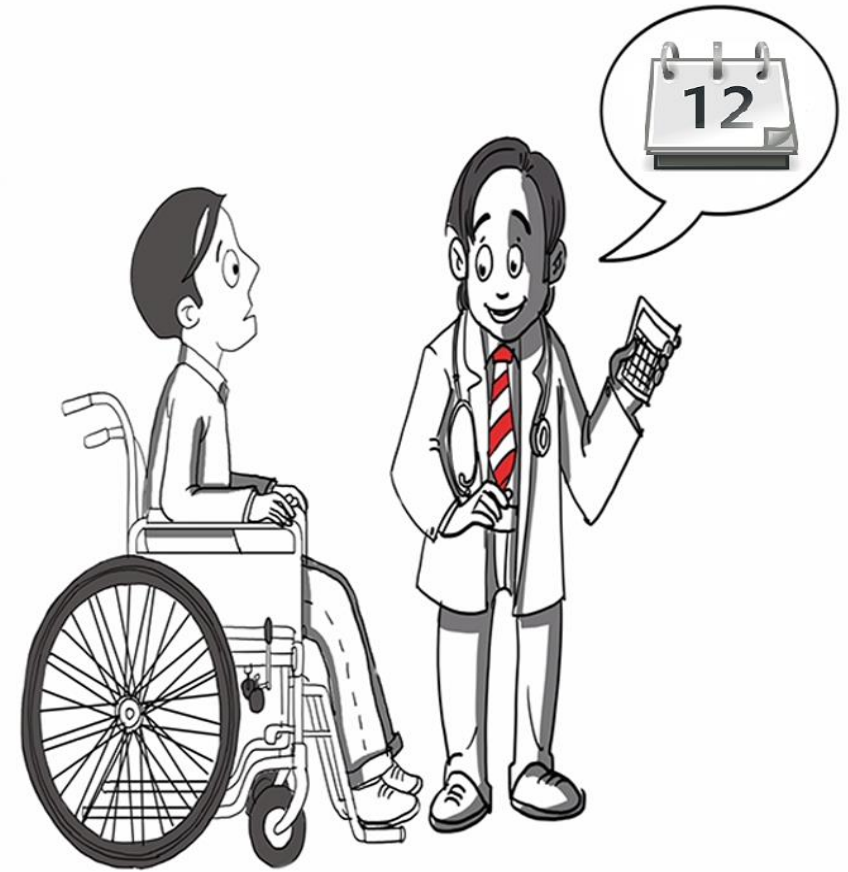




Introduction to Multiple Regression

U. DINESH KUMAR



What is Multiple Linear Regression

- Several independent variables may influence the change in response variable we are trying to study.
- When several independent variables are included in the equation, the regression is called multiple linear regression.

Multiple Regression Modeling Steps

1. Start with a hypothesis or belief.
2. Estimate unknown model parameters ($\beta_0, \beta_1, \beta_2, \dots$)
3. Specify probability distribution of random error term
 - Assumed to be a normal distribution
4. Check the assumptions of regression (normality, heteroscedasticity and multi-collinearity)
5. Evaluate Model
6. Use Model for Prediction & Estimation and other purposes.

Multiple Linear Regression Model

Predictive Analytics : QM901.1x
Prof U Dinesh Kumar, IIMB

- Relationship between 1 dependent & 2 or more independent variables is a linear function.

The diagram shows the Multiple Linear Regression equation:
$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_k X_{ki} + \varepsilon_i$$
 with the following labels and arrows:

- Population Y-intercept**: Points to β_0 .
- Population slopes**: Points to β_1 and β_2 .
- Random error**: Points to ε_i .
- Dependent (response) variable**: Points to Y_i .
- Independent (explanatory) variables**: Points to X_{1i} , X_{2i} , and X_{ki} .

Prediction Model

Predictive Analytics : QM901.1x
Prof U Dinesh Kumar, IIMB

The prediction equation obtained from sample data is:

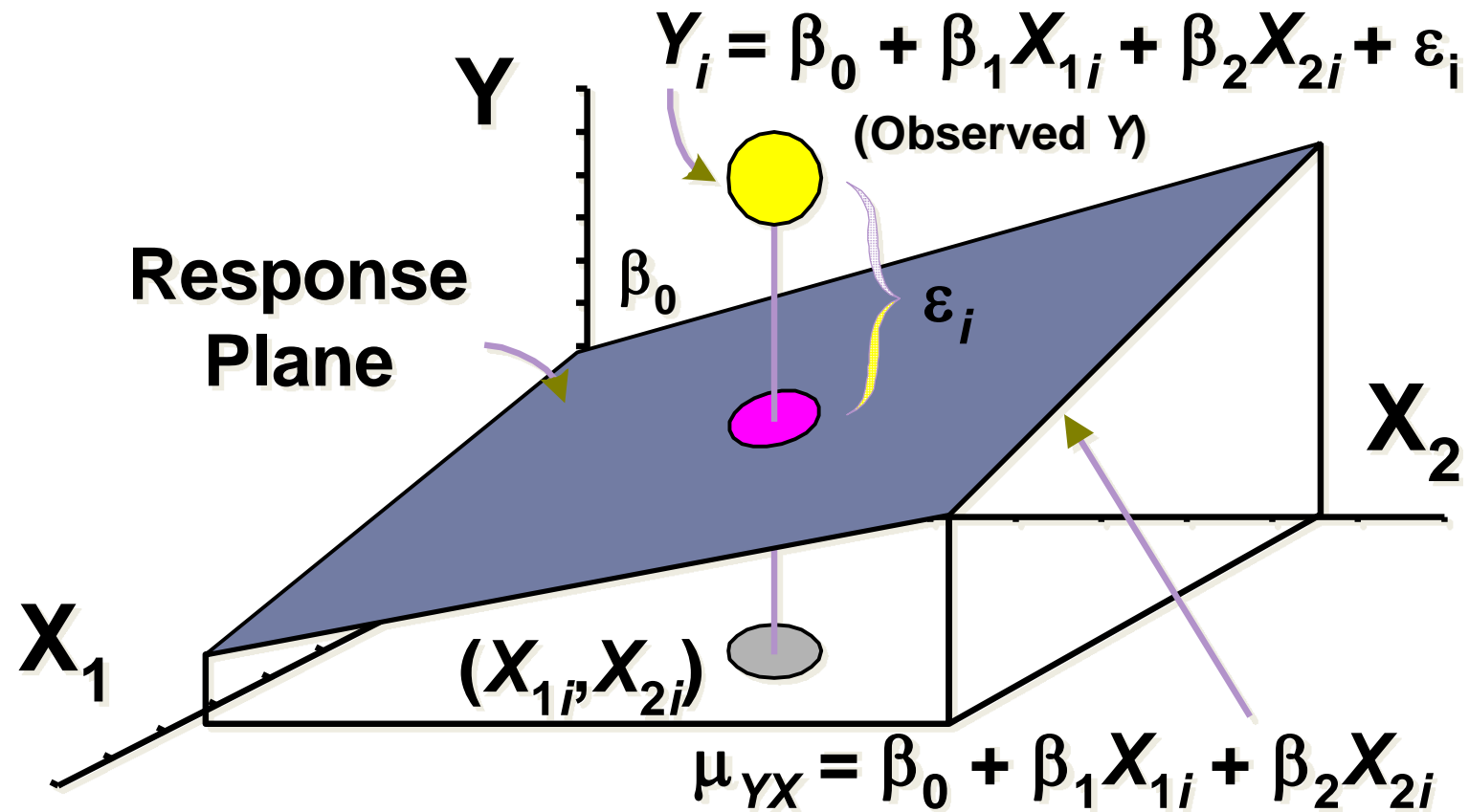
$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \cdots + \hat{\beta}_k X_{ki}$$

Multiple Regression Model

Y_i = Treatment Cost

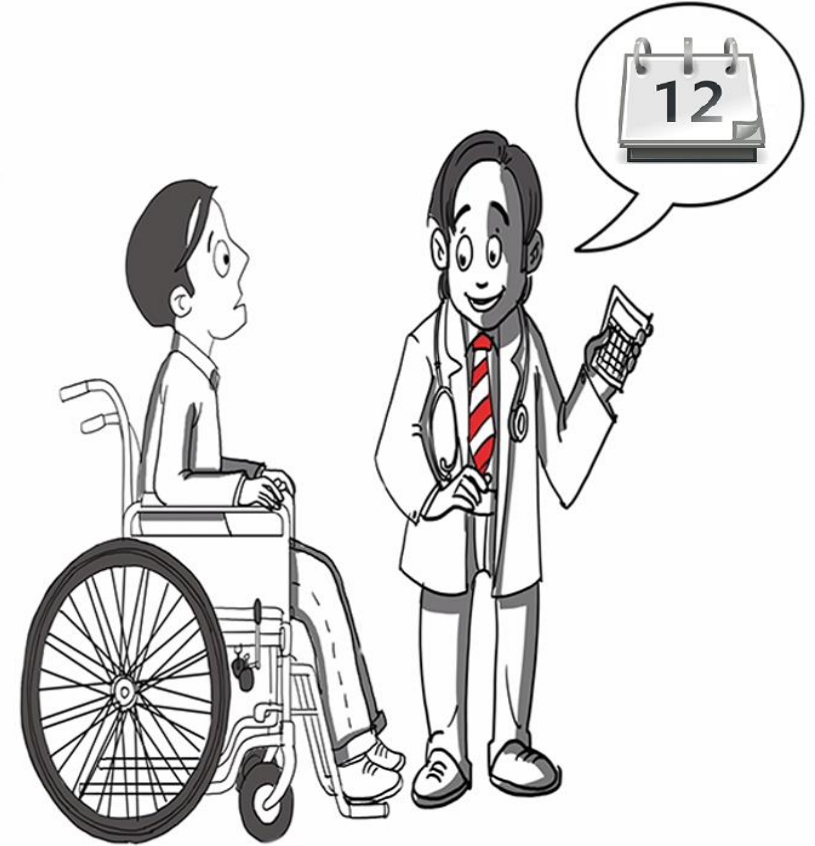
X_{1i} = Body Weight

X_{2i} = HR Pulse



Interpretation of Regression Coefficients

U. DINESH KUMAR



Estimating and Interpreting the β -Parameters

- Belief

$$Y = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k$$

- Fitted Model

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \dots + \hat{\beta}_k x_k$$

- Using least squares method

$$\text{Minimize } SSE = \sum (y - \hat{y})^2$$

Estimation of Parameters in Multiple Regression

Predictive Analytics : QM901.1x
Prof U Dinesh Kumar, IIMB

- The least squares function is given by

$$SSE = \sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^k \beta_j x_{ij} \right)^2 \quad \text{—————} \quad (1)$$

- The least squares estimates must satisfy

$$\frac{\partial SSE}{\partial \beta_0} = -2 \sum_{i=1}^n \left(y_i - \hat{\beta}_0 - \sum_{j=1}^k \hat{\beta}_j x_{ij} \right) = 0 \quad \text{—————} \quad (2)$$

and

$$\frac{\partial SSE}{\partial \beta_j} = -2 \sum_{i=1}^n \left(y_i - \hat{\beta}_0 - \sum_{j=1}^k \hat{\beta}_j x_{ij} \right) x_{ij} = 0 \quad \forall j \quad \text{—————} \quad (3)$$

Estimation of Parameters in Multiple Regression

Predictive Analytics : QM901.1x
Prof U Dinesh Kumar, IIMB

The least squares equations are

$$\begin{aligned} n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^n x_{i1} + \hat{\beta}_2 \sum_{i=1}^n x_{i2} + \cdots + \hat{\beta}_k \sum_{i=1}^n x_{ik} &= \sum_{i=1}^n y_i \\ \hat{\beta}_0 \sum_{i=1}^n x_{i1} + \hat{\beta}_1 \sum_{i=1}^n x_{i1}^2 + \hat{\beta}_2 \sum_{i=1}^n x_{i1} x_{i2} + \cdots + \hat{\beta}_k \sum_{i=1}^n x_{i1} x_{ik} &= \sum_{i=1}^n x_{i1} y_i \\ \vdots & \\ \hat{\beta}_0 \sum_{i=1}^n x_{ik} + \hat{\beta}_1 \sum_{i=1}^n x_{ik} x_{i1} + \hat{\beta}_2 \sum_{i=1}^n x_{ik} x_{i2} + \cdots + \hat{\beta}_k \sum_{i=1}^n x_{ik}^2 &= \sum_{i=1}^n x_{ik} y_i \end{aligned}$$

The solution to the above equations are the **least squares estimators** of the regression coefficients.

DAD Example

$$\text{Total Hospital Cost} = \beta_0 + \beta_1 \times \text{Body weight} + \beta_2 \times \text{Body height} + \beta_3 \times \text{HR Pulse} + \beta_4 \times \text{RR}$$

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.423 ^a	.179	.166	111968.4772 4327330000	.179	13.268	4	243	.000

a. Predictors: (Constant), RR, BODYHEIGHT, HRPULSE, BODYWEIGHT

b. Dependent Variable: TOTALCOSTTOHOSPITAL

MODEL

Predictive Analytics : QM901.1x
Prof U Dinesh Kumar, IIMB

Coefficients^a

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations			Collinearity Statistics	
	B	Std. Error	Beta			Zero-order	Partial	Part	Tolerance	VIF
(Constant)	-118569.892	70668.651		-1.678	.095					
1 BODYWEIGHT	2512.260	604.558	.474	4.156	.000	.348	.258	.242	.260	3.849
BODYHEIGHT	161.083	348.148	.051	.463	.644	.294	.030	.027	.273	3.664
HRPULSE	1537.105	445.829	.255	3.448	.001	-.009	.216	.200	.619	1.615
RR	2560.631	2038.500	.080	1.256	.210	.016	.080	.073	.828	1.208

a. Dependent Variable: TOTALCOSTTOHOSPITAL

$$\text{Treatment Cost} = -118569.892 + 2512.260 \times \text{Body weight} + 161.083 \times \text{Body height} + 1537.105 \times \text{HR Pulse} + 2560.631 \times \text{RR}$$

Interpretation of Coefficients

- (β_1) – Partial regression coefficient

Total cost of treatment increases by 2512.260 INR for every one kg increase in body weight while other parameters like body height, HR pulse and RR are controlled (kept constant).

$$\text{Treatment Cost} = -118569.892 + 2512.260 \times \text{Body weight} + 161.083 \times \text{Body height} + 1537.105 \times \text{HR Pulse} + 2560.631 \times \text{RR}$$

Interpretation of Coefficients

- (β_3) – Partial regression coefficient

Total cost of treatment increases by 1537.105 INR for every one unit increase in HR pulse while other parameters like body weight, Body height and RR are controlled.

$$\text{Treatment Cost} = -118569.892 + 2512.260 \times \text{Body weight} + 161.083 \times \text{Body height} + 1537.105 \times \text{HR Pulse} + 2560.631 \times \text{RR}$$

Standardized Beta (Beta Weights)

- The **beta weights** are the regression coefficients for standardized data.
- Standardized Beta is the average increase in the dependent variable when the independent variable is increased by one standard deviation and other independent variables are held constant.

Beta Weights

$$\text{Standardized Beta} = \beta_i \times \frac{S_x}{S_y}$$

S_x = Standard deviation of X

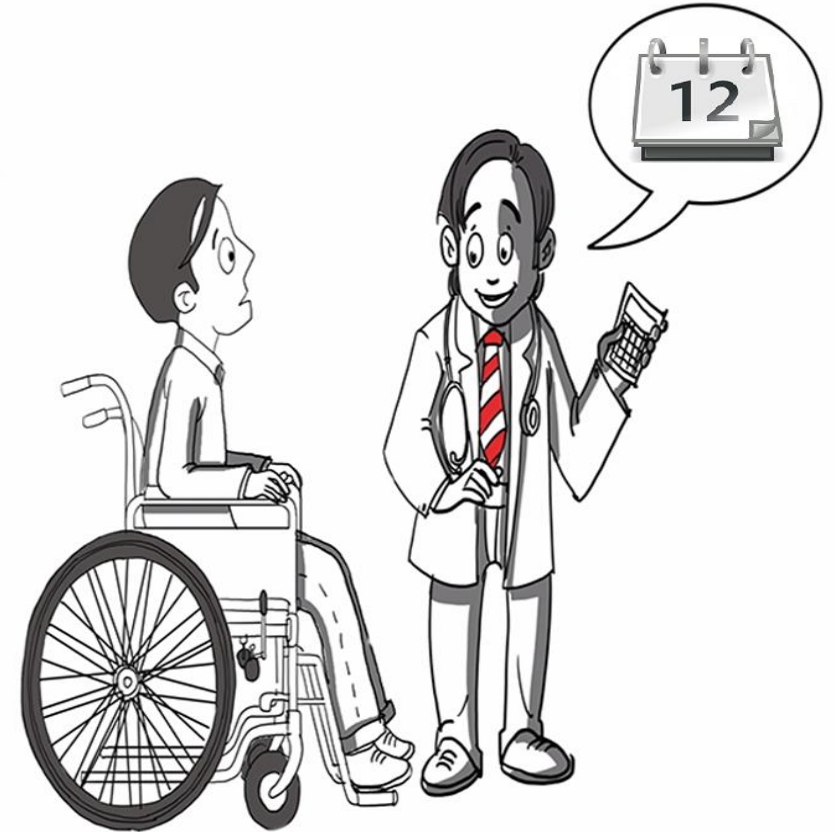
S_y = Standard deviation of Y

Coefficients^a

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations			Collinearity Statistics	
	B	Std. Error	Beta			Zero-order	Partial	Part	Tolerance	VIF
(Constant)	-118569.892	70668.651		-1.678	.095					
1 BODYWEIGHT	2512.260	604.558	.474	4.156	.000	.348	.258	.242	.260	3.849
BODYHEIGHT	161.083	348.148	.051	.463	.644	.294	.030	.027	.273	3.664
HRPULSE	1537.105	445.829	.255	3.448	.001	-.009	.216	.200	.619	1.615
RR	2560.631	2038.500	.080	1.256	.210	.016	.080	.073	.828	1.208

Partial and Part correlation

U. DINESH KUMAR

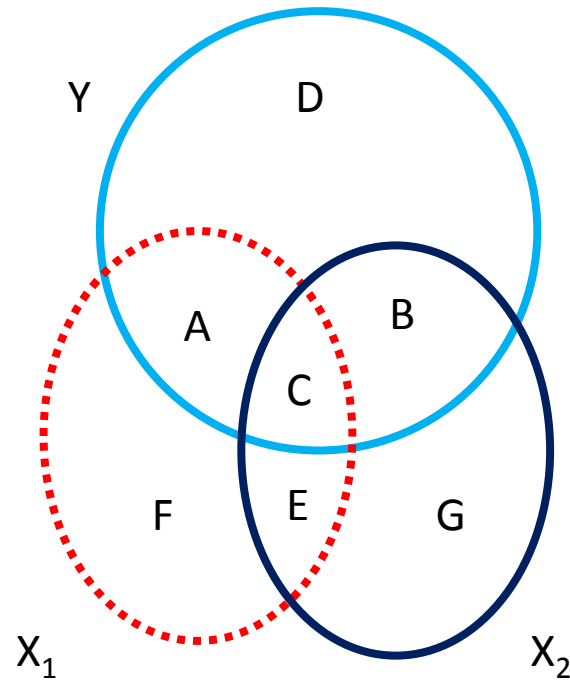


Partial Correlation

- **Partial correlation coefficient** measures the relationship between two variables (say Y and X_1) when the influence of all other variables (say X_2, X_3, \dots, X_k) connected with these two variables (Y and X_1) are removed.
- Partial correlation is the correlation between residualized response and residualized predictor.

Semi-Partial (Part) Correlation

- **Semi-partial (or part correlation) coefficient** measures the relationship between two variables (say Y and X_1) when the influence of all other variables (say X_2, X_3, \dots, X_k) connected with these two variables (Y and X_1) are removed from one of the variables (X_1).
- Only the predictor is residualized. The denominator of the coefficient (the total variance of the response variable, Y) remains the same no matter which predictor is being examined.



Variance in $Y = A + B + C + D$

Variance in Y explained by the model $= A + B + C$

Variance in Y not explained by the model $= E$

Semi partial (part) correlation for $X_1 = A / (A+B+C+D)$

Semi partial (part) correlation for $X_2 = B / (A+B+C+D)$

Partial correlation for $X_1 = A / (A+D)$

Partial correlation for $X_2 = B / (B + D)$

👉 IMPORTANT

In part correlation we do not residualize the response variable

Partial Correlation

$$r_{12,3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}}$$

Correlation between y1 and x2, when the influence of x3 is removed from both y1 and x2.

$$r_{13,2} = \frac{r_{13} - r_{12}r_{23}}{\sqrt{(1 - r_{12}^2)(1 - r_{23}^2)}}$$

Correlation between y1 and x3, when the influence of x2 is removed from both y1 and x3.

$$r_{23,1} = \frac{r_{23} - r_{12}r_{13}}{\sqrt{(1 - r_{12}^2)(1 - r_{13}^2)}}$$

Correlation between x2 and x3, when the influence of y1 is removed from both x2 and x3.

For Proof: Theory of Econometrics by A Koutsoyiannis

Semi-partial correlation (Part Correlation)

Predictive Analytics : QM901.1x
Prof U Dinesh Kumar, IIMB

- Semi-partial (or part) correlation $sr_{12,3}$ is the correlation between y_1 and x_2 when influence of x_3 is partialled out of x_2 (not on y_1).

$$sr_{12,3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1 - r_{23}^2}}$$

$$\text{Part Correlation} = \frac{\text{Partial Correlation}}{\sqrt{(1 - r_{13}^2)}}$$

Semi-partial (part) correlation

Square of Semi-partial (or part correlation) gives the increase in the R^2 value (coefficient of multiple determination) when an explanatory variable is added to the model.

$$\underline{Y = \beta_0 + \beta_1 \times \text{Body weight} + \beta_2 \times \text{HR Pulse}}$$

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.416 ^a	.173	.167	111913.7163

a. Predictors: (Constant), HRPULSE, BODYWEIGHT

$$0.348^2 = 0.121$$

Coefficients^a

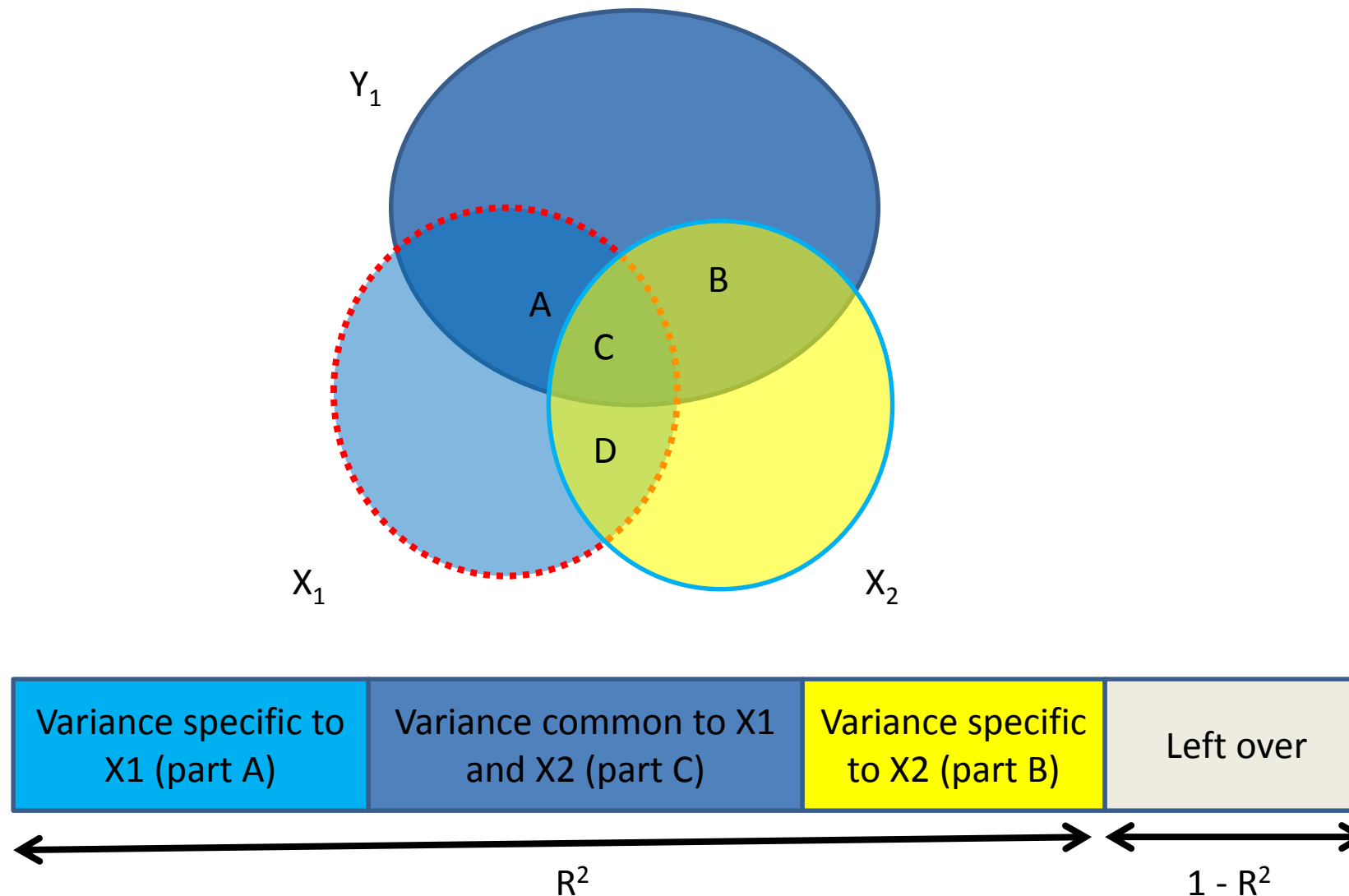
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations		
		B	Std. Error	Beta			Zero-order	Partial	Part
1	(Constant)	-55512.272	49018.015		-1.132	.259			
	BODYWEIGHT	2674.585	373.273	.504	7.165	.000	.348	.416	.416
	HRPULSE	1668.361	424.922	.276	3.926	.000	-.009	.243	.228

a. Dependent Variable: TOTALCOSTTOHOSPITAL

$$0.121 + (0.228)^2 = 0.173$$

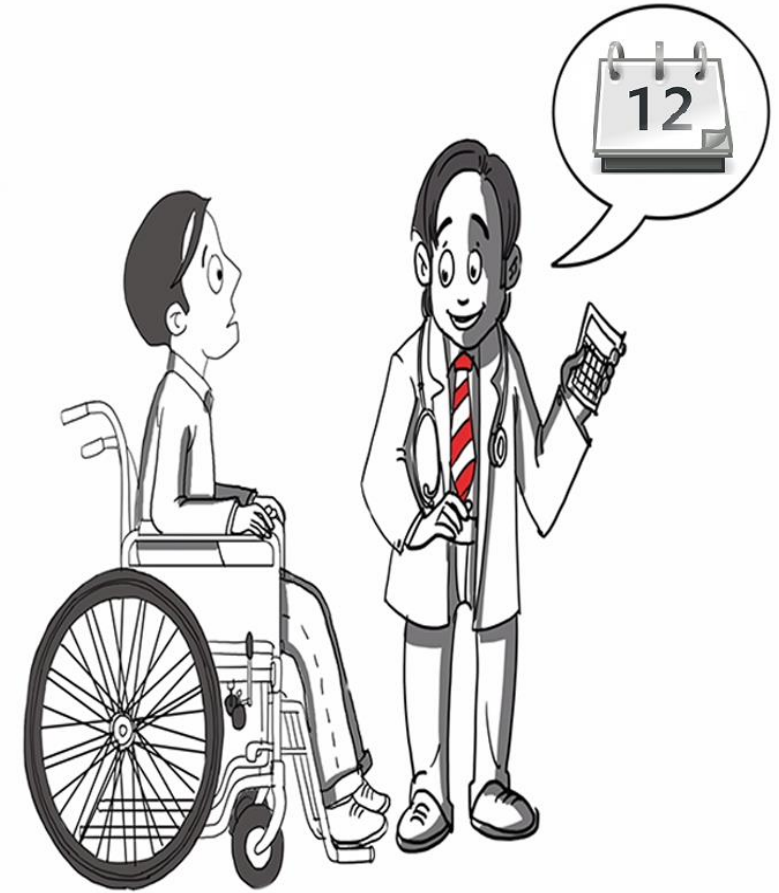
Change in R^2 when a new variable is added

Predictive Analytics : QM901.1x
Prof U Dinesh Kumar, IIMB



Model Diagnostics

U. DINESH KUMAR



Multiple Regression Model Diagnostics

Predictive Analytics : QM901.1x
Prof U Dinesh Kumar, IIMB

- Test for overall model fitness (R-Square and Adjusted R-Square)
- Test for overall model statistical significance (F test test)
- Test for portions of the model (Partial F-test)
- Test for statistical significance of individual explanatory variables (t test)
- Test for Normality and Homoscedasticity of residuals
- Test for Multi-collinearity and Auto Correlation

Co-efficient of multiple determination

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2}$$

R^2 = Multiple coefficient of multiple determination

R^2 is the percentage of the variation of y explained by
 x_1, x_2, \dots, x_k

$$\underline{Y = \beta_0 + \beta_1 \times \text{Body Weight} + \beta_2 \times \text{HR Pulse}}$$

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.416 ^a	.173	.167	111913.7163

a. Predictors: (Constant), HRPULSE, BODYWEIGHT

R^2 is 0.173
17.3% of variation in Y is
explained by Bodyweight and
HR Pulse

Coefficients^a

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
1 (Constant)	-55512.272	49018.015		-1.132	.259
BODYWEIGHT	2674.585	373.273	.504	7.165	.000
HRPULSE	1668.361	424.922	.276	3.926	.000

a. Dependent Variable: TOTALCOSTTOHOSPITAL

$$Y = -55512.272 + 2674.585 \times \text{Body Weight} + 1668.361 \times \text{HR Pulse}$$

Co-efficient of determination in Multiple Regression

- Coefficient of determination increases as the number of explanatory variables increases.
- In SSR/SST, the numerator, SSR, increases as the number of explanatory variables increases, whereas the denominator, SST, remains constant.
- Increase in R^2 can be deceptive, since more number of explanatory variables may over-fit the data.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.582 ^a	.338	.330	100326.5348

a. Predictors: (Constant), SL, HRPULSE, BODYWEIGHT

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	87329.003	47604.902		1.834	.068
	BODYWEIGHT	1982.956	346.170	.374	5.728	.000
	HRPULSE	1372.122	382.815	.227	3.584	.000
	SL	-719.406	92.216	-.421	-7.801	.000

a. Dependent Variable: TOTALCOSTTOHOSPITAL

Adjusted R^2

- Inclusion of additional explanatory variable will increase the R^2 value.
- By introducing an additional explanatory variable, we increase the numerator of the expression for R^2 while the denominator remains the same.
- To correct this defect, we adjust the R^2 by taking into account the degrees of freedom.

Adjusted R²

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

R_A^2 = Adjusted R - Square

n = number of observations

k = number of explanatory variables

$$\underline{Y = \beta_0 + \beta_1 \times \text{Body Weight} + \beta_2 \times \text{HR Pulse}}$$

Model Summary

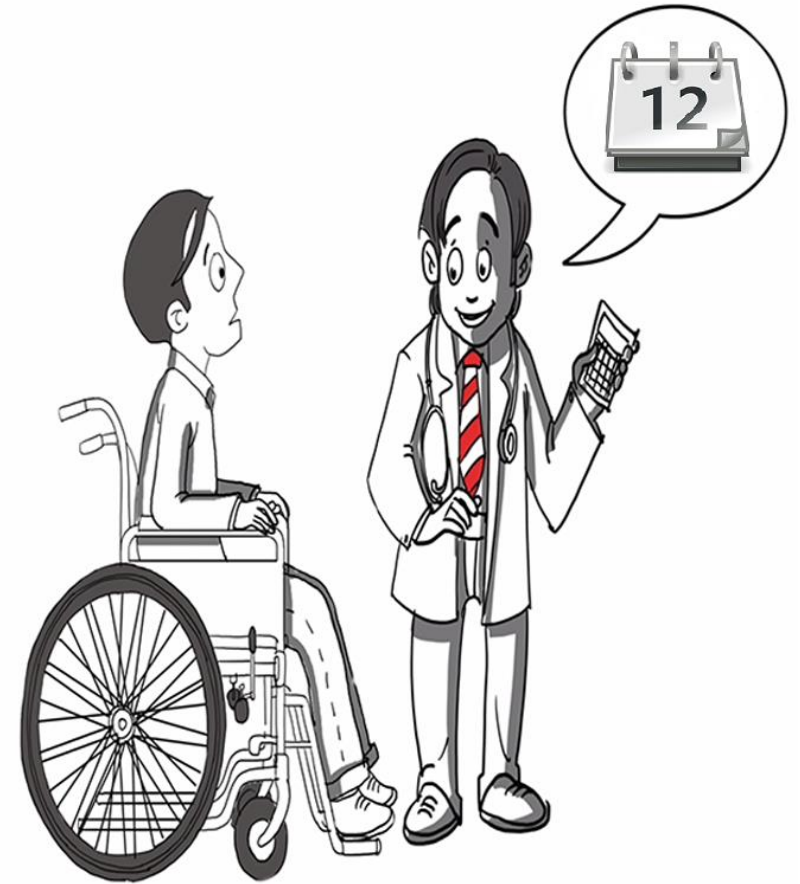
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.416 ^a	.173	.167	111913.7163

a. Predictors: (Constant), HRPULSE, BODYWEIGHT

Adjusted R² is 0.167

Tests for Model Significance

U. DINESH KUMAR



Testing For Overall Significance of Model – F Test

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_k = 0$$

H_A : Not all β values are zero

Test for overall significance of multiple regression model.

Checks if there is a statistically significant relationship between Y and any of the explanatory variables (X_1, X_2, \dots, X_k) .

F Statistic

$$F = MSR/MSE$$

Relationship between F and R²

$$F = \frac{R^2 / k}{(1 - R^2) / (n - k - 1)}$$

DAD CASE - ANOVA Table

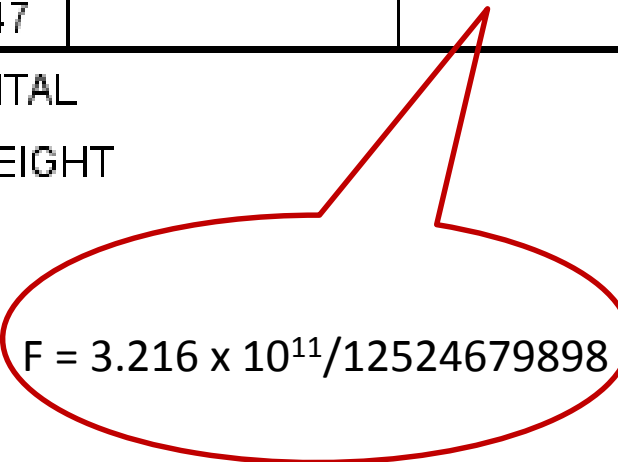
Predictive Analytics : QM901.1x
Prof U Dinesh Kumar, IIMB

ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	6.433E+11	2	3.216E+11	25.681	.000 ^b
	Residual	3.069E+12	245	12524679898		
	Total	3.712E+12	247			

a. Dependent Variable: TOTALCOSTTOHOSPITAL

b. Predictors: (Constant), HRPULSE, BODYWEIGHT


$$F = 3.216 \times 10^{11} / 12524679898$$

Test for Individual Variables

Testing for Significance of Individual Parameters

$$H_0 : \beta_i = 0$$

$$H_A : \beta_i \neq 0$$

$$t = \frac{\hat{\beta}_i}{S_e(\hat{\beta}_i)}$$

T- test:

By rejecting the null hypothesis, we can claim that there is a statistically significant relationship between the response variable Y and explanatory variable X_i .

SPSS Output for Coefficients

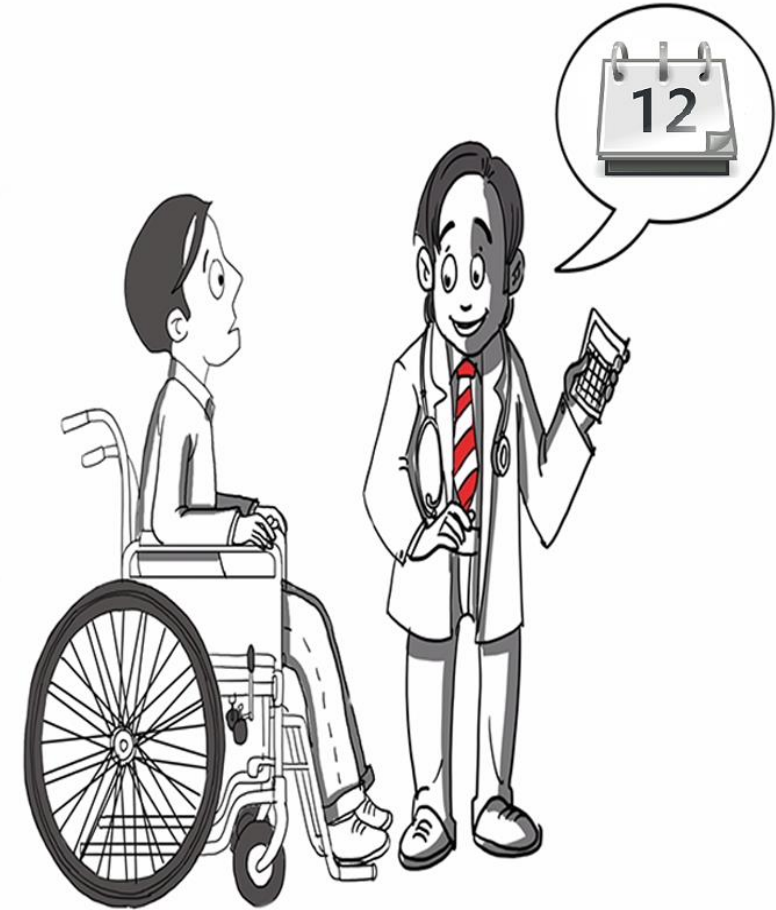
Coefficients^a

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
1	(Constant)	-118569.892	70668.651	-1.678	.095
	BODYWEIGHT	2512.260	604.558	.474	.000
	HRPULSE	1537.105	445.829	.255	.001
	BODYHEIGHT	161.083	348.148	.051	.644
	RR	2560.631	2038.500	.080	.210

a. Dependent Variable: TOTALCOSTTOHOSPITAL

Partial F-test and Variable Selection

U. DINESH KUMAR



Testing Model Portions – Partial F Test

- Examines the contribution of a set of explanatory variables to the regression model.
- Used for selecting explanatory variables in regression model building strategies such as stepwise regression.

Testing Model Portions – Partial F Test

Full Model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + \varepsilon$

Reduced Model ($r < k$): $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_r X_r + \varepsilon$

Test H_0 : $\beta_{r+1} = \dots = \beta_k = 0$

$$\text{Partial F} = \frac{(\text{SSE}_{\text{reduced}} - \text{SSE}_{\text{full}}) / (k - r)}{\text{MSE}_{\text{full}}}$$

k = Number of variables in the full model

r = Number of variables in the reduced model

Partial F-test Example (DAD)

Predictive Analytics : QM901.1x
Prof U Dinesh Kumar, IIMB

Full Model:

$$\text{Treatment cost} = \beta_0 + \beta_1 \times \text{Bodyweight} + \beta_2 \times \text{Bodyheight} + \beta_3 \times \text{HRpulse} + \beta_4 \times \text{RR} + \varepsilon$$

Reduced Model:

$$\text{Treatment cost} = \beta_0 + \beta_1 \times \text{Bodyweight} + \beta_2 \times \text{HRpulse} + \varepsilon$$

$$F = \frac{(SSE_{\text{reduced}} - SSE_{\text{full}}) / (4 - 2)}{MSE_{\text{full}}}$$

ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	6.654E+11	4	1.663E+11	13.268	.000 ^b
	Residual	3.046E+12	243	12536939896		
	Total	3.712E+12	247			

a. Dependent Variable: TOTALCOSTTOHOSPITAL

b. Predictors: (Constant), RR, BODYHEIGHT, HRPULSE, BODYWEIGHT

ANOVA^a

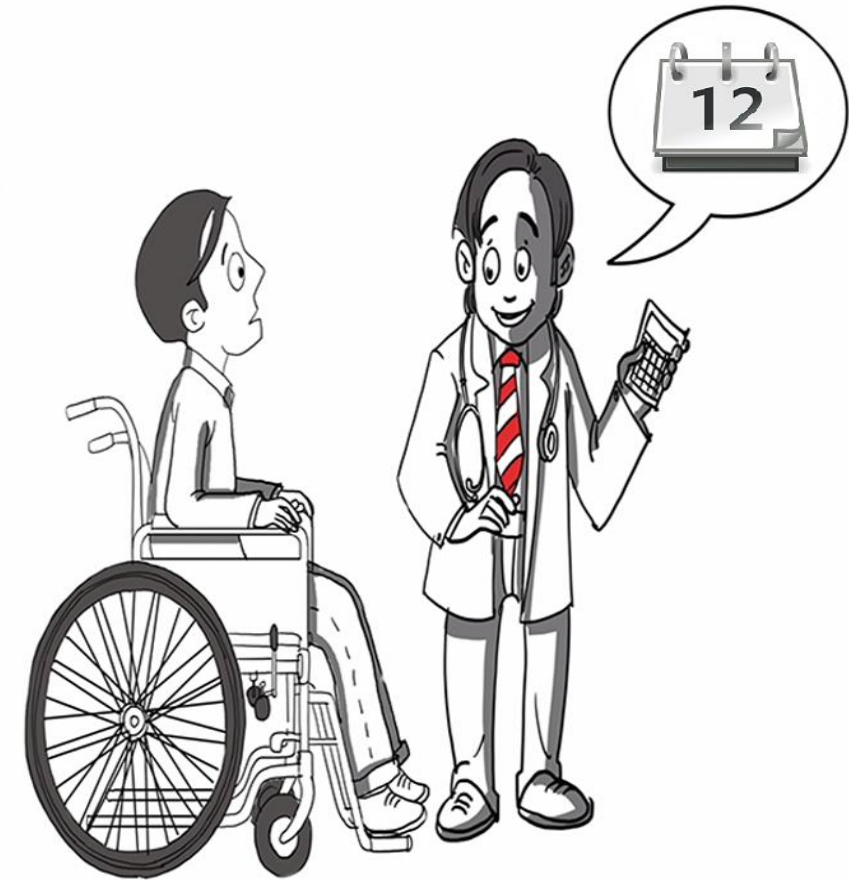
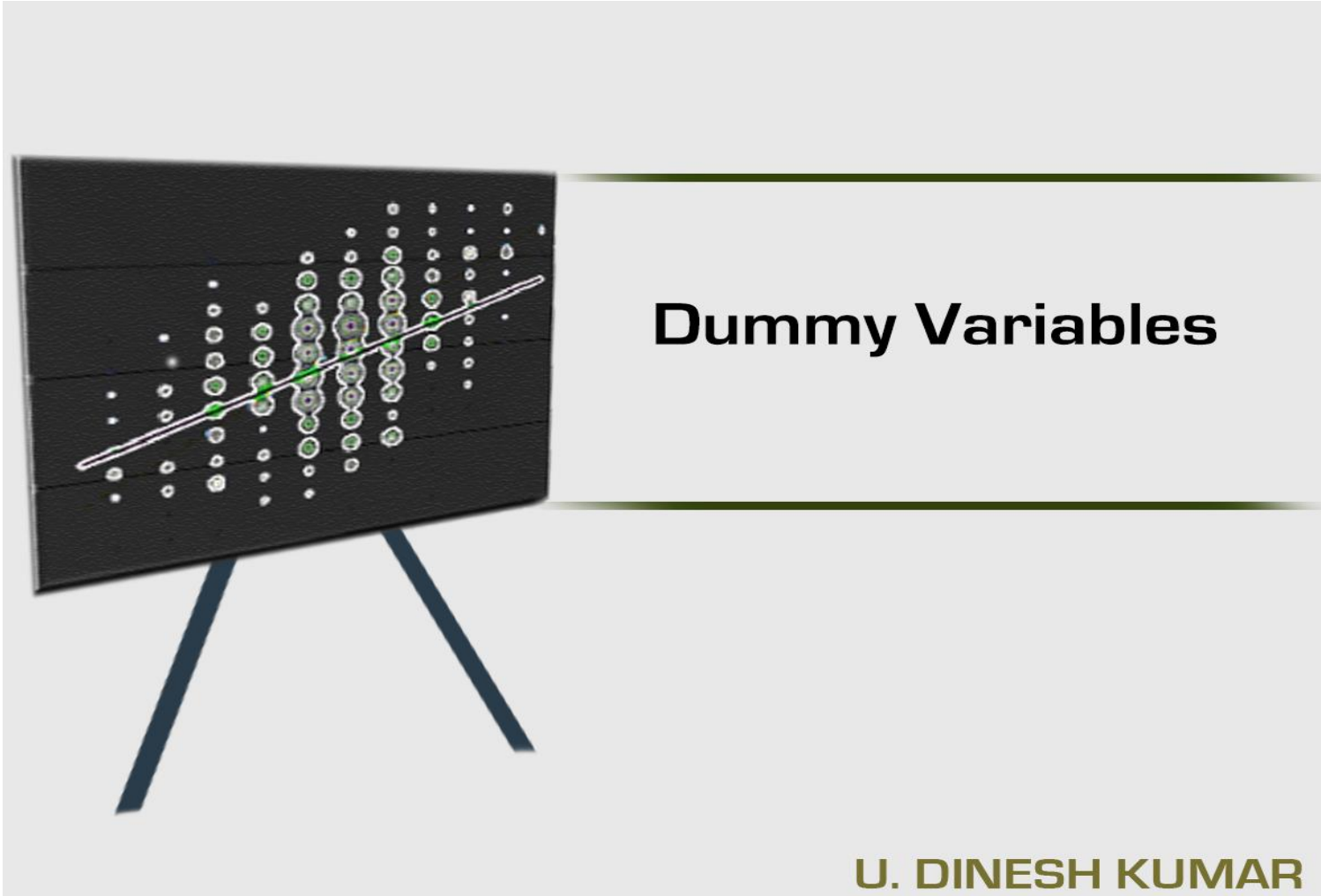
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	6.433E+11	2	3.216E+11	25.681	.000 ^b
	Residual	3.069E+12	245	12524679898		
	Total	3.712E+12	247			

a. Dependent Variable: TOTALCOSTTOHOSPITAL

b. Predictors: (Constant), HRPULSE, BODYWEIGHT

$$F = \frac{(SSE_{reduced} - SSE_{full}) / (4 - 2)}{MSE_{full}} = \frac{(3.069 \times 10^{12} - 3.046 \times 10^{12}) / 2}{1.253 \times 10^{10}} = 0.9172$$

$$p - value = 0.4700$$



Categorical Variables in Regression

- Many regression models are likely to have qualitative (categorical) variables as explanatory variable.
- Qualitative variables (categorical variables) in regression are replaced with dummy variables (or indicator variables) in regression model.
- A categorical variable with n levels are replaced with $(n-1)$ dummy variables. The category for which no dummy variable assigned is known as “**Base Category**”.

Dummy variable

- When there are more than one qualitative variable, it is advisable to use $(n-1)$ dummy variables for both qualitative variables along with the intercept.
- Use of n dummy variables along with intercept will result in multicollinearity, known as **dummy variable trap**.

Dummy variables in Regression

- The intercept, β_0 , is the mean value of the base category.
- The coefficients attached to dummy variables are called **differential intercept coefficients**.

Qualitative Variables – DAD Case

1. Gender (2 levels)
2. Marital Status (2 levels)
3. Key Complaints (6 levels)

Model with Dummy Variable

$$Y = \beta_0 + \beta_1 \times \text{Female}$$

Coefficients^a

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
1 (Constant)	211868.724	9421.879		22.487	.000
FEMALE	-39756.800	16385.377	-.153	-2.426	.016

a. Dependent Variable: TOTALCOSTTOHOSPITAL

Y for Male = 211868.724

Y for Female = 211868.724 – 39756.800 = 172111.924

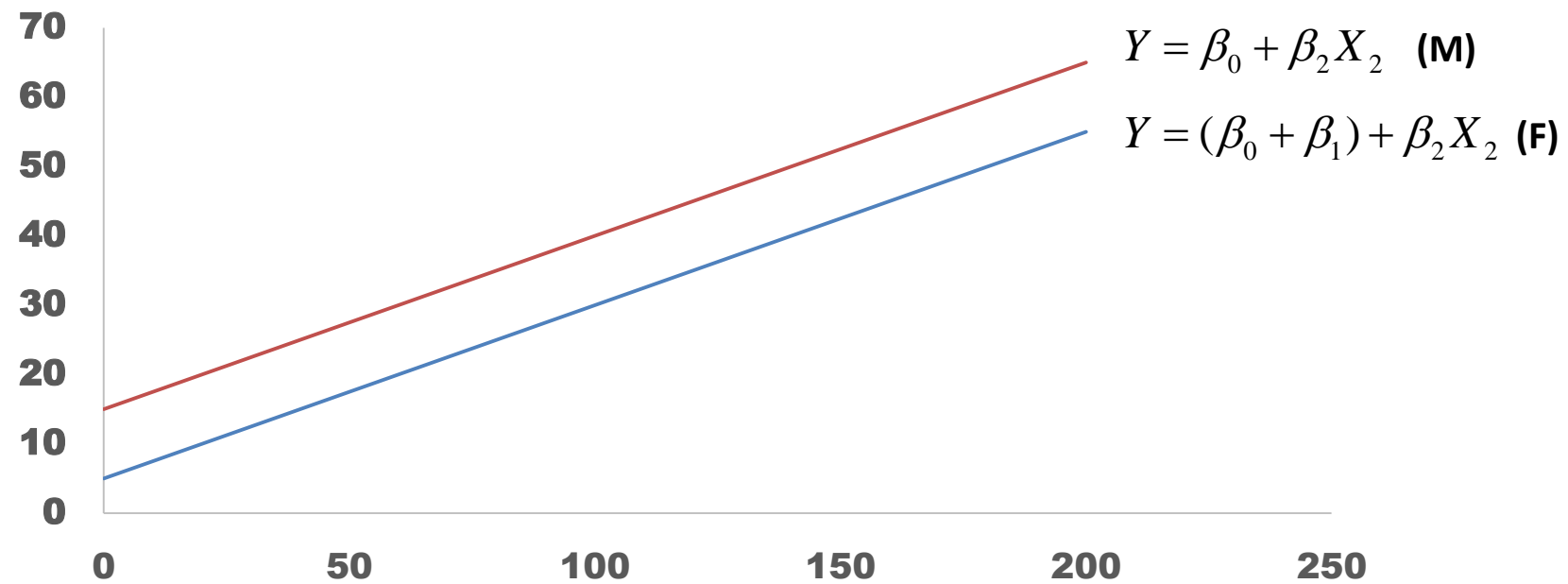
$$\underline{Y = \beta_0 + \beta_1 \times \text{Female} + \beta_2 \times \text{Body Weight}}$$

Coefficients^a

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
1 (Constant)	140891.968	15719.507		8.963	.000
FEMALE	-24698.709	15737.920	-.095	-1.569	.118
BODYWEIGHT	1758.786	320.899	.332	5.481	.000

a. Dependent Variable: TOTALCOSTTOHOSPITAL

Impact of Dummy Variable



Qualitative Variable with Multiple Levels

Predictive Analytics : QM901.1x
Prof U Dinesh Kumar, IIMB

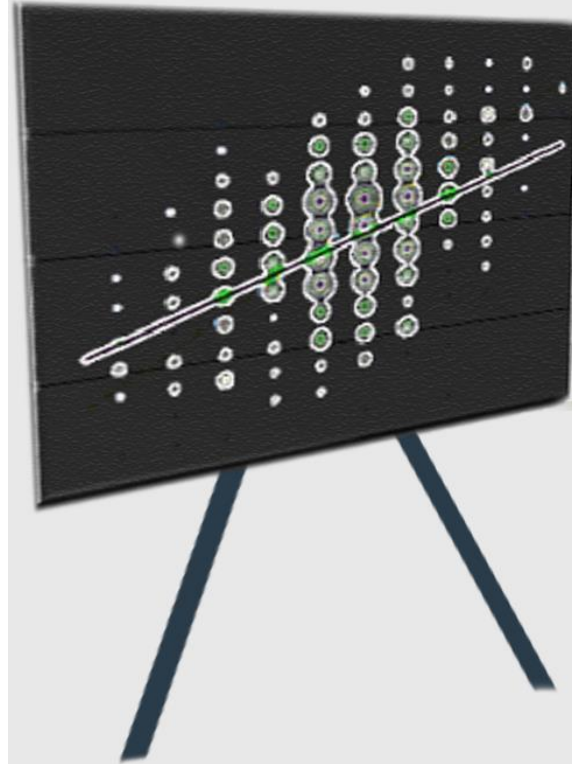
Key Complaints

- | | | |
|-----------|----------|---------|
| 1. ACHD | 2. CAD | 3. NONE |
| 4. OS-ASD | 5. OTHER | 6. RHD |

Coefficients^a

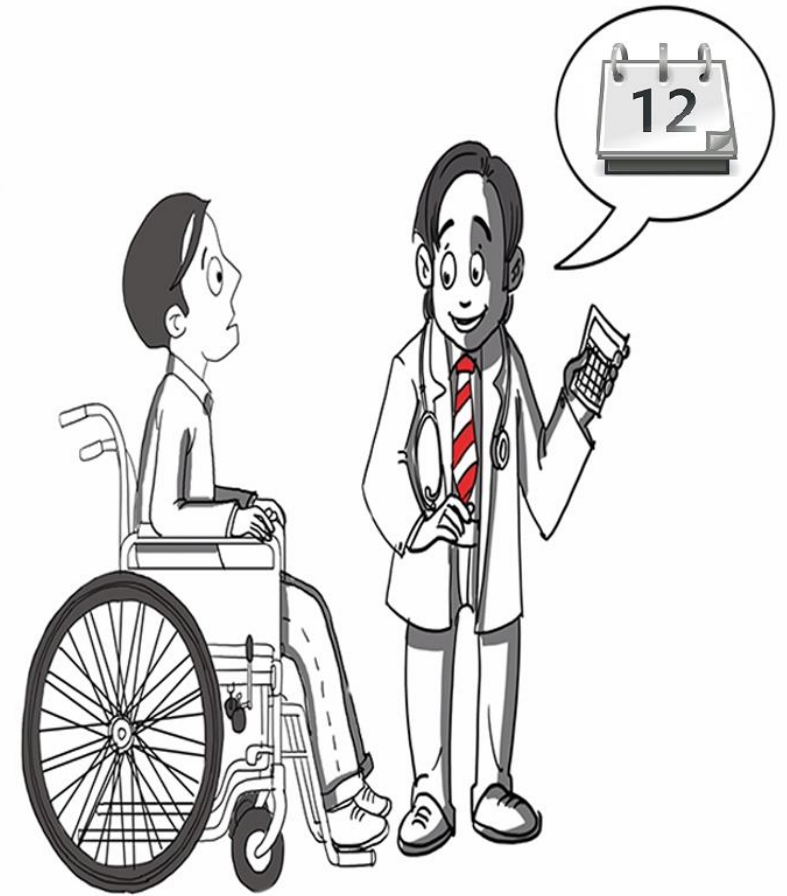
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	144998.351	18119.194		8.002	.000
	ACHD	-16491.016	30827.863	-.036	-.535	.593
	OSASD	-3857.177	33410.143	-.008	-.115	.908
	CAD	146464.013	23391.779	.494	6.261	.000
	RHD	110412.812	27980.022	.276	3.946	.000
	OTHERS	29746.856	21187.428	.119	1.404	.162

a. Dependent Variable: TOTALCOSTTOHOSPITAL



Interaction Variables

U. DINESH KUMAR



Regression Model Building

- Derived variables may help to explain the variation in the response variable.
- For example, consider a credit rating problem, which can be used for approving housing loan. A bank may use factors such as loan amount and value of the property etc.

Customer	1	2	3
Loan Amount	250,000	350,000	750,000

Derived Variables

Customer	1	2	3
Loan Amount	250,000	350,000	750,000
Value of Property	300,000	500,000	1,000,000
Loan / Value	0.83	0.70	0.75

Interaction Variables

Consider a regression model of type:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$$



**Interaction
variable**

The interaction variable is usually an interaction between a quantitative variable and qualitative variable

One of the popular applications of regression is to predict gender discrimination.

Consider a regression model with salary as response variable Y:

$$Y = \beta_0 + \beta_1 \text{ Gender} + \beta_2 \text{ Work Experience} + \beta_3 \text{ Gender x Work Experience}$$

Interaction Variable

Let Gender = 1 implies Female:

Then Y for Female is:

$$Y = \beta_0 + \beta_1 + (\beta_2 + \beta_3) \text{ Work Experience}$$

Y for Male is:

$$Y = \beta_0 + \beta_2 \times \text{Work Experience}$$

Conditional relationship in Interaction Variables

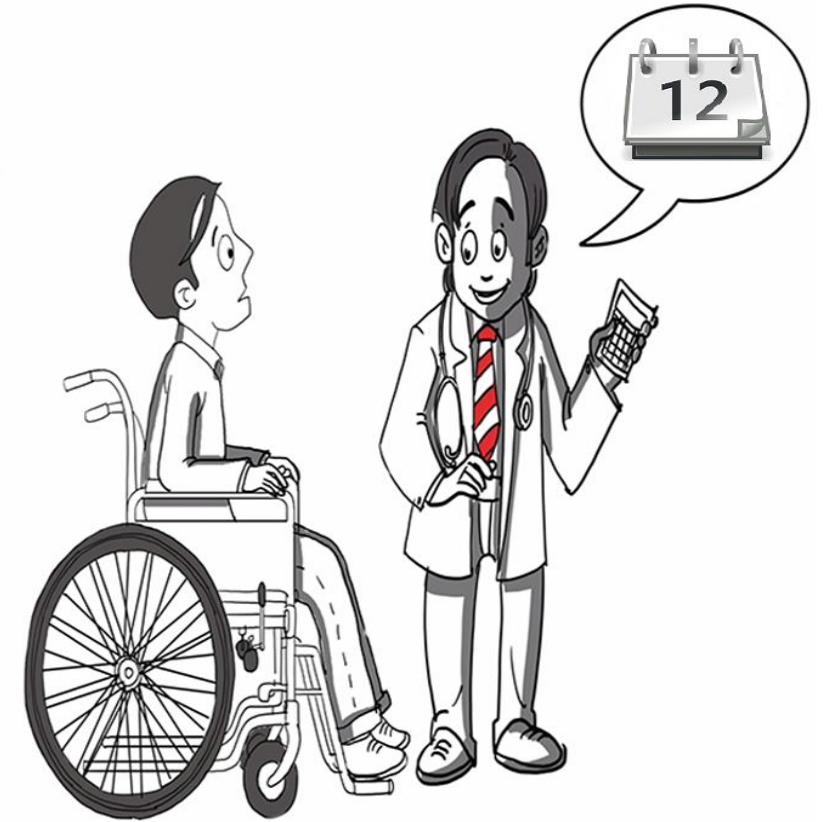
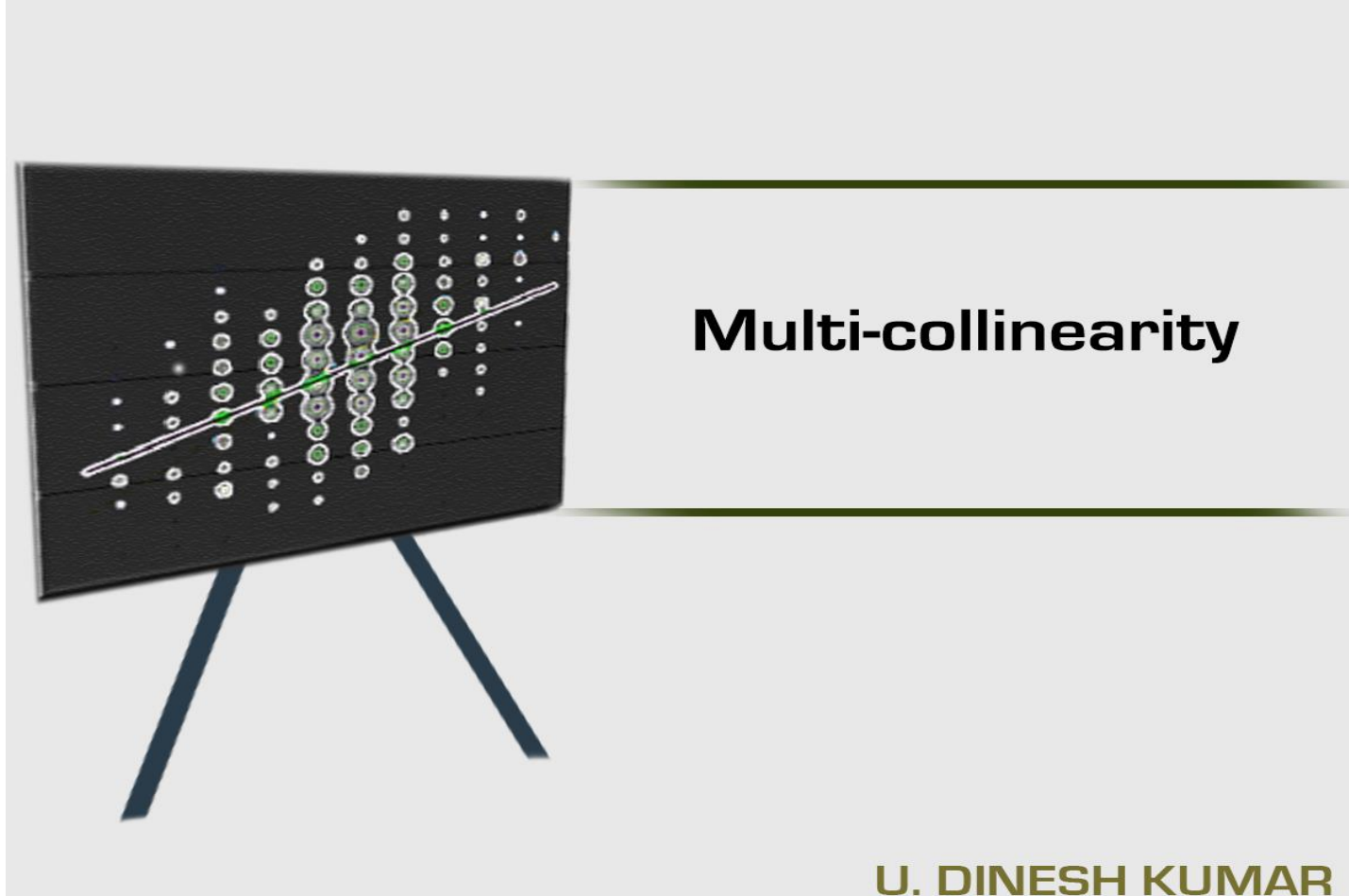
DAD Example – Interaction Variable

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	140567.363	17017.617		8.260	.000
	MARRIED	105081.981	63666.723	.426	1.651	.100
	BODYWEIGHT	863.168	678.558	.163	1.272	.205
	MarriedWeight	-773.453	1222.429	-.191	-.633	.528

a. Dependent Variable: TOTALCOSTTOHOSPITAL

$$Y = \beta_0 + \beta_1 \text{ Marital Status} + \beta_2 \text{ Body Weight} + \beta_3 \text{ Marital status} \times \text{Bodyweight}$$



Multicollinearity

- High correlation between explanatory variables is called multi-collinearity.
- Multi-collinearity leads to unstable coefficients.
- Always exists; matter of degree.

Multi-collinearity

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	140567.363	17017.617		8.260	.000
	MARRIED	105081.981	63666.723	.426	1.651	.100
	BODYWEIGHT	863.168	678.558	.163	1.272	.205
	MarriedWeight	-773.453	1222.429	-.191	-.633	.528

a. Dependent Variable: TOTALCOSTTOHOSPITAL

$$Y = \beta_0 + \beta_1 \text{ Marital Status} + \beta_2 \text{ Body Weight} + \beta_3 \text{ Marital status} \times \text{Bodyweight}$$

Symptoms of Multi-collinearity

- High R^2 but few significant t ratios.
- F-test rejects the null hypothesis, but none of the individual t-tests are rejected.
- Correlations between pairs of X variables are more than with Y variables.

Effects of Multi-collinearity

- The variances of regression coefficient estimators are inflated.
- The magnitudes of regression coefficient estimates may be different.
- Adding and removing variables produce large changes in the coefficient estimates.
- Regression coefficient may have opposite sign.

Identifying Multi-collinearity Variance Inflation factor

- The variance inflation factor (VIF) is a relative measure of the increase in the variance in standard error of beta coefficient because of collinearity.
- A VIF greater than 10 indicates that collinearity is very high. A VIF value of more than 4 is not acceptable.

Variance of Beta Estimates

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

$$Var(\hat{\beta}_1) = \frac{S_e^2}{\sum SS_{X_2} \times (1 - r_{12}^2)}$$

$$Var(\hat{\beta}_2) = \frac{S_e^2}{\sum SS_{X_1} \times (1 - r_{12}^2)}$$

Variance inflation factor

- Variance inflation factor associated with introducing a new variable X_j is given by:

$$VIF(X_j) = \frac{1}{1 - R_j^2}$$

R_j^2 is the coefficient of determination for the regression of X_j as dependent variable

The standard error of the corresponding Beta is inflated by \sqrt{VIF}

R_j	Tolerance ($1-R^2$)	VIF	Impact on $Se(\beta)$ \sqrt{VIF}
0	1.0	1.0	1.0
0.4	0.84	1.19	1.09
0.6	0.64	1.56	1.25
0.8	0.36	2.78	1.67
0.87	0.25	4.0	2.0
0.9	0.19	5.26	2.29

DAD CASELET

Predictive Analytics : QM901.1x
Prof U Dinesh Kumar, IIMB

Coefficients^a

		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	140567.363	17017.617		8.260	.000		
	BODYWEIGHT	863.168	678.558	.163	1.272	.205	.214	4.681
	MarriedWeight	-773.453	1222.429	-.191	-.633	.528	.038	26.172
	MARRIED	105081.981	63666.723	.426	1.651	.100	.053	19.032

a. Dependent Variable: TOTALCOSTTOHOSPITAL

For Body Weight Actual $t = 1.272 \times \sqrt{4.681} = 2.752$

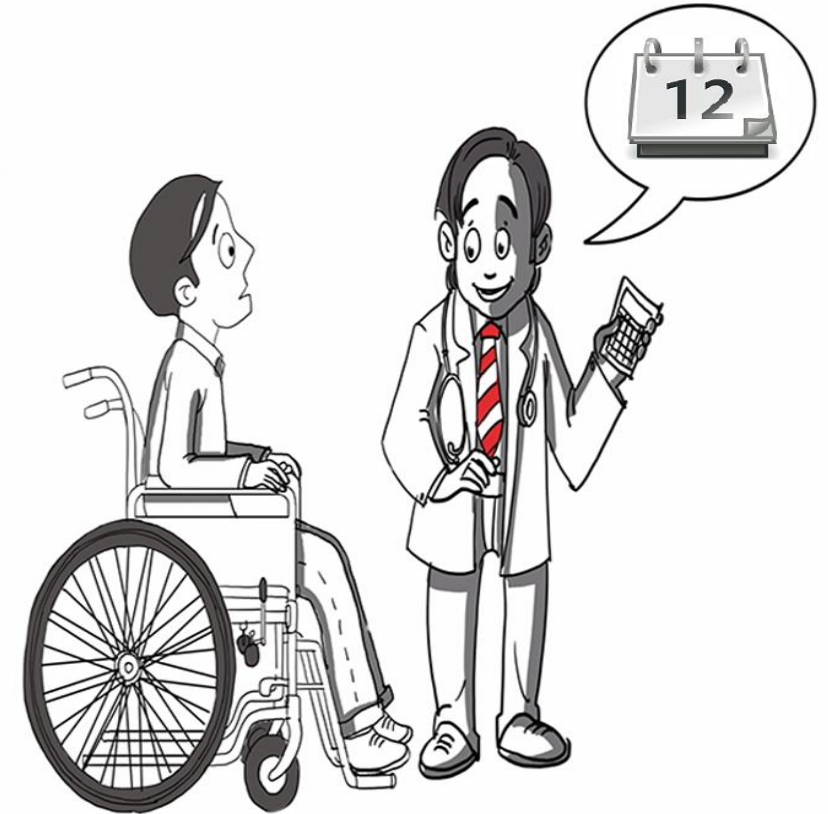
Corresponding p - value = 0.006

Remedies for Multi-collinearity

- Drop a variable (may result in specification bias).
- Principal component analysis (PCA)
- Partial Least Squares (PLS)

Regression Model Building

U. DINESH KUMAR



Forward Selection Method

- In forward selection procedure in which variables are sequentially entered into the model.
- The first variable considered for entry is the one with smallest p-value based on F-test.

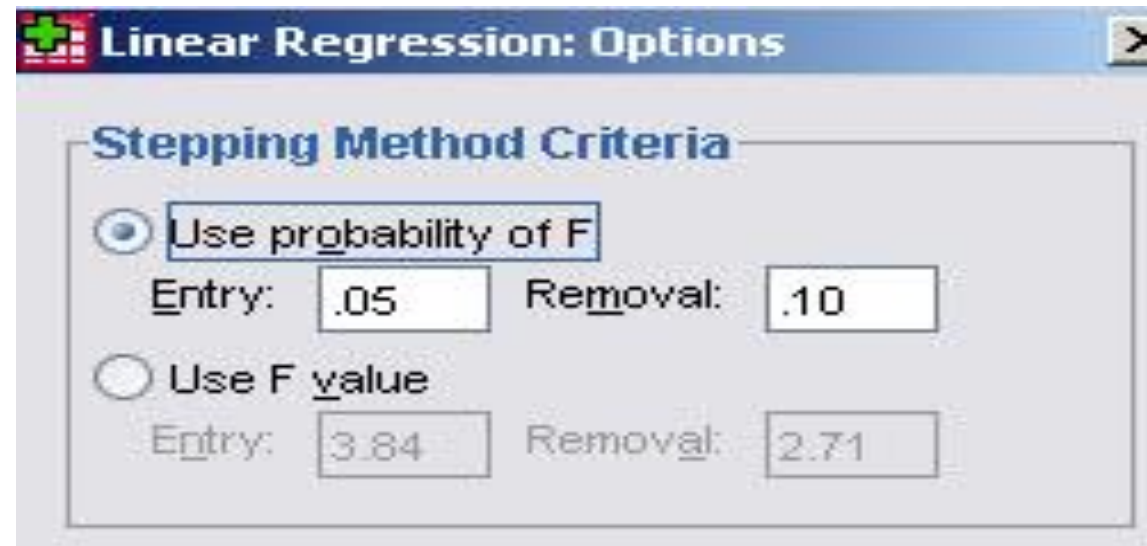


Backward elimination method

- A variable selection procedure in which all variables are entered into the equation and then sequentially removed starting with the most non-significant variable.
- At each step, the largest probability of F is removed.

Stepwise Regression

- ▶ The first variable considered for entry into the equation is the one with the largest F value or smallest p-value based on F-test.
- ▶ At each step, the independent variable not in the equation that has the smallest probability of F is entered. Variables already in the regression equation are removed if their probability of F becomes sufficiently large.



ANOVA^a

Predictive Analytics : QM901.1x
Prof U Dinesh Kumar, IIMB

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	5.937E+11	1	5.937E+11	46.839	.000 ^b
	Residual	3.118E+12	246	12675357019		
	Total	3.712E+12	247			
2	Regression	7.981E+11	2	3.990E+11	33.552	.000 ^c
	Residual	2.914E+12	245	11892977137		
	Total	3.712E+12	247			
3	Regression	8.782E+11	3	2.927E+11	25.206	.000 ^d
	Residual	2.834E+12	244	11613410239		
	Total	3.712E+12	247			
4	Regression	1.038E+12	4	2.594E+11	23.576	.000 ^e
	Residual	2.674E+12	243	11004485969		
	Total	3.712E+12	247			
5	Regression	1.082E+12	5	2.164E+11	19.914	.000 ^f
	Residual	2.630E+12	242	10866938886		
	Total	3.712E+12	247			

a. Dependent Variable: TOTALCOSTTOHOSPITAL

b. Predictors: (Constant), CAD

c. Predictors: (Constant), CAD, RHD

d. Predictors: (Constant), CAD, RHD, HRPULSE

e. Predictors: (Constant), CAD, RHD, HRPULSE, MARRIED

f. Predictors: (Constant), CAD, RHD, HRPULSE, MARRIED, OTHERS

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	172909.372	8083.124		21.391	.000
	CAD	118552.991	17322.405	.400	6.844	.000
2	(Constant)	160141.238	8413.770		19.033	.000
	CAD	131321.126	17059.656	.443	7.698	.000
	RHD	95269.926	22982.901	.239	4.145	.000
3	(Constant)	69764.074	35401.596		1.971	.050
	CAD	145960.413	17755.552	.492	8.221	.000
	RHD	103214.926	22911.748	.258	4.505	.000
	HRPULSE	936.320	356.507	.155	2.626	.009
4	(Constant)	1192.407	38881.960		.031	.976
	CAD	107402.104	20031.295	.362	5.362	.000
	RHD	85665.117	22774.163	.215	3.762	.000
	HRPULSE	1471.703	374.432	.244	3.930	.000
	MARRIED	67577.738	17745.840	.274	3.808	.000
5	(Constant)	-7164.969	38859.328		-.184	.854
	CAD	122458.276	21256.912	.413	5.761	.000
	RHD	103103.698	24223.806	.258	4.256	.000
	HRPULSE	1342.539	377.545	.222	3.556	.000
	MARRIED	71547.412	17743.873	.290	4.032	.000
	OTHERS	33998.332	16840.465	.136	2.019	.045

a. Dependent Variable: TOTALCOSTTOHOSPITAL

$$Y = - 7164.969 + 122458.276 \text{ CAD} + 103103.698 \text{ RHD} \\ + 1342.539 \text{ HR PULSE} + 71547.412 \text{ MARRIED} \\ + 33998.332 \text{ OTHERS}$$