CATEGORICAL VARIABLES AND INTERACTION TERMS

Categorical Variables → Binary Category

Example, Pink Slip = 1 if car has roadworthy
= 0 otherwise.

$Price_i = \beta_0 + \beta_1 (Pink\ Slip) + \varepsilon_i$ output →

DV: Price	Coeff	SE	t	P-value
Intercept	3978.3	1056.2	3.7	0.0002
Pink slip	1625.6	1203.7	1.3	0.179

$\hat{Price}_i = 3978 + 1626 (Pink\ Slip)$

Inference → A car with a pink slip would command a sale price $1,626 more than a car without pink slip, on average.

But **p value > 0.1** which means **p value / pink slip is not that important.**

So consider other values,

$Ln(Price) = \beta_0 + \beta_i (Age_i) + \beta_2 (Age)^2 + \beta_n Ln(Odometer_i) + \beta_4 (Pink\ Slip) + \varepsilon_i$

Output →

Ln(Price)	Coef	SE	t	P-value
Intercept	9.237	0.27	33.51	0.000
Age	-0.052	0.019	-3.78	0.003
Age2	0.001	0.000	4.72	0.000
Ln(Odometer)	-0.198	0.061	-3.24	0.0016
Pink slip	0.158	0.178	0.87	0.3846

$R^2 = 0.367$

Still pink slip have high p value

$Ln(Price) = 9.237 - 0.052(Age) + 0.001 (Age)^2 - 0.198 (odometer) + 0.158 (Pink\ Slip)$

Inference → A car with pink slip would command a sale price 15.8% higher than a car without a pink slip on average holding all other variables constant. because we log have log in dependent variable, so can express variable in percentage. by one.

Multi level Category

Age Category = 1 if age <= 5, = 2 if age > 5 & age <= 15, 3 if 15 < age <= 35
= 4 if age > 35, **to get this cut check Age Vs price (See below)**

$Ln(Price) = \beta_0 + \beta_1 (Age\ Cat) + \beta_2 ln(odometer) + \beta_3 (Pink\ Slip) + \varepsilon_i$

But the **problem is it is ordinal**, 4 > 3 > 2 > 1 but ~~doe~~ every time it will not mean older the car older the price or older the car higher the price, it cannot be ~~nomin~~ ordinal

So use dummy variable, & **Avoid AgeCat4, Dummy Variable trap** ← **AgeCat4 = 1 if age > 35, = 0 otherwise**

AgeCat1 = 1 if age <= 5, = 0 otherwise

AgeCat2 = 1 if 5 < Age <= 15, = 0 if Otherwise

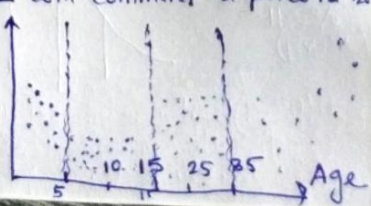
AgeCat3 = 1 if 15 < age <= 35, = 0 otherwise

$Ln(Price) = \beta_0 + \beta_1 (Age\ Cat2) + \beta_2 (Age\ Cat3) + \beta_3 (Age\ Cat4) + \beta_4 ln(Odometer) + \beta_5 (Pink\ Slip) + \varepsilon_i$

Ln(Price)	Coef	SE	t	P-value
Intercept	8.94	0.27	32.09	0.00
AgeCat2	-0.129	0.24	-0.52	0.60
AgeCat3	-0.735	0.26	-2.80	0.006
AgeCat4	0.479	0.32	1.45	0.150
Ln(Odometer)	-0.225	0.06	-3.69	0.004
Pink Slip	0.359	0.17	1.98	(0.502) ← decreased

Inference → On average, holding all other variable constant a car in age category 2 will command a price 12.9% lower than age category 1.

Age category plot price →



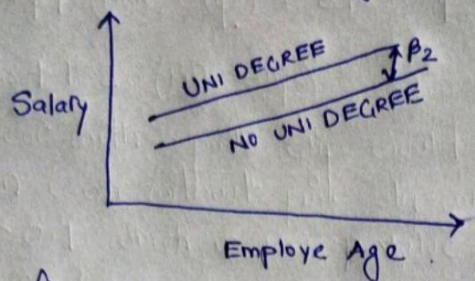
INTERACTION TERMS

Example – Build a model to explain the salary of all Google's employee.
Dependent Variable – Salary Independent Var – i) Employee Age
ii) University degree – 1 Yes, 0 No.

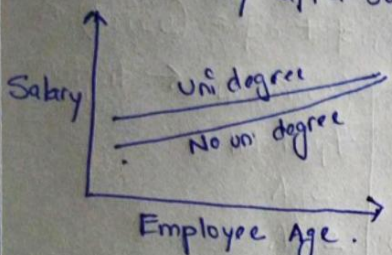
$Salary_i = \beta_0 + \beta_1 (Employee\ Age) + \beta_2 (Uni\ degree) + \varepsilon$.

β_2 will define how much a salary difference is there, if we have a uni degree with no uni degree.

If we plot,



But, in reality after some experience, university degree don't matter, its about experience.



So get the experience term, we need another variable.
like, Employee Age × Uni degree.

$Salary = \beta_0 + \beta_1 (Employee\ Age) + \beta_2 (Uni\ degree) + \beta_3 (Employee\ Age) \times (Uni\ degree) + \varepsilon_i$

The added term is called Interaction term.

Required when, X1 affects the relationship between X2 and Y.
Example, Age of employee affects the relationship between Uni degree & Salary.
Common misconception → An interaction term is required when X_1 & X_2 are correlated.

Model output →

Ln (Price)	Coeff	SE	t	P-value
Intercept	9.125	0.274	33.28	0.00
Age Cat 2	-0.181	0.23	-0.76	0.14
Age Cat 3	-0.800	0.25	-3.18	0.00
Age Cat 4	-0.390	0.42	-0.92	0.35
Ln (Odometer)	-0.290	0.059	-3.53	0.00
Pink slip	0.123	0.18	0.68	0.50
Pink slip × Age Cat 4	1.371	0.45	3.02	0.003

Why we took Age Cat 4?
→ Here come business/domain knowledge
In this case, we thought vintage cars might cost more.
But if vintage cars and also road worthy then it can add more. (we can drive that car now also).
→ P value is also showing it is important variable.

Interaction effect → Happens when one independent var interact with other var affect dependent variable.

Example →

Eating Pill \ Drinking diet juice	Yes	No
Yes	5 lbs	2 lbs
NO	2 lb	1 lb

Weight loss after 1 week

Highest weight loss is happening when we are taking both pill and drink together.

→ So we can find this through factor analysis.
→ Factor analysis extracts maximum common variance from all variable and put them into common score.
→ Methods which can be used as Factor analysis are –
① Principal component analysis (PCA)
② Maximum likelihood method.
③ Common factor analysis.