Assignment-1

AU1841131 Mansi Dobariya

→ In the study of starbucks debit card details a survey of 100 card purchases,

Answer-1 reg. model for predict amount of prepaid card

dependent variables:

Y: prepaid card in dollar

Independent variables:

X_1: age of customers

X_2: no. of days per month customer makes a purchase

X_3: no. of cups of coffee drinks per day

X_4: Income of customer ($1000)

→ no of predictors = 4

→ no of observations = 100

→ regression model,

b_0 = −96.0441

b_1 = −0.08729

b_2 = 0.28825

b_3 = 3.638967

b_4 = 3.06017

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4$$

$$\hat{y} = (-96.0441) + (-0.08729)\,x_1 + (0.28825)\,x_2 + (3.638967)\,x_3 + (3.06017)\,x_4$$

$\hat{y}$: predicted amount of prepaid card

Page No.:
Date:
YOUVA

→ Interpret intercept and reg. coefficients

- The sample Y intercept estimates amount of prepaid if age of customer is 0, no. of days per month customer makes purchase is 0, no. of cup of coffee per day is 0 and customers' income is 0 dollars. But,
- customers' age can't be zero here and also it's out of data range. same for income can be zero but here it's out of data range.
- Also, NO. of cup of coffee is zero and still amount of prepaid card has some value can't be same, its contradict so, this has no practical interpretetetation

- keeping constant the values of income (X_4) no. of cup of coffee per day (X_3) and no. of day per month customer (X_2) makes a purchase, 1 unit change in Age of customer (X_1) it will reflect in estimated amount of prepaid card by 0.08729 unit($) change.

- keeping constant customers' age (X_1), no. of cup of coffee per day (X_3) and ~~no. of day per month customer~~ customer's income (X_4), If 1 unit change in no. of day/month customer makes a purchase (X_2) it will reflect in estimated amount of prepaid card by 0.2882 unit($) change

- Keeping constant the values of age (X_1), no. of day/per month purchase (X_2) and income (X_4), If 1 unit change in no. of cup of coffee per day (X_3) drinks per day, it will reflect in estimated Amount of prepaid card by 3.6389 units change.

- Keeping constant the values of age (X_1), no. of day per month purchase (X_2) and no. of cup of coffee per day (X_3), If 1 unit change in income of customer (X_4), it will reflect in estimated amount of prepaid card by 3060.1 ($) unit change (3.0601 · 1000 $)

→ Interpreting Reg. statistics

Multiple R	0.939242
R^2	0.882176
adj R^2	0.877215
SE	18.37301

- multiple corr coefficient (R) indicates that, 0.93924 is greater than 0.75 multiple reg. model is strongly correlated with age, no. of day/month, no. of cup of coffee and income.
- coefficient of multiple determination (R^2) indicates, 0.882176 = 88.22% variation in amount of prepaid card is explained by variation in age no. of day/month, no. of cup of coffee and income.

(X_1, X_2, X_3 and X_4)

- adjusted R^2 (0.8772) = 87.72% of variation is explained by multiple reg. model adjusted for no. of Independent variable (4) and no. of observations (100).

- diff(R^2 - adj R^2) = 0.005 = 0.5% it's negligible so developed model is very significant

- standard error = 18.373 actual amount of prepaid card is differed by 18.373 from predicted amount of prepaid card.

Residuals
(-18.373, 18.373) ⇒ 67 out of 100
67% residuals are within interval
reg. model is appropriate for predicting prepaid amount of card

→ correlation Matrix

	Y	X_1	X_2	X_3	X_4
Y	1				
X_1	0.4302	1			
X_2	0.4025	-0.0175	1		
X_3	0.25331	0.2654	0.5119	1	
X_4	0.92418	0.11476	0.32992	0.09587	1

- amount of prepaid card is poorly positive correlate with age (X_1)
- amount of prepaid card is moderate positive correlated with no. of day per month customer makes purchase (X_2) (below avg)

to 0

- amount of prepaid card is moderate positive correlated with no. of cup of coffee per day (x_3) (below avg)
- amount of prepaid card is strongly positive correlated with income (x_4).
- age is negatively corelated with x_2 (no. of day(m)) with poor stregth, positively moderate core with x_3 (no. of cup of coffee) (below avg.) and positively poor correlated with x_4 (income)
- no. of day per month is positively (moderate) corelated with x_3 (no. of cup of coffee) (above avg), moderate positively correlated with income (x_4) (below avg).
- no. of cup of coffee is poorly positive correlated with income (x_4)

→ R (-0.0175, 0.2654, 0.11476, 0.51119, 0.3299, 0.09587)
no value is out of Range [-0.75, 0.75]
so, there's no multicollinearity problem

→ Test of for significance of overall Multiple Reg. model

significance $F = 3.28 \times 10^{-43}$
≈ 0.00

the overall significance of multiple reg model is tested with following Hypotheses:

$H_0 : \beta_1 = \beta_2 = 0 = \beta_3 = \beta_4$

H_a : Atleast one reg. coefficient is not equal to 0

- From ANOVA Table, it can be seen that significant F value is 0. At 5% significance level, $\alpha = 0.05$. significance F value is smaller than 0.05.
- This indicates null Hypothesis H_0 will be rejected and we conclude at least one of Independent variable is significantly related to amount of prepaid card.
- developed reg. model is statistically significant.

	P-value	Lower 95%	Upper 95%
x_1	0.6868	-0.51574	0.341173
x_2	0.6097	-0.8288	1.40539
x_3	0.0004	1.65499	5.62295
x_4	8.87×10^{-42}	2.8047	3.31568

$H_0: \beta_1 = 0$ and $H_a: \beta_1 \neq 0$

- Pvalue of x_1 is $0.6868 > 0.05$ so null Hypothesis is accepted and we conclude that there's a insignificant relationship between Y (amount of prepaid card) and age (x_1).
- P-value of x_2 $0.6097 > 0.05$ so H_0 is accepted and there's insignificant relationship between amount of prepaid card) and x_2 (no. of days per Month)
- Pvalue of x_3 $0.0004 < 0.05$ so H_0 is rejected and there's significant relationship between amount of prepaid card) and x_3 (no. of cup of coffee).
- Pvalue of x_4 $8.87 \times 10^{-42} << 0.05$ so H_0 is rejected and developed model is significant relationship between income and amount of card.

⇒ confidence Interval for x_1

$-0.51574 \leq \beta_1 \leq 0.341173$

includes "zero", reg. coefficient has a insignificant effect

taking into account effect of x_2, x_3 and x_4, estimated of 1 unit increase in age is to increase/decrease mean prepaid card amount by approx 0.5 to 0.4 bars 95% confidence interval correctly estimates relationship between variables

- confidence Interval for x_2

$-0.8289 \leq \beta_2 \leq 1.40539$

includes zero, reg. coefficient has a insignificant effect

taking into account effect of x_1, x_3, x_4, estimated of 1 unit increase in no. of days per month is to increase decrease mean prepaid card amount by approx 0.8 (1) to 1 bars. 95% confidence interval correctly estimates relationship between vars

- $1.65499 \leq \beta_3 \leq 5.62295$

excludes zero; reg. coefficient has a significant effect

1 unit increase in no. of cup of coffee is to increase mean prepaid card amount by approx 1 to 5 bars.

95% confidence interval correctly estimates relation

2.8047 ≤ β_3 ≤ 3.3156 ✓
excludes zero, reg ~~model~~ coefficient has a significant effect

Taking into account the effect of x_1, x_2, x_3, estimated of 1 unit increase income ~~prepaid card amount~~ is to increase mean prepaid card amount by app 2 to 3 bars. 95% confidence Interval correctly estimates relationship between variables.

→ Analysis of residual Plots

- There is no violation of variable, no pattern in relationship between Residuals and X1 (ages) or X2 (no. of day per month) or X3 (no. of cup of coffee) or X4 (income).
- healthy pattern appears in all of Four ~~three~~ plots.
- multiple reg. model is appropriate for predicting amount of prepaid cards.

→ Analysis of Line Fit plot
- There is no pattern (no curve) in relationship between Y and x_1, x_2, x_3.
Healthier plots they are.
- x_4 and Y is positively corelated with each other.

→ detection of outliers in standard Residuals.

[-2, 2] 99 out of 100 close in interval

2.344 (93rd) is outlier which is out of range.

→ Normal Pr Plot

→ Small sample sizes ($n \leq 75$) often produce normal probability plots that deviate substantially from linearity

→ Large sample size ($n \geq 95$) often produce plots which are much better behaved

→ Usually about $n = 85$ is required to produce stable and easily interpretable normal probability plots

Answer-2 reg. model for predict no. of day per month customer makes purchase

→ In the study of starbucks debit card details a survey of 100 card purchases,

→ Dependent variables:

y_{x_2} : no. of day/m customer makes purchase

Independent variables:

x_1 : age of customers

x_3 : no. of cups of coffee drinks per day

x_4 : income of customers ($1000)

→ no. of predictors = 3

no. of observations = 100

→ reg. model,

$$\hat{y}_{x_2} = 6.789028 + (-0.0880)\,x_1 + (0.969016)\,x_3 + (0.0808)\,x_4$$

$\hat{y}_{x_2}$ ⇒ estimated no. of day per month customer makes purchase.

→ Intercept and reg. coefficients

- The sample x_2 intercept estimates no. of day per month if age of customer is 0, no. of ~~day~~ cup of coffee per day is 0, and customers income is 0 dollars. But,
- customers' age can't be zero here and also its out of data range same income can be zero but out of data range

- also, no. of cup of coffee can't be zero because of contracting to no. of day per month purchase.
- this has no practical interpretation.

- keeping constant the value of income (x_4). no. of cup of coffee per day (x_3) ~~coefficient~~, If 1 unit change in age of customer (x_1) it will reflect in estimated no. of day per month by 0.088 unit change

- keeping constant age (x_1) and income (x_4), If 1 unit change in no. of cup of coffee per day (x_3), it will reflect in estimated no of day per month by 0.9690 unit change

- keeping constant age (x_1) and no. of cup of coffee per day (x_3), If 1 unit change in income (x_4) it will reflect in estimated no. of day per month by 0.0808 unit change

→ Interpreting Reg. statistics

Multiple R	0.6135
R^2	0.3763
adj R^2	0.3568
SE	3.3324

multiple corr2 coefficient (R) indicates that, 0.6135 is less than 0.75. multiple reg. model is moderate connected with age, no. of cup of coffee and income with above avg.

- coefficient of multiple determination (R^2) indicates, 0.3763 = 37.63 % variation in ~~correct~~ no. of day per month (X_2) is explained by variation in age, no. of cup of coffee, income (X_1, X_3, X_4)

- adjusted R^2 = 0.3568 = 35.68 % of variation is explained by multiple reg. model adjusted for no. of Independent variables (3) and no. of observations. (100.)

- Standard error = 3.3324 actual no. of day per month costumer makes purchase is differed by 3.3324 from predicted amount of ~~people~~ no. of day per month.

Residuals

(-3.3324, 3.3324) ⇒ 73 % residuals are within interval reg. model is appropriate for predicting no. of day per month user makes purchase

Page No.:
Date:

Correlation matrix

	y	x_2	x_1	x_3	x_4
No. of Day/month	y	1			
	x_2	-0.01757	1		
	x_1	0.511915	0.26543	1	
	x_3	0.3299	0.114762	0.0958	1

→ no. of day/month customer makes purchase is negatively poor correlated with age of customer (x_1)

→ no. of day/month customer makes purchase is positively moderate correlated with no. of cup of coffee (x_3) customer drinks per day with above avg.

→ no. of day per month customer makes purchase is positively moderate correlated with income of customer (x_4) with below avg

- age is poorely +ve correlated with x_3 and x_4
- income is poorely +ve correlated with x_3.
- Each Independent variables are poorely correlated so there's no problem of multicollinearity.

a → Test for significance of overall Multiple Reg Model.

significance level $F = 6.06 \times 10^{-10}$

$\simeq 0.000$

Hypothesis:

H_0 : $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$

H_a : atleast one reg. coefficient is not equal to 0.

- From ANOVA Table, it can be seen that significant F value is 0. At 5% significance level $\alpha = 0.05$. significance F value is smaller than 0.05
- This indicates null Hypothesis H_0 will be rejected and we conclude at least one of Independent variable is significantly related to no. of day per month (Y_{x_2})
- developed model is statistically significant.

→ 95% confidence interval correctly estimates relationship between variables

	P-value	Lower 95%	Upper 95%
x_1	0.023	−0.1636	−0.0124
x_3	6.25×10^{-09}	0.6675	1.2705
x_4	0.000359	0.03746	0.1241

→ Pvalue of x_1 is $0.023 < 0.05$ so H_0 is rejected and we conclude that there's significant relationship between Y_{x_2} and x_1 (age)

→ P value of x_3 is $0.00 < 0.05$ so H_0 is rejected and we conclude that there's singificant relationship between Y_{x_2} and x_3 (no. of cup of coffee / day)

→ Pvalue of x_4 is $0.000359 < 0.05$ so H_0 is rejected and there's significant relationship between Y_{x_2} and x_4 (income)

→ confidence interval for x_1

$-0.1636 \leq \beta_1 \leq -0.0124$

excludes "0", reg. coefficient has a significant effect

- Taking into accout effect of x_3, x_4 estimated of 1 unit increase in age is to reduce mean no. of day/month customer purchase by approx 0.0124 to 0.16 (approx) 0 bars.

→ confidence interval for x_3,

$$0.6675 \leq \beta_3 \leq 1.2705$$

excludes "0". reg. coefficient has a significant effect

- Taking into account effect of x_1, x_4 estimated of 1 unit increase in no. of cup of coffee is to increase mean no. of day/month customer purchase by approx 0 to 1 bars.

→ confidence interval for x_4,

$$0.0374 \leq \beta_4 \leq 0.124156$$

excludes "0" reg coefficient has a significant effect.

Taking into account effect of x_1, x_3 estimate of 1 unit increase in no. of cup of coffee is to increase mean no. of day per month customer purchase by approx 0.03 (0 to 0.124) bars approx 0 bars.

→ Analysis of residual plots.

- There is no violation of variable, no pattern in relationship between Residuals and x1 (ages) or x_3 (no. of cup of cotted) or x$ (income)
- healthy pattern appears in all of three plots.
- multiple reg. model is app. for predicting no. of day/Month customer purchase.

→ Analysis of Line Fit Plot

- there's little increasing pattern in graphs. It's positive related to Yx_2 and x_3 or x_4.
- negatively correlation between age(x_1) and no. of day per month customer purchase (x_2)
- x_3 and Yx_2 is moderate positively corr as shown in graph
- x_4 and Yx_2 is strongly positive corr with each other

→ detection of outliers in standard Residuals

[-2, 2] ⇒ 94 out of 1000 inside the range

2.027941 (5th, 13th, 15th, 23rd, 27th and 32th) are outlier which it outside the range

→ Normal pr Plot

- Smeell sample size ($n \leq 25$) often produce normed, probability plots that deviate substantially from linearity
- Large sample size ($n \geq 65$) often produce plots which are much better behaved.
- Usually about $n = 40$ is required to produce stable and easily interpretable normal pr. plots.