Simple Linear Regression Solutions

Sanjivani Panse – Data Science 27th Jan weekday batch

**1) Calories Consumed**

a) Iteration I

Correlation Coefficient

# since correleation coefficient is 0.946991, the correlation is strong positive

Summary of Regression model

Residuals:

Min 1Q Median 3Q Max

-158.67 -107.56 36.70 81.68 165.53

Coefficients:

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

(Intercept) -625.75236 100.82293 -6.206 4.54e-05 \*\*\*

`Calories Consumed` 0.42016 0.04115 10.211 2.86e-07 \*\*\*

---

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

Residual standard error: 111.6 on 12 degrees of freedom

Multiple R-squared: 0.8968, Adjusted R-squared: 0.8882

F-statistic: 104.3 on 1 and 12 DF, p-value: 2.856e-07

Slope = 0.42016

p-value for the slope of calories consumed is 2.86e-07 < 0.05. Hence, the slope is significant.

Intercept = -625.75236

p-value for the intercept of calories consumed is 4.54e-05 < 0.05. Hence, the intercept is significant.

R-squared = 0.8968 that is 89.68% which is good

Linear Model is

Weight gained = -625.75236 + 0.42016\*Calories Consumed

RMSE = 103.3025

Since RMSE is high we use transformation of X to X2 that is (Calories Consumed)2

b) Iteration 2

Squaring Calories Consumed

Correlation Coefficient

Correlation Coefficient = 0.9710636 which indicates strong positive correlation

Summary of Regression model

Residuals:

Min 1Q Median 3Q Max

-115.112 -73.801 1.557 74.638 108.083

Coefficients:

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

(Intercept) -1.438e+02 4.194e+01 -3.428 0.005 \*\*

calsq 8.353e-05 5.930e-06 14.085 7.96e-09 \*\*\*

---

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

Residual standard error: 82.95 on 12 degrees of freedom

Multiple R-squared: 0.943, Adjusted R-squared: 0.9382

F-statistic: 198.4 on 1 and 12 DF, p-value: 7.963e-09

Slope = 8.353e-05

p-value for the slope of calories consumed is 96e-09 < 0.05. Hence, the slope is significant.

R-squared = 0.943 that is 94.3% which is good

RMSE = 76.79385

#Conclusion : Since RMSE has reduced and Rsquared is also 94.3% we can say this model is a better fit

Model is

Weight gained = -143.7949 + 0.00008352629\*(Calories Consumed \* Calories Consumed)

**2) Delivery Time**

a) Iteration 1

Correlation Coefficient

# since correleation coefficient is 0.8259973, the correlation is strong positive

Summary of Regression model

Residuals:

Min 1Q Median 3Q Max

-5.1729 -2.0298 -0.0298 0.8741 6.6722

Coefficients:

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

(Intercept) 6.5827 1.7217 3.823 0.00115 \*\*

`Sorting Time` 1.6490 0.2582 6.387 3.98e-06 \*\*\*

---

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

Residual standard error: 2.935 on 19 degrees of freedom

Multiple R-squared: 0.6823, Adjusted R-squared: 0.6655

F-statistic: 40.8 on 1 and 19 DF, p-value: 3.983e-06

Slope = 1.6490

p-value for the slope of calories consumed is 3.98e-06 < 0.05. Hence, the slope is significant.

R-squared = 0.6823 that is 68.23% which is not so good

RMSE

#RMSE is 2.79165 which is good

We will try to improve R-squared and RMSE

b) Iteration 2

## Transform sorting time to log(sorting time)

# correlation coefficient = 0.8339325 shows marginal increase

c) Iteration 3

## Transform delivery time to log(delivery time)

# correlation coefficient = 0.8431773 shows marginal increase

d) Iteration 4

# Still for a better correlation, we take log(sort) and log(delivery time) both

# correlation coefficient = 0.8787271 shows very good increase

Run a linear regression on log(sort) and log(delivery time)

Summary of Regression model

Residuals:

Min 1Q Median 3Q Max

-0.23303 -0.09050 -0.00825 0.08897 0.36439

Coefficients:

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

(Intercept) 1.74199 0.13312 13.086 5.92e-11 \*\*\*

logsort 0.59752 0.07446 8.024 1.60e-07 \*\*\*

---

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

Residual standard error: 0.1558 on 19 degrees of freedom

Multiple R-squared: 0.7722, Adjusted R-squared: 0.7602

F-statistic: 64.39 on 1 and 19 DF, p-value: 1.602e-07

Slope = 0.59752

p-value for the slope of log of sorting time is 1.60e-07 < 0.05. Hence, the slope is significant.

R-squared = 0.7722 that is 77.22% which has improved

RMSE

#RMSE is 2.745829 which has reduced , hence model has improved well

# Final model is

#log(Delivery Time) = 1.74199 + 0.59752 \* log(Sorting Time)

**3) Employee Data**

a) Iteration 1

Correlation Coefficient

# since correleation coefficient is -0.9117216, the correlation is strong negative

Summary of Regression model

Residuals:

Min 1Q Median 3Q Max

-3.804 -3.059 -1.819 2.430 8.072

Coefficients:

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

(Intercept) 244.36491 27.35194 8.934 1.96e-05 \*\*\*

Salary\_hike -0.10154 0.01618 -6.277 0.000239 \*\*\*

---

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

Residual standard error: 4.469 on 8 degrees of freedom

Multiple R-squared: 0.8312, Adjusted R-squared: 0.8101

F-statistic: 39.4 on 1 and 8 DF, p-value: 0.0002386

Slope = -0.10154

p-value for the slope of Salary Hike is 0.000239 < 0.05. Hence, the slope is significant.

R-squared = 0.8312 that is 83.12% which has improved

RMSE

#RMSE is 3.997528 which is good

b) Iteration 2

Transform Salary Hike to 1/Salary Hike

Correlation Coefficient

# since correleation coefficient is 0.9301463, the correlation is strong positive for churn out rate and reciprocal of salary hike

Summary of Regression model

Residuals:

Min 1Q Median 3Q Max

-3.539 -2.730 -1.759 2.119 7.175

Coefficients:

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

(Intercept) -107.87 25.26 -4.270 0.00272 \*\*

sal\_rec 304462.43 42493.97 7.165 9.57e-05 \*\*\*

---

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

Residual standard error: 3.995 on 8 degrees of freedom

Multiple R-squared: 0.8652, Adjusted R-squared: 0.8483

F-statistic: 51.33 on 1 and 8 DF, p-value: 9.569e-05

Slope = 304462.43

p-value for the slope of Salary Hike is 9.57e-05 < 0.05. Hence, the slope is significant.

R-squared = 0.8652 that is 86.52% which has improved

RMSE

#RMSE is 3.57303 which has reduced, hence model has improved

# Final model is

# Churn\_out\_rate = -107.87 + 304462.43 \* (1/salary\_hike)

**4) Salary Hike Data**

a)\_Iteration 1

Correlation Coefficient

# since correleation coefficient is 0.9782416, the correlation is strong positive

Summary of Regression model

Residuals:

Min 1Q Median 3Q Max

-7958.0 -4088.5 -459.9 3372.6 11448.0

Coefficients:

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

(Intercept) 25792.2 2273.1 11.35 5.51e-12 \*\*\*

YearsExperience 9450.0 378.8 24.95 < 2e-16 \*\*\*

---

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

Residual standard error: 5788 on 28 degrees of freedom

Multiple R-squared: 0.957, Adjusted R-squared: 0.9554

F-statistic: 622.5 on 1 and 28 DF, p-value: < 2.2e-16

Slope = 9450

p-value for the slope of Salary Hike is 2e-16 < 0.05. Hence, the slope is significant.

R-squared = 0.957 that is 95.7% which has good

RMSE

#RMSE is 5592.044

## Even after trying various transformations the correlation coefficient ( and hence R2 ) does not improve

# Hence,

# Final model is

# Salary = 14928 + 40582 \* YearsExperience