Sample MLR Problem With Soln

ECO204@EWU, Summer 2025

Faculty : STH, TA: Habiba Afroz

2025-08-26

Contents

Package Installations

Install Following packages with the command `install.packages("package_name")` if not already installed, and then load the package using `library(package_name)`.

```
library(readxl)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
## The following objects are masked from 'package:stats':
##
##     filter, lag
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(ggplot2)
library(knitr)
library(broom)
library(stargazer)
```

```
##
## Please cite as:
##  Hlavac, Marek (2022). stargazer: Well-Formatted Regression and Summary Statistics Tables.
##  R package version 5.2.3. https://CRAN.R-project.org/package=stargazer
```

```r
options(scipen = 100) # to avoid scientific notation
```

Problem Setup

The owner of Showtime Movie Theaters, Inc., would like to predict weekly gross revenue as a function of advertising expenditures. We have historical data for a sample of eight weeks with the following variables:

- `revenue`: gross revenue ($1000)
- `tv`: television advertising ($1000)
- `newspaper`: newspaper advertising ($1000)
- `magazines`: magazine advertising ($1000)
- `leaflets`: leaflets advertising ($1000)
- `age-category`: age category (1 = teen, 2 = middle, 3 = aged)

```r
# Load the new dataset I gave the full path, you can set the directory first and then just the name
# of the file is enough ...

showtime_data <- read_excel("/home/tanvir/Documents/ownCloud/Git_Repos/EWU_repos/6_Summer_2025/ECO204/e

# Display the showtime_data
kable(showtime_data, caption = "Showtime Movie Theater Data Set")
```

```
## Warning in attr(x, "align"): 'xfun::attr()' is deprecated.
## Use 'xfun::attr2()' instead.
## See help("Deprecated")
```

```
## Warning in attr(x, "format"): 'xfun::attr()' is deprecated.
## Use 'xfun::attr2()' instead.
## See help("Deprecated")
```

Table 1: Showtime Movie Theater Data Set

revenue	tv	newspaper	magazines	leaflets	age_category
23.43	3.93	1.56	0.47	1.32	1
21.28	2.90	0.86	1.54	1.43	1
22.76	3.06	1.70	1.94	0.56	1
22.32	2.99	1.86	0.59	1.37	1
22.60	4.12	2.15	0.38	0.87	1
17.88	3.02	0.79	1.58	1.03	1
24.20	4.59	0.74	1.50	1.23	1
14.43	1.51	0.94	0.80	1.21	1
26.37	3.58	2.98	0.67	1.07	1
21.74	3.28	2.76	0.53	0.48	1
19.93	2.74	1.51	0.39	1.00	2

revenue	tv	newspaper	magazines	leaflets	age_category
33.32	4.81	2.83	1.99	1.45	2
25.28	3.62	2.42	1.30	0.25	2
17.38	1.10	2.54	0.97	1.23	2
23.78	2.87	2.99	0.32	0.71	2
30.91	4.43	2.53	1.09	1.23	2
24.26	2.46	2.37	1.67	1.31	2
20.02	2.46	0.83	1.29	0.71	2
26.16	4.03	1.63	1.70	1.00	2
18.80	1.43	1.85	1.26	1.40	2
19.58	2.38	0.74	1.82	1.30	3
30.23	4.26	2.34	0.59	0.29	3
17.91	1.16	0.95	1.87	1.13	3
21.39	3.86	0.83	0.35	1.23	3
21.19	2.35	1.80	0.69	0.38	3
26.64	4.67	1.40	0.83	0.27	3
26.51	4.11	1.38	0.85	0.44	3
20.05	1.80	1.98	0.55	1.10	3
28.67	4.43	2.41	0.31	0.77	3
22.90	2.04	2.00	1.13	0.33	3

```r
# Convert age-category to factor
showtime_data$age_category_factor <- factor(showtime_data$age_category,
                                            levels = c(1, 2, 3),
                                            labels = c("Teen", "Middle", "Aged"))
```

Multiple Linear Regression Model Estimation

Develop an estimated regression equation to predict weekly gross revenue with all other variables (except age-category) as the independent variables.

```r
# Multiple linear regression excluding age-category
model_full<- lm(revenue ~ tv + newspaper + magazines + leaflets, data = showtime_data)
summary(model_full)
```

```
##
## Call:
## lm(formula = revenue ~ tv + newspaper + magazines + leaflets,
##     data = showtime_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5977 -0.8777  0.1762  0.9504  2.9054
##
## Coefficients:
##             Estimate Std. Error t value       Pr(>|t|)
## (Intercept)   7.5221     1.7360   4.333        0.00021 ***
## tv            2.9633     0.2977   9.953 0.000000000353 ***
## newspaper     2.6515     0.4637   5.718 0.000005898790 ***
## magazines     1.8699     0.6176   3.028        0.00565 **
## leaflets     -0.4441     0.8419  -0.527        0.60256
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 1.714 on 25 degrees of freedom
## Multiple R-squared:  0.8624, Adjusted R-squared:  0.8404
## F-statistic: 39.16 on 4 and 25 DF,  p-value: 0.0000000002018
```

A slightly better output with the stargazer package

```
# you can play around with different options
stargazer(model_full, type = "text",  ci = TRUE, ci.level = 0.95)
```

```
##
## ===============================================
##                         Dependent variable:
##                     ---------------------------
##                                revenue
## -----------------------------------------------
## tv                            2.963***
##                             (2.380, 3.547)
##
## newspaper                     2.651***
##                             (1.743, 3.560)
##
## magazines                     1.870***
##                             (0.659, 3.080)
##
## leaflets                      -0.444
##                             (-2.094, 1.206)
##
## Constant                      7.522***
##                             (4.120, 10.925)
##
## -----------------------------------------------
## Observations                    30
## R2                             0.862
## Adjusted R2                    0.840
## Residual Std. Error       1.714 (df = 25)
## F Statistic           39.162*** (df = 4; 25)
## ===============================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```

ANOVA Table and Sum of Squares

Write the ANOVA table and calculate SST, SSR, SSE, MSR, and MSE. Here to get a proper ANOVA table we need to run a NUll model and Full model and use the general restricted-unrestricted way to get the table

```
# fit null model
model_null <- lm(revenue ~ 1, data = showtime_data)  # Null model with only intercept

# we already have the complete model

# anova function will compare between two models
anova(model_null, model_full)
```

```
## Analysis of Variance Table
##
## Model 1: revenue ~ 1
```

```
## Model 2: revenue ~ tv + newspaper + magazines + leaflets
##   Res.Df    RSS Df Sum of Sq      F          Pr(>F)
## 1     29 533.90
## 2     25  73.48  4    460.42 39.163 0.0000000002018 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Partial F-test (Restricted F test) for Magazines and Leaflets

Test whether both magazine and leaflets advertising should be dropped.

```
# Reduced model without magazines and leaflets
restricted_model <- lm(revenue ~ tv + newspaper, data = showtime_data)

# Partial F-test
anova_comparison <- anova(restricted_model, model_full)
print(anova_comparison)
```

```
## Analysis of Variance Table
##
## Model 1: revenue ~ tv + newspaper
## Model 2: revenue ~ tv + newspaper + magazines + leaflets
##   Res.Df     RSS Df Sum of Sq      F  Pr(>F)
## 1     27 100.427
## 2     25  73.479  2    26.947 4.5841 0.02014 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Prediction

Question: Predict gross revenue when TV = \$3500, Newspaper = \$2300, Magazines = \$1000, Leaflets = \$500.

```
# Create new data for prediction
new_data <- data.frame(tv = 3.5, newspaper = 2.3, magazines = 1.0, leaflets = 0.5)

# Make prediction
predicted_revenue <- predict(model_full, new_data)
```

Dummy Variable Regression Function

What is the population regression function with age-category dummy variables?

Answer:

$$Y_i = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + \epsilon_i$$

Where: - $D_{1i} = 1$ if i^{th} restaurant has majority Adults, 0 otherwise - $D_{2i} = 1$ if i^{th} restaurant has majority Aged, 0 otherwise - Teen is the reference category is the reference category

Dummy Variable Regression Estimation

In R we don't need to create dummy variables, we can directly run the regression using the factor variable. Run the regression and interpret the coefficients. Note that it automatically sets the first category (Teen) as the reference category. However you need to be careful what you want as a reference

```r
dummy_model <- lm(revenue ~ age_category_factor, data = showtime_data)
summary(dummy_model)
```

```
##
## Call:
## lm(formula = revenue ~ age_category_factor, data = showtime_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.2710 -3.7300  0.1575  2.4183  9.3360
##
## Coefficients:
##                           Estimate Std. Error t value            Pr(>|t|)
## (Intercept)                 21.701      1.367  15.869 0.00000000000000327 ***
## age_category_factorMiddle    2.283      1.934   1.181               0.248
## age_category_factorAged      1.806      1.934   0.934               0.359
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.324 on 27 degrees of freedom
## Multiple R-squared:  0.05433,    Adjusted R-squared:  -0.01572
## F-statistic: 0.7755 on 2 and 27 DF,  p-value: 0.4705
```

```r
# if you want to specify the base category, you can relvel the factor and run the regression again

# showtime_data$age_category_factor <- relevel(showtime_data$age_category_factor, ref = "Middle")
# dummy_model <- lm(revenue ~ age_category_factor, data = showtime_data)
```

Finally using the `group_by` and `summarise` functions from dplyr to show that the regression results are the same as the group means

Show that we don't need regression - just group means

```r
group_means <- showtime_data %>%
  group_by(age_category_factor) %>%
  summarise(mean_revenue = mean(revenue))

group_means
```

```
## # A tibble: 3 x 2
##   age_category_factor mean_revenue
##   <fct>                      <dbl>
## 1 Teen                        21.7
## 2 Middle                      24.0
## 3 Aged                        23.5
```

Interaction Model

Conduct multiple linear regression with newspaper advertising, age-category, and their interaction terms.

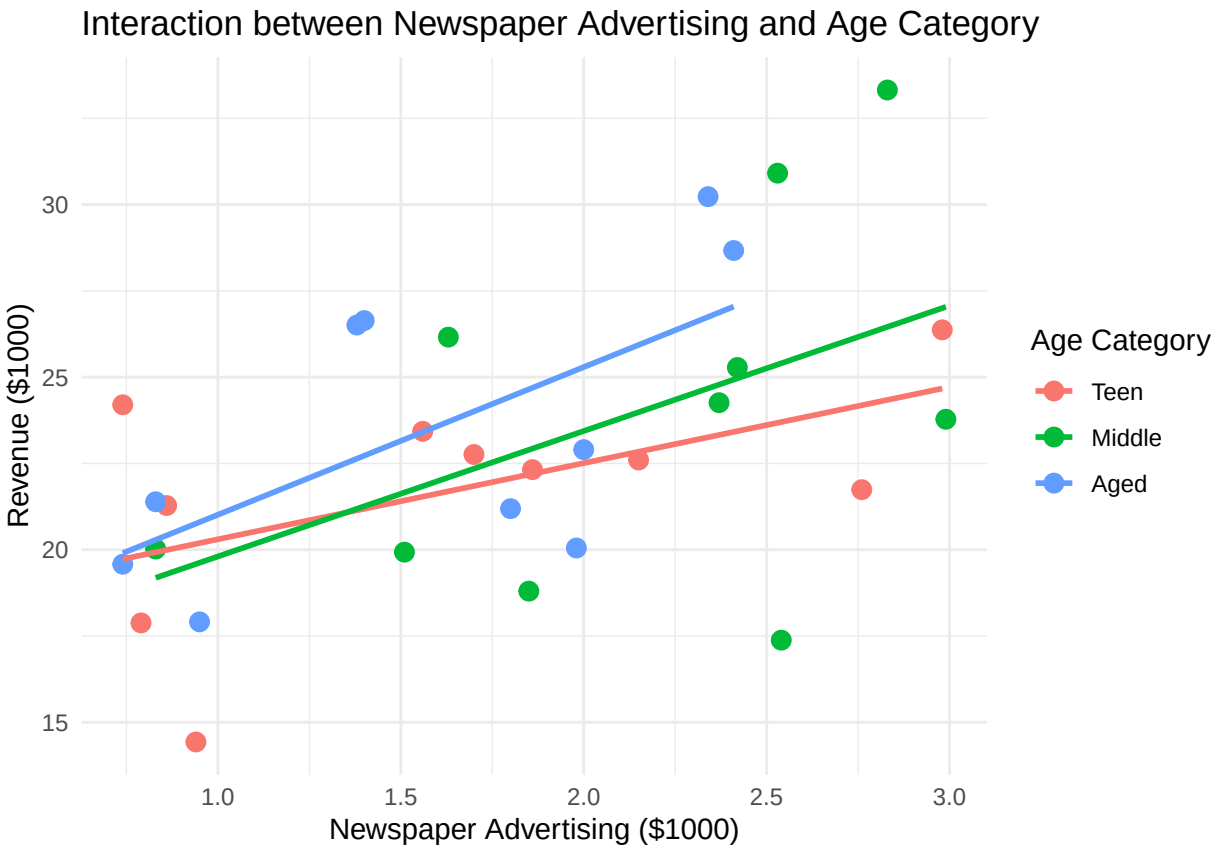
```r
#Interaction model with newspaper and age category
interaction_model <- lm(revenue ~ newspaper*age_category_factor, data = showtime_data)
summary(interaction_model)
```

```
##
## Call:
## lm(formula = revenue ~ newspaper * age_category_factor, data = showtime_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.0214 -2.4318  0.2175  1.8440  6.8646
##
## Coefficients:
##                                    Estimate Std. Error t value   Pr(>|t|)
## (Intercept)                          18.096      2.858   6.333 0.00000151 ***
## newspaper                             2.206      1.580   1.396      0.175
## age_category_factorMiddle            -1.926      5.154  -0.374      0.712
## age_category_factorAged              -1.360      4.554  -0.299      0.768
## newspaper:age_category_factorMiddle   1.428      2.480   0.576      0.570
## newspaper:age_category_factorAged     2.071      2.630   0.787      0.439
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.871 on 24 degrees of freedom
## Multiple R-squared:  0.3266, Adjusted R-squared:  0.1863
## F-statistic: 2.328 on 5 and 24 DF,  p-value: 0.07387
```

Optional

```
#Create a visualization of the interaction effect
ggplot(showtime_data, aes(x = newspaper, y = revenue, color = age_category_factor)) +
  geom_point(size = 3) +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Interaction between Newspaper Advertising and Age Category",
       x = "Newspaper Advertising ($1000)",
       y = "Revenue ($1000)",
       color = "Age Category") +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



The analysis reveals that:

1. **Multiple regression model** shows varying significance levels for different advertising types
2. **Overall model** is significant, indicating advertising expenditures do predict revenue

3. **Individual variables** have different levels of significance
4. **Interaction analysis** suggests newspaper advertising effectiveness varies by age group
5. **Different slopes** for each age category indicate targeted advertising strategies may be beneficial

This comprehensive analysis provides insights for optimizing advertising spend across different media and age demographics for Showtime Movie Theaters.