Homework 1 (88 pts)

Part 1: Overview of Dataset

This homework assignment is based on a project on life expectancy research.

Background

Life expectancy is a measure of the average lifespan of a person born into a nation. We plan to examine the factors that can increase or decrease life expectancy and their interactions. There are almost 200 nations in the world, each with unique living conditions, economic status, and healthcare treatments. Our chosen data set contains data for 178 countries, over the 15-year period from 2000-2015, as well as a variety of health-related and economic data about the nation for each year. It is important to note, some nations were excluded from the dataset because of repeated missing values due to difficulty finding data from smaller nations such as Togo, Vanuatu, and Cape Verde. The dataset contains four main groupings of predictor variables immunizations, mortalities, economic, and social. Data has been collected from multiple resources: the national health and economic dataset have been collected from Kaggle, and the number of medical professionals (doctors, nurses, and pharmacists, per 10,000 people in the nation) have been collected from the WHO.

- https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who
  - It contains life expectancy data for many countries from 2000 - 2015, as well as information about health factors and economic data
- https://www.who.int/data/gho/data/indicators/indicator-details/GHO/medical-doctors-(per-10-000-population)
- https://www.who.int/data/gho/data/indicators/indicator-details/GHO/pharmacists-(per-10-000-population), https://www.who.int/data/gho/data/indicators/indicator-details/GHO/nursing-and-midwifery-personnel-(per-10-000-population)

Variable Description

Variable Notation	Variable name	Variable Type	Description
X1	Medical doctors	Continuous	Number of doctors per 10000 persons.
X2	Nurses and Midwives	Continuous	Number of nurses and midwives per 10000 persons.
X3	Pharmacists	Continuous	Number of pharmacists per 10000 persons.
Y	Life Expectancy	Continuous	The life expectancy of the country in the year.

Part 2：Homework Types and Format Requirements

There are two kinds of problems, conceptual and application. The conceptual problem focuses on definition, notation, and formula. For this kind of problem, you are supposed to compute by hand (or basic arithmetic function in Excel or R), but not the function that directly shows the answer. Formula and working progress should be clearly shown. By default, all questions in the homework assignment are of this type.

The application problem focuses on R application skill and output interpretation. This problem usually contains the phrase "use R….", or "according to the R output". For this kind of problem, you don't need to compute the results by hand. Instead, get the result from R and proceed.

For instance, in Homework 1, Problem 1-3 are conceptual problems, and Problem 4 is application problem.

For Problems 1 to 3, you may use Excel or R to compute the residuals and sum of squares, means for the variables before computing the residual standard error. When computing the item, show the formula and detail and use the correct notation. You may not use the linear regression function, such as lm() to compute the numbers because the purpose of these problems is to get familiar with the formula and notation.

For Problem 4, you may use the linear regression function such as lm() to run the analysis, the purpose is to be familiar with the R output.

Part 3. Homework Problems

In this homework, we consider a simple linear regression $Y \sim X$, where $X = X_1$, the number of medical doctors.

Problem 1 (10 pts)

Estimate the parameters $(\beta_0, \text{and } \beta_1)$ for a linear regression to predict Y based on X. Complete the following with details.

(a) [1 pt] $\bar{X} = \frac{3460.34}{178} = \mathbf{19.440}$
(b) [1 pt] $\bar{Y} = \frac{12{,}752.2}{178} = \mathbf{71.642}$
(c) [1 pt] $\Sigma(X_i - \bar{X})(Y_i - \bar{Y}) = \mathbf{18{,}444.76}$
(d) [1 pt] $\Sigma(X_i - \bar{X})^2 = \mathbf{52{,}922.52}$
(e) [1 pt] $b_0 = \bar{Y} - b_1\bar{X} = \mathbf{64.8672}$
(f) [1 pt] $b_1 = \mathbf{0.3485}$
(g) [1 pt] $SSE = \Sigma\left(Y_i - \widehat{Y_i}\right)^2 = \mathbf{5{,}219.176}$
(h) [1 pt] $MSE = \frac{5{,}219.176}{178-2} = \mathbf{29.654}$
(i) [1 pt] $SST = \Sigma(Y_i - \bar{Y})^2 = \mathbf{11{,}647.61}$
(j) [1 pt] Verify that $SST = SSR + SSE$, where $SSR = \Sigma\left(\bar{Y} - \hat{Y}_i\right)^2$
$\mathbf{SSR = 6{,}428.438}$ **and** $\mathbf{SSE = 5{,}219.176 \Rightarrow SSR + SSE = 6{,}428.438 + 5{,}219.176 = 11{,}647.61}$

Problem 2 (8 pts)

In order to estimate the linear impact of X on Y, at a confidence of $(1-\alpha)\%$, you should use the critical value, or the t value denoted as $\boldsymbol{t\left(1-\frac{\alpha}{2}, n-2\right) = t\left(1-\frac{\alpha}{2}, 176\right)}$, which has a value of **1.654** (use basic R function or Excel for the exact value), at $\alpha = 0.1$, and **1.974** at $\alpha = 0.05$. The standard error of the estimation $s\{b_1\} = \sqrt{\frac{MSE}{\sum(X_i - \bar{X})^2}}$ (formula) = **0.0237** (value). The margin error, or $t * SE$, of the confidence interval is **0.0392** at $\alpha = 0.1$, and **0.0468** at $\alpha = 0.05$.

Problem 3 (10 pts)

Perform a hypothesis test on the linear impact of X on Y, with a T test with a significant value of 0.1.

Note:

- If a question doesn't specify the hypothesized value, it is two-sided test against 0.
- All hypothesis problem should include the following component: Ho/Ha defined in symbols (β, μ etc.), test statistic (notation and formulas), reject region defined on a critical value (p-value computed on a probability formula), and conclusion.

Hypotheses: $H_0: \beta_1 = 0$ vs $H_a: \beta_1 \neq 0$

Test Statistic: $t_s = \frac{b_1}{s\{b_1\}} = \frac{0.3485}{0.0237} = 14.70$

Rejection Region: Reject H_0 if $|t_s| > t\left(1-\frac{\alpha}{2}, 176\right) = 1.654$

P-value: 0

Problem 4 (6 pts)

Use R to obtain a summary of this SLR model. Highlight the following concepts on the output, the notation, the values, and finally an interpretation. Compute the item with R, or Excel if it is not directly available in the R model summary output.

For example, the point estimate of linear impact of X on Y

```
Coefficients:
             Estimate
(Intercept) 6.566373
x           0.037756
```

The point estimate of linear impact of X on Y: $\hat{\beta} = b_1 = 0.037756$, it means when X is increased by 1 unit, Y is increased by 0.037756 unit. It measures the linear impact of X on Y through the SLR model.

(a) [1 pt] The standard error of the point estimate of the linear impact of X on Y

```
Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-14.1102  -3.5062   0.4287   4.0203  11.7057

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 64.86623    0.61511  105.45   <2e-16 ***
x            0.34852    0.02367   14.72   <2e-16 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.446 on 176 degrees of freedom
Multiple R-squared:  0.5519,    Adjusted R-squared:  0.5494
F-statistic: 216.8 on 1 and 176 DF,  p-value: < 2.2e-16
```

This is a measure of the amount that the point estimate could deviate from the true parameter value.

(b) [1 pt] The residual standard error

```
Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-14.1102  -3.5062   0.4287   4.0203  11.7057

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 64.86623    0.61511  105.45   <2e-16 ***
x            0.34852    0.02367   14.72   <2e-16 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.446 on 176 degrees of freedom
Multiple R-squared:  0.5519,    Adjusted R-squared:  0.5494
F-statistic: 216.8 on 1 and 176 DF,  p-value: < 2.2e-16
```

This is a measure of how much observed values deviate from predicted values

(c) [1 pt] The degree of freedom of the residual (the interpretation of this concept will be covered later)

```
Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-14.1102  -3.5062   0.4287   4.0203  11.7057

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 64.86623    0.61511  105.45   <2e-16 ***
x            0.34852    0.02367   14.72   <2e-16 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.446 on 176 degrees of freedom
Multiple R-squared:  0.5519,    Adjusted R-squared:  0.5494
F-statistic: 216.8 on 1 and 176 DF,  p-value: < 2.2e-16
```

This is related to sample size.

(d) [1 pt] The mean square of the standard error

```
Analysis of Variance Table

Response: y
           Df Sum Sq Mean Sq F value    Pr(>F)
x           1 6428.4  6428.4  216.78 < 2.2e-16 ***
Residuals 176 5219.2    29.7
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Again, a measure of how much the observed responses deviate from the predicted responses.

(e) [2 pts] The standard deviation of the dependent variable Y, **denoted by s_y,** and briefly explain how it is related to the *total sum of variance*, SST=$\Sigma(Y_i - \bar{Y})^2$

$$s_y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2} = \sqrt{\frac{1}{n-1}SST} = \sqrt{\frac{1}{n-1}(SSR + SSE)} = \sqrt{\frac{1}{177}(6428.4 + 5219.2)} = 8.1121$$

```
Analysis of Variance Table

Response: y
           Df Sum Sq Mean Sq F value    Pr(>F)
x           1 6428.4  6428.4  216.78 < 2.2e-16 ***
Residuals 176 5219.2    29.7
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Problem 5 – True/False and Multiple Choice (6 pts)

(a) [1 pt] The tendency, or the form by which of the response variable, Y, varies with X can be estimated with a linear function **True** (T/F). The linear function has a true form of $\beta_0 + \beta_1 X$ in the population domain.

(b) [1 pt] At a general $X = X_h$ level, the predicted value is estimated by $\beta_0 + \beta_1 X_h$. Both β_0 and β_1 are variables and can be estimated by b_0 and b_1 on a sample **False** (T/F)

(c) [1 pt] The deviation between the actual response variable Y and the predicted Y, or $\hat{Y}$ at a given $X = X_h$ level is called the random error and is denoted by $\boldsymbol{\varepsilon}$ (ε / e), which can be estimated with a value denoted by $\boldsymbol{e}$ (ε / e) in a sample.

(d) [2 pts] This random error is assumed to have a distribution of $N(0, \sigma^2)$ **True** (T/F), where the standard deviation, σ, can be estimated by the standard error term denoted by $\boldsymbol{s}$ (s / s_y) computed from a sample.

(e) [1 pt] The actual response variable, $Y = \beta_0 + \beta_1 X + \varepsilon$, represents the linear relationship between X and Y. The two "ingredients" in this relationship can be identified as **B ($\boldsymbol{\beta_0 + \beta_1 X}$ and $\boldsymbol{\varepsilon}$).**

A. β_0 and β_1 B. $\beta_0 + \beta_1 X$ and ε C. X and Y

For the following problems, we use the life expectancy data and consider a simple linear regression $Y \sim X$, where $X = X_2$ is the number of nurses and midwives, and Y is the life expectancy.

In the confidence interval problems, note that components in a confidence interval include the point estimate, the critical value, the standard error, and the margin error. The result should be computed toward the end.

For example, compute a CI as $5 \pm 2 * 4 = 5 \pm 8 = (-3, 13)$

In computation problems, a basic rule is that you keep 3 or more significant decimal places for numbers during the working period and keep 2 or more significant decimal places at the number reported at the end.

In the fill-in-the-blank question, when denote or write the formula for a term, show **both the general and the specific form based on the question**. Remember to **fill your answer in the blanks or above the line** to be graded properly.

For example, the critical value for a one-sided t-test, $H_0\ \mu = 0, H_a: \mu > 0$ is denoted by $t(1 - \alpha, n - 1) = t(0.95, 30)$.
The test statistic, t_s, can be computed with a formula $t_s = \frac{\bar{Y}}{s/\sqrt{n}} = \frac{10}{20/\sqrt{25}}$, and a value of 2.5, where the general form is $t_s = \frac{\bar{Y}}{s/\sqrt{n}}$, and the specific form is $\frac{10}{20/\sqrt{25}}$,
The p value can be computed with a formula $\Pr(t > t_s \mid \mu_0 \text{ is true}) = \Pr(t > 2.5, \text{ given } \mu_0 = 0)$, where the general form is $\Pr(t > t_s \mid \mu_0 \text{ is true})$ and the specific form is $\Pr(t > 2.5, \text{ given } \mu_0 = 0)$,

Problem 6 (16 pts)

Use a significance level of 0.05, or confidence level of 0.95, and suppose the prediction is made at $X_h = \text{the mean of } X, \text{or } \bar{X}$. Complete the confidence interval questions. (1 pt each blank, no partial credit)

(a) [4 pts] To estimate the mean response value of Y, the point estimate can be estimated as $\widehat{Y} = b_0 + b_1\bar{X} = 66.067 + 0.118 \times 47.106$ (both the general formula and specific formula in this question)= **71.626** (computed as this value).

The standard error of this estimation is denoted as the formula $\sqrt{MSE\left(\frac{1}{n} + \frac{(X_h-\bar{X})^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}\right)} = \sqrt{37.2\left(\frac{1}{178} + \frac{0}{363{,}704.8}\right)}$ (both the general formula and specific formula in this question) and computed as **0.4572** (computed as the value). The t-value is denoted by $t\left(1-\frac{\alpha}{2}, n-2\right) = t\left(1-\frac{0.05}{2}, 176\right)$ (both the general formula and specific formula in this question) and computed as **1.974** (computed as this value).

(b) [4 pts] To predict the single response (the next observation value), the point estimate can be estimated as $\widehat{Y} = b_0 + b_1\bar{X} = 66.067 + 0.118 \times 47.106$ (both the general formula and specific formula in this question)= **71.626** (computed as this value).

The standard error of this estimation is denoted $\sqrt{MSE\left(1 + \frac{1}{n} + \frac{(X_h-\bar{X})^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}\right)} = \sqrt{37.2\left(1 + \frac{1}{178} + \frac{0}{363{,}704.8}\right)}$ (both the general formula and specific formula in this question)= **6.1163** (computed as the value). The t-value is denoted by $t\left(1-\frac{\alpha}{2}, n-2\right) = t\left(1-\frac{0.05}{2}, 176\right)$ (both the general formula and specific formula in this question) = **1.974** (computed as this value).

(c) [4 pts] To predict the mean of 3 responses (the average of the next m observation values, where X_h the same for all m observations), the point estimate can be estimated as $\widehat{Y} = b_0 + b_1\bar{X} = 66.067 + 0.118 \times 47.106$ (both the general formula and specific formula in this question)= **71.626** (computed as this value, as m=3).

The standard error of this estimation is denoted _____(both the general formula and specific formula in this question)= $\sqrt{MSE\left(\frac{1}{m} + \frac{1}{n} + \frac{(X_h-\bar{X})^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}\right)} = \sqrt{37.2\left(\frac{1}{3} + \frac{1}{178} + \frac{0}{363{,}704.8}\right)}$ (computed as the value). The t-value is denoted by **3.5509** (both the general formula and specific formula in this question) = **1.974** (computed as this value).

(d) [2 pts] Answer this question without computation, when estimate the mean response value given the $X_h = \text{median of } X$, the corresponding standard error is **bigger** (bigger than/smaller than/the same) when $X_h = \text{mean of } X$, **because the mean is not the same as the median.**

(e) [2 pts] Answer this question without computation, when estimating the mean of 10 responses, the corresponding standard error is **smaller** (bigger than/smaller than/the same as) when estimating the mean of 3 responses at the same X level, because **the standard error decreases as the number of observations increases**.

Problem 7 – Compare the ANOVA F-test and the T-test on the significance of a SLR model (19 pts)

We know that **both ANOVA F test and T-test can be used to address the significance of the linear impact of X on Y,** $H_0: \beta_1 = 0, H_a: \beta_1 \neq 0$. We have completed the T-test in the previous questions, now complete an ANOVA F-test.

Note: **the significance of a linear model**. In the simple linear regression model (SLR) with only one X, the test on the significance linear impact is equivalent to the test on the **significance of the linear model**. In the multiple linear regression model (MLR) with multiple Xs, the test on individual linear impact cannot imply the significance of the linear model.

The significance of the linear model, $H_0: \beta_1 = \beta_2 \ldots = \beta_k = 0, H_a,$ at least one β is different, can be tested with the Global F test, or the ANOVA F test on the entire model, $F_s = MSR/MSE$

(a) [12 pts] Perform a global ANOVA F test for the significance of the SLR model. The significance of the SLR model can be defined in symbols:

$H_0: \beta_1 = 0, \qquad Ha: \beta_1 \neq 0$

Or define in the following statements.

H_0: The predictor X has no impact on Y in a linear model $Y \sim X$.
H_0: The predictor X can be dropped from the linear model $Y \sim X$.

The Global F test computes the test statistic as the ration of MSR to the MSE in the regression model.

```
Call:
lm(formula = y ~ x2)

Residuals:
     Min       1Q   Median       3Q      Max
-16.0640  -3.7798   0.2097   4.6965  13.4882

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 66.06723    0.66064  100.00   <2e-16 ***
x2           0.11834    0.01012   11.69   <2e-16 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.103 on 176 degrees of freedom
Multiple R-squared:  0.4373,    Adjusted R-squared:  0.4341
F-statistic: 136.8 on 1 and 176 DF,  p-value: < 2.2e-16
```

```
Analysis of Variance Table

Response: y
           Df Sum Sq Mean Sq F value    Pr(>F)
x2          1 5093.0  5093.0  136.75 < 2.2e-16 ***
Residuals 176 6554.6    37.2
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

SST can be computed with the formula: $\sum_{i=1}^{n}(y_i - \bar{y})^2$ and a value of **11647.61**, the degree of freedom is computed with the formula: $n - 1$, and a value of **177**.

SSE can be computed with the formula: $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$, and a value of **6554.586**, the degree of freedom is computed with the formula: $n - k$, and a value of **176**.

SSR can be computed with the formula: $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$, and a value of **5093.026**, the degree of freedom is computed with the formula: $k - 1$, and a value of **1**.

(b) [2 pts] Compute the test statistic for the F test, F_s. The formula is $F_s = \frac{MSR}{MSE}$ and has a value of $\frac{\frac{5093.026}{1}}{\frac{6554.586}{176}} = 136.755$.

How is it related to the t_s in Problem 3, part (a) and Problem 3, part (b)?

F_s is used to test for significance of the full model, both linear relationship and significance of the slopes, while t_s only tests for the significance of the slope.

It is also fine if students recognize that X_1 was used in Problem 3 and X_2 was used in Problem 7. If they say it doesn't make sense to compare them or something like that, they can get full credit.

(c) [2 pts] the critical value of the F-test can be denoted by $F(1-\alpha, k-1, n-k) = F(0.95, 1, 176)$ and has a value of **3.8948**.

(d) [3 pts] Compute the p-value for the F-test with the formula $P(F_s > f_s) = P(F_s > 136.755)$, and the value of **0**. It is the **same as** (same as / different from) the p-value for the t-test in Problem 3.

Again, if students recognize that the models used different predictors and say the p-value is different, then that is fine as well and they should get full credit.

Problem 8 – Compare the ANOVA F test and the GLT test on the significance of the SLR model, i.e. $H_0: \beta_1 = 0, H_a: \beta_1 \neq 0$ (13 pts)

(a) [2 pts] General lineal test (GLT) constructs and then compares two models that establish under Ho and Ha. Specifically, the full model is established under H_a (Ho/Ha), and the reduced model is established under H_0 (Ho/Ha)

(b) [4 pts] The total error in the full model, SSE(Full) or SSE(F) has a value of **6554.586** and a degree of freedom of **176**. The total error in the reduced model, SSE(Reduced) or SSE(R) has a value of **11647.61** and a degree of freedom of **177**

(c) [7 pts] Discuss the connection between the Global F test and the GLT F test in the following perspectives.

(i) The null and alternative hypothesis

The GLT uses an F-test to analyze the variances in the different models and the Global F-test is the significant test of the SLR model to find the significant linear impact of an independent model (can sometimes be replaced by a t-test). The GLT uses the reduced model for the null hypothesis and the full model for the alternative hypothesis, while the Global F-test is $H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$ vs H_a: at least one β is different.

(ii) The test statistic

The test statistics from the GLT in simple linear regression (SLR) is identical to an ANOVA test statistic (from the Global F-test).

(iii) Situations when the two methods are equivalent, and situations when only GLT T test is appropriate.

For SLR, the two methods are equivalent. However, with multiple linear regression (MLR), the two methods are not equivalent and the GLT would be used.