BC2406 Business Analytics I: Predictive Techniques

Seminars 6
Linear Regression Model

Instructor: Prof. Lee Gun-woong

Nanyang Business School

Objectives

–Understand the linear regression model

–Understand how to setup a regression model

–Explain Ordinary Least Squares

–Interpret the model output

–Evaluate the model performance

Understanding Regression

Linear Regression Model

- Model: Representation of some phenomenon
  - Describe the relationship between variables

- Deterministic vs. Probabilistic Models
  - DM: Describe exact relationships (No errors)
    - BMI = $\frac{Weight\ in\ kg}{(Height\ in\ meter)^2}$
  - PM: Describe Deterministic Components + Error
    - Sales = f(Price, Popularity, …) + Error

- Regression analysis is a probabilistic model describing the relationship between two (or more) variables with unobserved errors.

Linear Regression Model

- Regression: specifies the relationship between a single dependent variable (Y) and one or more independent variables (x1, x2,…x_k)

- Y
  - Dependent/Response/Target/Outcome Variable
  - A Single Continuous Variable
    - App Sales

- X: x_1, x_2, …x_k
  - Independent/ExplanatoryVariables/Predictors
  - One or More Numerical Variables
    - Price, Review Score, #Ratings, …

- $Y=f(x_1, x_2, \ldots x_k)$
  - *Linear* Relationships

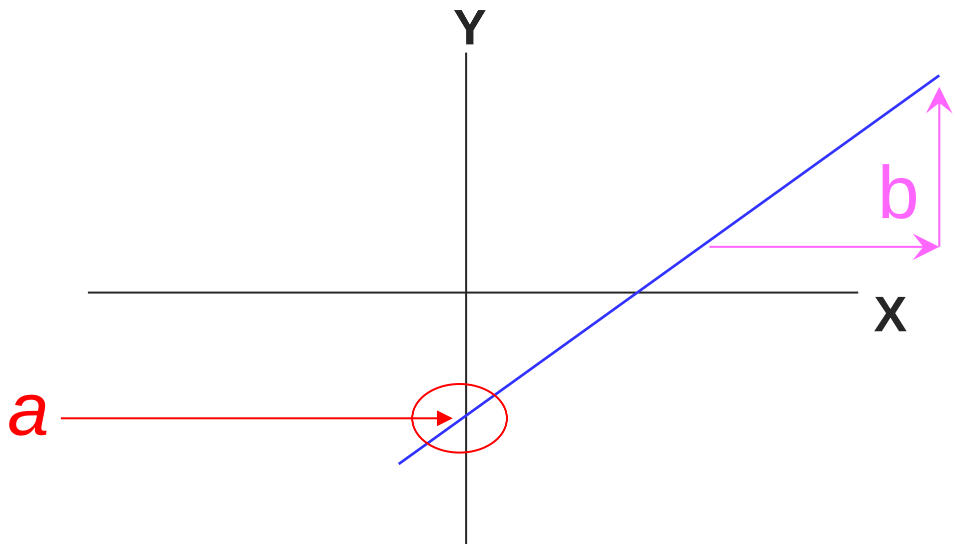
NANYANG TECHNOLOGICAL UNIVERSITY

Correlation vs. Regression

- In correlation, the two variables are treated as equals.
- In regression, one variable is considered independent (=predictor) variable (X) and the other the dependent (=outcome) variable Y.

Linear Regression Model

- Recall from basic algebra that lines can be defined in a *slope-intercept* form: $y = a + bx$
- What are we trying to estimate?
  - Slope: how much the line rises for each increase in x
  - Intercept: the point where the line crosses the y-axis (y when x=0)



NANYANG TECHNOLOGICAL UNIVERSITY

Linear Regression Model

- The first order linear model

$$y = \beta_0 + \beta_1 x + \varepsilon$$

y = Dependent Variable

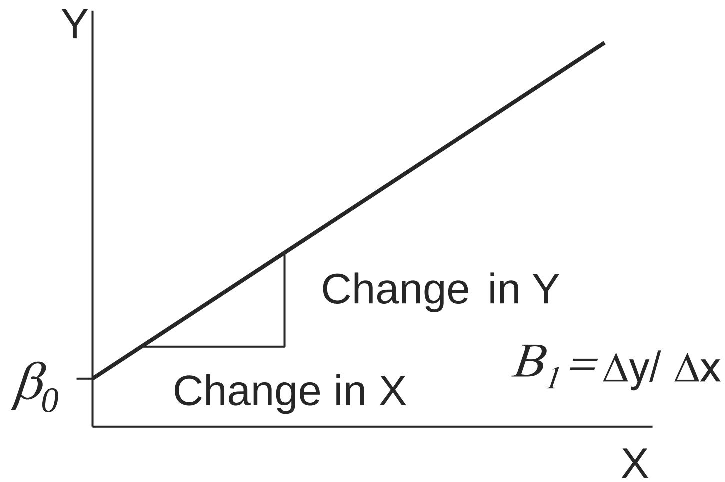
x = Independent Variable

β_0 = y-intercept

β_1 = slope of the line

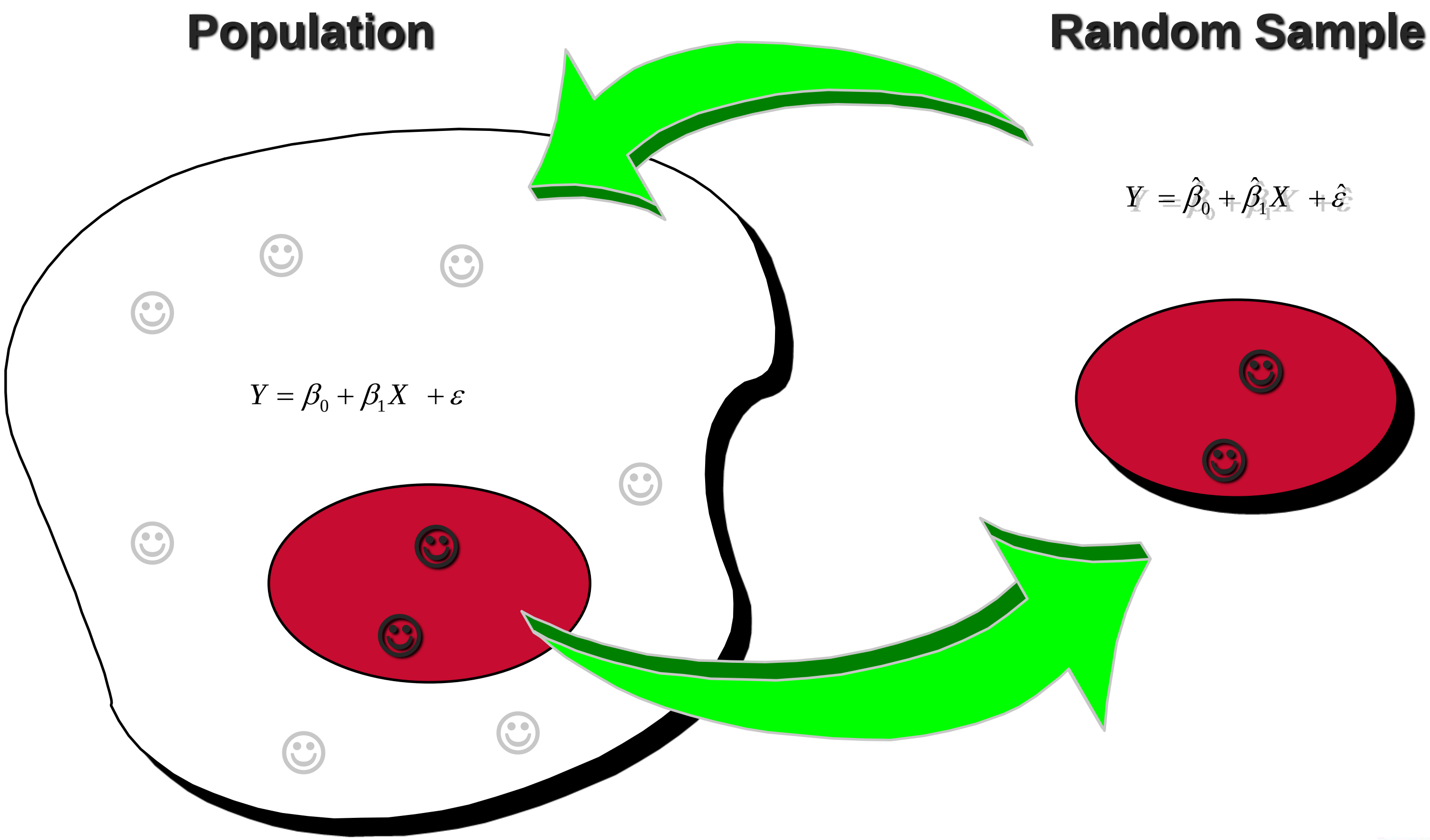
ε = error variable

β_0 and β_1 are not known exactly, but are estimated from sample data. Their estimates are denoted $\hat{\beta}_0$ and $\hat{\beta}_1$.



The output of a regression is a function that predicts the dependent variable based upon values of the independent variables.

Linear Regression Model



Regression Models

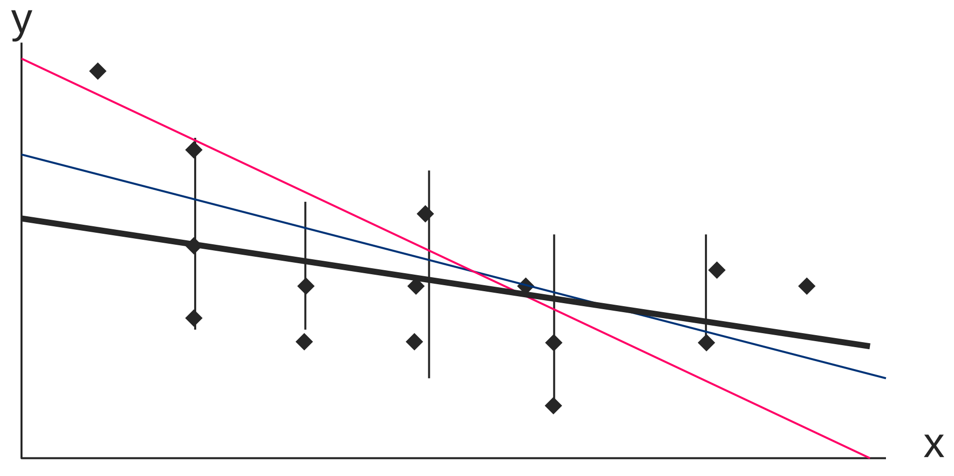
- A *simple regression* model
  - A model with only one independent variable
    - App Sales = f(*Price*)

- A *multiple regression* model
  - A model with multiple independent variables
    - App Sales = f(*Price, Review_Score, Review_Volume, …*)

Model Estimation

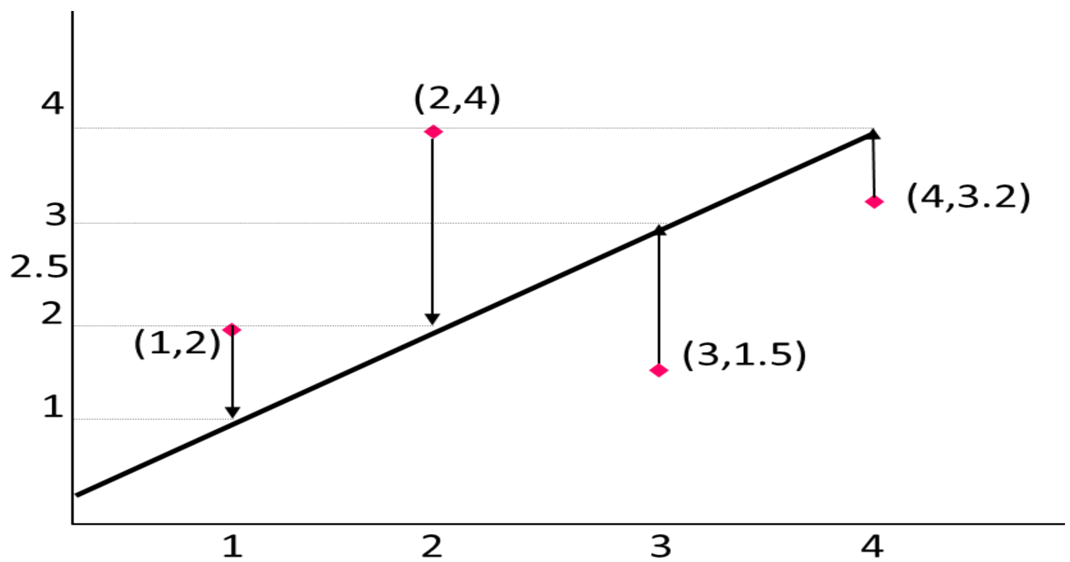
Estimating the Coefficients

- The estimates are determined by
  - producing a straight line that cuts into the data.
  - <u>Which straight line fits best?</u>



Idea behind Estimations

- The best line is the one that minimizes the sum of squared vertical differences between the points and the line.
- The smaller the sum of squared differences the better the fit of the line to the data.



Sum of squared differences = $(2 - 1)^2 + (4 - 2)^2 + (1.5 - 3)^2 + (3.2 - 4)^2 = 6.89$

This value is called the Sum of Squares of Error, or SSE.

NANYANG TECHNOLOGICAL UNIVERSITY

Ordinary Least Squares estimation(OLS)

- Function:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p$$

- OLS estimation

$$\min SSE = \min \sum \left(Y - \hat{Y} \right)^2$$

NANYANG TECHNOLOGICAL UNIVERSITY

Interpretation

Simple Regression Model Example

- A car manufacture wants to find the relationship between the Horse Power (*hp*) and Miles per Gallon (*mpg*)
- A random sample of 32 cars is selected, and the data recorded.
- Simple Regression Model

$$mpg = \beta_0 + \beta_1 hp + \varepsilon$$

- Find the regression line

	mpg	hp
Mazda RX4	21.0	110
Mazda RX4 Wag	21.0	110
Datsun 710	22.8	93
Hornet 4 Drive	21.4	110
Hornet Sportabout	18.7	175
Valiant	18.1	105
Duster 360	14.3	245
Merc 240D	24.4	62
Merc 230	22.8	95
Merc 280	19.2	123
Merc 280C	17.8	123
Merc 450SE	16.4	180
Merc 450SL	17.3	180
Merc 450SLC	15.2	180
Cadillac Fleetwood	10.4	205
Lincoln Continental	10.4	215
Chrysler Imperial	14.7	230
Fiat 128	32.4	66
Honda Civic	30.4	52
Toyota Corolla	33.9	65
Toyota Corona	21.5	97

NANYANG TECHNOLOGICAL UNIVERSITY

Estimation Output in R

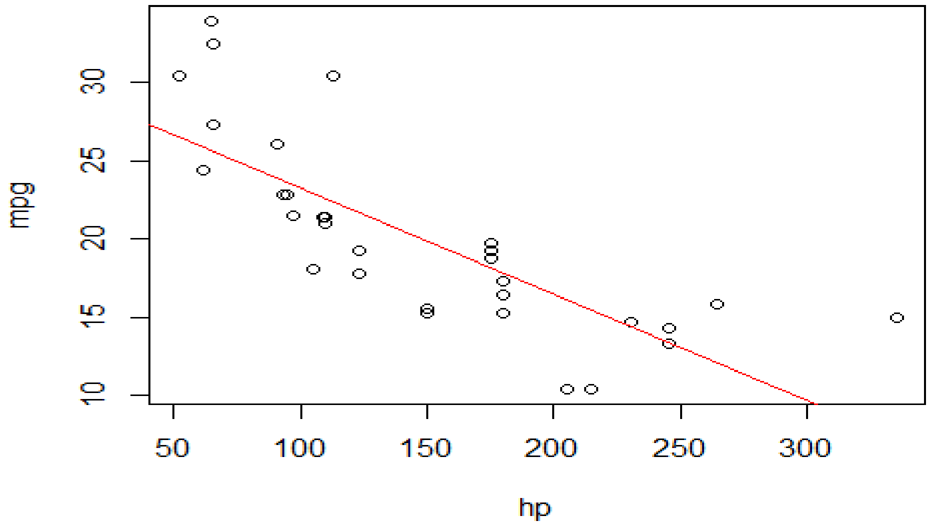
```
Call:
lm(formula = mpg ~ hp, data = mtcars)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-5.7121 -2.1122 -0.8854  1.5819  8.2360
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 30.09886    1.63392  18.421  < 2e-16 ***
hp          -0.06823    0.01012  -6.742 1.79e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 3.863 on 30 degrees of freedom
Multiple R-squared:  0.6024,    Adjusted R-squared:  0.5892
F-statistic: 45.46 on 1 and 30 DF,  p-value: 1.788e-07
```

Estimation Results



$$\hat{y} = 30.09886 - 0.06823x$$

This is the slope of the line. For each additional *hp*, the *mpg* decreases by an average of 0.06823

		Coefficients	Standard Error	t Value	P-value
Intercept	$\hat{\beta}_0$	30.09886	1.63392	18.421	2e-16
hp	$\hat{\beta}_1$	-0.06823	0.01012	-6.742	1.79e-07

Interpretation

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 30.09886    1.63392  18.421  < 2e-16 ***
hp          -0.06823    0.01012  -6.742 1.79e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

*** = 0.1%, ** = 1%, * = 5% Significant levels

- <u>Estimate</u>: the predicted coefficient for the intercept or slope of a predictor
  - A car has 30.09886 *mpg* when *hp* = 0 (No car exists with zero horse power)
  - A one-unit increase in *hp* decreases *mpg* by 0.06823
- <u>Standard Error</u>: a measure of the accuracy of predictions
- <u>t-value</u>: the estimate divided by its std. error.
- <u>p-value</u>: the probability that the true coefficient is zero given the value of the estimate (i.e., *the chance of having a zero effect on the dependent variable*)
- Small p-values suggest that the predictor is extremely unlikely to have no relationship with the dependent variable.
  - An independent variable is highly unlikely to have a zero effect on the dependent variable.

How to Interpret P-value

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 30.09886    1.63392  18.421  < 2e-16 ***
hp          -0.06823    0.01012  -6.742 1.79e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

P-value	Significant Level	Stars
p <= 0.001	**0.1%**	***
0.001 < p <= 0.01	**1%**	**
0.01 < p <= 0.05	**5%**	*
0.05 < p	**Insignificant**	.

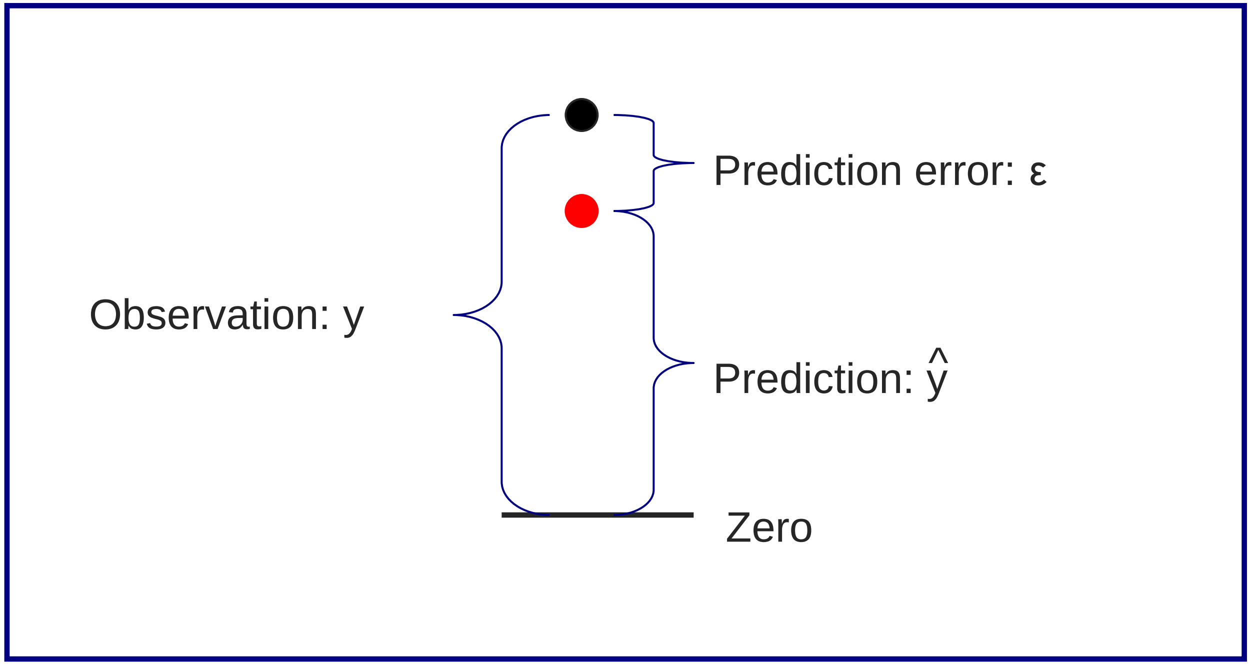
Note that some of p-values have stars (***), which correspond to the footnotes to indicate the significant level met by the estimate

- A negative association between *hp* and *mpg* is ***statistically significant*** at the 0.1% level (=***).
- Or, *hp* has a negative impact on *mpg* at the 0.1% significant level.

NANYANG TECHNOLOGICAL UNIVERSITY

Model Evaluation

Residuals

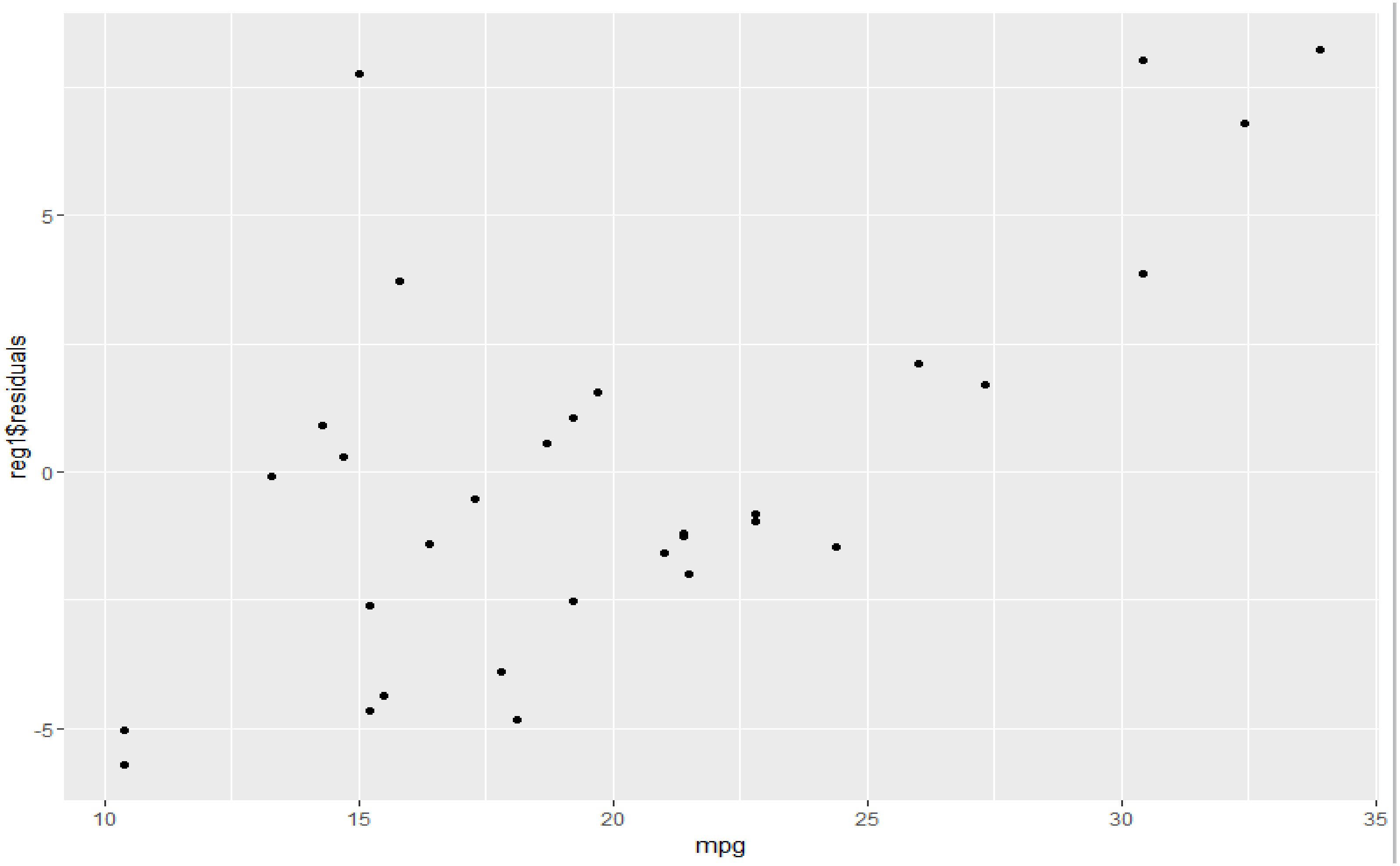


For each observation, the variation can be described as:

$$y = \hat{y} + \varepsilon$$

Actual = predicted + Error (Residual)

Plot the predicted values against the actual values



Example

```
Residuals:
    Min      1Q  Median      3Q     Max
-5.7121 -2.1122 -0.8854  1.5819  8.2360
```

- Summary statistics for the errors in the predictions
  - *Residual* (**ε**) *= Actual Value (y) – Predicted Value* ($\hat{y}$)
  - Maximum Residual
    - The model under-predicted by nearly 8 *mpg* for at least one observation.
  - Minimum Residual
    - The model over-predicted by nearly 6 *mpg* for at least one observation
  - Residuals in 1Q (25th percentile) and 3Q (75th percentile)
    - The majority of predictions (50 percent of residuals) were between *2.1122 mpg* over the actual value and *1.5819 mpg* under the actual value.

Evaluating Model Fit

R-squared Value

- R^2 is called the coefficient of determination.
  - Indicates the portion of the variance in the dependant variable that is predictable from the independent variable(s)
- R^2 measures how close the data is to the fitted regression line.
- R^2 takes on any value between zero and one.

  $R^2 = 1$: Perfect match between the line and the data points.

  $R^2 = 0$: There are no linear relationship between x and y.

Adjusted R-squared

- A better goodness of fit measure is the adjusted R^2, which is computed as follows:

$$Adjusted\ R^2 = 1 - \frac{n-1}{n-p-1}(1 - R^2)$$

  - p is the total number of explanatory variables in the model (not including the constant term)
  - n is the sample size
- The adjusted *R-squared* adjusts for the number of predictors in the model.
  - It decreases when a predictor improves the model by less than expected by chance.

NANYANG TECHNOLOGICAL UNIVERSITY

Model Specification Test

- F-Test
  - Check if the predictor are jointly significant
  - Test if $\beta_1 = \beta_2 = \ldots = 0$
  - Low p-value indicates that the predictors are jointly significant and the corresponding estimates are not zeros.

NANYANG TECHNOLOGICAL UNIVERSITY

Example

```
Residual standard error: 3.863 on 30 degrees of freedom
Multiple R-squared:  0.6024,    Adjusted R-squared:  0.5892
F-statistic: 45.46 on 1 and 30 DF,  p-value: 1.788e-07
```

- R-Squared value
  - About 60% of the variation in *mpg* is explained by *hp*. The rest (40%) remains unexplained by this model.
- F-Statistic
  - p-value of 1.788e-07 means the predictor, *hp*, is not highly likely to have a zero effect on *mpg*.

Evaluating Predictive Performance

Measuring Predictive Error

- Not the same as *R*-squared values.

- We want to know how well the model predicts **new data**, not how well it fits the data it was trained with

- Key component of most measures is difference between actual *y* and predicted *y* ("error")

NANYANG TECHNOLOGICAL UNIVERSITY

Mean Squared Error (MSE)

- Actual values: $a_1, a_2, \ldots, a_n$
- Predicted values: $p_1, p_2, \ldots, p_n$
- Error $e_i = (p_i - a_i)$
- Mean Square Error MSE = $[e_1^2 + \ldots + e_n^2]/n$

  = SSE / n

Mean Squared Error (MSE)

Predicted Model

$$\hat{mpg} = 30.09886 - 0.06823 * hp$$

ID	Car	*hp*	Actual *mpg*	Predicted *mpg*
1	Mazda RX4	110	21.0	22.59
2	Mazda RX4 Wag	110	21.0	22.59
3	Datsun 710	93	22.8	23.75
4	Hornet 4 Drive	110	21.4	22.59

MSE = $[(21.0 - 22.59)^2 + (21.0 - 22.59)^2 + (22.8 - 23.75)^2 + (21.4 - 22.59)^2] + \ldots / n$

, where n is the number of observations

Multiple Linear Regression

Multiple Linear Regression

- More than one predictor…

$$Y = \beta_0 + \beta_1 {*}X + \beta_2 {*}W + \beta_3 {*}Z + \ldots + \varepsilon$$

Each regression coefficient is the amount of change in the outcome variable that would be expected per one-unit change of the predictor, if all other variables in the model are held constant.

NANYANG TECHNOLOGICAL UNIVERSITY

Multiple Linear Regression…

- More predictors improve predictions
- Outcomes can be influenced by various factors
  - *mpg* could be affected by horse power (*hp)*
  - *mpg* could be affected by *cylinders (cyl)*
  - *mpg* could be affected by *transmission type (am)*
  - …
- Multiple Liner Regression Model for *mpg*
  - $mpg = \beta_0 + \beta_1 * hp + \beta_2 * cyl + \beta_3 * am + \varepsilon$

Estimation Output

```
Residuals:
   Min     1Q Median     3Q    Max
-4.864 -1.811 -0.158  1.492  6.013
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 30.88834    2.78422  11.094 9.27e-12 ***
hp          -0.03688    0.01452  -2.540  0.01693 *
cyl         -1.12721    0.63417  -1.777  0.08636 .
am           3.90428    1.29659   3.011  0.00546 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 2.807 on 28 degrees of freedom
Multiple R-squared:  0.8041,    Adjusted R-squared:  0.7831
F-statistic: 38.32 on 3 and 28 DF,  p-value: 4.791e-10
```

*** = 0.1%, ** = 1%, * = 5% Significant levels