Stat 401 B – Lecture 12

Multiple Regression

- A single numerical response variable, Y.
- Multiple numerical explanatory variables, X_1, X_2,..., X_k

Multiple Regression

$$Y = \mu_{Y|x_1, x_2, \ldots, x_k} + \varepsilon$$

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon$$

Example

- Y, Response – Effectiveness score based on experienced teachers' evaluations.
- Explanatory – Test 1, Test 2, Test 3, Test 4.

Response EVAL

Summary of Fit

RSquare	0.802861
RSquare Adj	0.759052
Root Mean Square Error	37.53627
Mean of Response	444.4783
Observations (or Sum Wgts)	23

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	4	103286.25	25821.6	18.3265
Error	18	25361.49	1409.0	**Prob > F**
C. Total	22	128647.74		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-193.4994	125.3074	-1.54	0.1399
Test1	1.1158539	0.319746	3.49	0.0026*
Test2	2.243267	0.628449	3.57	0.0022*
Test3	-1.367001	0.563965	-2.42	0.0261*
Test4	6.0482387	1.202281	5.03	<.0001*

Prediction Equation

- Predicted Evaluation = –193.50 + 1.116*Test1 + 2.243*Test2 – 1.367*Test3 + 6.048*Test4

Conditions

- The random error term, ε, is
  - Independent
  - Identically distributed
  - Normally distributed with standard deviation, σ.

Estimate of Error Variance, σ^2

$$MS_{\text{Error}} = \frac{SS_{\text{Error}}}{df_{\text{Error}}}$$

$$MS_{\text{Error}} = \frac{\sum (y - \hat{y})^2}{n - (k+1)}$$

$$MS_{\text{Error}} = \frac{25361.49}{18} = 1409.0$$

7

Estimate of Error Std Dev, σ

- Root Mean Square Error

$$\text{RMSE} = \sqrt{MS_{\text{Error}}}$$

$$\text{RMSE} = \sqrt{1409.0} = 37.54$$

8

Multiple R^2

$$R^2 = \frac{SS_{\text{Model}}}{SS_{\text{Total}}} = 1 - \frac{SS_{\text{Error}}}{SS_{\text{Total}}}$$

$$R^2 = \frac{103286.25}{128647.74} = 0.802861$$

9

Interpretation

- 80.3% of the variation in the evaluation scores can be explained by the model, i.e. the relationship with the explanatory variables.

Caution

- Including additional explanatory variables in a model can only increase the value of R^2, even if those explanatory variables have nothing to do with the response variable.

Adjusted R^2

$$\text{adj}R^2 = 1 - \frac{\text{MS}_{\text{Error}}}{\text{MS}_{\text{Total}}}$$

$$\text{adj}R^2 = 1 - \frac{(25361.49/18)}{(128647.74/22)} = 0.75905$$

Stat 401 B – Lecture 12

Test of Model Utility

- Is there any explanatory variable in the model that is helping to explain significant amounts of variation in the response?

13

Step 1: Hypotheses

$$\mathrm{H}_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$$

$$\mathrm{H}_A : \text{at least one parameter is not zero}$$

14

Step 2: Test Statistic

$$F = \frac{\mathrm{MS}_{\mathrm{Model}}}{\mathrm{MS}_{\mathrm{Error}}}$$

$$F = \frac{25821.6}{1409.0} = 18.3265$$

$$\mathrm{P-value} < 0.0001$$

15

Stat 401 B – Lecture 12

Step 3: Decision

- Reject the null hypothesis because the P-value is so small.

16

Step 4: Conclusion

- At least one of the tests is providing statistically significant information about the evaluation score.
- The model is useful. Maybe not the best, but useful.

17

Alternative Form

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

$$F = \frac{\left(0.802861 / 4\right)}{\left(0.197139 / 18\right)} = 18.3265$$

$$\text{P}-\text{value} < 0.0001$$

18