Final

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.

Inference Statistical Assumptions

When validating the in-sample fit of a linear regression model, what are two assumptions that must be validated using the model residuals, and how does one validate each of these assumptions?

Response:

By 'in-sample' I assume this inquiry to be about building models for inference.

Statistical inference is focused on a set of formal hypotheses, denoted by H_0 for the *null hypothesis* and H_1 for the *alternate hypothesis*, and a test statistic with a known sampling distribution. A test statistic will have a specified distribution, e.g. the t-statistic for an OLS regression parameter has a t-distribution with the degrees-of-freedom equal to $n - p$ where p is the number of model parameters for the dimension of the model. [1]

When we fit a statistical model, we have underlying assumptions about the probabilistic structures for that model. All of our statistical inference is derived from those probabilistic assumptions. Hence, if our estimated model, which is dependent upon the sample data, does not conform to these probabilistic assumptions, then our inference will be incorrect.

The two ways to validate a model in sample are examination of the R-Square and analysis of the residuals. For these investigations we assume:

- The relationship between the response and the regressors is linear, at least approximately.
- The errors are normally distributed (and uncorrelated)

For the Fitted Regression Model

- Write down the null and alternate hypotheses for the t-test for X1.
- Compute the t-statistic associated with the regression coefficient for X1.
- Write down the null and alternate hypotheses for the Overall F-test.
- Compute the F-statistic for the Overall F-test.
- Compute the R-Squared value.
- Compute the Adjusted R-Squared value.
- Compute the AIC value.
- Compute the BIC value.

Table 1: Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	677.74649	135.54930		
Error	18	153.76309	8.54239		
Corrected Total	23	831.50958			

Table 2: Estimator Performance

Source	
Root MSE	2.92274
Dependent Mean	34.62917
Coeff Var	8.44010
R-Square	
Adj R-Square	

[1] Dr. Chad Bhatti, Statistical Inference Versus Predictive Modeling in OLS Regression

Table 3: Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t-value	$Pr > \|t\|$
Intercept	1	9.55767	3.30844	2.89	0.0098
X1	1	2.14462	0.73434		
X2	1	7.11226	3.92381	1.81	0.0866
X3	1	0.41847	0.44652	0.94	0.3611
X4	1	-0.80007	3.82619	-0.21	0.8367
X5	1	1.23453	1.17283	1.05	0.3064

Response:

Write Down the Null and Alternative Hypothesis for the T-test for X1

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

Compute the T-statistic associated with the regression coefficient for X1

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

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:

$$t_0 = \frac{2.14462}{0.73434} = 2.920472806$$

Write down the null and alternate hypotheses for the Overall F-test.

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

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 $p = k$

Compute the F-statistic for the Overall F-test.

$$F_0 = \frac{\frac{677.74649}{5}}{\frac{153.76309}{(24-6)}} = 15.867835148$$

Compute the R-Squared value.

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:

$$R^2 = 1 - \frac{153.76309}{831.50958} = 0.815079593$$

Compute the Adjusted R-Squared value.

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:

$$R^2_{ADJ} = 1 - \frac{\frac{153.76309}{(24-6)}}{\frac{831.50958}{24-1}} = 0.763712813$$

Compute the AIC value.

In the case of ordinary least squares regression, the Akaike Information Criterion is:

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

$$AIC = 24 \times ln(\frac{153.76309}{24}) + 2 \times 6 = 56.576621062$$

Compute the BIC value.

In the case of ordinary least squares regression, the Bayesian Analogues is:

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

$$BIC = 24 \times ln(\frac{153.76309}{24}) + 6 \times ln(24) = 63.644944044$$

Constructing a Model

Suppose that you were asked to build a linear regression model for a continuous response variable denoted by Y using only the categorical predictor variable gender. In your sample data set gender takes three values: M for male, F for Female, and U for Unknown (Missing).

How would you specify this linear regression model?

We would choose to model this as an indicator variable. We have, due to M, F, and U, a three level variable. We would specify the model as:

$$Y = \beta_0 + \beta_1 \times \text{male} + \beta_2 \times \text{female} + \epsilon$$

In this situation, even though we have a three level indicator variable, we will model the unknown as our base category, against which the others are assessed in order to avoid the dummy variable trap.

How would you test to see if there is a 'gender effect' on Y?

As we're modeling an indicator variable, we will evaluate/test for gender effect by interpreting the model for unit change in Y due to unit change in each of the three variations of the model fit:

$$Y = \beta_0 + \beta_1 \times \text{male} + \epsilon, \text{where male=1}$$
$$Y = \beta_0 + \beta_2 \times \text{female} + \epsilon, \text{where female=1}$$
$$Y = \beta_0 + \epsilon$$

By interpreting these three, or graphing them, we can consider the effect of gender on Y.

Suppose that in addition to gender, your model had to include a continuous predictor variable (X1). Including this variable, how do you test for a 'gender effect' on Y?

We would add the continuous variable X1 to our model as follows:

$$Y = \beta_0 + \beta_1 \times \text{X1} + \beta_2 \times \text{male} + \beta_3 \times \text{female} + \epsilon$$

To text for effect, we would evaluate in the same fashion as the fitted models as above:

$$Y = \beta_0 + \beta_1 \times \text{X1} + \beta_2 \times \text{male} + \epsilon, \text{where male=1}$$
$$Y = \beta_0 + \beta_1 \times \text{X1} + \beta_3 \times \text{female} + \epsilon, \text{where female=1}$$
$$Y = \beta_0 + \beta_1 \times \text{X1} + \epsilon$$