

A Regression is used to understand cause - effect relationship.

8. Predictive Analytics

8.1. Linear Regression

* Simple Linear Regression

what happens to sales of a brand of shampoo when discount of 15% is offered in particular week.

We expect sales to go up

Here

The cause : reduction in price

effect : increase in sales

Also if I want to know by how much
the sales has increased?

That is the quantification of the impact

* Regression analysis is a statistical technique used to infer the magnitude and direction of a possible causal relationship between an observed pattern and variables assumed to have an impact on the observed pattern.

Statistical -

A mathematical approach that assumes that the pattern of interest and the variables that impact the pattern are all random samples from underlying population.

Magnitude - size of impact (2 times, 10 times)

Direction - Positive or Negative

Causal -

Magnitude of rainfall has impact on crop yield, but crop yield does not influence rainfall

Observed Pattern - Dependent variable,
Distribution

* Simple Linear Regression

e.g. You work for a hospital and are looking to understand factors that may influence the birth weight of a baby

→ Maternal Diet

→ Gestation period

→ Maternal Health Issues

→ Ethnicity

→ Age of the mother

Sample Data

Birth Weight	Weeks of Gestation	Years of Education	Race	Smoked During Pregnancy
7.5	38	12	African American	No
7.8	39	14	White	No
8.1	40	16	Asian	No
8.4	41	18	White	No
8.7	42	20	Asian	No
9.0	43	22	White	No
9.3	44	24	Asian	No
9.6	45	26	White	No
10.0	46	28	Asian	No
10.3	47	30	White	No
10.6	48	32	Asian	No
11.0	49	34	White	No
11.3	50	36	Asian	No
11.6	51	38	White	No
12.0	52	40	Asian	No
12.3	53	42	White	No
12.6	54	44	Asian	No
13.0	55	46	White	No
13.3	56	48	Asian	No
13.6	57	50	White	No
14.0	58	52	Asian	No
14.3	59	54	White	No
14.6	60	56	Asian	No
15.0	61	58	White	No
15.3	62	60	Asian	No
15.6	63	62	White	No
16.0	64	64	Asian	No
16.3	65	66	White	No
16.6	66	68	Asian	No
17.0	67	70	White	No
17.3	68	72	Asian	No
17.6	69	74	White	No
18.0	70	76	Asian	No
18.3	71	78	White	No
18.6	72	80	Asian	No
19.0	73	82	White	No
19.3	74	84	Asian	No
19.6	75	86	White	No
20.0	76	88	Asian	No
20.3	77	90	White	No
20.6	78	92	Asian	No
21.0	79	94	White	No
21.3	80	96	Asian	No
21.6	81	98	White	No
22.0	82	100	Asian	No
22.3	83	102	White	No
22.6	84	104	Asian	No
23.0	85	106	White	No
23.3	86	108	Asian	No
23.6	87	110	White	No
24.0	88	112	Asian	No
24.3	89	114	White	No
24.6	90	116	Asian	No
25.0	91	118	White	No
25.3	92	120	Asian	No
25.6	93	122	White	No
26.0	94	124	Asian	No
26.3	95	126	White	No
26.6	96	128	Asian	No
27.0	97	130	White	No
27.3	98	132	Asian	No
27.6	99	134	White	No
28.0	100	136	Asian	No
28.3	101	138	White	No
28.6	102	140	Asian	No
29.0	103	142	White	No
29.3	104	144	Asian	No
29.6	105	146	White	No
30.0	106	148	Asian	No
30.3	107	150	White	No
30.6	108	152	Asian	No
31.0	109	154	White	No
31.3	110	156	Asian	No
31.6	111	158	White	No
32.0	112	160	Asian	No
32.3	113	162	White	No
32.6	114	164	Asian	No
33.0	115	166	White	No
33.3	116	168	Asian	No
33.6	117	170	White	No
34.0	118	172	Asian	No
34.3	119	174	White	No
34.6	120	176	Asian	No
35.0	121	178	White	No
35.3	122	180	Asian	No
35.6	123	182	White	No
36.0	124	184	Asian	No
36.3	125	186	White	No
36.6	126	188	Asian	No
37.0	127	190	White	No
37.3	128	192	Asian	No
37.6	129	194	White	No
38.0	130	196	Asian	No
38.3	131	198	White	No
38.6	132	200	Asian	No
39.0	133	202	White	No
39.3	134	204	Asian	No
39.6	135	206	White	No
40.0	136	208	Asian	No
40.3	137	210	White	No
40.6	138	212	Asian	No
41.0	139	214	White	No
41.3	140	216	Asian	No
41.6	141	218	White	No
42.0	142	220	Asian	No
42.3	143	222	White	No
42.6	144	224	Asian	No
43.0	145	226	White	No
43.3	146	228	Asian	No
43.6	147	230	White	No
44.0	148	232	Asian	No
44.3	149	234	White	No
44.6	150	236	Asian	No
45.0	151	238	White	No
45.3	152	240	Asian	No
45.6	153	242	White	No
46.0	154	244	Asian	No
46.3	155	246	White	No
46.6	156	248	Asian	No
47.0	157	250	White	No
47.3	158	252	Asian	No
47.6	159	254	White	No
48.0	160	256	Asian	No
48.3	161	258	White	No
48.6	162	260	Asian	No
49.0	163	262	White	No
49.3	164	264	Asian	No
49.6	165	266	White	No
50.0	166	268	Asian	No
50.3	167	270	White	No
50.6	168	272	Asian	No
51.0	169	274	White	No
51.3	170	276	Asian	No
51.6	171	278	White	No
52.0	172	280	Asian	No
52.3	173	282	White	No
52.6	174	284	Asian	No
53.0	175	286	White	No
53.3	176	288	Asian	No
53.6	177	290	White	No
54.0	178	292	Asian	No
54.3	179	294	White	No
54.6	180	296	Asian	No
55.0	181	298	White	No
55.3	182	300	Asian	No
55.6	183	302	White	No
56.0	184	304	Asian	No
56.3	185	306	White	No
56.6	186	308	Asian	No
57.0	187	310	White	No
57.3	188	312	Asian	No
57.6	189	314	White	No
58.0	190	316	Asian	No
58.3	191	318	White	No
58.6	192	320	Asian	No
59.0	193	322	White	No
59.3	194	324	Asian	No
59.6	195	326	White	No
60.0	196	328	Asian	No
60.3	197	330	White	No
60.6	198	332	Asian	No
61.0	199	334	White	No
61.3	200	336	Asian	No
61.6	201	338	White	No
62.0	202	340	Asian	No
62.3	203	342	White	No
62.6	204	344	Asian	No
63.0	205	346	White	No
63.3	206	348	Asian	No
63.6	207	350	White	No
64.0	208	352	Asian	No
64.3	209	354	White	No
64.6	210	356	Asian	No
65.0	211	358	White	No
65.3	212	360	Asian	No
65.6	213	362	White	No
66.0	214	364	Asian	No
66.3	215	366	White	No
66.6	216	368	Asian	No
67.0	217	370	White	No
67.3	218	372	Asian	No
67.6	219	374	White	No
68.0	220	376	Asian	No
68.3	221	378	White	No
68.6	222	380	Asian	No
69.0	223	382	White	No
69.3	224	384	Asian	No
69.6	225	386	White	No
70.0	226	388	Asian	No
70.3	227	390	White	No
70.6	228	392	Asian	No
71.0	229	394	White	No
71.3	230	396	Asian	No
71.6	231	398	White	No
72.0	232	400	Asian	No
72.3	233	402	White	No
72.6	234	404	Asian	No
73.0	235	406	White	No
73.3	236	408	Asian	No
73.6	237	410	White	No
74.0	238	412	Asian	No
74.3	239	414	White	No
74.6	240	416	Asian	No
75.0	241	418	White	No
75.3	242	420	Asian	No
75.6	243	422	White	No
76.0	244	424	Asian	No
76.3	245	426	White	No
76.6	246	428	Asian	No
77.0	247	430	White	No
77.3	248	432	Asian	No
77.6	249	434	White	No
78.0	250	436	Asian	No
78.3	251	438	White	No
78.6	252	440	Asian	No
79.0	253	442	White	No
79.3	254	444	Asian	No
79.6	255	446	White	No
80.0	256	448	Asian	No
80.3	257	450	White	No
80.6	258	452	Asian	No
81.0	259	454	White	No
81.3	260	456	Asian	No
81.6	261	458	White	No
82.0	262	460	Asian	No
82.3	263	462	White	No
82.6	264	464	Asian	No
83.0	265	466	White	No
83.3	266	468	Asian	No
83.6	267	470	White	No
84.0	268	472	Asian	No
84.3	269	474	White	No
84.6	270	476	Asian	No
85.0	271	478	White	No
85.3	272	480	Asian	No
85.6	273	482	White	No
86.0	274	484	Asian	No
86.3	275	486	White	No
86.6	276	488	Asian	No
87.0	277	490	White	No
87.3	278	492	Asian	No
87.6	279	494	White	No
88.0	280	496	Asian	No
88.3	281	498	White	No
88.6	282	500	Asian	No
89.0	283	502	White	No
89.3	284	504	Asian	No
89.6	285	506	White	No
90.0	286	508	Asian	No
90.3	287	510	White	No
90.6	288	512	Asian	No
91.0	289	514	White	No
91.3	290	516	Asian	No
91.6	291	518	White	No
92.0	292	520	Asian	No
92.3	293	522	White	No
92.6	294	524	Asian	No
93.0	295	526	White	No
93.3	296	528	Asian	No
93.6	297	530	White	No
94.0	298	532	Asian	No
94.3	299	534	White	No
94.6	300	536	Asian	No
95.0	301	538	White	No
95.3	302	540	Asian	No
95.6	303	542	White	No
96.0	304	544	Asian	No

Ques

What are possible ways to accessing these relationships?

- Graphical Visualization
- Correlations
- Run Regression Model

A better way to identify the relationship between these variables is to use a regression technique.

Why Regression?

- Multiple factor impact on the effect
- Statistical Significance of the impact

* Simple Linear Regression

we are trying to find relationship between baby birth weight and gestation period

Birthweight = f (gestation weeks)

$$Y = mX + C$$

$m \Rightarrow$ slope

$C \Rightarrow$ intercept

Slope \Rightarrow rate of change of Y when X changes

Intercept \Rightarrow value of Y when $X=0$

$$Y = \beta_0 + \beta_1 x + e$$

where $\beta_0 \Rightarrow$ Intercept

$\beta_1 \Rightarrow$ Slope

$e \Rightarrow$ Error

Terminology

Dependent Variable / Predicted / Target Variable

Independent Variable / Predictor / X

Beta coefficient(s):

The estimate of magnitude of impact of changes in the predictors on the predicted variable.

Error:

The impact of the unobserved variables on the dependent variable, usually calculated as the difference between the predicted value of y given the estimated regression function and the actual value of y .

The ordinary Least squares Regression (OLS) technique estimates coefficients on the variables hypothesized to have an impact on the variable of interest by identifying the line that minimizes the sum of squared differences between points on the estimated line and on the actual values of the independent variable.

- Coefficients : Betas

Minimizes : Least

Sum of Squared Differences : Square of Residuals

Estimated Line : Regression Line

Actual Values : values in dataset

OLS Estimates

The OLS Regression find that line by looking at the residuals (or the difference between the points on each line and ~~not~~ actual y) and minimizing the sum of their squares.

$$Q = \sum_{i=1}^N (y_i - b_0 - b_1 x_i)^2$$

Using differential calculus we get

$$b_0 = \frac{\sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i}{n \sum x_i^2 - (\sum x_i)^2}$$

$$b_1 = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - (\sum x_i)^2}$$

How to measure Model fit? (Validating the model)

→ R^2

→ Fit chart - Actual vs Fitted values

→ MAPE - Mean Absolute Percentage Error

* Higher the R^2 better the model

Regression Assumptions

1. Model is linear in parameters
2. The data is a random sample of population
 - The errors are statistically independent from one another
3. The expected value of the errors is always zero.
4. The independent variables are not too strongly collinear
5. The independent variables are measured precisely

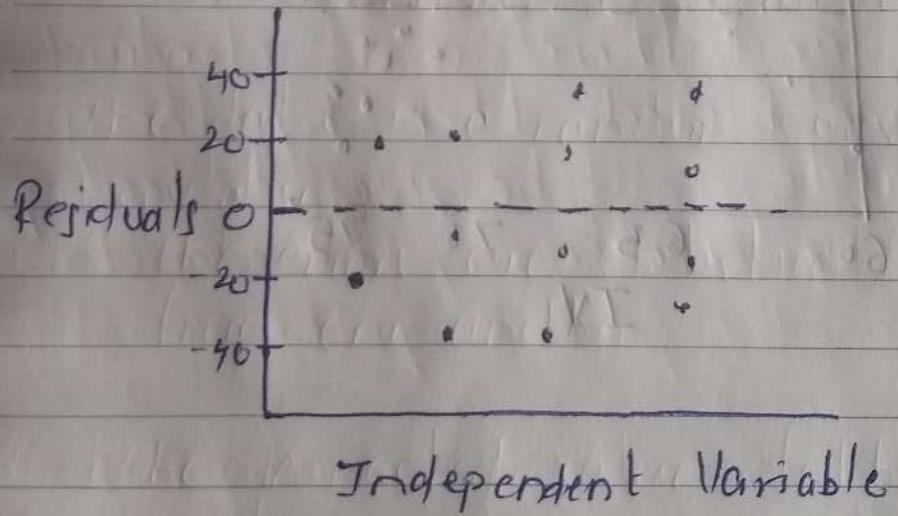
6. The residuals have constant variance
7. The errors are normally distributed
8. The model is correctly specified

If all these conditions hold,
then OLS estimates are BLUE -
Best Linear Unbiased Estimators

Checking If Assumptions are Valid

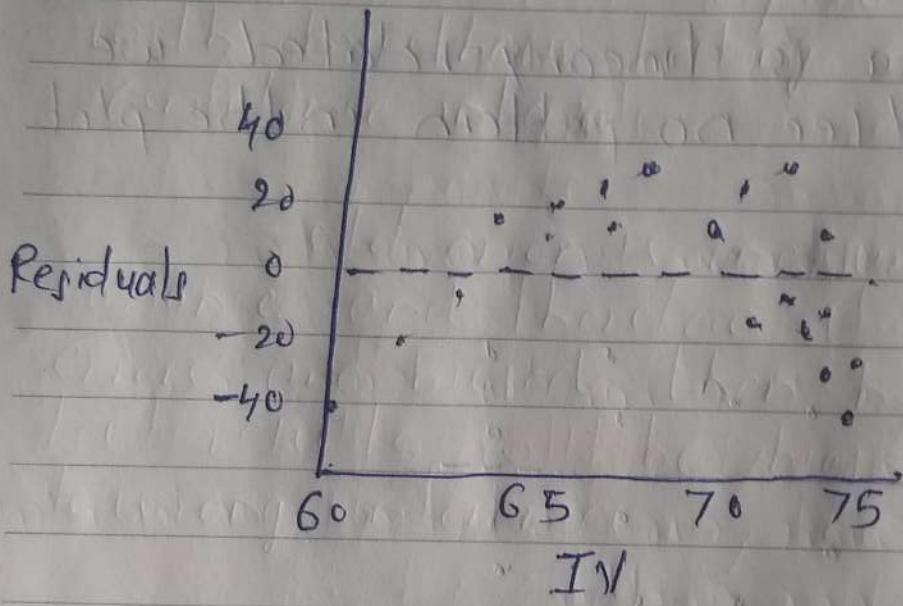
1) check for Linearity - Plot the residuals against each IV

- If data is linearly related, we should see no pattern in the plot



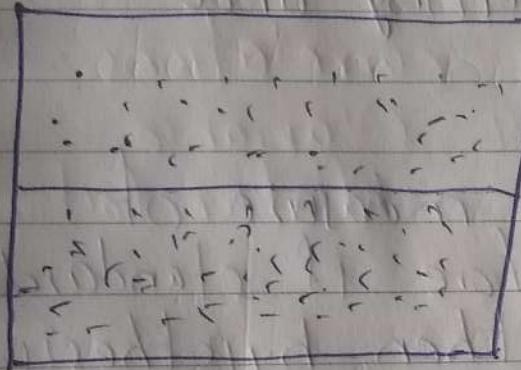
If the relationship is non-linear

we will see some pattern in the plot



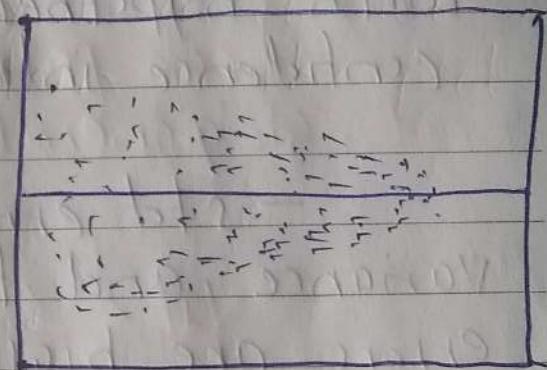
2. The residuals should have constant variance - homoscedasticity

Plot the residuals ~~again~~ against predicted Y



Homoscedasticity

No pattern in
data
Variance is constant



Heteroscedasticity

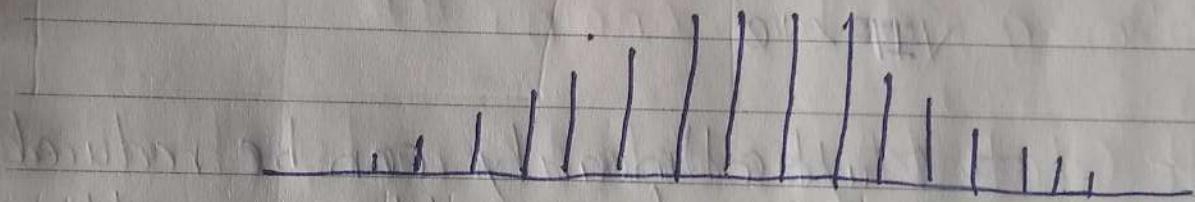
We can see
pattern in the
data

- The presence of heteroscedasticity does not imply bias in the estimates
- Heteroscedasticity leads to bias in the standard errors, leading to issues with hypothesis testing and confidence intervals
 - std error is measure of variance, and therefore if standard errors are biased, then hypothesis test results will be biased leading to wrong inferences

3. The residuals are normally distributed

Histogram or probability plot for residuals

Distribution of errors



If the residuals are not normally distributed?

Hypothesis test outcomes may be invalid, though less of an issue with large samples

4. The IVs are not too correlated -
multicollinearity

- The IVs should not be highly correlated to one another (40% and 10%)
- check pairwise correlations or generate VIF (Variance Inflation factor)

$$VIF < 10$$

- Multicollinearity can be reduced by Transforming the variable or by combining the variable

R Code

Age — Old / Middle / Young

Gender — Male / Female

Own Home — Own / Rent

Married — Single / Married

Location — Far / Close

Salary —

children — 0 / 1 / 2 / 3

History — High / Low / NA / Medium
no. of purchase

Catalogs — 6 to 24

Amount Spent —

```
setwd (" ")
```

```
data <- read.csv ("DirectMarketing.csv")
```

```
library (dplyr)
```

```
library (ggplot2)
```

```
library (car)
```

```
head (data)
```

```
## Do exploratory Analysis ##
```

```
plot (data$Age , data$AmountSpent, col = "red")
```

```
# Middle f old Age have similar spent
```

```
# Combine middle f old levels together
```

```
data$Age1 <- ifelse (data$Age1 == "Young",  
"Middle-old", as.character (data$Age))
```

```
data$Age1 <- as.factor (data$Age1)
```

```
summary (data$Age1)
```

plot (data\$Age1, data\$AmountSpent)

Gender

plot (data\$Gender, data\$AmountSpent)

own Home

summary (data\$ownHome)

plot (data\$ownHome, data\$AmountSpent,
col = "red")

Married

summary (data\$Married)

plot (data\$Married, data\$AmountSpent,
col = "red")

Location

summary (data\$Location)

plot (data\$Location, data\$AmountSpent,
col = "red")

Salary

summary (data\$Salary)

plot (data\$Salary, data\$AmountSpent)

Might be heteroscedasticity

children

summary (data\$children)

data\$children <- as.factor (data\$children)

plot (data\$children, data\$AmountSpent,
col = "red")

data\$children1

<- ifelse (data\$children == 3 | data\$children == 2,
"3-2", as.character (data\$children))

data\$children1 <- as.factor (data\$children1)

summary (data\$children1)

plot (Data\$children1, Data\$Amountspent,
col = "red")

History

summary (data\$History)

Impute missing values

tapply (data\$Amountspent, data\$History,
mean)

ind <- which (is.na (data\$History))

mean (data [ind, "Amountspent"])

```
data %>% filter(History == "Medium") %>%  
  select(AmountSpent) -> Amt_M
```

```
P <- ggplot(data = Amt_M, aes(x = AmountSpent))  
  + geom_histogram()
```

```
q <- ggplot(Amt_M, aes(x = AmountSpent))
```

```
  + geom_histogram()
```

```
  + geom_histogram()
```

```
# Create a category called missing
```

```
data$History1 <- ifelse(is.na(data$History),  
  "Missing", as.factor(data$History))  
data$History1 <- as.factor(data$History1)
```

```
summary(data$History1)
```

```
data$History1 <- factor(data$History1,  
                         labels = c("High", "Low", "Medium",  
                         "Missing"))
```

```
# catalogues
```

```
summary (data$catalogs)
```

```
data1 <- data [, -c(1,7,8)]
```

```
mod1 <- lm (AmountSpent ~ ., data = data1)
```

```
summary (mod1)
```

```
mod2
```

```
<- lm (formula = AmountSpent ~ Gender + Location  
      + Salary + catalogs + children1 + History1,  
      data = data1)
```

```
summary (mod2)
```

summary (data1)

Remove insignificant variables

History Missing

Gender Male

Create Dummy Variables

```
data$Male.d <- ifelse (data1$Gender == "Male",  
1, 0)
```

```
data$Male.f <- ifelse (data1$Gender == "Female",  
1, 0)
```

```
data1$Missing.d <- ifelse (data1$History1 ==  
"Missing", 1, 0)
```

data\$ Missing -

```
data$Low.d <- ifelse (data1$History1 == "Low",  
1, 0)
```

```
data$ Mid.d <- ifelse (data1$History1 == "Medium",  
1, 0)
```

```
data$High-d <- ifelse (data$History1 == "High",  
1, 0)
```

```
mod3 <- lm (formula = AmountSpent ~ Male_d +  
Location + Salary + Catalogs + Children1  
+ Med_d + Low_d, data = data1)
```

```
summary (mod3)
```

```
mod4  
<- lm (Formula = AmountSpent ~ Location + salary +  
catalogs + Children1 + Med_d + Low_d,  
data = data1)
```

```
summary (mod4)
```

```
# Signs
```

tapply(data1\$AmountSpent, data1\$History,
mean)

data1 %>% filter(History1 == "Medium",
History1 == "Low") %>% summarize
(Mean = mean(AmountSpent)) # In Line

tapply(data1\$AmountSpent, data1\$Location,
mean) # In line

Assumption Checks

hist(mod4\$residuals)

qqplot(mod4\$residuals)

Non normal behaviour observed

45° line denotes the behaviour if
data has been normal

Multicollinearity check

vif(mod4)

Constant variance check

plot4(mod4\$fitted.values, mod4\$residuals) # funnel shape

Remedies Remedies : Apply log transform to y variable

mod5

$\leftarrow \text{lm}(\text{formula} = \log(\text{AmountSpent}) \sim \text{Location} + \text{Salary} + \text{catalogs} + \text{children} + \text{Med_d} + \text{Low_d}, \text{data} = \text{data1})$

summary(mod5)

qqplot(mod5\$residuals) # looks okay

plot(mod5\$fitted.values, mod5\$residuals)
still funnel

Apply square root transform

mod6

```
<- lm (formula = sqrt(Amount_spent) ~ Location +
      salary + catalogs + Children1 + med_d +
      Low_d, data = data1)
```

summary (mod6)

qqplot (mod6\$residuals)

plot (mod6\$fitted.values, mod6\$residuals)

seems okay

vif (mod6)

I can also try cube root transform or Normal Logarithmic Transformation

```
predicted <- mod6$fit$predicted.values
```

```
actual <- sqrt(data1$AmountSpent)
```

```
dat <- data.frame(predicted, actual)
```

```
p <- ggplot(dat, aes(x = row(dat)[, 2],  
y = predicted))
```

```
+ geom_line(colour = "blue") + geom_line(data = dat,  
aes(y = actual), colour = "black")
```