HUL 315

Assignment - 1

Group Members

Lakshya Kumar	2022TT12183
Raunit Kumar Singh	2022TT12160
Yuvraj	2022MS11902
Keshav chachan	2022TT12148

Indian Institute of Technology, Delhi

September 6, 2025

Problem 1

Consider the following equation:

$$x + y + z = 100,$$

where $x, y, z \in \mathbb{Z}_{\geq 0}$ (the set of non-negative integers including zero). How many solutions exist for the above equation?

Solution:
We are asked to find the number of non-negative integer solutions to

$$x + y + z = 100.$$

This is a classic **stars-and-bars problem** in combinatorics. The general formula for the number of non-negative integer solutions of

$$x_1 + x_2 + \cdots + x_n = k$$

is

$$\binom{n+k-1}{n-1}.$$

Here, $n = 3$ (for x, y, z) and $k = 100$. Therefore, the number of solutions is

$$\binom{3+100-1}{3-1} = \binom{102}{2}.$$

We can compute this as:

$$\binom{102}{2} = \frac{102 \cdot 101}{2} = 5151.$$

Answer: There are **5151** non-negative integer solutions.

Problem 2

n friends place n pieces of paper with their names in a hat. After shuffling them each one of them takes one piece. How many possible outcomes exist such that no one takes their own name out of the hat?

Solution:
This problem asks for the number of permutations where no element appears in its original position. This is known as a **derangement**.
Let the set of friends be $F = \{f_1, \ldots, f_n\}$ and the set of name-slips be $P = \{p_1, \ldots, p_n\}$. A total outcome is a permutation of the slips. The condition is that friend f_i does not draw slip p_i for any i.
Let $!n$ or D_n denote the number of derangements of n items. This value can be calculated using the principle of inclusion-exclusion. The formula is:

$$D_n = !n = n! \sum_{k=0}^{n} \frac{(-1)^k}{k!} = n! \left(1 - \frac{1}{1!} + \frac{1}{2!} - \cdots + (-1)^n \frac{1}{n!}\right).$$

This counts all permutations ($n!$) and subtracts the cases where at least one person gets their

own name, then adds back the cases where at least two people do (as they were subtracted twice), and so on.
Answer: The number of outcomes is $!n$, the subfactorial of n.

Problem 3

Three numbers x, y, z are randomly selected **with replacement** from the set $\mathbb{Z}_{\geq 0}$ such that all of them are less than 100. Find the probability that

$$x < y < z.$$

Solution:
First, we determine the **total number of outcomes**. Since each of x, y, z can take any value from 0 to 99 independently (with replacement), the total number of ordered triples is

$$|\Omega| = 100 \times 100 \times 100 = 100^3.$$

$$where\ \Omega = \{(x, y, z) \mid x \in S,\ y \in S,\ z \in S\}, \quad S = \{0, 1, 2, \ldots, 99\}.$$

Now, we need to find the probability of the event

$$E = \{(x, y, z) \in \Omega \mid x < y < z\}.$$

Since $x < y < z$, all three numbers must be **distinct**. There are $\binom{100}{3}$ ways to choose 3 distinct numbers from 0 to 99. For each such set of 3 distinct numbers, there is exactly **one ordering** that satisfies $x < y < z$. Therefore, the number of favorable outcomes is

$$|E| = \binom{100}{3}.$$

So, the probability P(E) is

$$P(x < y < z) = \frac{|E|}{|\Omega|} = \frac{\binom{100}{3}}{100^3}.$$

Answer:

$$\boxed{\frac{\binom{100}{3}}{100^3} \approx 0.1617}.$$

Problem 4

A contestant in a TV game show is told that a valuable prize lies behind one of the closed doors marked A, B, and C. A "booby prize" in the form of a goat is behind another door, and the third door has nothing at all. The contestant chooses one door. The game master then opens one of the other doors to reveal the goat. The contestant is given the option to switch from the first choice to the remaining unopened door. Does making the switch increase the probability of winning? Show your work.

Solution:
Sticking Strategy: The initial probability of picking the prize is $\frac{1}{3}$. If the contestant sticks

with their original choice, their probability of winning remains $\frac{1}{3}$.

Switching Strategy: We analyze two scenarios based on the initial choice:

1. **The contestant initially chose the prize door (Probability $\frac{1}{3}$).** In this case, the other two doors hide the goat and nothing. The host must open the door with the goat. If the contestant switches, they will switch to the door with nothing and lose. The probability of winning by switching is 0.

2. **The contestant initially chose a non-prize door (Probability $\frac{2}{3}$).** In this case, the prize is behind one of the two unchosen doors. The host knows where the prize is and must open the door with the goat. The remaining unopened door must contain the prize. Therefore, switching guarantees a win. The probability of winning by switching is 1.

The overall probability of winning by switching is the sum of the probabilities of these two cases:

$$P(\text{win by switching}) = \left(\frac{1}{3} \times 0\right) + \left(\frac{2}{3} \times 1\right) = \frac{2}{3}.$$

Answer: Yes, making thea switch increases the probability of winning from $\mathbf{\frac{1}{3}}$ to $\mathbf{\frac{2}{3}}$.

Problem 5

The Magenta line metro starts with no passengers. At every station, 0, 1, or 2 passengers could board with probabilities 0.3, 0.5, and 0.2, respectively. An onboard passenger could alight with a probability of 0.1 at any station. Find the expected value and the variance of the number of passengers on the train after leaving the second station.

Solution:
Let B_i be the number of passengers boarding at station i and N_i be the number of passengers on the train after it leaves station i. We are given $N_0 = 0$.

Properties of Boarding (B_i):

$$E[B_i] = (0)(0.3) + (1)(0.5) + (2)(0.2) = 0.9.$$

$$E[B_i^2] = (0^2)(0.3) + (1^2)(0.5) + (2^2)(0.2) = 1.3.$$

$$\text{Var}(B_i) = E[B_i^2] - (E[B_i])^2 = 1.3 - 0.9^2 = 0.49.$$

After Station 1: Since $N_0 = 0$, no one can alight. Thus, $N_1 = B_1$.

$$E[N_1] = E[B_1] = 0.9 \quad \text{and} \quad \text{Var}(N_1) = \text{Var}(B_1) = 0.49.$$

After Station 2: Let R_2 be the number of passengers from station 1 who remain. Each of the N_1 passengers remains with probability $1 - 0.1 = 0.9$. The number of passengers after station 2 is $N_2 = R_2 + B_2$.

- **Expected Value** $E[N_2]$**:** Using the Law of Total Expectation,

$$E[N_2] = E[R_2]+E[B_2] = E[E[R_2|N_1]]+E[B_2] = E[0.9N_1]+0.9 = 0.9E[N_1]+0.9 = 0.9(0.9)+0.9 = 1.71.$$

- **Variance Var**(N_2)**:** Since boarding is independent of alighting, $\text{Var}(N_2) = \text{Var}(R_2) + \text{Var}(B_2)$. We find $\text{Var}(R_2)$ using the Law of Total Variance:

$$\text{Var}(R_2) = E[\text{Var}(R_2|N_1)] + \text{Var}(E[R_2|N_1]).$$

The conditional distribution $R_2|N_1$ is Binomial$(N_1, 0.9)$.

$$E[\text{Var}(R_2|N_1)] = E[N_1(0.9)(0.1)] = 0.09E[N_1] = 0.09(0.9) = 0.081.$$

$$\text{Var}(E[R_2|N_1]) = \text{Var}(0.9N_1) = 0.9^2\text{Var}(N_1) = 0.81(0.49) = 0.3969.$$

$$\text{Var}(R_2) = 0.081 + 0.3969 = 0.4779.$$

$$\text{Var}(N_2) = \text{Var}(R_2) + \text{Var}(B_2) = 0.4779 + 0.49 = 0.9679.$$

Answer: The expected value is **1.71** and the variance is **0.9679**.

Problem 6

In a batch of 1000 boxes of coconut water, there are 32 defective ones. Suppose 100 boxes are taken at random for testing. Let X be the number of defective boxes found during the testing. Find the p.m.f. for X.

Solution:
This scenario involves sampling without replacement from a finite population with two categories (defective and non-defective). Therefore, the random variable X follows a **hypergeometric distribution**.
The parameters are:

- Population size, $N = 1000$
- Number of defective items in population, $K = 32$
- Sample size, $n = 100$

The probability mass function (p.m.f.) for a hypergeometric distribution is given by:

$$P(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}.$$

Substituting the given values:

$$P(X = k) = \frac{\binom{32}{k}\binom{1000-32}{100-k}}{\binom{1000}{100}} = \frac{\binom{32}{k}\binom{968}{100-k}}{\binom{1000}{100}}.$$

The possible values for k (the number of defective boxes found in the sample) are integers from $\max(0, n - (N - K))$ to $\min(n, K)$, which is $k \in \{0, 1, \ldots, 32\}$.

Answer: The p.m.f. for X is $P(X = k) = \frac{\binom{32}{k}\binom{968}{100-k}}{\binom{1000}{100}}$ for $k \in \{0, 1, \ldots, 32\}$.

Problem 7

The p.m.f. for a discrete random variable X, which can only take natural number values, is given by $f(X = x) = \frac{1}{x(x+1)}$ for $x \in \{1, 2, 3, \ldots\}$. Find $\mathbb{E}[X]$.

Solution:
The expected value, $\mathbb{E}[X]$, of a discrete random variable is the sum of each possible value multiplied by its probability:

$$\mathbb{E}[X] = \sum_{x=1}^{\infty} x \cdot P(X = x).$$

Substituting the given p.m.f.:

$$\mathbb{E}[X] = \sum_{x=1}^{\infty} x \cdot \frac{1}{x(x+1)} = \sum_{x=1}^{\infty} \frac{x}{x(x+1)} = \sum_{x=1}^{\infty} \frac{1}{x+1}.$$

Let's write out the terms of this series to understand its behavior:

$$\mathbb{E}[X] = \frac{1}{1+1} + \frac{1}{2+1} + \frac{1}{3+1} + \cdots = \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \ldots$$

This series is the well-known **harmonic series**, $\sum_{n=1}^{\infty} \frac{1}{n}$, but with the first term ($n = 1$) removed. The harmonic series is a classic example of a divergent series, meaning its sum approaches infinity. Since our series for $\mathbb{E}[X]$ is just the harmonic series minus 1, it also diverges to infinity.
Answer: The expected value $\mathbb{E}[\mathbf{X}]$ is infinite (∞).

Problem 8

A factory is believed to produce a finite number N of consecutively numbered tanks each month. You draw a simple random sample without replacement of size n and observe serial numbers $X_1, X_2, \ldots, X_n$. Consider the following two estimators of N:

$$\hat{N}_{mn} = 2\bar{X} - 1 \qquad \hat{N}_{mx} = \frac{n+1}{n} \max\{X_1, X_2, \ldots, X_n\} - 1$$

Solution (Part a):
Estimator (i): $\hat{N}_{mn} = 2\bar{X} - 1$
- Expectation of one draw:

$$E[X_i] = \frac{N+1}{2}, \quad Var(X_i) = \frac{N^2 - 1}{12}.$$

- Because sample is without replacement,

$$Var(\bar{X}) = \frac{Var(X_i)}{n} \cdot \frac{N-n}{N-1} = \frac{(N^2-1)}{12n} \cdot \frac{N-n}{N-1}.$$

- Simplify:

$$Var(\bar{X}) = \frac{(N+1)(N-n)}{12n}.$$

- Now for the estimator:

$$E[\hat{N}_{mn}] = 2E[\bar{X}] - 1 = 2 \cdot \frac{N+1}{2} - 1 = N,$$

so it is unbiased.

$$Var(\hat{N}_{mn}) = 4Var(\bar{X}) = \frac{(N+1)(N-n)}{3n}.$$

—

Estimator (ii): $\hat{N}_{mx} = \frac{n+1}{n}M - 1$**, where** $M = \max\{X_1, \ldots, X_n\}$

- PMF of maximum:

$$P(M = k) = \frac{\binom{k-1}{n-1}}{\binom{N}{n}}, \quad k = n, \ldots, N.$$

- Expected value of M (using hockey-stick identity):

$$E[M] = \frac{n}{\binom{N}{n}} \sum_{k=n}^{N} \binom{k}{n} = \frac{n}{\binom{N}{n}} \binom{N+1}{n+1} = \frac{n(N+1)}{n+1}.$$

- Hence,

$$E[\hat{N}_{mx}] = \frac{n+1}{n}E[M] - 1 = \frac{n+1}{n} \cdot \frac{n(N+1)}{n+1} - 1 = N.$$

- Variance of M:

$$Var(M) = \frac{n(N-n)(N+1)}{(n+1)^2(n+2)}.$$

- Therefore,

$$Var(\hat{N}_{mx}) = \left(\tfrac{n+1}{n}\right)^2 Var(M) = \frac{(N-n)(N+1)}{n(n+2)}.$$

Finally:

$$E[\hat{N}_{mn}] = N, \quad Var(\hat{N}_{mn}) = \frac{(N+1)(N-n)}{3n}$$

$$E[\hat{N}_{mx}] = N, \quad Var(\hat{N}_{mx}) = \frac{(N-n)(N+1)}{n(n+2)}$$

Solution (Part b):

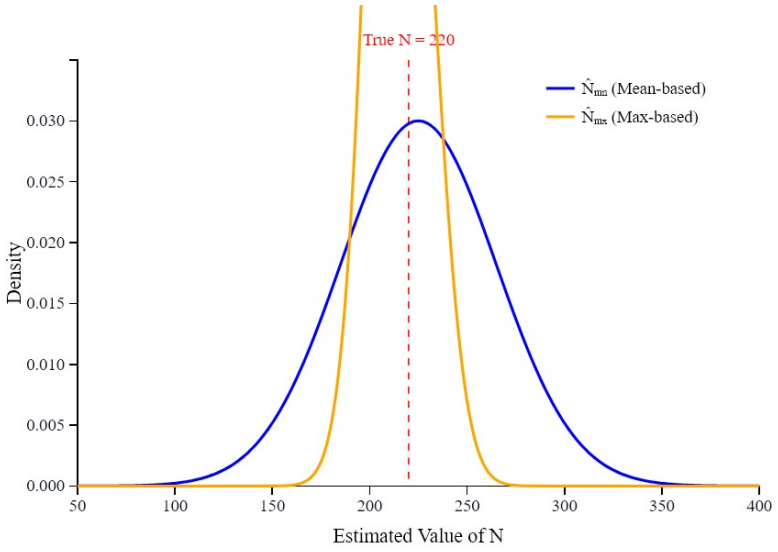
Listing 1: R Code: Simulation of Estimators

```
set.seed(123)
N <- 220; n <- 10; R <- 10000

simulate_once <- function() {
  sample_vals <- sample(1:N, n, replace = FALSE)
  mn_hat <- 2*mean(sample_vals) - 1
  mx_hat <- ((n+1)/n)*max(sample_vals) - 1
```

```
  return(c(mn_hat, mx_hat))
}

results <- replicate(R, simulate_once())
mn_hats <- results[1, ]; mx_hats <- results[2, ]

mean_mn <- mean(mn_hats); var_mn <- var(mn_hats)
bias_mn <- mean_mn - N; mse_mn <- var_mn + bias_mn^2

mean_mx <- mean(mx_hats); var_mx <- var(mx_hats)
bias_mx <- mean_mx - N; mse_mx <- var_mx + bias_mx^2

cat("Nmn: mean =", mean_mn, " var =", var_mn,
    " bias =", bias_mn, " MSE =", mse_mn, "\n")
cat("Nmx: mean =", mean_mx, " var =", var_mx,
    " bias =", bias_mx, " MSE =", mse_mx, "\n")

hist(mn_hats, probability=TRUE, col=rgb(1,0,0,0.3),
     main="Empirical Distributions of Estimators", xlab="Estimate")
hist(mx_hats, probability=TRUE, col=rgb(0,0,1,0.3), add=TRUE)
abline(v=N, col="black", lwd=2)
legend("topright", legend=c("Nmn","Nmx","True N"),
       fill=c(rgb(1,0,0,0.3), rgb(0,0,1,0.3), "black"))
```

Listing 2: Console Output: Empirical Stats

```
> source("A1-Q8b.R")
Nmn: mean = 220.47  var = 1525.66  bias = 0.47  MSE = 1525.88
Nmx: mean = 220.09  var = 378.85   bias = 0.09  MSE = 378.86
```



Solution (Part c):
Both estimators are unbiased, confirmed both analytically and by simulation. However:

- $\hat{N}_{mn}$ has variance ≈ 1526, - $\hat{N}_{mx}$ has variance ≈ 379,
so $\hat{N}_{mx}$ is much more efficient. The distribution of $\hat{N}_{mn}$ is wide and symmetric, while $\hat{N}_{mx}$ is more concentrated but slightly left-skewed (extreme-value behavior). Hence in practice $\hat{N}_{mx}$ is the superior estimator.

Problem 9

(a) (7 points) Write a code using either Stata / R / Python to create a vector x of dimension (1000×1) with realizations of a random variable distributed uniformly over the unit interval $[0, 1]$. Now draw a random sample of 100 elements from the vector x and calculate the sample mean. Repeat the exercise 1000 times and store the sample means in another vector y of dimension (1000×1). Task: Plot the distribution of x and y in the same diagram.

Solution (Part a):
Let X be a random variable with realizations drawn from the uniform distribution over the unit interval $[0, 1]$, i.e.

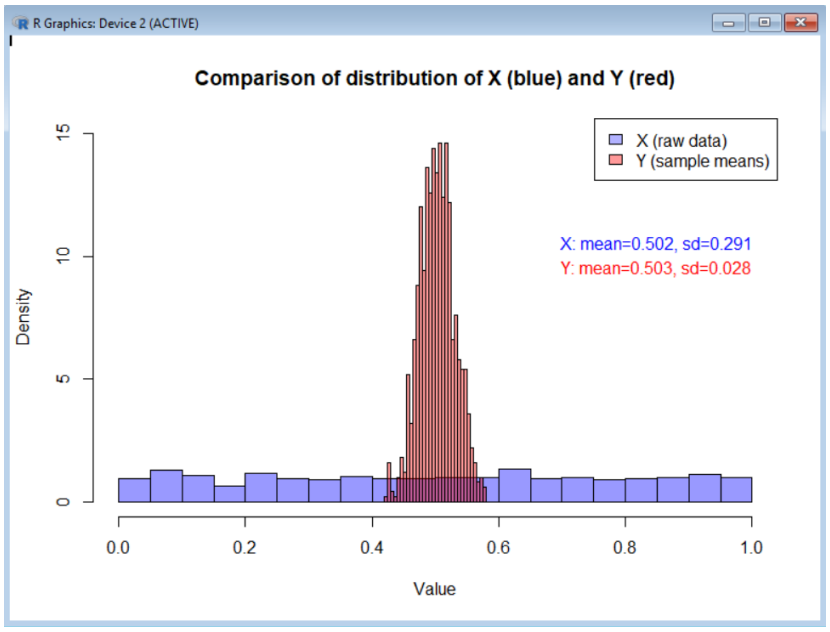
$$X \sim \text{Uniform}(0, 1).$$

We create a vector $\mathbf{x} \in \mathbb{R}^{1000\times 1}$ containing 1000 independent realizations of X. From $\mathbf{x}$, we draw random samples of size 100 (without replacement) and compute their sample means. Repeating this process 1000 times, we create another vector $\mathbf{y} \in \mathbb{R}^{1000\times 1}$, where each element of $\mathbf{y}$ represents one sample mean.

Listing 3: R Code

```
set.seed(2160)

# vector x, dim : (1000 x 1), values from a uniform distribution
   between 0 and 1
x <- runif(1000, min = 0, max = 1)

# vector y, dim : (1000 x 1), values from the mean of x
y <- numeric(1000)
for (i in 1:1000) {
  r_sample <- sample(x, size = 100, replace = FALSE)
  y[i] <- mean(r_sample)
}

# compute stats
mean_x <- mean(x); sd_x <- sd(x)
mean_y <- mean(y); sd_y <- sd(y)

# plot histograms
hist(x, breaks = 30, probability = TRUE, col = rgb(0,0,1,0.4),
     main = "Comparison of distribution of X (blue) and Y (red)", xlab
         = "Value", ylim = c(0, 15))
hist(y, breaks = 30, probability = TRUE, col = rgb(1,0,0,0.4), add =
   TRUE)

# legend
legend("topright", legend = c("X (raw data)", "Y (sample means)"),
```

```
        fill = c(rgb(0,0,1,0.3), rgb(1,0,0,0.4)))

# add mean and sd text to plot
text(0.85, 10.5, paste0("X: mean=", round(mean_x,3), ", sd=", round(sd
    _x,3)), col = "blue")
text(0.85, 9.5, paste0("Y: mean=", round(mean_y,3), ", sd=", round(sd_
    y,3)), col = "red")
```



(b) (3 points) Interpret your results.

Solution (Part b):
The results clearly illustrate the Central Limit Theorem (CLT). The original data x, drawn from a uniform distribution on $[0, 1]$, has a flat shape with theoretical mean $\mu = 0.5$ and standard deviation $\sigma = 0.2887$, closely matching the simulated values ($\hat{\mu}_x = 0.502$, $\hat{\sigma}_x = 0.2910$). When we take repeated samples of size $n = 100$ and compute their means to form vector y, the resulting distribution is approximately bell-shaped, with mean close to 0.5 ($\hat{\mu}_y = 0.5030$) and a much smaller spread ($\hat{\sigma}_y = 0.0280$), consistent with the theoretical $\sigma/\sqrt{n} \approx 0.0289$. This shows that the sample mean is an unbiased estimator of the population mean, that the distribution of sample means tends toward normality even if the original data is not, and that averaging reduces variability. Overall, the experiment verifies the CLT: as sample size increases, the distribution of sample means approaches a normal distribution with mean μ and variance σ^2/n, regardless of the underlying distribution.

Problem 10

(a) (5 points) Plot the variables x and y (in their respective axes) using a scatter plot for each distinct string values of the variable dataset.

Listing 4: R Code

```r
df <- read.csv("datadino.csv")
datasets <- unique(df$dataset)

n <- length(datasets) # total dataplots
row <- ceiling(sqrt(n)) # no. of rows
col <- ceiling(n / row) # no. of columns
par(mfrow = c(row, col))

for (d in datasets) {
  subset_df <- subset(df, dataset == d)
  plot(subset_df$x, subset_df$y,
       main = d,
       xlab = "x", ylab = "y",
       pch = 19, col = "blue")
}
```

Solution (Part a):

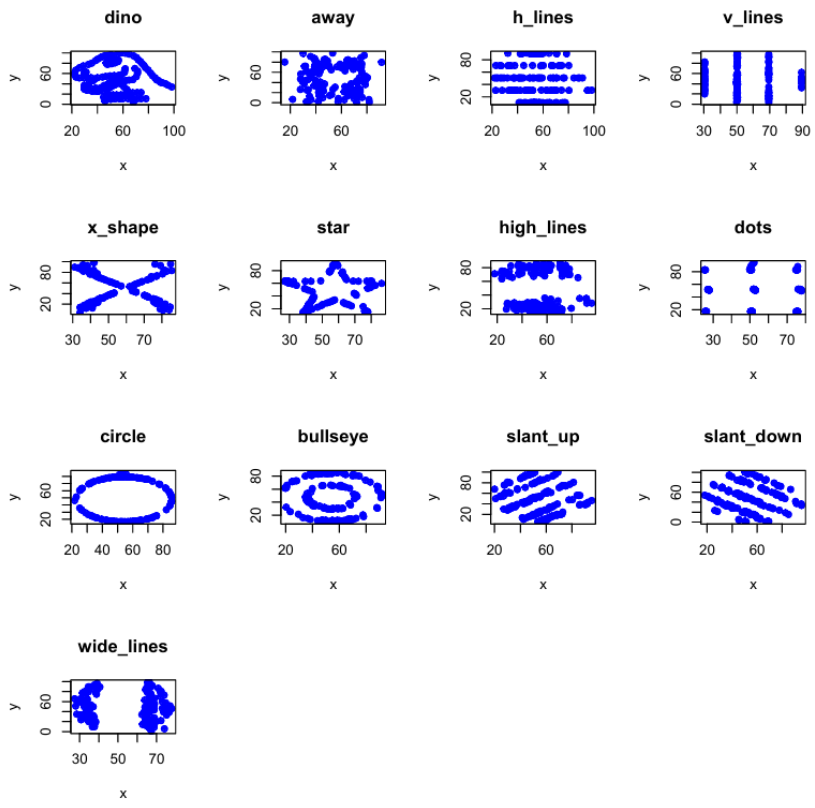


Figure 1: Scatter plots of variables x and y for each distinct dataset in `datadino.csv`

(b) (3 points) Compute the average values of x and y, the standard deviations of x and y, and the correlation coefficient between x and y for each distinct string value of the variable

dataset.

Solution (Part b):

Listing 5: R Code

```r
df <- read.csv("datadino.csv")

result <- do.call(rbind, lapply(split(df, df$dataset), function(subset
    _df) {
  data.frame(
    dataset = unique(subset_df$dataset),
    mean_x = mean(subset_df$x),
    mean_y = mean(subset_df$y),
    std_x  = sd(subset_df$x),
    std_y  = sd(subset_df$y),
    corr_xy = cor(subset_df$x, subset_df$y)
  )
}))

print(result)
```

round-mode=places, round-precision=4

Dataset	Mean x	Mean y	Std. x	Std. y	Corr(x, y)
away	54.26610	47.83472	16.76982	26.93974	-0.06412835
bullseye	54.26873	47.83082	16.76924	26.93573	-0.06858639
circle	54.26732	47.83772	16.76001	26.93004	-0.06834336
dino	54.26327	47.83225	16.76514	26.93540	-0.06447185
dots	54.26030	47.83983	16.76774	26.93019	-0.06034144
h_lines	54.26144	47.83025	16.76590	26.93988	-0.06171484
high_lines	54.26881	47.83545	16.76670	26.94000	-0.06850422
slant_down	54.26785	47.83590	16.76676	26.93610	-0.06897974
slant_up	54.26588	47.83150	16.76885	26.93861	-0.06860921
star	54.26734	47.83955	16.76896	26.93027	-0.06296110
v_lines	54.26993	47.83699	16.76996	26.93768	-0.06944557
wide_lines	54.26692	47.83160	16.77000	26.93790	-0.06657523
x_shape	54.26015	47.83972	16.76996	26.93000	-0.06558334

Table 1: Computed summary statistics (mean, standard deviation, and correlation) of x and y for each dataset in `datadino.csv`

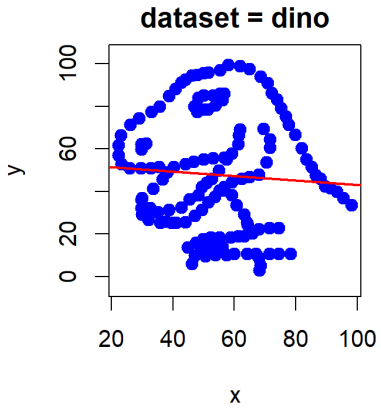
(c) (5 points) Plot the regression line of y on x and a constant on the same plots in part (a).

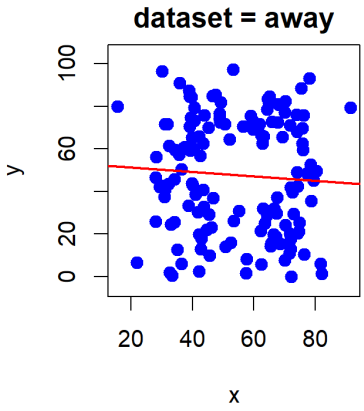
Solution (Part c):

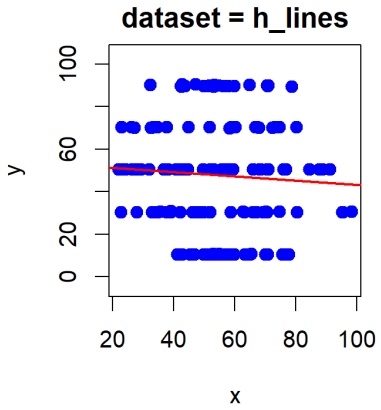
Listing 6: R Code

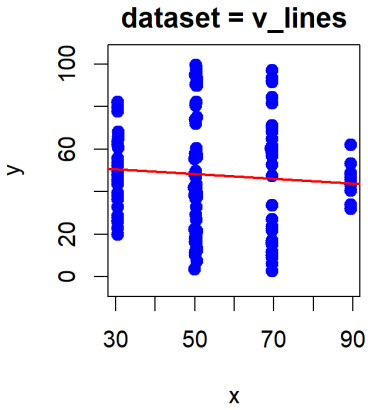
```r
data <- read.csv("datadino.csv")

# different datasets
```

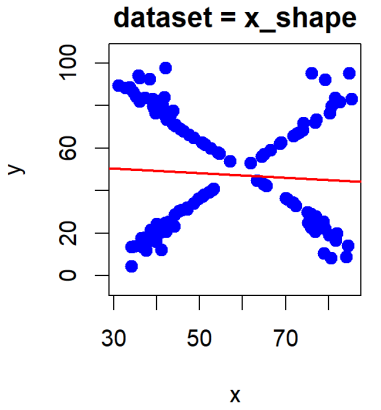
```r
datasets <- unique(data$dataset)
n <- length(datasets)

# 2 plot in one row for big plot
mcol <- 2
mrow <- ceiling(n / mcol)

# Exporting png
png("datadino_plots.png", width = 2000, height = 5000, res = 200)

par(mfrow = c(mrow, mcol), mar = c(4, 4, 2, 1), cex = 1.2, pty = "s")

for (d in datasets) {
  subset_data <- subset(data, dataset == d)

  plot(subset_data$x, subset_data$y,
       main = paste("dataset =", d),
       xlab = "x", ylab = "y",
       pch = 19, col = "blue", cex = 1.2,
       ylim = c(min(data$y) - 5, max(data$y) + 5))

  # regression
  model <- lm(y ~ x, data = subset_data)
  abline(model, col = "red", lwd = 2)
}

dev.off()
```

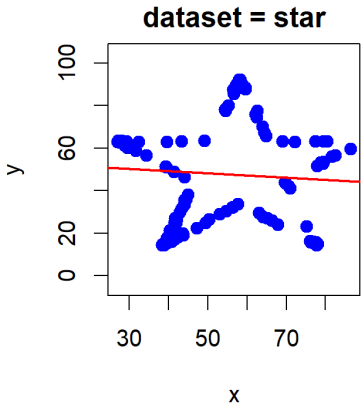

dataset = dino
y
x
0
20
40
60
80
100

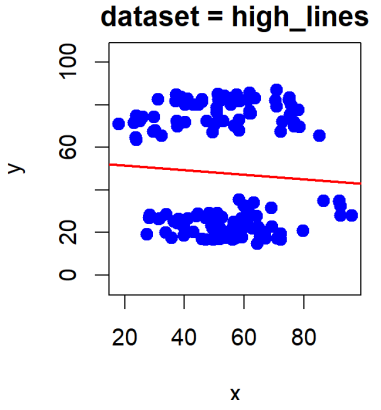

dataset = away
y
x
0
20
40
60
80
100

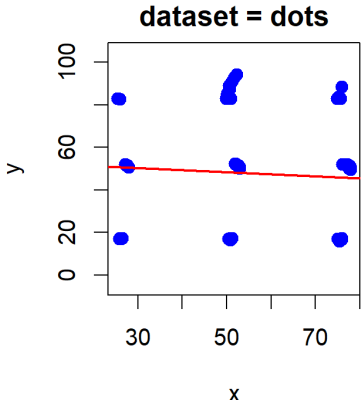

dataset = h_lines
y
x
0
20
40
60
80
100


dataset = v_lines
y
x
0
20
30
50
70
90
100


dataset = x_shape
y
x
0
20
30
50
70
100


dataset = star
y
x
0
20
30
50
70
100


dataset = high_lines
y
x
0
20
40
60
80
100


dataset = dots
y
x
0
20
30
50
70
100

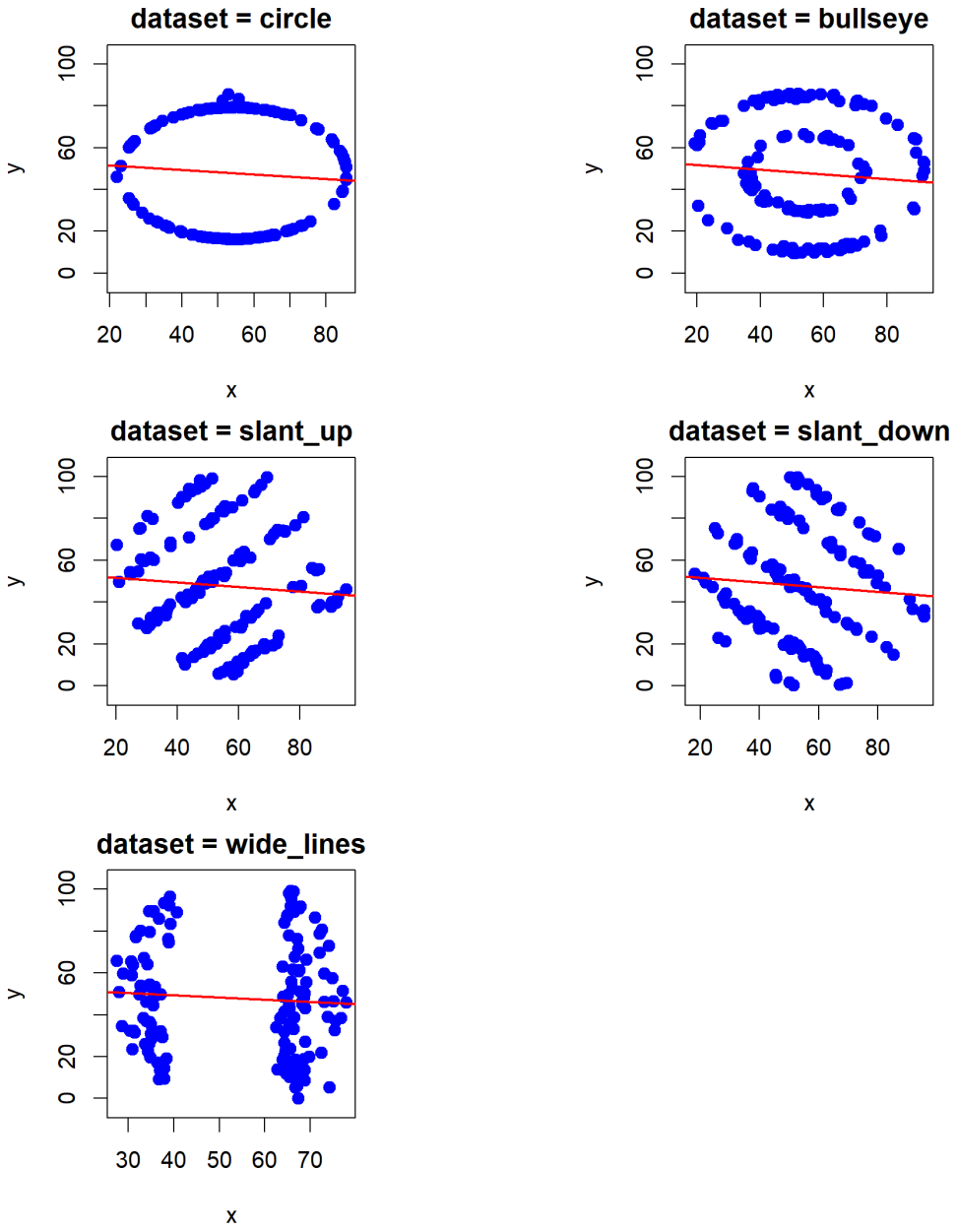


Figure 2: Scatter plots with regression lines for each dataset in `datadino.csv`.

(d) (2 points) Summarize your findings in two sentences.

Solution (Part d):
Even though the datasets have nearly the same averages, spreads, and correlations, the scatterplots look completely different. This makes it clear that looking at numbers alone can hide important patterns, so plotting the data is just as important as calculating summaries.

Problem 11

Read the data in the file energy gdp.csv in STATA (or any other statistical software of your choice).
(a) (5 points) Estimate the following simple linear regression and report the point estimates of the parameters α and β:

$$\log(\text{Energy}) = \alpha + \beta \log(\text{Real GDP}) + \varepsilon$$

Solution (Part a):

Listing 7: R Code: Estimates of Parameters

```
# Load the data
data <- read.csv("../data/energy_gdp.csv")

# Take logs of GDP and Energy
data$log_rgdp <- log(data$rgdp)
data$log_energy <- log(data$energy)

# Run simple linear regression
model <- lm(log_energy ~ log_rgdp, data = data)

# Display regression summary
summary(model)

# Extract point estimates of alpha (intercept) and beta (slope)
alpha <- coef(model)[1]
beta <- coef(model)[2]

cat("Alpha (Intercept):", alpha, "\n")
cat("Beta (Slope):", beta, "\n")
```

Listing 8: Console Output: Alpha and Beta Values

```
> source("A1-Q11a.R")
Alpha (Intercept): 1.988607
Beta (Slope): 0.7439504
> summary(model)

Call:
lm(formula = log_energy ~ log_rgdp, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-5.7713 -0.1425  0.4027  0.6198  1.1815

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.9886     2.1762   0.914  0.37291
log_rgdp      0.7440     0.2028   3.669  0.00176 **
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .
   0.1          1
```

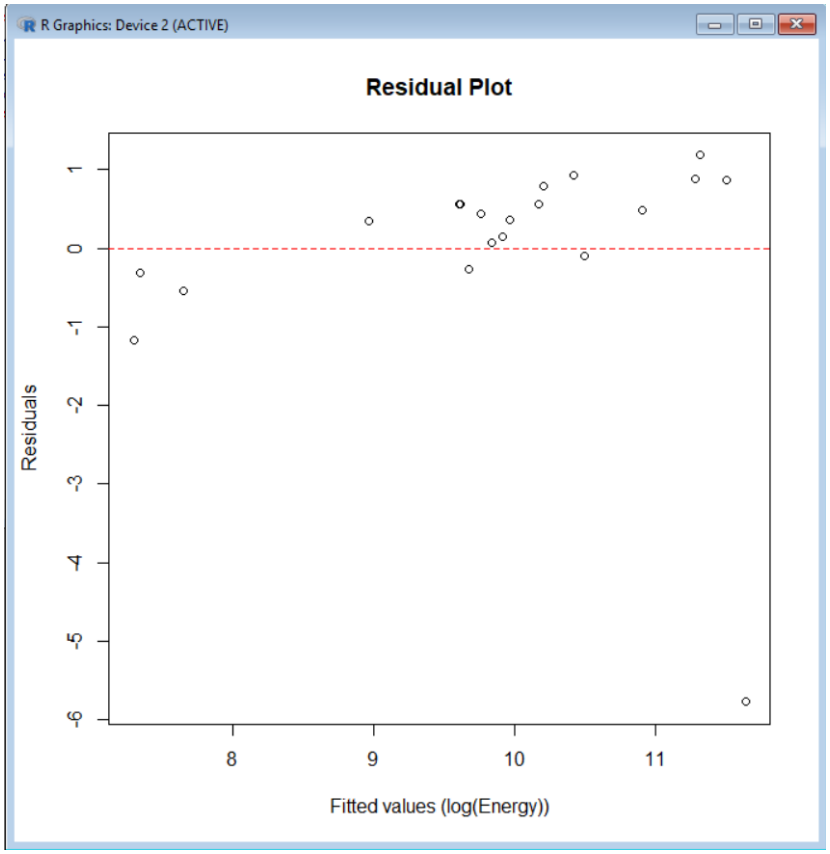
```
Residual standard error: 1.513 on 18 degrees of freedom
Multiple R-squared:  0.4279,    Adjusted R-squared:  0.3961
F-statistic: 13.46 on 1 and 18 DF,  p-value: 0.001756

>
```

(b) (5 points) Plot the residuals. Does anything strike you?

Solution (Part b):

Listing 9: R Code: Residual Plot

```r
# Load the data
data <- read.csv("../data/energy_gdp.csv")

# Take logs of GDP and Energy
data$log_rgdp <- log(data$rgdp)
data$log_energy <- log(data$energy)

# Run simple linear regression
model <- lm(log_energy ~ log_rgdp, data = data)

# Extract residuals
residuals <- resid(model)

# Plot residuals vs. fitted values
plot(fitted(model), residuals,
     xlab = "Fitted values (log(Energy))",
     ylab = "Residuals",
     main = "Residual Plot")
abline(h = 0, col = "red", lty = 2)
```



(c) (5 points) Test the hypothesis that the coefficient on log(Real GDP) is equal to one.

Solution (Part c):

Listing 10: R Code: Hypothesis Test

```r
# Load the data
data <- read.csv("../data/energy_gdp.csv")

# Take logs
data$log_rgdp <- log(data$rgdp)
data$log_energy <- log(data$energy)

# Run regression
model <- lm(log_energy ~ log_rgdp, data = data)

# Extract coefficient and standard error for log_rgdp
beta_hat <- coef(summary(model))["log_rgdp", "Estimate"]
se_beta <- coef(summary(model))["log_rgdp", "Std. Error"]

# Hypothesis test: H0: beta = 1
t_stat <- (beta_hat - 1) / se_beta
df <- model$df.residual
p_value <- 2 * pt(-abs(t_stat), df)
```

```
cat("Beta estimate:", beta_hat, "\n")
cat("t-statistic:", t_stat, "\n")
cat("p-value:", p_value, "\n")
```

Listing 11: Console Output: Hypothesis test

```
> source("A1-Q11c.R")
Beta estimate: 0.7439504
t-statistic: -1.262793
p-value: 0.2227783
>
```

(d) (5 points) For one country, the energy data is recorded incorrectly. Multiply it by 1000 and repeat the exercise in parts (a), (b), and (c).

Solution (Part d (i)):

Listing 12: R Code: Re-estimation after correcting energy data

```
# Multiply the last observation of energy by 1000
data$energy[nrow(data)] <- data$energy[nrow(data)] * 1000

# Re-estimate the regression model
model2 <- lm(log(energy) ~ log(rgdp), data = data)

# Extract coefficients
alphanew <- coef(model2)[1]
betanew  <- coef(model2)[2]

cat("AlphaNew (Intercept):", alphanew, "\n")
cat("BetaNew (Slope):", betanew, "\n")

# R-squared for model2
rsq_new <- summary(model2)$r.squared
cat("R-squared (New):", rsq_new, "\n")
```

Listing 13: Console Output: Re-estimation after correcting energy data

```
> source("A1-Q11d_part_a.R")
AlphaNew (Intercept): -0.7782875
BetaNew (Slope): 1.037499
R-squared (New): 0.9675033
>
```

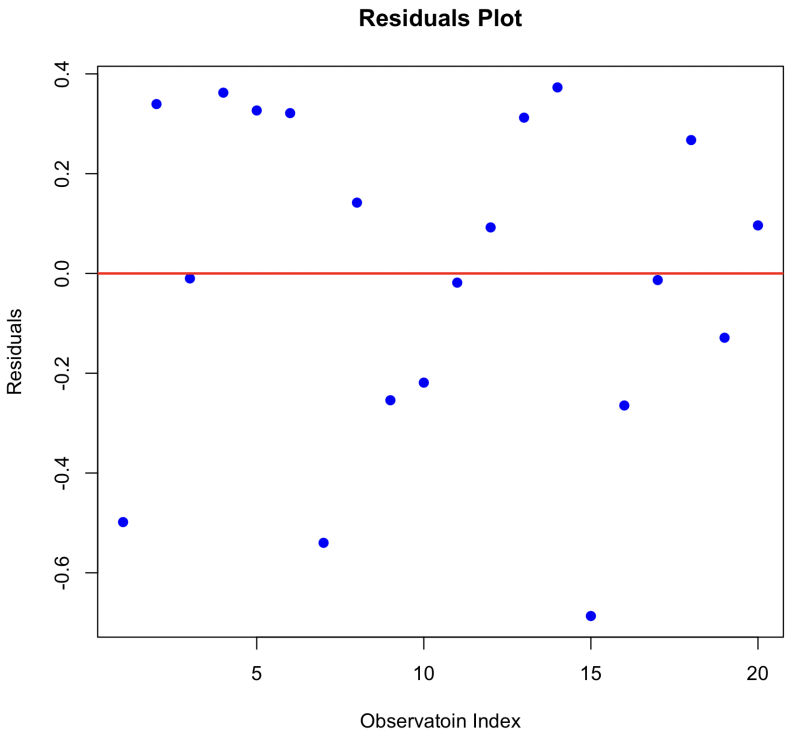
(d) Plot the residuals from the new regression model.

Solution (Part d (ii)):

Listing 14: R Code: Residual Plot after correcting energy data

```
# Extract residuals from the new model
residualnew <- resid(model2)

# Plot residuals
```

```r
plot(resid(model2),
     type = "p", pch = 19, col = "blue",
     xlab = "Observation Index",
     ylab = "Residuals",
     main = "Residuals Plot")

# Add horizontal line at 0
abline(h = 0, col = "red", lwd = 2)
```



(d)(c) Test the hypothesis that the coefficient on log(Real GDP) is equal to one after correcting the energy data.

Solution (Part d (iii)):

Listing 15: R Code: Hypothesis Test after correcting energy data

```r
# Standard error of beta (log(rgdp))
se_betanew <- summary(model2)$coefficients["log(rgdp)", "Std. Error"]

# Test statistic for H0: beta = 1
t_statnew <- (betanew - 1) / se_betanew
cat("t_statnew:", t_statnew, "\n")

# Degrees of freedom
dfnew <- df.residual(model2)
```

```r
# Two-sided p-value
p_valuenew <- 2 * pt(-abs(t_statnew), dfnew)
cat("p_valuenew:", p_valuenew, "\n")
```

Listing 16: Console Output: Hypothesis Test after correcting energy data

```
t_statnew: 0.836721
p_valuenew: 0.4137204
```

(e) (5 points) Estimate the equation $\log(\text{Real GDP}) = \delta + \gamma \log(\text{Energy}) + \epsilon$ and report the point estimates.

Solution (Part e):

Listing 17: R Code: Regression of log(Real GDP) on log(Energy)

```r
# Correct the last energy observation
data$energy[nrow(data)] <- data$energy[nrow(data)] * 1000

# Estimate the model
model3 <- lm(log(rgdp) ~ log(energy), data = data)

# Extract coefficients
delta <- coef(model3)[1]
gamma <- coef(model3)[2]

cat("Delta (Intercept):", delta, "\n")
cat("Gamma (Slope):", gamma, "\n")

# R-squared
r2_model3 <- summary(model3)$r.squared
cat("R^2 (Corrected Model3):", r2_model3, "\n")
```

Listing 18: Console Output: Regression Results

```
Model 3 - Delta (Intercept): 1.070317
Model 3 - Gamma (Slope): 0.9325338
Model 3 - R^2: 0.9675033
```

(f) (2 points) Report the R^2 values from both models.

Solution (Part f):

- **Model 2:** log(Energy) on log(Real GDP), $R^2 = 0.9675033$
- **Model 3:** log(Real GDP) on log(Energy), $R^2 = 0.9675033$

(g) (5 points) Proof that $\hat{\beta} \cdot \hat{\gamma} = R^2$ and numerical confirmation.
Solution (Part g):

Consider two random variables $X = \log(\text{Real GDP})$ and $Y = \log(\text{Energy})$. In the simple regression of Y on X, the slope coefficient is

$$\hat{\beta} = \frac{\text{Cov}(X, Y)}{\text{Var}(X)}.$$

In the reverse regression of X on Y, the slope coefficient is

$$\hat{\gamma} = \frac{\text{Cov}(X, Y)}{\text{Var}(Y)}.$$

Multiplying the two slope estimates gives

$$\hat{\beta}\hat{\gamma} = \frac{\text{Cov}(X, Y)^2}{\text{Var}(X)\text{Var}(Y)}.$$

The squared correlation coefficient is defined as

$$R^2 = \left(\frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}} \right)^2 = \frac{\text{Cov}(X, Y)^2}{\text{Var}(X)\text{Var}(Y)}.$$

Hence, we obtain the general result

$$\hat{\beta} \cdot \hat{\gamma} = R^2.$$

Numerical confirmation:

$$\hat{\beta} = 1.037499, \quad \hat{\gamma} = 0.9325338$$

$$\hat{\beta} \cdot \hat{\gamma} \approx 1.037499 \times 0.9325338 = 0.9675$$

$$R^2 \text{ (from Model 2 and Model 3)} = 0.9675033$$

Thus, the product $\hat{\beta}\hat{\gamma}$ equals the R^2, confirming the theoretical result.

(h) (5 points) Suppose energy is measured in BTU's instead of kilograms of coal. This implies that the original series must be multiplied by 60. We re-estimate the model:

$$\log(\text{Energy}_{BTU}) = \alpha + \beta \log(\text{Real GDP}) + \epsilon$$

Solution (Part h):

Listing 19: R Code: Regression with Energy in BTUs

```
# Convert energy to BTU after correcting
data$energy_btu <- data$energy * 60

# Estimate the model
model_btu <- lm(log(energy_btu) ~ log(rgdp), data = data)

# Extract coefficients
alpha_btu <- coef(model_btu)[1]
```

```r
beta_btu  <- coef(model_btu)[2]

cat("BTU Model - Alpha (Intercept):", alpha_btu, "\n")
cat("BTU Model - Beta (Slope):", beta_btu, "\n")

# R-squared
rsq_btu <- summary(model_btu)$r.squared
cat("BTU Model - R^2:", rsq_btu, "\n")
```

Listing 20: Console Output: Regression Results (BTU Model)

```
BTU Model - Alpha (Intercept): 3.316057
BTU Model - Beta (Slope): 1.037499
BTU Model - R^2: 0.9675033
```

Interpretation: Multiplying energy by 60 (to convert to BTUs) only shifts the intercept upward by $\log(60) \approx 4.094$, while the slope β and the R^2 remain unchanged.
Explanation: Why $\hat{\alpha}$ changes but $\hat{\beta}$ remains the same
Consider the regression model:

$$\log(\text{Energy}) = \alpha + \beta \log(\text{Real GDP}) + \epsilon$$

Now suppose Energy is rescaled by a constant factor $k = 60$, i.e.,

$$\text{Energy}^* = k \cdot \text{Energy}.$$

Taking logs:

$$\log(\text{Energy}^*) = \log(k \cdot \text{Energy}) = \log(k) + \log(\text{Energy}).$$

Substitute the original regression equation:

$$\log(\text{Energy}^*) = \log(k) + \alpha + \beta \log(\text{Real GDP}) + \epsilon.$$

Thus, in the new regression with Energy measured in BTUs:

$$\hat{\alpha}^* = \hat{\alpha} + \log(k), \quad \hat{\beta}^* = \hat{\beta}.$$

Answer: The slope coefficient $\hat{\beta}$ is unchanged because scaling the dependent variable by a constant does not affect the relationship between log(Energy) and log(Real GDP). However, the intercept $\hat{\alpha}$ increases by $\log(60) \approx 4.094$, which accounts for the change in measurement units.

Problem 12

The data, Earnings and Height.csv, consists of earnings, height, and other characteristics of a random sample of U.S. workers, taken from the US National Health Interview Survey for 1994. It is a subset of the data used in Anne Case and Christina Paxson's paper "Stature and Status: Height, Ability, and Labor Market Outcomes," *Journal of Political Economy*, 2008,116(3): 499–532. The dataset contains information on **17,870** workers.

(a) (1 point) What is the median value of height in the sample?

Solution (Part a):
The median height is **67 inches**.

Listing 21: R Code: Summary and Median Height

```
library(readxl)

file_path <- "../data/Earnings_and_Height.xlsx"
data <- read_excel(file_path)

cat("Rows:", nrow(data), "  Columns:", ncol(data), "\n")
str(data)
cat("\nSummary of 'height' column:\n")
print(summary(data$height))

# compute overall height
median_height <- median(data$height)
cat("\nMedian height (overall):", median_height, "inches\n")
```

Listing 22: Console Output: Summary and Median Height

```
> source("A1-Q12a.R")
Rows: 17870   Columns: 11
tibble [17,870 x 11] (S3: tbl_df/tbl/data.frame)
 $ sex       : num [1:17870] 0 0 0 0 0 0 0 0 0 0 ...
 $ age       : num [1:17870] 48 41 26 37 35 25 29 44 50 38 ...
 $ mrd       : num [1:17870] 1 6 1 1 6 6 1 4 6 1 ...
 $ educ      : num [1:17870] 13 12 16 16 16 15 16 18 14 12 ...
 $ cworker   : num [1:17870] 1 1 1 1 1 1 1 3 2 4 ...
 $ region    : num [1:17870] 3 2 1 2 1 4 2 4 3 3 ...
 $ race      : num [1:17870] 1 1 1 1 1 1 1 1 1 1 ...
 $ earnings  : num [1:17870] 84055 14021 84055 84055 28560 ...
 $ height    : num [1:17870] 65 65 60 67 68 63 67 65 67 66 ...
 $ weight    : num [1:17870] 133 155 108 150 180 101 150 125 129 110
   ...
 $ occupation: num [1:17870] 1 1 1 1 1 1 1 1 1 1 ...

Summary of 'height' column:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  48.00   64.00   67.00   66.96   70.00   84.00

Median height (overall): 67 inches
```

(b) Do the following – i. (1 point) Estimate average earnings for workers whose height is at most 67 inches. ii. (1 point) Estimate average earnings for workers whose height is greater than 67 inches. iii. (2 points) On average, do taller workers earn more than shorter workers? How much more? What is a 95% confidence interval for the difference in average earnings?

Solution (Part b):

(i) Average earnings for workers with height $\leq$ 67:
Average earnings for workers with height <= 67: **44488.44**

Listing 23: R Code: Average earnings for height $\leq$ 67

```
library(readxl)

# Load dataset
file_path <- "../data/Earnings_and_Height.xlsx"
data <- read_excel(file_path)

# Avg. earnings
short_workers <- subset(data, height <= 67)
avg_earnings_short <- mean(short_workers$earnings)
cat("Average earnings for workers with height <= 67: $", round(avg_
    earnings_short, 2), "\n")
```

Listing 24: Console Output

```
> source("A1-Q12b1.R")
Average earnings for workers with height <= 67: $ 44488.44
```

(ii) Average earnings for workers with height $>$ 67:
Average earnings for workers with height $>$ 67: **49987.88**

Listing 25: R Code: Average earnings for height $>$ 67

```
library(readxl)

# Load dataset
file_path <- "../data/Earnings_and_Height.xlsx"
data <- read_excel(file_path)

# Avg. earnings
tall_workers <- subset(data, height > 67)
avg_earnings_tall <- mean(tall_workers$earnings)
cat("Average earnings for workers with height > 67: $", round(avg_
    earnings_tall, 2), "\n")
```

Listing 26: Console Output

```
> source("A1-Q12b2.R")
Average earnings for workers with height > 67: $ 49987.88
```

(iii) Difference in earnings and 95% confidence interval:
According to the findings, employees who are taller than 67 inches make an average $49,987.88, while shorter workers earn on average $44,488.44. This gives an estimated difference of $5,499.44 in favor of taller workers.
The Welch two-sample t-test strongly rejects the null hypothesis of equal means ($p < 2.2 \times 10^{-16}$). The 95% confidence interval for the difference in average earnings is [$4,706.24, $6,292.64], which does not include zero. This shows that taller people have a statistically significant and economically significant higher earnings.

Listing 27: R Code: Difference and 95% CI

```r
library(readxl)
# Load dataset
file_path <- "../data/Earnings_and_Height.xlsx"
data <- read_excel(file_path)

# divide in two sub datase0st
tall <- subset(data, height > 67)$earnings
short <- subset(data, height <= 67)$earnings

# calculating avg. of two datasets
mean_tall <- mean(tall)
mean_short <- mean(short)
diff_means <- mean_tall - mean_short

cat("Average earnings (tall):   $", round(mean_tall, 2), "\n")
cat("Average earnings (short):  $", round(mean_short, 2), "\n")
cat("Difference (tall - short): $", round(diff_means, 2), "\n")

# 95% CI for difference in means
t_test <- t.test(tall, short, var.equal = FALSE)  # Welch's t-test
print(t_test)
```

Listing 28: Console Output

```
> source("A1-Q12b3.R")
Average earnings (tall):   $ 49987.88
Average earnings (short):  $ 44488.44
Difference (tall - short): $ 5499.44

        Welch Two Sample t-test

data:  tall and short
t = 13.59, df = 16624, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 4706.237 6292.643
sample estimates:
mean of x mean of y
 49987.88  44488.44
```

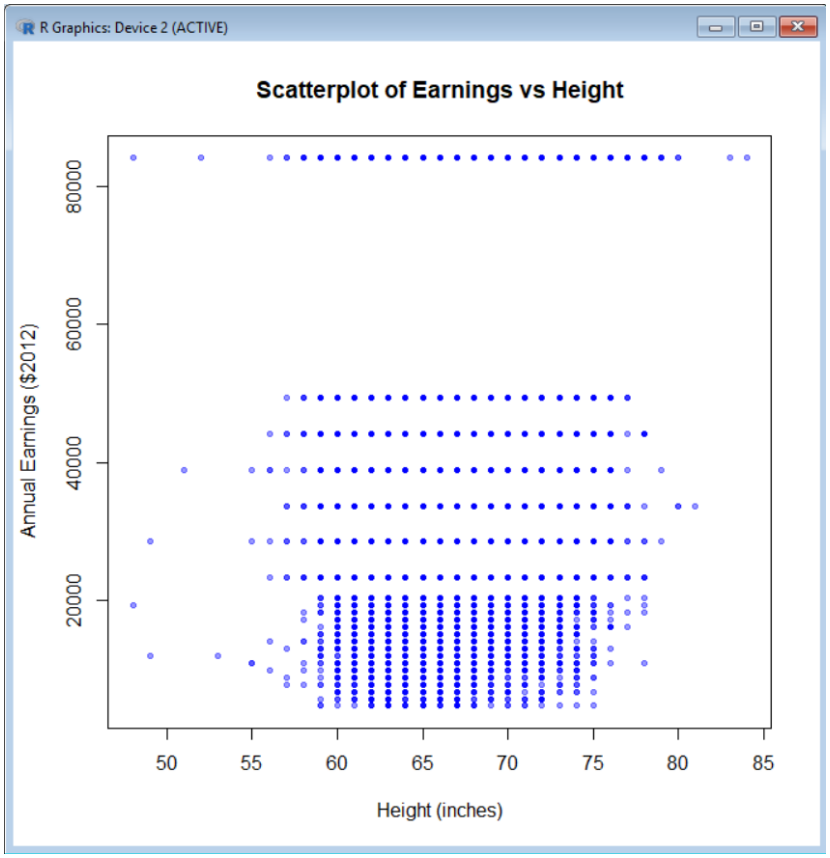
(c) (3 points) Construct a scatterplot of annual earnings (Earnings) on height (Height). Notice that the points on the plot fall along horizontal lines.

Solution (Part c):

Listing 29: R Code: Scatterplot of Earnings vs Height

```r
library(readxl)
# Load dataset
file_path <- "../data/Earnings_and_Height.xlsx"
```

```r
data <- read_excel(file_path)

# Scatterplot (earnings vs height)
plot(data$height, data$earnings,
     xlab = "Height (inches)",
     ylab = "Annual Earnings ($2012)",
     main = "Scatterplot of Earnings vs Height",
     pch = 20, col = rgb(0,0,1,0.4))
```



Explanation: The points fall along horizontal lines because the earnings variable is recorded in discrete dollar amounts (not continuous). Many workers have the same earnings, so multiple points with different heights share the same y-value, creating horizontal alignments in the scatterplot.

(d) Run a regression of Earnings on Height. i. (1 point) What is the estimated slope? ii. (1 point) Use the estimated regression to predict earnings for a worker who is 67 inches tall, for a worker who is 70 inches tall, and for a worker who is 65 inches tall.

Solution (Part d):

Listing 30: R Code: A1-Q12d.R

```r
# A1-Q12d.R
```

```r
slope_inches <- coef(model)[2]
slope_cm <- slope_inches / 2.54
intercept_cm <- coef(model)[1]
R2 <- summary(model)$r.squared
SER <- summary(model)$sigma

cat("Estimated slope (beta_1):", round(slope_inches, 4), "\n")

new_heights <- data.frame(height = c(67, 70, 65))
new_heights$Predicted_Earnings <- predict(model, newdata = new_heights
    )
print(new_heights)
```

Listing 31: Console Output

```
> source("A1-Q12d.R")
Estimated slope (beta_1): 707.6716
  Height Predicted_Earnings
1     67           46901.26
2     70           49024.28
3     65           45485.92
```

(e) Suppose height were measured in centimetres instead of inches. Answer the following questions about the Earnings on Height (in cm) regression. i. (2 points) What is the estimated slope of the regression? ii. (1 points) What is the estimated intercept? iii. (1 points) What is the R^2? iv. (1 points) What is the standard error of the regression?

Solution (Part e):

Listing 32: R Code: Regression in Centimetres

```r
slope_inches <- coef(model)[2]
slope_cm <- slope_inches / 2.54
intercept_cm <- coef(model)[1]
R2 <- summary(model)$r.squared
SER <- summary(model)$sigma

cat("Slope (in cm):", round(slope_cm, 4), "\n")
cat("Intercept:", round(intercept_cm, 2), "\n")
cat("R^2:", round(R2, 4), "\n")
cat("Standard Error of Regression:", round(SER, 2), "\n")
```

Listing 33: Console Output

```
> source("A1-Q12e.R")
Slope (in cm): 278.6108
Intercept: -512.73
R^2: 0.0109
Standard Error of Regression: 26777.24
```

(f) Run a regression of Earnings on Height, using data for female workers only. i. (2 points) What is the estimated slope? ii. (3 points) A randomly selected woman is 1 inch taller than

the average woman in the sample. Would you predict her earnings to be higher or lower than the average earnings for women in the sample? By how much?

Solution (Part f):

Listing 34: R Code: Female Earnings on Height

```
library(readxl)
# Load dataset
file_path <- "../data/Earnings_and_Height.xlsx"

data <- read_excel(file_path)

# Filter to females only (sex = 0)
female_data <- subset(data, sex == 0)

# Regression: Earnings on Height
model_female <- lm(earnings ~ height, data = female_data)
summary(model_female)

# Extract slope
slope <- coef(model_female)[2]
cat("Estimated slope (beta_1):", round(slope, 4), "\n")

# Prediction for "1 inch taller than average woman"
mean_height_female <- mean(female_data$height)
mean_earnings_female <- mean(female_data$earnings)

# Predicted earnings at mean height
pred_mean <- predict(model_female,
                     newdata = data.frame(height = mean_height_female)
                        )

# Predicted earnings at mean height + 1
pred_plus1 <- predict(model_female,
                      newdata = data.frame(height = mean_height_female
                          + 1))

diff <- pred_plus1 - pred_mean

cat("\nAverage female height:", round(mean_height_female, 2), "inches\
    n")
cat("Average female earnings (sample): $", round(mean_earnings_female,
     2), "\n")
cat("Predicted increase in earnings for +1 inch:", round(diff, 2), "\n
    ")
```

Listing 35: Console Output

```
> source("A1-Q12f.R")
Estimated slope (beta_1): 511.2222

Average female height: 64.49 inches
```

```
Average female earnings (sample): $ 45621
Predicted increase in earnings for +1 inch: 511.22
>
```

Explanation:
i. The estimated slope of the regression for female workers is **511.22** dollars per inch. This means that, on average, for each additional inch of height, a woman's earnings are predicted to increase by **$511.22**.
ii. A woman who is 1 inch taller than the average female height (**64.49 inches**) would be predicted to earn **higher** than the average female earnings. The predicted increase in earnings is approximately **$511.22**. Thus, height has a positive association with earnings among female workers, just like in the overall sample.

(g) (5 points) Do you think that height is uncorrelated with other factors that cause earning? That is, do you think that the regression error term, ui has a conditional mean of 0 given Height (Xi)?

Solution (Part g):
The assumption of regression demands:

$$E[u_i \mid \text{Height}_i] = 0$$

That is, height should be uncorrelated with left-out determinants of wages.
In reality, this is unlikely to be so. Height can be related to:

- Child health and nutrition, which, in turn, influence productivity.
- Socioeconomic status (families with more wealth have both taller kids and better-paying jobs).
- Discrimination and social judgments, whereby taller people are seen to be more authoritative or competent.

Thus, it is reasonable that height is related to unobserved determinants of earnings, so the regression error term does not have a zero conditional mean. This suggests that the OLS estimate of could be biased because of omitted variable bias.