BUDT 730

Data, Models and Decisions

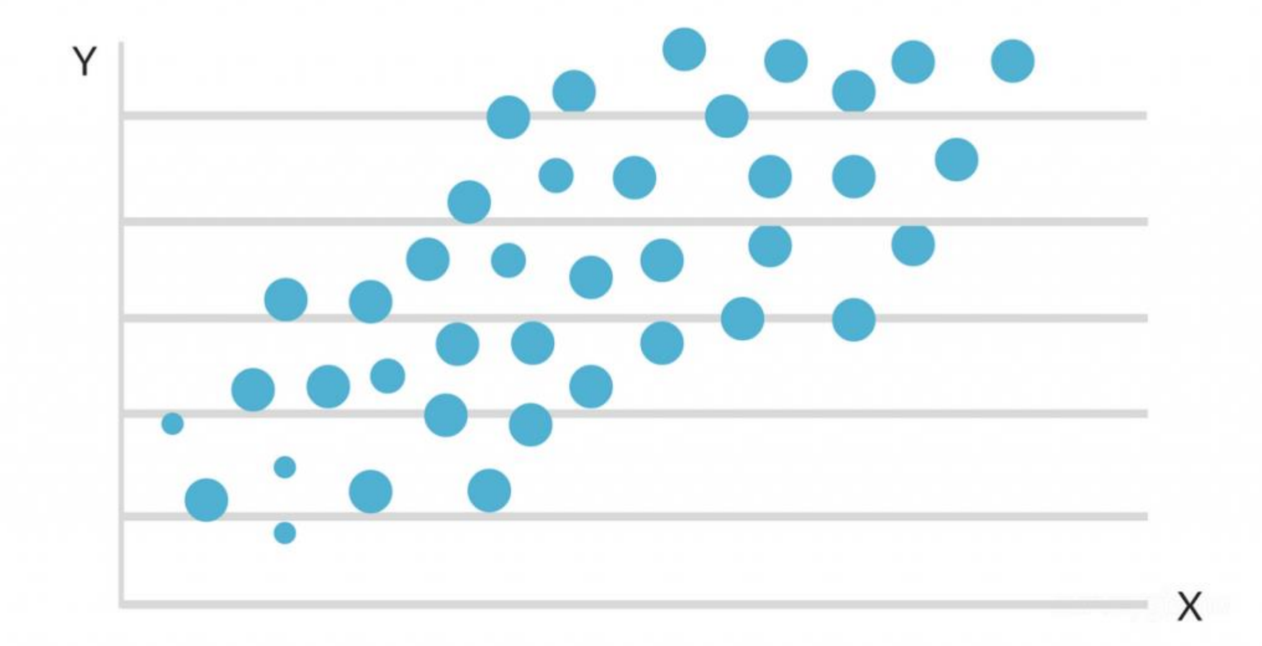
Lecture 11

Regression Analysis (3)

Interpretation of Regression Model

Prof. Sujin Kim

Regression Analysis



Simple Regression

Data file: Airline_data.xlsx

Example: Southwest Airline Data

S_CODE	S_CITY	E_CODE	E_CITY	COUPON	NEW	VACATION	SW	HI	S_INCOME	E_INCOME
*	Dallas/Fort	*	Amarillo	1.00	3	No	Yes	5291.99	$28,637	$21,112
*	Atlanta	*	Baltimore/Wash	1.06	3	No	No	5419.16	$26,993	$29,838
*	Boston	*	Baltimore/Wash	1.06	3	No	No	9185.28	$30,124	$29,838
ORD	Chicago	*	Baltimore/Wash	1.06	3	No	Yes	2657.35	$29,260	$29,838
MDW	Chicago	*	Baltimore/Wash	1.06	3	No	Yes	2657.35	$29,260	$29,838
*	Cleveland	*	Baltimore/Wash	1.01	3	No	Yes	3408.11	$26,046	$29,838
*	Dallas/Fort	*	Baltimore/Wash	1.28	3	No	No	6754.48	$28,637	$29,838
							es	5584.00	$26,752	$29,838

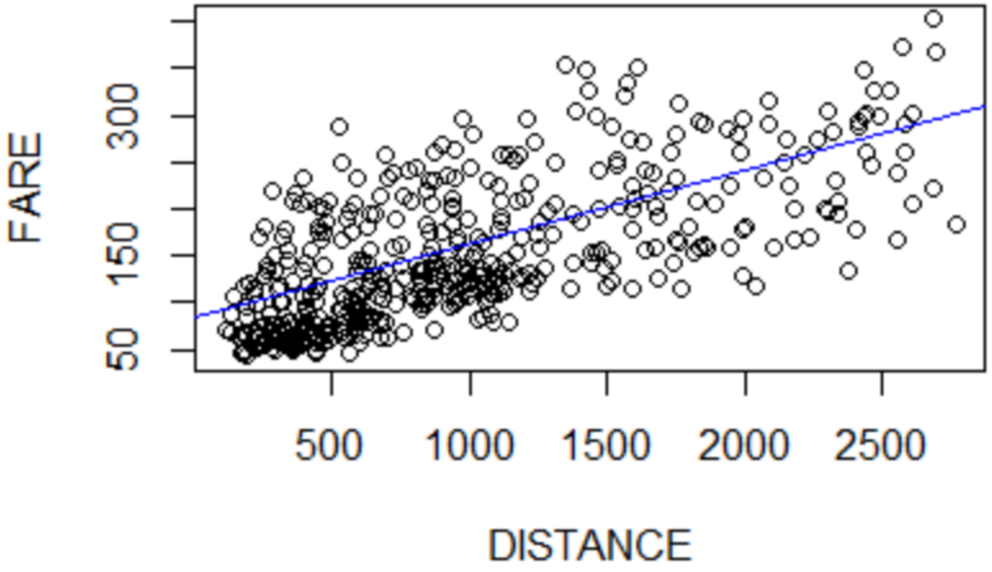
E_POP	SLOT	GATE	DISTANCE	PAX	FARE
205711	Free	Free	312	7864	$64.11
7145897	Free	Free	576	8820	$174.47
7145897	Free	Free	364	6452	$207.76
7145897	Controlled	Free	612	25144	$85.47
7145897	Free	Free	612	25144	$85.47
7145897	Free	Free	309	13386	$56.76
7145897	Free	Free	1220	4625	$228.00
7145897	Free	Free	921	5512	$116.54
7145897	Free	Free	1249	7811	$172.63
7145897	Free	Free	964	4657	$114.76
7145897	Free	Free	2104	4489	$158.20
7145897	Free	Free	2329	7349	$228.99
7145897	Free	Free	587	5654	$79.17
7145897	Free	Free	992	3525	$132.05

Dependent variable:
Which variable would you like to analyze?

We would like to investigate which variables relate to “Fare”.

Simple Regression Model with Distance

```
> attach(Airline_data)
> slr<-lm(FARE~DISTANCE)
> summary(slr)

Call:
lm(formula = FARE ~ DISTANCE)

Residuals:
    Min      1Q  Median      3Q     Max 
-137.59  -45.36  -10.52   40.49  163.41 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 83.976532   4.051412   20.73   <2e-16 ***
DISTANCE     0.078819   0.003463   22.76   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1

Residual standard error: 56.48 on 636 degrees of freedom
Multiple R-squared:  0.4489,	Adjusted R-squared:  0.4481 
F-statistic: 518.1 on 1 and 636 DF,  p-value: < 2.2e-16

> plot(FARE~DISTANCE)
> abline(slr,col="blue")
```



Model Coefficients: $Y = a + bX + \varepsilon$

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 83.976532   4.051412   20.73   <2e-16 ***
DISTANCE     0.078819   0.003463   22.76   <2e-16 ***
```

- Interpretation of b
  - **On average ,**
  - **one unit increase in X is associated with b units increase in Y**
- Recall our regression formula for Southwest example:

  Fare = 83.98 + 0.0788 * Distance

- Interpretation of b:
  - <u>On average</u>, one mile increase in distance is associated with 7.88 cents increase in fare.
  - In more plain English,"for each additional mile travelled, the average fare increases by 7.88 cents."
- The interpretation of the intercept, a is less important.
  - Often it is hard to find a meaningful interpretation.

Fitting a Regression Model

- The **least-squares estimation (LSE) method** generates the best-fitting line through the observed values, minimizing the sum of squared errors
  - The **sum of squared errors (SSE**) is also known as the **residual sum of squares** (**RSS**)
- Why the method of least-squares?
  - The best unbiased estimator:
    - ‘Best’ means the smallest variance
    - ‘Unbiased means no bias
- Key results from R:
  - Standard error of the estimate, R^2
  - Model coefficients
  - Overall fit: F- test

Measure of Error: Standard Error

- The magnitude of the residuals describe how useful the regression line is for predicting Y from X
- **Standard error of the estimate** (s_e)
  - Essentially the standard deviation of the residuals

$$s_e = \sqrt{\frac{\sum e_i^2}{n-2}}$$

$e_i = Y_i - \hat{Y}_i$ (observed – fitted) is the residual of the i^{th} observation

  - Measures how tightly the data fits around the regression line
  - The smaller the better
  - The regression line minimizes s_e

```
Residual standard error: 56.48 on 636 degrees of freedom
Multiple R-squared:  0.4489,	Adjusted R-squared:  0.4481
F-statistic: 518.1 on 1 and 636 DF,  p-value: < 2.2e-16
```

Sample size = 658

Measure of Model Fit : R^2

- R^2 is perhaps the most commonly used measure for statistical models
- It measures the proportion of total variation of Y that is explained by the regression model
- Essentially, how much better does the model explain Y than simply using the mean of Y?

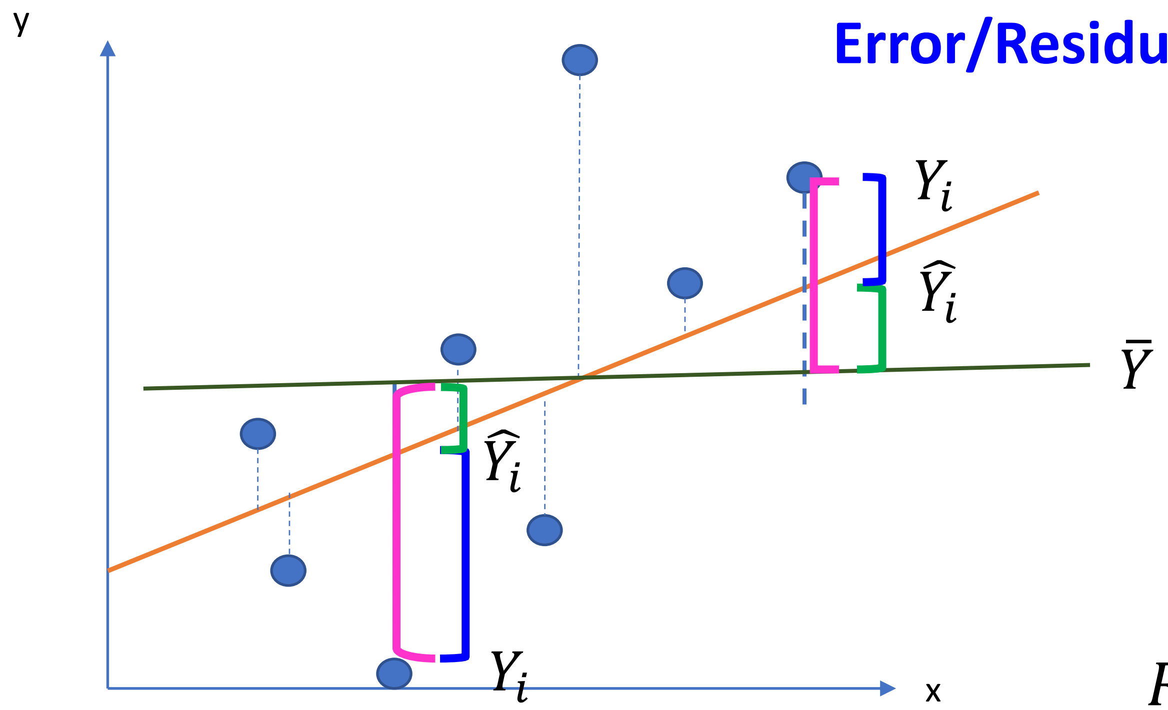
```
Residual standard error: 56.48 on 636 degrees of freedom
Multiple R-squared:  0.4489,	Adjusted R-squared:  0.4481
F-statistic: 518.1 on 1 and 636 DF,  p-value: < 2.2e-16
```

Measure of Model Fit : R^2

Total variation: $Y_i - \bar{Y} = (\hat{Y}_i - \bar{Y}) + (Y_i - \hat{Y}_i)$

Variation due to the regression (explained): $= \hat{Y}_i - \bar{Y}$

Error/Residual (unexplained): $e_i = Y_i - \hat{Y}_i$



$$\underbrace{\sum (Y_i - \bar{Y})^2}_{STT} = \underbrace{\sum (\hat{Y}_i - \bar{Y})^2}_{SSR} + \underbrace{\sum e_i^2}_{RSS}$$

$$STT = SSR + RSS$$

$$R^2 = \frac{\sum (Y_i - \bar{Y})^2 - \sum (Y_i - \hat{Y}_i)^2}{\sum (Y_i - \bar{Y})^2} = \frac{SSR}{STT}$$

SST, SSR & RSS

- The sum of the squared *total variation* (*SST*):

$$\text{SST} = \sum(Y_i - \bar{Y})^2$$

- The residual sum of the squares (RSS or SSE) - The part *unexplained* by the regression equation:

$$\text{RSS} = \sum e_i^2$$

- The sum of the squares due to regression (SSR) - The part that is *explained*:

$$SSR = \sum\left(\hat{Y}_i - \bar{Y}\right)^2 = SST - RSS$$

$$R^2 = \frac{SSR}{SST} = 1 - \frac{RSS}{SST}$$

Measure of Model Fit : R^2

$$R^2 = \frac{SSR}{SST}$$

- R^2 is always between 0 and 1 – the larger the better
  - When the residuals are small, R^2 is close to 1
  - When the residuals are large, R^2 is close to 0
- <u>Improves if you add additional variables to the model</u>

Measure of Model Fit: Adjusted R^2

- Adjusted R^2 for multiple regression (regression model with multiple independent variables)
  - Adjusted R^2 is an alternative measure that adjusts R^2 for the number of variables in the equation; it is used to monitor whether extra variables are actually helping
  - Does not always improve when additional variables are added
  - Is always between 0 and 1 – the larger the better
  - Does *not* have the interpretation of proportion of variation in *Y* explained by the model

Model Coefficients: $Y = a + bX + \varepsilon$
Sampling Distribution of the Slope

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 83.976532   4.051412   20.73   <2e-16 ***
DISTANCE     0.078819   0.003463   22.76   <2e-16 ***
```

- Since the slope (*b*) is obtained from a sample, it is a sample statistic and consequently, it is a random variable.
- It has a probability distribution
- Its expected value is the **population slope** (β): $E[b] = \beta$
- It can be mathematically derived that the sampling distribution of:

$$T = \frac{b - \beta}{s_b}$$

is a t-distribution with $n - k - 1$ degrees of freedom (k=# of independent variables used in the regression model)

Sampling Distribution of the Regression Coefficients

In words:

- The point estimate *b* is **unbiased** → E[b] = β
- The sampling distribution of *b* is t-distribution (so it is symmetric and bell-shaped)
- *t-value* represents the normalized error (standard) between the point estimate (*b*) and the true population parameter (β)

Testing Usefulness of a Predictor Variable

- If X is not a useful predictor for Y, then β must be equal to __
- If it is a useful predictor for Y, then _______
- However, β is unknown
- We use the estimate to conduct a hypothesis test to check whether ______________
- The hypothesis test:

Testing Usefulness of a Predictor Variable

- If X is not a useful predictor for Y, then β must be equal to <u>0</u>
- If it is a useful predictor for Y, then $\underline{\beta \neq 0.}$
- However, β is unknown
- We use the estimate to conduct a hypothesis test to check whether β= 0 or not.
- The hypothesis test:
  - $H_0: \beta = 0$
  - $H_1: \beta \neq 0$

Testing Usefulness of a Predictor Variable

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 83.976532   4.051412   20.73   <2e-16 ***
DISTANCE     0.078819   0.003463   22.76   <2e-16 ***
```

- Test statistic: $T = b / s_b$
  - Confidence interval: $b \pm$ t-multiple $* s_b$
- Interpretations
  - **Small p value**: Reject H_0
    - Strong evidence that $\beta \neq 0$
    - **Independent variable is meaningful** to add the variable to the model.
  - Large p-value: Do not reject H_0
    - Little evidence that $\beta \neq 0$
    - Independent variable provides little to no value to the model strong evidence to reject Ho- We may remove the variable.

F Test for Overall Significance of the Model

Question: Is the overall model significant?

```
Residual standard error: 56.48 on 636 degrees of freedom
Multiple R-squared:  0.4489,	Adjusted R-squared:  0.4481
F-statistic: 518.1 on 1 and 636 DF,  p-value: < 2.2e-16
```

- F-test shows if there is a linear relationship between all of the X variables considered together and Y.

- Hypotheses:

  H_0: $\beta_1 = \beta_2 = \cdots = \beta_k = 0$ (all coefficients are zero, no linear relationship)

  H_1: at least one $\beta_i \neq 0$ (at least one coefficient not zero, at least one independent variable relates with Y)

Small p-value => The model is overall significant

"Your model provides a better fit than the intercept-only model (base model with no independent variables".

F Test for Overall Significance (Optional)

- Test statistic:

$$F_{STAT} = \frac{MSR}{MSE} = \frac{\frac{SSR}{k}}{\frac{SSE}{n-k-1}}$$

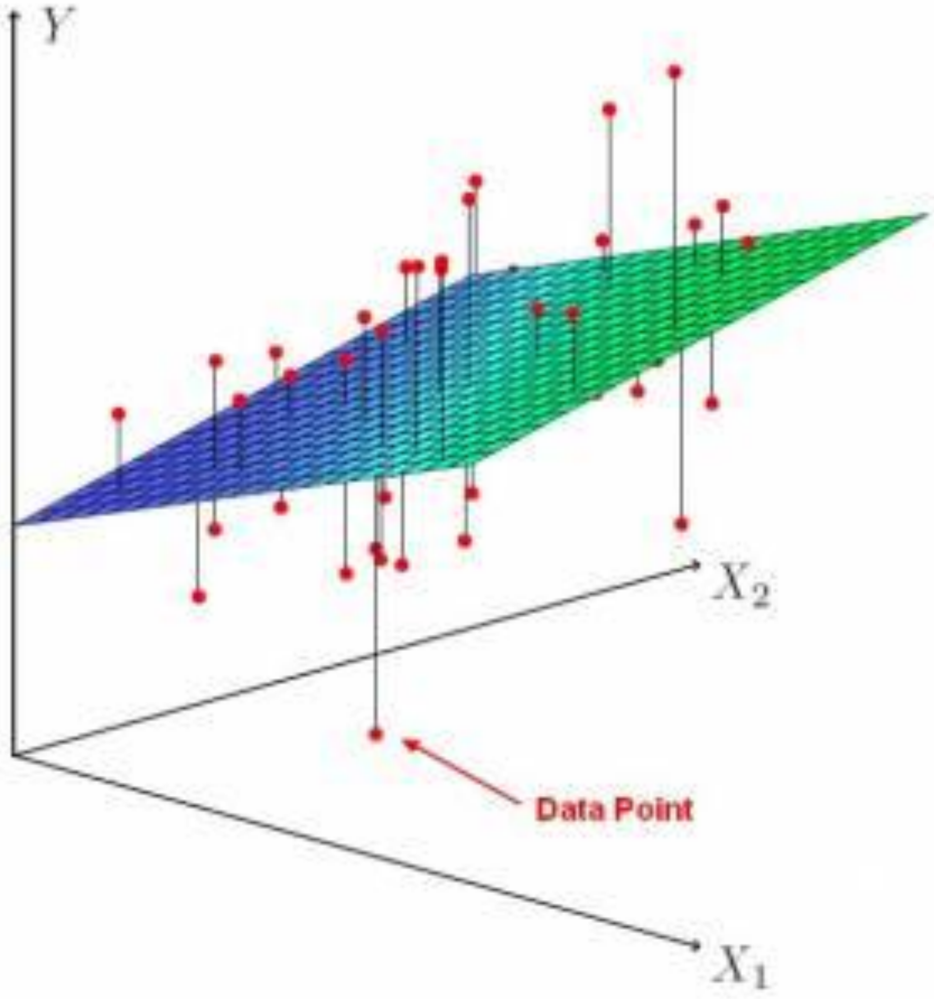
where F_{STAT} has numerator d.f. = k and

denominator d.f. = (n – k - 1)

Regression Analysis

Multiple Regression

- Interpretation
- Variable Transformation



Multiple Regression

- Oftentimes, a single independent variable is not sufficient to produce a good fit
- When we include more than one independent variable to obtain a better fit, we have a **multiple regression** model
  - The regression equation is still estimated by least squares, but now there is a slope term for each independent variable

$$Y = a + b_1 X_1, + \ldots + b_k X_k + \varepsilon$$

Interpretation in Multiple Regression

- Multiple regression output is similar to the simple case
  - The standard error of the estimate is interpreted the same, but the denominator is adjusted for the number of estimated independent variables ($n - k - 1$)
  - R^2 is also the same, but the drawback is that it only <u>increases with the number of variables in the model</u>

Interpretation in Multiple Regression

- Model coefficients
  - When interpreting a change in Y as a function of a change in an X, we must include '**all else being held constant**'
  - Interpretation of b_i
    - **On average ,**
    - **one unit increase in X_i is associated with b_i units increase in Y**
    - **if all else held equal (or all else being held constant)**

Multiple Regression: Southwest

Simple Regression

```
Residual standard error: 56.48 on 636 degrees of freedom
Multiple R-squared:  0.4489,	Adjusted R-squared:  0.4481
F-statistic: 518.1 on 1 and 636 DF,  p-value: < 2.2e-16
```

```
> mlr<-lm(FARE~DISTANCE+PAX)
> summary(mlr)

Call:
lm(formula = FARE ~ DISTANCE + PAX)

Residuals:
    Min      1Q  Median      3Q     Max 
-137.54  -45.26  -11.44   40.21  162.39 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) 85.8780647  4.7760691  17.981   <2e-16 ***
DISTANCE     0.0785506  0.0034823  22.557   <2e-16 ***
PAX         -0.0001283  0.0001705  -0.752    0.452    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 56.5 on 635 degrees of freedom
Multiple R-squared:  0.4494,	Adjusted R-squared:  0.4477
F-statistic: 259.2 on 2 and 635 DF,  p-value: < 2.2e-16
```

Interpretation in Multiple Regression

- Assume that the multiple regression model is valid (this this later!)

$$FARE = 85.89 + 0.07855\ DISTANCE - 0.0001283\ PAX$$

- <u>On average</u>, one passenger increase in PAX is associated with .01283 cents decrease in FARE <u>if DISTANCE remains unchanged</u>

- What if DISTANCE also changes?
  - The change in FARE is no longer simply b_{PAX}, but can be easily calculated from the above equation. In general,

$$\Delta\ FARE = 0.07855\ \Delta\ DISTANCE - 0.0001283\ \Delta\ PAX$$

Δvariable: difference in the variable

Multiple Regression

- Observation
  - The coefficient of PAX is much smaller (in magnitude) than the coefficient of DISTANCE. Is PAX less important for predicting/explaining FARE than DISTANCE?

  - NO! In general, "Importance" of a variable not linked to the size of regression coefficient
  - However, the p-value of PAX is large (.4520), so in this model we can conclude that PAX is less important based on the p-value, not based on the coefficient.

Next : Variable Transformations

- Several types of independent variables can be used in regression equations:
  - Dummy variables
  - Interaction variables
  - Nonlinear transformations
- We should be selective, and not include too many different types in a particular regression model
  - Only a few might improve the fit!