Assignment 4: Statistical Inference in Linear Regression

Andrew G. Dunn[1]

[1]andrew.g.dunn@u.northwestern.edu

Andrew G. Dunn, Northwestern University Predictive Analytics Program

Prepared for PREDICT-410: Regression & Multivariate Analysis.

Formatted using markdown, pandoc, and LaTeX. References managed using Bibtex, and pandoc-citeproc.

Model 1

Consider the following diagnostic output for a regression model:

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	2126.00904	531.50226		< 0.0001
Error	67	630.35953	9.40835		
Corrected Total	71	2756.36857			

Table 1: Analysis of Variance

Source	
Root MSE	3.06730
Dependent Mean	37.26901
Coeff Var	8.23017
R-Square	
Adj R-Square	

Table 2: Estimator Performance

Variable	DF	Parameter Estimate	Standard Error	t-value	$Pr > \|t\|$
Intercept	1	11.33027	1.99409	5.68	< 0.0001
X1	1	2.18604	0.41043		< 0.0001
X2	1	8.27430	2.33906	3.54	0.0007
X3	1	0.49182	0.26473	1.86	0.0676
X4	1	-0.49356	2.29431	-0.22	0.8303

Table 3: Parameter Estimates

Number in Model	C(p)	R-Square	AIC	BIC	Variables in Model
4	5.000	0.7713	166.2129	168.9481	X1 X2 X3 X4

Table 4: Model Quality

How many observations are within the sample data?

There are 72 observations within the sample data. We look at the Corrected Total and know that 1 degree of freedom is devoted towards estimating the intercept parameter.

Write out the null and alternate hypotheses for the t-test for β_1

We find in [1] on page 24, a discussion of testing the significance of regression. The hypothesis for β_1 is meant to be an implication of whether there is a linear relationship between the dependent and independent variable.

$$H_0 : \beta_1 = 0 \text{ versus } H_1 : \beta_1 \neq 0$$

Compute the t-statistic for β_1

We find in [1] on page 25 the equation for the t-statistic as:

$$t_0 = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)}$$

where t_i has degrees of freedom equal to the sample size minus the number of model parameters, i.e. $df = n - dim(Model)$.

With the table output of this model we have a fairly easy job of computing the t-statistic:

$$t_0 = \frac{2.18604}{0.41043} = 5.32621$$

Compute the R-Square value for Model 1

We consider:

- The Total Sum of Squares is the total variation in the sample
- The Regression Sum of Squares is the variation in the sample that has been explained by the regression model
- The Error Sum of Squares is the variation in the sample that cannot be explained

SST	$\sum_i^n (Y_i - \bar{Y})^2$	Total Sum of Squares
SSR	$\sum_i^n (\hat{Y}_i - \bar{Y})^2$	Regression Sum of Squares
SSE	$\sum_i^n (Y_i - \hat{Y})^2$	Error Sum of Squares

Where the Coefficient of Determination - R-Squared is:

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

With the table output of this model we have a fairly easy job of computing the R-square value:

$$R^2 = 1 - \frac{630.35953}{2756.36857} = 0.7713$$

Compute the Adjusted R-Square value for Model 1

We consider:

$$R^2_{ADJ} = 1 - \frac{\frac{SSE}{(n-k-1)}}{\frac{SST}{(n-1)}} = 1 - \frac{\frac{SSE}{(n-p)}}{\frac{SST}{n-1}}$$

The standard regression notation uses k for the number of predictor variables included in the regression model and p for the total number of parameters in the model. When the model includes an intercept term, then $p = k + 1$. When the model does not include an intercept term, then $p = k$.

With the table output of this model we have a fairly easy job of computing the Adjusted R-Square value:

$$R^2_{ADJ} = 1 - \frac{\frac{630.35953}{(72-5)}}{\frac{2756.36857}{72-1}} = 0.75765$$

Write out the null and alternative hypotheses for the Overall F-test

We find in [1] on page 83 a discussion of testing for significance of regression. This is a test to determine if there is a linear relationship between the dependent variable and any of the regressors variables.

$$H_0 : \beta_1 = \ldots = \beta_k = 0 \text{ versus } H_1 : \beta_i \neq 0$$

for some $i \in 1, \ldots, k$

Compute the F-statistic for the Overall F-test

We find in [1] on page 85 the equation for the f-statistic as:

$$F_0 = \frac{\frac{SSR}{k}}{\frac{SSE}{(n-p)}}$$

which has a F-distribution with $(k, n-p)$ degrees-of-freedom for a regression model with k predictor variables and p total parameters. When the regression model includes and intercept, then $p = k + 1$. If the regression model does not include an intercept, then $p = k$

We compute the F-statistic by:

$$F_0 = \frac{\frac{2126.00904}{4}}{\frac{630.35953}{(72-5)}} = 56.4926$$

Model 2

Consider the following diagnostic output for a regression model:

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	2183.75946	363.95991	41.32	< 0.0001
Error	65	572.60911	8.80937		
Corrected Total	71	2756.63857			

Table 6: Analysis of Variance

Source	
Root MSE	2.96806
Dependent Mean	37.26901
Coeff Var	7.96388
R-Square	0.7923
Adj R-Square	0.7731

Table 7: Estimator Performance

Variable	DF	Parameter Estimate	Standard Error	t-value	$Pr > \|t\|$
Intercept	1	14.39017	2.89157	4.98	< 0.0001
X1	1	1.97132	0.43653	4.52	< 0.0001
X2	1	9.13895	2.30071	3.97	0.0002
X3	1	0.56485	0.26266	2.15	0.0352
X4	1	0.33371	2.42131	0.14	0.8908
X5	1	1.90698	0.76459	2.49	0.0152
X6	1	-1.04330	0.64759	-1.61	0.1120

Table 8: Parameter Estimates

Number in Model	C(p)	R-Square	AIC	BIC	Variables in Model
6	7.000	0.7923	163.2947	166.7792	X1 X2 X3 X4 X5 X6

Table 9: Model Quality

Consider Model 1 and Model 2 as a pair of models. Does Model 1 nest Model 2, or does Model 2 nest Model 1?

If we consider the two models as:

$$Y_{\text{Model 1}} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$$

$$Y_{\text{Model 2}} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6$$

We notice that the predictor variables in Model 1 are a subset of the predictor variables in Model 2. We can say that Model 1 *nests* Model 1, or that Model 1 is *nested* by Model 2.

For discussion of nested models, we use the terminology *full model* (FM) and *reduced model* (RM). In this terminology RM is a subset of FM, alternatively $RM \subset FM$.

Write out the null and alternate hypotheses for a nested F-test using Model 1 and Model 2

We will write a null hypothesis to test multiple predictor variables:

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_6 = 0 \text{ versus } H_1 : \beta_i \neq 0$$

for some $i \in 1..6$.

Compute the F-statistic for a nested F-test using Model 1 and Model 2

The test Statistic for the Nested F-Test:

$$F_0 = \frac{\frac{[SSE(RM) - SSE(FM)]}{(dim(FM) - dim(RM))}}{\frac{SSE(FM)}{[n - dim(FM)]}}$$

We compute the nested F-test statistic as:

$$F_0 = \frac{\frac{[630.35953 - 572.60911)]}{(7-5)}}{\frac{572.60911}{[72-7]}} = 3.2777$$

Additional Consideration

Compute the AIC values for both Model 1 and Model 2

In the case of ordinary least squares regression, the Akaike Information Criterion is [1] page 366:

$$AIC = n \times ln(\frac{SSE}{n}) + 2p$$

For Model 1:

$$AIC = 72 \times ln(\frac{630.35953}{72}) + 2 \times 5 = 166.2129$$

For Model 2:

$$AIC = 72 \times ln(\frac{572.60911}{72}) + 2 \times 7 = 163.2946$$

Compute the BIC values for both Model 1 and Model 2

In the case of ordinary least squares regression, the Bayesian Analogues is [1] page 366:

$$BIC = n \times ln(\frac{SSE}{n}) + p \times ln(n)$$

For Model 1:

$$BIC = 72 \times ln(\frac{630.35953}{72}) + 5 \times ln(72) = 177.5962$$

For Model 2: [1]

$$BIC = 72 \times ln(\frac{572.60911}{72}) + 7 \times ln(72) = 179.2313$$

Compute the Mallow's C_p values for both Model 1 and Model 2

Mallow's C_p Statistic [1] on page 334:

$$C_p = \frac{SSR_p}{MSE} - n + 2 \times p$$

For Model 1:

$$C_p = \frac{630.35953}{9.40835} - 72 + 2 \times 5 = 5.0000$$

For Model 2:

$$C_p = \frac{572.60911}{8.80937} - 72 + 2 \times 7 = 7.0000$$

Verify the t-statistics for the remaining coefficients in Model 1

We find in [1] on page 25 the equation for the t-statistic as:

$$t_0 = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)}$$

Parameter: Intercept

$$t_{intercept} = \frac{11.33027}{1.99409} = 5.6819$$

Parameter: X2

$$t_{X2} = \frac{8.27430}{2.33906} = 3.5374$$

Parameter: X3

$$t_{X3} = \frac{0.49182}{0.26473} = 1.8578$$

Parameter: X4

$$t_{X4} = \frac{-0.49356}{2.29431} = -0.2151$$

References

[1]J. Fox, “Regression diagnostics: An introduction (quantitative applications in the social sciences).” Sage Publications, Inc., Thousand Oaks, CA, 1991.