example_questions_key

October 8, 2025

1 Example Questions: Answer Key

---

1. **Linear Regression**

SNo.	X1	X2	X3	Y
1	2.258	0.917	39	150.74
2	1.315	0.595	22	126.33
3	4.996	0.843	37	151.13
4	4.647	0.474	17	124.04
5	1.653	0.123	7	99.04
6	4.939	0.995	39	158.37
7	2.762	0.026	10	98.63
8	3.214	0.180	37	121.06
9	3.229	0.222	8	106.24
10	0.464	0.345	27	117.56
11	2.567	0.463	6	113.86
12	4.755	0.700	99	180.27
13	4.770	0.811	3	130.01
14	4.753	0.663	12	129.07
15	2.150	0.001	67	129.14
16	1.162	0.171	9	101.33
17	4.313	0.351	49	136.73
18	1.676	0.157	42	120.46
19	1.056	0.449	62	142.75
20	3.412	0.855	57	160.40
21	4.385	0.644	26	135.69
22	3.261	0.527	58	147.32
23	0.046	0.447	25	119.88
24	0.817	0.782	73	162.26
25	1.650	0.512	28	126.90
26	3.616	0.201	91	153.45
27	3.740	0.284	83	152.45
28	3.595	0.748	81	170.04
29	1.616	0.440	13	115.34
30	2.856	0.480	39	133.81

For the above-detailed dataset with exogenous variables X_1, X_2, X_3 and endogenous variable Y, develop the following linear regression model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$ in R, and answer the questions below

```
{tip} Save the dataset in an Excel/CSV file and import it into R using functions
like `read.csv()` or `readxl::read_excel()`.
```

a. Comment on the estimated value and significance of each coefficient - $\beta_0, \beta_1, \beta_2, \beta_3$
b. Fill in the above table, computing fitted values and corresponding errors
c. Compute the following model statistics

- Sum Squares Total
- Sum Squares Regression
- Sum Squares Error
- R-squared
- Adjusted R-squared

d. Perform ex-post analysis (compute correlation between exogenous variables; draw residuals plot) to comment upon the validity of the model.

[2]:
```r
# --- TO BE PERFORMED IN R ---
# Packages
library(dplyr)
library(ggplot2)
options(repr.plot.width = 12, repr.plot.height = 8)
# Dataset
data <- data.frame(
  X1 = c(2.258, 1.315, 4.996, 4.647, 1.653, 4.939, 2.762, 3.214, 3.229, 0.464,
         2.567, 4.755, 4.770, 4.753, 2.150, 1.162, 4.313, 1.676, 1.056, 3.412,
         4.385, 3.261, 0.046, 0.817, 1.650, 3.616, 3.740, 3.595, 1.616, 2.856),
  X2 = c(0.917, 0.595, 0.843, 0.474, 0.123, 0.995, 0.026, 0.180, 0.222, 0.345,
         0.463, 0.700, 0.811, 0.663, 0.001, 0.171, 0.351, 0.157, 0.449, 0.855,
         0.644, 0.527, 0.447, 0.782, 0.512, 0.201, 0.284, 0.748, 0.440, 0.480),
  X3 = c(39, 22, 37, 17, 7, 39, 10, 37, 8, 27,
         6, 99, 3, 12, 67, 9, 49, 42, 62, 57,
         26, 58, 25, 73, 28, 91, 83, 81, 13, 39),
  Y  = c(150.74, 126.33, 151.13, 124.04, 99.04, 158.37, 98.63, 121.06, 106.24,
 ↪117.56,
         113.86, 180.27, 130.01, 129.07, 129.14, 101.33, 136.73, 120.46, 142.
 ↪75, 160.40,
         135.69, 147.32, 119.88, 162.26, 126.90, 153.45, 152.45, 170.04, 115.
 ↪34, 133.81)
)
# Model
model <- lm(Y ~ X1 + X2 + X3, data = data)
summary(model)
```

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

Residuals:
      Min        1Q    Median        3Q       Max
-0.068356 -0.004011  0.002531  0.008781  0.022472

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 8.727e+01  8.382e-03 10412.4   <2e-16 ***
X1          1.656e+00  2.304e-03   718.6   <2e-16 ***
X2          4.085e+01  1.231e-02  3318.7   <2e-16 ***
X3          5.710e-01  1.153e-04  4950.7   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01696 on 26 degrees of freedom
Multiple R-squared:      1,	Adjusted R-squared:      1
F-statistic: 1.534e+07 on 3 and 26 DF,  p-value: < 2.2e-16
```

[5]:

```r
# --- TO BE PERFORMED BY HAND ---
# Fitted values and Residuals
data$fitted <- round(fitted(model), 5)
data$resid  <- round(resid(model), 5)
data
# Statistics
sst <- sum((data$Y - mean(data$Y))^2)
ssr <- sum((model$fitted.values - mean(data$Y))^2)
sse <- sum((data$Y - model$fitted.values)^2)
rse <- sqrt(sse / (nrow(data) - 2))
Rsquared <- ssr / sst
AdjRsquared <- 1 - (1 - Rsquared) * (nrow(data) - 1) / (nrow(data) - 2)
cat("SST                :", round(sst, 2), "\n")
cat("SSR                :", round(ssr, 2), "\n")
cat("SSE                :", round(sse, 2), "\n")
cat("RSE                :", round(rse, 2), "\n")
cat("R-squared          :", round(Rsquared, 3), "\n")
cat("Adjusted R-squared:", round(AdjRsquared, 3), "\n")
# Residuals Plot
cat("Cor: X1-X2", round(cor(data$X1, data$X2), 2), "\n")
cat("Cor: X1-X3", round(cor(data$X1, data$X3), 2), "\n")
cat("Cor: X2-X3", round(cor(data$X2, data$X3), 2), "\n")
ggplot(data, aes(x = fitted, y = resid)) +
  geom_point(color = "steelblue", size = 3) +
  labs(
```

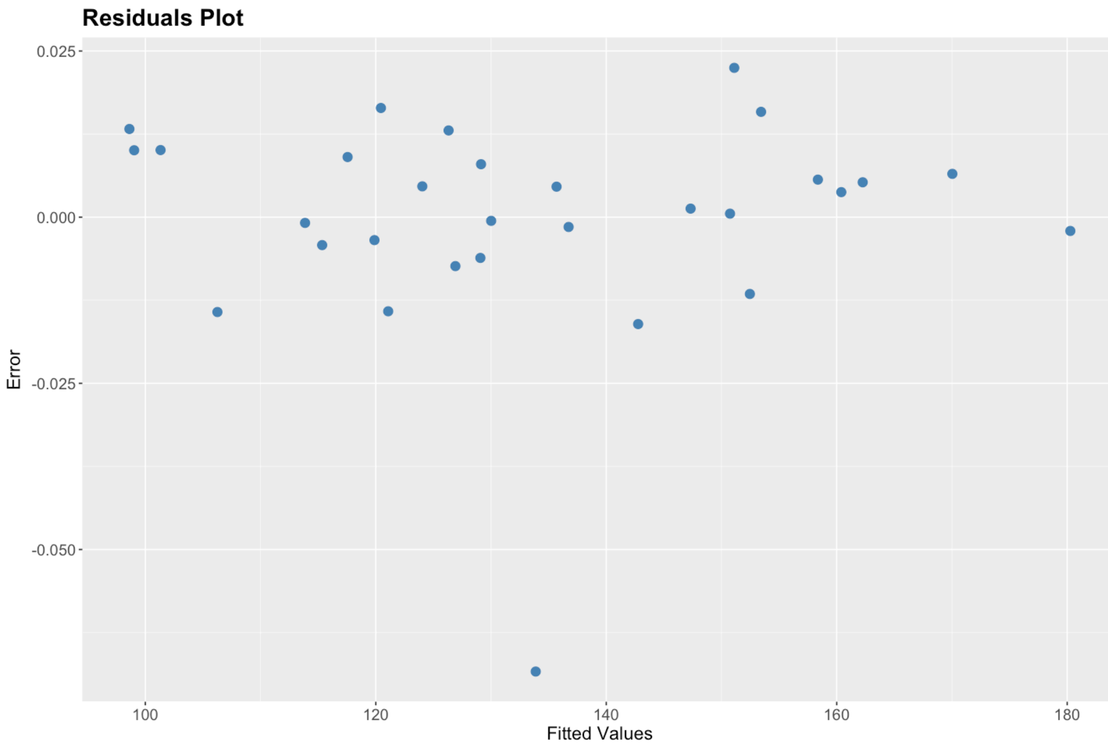
```
    title = "Residuals Plot",
    x = "Fitted Values",
    y = "Error"
  ) +
  theme(
    plot.title = element_text(size = 18, face = "bold"),
    axis.title = element_text(size = 14),
    axis.text = element_text(size = 12)
  )
```

A data.frame: 30 × 6

X1 <dbl>	X2 <dbl>	X3 <dbl>	Y <dbl>	fitted <dbl>	resid <dbl>
2.258	0.917	39	150.74	150.73947	0.00053
1.315	0.595	22	126.33	126.31695	0.01305
4.996	0.843	37	151.13	151.10753	0.02247
4.647	0.474	17	124.04	124.03536	0.00464
1.653	0.123	7	99.04	99.02993	0.01007
4.939	0.995	39	158.37	158.36435	0.00565
2.762	0.026	10	98.63	98.61673	0.01327
3.214	0.180	37	121.06	121.07416	-0.01416
3.229	0.222	8	106.24	106.25427	-0.01427
0.464	0.345	27	117.56	117.55095	0.00905
2.567	0.463	6	113.86	113.86086	-0.00086
4.755	0.700	99	180.27	180.27207	-0.00207
4.770	0.811	3	130.01	130.01055	-0.00055
4.753	0.663	12	129.07	129.07613	-0.00613
2.150	0.001	67	129.14	129.13203	0.00797
1.162	0.171	9	101.33	101.31990	0.01010
4.313	0.351	49	136.73	136.73147	-0.00147
1.676	0.157	42	120.46	120.44357	0.01643
1.056	0.449	62	142.75	142.76608	-0.01608
3.412	0.855	57	160.40	160.39623	0.00377
4.385	0.644	26	135.69	135.68541	0.00459
3.261	0.527	58	147.32	147.31871	0.00129
0.046	0.447	25	119.88	119.88345	-0.00345
0.817	0.782	73	162.26	162.25475	0.00525
1.650	0.512	28	126.90	126.90736	-0.00736
3.616	0.201	91	153.45	153.43415	0.01585
3.740	0.284	83	152.45	152.46155	-0.01155
3.595	0.748	81	170.04	170.03348	0.00652
1.616	0.440	13	115.34	115.34420	-0.00420
2.856	0.480	39	133.81	133.87836	-0.06836

```
SST                  : 13242.6
SSR                  : 13242.6
SSE                  : 0.01
RSE                  : 0.02
R-squared            : 1
```

```
Adjusted R-squared: 1
Cor: X1-X2 0.36
Cor: X1-X3 0.14
Cor: X2-X3 0.12
```



---

2. **Logistic Regression**

For the following dataset with exogenous variables X_1, X_2, X_3 and binary endogenous variable Y, develop the following logistic regression model in R

`{tip} Save the dataset in an Excel/CSV file and import it into R using functions like `read.csv()` or `readxl::read_excel()`.`

$$\log\left(\frac{\hat{P}_{Y=S}}{1-\hat{P}_{Y=S}}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

a. Comment on the estimated value and significance of each coefficient $\beta_0, \beta_1, \beta_2, \beta_3$.

b. Compute estimated probabilities ($\hat{P}_{Y=S}$ and $\hat{P}_{Y=F}$)

c. Compute the following model statistics

- Log-Likelihood for
  - Equally-Likely Model

    - Market Share Model
- McFadden R-squared for the Estimated Model vs.
    - Equally-Likely Model
    - Market Share Model
- Adjusted McFadden R-squared for the Estimated Model vs.
    - Equally-Likely Model
    - Market Share Model

SNo	X_1	X_2	X_3	Y
1	1.8	10	0	F
2	2.1	14	0	F
3	2.3	13	1	F
4	2.5	15	0	F
5	2.7	18	1	S
6	2.9	16	0	F
7	3.0	20	1	S
8	3.1	17	0	F
9	3.2	22	1	S
10	3.3	21	0	F
11	3.4	19	1	S
12	3.5	25	0	S
13	3.6	23	1	S
14	3.7	26	0	S
15	3.8	24	1	S
16	4.0	28	0	S
17	2.2	12	1	F
18	2.6	15	1	S
19	2.8	19	0	F
20	3.0	18	1	S
21	3.2	22	0	S
22	3.4	20	1	S
23	3.6	27	0	S
24	3.8	29	1	S
25	2.4	14	0	F
26	2.7	16	1	F
27	2.9	18	1	S
28	3.1	21	0	S
29	3.5	23	1	S
30	3.9	30	0	S

[3]:
```
# --- TO BE PERFORMED IN R ---
# Packages
library(dplyr)
library(tidyr)
```

```r
library(mlogit)
# Dataset
data <- data.frame(
  ID = 1:30,
  X1 = c(1.8, 2.1, 2.3, 2.5, 2.7, 2.9, 3.0, 3.1, 3.2, 3.3,
         3.4, 3.5, 3.6, 3.7, 3.8, 4.0, 2.2, 2.6, 2.8, 3.0,
         3.2, 3.4, 3.6, 3.8, 2.4, 2.7, 2.9, 3.1, 3.5, 3.9),
  X2 = c(10, 14, 13, 15, 18, 16, 20, 17, 22, 21,
         19, 25, 23, 26, 24, 28, 12, 15, 19, 18,
         22, 20, 27, 29, 14, 16, 18, 21, 23, 30),
  X3 = c(0, 0, 1, 0, 1, 0, 1, 0, 1, 0,
         1, 0, 1, 0, 1, 0, 1, 1, 0, 1,
         0, 1, 0, 1, 0, 1, 1, 0, 1, 0),
  Y  = c("F", "F", "F", "F", "S", "F", "S", "F", "S", "F",
         "S", "S", "S", "S", "S", "S", "F", "S", "F", "S",
         "S", "S", "S", "S", "F", "F", "S", "S", "S", "S")
)
data$S <- ifelse(data$Y == "S", 1, 0)
data$F <- ifelse(data$Y == "F", 1, 0)
# Long-Form
long_data <- data %>%
  select(ID, X1, X2, X3, S, F) %>%
  pivot_longer(
    cols = c(S, F),
    names_to = "alt",
    values_to = "bin"
  ) %>%
  mutate(bin = bin == 1)
# Model Data
model_data <- mlogit.data(
  long_data,
  choice = "bin",
  shape = "long",
  chid.var = "ID",
  alt.var = "alt"
)
# Model
model <- mlogit(bin ~ 1 | X1 + X2 + X3, data = model_data)
summary(model)
```

```
Call:
mlogit(formula = bin ~ 1 | X1 + X2 + X3, data = model_data, method = "nr")

Frequencies of alternatives:choice
      F       S
0.36667 0.63333
```

```
nr method
9 iterations, 0h:0m:0s
g'(-H)^-1g = 5.18E-05
successive function values within tolerance limits

Coefficients :
              Estimate Std. Error z-value Pr(>|z|)
(Intercept):S -28.7314    16.7510 -1.7152  0.08631 .
X1:S          -10.4524    11.4175 -0.9155  0.35995
X2:S            2.9522     2.1150  1.3959  0.16275
X3:S           10.8344     6.6282  1.6346  0.10213
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Log-Likelihood: -3.5054
McFadden R^2:  0.82219
Likelihood ratio test : chisq = 32.419 (p.value = 4.2711e-07)
```

[ ]:
```
# --- TO BE PERFORMED BY HAND ---
# Fitted Probabilities
data$S <- ifelse(data$Y == "S", 1, 0)
data$F <- ifelse(data$Y == "F", 1, 0)
data$`P(Y=S)` <- fitted.values(model) * data$S + (1 - fitted.values(model)) * 
  data$F
data$`P(Y=F)` <- 1 - data$`P(Y=S)`
data
# Statistics
LL_EL  <- sum(data$S * log(1/2) + data$F * log(1/2))
LL_MS  <- sum(data$S * log(0.633) + data$F * log(0.367))
LL_MNL <- sum(data$S * log(data$`P(Y=S)`) + data$F * log(data$`P(Y=F)`))
R2_EL  <- 1 - (LL_MNL / LL_EL)
R2_MS  <- 1 - (LL_MNL / LL_MS)
AR2_EL <- 1 - ((LL_MNL - 4) / LL_EL)
AR2_MS <- 1 - ((LL_MNL - 4) / LL_MS)
cat("\n--- Log-likelihoods ---\n")
cat(sprintf("EL  : %0.3f\n", LL_EL))
cat(sprintf("MS  : %0.3f\n", LL_MS))
cat(sprintf("MNL : %0.3f\n", LL_MNL))
cat("\n--- McFadden R^2 ---\n")
cat(sprintf("R2 vs EL : %0.4f\n", R2_EL))
cat(sprintf("R2 vs MS : %0.4f\n", R2_MS))
cat("\n--- Adjusted McFadden R^2 ---\n")
cat(sprintf("Adj R2 vs EL : %0.4f (K = %d)\n", AR2_EL, 4))
cat(sprintf("Adj R2 vs MS : %0.4f (K = %d)\n", AR2_MS, 4))
```

```
--- Log-likelihoods ---
EL  : -20.794
```

```
MS  : -19.715
MNL : -3.505

--- McFadden R^2 ---
R2 vs EL : 0.8314
R2 vs MS : 0.8222

--- Adjusted McFadden R^2 ---
Adj R2 vs EL : 0.6391 (K = 4)
Adj R2 vs MS : 0.6193 (K = 4)
```

A data.frame: 30×9

ID	X1	X2	X3	Y	S	F	P(Y=S)	P(Y=F)
<int>	<dbl>	<dbl>	<dbl>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>
1	1.8	10	0	F	0	1	0.00000	1.00000
2	2.1	14	0	F	0	1	0.00009	0.99991
3	2.3	13	1	F	0	1	0.02768	0.97232
4	2.5	15	0	F	0	1	0.00003	0.99997
5	2.7	18	1	S	1	0	0.99911	0.00089
6	2.9	16	0	F	0	1	0.00001	0.99999
7	3.0	20	1	S	1	0	0.99994	0.00006
8	3.1	17	0	F	0	1	0.00002	0.99998
9	3.2	22	1	S	1	0	1.00000	0.00000
10	3.3	21	0	F	0	1	0.22654	0.77346
11	3.4	19	1	S	1	0	0.93443	0.06557
12	3.5	25	0	S	1	0	0.99979	0.00021
13	3.6	23	1	S	1	0	1.00000	0.00000
14	3.7	26	0	S	1	0	0.99991	0.00009
15	3.8	24	1	S	1	0	1.00000	0.00000
16	4.0	28	0	S	1	0	0.99999	0.00001
17	2.2	12	1	F	0	1	0.00421	0.99579
18	2.6	15	1	S	1	0	0.31215	0.68785
19	2.8	19	0	F	0	1	0.12941	0.87059
20	3.0	18	1	S	1	0	0.97988	0.02012
21	3.2	22	0	S	1	0	0.94101	0.05899
22	3.4	20	1	S	1	0	0.99635	0.00365
23	3.6	27	0	S	1	0	1.00000	0.00000
24	3.8	29	1	S	1	0	1.00000	0.00000
25	2.4	14	0	F	0	1	0.00000	1.00000
26	2.7	16	1	F	0	1	0.75342	0.24658
27	2.9	18	1	S	1	0	0.99283	0.00717
28	3.1	21	0	S	1	0	0.70320	0.29680
29	3.5	23	1	S	1	0	1.00000	0.00000
30	3.9	30	0	S	1	0	1.00000	0.00000