Interpreting Regression Results

From the Problem Solving with Advanced Analytics Course as part of the Predictive Analytics for Business Nanodegree Program from Udacity

Getting Started

To the right you see the results of a linear regression. Results are similarly reported by almost any regression tool. Do not be intimidated by the complexity of the presentation of the results; we'll walk you through the most important values and how to interpret and apply them. For purposes of this exercise, the focus will be on three values that are of particular importance: the coefficient estimates, the p-values, and the R-squared.

Report

Report for Linear Model X

Basic Summary

Call:
lm(formula = Average.Number.of.Tickets ~ Number.of.Employees + Value.of.Contract + Industry, data = the.data)

Residuals:

Min	1Q	Median	3Q	Max
-9.711	-3.149	-0.685	6.287	9.974

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-1.845e+01	4.7645757	-3.872	0.0015	**
Number.of.Employees	1.116e-01	0.0162071	6.886	1e-05	***
Value.of.Contract	4.858e-04	0.0001107	4.387	0.00053	***
IndustryRetail	-8.725e+00	3.7557388	-2.323	0.03463	*
IndustryServices	1.249e+01	3.7000971	3.377	0.00415	**

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.3709 on 15 degrees of freedom
Multiple R-squared: 0.9651, Adjusted R-Squared: 0.9558
F-statistic: 103.7 on 4 and 15 DF, p-value: 9.691e-11

Coefficient Estimates

Coefficients:

	Estimate
(Intercept)	-1.845e+01
Number.of.Employees	1.116e-01
Value.of.Contract	4.858e-04
IndustryRetail	-8.725e+00
IndustryServices	1.249e+01

Remember the regression equation? Y = B0 + B1X1 + B2X2…? These coefficients are the estimates of the B's. They represent the relationship between each predictor variable and the target variable. For example, the coefficient on the **number.of.employees** means that each additional employee will lead to roughly an additional 0.1 tickets, *holding all other variables constant*. A simpler way to think about this is that we can expect about 1 ticket for every additional 10 employees.

P Value

Coefficients:

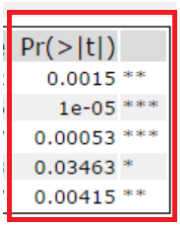
	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-1.845e+01	4.7645757	-3.872	0.0015	**
Number.of.Employees	1.116e-01	0.0162071	6.886	1e-05	***
Value.of.Contract	4.858e-04	0.0001107	4.387	0.00053	***
IndustryRetail	-8.725e+00	3.7557388	-2.323	0.03463	*
IndustryServices	1.249e+01	3.7000971	3.377	0.00415	**

The **p value** is the probability that the observed results (the coefficient estimate) occurred by chance, and that there is no actual relationship between the predictor and target variable. In other words, the p-value is the probability that the coefficient is zero. The lower the p-value the higher the probability that a relationship exists between the predictor and target variable. If the p-value is high, we should not rely

on the coefficient estimate. When a predictor variable has a p-value below 0.05, the relationship between it and the target variable is considered to be statistically significant.

Statistical Significance - "Statistical significance is a result that is not likely to occur randomly, but rather is likely to be attributable to a specific cause." - Investopedia

Apart from the p-value, the stars to the right indicate statistical significance as well. More *'s means more significance. In our example, we see that our predictor variables are significant, with the number of employees and the value of the contract being the most significant. Generally, we'll want to remove variables from the model that are not statistically significant predictors of the target variable.



Pr(>|t|)	
0.0015	**
1e-05	***
0.00053	***
0.03463	*
0.00415	**

R-Squared

In the course example, the R-squared value is 0.9651, and the adjusted R-squared value is 0.9558. In a real life problem, we might run the model with different predictor variables, or see if we had additional information to add to the model.

Coefficients:

	Estimate	Std. Error
(Intercept)	-1.845e+01	4.7645757
Number.of.Employees	1.116e-01	0.0162071
Value.of.Contract	4.858e-04	0.0001107
IndustryRetail	-8.725e+00	3.7557388
IndustryServices	1.249e+01	3.7000971

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.3709 on 15 degrees of freedom
Multiple R-squared: 0.9651, Adjusted R-Squared: 0.9558
F-statistic: 103.7 on 4 and 15 DF, p-value: 9.691e-11

Remember, R-squared ranges from 0 to 1 and represents the amount of variation in the target variable explained by the variation in the predictor variables. The higher the r-squared, the higher the explanatory power of the model.