Assignment 1

Simon Clarke Xia Yu David Ewing

2025-03-10

Braking Distance

In this question, do not use the `lm` function or a module that provides an implementation of k-NN. You are allowed to use elementary statistical objects like mean, variance, etc.

We will be predicting the distance that a car takes to stop when driving at a certain speed. The dataset is from 1930, so it might be slightly outdated. Units are miles per hour (speed) and feet (distance).

Data Preparation

```
# Load and preprocess dataset
```

Linear Regression (Without lm)

```
# Compute slope and intercept for simple linear regression
```

Using the linear regression model, predict the braking distance for a car going at 30 km/h and include an 80% prediction interval.

```
# Prediction for 30 km/h
```

k-NN Model

```
# Fit and predict using k-NN model
```

Filipino Household Income

In this question, you are allowed to use `lm`*.*

Load the dataset `income.csv`, which contains more than 40000 entries of data about households in the Philippines. There is a lot to learn in this dataset, but in this assignment we are going to focus on predicting household income (`Total.Household.Income`) from the number of underage children living in the household (`Members.with.age.5...17.years.old`).

1. For this assignment we will just be looking at the two features mentioned, so feel free to remove unnecessary columns and rename the feature names if that makes your life easier (I will just refer to these two features by `income` and `children` from now on).
2. Split the dataset randomly into a training set (80%) and test set (20%).
3. Perform a linear regression of type `income = b0 + b1 * children` for analysing the influence on `children` on `income`.
    1. What is the specific form of the affine-linear model, i.e. what are `b0` and `b1`.?
    2. What is the predicted mean income of a household with n children, for {0,1,...,8}n {0,1,...,8}? What are the associated 90% prediction intervals? Summarise all of this in a table.
    3. Using your test set, check how many percent of datapoints lie within the 90% prediction intervals.
4. Do all steps of part 3. again, but this time you will be predicting `log_income = log(income)` instead of `income`.
5. *Opportunity for showing extra effort:* Reflect on your results. What did you observe, and what do you think are the reasons for that?

Data Preparation

Data Load and Variable Rename

```
# Load the dataset 'income.csv' into R as a data frame called 'income_data'
df_total <- read.csv("income.csv")

# data column name exploration
str(df_total)
```

```
## 'data.frame':    41544 obs. of  60 variables:
##  $ Total.Household.Income                        : int  480332 198235 82785 107589 189322 152883 1986
##  $ Region                                        : chr  "CAR" "CAR" "CAR" "CAR" ...
##  $ Total.Food.Expenditure                        : int  117848 67766 61609 78189 94625 73326 104644 9
##  $ Main.Source.of.Income                         : chr  "Wage/Salaries" "Wage/Salaries" "Wage/Salarie
##  $ Agricultural.Household.indicator              : int  0 0 1 0 0 0 0 1 0 0 ...
##  $ Bread.and.Cereals.Expenditure                 : int  42140 17329 34182 34030 34820 29065 40992 371
##  $ Total.Rice.Expenditure                        : int  38300 13008 32001 28659 30167 25190 36312 281
##  $ Meat.Expenditure                              : int  24676 17434 7783 10914 18391 15336 12968 1464
##  $ Total.Fish.and..marine.products.Expenditure   : int  16806 11073 2590 10812 11309 8572 12310 15896
##  $ Fruit.Expenditure                             : int  3325 2035 1730 690 1395 2614 2565 3365 1370 3
##  $ Vegetables.Expenditure                        : int  13460 7833 3795 7887 11260 9035 15620 10520 5
##  $ Restaurant.and.hotels.Expenditure             : int  3000 2360 4545 6280 6400 0 6200 1130 10550 15
##  $ Alcoholic.Beverages.Expenditure               : int  0 960 270 480 1040 180 1920 480 0 0 ...
##  $ Tobacco.Expenditure                           : int  0 2132 4525 0 0 240 0 0 0 0 ...
##  $ Clothing..Footwear.and.Other.Wear.Expenditure: int  4607 8230 2735 1390 4620 1930 7930 4085 2780
##  $ Housing.and.water.Expenditure                 : int  63636 41370 14340 16638 31122 22782 24126 407
##  $ Imputed.House.Rental.Value                    : int  30000 27000 7200 6600 16800 6600 12000 19800
##  $ Medical.Care.Expenditure                      : int  3457 3520 70 60 140 95 340 75 200 1786 ...
##  $ Transportation.Expenditure                    : int  4776 12900 324 6840 6996 4044 12696 4140 7200
##  $ Communication.Expenditure                     : int  2880 5700 420 660 2100 1500 1848 3000 1800 72
##  $ Education.Expenditure                         : int  36200 29300 425 300 0 0 0 50 8000 13180 ...
```

```
##  $ Miscellaneous.Goods.and.Services.Expenditure : int  34056 9150 6450 3762 8472 5394 6126 5562 6510
##  $ Special.Occasions.Expenditure                : int  7200 1500 500 500 1000 600 6400 1500 500 4000
##  $ Crop.Farming.and.Gardening.expenses          : int  19370 0 0 15580 18887 0 72290 51840 0 0 ...
##  $ Total.Income.from.Entrepreneurial.Acitivites : int  44370 0 0 15580 75687 0 72290 51840 0 312974
##  $ Household.Head.Sex                           : chr  "Female" "Male" "Male" "Male" ...
##  $ Household.Head.Age                           : int  49 40 39 52 65 46 45 33 17 53 ...
##  $ Household.Head.Marital.Status                : chr  "Single" "Married" "Married" "Married" ...
##  $ Household.Head.Highest.Grade.Completed       : chr  "Teacher Training and Education Sciences Prog
##  $ Household.Head.Job.or.Business.Indicator     : chr  "With Job/Business" "With Job/Business" "With
##  $ Household.Head.Occupation                    : chr  "General elementary education teaching profes
##  $ Household.Head.Class.of.Worker               : chr  "Worked for government/government corporation
##  $ Type.of.Household                            : chr  "Extended Family" "Single Family" "Single Fam
##  $ Total.Number.of.Family.members               : int  4 3 6 3 4 4 5 5 2 6 ...
##  $ Members.with.age.less.than.5.year.old        : int  0 0 0 0 0 0 1 1 0 0 ...
##  $ Members.with.age.5...17.years.old            : int  1 1 4 3 0 0 0 1 2 0 ...
##  $ Total.number.of.family.members.employed      : int  1 2 3 2 2 3 1 0 0 1 ...
##  $ Type.of.Building.House                       : chr  "Single house" "Single house" "Single house"
##  $ Type.of.Roof                                 : chr  "Strong material(galvanized,iron,al,tile,conc
##  $ Type.of.Walls                                : chr  "Strong" "Strong" "Light" "Light" ...
##  $ House.Floor.Area                             : int  80 42 35 30 54 40 35 35 35 70 ...
##  $ House.Age                                    : int  75 15 12 15 16 7 18 48 8 12 ...
##  $ Number.of.bedrooms                           : int  3 2 1 1 3 2 1 2 1 3 ...
##  $ Tenure.Status                                : chr  "Own or owner-like possession of house and lo
##  $ Toilet.Facilities                            : chr  "Water-sealed, sewer septic tank, used exclus
##  $ Electricity                                  : int  1 1 0 1 1 1 1 1 1 1 ...
##  $ Main.Source.of.Water.Supply                  : chr  "Own use, faucet, community water system" "Ow
##  $ Number.of.Television                         : int  1 1 0 1 1 1 1 1 1 1 ...
##  $ Number.of.CD.VCD.DVD                         : int  1 1 0 0 0 0 0 1 0 0 ...
##  $ Number.of.Component.Stereo.set               : int  0 1 0 0 0 0 1 0 0 1 ...
##  $ Number.of.Refrigerator.Freezer               : int  1 0 0 0 1 0 0 0 0 1 ...
##  $ Number.of.Washing.Machine                    : int  1 1 0 0 0 1 0 1 0 0 ...
##  $ Number.of.Airconditioner                     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Number.of.Car..Jeep..Van                     : int  0 0 0 0 0 0 0 0 0 1 ...
##  $ Number.of.Landline.wireless.telephones       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Number.of.Cellular.phone                     : int  2 3 0 1 3 4 2 2 2 4 ...
##  $ Number.of.Personal.Computer                  : int  1 1 0 0 0 0 0 0 0 1 ...
##  $ Number.of.Stove.with.Oven.Gas.Range          : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Number.of.Motorized.Banca                    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Number.of.Motorcycle.Tricycle                : int  1 2 0 0 1 1 1 1 0 0 ...
```

```r
# remove unnecessary columns and rename the feature names as `income` and `children`.
income <- df_total$Total.Household.Income
children <- df_total$Members.with.age.5...17.years.old
```

Initial Exploration Of Target Variable 'income' And 'children'

```r
# Print the first few rows of the data
cat("First few rows of Household Income:\n", head(income), "\n\n")
```

```
## First few rows of Household Income:
##  480332 198235 82785 107589 189322 152883
```

```r
cat("First few rows of Number of Children:\n", head(children), "\n\n")
```

```
## First few rows of Number of Children:
##  1 1 4 3 0 0
```

```r
# Check for missing values
cat("Number of missing values in Household Income:", sum(is.na(income)), "\n")
```

```
## Number of missing values in Household Income: 0
```

```r
cat("Number of missing values in Number of Children:", sum(is.na(children)), "\n")
```

```
## Number of missing values in Number of Children: 0
```

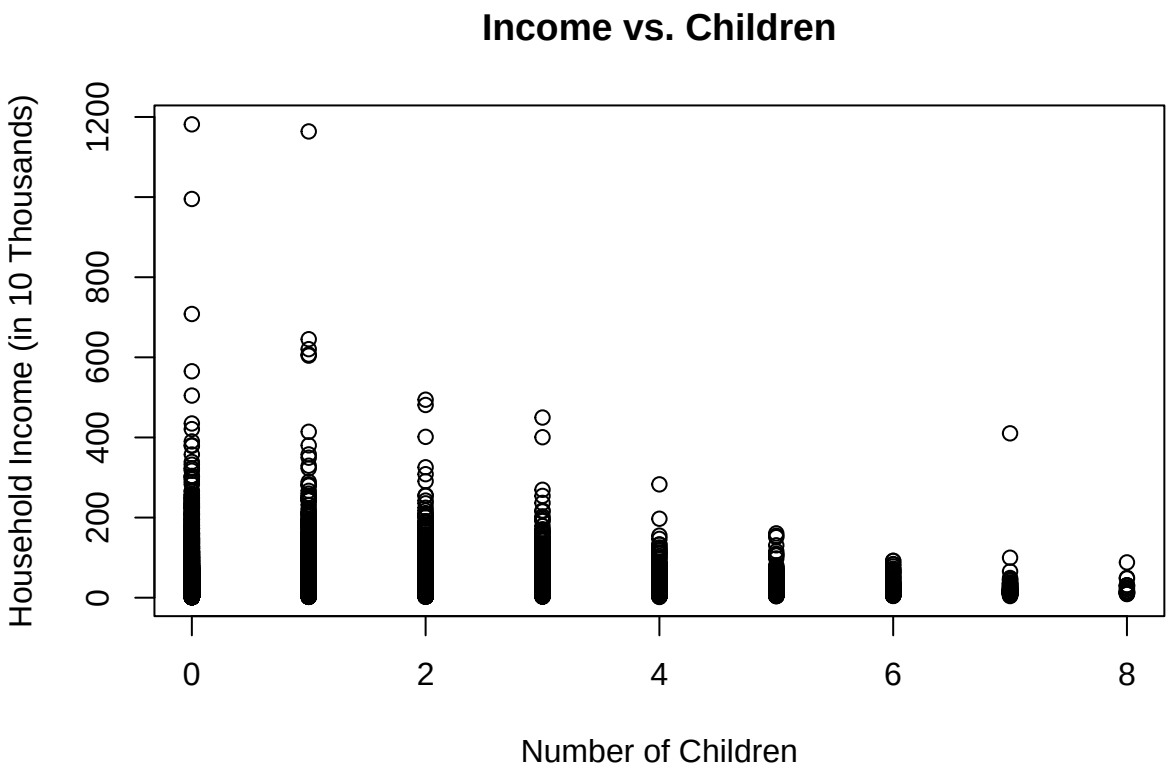
```r
# Print summaries in a concise way without storing the summaries separately
cat("\nSummary of Household Income:\n",
    paste(names(summary(income)), summary(income), sep = " = ", collapse = "\n"), "\n\n")
```

```
##
## Summary of Household Income:
##  Min. = 11285
## 1st Qu. = 104895
## Median = 164079.5
## Mean = 247555.584801656
## 3rd Qu. = 291138.5
## Max. = 11815988
```

```r
cat("Summary of Number of Children:\n",
    paste(names(summary(children)), summary(children), sep = " = ", collapse = "\n"), "\n")
```

```
## Summary of Number of Children:
##  Min. = 0
## 1st Qu. = 0
## Median = 1
## Mean = 1.36257943385326
## 3rd Qu. = 2
## Max. = 8
```

```r
# Scatter plot of income vs. children
plot(children, income/10000, xlab = "Number of Children", ylab = "Household Income (in 10 Thousands)", m
```



Linear Regression

Training Set And Testing Set Preparation

```
# Set a seed for reproducibility
set.seed(888)  # seed number could be modified, here I use seed for reduce RAM pressure \
# \in compution when smaple size is over 40000.
# Combine income and children into a data frame
data <- data.frame(income = income, children = children)
# Split the data into training (80%) and test sets (20%)
train_data <- data[sample(1:nrow(data), size = 0.8 * nrow(data)), ]
test_data <- data[-sample(1:nrow(data), size = 0.8 * nrow(data)), ]
```

Model Trainning

```
# Fit a linear regression model
model <- lm(income ~ children, data = train_data)

# Print the summary of the model
summary(model)
```

```
##
```

```
## Call:
## lm(formula = income ~ children, data = train_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
##  -251450  -140741   -79204    45428 11553253
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   262735       2181  120.45   <2e-16 ***
## children      -11256       1112  -10.12   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 285600 on 33233 degrees of freedom
## Multiple R-squared:  0.003072,   Adjusted R-squared:  0.003042
## F-statistic: 102.4 on 1 and 33233 DF,  p-value: < 2.2e-16
```

Model Explanation

1. What is the specific form of the affine-linear model, i.e. what are `b0` and `b1`?

- The affine-linear model |SLR model (simple linear regression model) has the general form: income $= b_0 + b_1 \cdot$ children
- b_0 is the **intercept** (the value of income when children $= 0$). Our model's $b_0 = 262735$
- b_1 is the **slope** (the change in income for a one-unit increase in children). Our model's $b_1 = -11256$
- So the specific form of our SLR model is:
  income $= 262735 - 11256 \cdot$ children

2. What is the predicted mean income of a household with n children, for n $\in \{0, 1, ..., 8\}$? What are the associated 90% prediction intervals? Summarize all of this in a table.

```
n <- array(0:8)
b_0 <- 262735
b_1 <- -11256
income_n <- b_0+b_1*n
cat(paste0("The predicted mean income of a household with 0-8 children = ",mean(income_n)),"\n")
```

```
## The predicted mean income of a household with 0-8 children = 217711
```

When the sample number n $= 8$, which is very small, we could .... (to be continue)

2. Using your test set, check how many percent of datapoints lie within the 90% prediction intervals.

Predicting Possum Age

Data and Initial Analysis

```
# Load dataset and visualize
```

Data Preparation

```
# Preprocess dataset
```

Feature Selection and Model Training

```
# Forward feature selection and model training
```

Model Evaluation

```
# Compute evaluation metrics
```

Further Exploration

```
# Additional analysis or research questions
```