Part I

1

1. Yes, X is a statistically significant predictor of Y.

For every X increase, Y increases by 3.43

1. SSR = 672.9797

SSE = 32.22028

SST = SSR + SSE = 705.19998

K = 1

Note: K is the number of independent variable(s) you included in the regression

MSR = 672.9797 / 1 = 672.9797

n = 5

MSE = SSE / (n - k - 1)

= 32.22028 / (5 - 1 - 1)

= 32.22028 / 3

= 10.7400933333

= 10.74

1. F statistics = MSR / MSE

= 672.9797 / 10.7400933333

= 62.660507606

= 62.66

Conclusion: No, F is statistically insignificant.

If F > critical value, we reject the null hypothesis. This case 62.66 > 3.5, therefore the model is statistically insignificant. The variances of the 4 observations are unequal.

|  |  |  |
| --- | --- | --- |
| X | Y | Ŷ |
| 5 | 23 | 20.16667 |
| 5 | 16 | 20.16667 |
| 6 | 25 | 23 |
| 8 | 28 | 28.6667 |

Observation 1 and 2, (x = 5)

Y-Hat = b0 + b1x

= 6 + 2.83 \* 5

= 20.16667

Observation 3 (x = 6)

Y-Hat = b0 + b1x

= 6 + 2.83 \* 6

= 23

Observation 3 (x = 8)

Y-Hat = b0 + b1x

= 6 + 2.83 \* 8

= 28.66667

1. R2 = SSR / SST

= 0.9543

Yes, R Square equals 0.954, which is a very good fit. 95% of the variation in the dependent variable, Y is explained by the independent variable, X.

The closer to 1, the better the regression line (read on) fits the data.

2. It is clear from the scatter plot that as X increases, Y also increases. It seems to be the case that the points follow a linear pattern well, then we say that there is a high linear or linear positive correlation.

3.

1. No, both Experience and Experience squared are statistically insignificant predictors in this model. The reason being that their p-values are higher than 0.05 in this case 0.728582 for Experience and 0.773812 for Experience squared.
2. No, model A is the better model, as its p-value is less than 0.05 and as such it should be used for predictive analytics purposes.
3. The regression line is: Y = 8.670659 + 1.886228 \* Experience + 0.080838 \* Experience squared. In other words, for each increase in Experience, Y increases by 1.886228. For each increase in Experience squared, Y increases by 0.080838. This is valuable information.
4. We cannot interpret the coefficients from model B as casual parameters. Since there are no significant correlation between the dependent and independent variables there might be other variables omitted from our model.

Part II

4. The L.I.N.E assumptions are the four assumptions of linear regression. Namely

* Linearity of residuals

There exists a linear relationship between the independent variable, x, and the dependent variable, y.

* Independence of residuals

The residuals are independent. In particular, there is no correlation between consecutive residuals in time series data.

* Normal distribution of residuals

The residuals of the model are normally distributed.

* Equal variance of residuals

The residuals have constant variance at every level of x.

5.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| *Obs* | *Beer (x)* | *Cigarettes (y)* | *xy* | *x2* | *x̄* | *(x-x̄)* | *(x-x̄)2* |
| 1 | 8 | 16 | 128 | 64 | 5 | 3 | 9 |
| 2 | 6 | 13 | 78 | 36 | 5 | 1 | 1 |
| 3 | 2 | 4 | 8 | 4 | 5 | -3 | 9 |
| 4 | 4 | 9 | 36 | 16 | 5 | -1 | 1 |
| **4** | **20** | **42** | **250** | **120** | **5** |  | **20** |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
| *Obs* | *Beer (x)* | *Cigarettes (y)* | *Y2* | *Ȳ* | *Ŷ* | *(Y-Ŷ)* | *(Y-Ŷ)2* |
| 1 | 8 | 16 | 256 | 10.5 | 16.5 | -0.5 | 0.25 |
| 2 | 6 | 13 | 169 | 10.5 | 12.5 | 0.5 | 0.25 |
| 3 | 2 | 4 | 16 | 10.5 | 4.5 | -0.5 | 0.25 |
| 4 | 4 | 9 | 81 | 10.5 | 8.5 | 0.5 | 0.25 |
| **4** | **20** | **42** | **522** | **10.5** |  |  | **1** |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
| *b1* | *b0* |  |  |  |  |  |  |
| **2** | **0.5** |  |  |  |  |  |  |

Calculations

b0 represents the y intercept

b1 represents the slope of the best fit line

