The Proof of the Pudding is in the Eating

How do you know if you have a useful predictive model?

One might think that a useful model would have a high R^2. But, is a high R^2 necessary to have a useful model? In the introduction to Chapter 14, I mentioned a story relating my first experience using regression analysis in predicting the speed of a horse in a race. The R^2 of that model was roughly 0.35. Yet the model predicted well enough to allow us to have a profitable betting experience. Later, this same friend and I would start a company that predicted stock prices. The R^2 associated with many of our models was less than 0.1. Yet, we were able to profitably trade substantial volumes of stock with these models.

To judge how effective a model is, you need to use it for its intended purpose. Thus, there are two questions for any predictive model. First, can you predict better with the model than without it? Second, can your model's predictions achieve the goals you have for the model? If the answer is yes to both of these questions, you have a useful model regardless of the value of R^2. Also note, a model with a large value of R^2 may not be a useful model by the two preceding criteria.

Adjusted R^2 (R_a^2)

Adding more independent variables to a regression model will always increase the R^2 value. R^2 will never decrease as variables are added because the SSE can never become smaller with the addition of independent variables, and the Total SS is always the same for a given set of responses. Since R^2 can be made larger by including a large number of independent variables, it is sometimes suggested that a modified measure be used that adjusts for the number of independent variables in the model. The **adjusted coefficient of determination** (denoted by R_a^2) adjusts R^2 by dividing each sum of squares by its associated degrees of freedom. Thus, R_a^2 is given by the following formula.

Adjusted R^2

The adjusted R^2 statistic takes into account the number of independent variables in the model by dividing each sum of squares by its associated degrees of freedom.

$$R_a^2 = 1 - \left(\frac{n-1}{n-k-1}\right)\frac{\text{SSE}}{\text{Total SS}}$$

where n is the number of observations and k is the number of independent variables in the model.

DEFINITION

For example, if one were to fit a simple linear regression model to *List Price* using only *Square Footage* as the independent variable, we would get $R^2 = 0.8799$ and $R_a^2 = 0.8761$. However, when we add the other two independent variables to the model, we have $R^2 = 0.9634$ and $R_a^2 = 0.9598$. In this case, the value of R^2 increased by 0.0835, indicating that adding the other variables to the model helped explain more variability in *List Price*. On the other hand, the value of R_a^2 increased by slightly more (0.0837). R_a^2 is commonly used as a method of comparison between multiple regression models when one is attempting to find the model that best fits the data. Unlike the R^2 value, the adjusted coefficient of determination may actually become smaller when another independent variable is added to the model. Thus, the adjusted R^2 value is most useful when comparing multiple regression models with different numbers of independent variables.

14.2 Exercises

Basic Concepts

1. What does R^2 represent?
2. What range of values can the coefficient of determination take on?
3. Can you think of a way a model might have a large R^2 and not be useful for prediction? Explain.
4. Explain the difference between R^2 and adjusted R^2.
5. Explain why the adjusted R^2 statistic is sometimes a better measure to use to evaluate the fit of a regression model.
6. Will there ever be a situation in which the adjusted R^2 statistic is greater than the R^2 statistic? Explain your answer.

Exercises

Data

stat.hawkeslearning.com
Discovering Statistics and Data, Fourth Edition > Data Sets > Mount Pleasant Real Estate Data.

7. Using the Mount Pleasant Real Estate data set, construct a multiple regression model relating housing prices (in thousands of dollars) to the number of bedrooms in the house, labeled *Bedrooms*, and the size of the lot on which the house was built, labeled *Acreage*.

a. Write the estimated regression equation.

b. Identify the values of SSR, SSE, and Total SS from the table.

c. What is the coefficient of determination for this model? Interpret this value in terms of the problem.

d. What is R_a^2 ? Interpret this value.

e. Compare the R^2 and R_a^2 values. Which value should be used to evaluate the fit of the multiple regression model? Explain why.

8. Add an additional variable, *Square Footage*, to the housing price model from Exercise 7.

a. Write the estimated regression equation.

b. What is R_a^2 for this model?

c. How does the adjusted R^2 value for this model compare to the adjusted R^2 value for the model in Exercise 7?

d. Do you think adding the additional independent variable, *Square Footage*, improved the model? Explain your answer.

9. The owner of a new pizzeria in town wants to study the relationship between weekly revenues and advertising expenditures. Both measures were recorded in thousands of dollars. The computer output for the simple linear regression model is given below.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.858179902
R Square	0.736472743
Adjusted R Square	0.692551534
Standard Error	1.058296197
Observations	8

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	18.78005496	18.78005496	16.76804334	0.006394067
Residual	6	6.719945042	1.11999084		
Total	7	25.5			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	74.69887795	7.104358625	10.51451396	4.34789E-05	57.31513863	92.08261726
Advertising Expenditures	1.854820243	0.452960815	4.094880138	0.006394067	0.746465058	2.963175428

a. Write the estimated regression equation.

b. What is the coefficient of determination for this model? Interpret this value.

c. What is the value of the adjusted R^2 statistic? Is this statistic useful for the pizzeria owner as he studies this model? Explain.

d. Do you believe this model is useful in explaining revenues based on advertising expenditures? Explain your answer.

10. How could the restaurant owner improve this model? Are there other independent variables that he should consider including? The owner of the pizzeria discussed in Exercise 9 wishes to build on the model relating revenues to advertising expenditures by breaking the advertising expenditures into three categories: television advertising, newspaper advertising, and direct mail advertising.

a. Write the new regression model in terms of television, newspaper, and mail expenditures. Assume the coefficients have not yet been estimated.

b. Consider the following summary output for the new model. Write the estimated multiple regression equation.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.967040091
R Square	0.935166537
Adjusted R Square	0.88654144
Standard Error	0.64289449
Observations	8

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	23.8467467	7.948915566	19.23217829	0.007708883
Residual	4	1.653253302	0.413313326		
Total	7	25.5			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	73.93199827	4.523870838	16.34264127	8.20538E-05	61.37171922	86.49227731
Television	2.383047934	0.318133378	7.490719616	0.001698799	1.499768074	3.266327793
Newspaper	1.454439994	0.355820285	4.087569076	0.015004989	0.466524505	2.442355483
Mail	1.815990841	0.276487962	6.568064755	0.002780349	1.048337191	2.58364449

c. Interpret the coefficient for television advertising expenditures. Remember that revenues and expenditures are in thousands of dollars.

d. What is the adjusted coefficient of determination? Interpret this value.

e. How does the coefficient of determination of this model compare to the coefficient of determination for the simple linear regression model in Exercise 9? Does this appear to be a more useful model? Explain.

f. What is the value of the R^2 statistic for this model? Should we use the R^2 value or the adjusted R^2 value when evaluating the usefulness of this model? Explain why.