LAB SESSION 9 – MULTIPLE LINEAR REGRESSION

Analytics Primer

MULTIPLE LINEAR REGRESSION

Inference

Example

- A real estate company is trying to model housing prices (in dollars) of their customers with the variables:
  - x_1: Size of Home (square feet)
  - x_2: Age of Home (years)
  - x_3: Acreage of Land (acres)
  - x_4: Number of Bedrooms
- Using a sample of 105 houses they derive the following model:

$$\hat{y} = 24{,}312 + 86.5x_1 - 324x_2 + 9{,}610x_3 + 3{,}617x_4$$

$SSE = 27695831$ $\quad$ $SSR = 45963293$ $\quad$ $TSS = 73659124$

Example

$$\hat{y} = 24{,}312 + 86.5x_1 - 324x_2 + 9{,}610x_3 + 3{,}617x_4$$

$SSE = 27695831$ $\quad SSR = 45963293$ $\quad TSS = 73659124$

1. Test the overall significance of the model.

$$H_0:\ \beta_1 = \beta_2 = \beta_3 = \cdots = \beta_k = 0$$

$$H_a:\ \text{At least one coefficient is nonzero}$$

$$MSR = \frac{45963293}{4} = 11490823.25$$

$$MSE = \frac{27695831}{105 - 4 - 1} = 276958.31$$

$$F = \frac{MSR}{MSE} = 41.49$$

P-value < 0.05 → REJECT H_0

Example

$$\hat{y} = 24{,}312 + 86.5x_1 - 324x_2 + 9{,}610x_3 + 3{,}617x_4$$

$SSE = 27695831$ $\quad$ $SSR = 45963293$ $\quad$ $TSS = 73659124$

2. Test the individual significance of the variable x_3.

$$s_{\hat{\beta}_3} = 3313$$

$H_0: \beta_3 = 0$

$H_a: \beta_3 \neq 0$

$$t = \frac{9610 - 0}{3313} = 2.9$$

P-value = (0.002, 0.01) → REJECT H_0

Example

$$\hat{y} = 24{,}312 + 86.5x_1 - 324x_2 + 9{,}610x_3 + 3{,}617x_4$$

$SSE = 27695831$ $\qquad SSR = 45963293$ $\qquad TSS = 73659124$

3. Test the individual significance of the remaining variables. Should any be removed from the model?

$s_{\widehat{\beta}_1} = 7109$ P-value ~ 1 → DO NOT REJECT H_0

$s_{\widehat{\beta}_2} = 15$ P-value < 0.001 → REJECT H_0

$s_{\widehat{\beta}_4} = 3480$ P-value = (0.3, 0.4) → DO NOT REJECT H_0

MULTIPLE LINEAR REGRESSION

Categorical Predictors

Example

- Develop both effects coding and dummy / reference coding for a categorical variable with 4 categories.

	x_1	x_2	x_3
A	1	0	0
B	0	1	0
C	0	0	1
D	-1	-1	-1

	x_1	x_2	x_3
A	1	0	0
B	0	1	0
C	0	0	1
D	0	0	0

Example

- A real estate company is trying to model housing prices (in dollars) of their customers with the variables:
  - x_1: Size of Home (square feet)
  - x_2: Age of Home (years)
  - x_3: Acreage of Land (acres)
  - x_4: Number of Bedrooms
  - x_5: Located on golf course
- Using a sample of 105 houses they derive the following model:

$$\hat{y} = 24{,}312 + 86.5x_1 - 324x_2 + 9{,}610x_3 + 3{,}617x_4 + 12{,}550x_5$$

$$s_{\hat{\beta}_5} = 4532$$

Example

$$\hat{y} = 24{,}312 + 86.5x_1 - 324x_2 + 9{,}610x_3 + 3{,}617x_4 + 12{,}550x_5$$

$$s_{\hat{\beta}_5} = 4532$$

1. How would you code the variable summarizing whether a house was on the golf course?

$$x_5 = \begin{cases} 1 & \text{if on golf course} \\ 0 & \text{if not on golf course} \end{cases}$$

Example

$$\hat{y} = 24{,}312 + 86.5x_1 - 324x_2 + 9{,}610x_3 + 3{,}617x_4 + 12{,}550x_5$$

$$s_{\hat{\beta}_5} = 4532$$

2. What is the interpretation of the coefficient on the variable x_5?
    - The **average** increase in home price for home on a golf course compared to not is $12,550, **all else equal**.

Example

$$\hat{y} = 24{,}312 + 86.5x_1 - 324x_2 + 9{,}610x_3 + 3{,}617x_4 + 12{,}550x_5$$

$$s_{\hat{\beta}_5} = 4532$$

3. Calculate the test of significance for the variable x_5.

$H_0: \beta_5 = 0$

$H_a: \beta_5 \neq 0$

$$t = \frac{12550 - 0}{4532} = 2.77$$

P-value = (0.002, 0.01) → REJECT H_0

MULTIPLE LINEAR REGRESSION

Polynomial Predictors

More Examples

- The plot is fitted with a quadratic model for x predicting y. From the above plot, what can you determine about the sign of the coefficient estimate for the quadratic term of x?