MBA 753 : Causal Inference Methods for Business Analytics

Dr. Nivedita Bhaktha

05.08.2024

Agenda

- Fundamentals of multiple linear regression
- Categorical predictors and its interpretation
- Interaction effects and its interpretation

Fundamentals of Multiple Linear Regression

Multiple linear regression model

- Extends SLR models to accommodate multiple independent variables that are associated with the single dependent variable

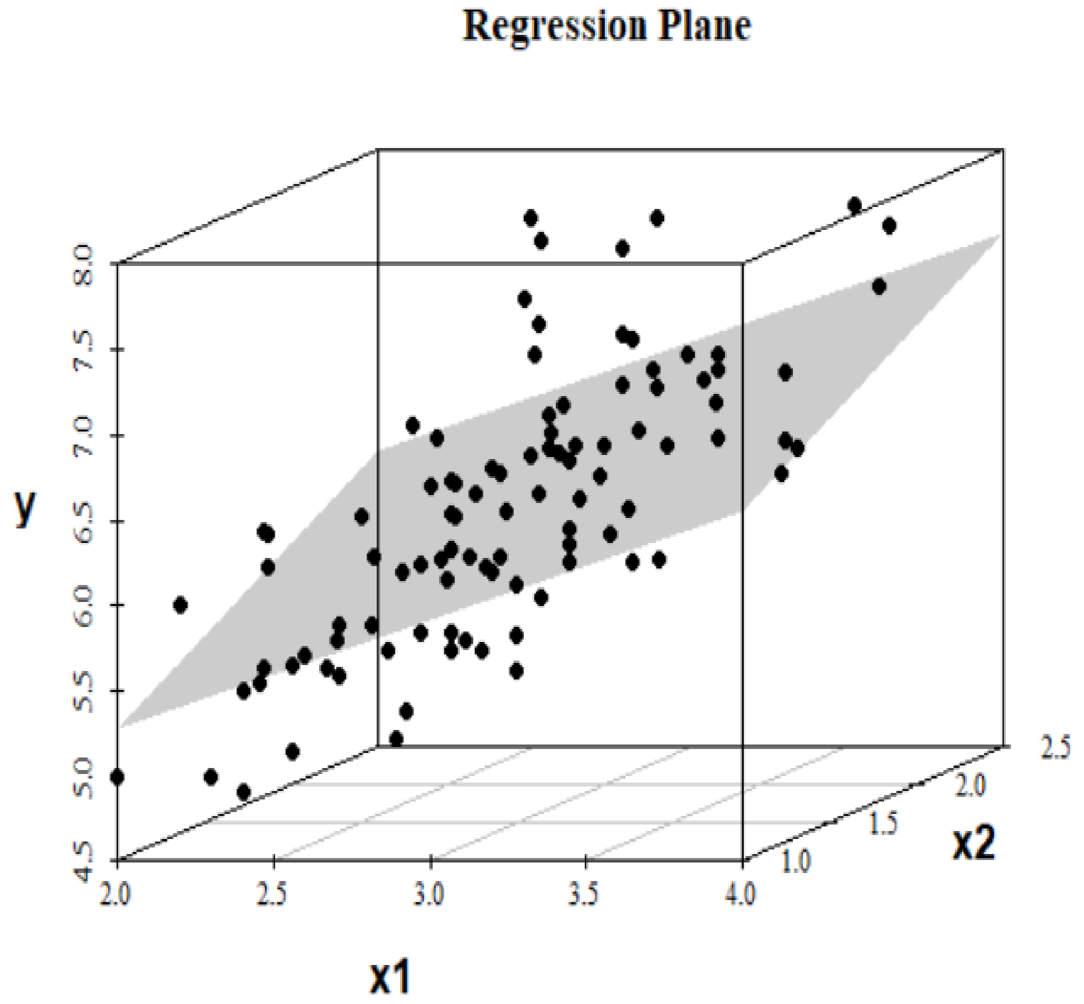
$$\mathbf{y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} \ + \ \varepsilon_i; \quad i = 1, \ldots, n}$$

- Describes the relationship in the population
- Parameters are estimated through sample and assumptions

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \cdots + \hat{\beta}_k x_{ki}$$

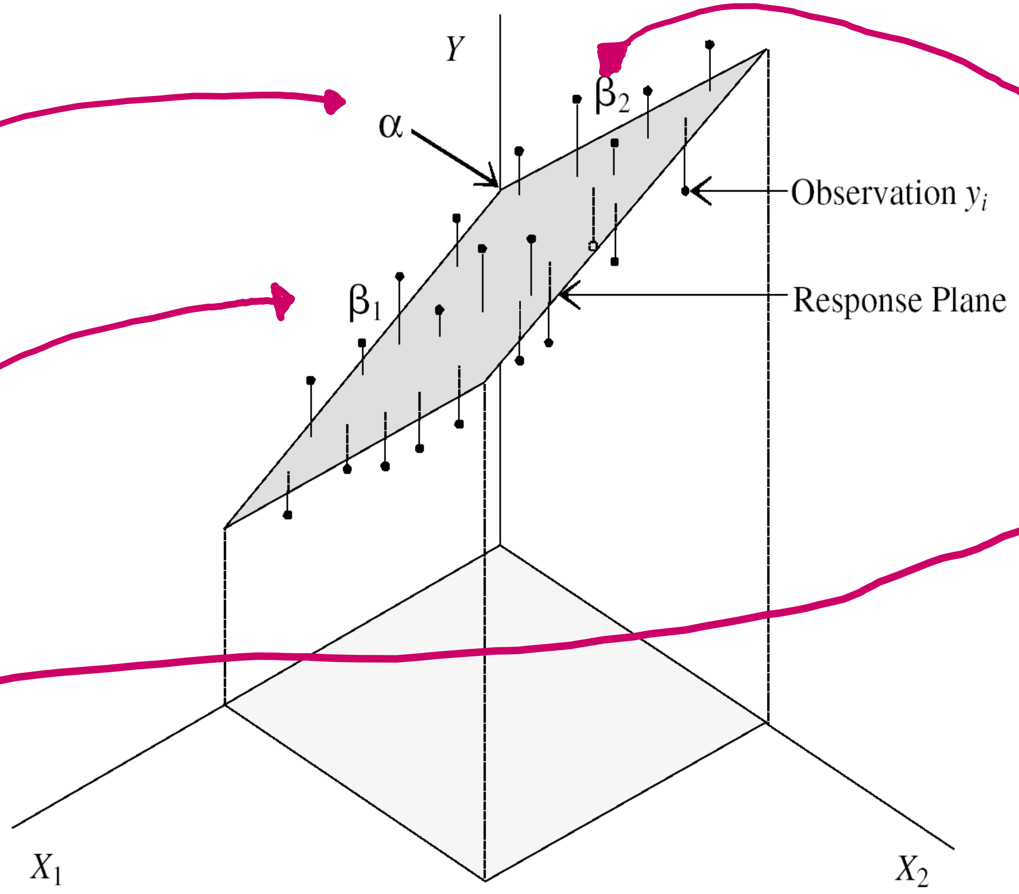
MLR model -II

- Residuals: observed – predicted
- Consider a simple case with two explanatory variables:
- *Idea*: Fit a regression plane in 3D space
- OLS estimation is used



MLR model -III

- Intercept α predicts where the regression *plane* crosses the Y axis
- Slope for variable X_1 (β_1) predicts the change in Y per unit X_1 holding X_2 constant
- The slope for variable X_2 (β_2) predicts the change in Y per unit X_2 holding X_1 constant



MLR - Interpretation

- Ceteris Paribus - all else being equal

$$\Delta\hat{y} = \hat{\beta}_1 \Delta x_1 + \hat{\beta}_2 \Delta x_2$$

- $\hat{\beta}_1$ and $\hat{\beta}_2$ are called partial effects

- Interpret: $\hat{y} = 27 + 9x_1 + 12x_2$ where $\hat{y}$ is the predicted sales ($1000s), x_1 is the capital investments ($1000s) and x_2 is the marketing expenditure ($1000s)

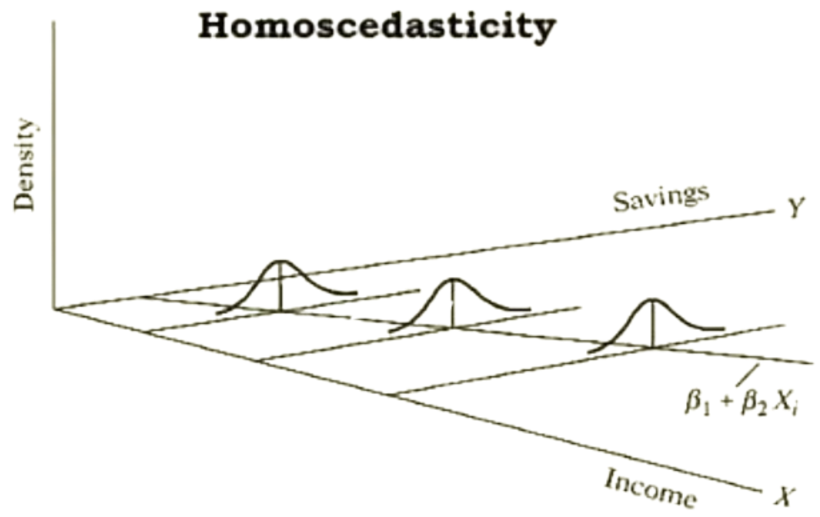
MLR - Assumptions

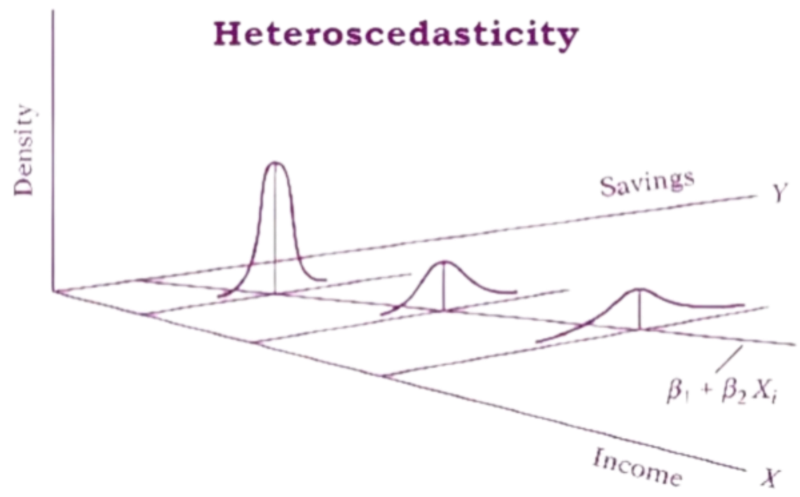
- Random and independent samples: $(y_i, x_{1i}, x_{2i}, \ldots, x_{ki})$; *i* = 1, …, *n*
- Linearity in parameter: $y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + v_i$
- Zero conditional mean and normal distribution for error
  - We assume $\varepsilon_i \sim N(0, \sigma^2)$ $and\ cov(\varepsilon_i, \varepsilon_j) = 0$
  - The conditional mean: $E(\varepsilon \mid x_1, \ldots, x_k) = E(\varepsilon) = 0$ for all $(x_1, \ldots, x_k)$
  - Error and predictors are independent: $cov(\varepsilon,\ x_1) = \cdots = cov(\varepsilon,\ x_k) = 0$

MLR – Assumptions II

- Homoscedasticity: variance of error does not change with Xs

$$var(\varepsilon \mid x_1, \dots, x_k) = var(\varepsilon) = \sigma^2 < \infty$$





MLR – Assumptions III

- Multicollinearity: No perfect collinearity
  - No perfect relationship among the predictors
    $$Cor(x_i, x_j) \neq \pm 1$$
  - None of the predictors are constant
  - In case of perfect collinearity, OLS estimation will not work
  - Even high correlation among predictors leads to unstable coefficients
    - Example: Using both total number of rooms and number of bedrooms as explanatory variables in same model

Goodness of Fit

- $SST = SSR + SSE$
  - SST: *Total sum of squares*
  - SSR: *Sum of squares due to regression*
  - SSE: *Error sum of squares*
- R^2: Proportion of variance in Y accounted for by the set of Xs

- Any additional independent variable in the model will increase SSR i.e.

Goodness of fit - II

- Adjusted R^2
  - Modified version of R^2 that adjusts for non-significant predictors
  - Corrects for overestimation by taking into account -
    - Sample size
    - Number of independent variables
  - Adjusted R^2 might decrease if a specific effect does not improve the model

$$R^2_{adj} = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)}$$

Significance Tests

- Individual regression coefficients

$H_0: \beta_j = a$ v/s $H_1: \beta_j \neq a$; $j = 1, \dots, K$ Generally $a = 0$

we can use t-test statistic such that

$$t_0 = \frac{\hat{\beta}_j - a}{\widehat{SE}(\hat{\beta}_j)} \sim t_{n-K-1}$$

- Overall regression significance

$H_0: \beta_1 = \dots = \beta_K = 0$ v/s H_1: atleast one $\beta_j \neq 0$; $j = 1, \dots, K$

we can use F-statistic such that

$$F_0 = \frac{SSR/K}{SSE/n-K-1} \sim F_{K, n-K-1}$$

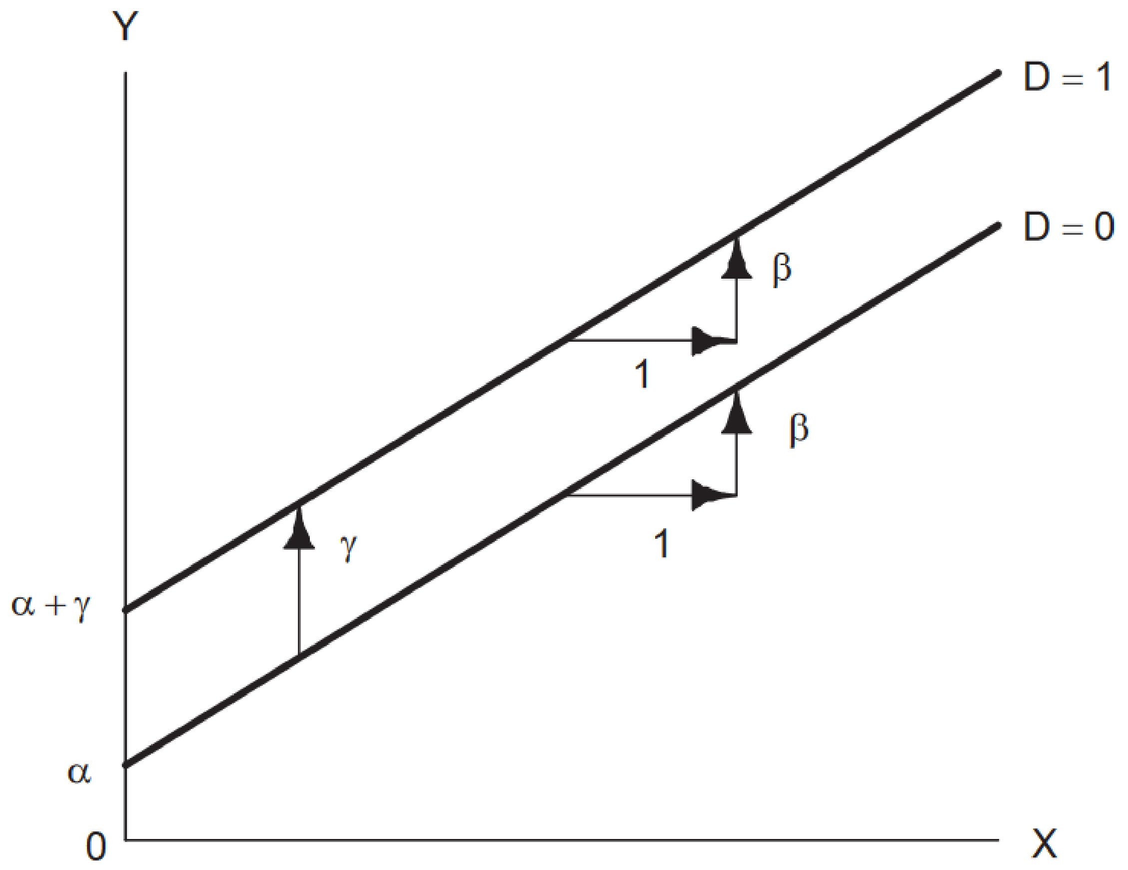
MLR – R output

```
Call:
lm(formula = sales ~ youtube + facebook + newspaper, data = marketing)

Residuals:
     Min       1Q   Median       3Q      Max 
-10.5932  -1.0690   0.2902   1.4272   3.3951 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3.526667   0.374290   9.422   <2e-16 ***
youtube      0.045765   0.001395  32.809   <2e-16 ***
facebook     0.188530   0.008611  21.893   <2e-16 ***
newspaper   -0.001037   0.005871  -0.177     0.86    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.023 on 196 degrees of freedom
Multiple R-squared:  0.8972,	Adjusted R-squared:  0.8956 
F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

Categorical Predictors

Categorical Predictor Variables

- Categorical independent variables can be incorporated into a regression model by converting them into 0/1 ("dummy") variables
  - Involves categorical *X* variable with two levels
  - Assumes only intercept is different
    - Slopes are constant across categories



Dummy Regressors

- Dummy regressors are easily extended to explanatory variables with more than two categories
  - A variable with m categories has m – 1 regressors
  - As with the two-category case, one of the categories is a reference group (coded 0 for all dummy regressors)

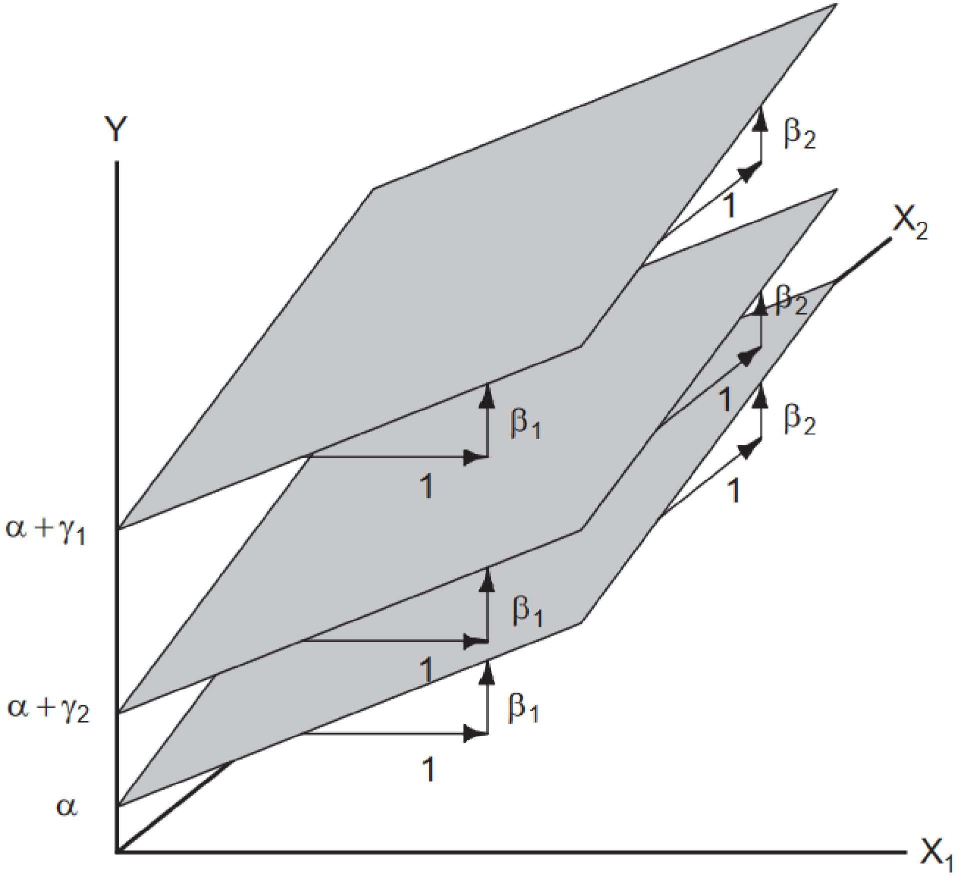
	D_1	D_2
Blue Collar	1	0
Professional	0	1
White Collar	0	0

Dummy Regression Model

$$Y_i = \alpha + \beta X_i + \gamma_1 D_{i1} + \gamma_2 D_{i2} + \varepsilon_i$$

- This gives three parallel regression lines

$$\text{Blue Collar: } Y_i = (\alpha + \gamma_1) + \beta X_i + \varepsilon_i$$
$$\text{Professional: } Y_i = (\alpha + \gamma_2) + \beta X_i + \varepsilon_i$$
$$\text{White Collar: } Y_i = \alpha + \beta X_i + \varepsilon_i$$



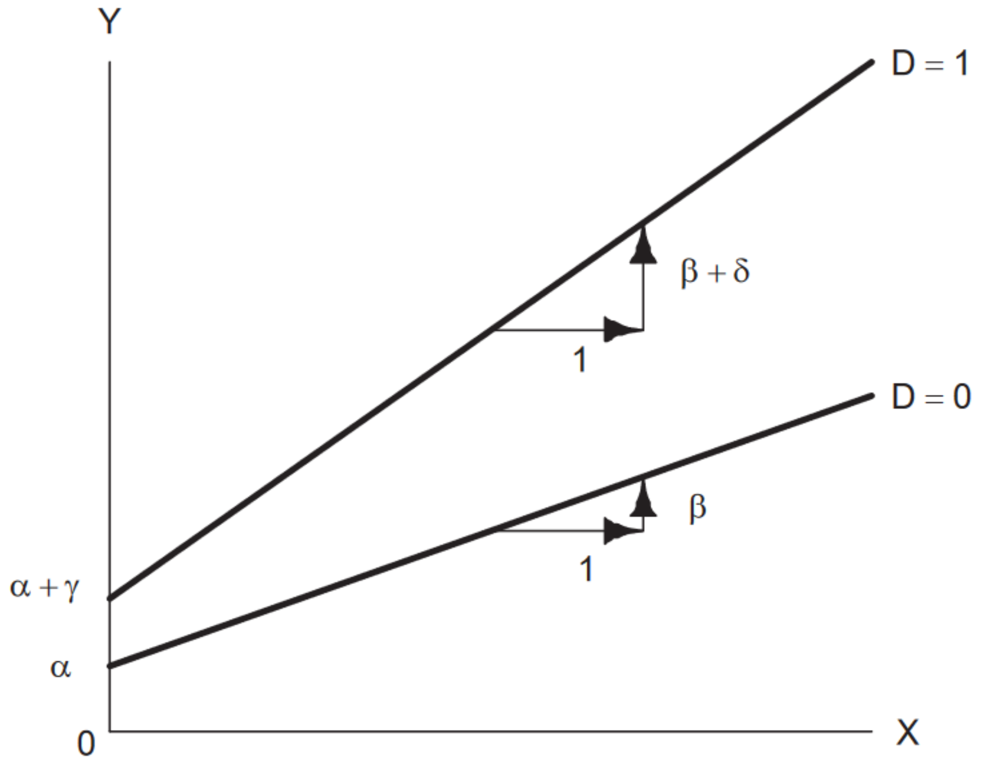
Interaction Effect and its Interpretation

Interaction Effect

- Two predictor variables "interact" when the partial effect of one variable depends on the value of another variable
  - For example, testing whether age effects are different for men (coded 1) and women (coded 0)
  - Separate models cannot test for differences among groups
  - Testing for differences in slope

$$Y_i = \alpha + \beta X_i + \gamma D_i + \delta(X_i D_i) + \varepsilon_i$$

$$\text{income}_i = \alpha + \beta \ \text{age}_i + \gamma \ \text{men}_i + \delta(\text{age}_i \times \text{men}_i) + \varepsilon_i$$



Interaction Interpretation

- When the interaction effect is significant
  - The unique partial effects (for example that of age and gender) are no longer interpretable just by themselves

- Omitting interaction effects can lead to erroneous conclusions

Recap

Summary

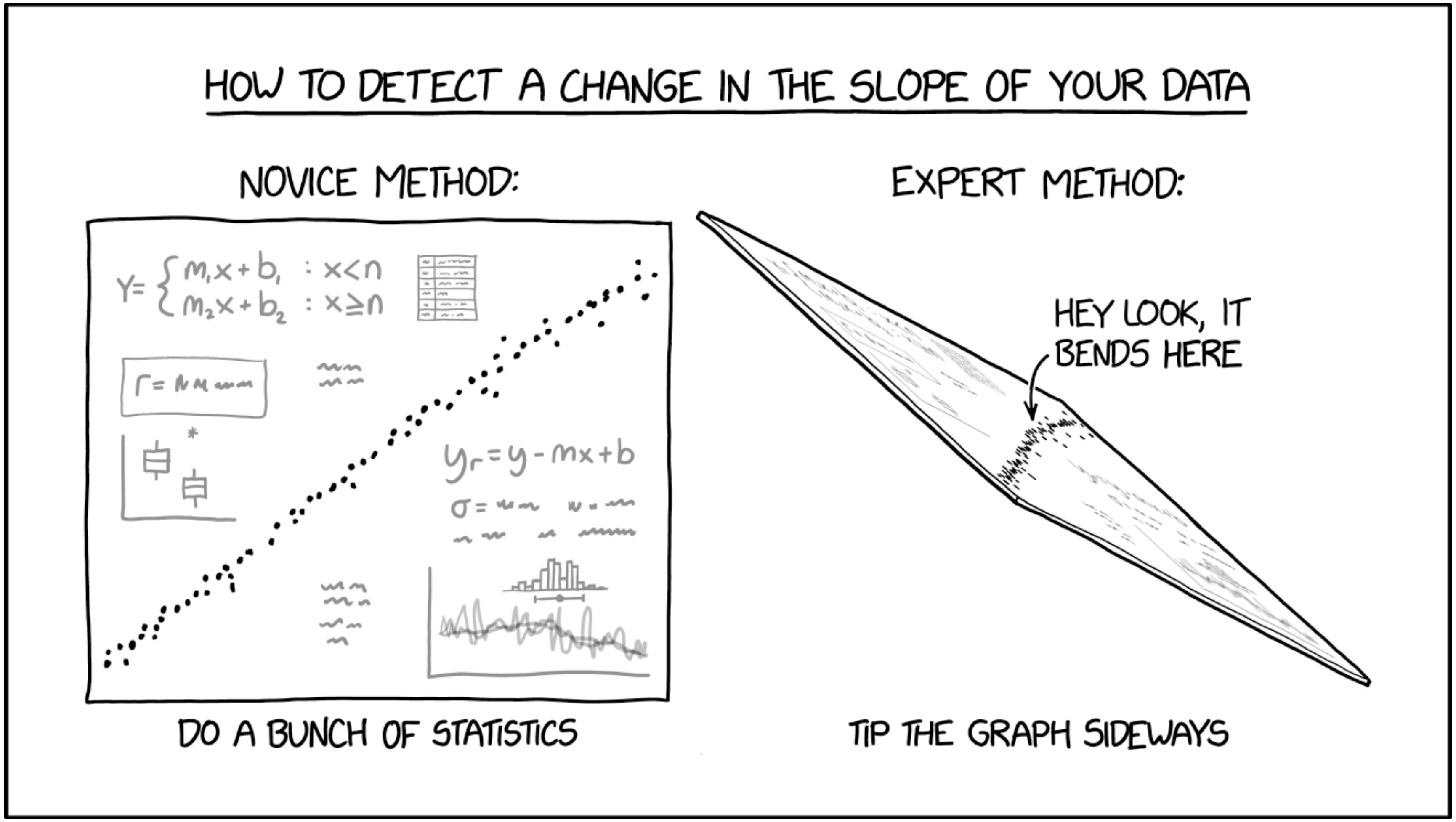
- MLR – fitting the best regression space
- Partial effects are estimated assuming ceteris paribus
- A categorical predictor with m groups will have m-1 regressors
- Interaction effects - when effect of one predictor depends on the values of the other

- Objectives achieved:
  - Can understand and interpret “effects” in a MLR model with categorical predictors and interactions
  - Can fit models and perform model diagnostics in a statistical software

References

- Model diagnostics for MLR in R: https://sscc.wisc.edu/sscc/pubs/RegDiag-R/index.html
- Stock, J. H., Watson, M. W., Wooldridge, J. M., & Wooldridge, J. M. Introductory Econometrics: A Modern Approach (4th Edition International).
- Neter J, Kutner MH, Nachtsheim CJ, Wasserman W. Applied linear statistical models. McGraw Hill Education.
- Scott Cunningham, Causal Inference: The Mix Tape, Yale University Press.

Thank You ☺



HOW TO DETECT A CHANGE IN THE SLOPE OF YOUR DATA
NOVICE METHOD:
EXPERT METHOD:
$Y = \begin{cases} m_1x+b_1 & : x<n \\ m_2x+b_2 & : x\geq n \end{cases}$
$y_r = y - mx + b$
HEY LOOK, IT BENDS HERE
DO A BUNCH OF STATISTICS
TIP THE GRAPH SIDEWAYS