Session 15: Linear Models in R

Agenda

Sl. No.	Topics For The Agenda
1.	Introduction To Regression
2.	Why Do A Regression Analysis?
3.	Regression Analysis
4.	Where Is It Used?
5.	Types Of Regression
6.	Some More Types Of Regression
7.	Determine The Best Fit
8.	OLS Regression – Least Squares
9.	Dependent And Independent Variable(s)
10.	Ordinary Least Squares Method

Sl. No.	Topics For The Agenda
11.	Least Squares Method
12.	Example
13.	So What Are a and b?
11.	OLS Estimators
12.	Steps To Implement A Regression Model
13.	Fitting Regression Models with lm()
14.	Single Linear Regression – Example
15.	Coaching Data Set
16.	Single Linear Regression
17.	Single Linear Regression – Output
18.	Simple Regression Analysis
19.	Exercise

Introduction To Regression

- Regression is a tool for finding a relationship between a dependent variable and one or more independent variables in a study.
- The relationship can be linear or non-linear.
- The basic function of regression is to identify statistically significant independent variables and estimate the model parameters.

Why Do A Regression Analysis?

- How MRP affects the purchase decision?
- Which promotion is more effective?
- What is the risk associated with price increase on customer retention?
- Which customer is likely to default?
- What percentage of loans is likely to result in a loss?
- How to identify the most profitable customer?

Regression Analysis

Example

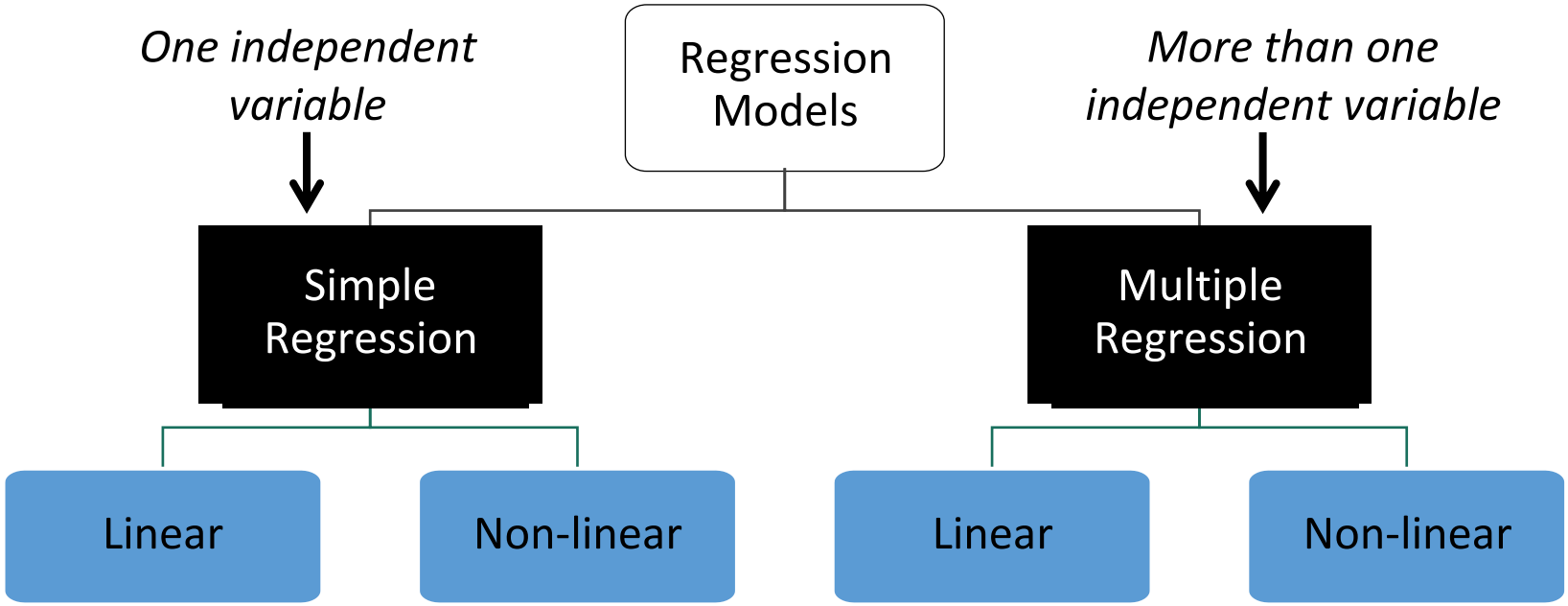
- What's the relationship between exercise duration and calories burned? Is it linear or curvilinear?
- Does exercise have less impact on the number of calories burned after a certain point?
- How does effort (the percentage of time at the target heart rate, the average walking speed) factor in?
- Are these relationships the same for young and old, male and female, heavy and slim?

Where Is It Used?

Every functional area of management uses regression:

- Finance: Chance of bankruptcy, credit risk fraud.
- Marketing: Sales, market share, customer satisfaction, customer churn, customer retention, customer life time value.
- Operations: Inventory, productivity, efficiency.
- HR: Job satisfaction, attrition.
- Healthcare : New plans, health insurance

Types Of Regression



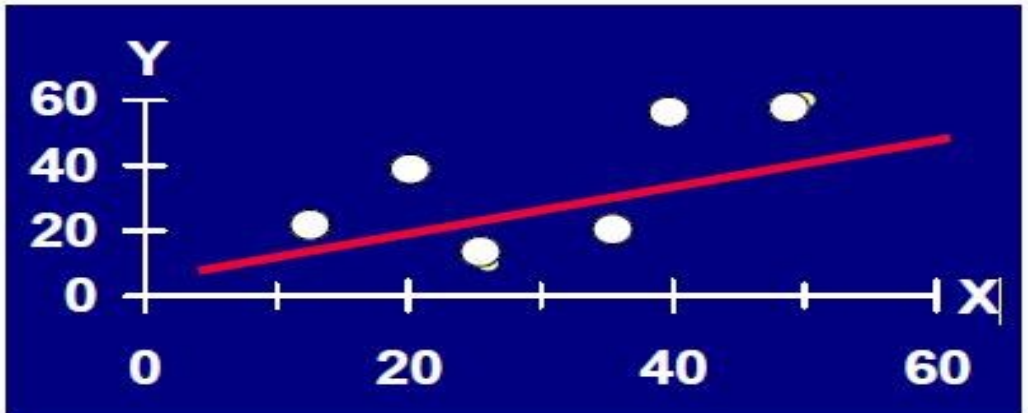
Types Of Regression (Contd.)

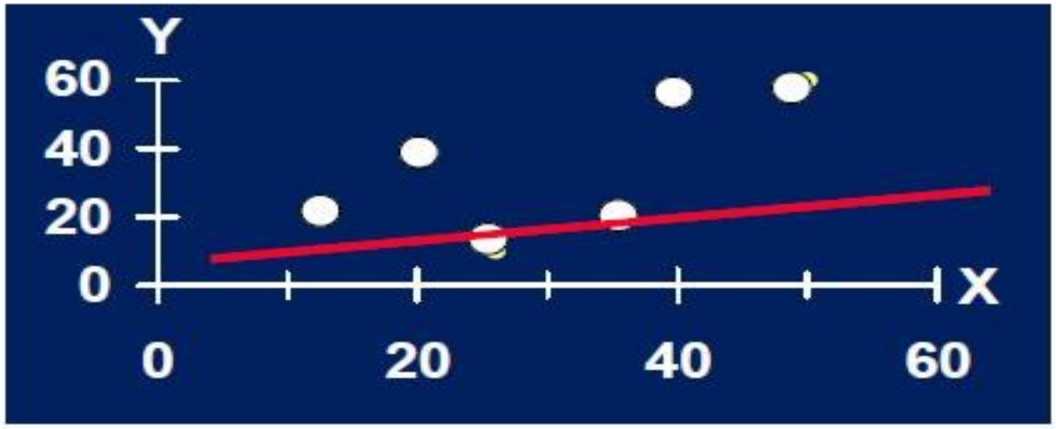
Type of regression	Typical use
Simple linear	Predicting a quantitative response variable from a quantitative explanatory variable
Polynomial	Predicting a quantitative response variable from a quantitative explanatory variable, where the relationship is modeled as an nth order polynomial
Multiple linear	Predicting a quantitative response variable from two or more explanatory variables
Multivariate	Predicting more than one response variable from one or more explanatory variables
Logistic	Predicting a categorical response variable from one or more explanatory variables
Poisson	Predicting a response variable representing counts from one or more explanatory variables

Some More Types Of Regression

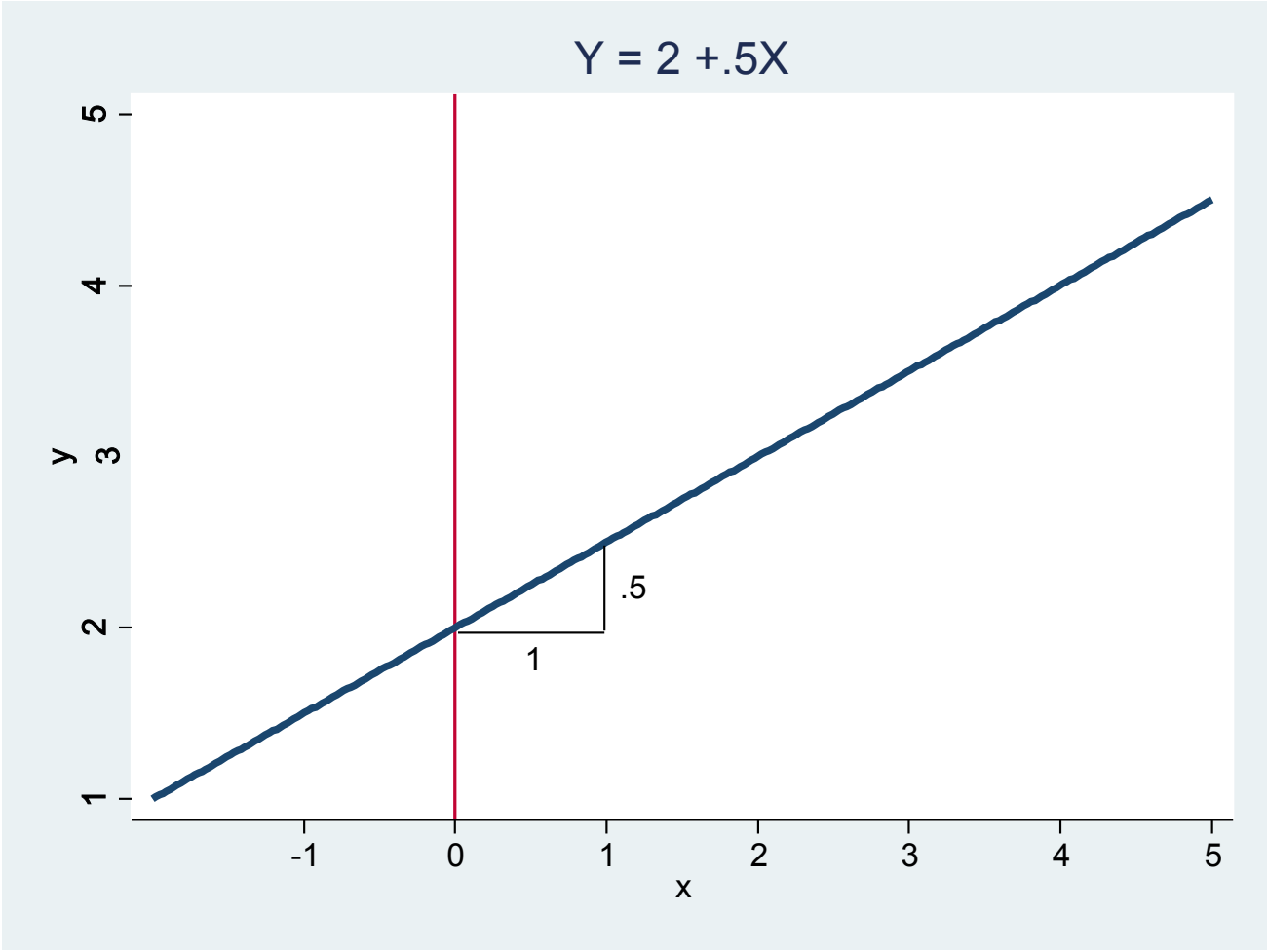
Type of regression	Typical use
Cox proportional hazards	Predicting time to an event (death, failure, relapse) from one or more explanatory variables
Time-series	Modeling time-series data with correlated errors
Nonlinear	Predicting a quantitative response variable from one or more explanatory variables, where the form of the model is nonlinear
Nonparametric	Predicting a quantitative response variable from one or more explanatory variables, where the form of the model is derived from the data and not specified a priority
Robust	Predicting a quantitative response variable from one or more explanatory variables using an approach that's resistant to the effect of influential observations

Determine The Best Fit





OLS Regression – Least Squares



Dependent And Independent Variable(s)

- Dependent variable (DV) = response variable = left-hand side (LHS) variable
- Independent variables (IV) = explanatory variables = right-hand side (RHS) variables = regressor (excluding a or b0)
- a (b0) is an estimator of parameter α, $\beta 0$
- b (b1) is an estimator of parameter β, $\beta 1$
- a and b are the intercept and slope

Ordinary Least Squares Method

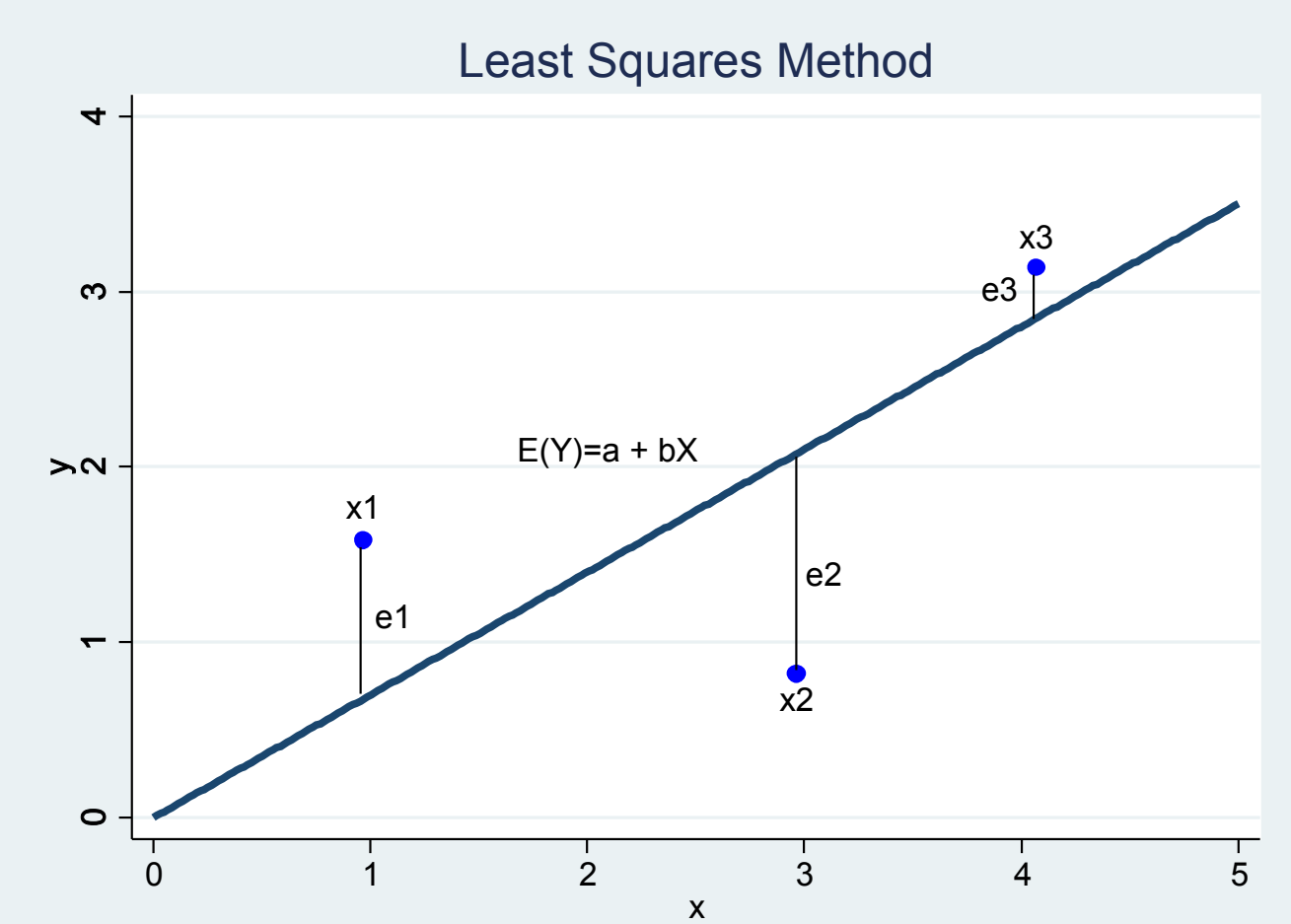
How to draw such a line based on data points observed?

- Suppose a imaginary line of y= a + bx
- Imagine a vertical distance (or error) between the line and a data point. E=Y-E(Y)
- This error (or gap) is the deviation of the data point from the imaginary line, regression line

What is the best values of a and b?

- A and b that minimizes the sum of such errors (deviations of individual data points from the line)

Ordinary Least Squares Method (Contd.)



Ordinary Least Squares Method (Contd.)

- Deviation does not have good properties for computation
- Why do we use squares of deviation? (e.g., variance)
- Let us get a and b that can minimize the sum of squared deviations rather than the sum of deviations.
- This method is called least squares

Ordinary Least Squares Method (Contd.)

- Least squares method minimizes the sum of squares of errors (deviations of individual data points form the regression line)
- Such a and b are called least squares estimators (estimators of parameters α and β).
- The process of getting parameter estimators (e.g., a and b) is called estimation
- "Regress Y on X"
- Lest squares method is the estimation method of ordinary least squares (OLS)

Ordinary Least Squares Method (Contd.)

- Ordinary least squares (OLS)
- Linear regression model
- Classical linear regression model
  - Linear relationship between Y and Xs
  - Constant slopes (coefficients of Xs)
  - Least squares method
  - Xs are fixed; Y is conditional on Xs
  - Error is not related to Xs
  - Constant variance of errors

Least Squares Method

$$Y = \alpha + \beta X + \varepsilon$$

$$E(Y) = \hat{Y} = a + bX$$

$$\varepsilon = Y - \hat{Y} = Y - (a + bX) = Y - a - bX$$

$$\varepsilon^2 = (Y - \hat{Y})^2 = (Y - a - bX)^2$$

$$(Y - a - bX)^2 = Y^2 + a^2 + b^2 X^2 - 2aY - 2bXY + 2abX$$

$$\sum \varepsilon^2 = \sum (Y - \hat{Y})^2 = \sum (Y - a - bX)^2$$

$$Min \sum \varepsilon^2 = Min \sum (Y - a - bX)^2$$

How to get *a* and *b* that can minimize the sum of squares of errors?

Example

No	x	y	x-xbar	y-ybar	(x-xb)(y-yb)	(x-xb
1	43	128	-14.5	-8.5	123.25	2
2	48	120	-9.5	-16.5	156.75	
3	56	135	-1.5	-1.5	2.25	
4	61	143	3.5	6.5	22.75	
5	67	141	9.5	4.5	42.75	
6	70	152	12.5	15.5	193.75	1
Mean	57.5	136.5				
Sum	345	819			541.5	

$$b = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sum (X - \bar{X})^2} = \frac{SP_{xy}}{SS_x} = \frac{541.5}{561.5} = .9644$$

$$a = \bar{Y} - b\bar{X} = 136.5 - .9644 \times 57.5 = 81.0481$$

So What Are a and b?

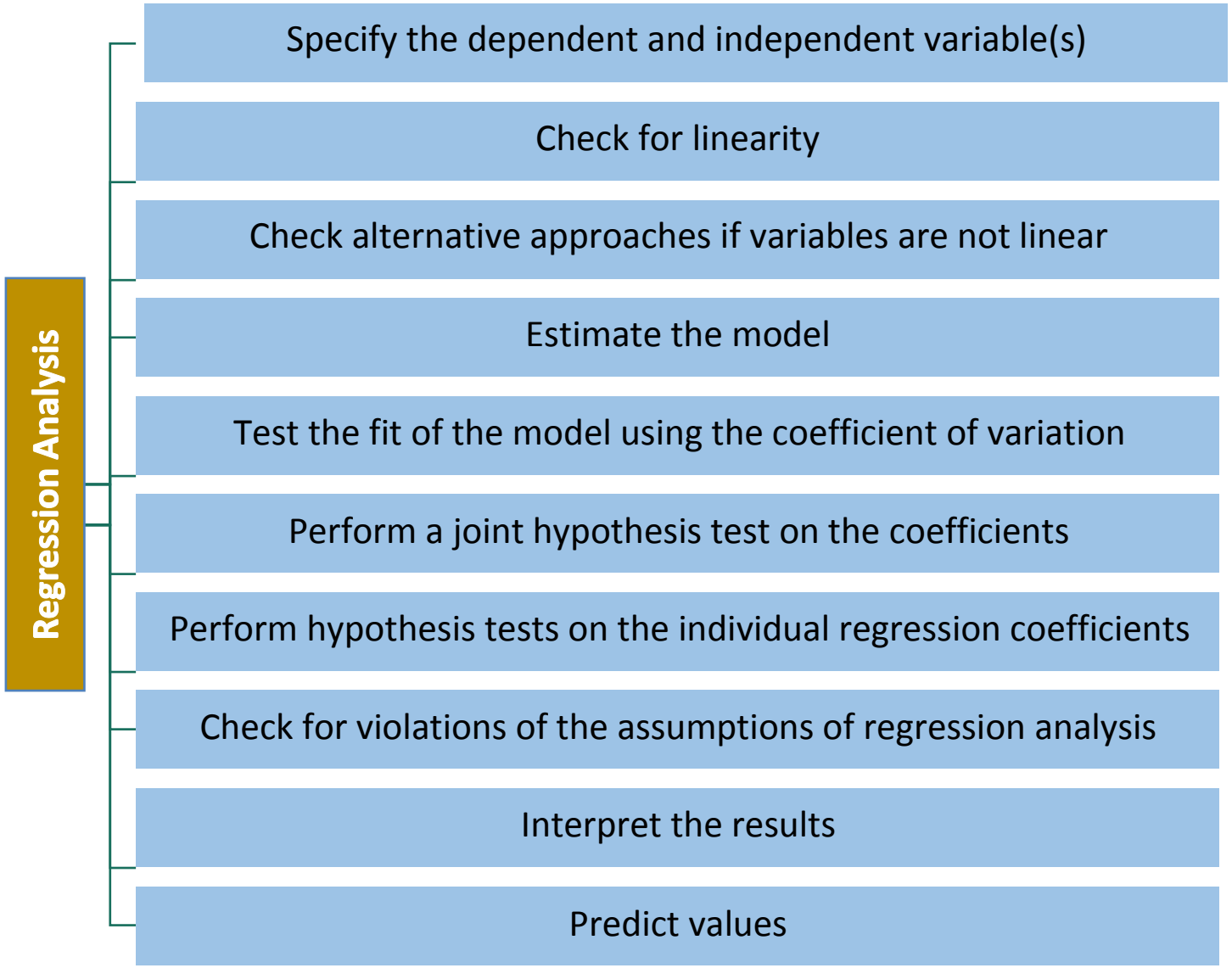
- a is an estimator of its parameter α
- a is the intercept, a point of y where the regression line meets the y axis
- b is an estimator of its parameter β
- b is the slope of the regression line
- b is constant regardless of values of Xs
- For unit increase in x, the expected change in y is b, holding other things (variables) constant.
- For unit increase in x, we expect that y increases by b, holding other things (variables) constant.

OLS Estimators

The outcome of least squares method is OLS parameter estimators *a* and *b*.

- OLS estimators are linear
- OLS estimators are unbiased (precise)
- OLS estimators are efficient (small variance)

Steps To Implement A Regression Model



Fitting Regression Models with lm() (Contd.)

Each of these functions listed below is applied to the object returned by lm() in order to generate additional information based on that fitted model.

Function	Actions
summary()	Displays detailed results for the fitted model
coefficients()	Lists the model parameters (intercept and slopes) for the fitted model
confint()	Provides confidence intervals for the model parameters (95 percent by default)
fitted()	Lists the predicted values in a fitted model
residuals()	Lists the residual values in a fitted model
anova()	Generates an ANOVA table for a fitted model, or an ANOVA table comparing two or more fitted models
vcov()	Lists the covariance matrix for model parameters
AIC()	Prints Akaike's Information Criterion
plot()	Generates diagnostic plots for evaluating the fit of a model
predict()	Uses a fitted model to predict response values for a new dataset

Simple Regression Analysis

Single Linear Regression - Example

Simple linear regression $Y = \beta_o + \beta_1 X$

Description

A coaching institute wants to start new stream of coaching. It would like to estimate the number of applications it can expect for the next year admission. Below are the data points given by the coaching center

- Year – Year
- App_num– number of applications when a new course was introduced
  Avg_rate_Pmt – Average Placement Rate
- Graduation_num – Number of under grad final year students who would graduate in the current year

Coaching Data Set

App_num	Avg_rate_Pmt	Graduation_num
5945	61	13742
6500	50	14744
5888	53	13588
4000	55	13000
4700	50	12500
6300	44	12800
6200	45	13100
7000	44	13850
5000	43	13900
3000	57	12000
1000	62	11000
4000	55	11531
4600	54	12788
3000	62	13000
1000	79	13500

- Given the record for last 15 years, predict the no of applications using the Placement Rate.
- What is the expected number of applications (App_num) given this year's placement rate (Avg_rate_Pmt) of 70%.

Single Linear Regression

Linear regression with a single predictor

- Reading the csv into a R dataframe

  s <- read.csv("student.csv",header=TRUE)

  attach(s)

- Fit the regression model using the function lm()

  *result <- lm(app_num~*avg_rate_pmt*, data=s)*

- Use the function summary () to get some results

  summary(result)

Single Linear Regression - Output

```
Call:
lm(formula = app_num ~ avg_rate_pmt, data = s)

Residuals:
   Min     1Q  Median     3Q    Max
-359.60 -142.76   23.68  132.09  361.75

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3476.647    307.955   11.29 4.33e-08 ***
avg_rate_pmt  -16.674      5.539   -3.01     0.01 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 209.5 on 13 degrees of freedom
Multiple R-squared:  0.4108,            Adjusted R-squared:  0.3654
F-statistic: 9.062 on 1 and 13 DF,  p-value: 0.01004
```

Simple Regression Analysis

Output Interpretation

- From this output, we have determined that the intercept is 3476.647 and the coefficient for the placement rate is -16.674

- Therefore, the complete regression equation is

apps = 3476.647 + (- 16.674* avg_rate_pmt)

- This equation tells us that the predicted number of applications for the coaching will decrease by 16.674 students for every one percent increase in the placement rate.
- What is the expected number of applications (APPLICANTS), given this year's placement rate (avg_rate_pmt) of 60% ?

apps= 3476.647 + (- 16.674* .6) = 13006.4

Simple Regression Analysis (contd.)

Intercept = Also known as the y intercept, it is simply the value at which the fitted line crosses the y-axis. The intercept is the expected mean value of Y when all **X=0**.

Std. Error = represents the average distance that the observed values fall from the regression line. Conveniently, it tells you how wrong the regression model is on average using the units of the response variable. Smaller values are better because it indicates that the observations are closer to the fitted line.

t-statistic = is just *the estimated coefficient divided by its own standard error*. Thus, it measures "how many standard deviations from zero" the estimated coefficient is, and it is used to test the hypothesis that the true value of the coefficient is non-zero, in order to confirm that the independent variable really belongs in the model.

p-value = is the probability of observing a t-statistic that large or larger in magnitude given the null hypothesis that the true coefficient value is zero. If the p-value is greater than 0.05--which occurs roughly when the t-statistic is less than 2 in absolute value--this means that the coefficient may be only "accidentally" significant.

Place_Rate (in this case) = Since X1 is a continuous variable, B1 represents the difference in the predicted value of Y for each one-unit difference in X1. This means that if X1 differed by one unit, Y will differ by B1 units, on average.

Residual standard error is an estimate of the parameter σ . The assumption in ordinary least squares is that the residuals are individually described by a Gaussian (normal) distribution with mean 0 and standard deviation σ . The σ relates to the constant variance assumption; each residual has the same variance and that variance is equal to σ 2.

F statistic tests whether the predictor variables taken together, predict the response variable above chance levels, because there's only one predictor variable in simple regression.

Coefficient of variation (also known as R2) is used to determine how closely a regression model "fits" or explains the relationship between the independent variable (X) and the dependent variable (Y). R2 can assume a value between 0 and 1; the closer R2 is to 1, the better the regression model explains the observed data.

Adjusted R 2 is the same thing as R 2, but adjusted for the complexity of the model, i.e. the number of parameters in the model. If we have a model with a single parameter, it will have a certain R 2 . If we add another parameter to this model, the R 2 of the new model has to increase, even if the added parameter has no statistical power. The adjusted R 2 tries to account for this, by including information on the number of parameters in the model.

Simple Regression Analysis (contd.)

- Residuals and Fitted Values
- Fitted values obtained using the function fitted()
- Residuals obtained using the function resid()
- Create a table with fitted values and residuals
- s_fitted<-data.frame(s, fitted.value=fitted(result),residual=resid(result))

Simple Regression Analysis (contd.)

		app_num	avg_rate_pmt	graduation_num	Fitted value	residua
	1	2342	53	1334	2592.902	-250.9(
	2	2630	39	1276	2826.344	-196.34
	3	2988	51	1385	2626.251	361.74
	4	2400	59	1000	2492.856	-92.855
	5	2900	48	1160	2676.275	223.72
	6	2800	49	1820	2659.600	140.39
	7	2700	54	1410	2576.228	123.77
	8	2300	49	1685	2659.600	-359.6(
	9	2800	42	1630	2776.321	23.678
	10	2700	53	1590	2592.902	107.09
	11	2600	67	1402	2359.460	240.53
	12	2400	66	1251	2376.135	23.865
	13	2600	49	1258	2659.600	-59.60(
	14	2200	67	1375	2359.460	-159.46
	15	2100	75	1743	2226.065	-126.06

Simple Regression Analysis (contd.)

Confidence & Prediction

Predicted values are obtained using the function predict()

Obtaining the confidence bands:

predict(result, interval="confidence")

	fit	lwr	upr
1	2592.902	2474.201	2711.603
2	2826.344	2604.749	3047.940
3	2626.251	2501.129	2751.373
4	2492.856	2365.316	2620.396
5	2676.275	2534.318	2818.231
6	2659.600	2524.077	2795.123
7	2576.228	2459.024	2693.432
8	2659.600	2524.077	2795.123
9	2776.321	2584.289	2968.353
10	2592.902	2474.201	2711.603
11	2359.460	2171.829	2547.092
12	2376.135	2197.708	2554.561
13	2659.600	2524.077	2795.123
14	2359.460	2171.829	2547.092
15	2226.065	1956.856	2495.274

Simple Regression Analysis (Contd.)

- Summary consists of the below functions:
- names(result)

```
> names(result)
 [1] "coefficients"  "residuals"     "effects"       "rank"          "fitted.values"
 [6] "assign"        "qr"            "df.residual"   "xlevels"       "call"
[11] "terms"         "model"
>
```

```
> result$coefficients
 (Intercept) avg_rate_pmt
  3476.64722    -16.67443
> result$residuals
        1         2         3         4         5         6         7         8         9        10        11        12
-250.90235 -196.34439  361.74879  -92.85576  223.72549  140.39993  123.77208 -359.60007   23.67890  107.09765  240.53969   23.86526
       13        14        15
 -59.60007 -159.46031 -126.06485
> result$effects
 (Intercept) avg_rate_pmt
-9930.329300  -630.755054   422.433121   -53.952101   292.577579   206.529426   176.288662  -293.470574   108.866495   162.336815

  257.662677    43.710830     6.529426  -142.337323  -130.722545
> result$rank
[1] 2
> result$fitted.values
      1       2       3       4       5       6       7       8       9      10      11      12      13      14      15
2592.902 2826.344 2626.251 2492.856 2676.275 2659.600 2576.228 2659.600 2776.321 2592.902 2359.460 2376.135 2659.600 2359.460 2226.065
> result$assign
[1] 0 1
```

Exercise

Use the below data find the relationship between age (in months) and height (Average height in cm)

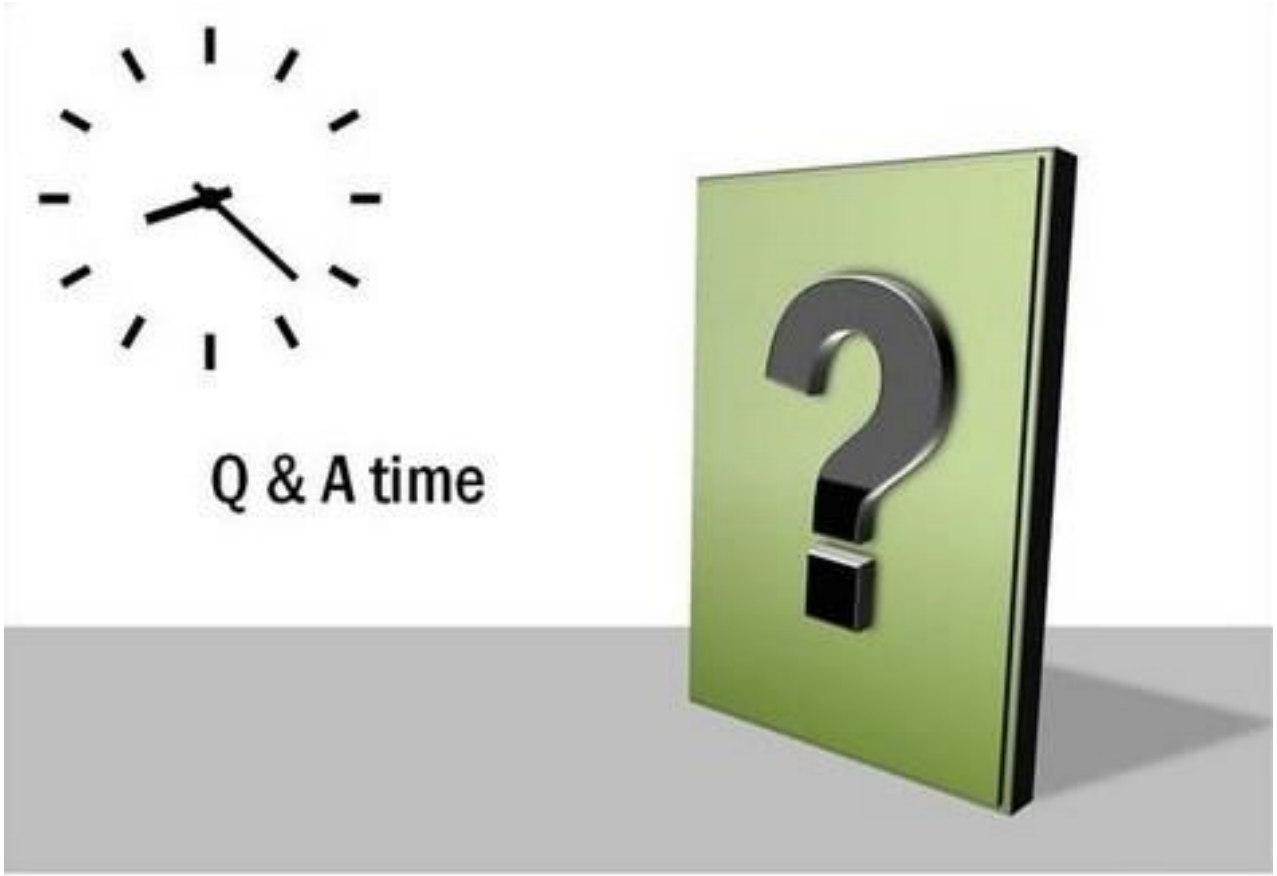
- age=16:27
- height=c(61.1,61.2,61.8,62.8,63.5,76.1,77,78.1,78.2,78.8,79.7,79.9)

1) check that age and height have the same number of elements
2) Create a scatterplot to determine the relationship between age and height
3) Create a "linear model" to fits the data above
4) Find the equation of the line of best fit

Next Class
Correlation & Multiple Regression

Sl. No.	Topics For The Agenda
1.	Correlation
2.	How Is Relationship Measured
3.	Scatterplot To Analyze Correlation
4.	Strength Of Linear Association
5.	Other Strengths
6.	Examples
7.	Limitations Of Correlation
8.	Causation
9.	Least-squares Or Regression Line
10.	Linear Regression Model
11.	Estimated Regression Line

Sl. No.	Topics For The Agenda
12.	Correlation Coefficient, R
13.	Coefficient Of Determination
14.	Difference Between Correlation And Regression
15.	Limitations Of The Correlation Coefficient
16.	Multiple Linear Regression
17.	Assumptions Of A Linear Model
18.	Regression Diagnostics
19.	Detection of Collinearity: Simple Signs
20.	Detecting Multicollinearity
21.	Exercise


Q & A time