**QMB Regression Project**

**Group Members:**

**Prasanna Eddala**

**Data source:** For the project, I have taken Used cars data. The data is collected from 1998 to 2019 years used cars with respect to kilometers driven and transmission. Here we will analyze how the selling price is affecting based on the kilometers driven and the year when the car is purchased from a store.

The link to the data set is [https://www.kaggle.com/datasets/nehalbirla/vehicle-dataset-from-cardekho](https://www.kaggle.com/datasets/nehalbirla/vehicle-dataset-from-cardekh)

**Description of the data**: In the data set we have a total of 4340 observations with 8 columns. **The columns in the dataset are as follows**:

1. Name

2. Year

3. Selling\_price

4. km\_driven

5. Fuel

6. seller\_type

7. Transmission

8. Owner

|  |  |
| --- | --- |
| **Variable** | **Definition** |
| Selling\_price (Dependent) y | Selling price of the car  It is a numeric variable. |
| Km\_driven (X1) | It describes the kilometers driven since buying the new car  It is a numeric variable. |
| Year (X2) | It says the year of car purchase. |
| Transmission (X3) | It tells whether the car is manual or automatic. |

Dependent and independent variables in our data:

* In our data we have more than 100 samples. We have sufficient data to fit a model.
* Regression analyses of the following model combinations:

Model 1:

Code:

#y=selling\_price

#X1=km\_driven

#X2=year

#X3=transmission

M1=lm(selling\_price~km\_driven,data=cars\_data)

summary(M1)

Output:

Call:

lm(formula = selling\_price ~ km\_driven, data = cars\_data)

Residuals:

Min 1Q Median 3Q Max

-382624 -231602 -78430 153350 921753

Coefficients:

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

(Intercept) 576271.424 85114.177 6.771 1.65e-08 \*\*\*

km\_driven -2.401 1.184 -2.028 0.0481 \*

---

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

Residual standard error: 321500 on 48 degrees of freedom

Multiple R-squared: 0.07895, Adjusted R-squared: 0.05976

F-statistic: 4.114 on 1 and 48 DF, p-value: 0.04809

The regression equation is y(selling\_price)= 576271.424 (Intercept) -2.401(km\_driven)

Analysis:

In the above output, the regression model for simple regression of the dependent variable selling\_price(y) and km\_driven is the independent variable(x1). Here if we observe, the p-value is 1.65e-08 which is less than 0.05.So, we can reject Null hypothesis. Coming to the km\_driven variable, the coefficient is -2.401 and the p-value is 0.0481 where we can clearly reject the null hypothesis. The residual standard error is 321500 on 48 degrees of freedom and the adjected R-squared value is 0.05976.

Chart, scatter chart

Description automatically generatedChart, line chart

Description automatically generated

Chart, scatter chart

Description automatically generated

LINE Assumptions:

1. If we observe the scatter plot, it is not following the linearity.
2. From the Q-Q plot we can say it is not following the normality.
3. From the last graph we can say we are finding the patterns and it is not following the equality of variance.
4. So, we can clearly say that this model doesn’t fit best to our data.

Model 2:

Code:

M2=lm(selling\_price~year,data=cars\_data)

summary(M2)

Output:

Call:

lm(formula = selling\_price ~ year, data = cars\_data)

Residuals:

Min 1Q Median 3Q Max

-443267 -136911 -64091 34479 856115

Coefficients:

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

(Intercept) -79868902 18346345 -4.353 6.99e-05 \*\*\*

year 39897 9115 4.377 6.48e-05 \*\*\*

---

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

Residual standard error: 283200 on 48 degrees of freedom

Multiple R-squared: 0.2853, Adjusted R-squared: 0.2704

F-statistic: 19.16 on 1 and 48 DF, p-value: 6.478e-05

The regression equation is y(selling\_price)= -79868902 (Intercept) + 39897 (year)

Analysis:

In the above output, the regression model for simple regression of the dependent variable selling\_price(y) and year is the independent variable(x2). Here if we observe, the p-value is 6.99e-05 which is less than 0.05. So, we can reject Null hypothesis. Coming to the year variable, the coefficient is 39897and the p-value is 6.48e-05 where we can clearly reject the null hypothesis. The residual standard error is 283200 on 48 degrees of freedom and the adjected R-squared value is 0.2704.

Chart, scatter chart

Description automatically generatedChart, line chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

LINE Assumptions:

1. If we observe the scatter plot, it is not following the linearity.
2. From the Q-Q plot we can say it is not following the normality.
3. From the last graph we can say we are finding the patterns and it is not following the equality of variance.
4. So, we can clearly say that this model doesn’t fit best to our data.

Model 3:

Code:

M3=lm(selling\_price~transmission,data=cars\_data)

summary(M3)

Output:

Call:

lm(formula = selling\_price ~ transmission, data = cars\_data)

Residuals:

Min 1Q Median 3Q Max

-657500 -185348 -85348 162152 1064652

Coefficients:

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

(Intercept) 947500 148309 6.389 6.37e-08 \*\*\*

transmissionManual -562152 154623 -3.636 0.000675 \*\*\*

---

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

Residual standard error: 296600 on 48 degrees of freedom

Multiple R-squared: 0.2159, Adjusted R-squared: 0.1996

F-statistic: 13.22 on 1 and 48 DF, p-value: 0.000675

Analysis:

In the above output, the regression model for simple regression of the dependent variable selling\_price(y) and transmission is the independent variable(x3). Here if we observe, the p-value is 6.37e-08 which is less than 0.05. So, we can reject Null hypothesis. Coming to the transmission variable, the coefficient is -562152 and the p-value is 0.000675 where we can clearly reject the null hypothesis. The residual standard error is 296600 on 48 degrees of freedom and the adjected R-squared value is 0.1996.

Chart, line chart

Description automatically generatedChart, line chart

Description automatically generated

Chart

Description automatically generated

LINE Assumptions:

1. If we observe the scatter plot, it is not following the linearity.
2. From the Q-Q plot we can say it is not following the normality.
3. From the last graph we can say we are finding the patterns and it is not following the equality of variance.
4. So, we can clearly say that this model doesn’t fit best to our data.

Model 4:

Code:

M4=lm(selling\_price~km\_driven+year,data=cars\_data)

summary(M4)

Output:

Call:

lm(formula = selling\_price ~ km\_driven + year, data = cars\_data)

Residuals:

Min 1Q Median 3Q Max

-446522 -138703 -63519 44659 856024

Coefficients:

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

(Intercept) -7.881e+07 2.154e+07 -3.658 0.000641 \*\*\*

km\_driven -1.175e-01 1.222e+00 -0.096 0.923809

year 3.938e+04 1.069e+04 3.685 0.000591 \*\*\*

---

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

Residual standard error: 286200 on 47 degrees of freedom

Multiple R-squared: 0.2854, Adjusted R-squared: 0.255

F-statistic: 9.385 on 2 and 47 DF, p-value: 0.000372

Analysis:

In the above output, the regression model for simple regression of the dependent variable selling\_price(y) and here two independent variables. Here if we observe, the p-value is 0.000641 which is less than 0.05. So, we can reject Null hypothesis. Coming to the transmission variable, the coefficient is 3.938e+04 and the p-value is 0.000591 where we can clearly reject the null hypothesis. The residual standard error is 286200 on 48 degrees of freedom and the adjected R-squared value is 0.255.

Chart, scatter chart

Description automatically generated Chart, line chart

Description automatically generated

Chart, scatter chart

Description automatically generated

LINE Assumptions:

1. If we observe the scatter plot, it is not following the linearity.
2. From the Q-Q plot we can say it is not following the normality.
3. From the last graph we can say we are finding the patterns and it is not following the equality of variance.
4. So, we can clearly say that this model doesn’t fit best to our data.

Model 5:

Code:

M5=lm(selling\_price~km\_driven+transmission,data=cars\_data)

summary(M5)

Output:

Call:

lm(formula = selling\_price ~ km\_driven + transmission, data = cars\_data)

Residuals:

Min 1Q Median 3Q Max

-601522 -188759 -56535 176107 961540

Coefficients:

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

(Intercept) 1.115e+06 1.568e+05 7.115 5.47e-09 \*\*\*

km\_driven -2.555e+00 1.040e+00 -2.457 0.017740 \*

transmissionManual -5.758e+05 1.472e+05 -3.912 0.000294 \*\*\*

---

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

Residual standard error: 282200 on 47 degrees of freedom

Multiple R-squared: 0.3052, Adjusted R-squared: 0.2756

F-statistic: 10.32 on 2 and 47 DF, p-value: 0.0001923

Analysis:

In the above output, the regression model for simple regression of the dependent variable selling\_price(y) and here two independent variables. Here if we observe, the p-value is 5.47e-09 which is less than 0.05. So, we can reject Null hypothesis. Coming to the transmission variable, the coefficient is -5.758e+05 and the p-value is 0.000294 where we can clearly reject the null hypothesis. The residual standard error is 282200 on 48 degrees of freedom and the adjected R-squared value is 0.2756.

Chart, scatter chart

Description automatically generated Chart, line chart

Description automatically generated

Chart, scatter chart

Description automatically generated

LINE Assumptions:

1. If we observe the scatter plot, it is not following the linearity.
2. From the Q-Q plot we can say it is not following the normality.
3. From the last graph we can say we are finding the patterns and it is not following the equality of variance.
4. So, we can clearly say that this model doesn’t fit best to our data.

Model 6:

Code:

M6=lm(selling\_price~year+transmission,data=cars\_data)

summary(M6)

Output:

Call:

lm(formula = selling\_price ~ year + transmission, data = cars\_data)

Residuals:

Min 1Q Median 3Q Max

-583695 -104471 -26350 90554 842779

Coefficients:

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

(Intercept) -83896742 15000655 -5.593 1.11e-06 \*\*\*

year 42174 7456 5.656 8.89e-07 \*\*\*

transmissionManual -603868 120757 -5.001 8.39e-06 \*\*\*

---

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

Residual standard error: 231200 on 47 degrees of freedom

Multiple R-squared: 0.5335, Adjusted R-squared: 0.5136

F-statistic: 26.87 on 2 and 47 DF, p-value: 1.654e-08

Analysis:

In the above output, the regression model for simple regression of the dependent variable selling\_price(y) and independent variables. Here if we observe, the p-value is 1.11e-06 which is less than 0.05. So, we can reject Null hypothesis. Coming to the transmission variable, the coefficient is -603868 and the p-value is 8.39e-06 where we can clearly reject the null hypothesis. The residual standard error is 231200 on 47 degrees of freedom and the adjected R-squared value is 0.5136.

Chart, scatter chart

Description automatically generated Chart, line chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

LINE Assumptions:

1. If we observe the scatter plot, it is not bad compared to all models.
2. From the Q-Q plot we can say it is somewhat following the normality.
3. From the last graph we can say that we are finding the patterns and only some dots are lying on the line which are following the equality of variance.
4. So, we can say that this model can fit best to our data compared to remaining models.

Model 7:

Code:

M7=lm(selling\_price~km\_driven+year+transmission,data=cars\_data)

summary(M7)

Output:

Call:

lm(formula = selling\_price ~ km\_driven + year + transmission,

data = cars\_data)

Residuals:

Min 1Q Median 3Q Max

-581470 -107178 -29367 101896 840098

Coefficients:

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

(Intercept) -8.249e+07 1.761e+07 -4.685 2.51e-05 \*\*\*

km\_driven -1.571e-01 9.982e-01 -0.157 0.876

year 4.148e+04 8.735e+03 4.748 2.03e-05 \*\*\*

transmissionManual -6.040e+05 1.220e+05 -4.950 1.04e-05 \*\*\*

---

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

Residual standard error: 233700 on 46 degrees of freedom

Multiple R-squared: 0.5337, Adjusted R-squared: 0.5033

F-statistic: 17.55 on 3 and 46 DF, p-value: 9.781e-08

Analysis:

In the above output, the regression model for simple regression of the dependent variable selling\_price(y) and independent variable. Here if we observe, the p-value is 2.51e-05 which is less than 0.05. So, we can reject Null hypothesis. Coming to the transmission variable, the coefficient is -6.040e+05 and the p-value is 1.04e-05 where we can clearly reject the null hypothesis. The residual standard error is 233700 on 46 degrees of freedom and the adjected R-squared value is 0.5033.

Chart, scatter chart

Description automatically generated Chart, line chart

Description automatically generated

Chart, scatter chart

Description automatically generated

LINE Assumptions:

1. If we observe the scatter plot, it is not following the linearity.
2. From the Q-Q plot we can say it is not following the normality.
3. From the last graph we can say we are finding the patterns and it is not following the equality of variance.
4. So, we can clearly say that this model doesn’t fit best to our data.

Model 8:

Code:

M8=lm(selling\_price~km\_driven+year,km\_driven:year,data=cars)

summary(M8)

Output:

Call:

lm(formula = selling\_price ~ km\_driven + year, data = cars, subset = km\_driven:year)

Residuals:

Min 1Q Median 3Q Max

-639092 -211324 -86068 59384 8261035

Coefficients:

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

(Intercept) -1.014e+08 5.244e+06 -19.341 <2e-16 \*\*\*

km\_driven -1.357e-01 2.419e-01 -0.561 0.575

year 5.063e+04 2.602e+03 19.461 <2e-16 \*\*\*

---

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

Residual standard error: 484100 on 2331 degrees of freedom

(65660 observations deleted due to missingness)

Multiple R-squared: 0.169, Adjusted R-squared: 0.1683

F-statistic: 237.1 on 2 and 2331 DF, p-value: < 2.2e-16

Analysis:

In the above output, the regression model for simple regression of the dependent variable selling\_price(y) and independent variable. Here if we observe, the p-value is <2e-16 which is less than 0.05. So, we can reject Null hypothesis. Coming to the coefficient is 5.063e+04 and the p-value is <2e-16 where we can clearly reject the null hypothesis. The residual standard error is 484100 on 2311 degrees of freedom and the adjected R-squared value is 0.1683.

Chart, line chart

Description automatically generatedChart, scatter chart

Description automatically generated

LINE Assumptions:

1. From the Q-Q plot we can say it is not following the normality.
2. From the last graph we can say we are finding the patterns and it is not following the equality of variance.
3. So, we can clearly say that this model doesn’t fit best to our data.

Model 9:

Code:

M9=lm(selling\_price~km\_driven+I(km\_driven^2),data=cars)

summary(M9)

Output:

Call:

lm(formula = selling\_price ~ km\_driven + I(km\_driven^2), data = cars)

Residuals:

Min 1Q Median 3Q Max

-1682876 -283791 -127741 104335 8217717

Coefficients:

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

(Intercept) 7.356e+05 1.773e+04 41.496 < 2e-16 \*\*\*

km\_driven -4.193e+00 3.000e-01 -13.978 < 2e-16 \*\*\*

I(km\_driven^2) 7.039e-06 9.235e-07 7.621 3.06e-14 \*\*\*

---

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

Residual standard error: 564100 on 4337 degrees of freedom

Multiple R-squared: 0.0497, Adjusted R-squared: 0.04926

F-statistic: 113.4 on 2 and 4337 DF, p-value: < 2.2e-16

Analysis:

In the above output, the regression model for simple regression of the dependent variable selling\_price(y) and independent variables. Here if we observe, the p-value is < 2e-16 which is less than 0.05. So, we can reject Null hypothesis. Coming to the coefficient is 7.039e-06 and the p-value is 3.06e-14 where we can clearly reject the null hypothesis. The residual standard error is 564100 on 4337 degrees of freedom and the adjected R-squared value is 0.04926.

Chart, scatter chart

Description automatically generated Chart, line chart

Description automatically generated

Chart, scatter chart

Description automatically generated

LINE Assumptions:

1. If we observe the scatter plot, it is not following the linearity.
2. From the Q-Q plot we can say it is not following the normality.
3. From the last graph we can say we are finding the patterns and it is not following the equality of variance.
4. So, we can clearly say that this model doesn’t fit best to our data.

Model 10:

Code:

M10=lm(selling\_price~year+I(year^2),data=cars)

summary(M10)

Output:

Call:

lm(formula = selling\_price ~ year + I(year^2), data = cars)

Residuals:

Min 1Q Median 3Q Max

-869177 -198477 -71949 35951 8227530

Coefficients:

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

(Intercept) 1.350e+10 1.280e+09 10.55 <2e-16 \*\*\*

year -1.348e+07 1.273e+06 -10.59 <2e-16 \*\*\*

I(year^2) 3.366e+03 3.164e+02 10.64 <2e-16 \*\*\*

---

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

Residual standard error: 520000 on 4337 degrees of freedom

Multiple R-squared: 0.1924, Adjusted R-squared: 0.192

F-statistic: 516.6 on 2 and 4337 DF, p-value: < 2.2e-16

Analysis:

In the above output, the regression model for simple regression of the dependent variable selling\_price(y) and independent variables. Here if we observe, the p-value is <2e-16 which is less than 0.05. So, we can reject Null hypothesis. Coming to the transmission variable, the coefficient is 3.366e+03 and the p-value is <2e-16 where we can clearly reject the null hypothesis. The residual standard error is 520000 on 4337 degrees of freedom and the adjected R-squared value is 0.192.

Chart, scatter chart

Description automatically generated Chart, line chart

Description automatically generated

Chart, scatter chart

Description automatically generated

LINE Assumptions:

1. If we observe the scatter plot, it is not following the linearity.
2. From the Q-Q plot we can say it is not following the normality.
3. From the last graph we can say we are finding the patterns and it is not following the equality of variance.
4. So, we can clearly say that this model doesn’t fit best to our data.

**Conclusion:**

The best model for our data set is model 6 compared to all remaining models. Because, the residual standard error is very less and also Adjusted R-squared is more. From the graphs , we can clearly say that only model 6 is somewhat better in following line assumptions compared to all models .

Predictions:

new\_carsdata=data.frame(M6$km\_driven=='10000')

# predict is a point in time

predict(M6,new\_carsdata,interval="predict")

# confidence is a long term average

predict(M6,new\_carsdata,interval="confidence")

Output:

# predict is a point in time

> predict(M6,new\_carsdata,interval="predict")

fit lwr upr

1. 312615.85 -133671.744 758903.5

# confidence is a long term average

> predict(M6,new\_carsdata,interval="confidence")

fit lwr upr

1. 312615.85 245544.32 379687.39

Analysis:

From the above predictions, we can predict the selling price of used car for 10000 kilometers driven car. But we cannot say exactly that this is best fit because the range of predict is in negative that is -133671.744 758903.5 and also it has wider difference in range values. For the confidence interval we got 245544.32 379687.39.Finally if we compared all the ten models we got model 6 is best fit .