Ex 4.38

For independent observations $y_1, ..., y_n$ having the geometric distribution $f(y) = (1-\pi)^{y-1}\pi$, with $y = 1, 2, 3, ...$:

a. Find a sufficient statistic for π.

b. Derive the ML estimator of π.

Solution

a. **Find a sufficient statistic for π**

A statistic $T(y)$ is **sufficient** for θ if it contains all the information needed to compute any estimate of the parameter. To find a sufficient statistic for π based on the geometric distribution $f(y) = (1-\pi)^{y-1}\pi$ for $y = 1, 2, 3, ...$ we proceed as follows:

1. **Calculate the likelyhood**:

   Given independent observations $y_1, y_2, \ldots, y_n$ from the geometric distribution:

   $$f(y; \pi) = (1-\pi)^{-1}\pi, \quad y = 1, 2, 3, ...$$

   The joint probability mass function (likelihood function) for all observations is the product of individual probabilities:

   $$L(\pi) = \prod_{i=1}^{n} \left[(1-\pi)^{y_i - 1}\pi\right]$$

   With some algebra we can separate the terms involving π:

   $$L(\pi; y) = \pi^n \prod_{i=1}^{n} (1-\pi)^{y_i - 1} = \pi^n (1-\pi)^{\sum_{i=1}^{n}(y_i - 1)} = \pi^n (1-p)^{T(y)-n}, \qquad \text{where} \;\; T(y) = \sum_{i=1}^{n} y_i$$

2. **Factorization Criterion:**

   The **Neyman–Fisher factorization theorem** states that a statistic $T(y)$ is sufficient for parameter π if the likelihood can be factorized into:

   $$L(y; \pi) = h(y) \cdot g(T(y); \pi)$$

   where:

   - $g(T(y); \pi)$ depends on the data only through $T(y)$, and the parameter π,
   - $h(y)$ does not depend on π

   From our likelihood function:

   $$L(\pi; y) = \underbrace{1}_{h(y)} \cdot \underbrace{\pi^n (1-\pi)^{T(y)-n}}_{g(T(y);\pi)}$$

   Here, $g(T(y), \pi)$ depends on the data only through $T(y) = \sum y_i$ and π, and $h(y) = 1$ does not depend on π.

By the factorization theorem, a sufficient statistic for π is:

$$\boxed{T(y) = \sum_{i=1}^{n} y_y}$$

b. **Derive the ML estimator of π**

To derive the Maximum Likelihood Estimator (MLE) of π, we firstly have to calculate the Log-Likelihood function:

$$\ell(\pi) = \ln L(\pi; y) = n \ln \pi + (T(y) - n) \ln(1 - \pi)$$

Now we can find the maximum setting its derivative to zero:

$$\frac{d\ell}{d\pi} = \frac{n}{\pi} - \frac{T(y) - n}{1 - \pi} = 0 \Rightarrow \quad n(1-\pi) - (T(y)-n)\pi = 0 \quad \Rightarrow \quad n - n\pi - T(y)\pi + n\pi = n - T(y)\pi = 0 \quad \Rightarrow \quad T(y)\pi = n \Rightarrow$$

The maximum likelihood estimator of π is:

$$\boxed{\hat{\pi} = \frac{n}{\sum_{i=1}^{n} y_i}}$$

Ex 6.12

For the `UN` data file at the book's website (see Exercise 1.24), construct a multiple regression model predicting `Internet` using all the other variables. Use the concept of multicollinearity to explain why adjusted R^2 is not dramatically greater than when `GDP` is the sole predictor. Compare the estimated `GDP` effect in the bivariate model and the multiple regression model and explain why it is so much weaker in the multiple regression model.

Solution First, let's load the data into a variable and inspect its structure:

```r
url <- "https://stat4ds.rwth-aachen.de/data/UN.dat"
UN <- read.table(url, header = TRUE)

summary(UN)
```

```
##     Nation               GDP             HDI              GII        
##  Length:42          Min.   : 4.40   Min.   :0.5000   Min.   :0.0300  
##  Class :character   1st Qu.:13.18   1st Qu.:0.7400   1st Qu.:0.0850  
##  Mode  :character   Median :27.45   Median :0.8600   Median :0.1850  
##                     Mean   :26.83   Mean   :0.8045   Mean   :0.2414  
##                     3rd Qu.:40.33   3rd Qu.:0.8975   3rd Qu.:0.3875  
##                     Max.   :62.90   Max.   :0.9400   Max.   :0.5600  
##    Fertility          CO2             Homicide           Prison     
##  Min.   :1.200   Min.   : 0.500   Min.   : 0.300   Min.   : 30.0  
##  1st Qu.:1.700   1st Qu.: 4.025   1st Qu.: 0.900   1st Qu.: 82.0  
##  Median :1.900   Median : 6.900   Median : 1.550   Median :119.5  
##  Mean   :2.038   Mean   : 6.695   Mean   : 4.257   Mean   :153.9  
##  3rd Qu.:2.200   3rd Qu.: 8.975   3rd Qu.: 3.650   3rd Qu.:188.8  
##  Max.   :6.000   Max.   :17.000   Max.   :30.900   Max.   :716.0  
##     Internet    
##  Min.   :11.00  
##  1st Qu.:46.00  
```

```
##   Median :67.00
##   Mean   :63.86
##   3rd Qu.:84.00
##   Max.   :95.00
```

Next, we fit a **multiple regression model** using `Internet` as the response variable and all other variables as predictors:

```r
fit1 <- lm(Internet ~ GDP + HDI + GII + Fertility + CO2 + Homicide + Prison, data = UN)

summary(fit1)
```

```
##
## Call:
## lm(formula = Internet ~ GDP + HDI + GII + Fertility + CO2 + Homicide +
##     Prison, data = UN)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.819  -5.827  -2.182   7.166  26.354
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  11.158310  38.773097   0.288  0.77526
## GDP           0.440903   0.290680   1.517  0.13856
## HDI          55.851013  46.652218   1.197  0.23952
## GII         -72.428931  25.323061  -2.860  0.00719 **
## Fertility     4.092148   3.065379   1.335  0.19076
## CO2           0.310113   0.654899   0.474  0.63886
## Homicide      0.377324   0.299751   1.259  0.21668
## Prison        0.009091   0.018347   0.495  0.62344
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.53 on 34 degrees of freedom
## Multiple R-squared:  0.8477, Adjusted R-squared:  0.8164
## F-statistic: 27.04 on 7 and 34 DF,  p-value: 3.947e-12
```

For comparison, we fit a **bivariate regression model** using only `GDP` as the predictor for `Internet`:

```r
fit2 <- lm(Internet ~ GDP, data = UN)

summary(fit2)
```

```
##
## Call:
## lm(formula = Internet ~ GDP, data = UN)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -27.130  -5.729   2.124  10.092  20.165
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  26.1341     3.7490   6.971 2.06e-08 ***
## GDP           1.4060     0.1217  11.555 2.55e-14 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.95 on 40 degrees of freedom
## Multiple R-squared:  0.7695, Adjusted R-squared:  0.7637
## F-statistic: 133.5 on 1 and 40 DF,  p-value: 2.549e-14
```

From the summaries, we observe the following:

- Adjusted R^2 for the **multiple regression model**: 0.8164
- Adjusted R^2 for the **bivariate regression model**: 0.7637

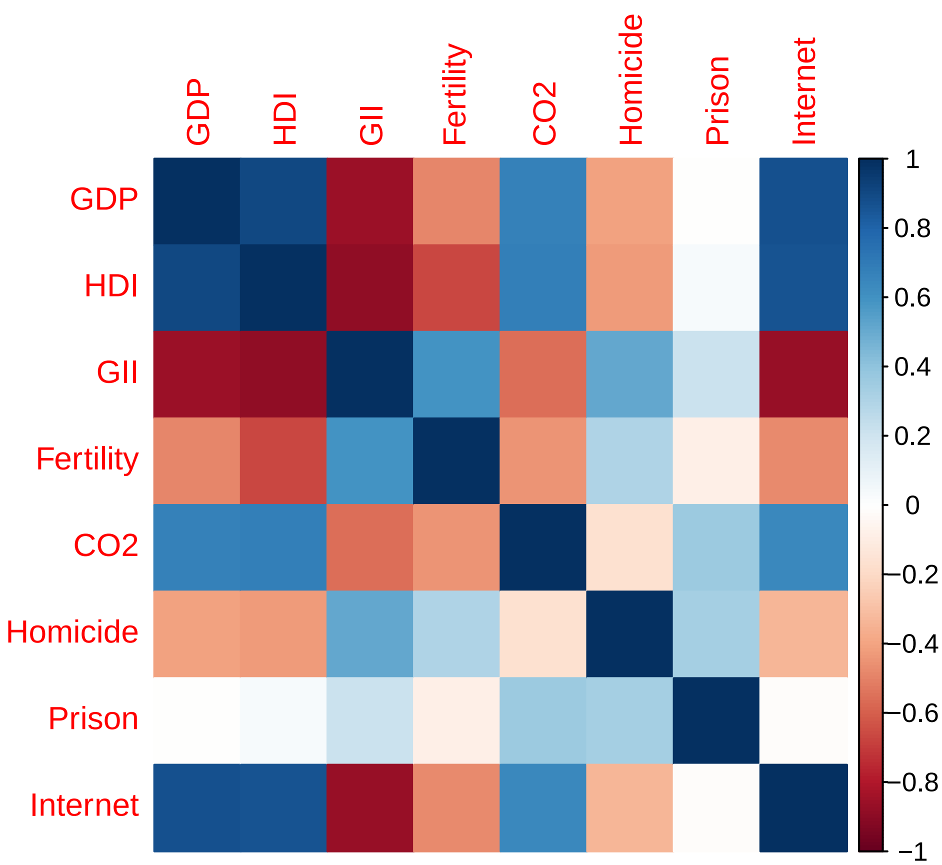
Although the multiple regression model includes additional predictors, the improvement in adjusted R^2 is relatively small. This indicates that **`GDP` alone explains most of the variability in `Internet` usage**. To better understand this, we investigate the correlation structure between the predictors.

We plot the correlation matrix to examine the relationships between the predictors and `Internet`:

```
library(corrplot)
```

```
## corrplot 0.95 loaded
```

```
M = cor(UN[-1], use = "complete.obs")
corrplot(M, method = 'color')
```



The correlation plot reveals that:

- `GDP` is highly positively correlated with `Internet`, which explains its dominance in both models.
- Several other predictors, such as `HDI` and `Fertility`, are also strongly correlated with `Internet` and with GDP. This multicollinearity reduces the unique contribution of each variable in the multiple regression model.

The significant decrease in the `GDP` coefficient from the bivariate model to the multiple regression model is due to multicollinearity between `GDP` and the other predictor variables (such as `HDI`, `GII`, `Fertility`, etc.). In practice, `GDP` shares a substantial portion of its variance with these variables, making it difficult to isolate the unique effect of `GDP` on `Internet` usage in the context of the multiple regression model. This is visible observing the `GDP` coefficients of the two models.

```r
cat("Coefficient of GDP in the bivariate regression model:\n\n")
```

```
## Coefficient of GDP in the bivariate regression model:
```

```r
print(summary(fit2)$coefficients["GDP", ])
```

```
##     Estimate   Std. Error      t value     Pr(>|t|)
## 1.405951e+00 1.216743e-01 1.155504e+01 2.548673e-14
```

```r
cat("\nCoefficient of GDP in the multiple regression model:\n\n")
```

```
##
## Coefficient of GDP in the multiple regression model:
```

```r
print(summary(fit1)$coefficients["GDP", ])
```

```
##   Estimate Std. Error    t value   Pr(>|t|)
##  0.4409033  0.2906796  1.5168016  0.1385605
```