residuals",
     ylim = c(-3.5, 3.5))
abline(h = 0, lty = 2)
abline(h = c(-3, 3), lty = 3)
```


```r
cat(
  "The residual plot for the quadratic model shows\n",
  "no clear pattern or curvature, suggesting that\n",
  "linearity is reasonably addressed.\n\n",

  "The residuals appear to be evenly spread across age,\n",
  "indicating that the constant variance assumption holds.\n\n",

  "All standardized residuals are within ±3, so there\n",
  "are no major outliers or influential points.\n\n",

  "Conclusion: There are no noticeable violations\n",
  "of regression assumptions in the quadratic model."
)
```

```
## The residual plot for the quadratic model shows
##  no clear pattern or curvature, suggesting that
##  linearity is reasonably addressed.
##
##  The residuals appear to be evenly spread across age,
##  indicating that the constant variance assumption holds.
##
##  All standardized residuals are within ±3, so there
##  are no major outliers or influential points.
##
##  Conclusion: There are no noticeable violations
```

```
##  of regression assumptions in the quadratic model.
```

Part E

```r
# Compute correlation between X and X^2
cor_age_age2 <- cor(data$age, data$age2)
cat("Correlation between age and age²:", round(cor_age_age2, 4), "\n")
```

```
## Correlation between age and age²: 0.9891
```

```r
cat(
  "The correlation between age and age² is very close to 1.\n",
  "This indicates strong collinearity between the two predictors.\n\n",
  "When two variables are highly correlated, including both\n",
  "in a regression model may lead to multicollinearity.\n",
  "This can inflate standard errors and make it harder\n",
  "to interpret the individual effect of each predictor."
)
```

```
## The correlation between age and age² is very close to 1.
##  This indicates strong collinearity between the two predictors.
##
##  When two variables are highly correlated, including both
##  in a regression model may lead to multicollinearity.
##  This can inflate standard errors and make it harder
##  to interpret the individual effect of each predictor.
```

Part F

```r
# Center the age variable
data$x <- data$age - mean(data$age)

# Compute x²
data$x2 <- data$x^2

# Correlation between x and x²
cor_x_x2 <- cor(data$x, data$x2)

# Print result and compare
cat("Correlation between x and x²:", round(cor_x_x2, 4), "\n\n")
```

```
## Correlation between x and x²: 0.1916
```

```r
cat(
  "Correlation between x and x²: 0.1916\n\n",
  "Compared to part (e), the correlation dropped\n",
  "from 0.9891 to 0.1916 after centering age.\n\n",
  "This reduces multicollinearity between\n",
  "the linear and quadratic terms.\n\n"
)
```

```
## Correlation between x and x²: 0.1916
##
##  Compared to part (e), the correlation dropped
##  from 0.9891 to 0.1916 after centering age.
##
##  This reduces multicollinearity between
##  the linear and quadratic terms.
```

Part G

```r
# Fit quadratic model using centered x and x^2
model_centered <- lm(steroid ~ x + x2, data = data)

# Get summary
s_centered <- summary(model_centered)

# Extract coefficients
b0_c <- s_centered$coefficients["(Intercept)", "Estimate"]
b1_c <- s_centered$coefficients["x", "Estimate"]
b2_c <- s_centered$coefficients["x2", "Estimate"]

# Extract R-squared, adjusted R-squared, and p-value for  1
r2_c <- s_centered$r.squared
adj_r2_c <- s_centered$adj.r.squared
pval_b1_c <- s_centered$coefficients["x", "Pr(>|t|)"]

# Print result
cat("Fitted equation:\n")
```

```
## Fitted equation:
```

```r
cat("Ŷ = ", round(b0_c, 5), " + ", round(b1_c, 5), "* x + ", round(b2_c, 5), "* x²\n\n")
```

```
## Ŷ =  16.0324  +  2.13804 * x +  0.08365 * x²
```

```r
cat("R-squared: ", round(r2_c, 4), "\n")
```

```
## R-squared:  0.9796
```

```r
cat("Adjusted R-squared: ", round(adj_r2_c, 4), "\n")
```

```
## Adjusted R-squared:  0.979
```

```r
cat("P-value for  1 (x): ", format.pval(pval_b1_c, digits = 3), "\n\n")
```

```
## P-value for  1 (x):  <2e-16
```

```r
cat("Compared to part (c):\n",
  "- R² and adjusted R² remain the same because the model fit\n",
  "  hasn't changed, just the variable scale.\n",
  "- The coefficient for  1 is easier to interpret now,\n",
  "  since x is centered.\n",
  "- The p-value for  1 is now highly significant, showing\n",
  "  improved stability due to reduced multicollinearity.\n"
)
```

```
## Compared to part (c):
##  - R² and adjusted R² remain the same because the model fit
##    hasn't changed, just the variable scale.
##  - The coefficient for  1 is easier to interpret now,
##    since x is centered.
##  - The p-value for  1 is now highly significant, showing
##    improved stability due to reduced multicollinearity.
```

Problem 2

Part A

```r
# Load data
data2 <- read.csv("hw07pr02.csv", header = TRUE)

# Fit regression model with price and discount only
model_a <- lm(market.share ~ price + discount, data = data2)

summary(model_a)
```

```
##
## Call:
## lm(formula = market.share ~ price + discount, data = data2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.37824 -0.18357 -0.06490  0.08805  1.87274
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.8684     0.4572   6.274 3.37e-07 ***
## price        -0.1391     0.1904  -0.730   0.4700
## discount      0.2749     0.1214   2.265   0.0298 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3684 on 35 degrees of freedom
## Multiple R-squared:  0.1347, Adjusted R-squared:  0.08523
## F-statistic: 2.724 on 2 and 35 DF,  p-value: 0.07954
```

```r
cat(
  "Fitted equation:\n",
  "Ŷ = 2.8684 - 0.1391 * price + 0.2749 * discount\n\n",

  "Interpretation:\n",
  "  (price): For every $1 increase in price, the market share\n",
  "is expected to decrease by 0.1391 units,\n",
  "holding discount constant. (p = 0.470 - not significant)\n\n",

  "  (discount): Discounted products have, on average,\n",
  "0.2749 units higher market share than non-discounted products,\n",
  "holding price constant. (p = 0.0298 - statistically significant)\n\n",

  "Group-specific equations:\n",
  "Non-discounted (discount = 0):\n",
  "Ŷ = 2.8684 - 0.1391 * price\n\n",
  "Discounted (discount = 1):\n",
  "Ŷ = 3.1433 - 0.1391 * price\n"
)
```

```
## Fitted equation:
##  Ŷ = 2.8684 - 0.1391 * price + 0.2749 * discount
##
##  Interpretation:
##    (price): For every $1 increase in price, the market share
##  is expected to decrease by 0.1391 units,
##  holding discount constant. (p = 0.470 - not significant)
##
##    (discount): Discounted products have, on average,
##  0.2749 units higher market share than non-discounted products,
##  holding price constant. (p = 0.0298 - statistically significant)
##
##  Group-specific equations:
##  Non-discounted (discount = 0):
##  Ŷ = 2.8684 - 0.1391 * price
##
##  Discounted (discount = 1):
##  Ŷ = 3.1433 - 0.1391 * price
```

Part B

```r
# Fit model with interaction between price and discount
model_b <- lm(market.share ~ price * discount, data = data2)

# Show summary output
summary(model_b)
```

```
##
## Call:
## lm(formula = market.share ~ price * discount, data = data2)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.73574 -0.14380  0.03335  0.11914  1.05201
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -2.1049     1.1809  -1.782 0.083617 .
## price            1.9755     0.5011   3.942 0.000382 ***
## discount         5.7756     1.2445   4.641  5.0e-05 ***
## price:discount  -2.3345     0.5265  -4.434  9.2e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2975 on 34 degrees of freedom
## Multiple R-squared:  0.4517, Adjusted R-squared:  0.4033
## F-statistic: 9.336 on 3 and 34 DF,  p-value: 0.0001208
```

```r
cat(
  "Fitted equation:\n",
  "Ŷ = -2.1049 + 1.9755 * price + 5.7756 * discount\n",
  "     - 2.3345 * (price × discount)\n\n",

  "Interpretation of   (interaction):\n",
  "The effect of price on market share differs depending\n",
  "on whether the product is discounted. Specifically,\n",
  "the slope for price decreases by 2.3345 units when the\n",
  "product is discounted. (p < 0.001 - highly significant)\n\n",

  "Group-specific fitted equations:\n",
  "Non-discounted (discount = 0):\n",
  "Ŷ = -2.1049 + 1.9755 * price\n\n",
  "Discounted (discount = 1):\n",
  "Ŷ = (-2.1049 + 5.7756) + (1.9755 - 2.3345) * price\n",
  "Ŷ = 3.6707 - 0.3590 * price\n"
)
```

```
## Fitted equation:
##  Ŷ = -2.1049 + 1.9755 * price + 5.7756 * discount
##       - 2.3345 * (price × discount)
##
##  Interpretation of   (interaction):
##  The effect of price on market share differs depending
##  on whether the product is discounted. Specifically,
##  the slope for price decreases by 2.3345 units when the
##  product is discounted. (p < 0.001 - highly significant)
##
##  Group-specific fitted equations:
##  Non-discounted (discount = 0):
##  Ŷ = -2.1049 + 1.9755 * price
##
##  Discounted (discount = 1):
##  Ŷ = (-2.1049 + 5.7756) + (1.9755 - 2.3345) * price
##  Ŷ = 3.6707 - 0.3590 * price
```

Part C

```r
# Fit the full model with all predictors
model_c <- lm(market.share ~ price + discount + promotion + rating, data = data2)

# Show fitted equation
summary(model_c)
```

```
##
## Call:
## lm(formula = market.share ~ price + discount + promotion + rating,
##     data = data2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.45663 -0.19002 -0.05333  0.15586  1.51755
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.7262350  0.4333347   6.291  4.1e-07 ***
## price       -0.2360562  0.1805194  -1.308   0.2000
## discount     0.2886600  0.1142393   2.527   0.0165 *
## promotion    0.1482332  0.1184461   1.251   0.2196
## rating       0.0006831  0.0003087   2.213   0.0340 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3401 on 33 degrees of freedom
## Multiple R-squared:  0.3047, Adjusted R-squared:  0.2204
## F-statistic: 3.615 on 4 and 33 DF,  p-value: 0.015
```

```r
cat(
  "Fitted equation:\n",
  "Ŷ = 2.7262 - 0.2361 * price + 0.2887 * discount\n",
  "     + 0.1482 * promotion + 0.0006831 * rating\n"
)
```

```
## Fitted equation:
##  Ŷ = 2.7262 - 0.2361 * price + 0.2887 * discount
##       + 0.1482 * promotion + 0.0006831 * rating
```

Part D

```r
# Load leaps package
library(leaps)
```

```
## Warning: package 'leaps' was built under R version 4.4.1
```

```r
# Run subset selection
subset_model <- regsubsets(market.share ~ price + discount + promotion + rating,
                           data = data2, nbest = 3)

# Display summary
subset_summary <- summary(subset_model)

# Show results
print(data.frame(
  Variables = apply(subset_summary$which, 1, function(x) paste(names(x)[x][-1], collapse = ", ")),
  R2 = round(subset_summary$rsq, 4),
  Adj_R2 = round(subset_summary$adjr2, 4),
  Cp = round(subset_summary$cp, 4)
))
```

```
##                               Variables     R2 Adj_R2     Cp
## 1                              discount 0.1215 0.0971 7.6943
## 2                                rating 0.0973 0.0722 8.8421
## 3                             promotion 0.0744 0.0487 9.9277
## 4                      discount, rating 0.2470 0.2039 3.7385
## 5                   discount, promotion 0.1732 0.1260 7.2398
## 6                     promotion, rating 0.1412 0.0922 8.7569
## 7               price, discount, rating 0.2717 0.2074 4.5662
## 8           discount, promotion, rating 0.2687 0.2041 4.7099
## 9            price, discount, promotion 0.2015 0.1311 7.8952
## 10 price, discount, promotion, rating 0.3047 0.2204 5.0000
```

```r
cat(
  "Model selection summary:\n\n",

  "Best model by Adjusted R-squared:\n",
  "  Variables: price, discount, rating\n",
  "  Adjusted R² = 0.2074\n\n",

  "Best model by Mallows' Cp:\n",
  "  Variables: discount, rating\n",
  "  Cp = 3.7385 (closest to p + 1 = 3)\n\n",

  "Conclusion:\n",
  "  The model with discount and rating has the most favorable Cp,\n",
  "  indicating minimal bias and good fit.\n",
  "  However, the model with price, discount, and rating has the highest\n",
  "  adjusted R² and may explain more variance.\n"
)
```

```
## Model selection summary:
##
##  Best model by Adjusted R-squared:
##    Variables: price, discount, rating
##    Adjusted R² = 0.2074
##
##  Best model by Mallows' Cp:
##    Variables: discount, rating
```

```
##    Cp = 3.7385 (closest to p + 1 = 3)
##
##  Conclusion:
##    The model with discount and rating has the most favorable Cp,
##    indicating minimal bias and good fit.
##    However, the model with price, discount, and rating has the highest
##    adjusted R² and may explain more variance.
```

Part E

```r
cat(
  "Total models checked in subset selection:\n",
  "For 4 predictors, all possible non-empty subsets = 15\n"
)
```

```
## Total models checked in subset selection:
##  For 4 predictors, all possible non-empty subsets = 15
```

Part F

```r
model_f <- lm(market.share ~ price + discount + rating, data = data2)
summary(model_f)
```

```
##
## Call:
## lm(formula = market.share ~ price + discount + rating, data = data2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.49562 -0.19618 -0.02018  0.10332  1.53481
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.6590792  0.4335651   6.133 5.79e-07 ***
## price       -0.1916380  0.1784635  -1.074   0.2905
## discount     0.3104420  0.1138419   2.727   0.0100 *
## rating       0.0007680  0.0003037   2.529   0.0162 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3429 on 34 degrees of freedom
## Multiple R-squared:  0.2717, Adjusted R-squared:  0.2074
## F-statistic: 4.228 on 3 and 34 DF,  p-value: 0.01211
```

```r
anova(model_f)
```

```
## Analysis of Variance Table
##
## Response: market.share
```

```
##           Df Sum Sq Mean Sq F value  Pr(>F)
## price      1 0.0432 0.04325  0.3679 0.54821
## discount   1 0.6959 0.69594  5.9194 0.02039 *
## rating     1 0.7519 0.75191  6.3955 0.01625 *
## Residuals 34 3.9974 0.11757
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Manually confirm stats for: model_f = price + discount + rating
model_f <- lm(market.share ~ price + discount + rating, data = data2)
model_full <- lm(market.share ~ price + discount + promotion + rating, data = data2)

# Sample size and number of predictors
n <- nrow(data2)
p <- 3   # number of predictors in model_f
p_full <- 4

# Extract RSS (Residual Sum of Squares)
rss <- sum(resid(model_f)^2)

# Total Sum of Squares (SST)
sst <- sum((data2$market.share - mean(data2$market.share))^2)

# R-squared
r_squared <- 1 - rss / sst

# Adjusted R-squared
adj_r_squared <- 1 - (rss / (n - p - 1)) / (sst / (n - 1))

# Get MSE from full model
mse_full <- summary(model_full)$sigma^2  # sigma is residual std error

# Mallows' Cp
cp <- rss / mse_full - (n - 2 * p)

# Show results
cat("Manual calculations:\n")
```

```
## Manual calculations:
```

```r
cat("R-squared:         ", round(r_squared, 4), "\n")
```

```
## R-squared:          0.2717
```

```r
cat("Adjusted R-squared:", round(adj_r_squared, 4), "\n")
```

```
## Adjusted R-squared: 0.2074
```

```r
cat("Mallows' Cp:       ", round(cp, 4), "\n")
```

```
## Mallows' Cp:        2.5662
```