**Simpler Linear Model**

The intercept (B1) in the regression model represents the estimated average life expectancy when alcohol consumption is zero, and it is calculated to be 67.73 years. The coefficient for Alcohol (B2) indicates that for every one-unit increase in alcohol consumption, there is an expected increase of 0.9441 years in life expectancy, on average. The statistical significance of the Alcohol coefficient is confirmed by its very small p-value (2.5e-11), providing strong evidence to reject the null hypothesis that the Alcohol coefficient is zero. This suggests that the Alcohol variable is statistically significant in predicting life expectancy.

Table : Simpler Linear Model

|  |  |
| --- | --- |
|  | (1) |
| (Intercept) | 67.731 |
|  | (0.756) |
| Alcohol | 0.944 |
|  | (0.132) |
| Num.Obs. | 171 |
| R2 | 0.232 |
| R2 Adj. | 0.228 |
| AIC | 1161.1 |
| BIC | 1170.5 |
| Log.Lik. | -577.543 |
| F | 51.123 |
| RMSE | 7.09 |

**Goodness of Fit of the Model**

The model's goodness of fit is assessed through the R-squared value, which is 0.2322. This value signifies that approximately 23.22% of the variance in life expectancy can be explained by the linear relationship with alcohol consumption. Additionally, the adjusted R-squared, at 0.2277, indicates that the model is not overfitting and provides a reasonable fit to the data.

**Potential Bias in the Alcohol Coefficient**

A potential source of bias in the alcohol coefficient arises from the simplicity of the model. The model assumes a linear relationship between alcohol consumption and life expectancy while not considering other potentially influential factors such as BMI, Total\_Expenditure, or Schooling. Omitting these relevant variables may lead to biased estimates of the true impact of alcohol on life expectancy. To mitigate this bias, a more comprehensive approach involving a multiple regression model with additional predictors could provide a more accurate understanding of the relationship, accounting for potential confounding factors.

**Multilinear Regression Mode I**

The multiple linear regression model includes three predictor variables: Alcohol, Schooling, and BMI. The intercept (B1) is 42.24690, representing the estimated average life expectancy when Alcohol, Schooling, and BMI are all zero.

Table : Multilinear Regression Mode I

|  |  |
| --- | --- |
|  | (1) |
| (Intercept) | 42.247 |
|  | (1.941) |
| Alcohol | 0.049 |
|  | (0.111) |
| Schooling | 2.054 |
|  | (0.177) |
| BMI | 0.061 |
|  | (0.021) |
| Num.Obs. | 171 |
| R2 | 0.646 |
| R2 Adj. | 0.639 |
| AIC | 1032.8 |
| BIC | 1048.5 |
| Log.Lik. | -511.421 |
| F | 101.456 |
| RMSE | 4.82 |

The coefficients for each predictor are as follows:

* **Alcohol (B2):** The coefficient is 0.04870 with a p-value of 0.66116, suggesting that alcohol consumption is not statistically significant in predicting life expectancy when accounting for Schooling and BMI. The small coefficient indicates a negligible change in life expectancy for a one-unit increase in alcohol consumption.
* **Schooling (B3):** The coefficient is 2.05413 with a very small p-value (< 2e-16), indicating that Schooling is highly statistically significant in predicting life expectancy. For every one-unit increase in Schooling, there is an expected increase of 2.05413 years in life expectancy, holding Alcohol and BMI constant.
* **BMI (B4):** The coefficient is 0.06135 with a p-value of 0.00368, indicating that BMI is statistically significant in predicting life expectancy. For every one-unit increase in BMI, there is an expected increase of 0.06135 years in life expectancy, holding Alcohol and Schooling constant.

**Goodness of Fit**

The R-squared value is 0.6457, suggesting that approximately 64.57% of the variance in life expectancy is explained by the linear relationship with Alcohol, Schooling, and BMI. The adjusted R-squared is 0.6393, which accounts for the number of predictors and indicates a well-fitting model. The residual standard error is 4.873, representing the average deviation of the observed values from the predicted values.

**F-Test for Model Comparison**

The F-test compares the fit of the simple model to the multiple model 1. The p-value (< 2.2e-16) is highly significant, indicating that adding Schooling and BMI to the model significantly improves its explanatory power. This supports the inclusion of Schooling and BMI in the model as they contribute to a better understanding of the factors influencing life expectancy.

Table : F-Test for Model Comparison

| Res.Df | RSS | Df | Sum of Sq | F | Pr(>F) |
| --- | --- | --- | --- | --- | --- |
| 169 | 8592.719 | NA | NA | NA | NA |
| 167 | 3965.213 | 2 | 4627.506 | 97.44667 | 0 |

The multiple linear regression model indicates that Schooling and BMI are significant predictors of life expectancy, while the contribution of Alcohol is not statistically significant. The model, with a high R-squared value, demonstrates a good fit to the data, and the F-test supports the enhanced explanatory power gained by including Schooling and BMI in the model.

**Multilinear Regression Model II**

The multiple linear regression model incorporates Alcohol, Schooling, a quadratic term for Schooling, Schooling squared, and BMI as predictors. The intercept (B1) estimates the average life expectancy when Alcohol, Schooling, and BMI are all zero, yielding a value of 45.89101 years.

Table : Multilinear Regression Model II

|  |  |
| --- | --- |
|  | (1) |
| (Intercept) | 45.891 |
|  | (5.497) |
| Alcohol | 0.034 |
|  | (0.113) |
| Schooling | 1.441 |
|  | (0.883) |
| I((Schooling)^2) | 0.025 |
|  | (0.035) |
| BMI | 0.063 |
|  | (0.021) |
| Num.Obs. | 171 |
| R2 | 0.647 |
| R2 Adj. | 0.638 |
| AIC | 1034.3 |
| BIC | 1053.2 |
| Log.Lik. | -511.162 |
| F | 75.990 |
| RMSE | 4.81 |

The coefficient for Alcohol (B2) is 0.03353, and its non-significant p-value of 0.7673 suggests that alcohol consumption does not have a statistically significant impact on life expectancy when accounting for the other variables.

The linear effect of Schooling (B3) is represented by a coefficient of 1.44135. Although positive, the p-value of 0.1044 implies that the linear relationship between Schooling and life expectancy is not statistically significant at the conventional significance level of 0.05.

The quadratic effect of Schooling, Schooling squared has a coefficient of 0.02452, and its non-significant p-value (0.4795) suggests that a quadratic relationship does not significantly contribute to explaining life expectancy in this model.

On the other hand, BMI (B5) has a statistically significant coefficient of 0.06261 (p-value: 0.0032), indicating that an increase in BMI is associated with a corresponding increase in life expectancy.

Regarding the goodness of fit, the R-squared value of 0.6468 signifies that approximately 64.68% of the variance in life expectancy is explained by the model. The adjusted R-squared, at 0.6383, considers the number of predictors and supports the model's reasonable fit to the data. The residual standard error of 4.88 represents the average deviation between observed and predicted values.

The model underscores the significance of BMI as a predictor of life expectancy, while suggesting that the linear and quadratic effects of Schooling are not statistically significant in this particular context. The model as a whole provides a reasonable explanation for a substantial portion of the variance in life expectancy.

**Comparing Multilinear Regression Model, I and II**  
The comparison of Model I and Model II, considering the multiple R-squared and adjusted R-squared values, indicates a marginal improvement with the addition of the squared term for Schooling in Model II. The multiple R-squared value slightly increases from 0.6457 in Model I to 0.6468 in Model II. However, the adjusted R-squared values remain the same at 0.6383 for both models. Since the adjusted R-squared considers the number of predictors and guards against overfitting, the results suggest that the inclusion of the squared term for Schooling does not significantly enhance the explanatory power of the model beyond Model I. The improvement observed in the multiple R-squared may be due to the increased flexibility in capturing nonlinear relationships but is not supported by the adjusted R-squared, indicating a more conservative assessment of model fit.

**Logarithmic Model**

The logarithmic model provides insights into the relationships between predictors and the log of life expectancy. The intercept (of approximately 3.4008213 represents the expected log of life expectancy when all predictors are zero. However, the coefficient for Alcohol is not statistically significant (p-value = 0.31715), indicating that Alcohol may not have a significant effect on the log of life expectancy. In contrast, both log (Schooling) and BMI exhibit statistically significant positive associations. For each one-unit increase in log (Schooling), there is an expected increase of approximately 0.3211605 in the log of life expectancy, and for each one-unit increase in BMI, there is an expected increase of approximately 0.0010192 in the log of life expectancy.

The goodness of fit metrics further supports the model's efficacy. The residuals show relatively small errors in predictions. The R-squared (0.5967) and Adjusted R-squared (0.5895) values indicate that approximately 59.67% of the variance in the log-transformed life expectancy is explained by the model. The F-statistic (82.37) with an extremely low p-value (< 2.2e-16) signifies the overall statistical significance of the model. This implies that at least one of the predictors significantly influences the log of life expectancy. In conclusion, the logarithmic model underscores the importance of log (Schooling) and BMI in explaining variations in life expectancy, while suggesting that Alcohol may not be a significant predictor in this context.

**Discussion and Conclusion of the superior model**

Each estimated model showcases unique strengths and limitations, tying the choice of a superior model to particular research goals. The simple linear model focused on alcohol consumption establishes a clear-cut relationship with life expectancy; however, its simplicity might disregard other factors that could potentially wield influence.

Particularly, the Multilinear Model II enhanced with a quadratic term for schooling: these multilinear models strive to capture more nuanced relationships; they present an exceptionally flexible framework. However justifiable only if the quadratic element is statistically significant – this augmentation in complexity could pose challenges towards the interpretability of the model.

Transforming life expectancy and predictors into logarithmic scales, the logarithmic model captures proportional relationships with precision; it offers a nuanced perspective--particularly valuable when maneuvering through socioeconomic indicators. Nevertheless, because of its use of the logarithmic scale interpretation does become complex.

The multilinear models and the logarithmic model each possess merits, when we consider the trade-off between model complexity and interpretability along with analysis context. Notably, the logarithmic model excels in capturing proportional relationships which provides a nuanced comprehension of data. Yet, one must balance its advantages against the challenges it presents in terms of interpretability. The robustness of the logarithmic model could become apparent through further validation and testing against additional datasets: thus, ultimately selecting the superior model hinges on specific research objectives - maintaining a balance between model complexity and explanatory power.

**Appendix**

> summary(model\_simple)

Call:

lm(formula = Life\_Expectancy ~ Alcohol, data = df)

Residuals:

Min 1Q Median 3Q Max

-24.278 -3.917 1.060 5.467 13.242

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 67.7305 0.7564 89.55 < 2e-16 \*\*\*

Alcohol 0.9441 0.1320 7.15 2.5e-11 \*\*\*

---

Signif. codes:

0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.131 on 169 degrees of freedom

Multiple R-squared: 0.2322, Adjusted R-squared: 0.2277

F-statistic: 51.12 on 1 and 169 DF, p-value: 2.498e-11

> summary(model\_multiple)

Call:

lm(formula = Life\_Expectancy ~ Alcohol + Schooling + BMI, data = df)

Residuals:

Min 1Q Median 3Q Max

-16.3144 -2.8698 0.3882 3.0997 14.7307

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 42.24690 1.94126 21.763 < 2e-16 \*\*\*

Alcohol 0.04870 0.11092 0.439 0.66116

Schooling 2.05413 0.17730 11.586 < 2e-16 \*\*\*

BMI 0.06135 0.02083 2.946 0.00368 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.873 on 167 degrees of freedom

Multiple R-squared: 0.6457, Adjusted R-squared: 0.6393

F-statistic: 101.5 on 3 and 167 DF, p-value: < 2.2e-16

> print(anova\_result)

Analysis of Variance Table

Model 1: Life\_Expectancy ~ Alcohol

Model 2: Life\_Expectancy ~ Alcohol + Schooling + BMI

Res.Df RSS Df Sum of Sq F Pr(>F)

1 169 8592.7

2 167 3965.2 2 4627.5 97.447 < 2.2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> summary(model\_quadratic)

Call:

lm(formula = Life\_Expectancy ~ Alcohol + Schooling + I((Schooling)^2) +

BMI, data = df)

Residuals:

Min 1Q Median 3Q Max

-16.0643 -2.8445 0.4559 3.0709 13.6225

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 45.89101 5.49711 8.348 2.59e-14 \*\*\*

Alcohol 0.03353 0.11313 0.296 0.7673

Schooling 1.44135 0.88267 1.633 0.1044

I((Schooling)^2) 0.02452 0.03460 0.709 0.4795

BMI 0.06261 0.02093 2.991 0.0032 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.88 on 166 degrees of freedom

Multiple R-squared: 0.6468, Adjusted R-squared: 0.6383

F-statistic: 75.99 on 4 and 166 DF, p-value: < 2.2e-16

> summary(log\_model)

Call:

lm(formula = log(Life\_Expectancy) ~ Alcohol + log(Schooling) +

BMI, data = df)

Residuals:

Min 1Q Median 3Q Max

-0.284775 -0.037745 0.005375 0.046391 0.278812

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 3.4008213 0.0730552 46.551 < 2e-16 \*\*\*

Alcohol 0.0016833 0.0016777 1.003 0.31715

log(Schooling) 0.3211605 0.0315668 10.174 < 2e-16 \*\*\*

BMI 0.0010192 0.0003244 3.142 0.00199 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.07588 on 167 degrees of freedom

Multiple R-squared: 0.5967, Adjusted R-squared: 0.5895

F-statistic: 82.37 on 3 and 167 DF, p-value: < 2.2e-16

**R-codes**

Loading the data

library(readxl)

df<- read\_excel("2013lifeexpectancy.xls")

View(df)

# Simple linear Model

# Fit the simple linear regression model

model\_simple <- lm(Life\_Expectancy ~ Alcohol, data = df)

# Print the summary of the model

summary(model\_simple)

library(modelsummary)

modelsummary(model\_simple, output = "markdown")

# Multilinear regression model I

# Fit the multiple linear regression model

model\_multiple <- lm(Life\_Expectancy ~ Alcohol + Schooling + BMI, data = df)

# Print the summary of the model

summary(model\_multiple)

modelsummary(model\_multiple, output = "markdown")

# F-test

# Fit the simple model without new variables

model\_simple <- lm(Life\_Expectancy ~ Alcohol, data = df)

# Perform ANOVA test to compare models

anova\_result <- anova(model\_simple, model\_multiple)

print(anova\_result)

# Print the ANOVA results using kable

kable(anova\_result, format = "markdown", caption = "ANOVA Test Results for Model Comparison")

# Multilinear regression model II

# Fit the multiple linear regression model with a quadratic term for Schooling

model\_quadratic <- lm(Life\_Expectancy ~ Alcohol + Schooling + I((Schooling)^2) + BMI, data = df)

# Print the summary of the model

summary(model\_quadratic)

modelsummary(model\_quadratic, output = "markdown")

# Logarithmic model

# Fit the logarithmic model

log\_model <- lm(log(Life\_Expectancy) ~ Alcohol + log(Schooling) + BMI, data = df)

# Print the summary of the logarithmic model

summary(log\_model)