Assignment 5: Variable Selection and Validation

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.

Dummy Coding of Categorical Variables

Examine a Categorical Variable

We'll choose to examine the categorical variable OverallQual as it was the variable that offered the best performance when doing a simple linear regression to SalePrice. We realize this variable is a Lickert scale [1], 10 being Very Excellent. and 1 being Very Poor. To investigate, we perform a sort and means procedure, obtaining:

N	OverallQual	SalePrice Mean
4	1	48725
13	2	52325.31
40	3	83185.98
226	4	106485.10
825	5	134752.52
732	6	162130.32
602	7	205025.76
350	8	270913.59
107	9	368336.77
31	10	450217.32

Table 1: Sorted Means procedure with OveralQual

We run a simple linear regression model:

$$\text{SalePrice} = \beta_0 + \beta_1 \text{OverallQual} + \epsilon$$

And we get the parameter estimation and diagnostic information:

$$\text{SalePrice} = 45251 \times \text{OverallQual} - 95004$$

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-95004	3933.82223	-24.15	<0.0001
OverallQual	1	45251	628.80511	71.96	<0.0001

Table 2: Model Parameter Estimates for SalePrice = OverallQual

Source	
Root MSE	48019
R-Square	0.6388
Adj R-Square	0.6387

Source	
F Value	5178.75

Table 3: Model Estimator Performance for SalePrice = OverallQual

We'll compute the pass-through points for the Model SalePrice $= \beta_0 + \beta_1 \text{OverallQual} + \epsilon$.

$$\text{SalePrice} = 45251 \times 1 - 95004 = -49753$$
$$\text{SalePrice} = 45251 \times 2 - 95004 = -4502$$
$$\text{SalePrice} = 45251 \times 3 - 95004 = 40749$$
$$\text{SalePrice} = 45251 \times 4 - 95004 = 86000$$
$$\text{SalePrice} = 45251 \times 5 - 95004 = 131251$$
$$\text{SalePrice} = 45251 \times 6 - 95004 = 176502$$
$$\text{SalePrice} = 45251 \times 7 - 95004 = 221753$$
$$\text{SalePrice} = 45251 \times 8 - 95004 = 267004$$
$$\text{SalePrice} = 45251 \times 9 - 95004 = 312255$$
$$\text{SalePrice} = 45251 \times 10 - 95004 = 357506$$

The predicted model appears to only be relatively close at points 3, 5, 6, 8.

Indicator Coding a Categorical Variable

OverallQual is an interesting parameter, because it is a 10-way Lickert some would choose to incorporate the parameter into a model as a continuous variable. Above we model it as a continuous parameter, here will we dummy code it and model it as an indicator variable. Although OverallQual ranges from 1-10, we can use nine variables to investigate, this is simpler to consider as a table:

OverallQual	$x1$	$x2$	$x3$	$x4$	$x5$	$x6$	$x7$	$x8$	$x9$
1	1	0	0	0	0	0	0	0	0
2	0	1	0	0	0	0	0	0	0
3	0	0	1	0	0	0	0	0	0
4	0	0	0	1	0	0	0	0	0
5	0	0	0	0	1	0	0	0	0
6	0	0	0	0	0	1	0	0	0
7	0	0	0	0	0	0	1	0	0
8	0	0	0	0	0	0	0	1	0
9	0	0	0	0	0	0	0	0	1
10	0	0	0	0	0	0	0	0	0

Table 4: Modeling 10-Way Lickert OverallQual with 9 Indicator Variables

We will use the data procedure to dummy code OverallQual as an indicator variable. To examine our progress we evaluate a proc freq of the OverallQual:

```
data ames_dummy_oc;
  set ames;
  keep SalePrice OverallQual oc_1 oc_2 oc_3 oc_4 oc_5 oc_6 oc_7 oc_8 oc_9 oc_10;
  if OverallQual in (1 2 3 4 5 6 7 8 9 10) then do;
    oc_1 = (OverallQual eq 1);
    oc_2 = (OverallQual eq 2);
    oc_3 = (OverallQual eq 3);
    oc_4 = (OverallQual eq 4);
    oc_5 = (OverallQual eq 5);
    oc_6 = (OverallQual eq 6);
    oc_7 = (OverallQual eq 7);
    oc_8 = (OverallQual eq 8);
    oc_9 = (OverallQual eq 9);
    oc_10 = (OverallQual eq 10);
    end;

proc freq data=ames_dummy_oc;
  tables OverallQual oc_1 oc_2 oc_3 oc_4 oc_5 oc_6 oc_7 oc_8 oc_9 oc_10;
```

OverallQual	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	4	0.14	4	0.14
2	13	0.44	17	0.58
3	40	1.37	57	1.95
4	226	7.71	283	9.66
5	825	28.16	1108	37.82
6	732	24.98	1840	62.80
7	602	20.55	2442	83.34
8	350	11.95	2792	95.29
9	107	3.65	2899	98.94
10	31	1.06	2930	100.00

Table 5: Frequency OverallQual

For brevity we examine the oc_1 and oc_2 variables:

oc_1	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	2926	99.86	2926	99.86
1	4	0.14	2930	100.00

Table 6: Frequency oc_1, Indicator Variable for OverallQual 1

oc_2	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	2917	99.56	2917	99.56
1	13	0.44	2930	100.00

Table 7: Frequency oc_2, Indicator Variable for OverallQual 2

We notice that the oc_1 and oc_2 variables are properly coded to match up with the OverallQual frequency table, with oc_1 have 1 coded 4 times and oc_2 having 1 coded 13 times respectively.

We now build a model, but we hold oc_10 to be the basis of interpretation:

$$\text{SalePrice} = \beta_0 + \beta_1 \text{oc\_1} + \beta_2 \text{oc\_2} + \beta_3 \text{oc\_3} + \beta_4 \text{oc\_4} + \beta_5 \text{oc\_5} + \beta_6 \text{oc\_6} + \beta_7 \text{oc\_7} + \beta_8 \text{oc\_8} + \beta_9 \text{oc\_9} + \epsilon$$

```
proc reg data=ames_dummy_ec;
  model saleprice = oc_1 oc_2 oc_3 oc_4 oc_5 oc_6 oc_7 oc_8 oc_9;
```

Resulting in the parameter estimations and model diagnostics:

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	450217	7841.25572	57.42	<.0001
oc_1	1	-401492	23195	-17.31	<.0001
oc_2	1	-397892	14426	-27.58	<.0001
oc_3	1	-367031	10447	-35.13	<.0001
oc_4	1	-343732	8361.76503	-41.11	<.0001
oc_5	1	-315465	7987.21777	-39.50	<.0001
oc_6	1	-288087	8005.57159	-35.99	<.0001
oc_7	1	-245192	8040.61424	-30.49	<.0001
oc_8	1	-179304	8181.14487	-21.92	<.0001
oc_9	1	-81881	8904.98663	-9.19	<.0001

Table 8: Model Parameter Estimates for SalePrice = OverallQual Indicator Variables

Source	
Root MSE	43658
R-Square	0.7023
Adj R-Square	0.7013
F Value	765.22

Table 9: Model Estimator Performance for SalePrice = OverallQual Indicator Variables

We notice that our Adj. R-Square value has increased from 0.6388 to 0.7013.

The resulting Model is:

$$\begin{aligned}\text{SalePrice} = 450217 - 401492 \times \text{oc\_1} - 397892 \times \text{oc\_2} \\ -367031 \times \text{oc\_3} - 343732 \times \text{oc\_4} - 315465 \times \text{oc\_5} \\ -288087 \times \text{oc\_6} - 245192 \times \text{oc\_7} - 179304 \times \text{oc\_8} \\ -81881 \times \text{oc\_9} + \epsilon\end{aligned}$$

If the house is of OverallQual 1, then the above model becomes:

$$\text{SalePrice} = 450217 - 401492$$

That is to say, if the OverallQual is 1, then the SalePrice in this model is 48725. Looking back at our sorted mean we see the SalePrice for OverallQual of 1 had a mean of 48725.

If the house is of OverallQual 2, then the above model becomes:

$$\text{SalePrice} = 450217 - 397892$$

That is to say, if the OverallQual is 2, then the SalePrice in this model is 52325 . Looking back at our sorted mean we see the SalePrice for OverallQual of 2 had a mean of 52325.31.

If the house is of OverallQual 3, then the above model becomes:

$$\text{SalePrice} = 450217 - 367031$$

That is to say, if the OverallQual is 3, then the SalePrice in this model is 83186. Looking back at our sorted mean we see the SalePrice for OverallQual of 3 had a mean of 83185.98.

If the house is of OverallQual 4, then the above model becomes:

$$\text{SalePrice} = 450217 - 343732$$

That is to say, if the OverallQual is 4, then the SalePrice in this model is 106485. Looking back at our sorted mean we see the SalePrice for OverallQual of 4 had a mean of 106485.10.

If the house is of OverallQual 5, then the above model becomes:

$$\text{SalePrice} = 450217 - 315465$$

That is to say, if the OverallQual is 5, then the SalePrice in this model is 134752. Looking back at our sorted mean we see the SalePrice for OverallQual of 5 had a mean of 134752.52.

If the house is of OverallQual 6, then the above model becomes:

$$\text{SalePrice} = 450217 - 288087$$

That is to say, if the OverallQual is 6, then the SalePrice in this model is 162130. Looking back at our sorted mean we see the SalePrice for OverallQual of 6 had a mean of 162130.32.

If the house is of OverallQual 7, then the above model becomes:

$$\text{SalePrice} = 450217 - 245192$$

That is to say, if the OverallQual is 7, then the SalePrice in this model is 205025. Looking back at our sorted mean we see the SalePrice for OverallQual of 7 had a mean of 205025.76.

If the house is of OverallQual 8, then the above model becomes:

$$\text{SalePrice} = 450217 - 179304$$

That is to say, if the OverallQual is 8, then the SalePrice in this model is 270913. Looking back at our sorted mean we see the SalePrice for OverallQual of 8 had a mean of 270913.59.

If the house is of OverallQual 9, then the above model becomes:

$$\text{SalePrice} = 450217 - 81881$$

That is to say, if the OverallQual is 1, then the SalePrice in this model is 368336. Looking back at our sorted mean we see the SalePrice for OverallQual of 9 had a mean of 368336.77.

If the house is of OverallQual 10, then the above model becomes:

$$\text{SalePrice} = 450217$$

That is to say, if the OverallQual is 10, then the SalePrice in this model is 450217. Looking back at our sorted mean we see the SalePrice for OverallQual of 10 had a mean of 450217.32.

Overall, due to the expression of the concept in the beginning of this section, we should have expected these results from the beginning. The model passes through the means of each OverallQual step.

Dummy Code Hypothesis Testing

$$H_0 : \beta_{1..9} = 0 \text{ versus } H_1 : \beta_{1..9} \neq 0$$

Report hypothesis test for each beta Discuss results for each test

Dummy Code another Categorical Variable

Data step, one dummy code for each category

Automated Variable Selection

Obtaining the "Best" Model

I'm going to pre-select continuous variables based upon our previous investigation.

proc reg data=part2 outest=rsqest; model saleprice = ... / selection=adjrsq aic bic;

proc print data=rsqest;

proc reg data=part2; model saleprice = ... / selection=cp aic bic;

adjusted R-Squared, Mallow's Cp, AIC, Forward, Backward and Stepwise, in six separate modeling steps

Report summary tables for each technique

Did the different techniques select the same model?

Discuss any observations

If Dummy Variable Inclusion, Must Include all Dummy Variable for that Parameter

Select one of the six models that incorporated a dummy variable, refit the model after adding in the other dummy variables from that parameter

Report the model, interpret the coefficients.

Discuss any observations

Validation Framework

Create a Training and Test Data set

data temp; set mydata.ames_housing_data; * generate a uniform(0,1) random variable with seed set to 123; u = uniform(123); if (u < 0.70) then train = 1; else train = 0; if (train=1) then train_response=SalePrice; else train_response=.; run;

Obtaining the "Best" Model

With train_response re runn: adjusted R-Squared, Mallow's Cp, AIC, Forward, Backward and Stepwise, in six separate modeling steps

Report summary tables for each technique

Did the different techniques select the same model?

Discuss any observations, specifically how do these models compare to the models that were run against salePrice

Comparing Models with Training and Test Data

Identify each of the 6 models in a table: model_{AdjRSqr}... For each, obtain Adjusted R-Square, BIC, MSE, MAE on training set (proc reg)

Next, use a new SAS data step and a PROC MEANS statement to calculate the average squared error (MSE) and the average absolute error (MAE) for the test sample and the validation sample. Which model fits the best based on these statistics? Did the model that fit best in-sample predict the best out-of-sample?

proc reg data=part8; model train_response = extcond_ta extcond_fa/ selection=forward; output out=part9 predicted=yhat;

data part9b; set part9; mae = abs(yhat - train_response);

proc means data=part9b; var mae; title 'MAE Calculation';

Operational Validation

We have validated these models in the statistical sense, but what about the business sense? Do MSE or MAE easily translate to the development of a business policy? To do this, you will need to create a new datastep after saving the predicted values from the model. Define the variable "Prediction_Grade" (define the variable using format $7.). Let's consider the predicted value to be "Grade 1" if it is within ten percent of the actual value, "Grade 2" if it is within fifteen percent of the actual value, and "Grade 3" otherwise. How accurate are the models under this definition of predictive accuracy? Use PROC FREQ to provide a table of the model's operational accuracy.

proc reg data=part8; model train_response = extcond_ta extcnd_fa/ selection=forward; output out=part10 predicted=yhat;

proc print data=part10 (obs=10);

data part10b; set part10; if train_response = . then delete;

length prediction_grade $7.;

pct_diff = abs((yhat - train_response) / train_response);

```
if pct_diff LE 0.10 then prediction_grade = 'Grade 1';
  else if pct_diff GT 0.10 and pct_diff LE 0.15 then prediction_grade = 'Grade 2';
  else prediction_grade = 'Grade 3';
```

proc print data=part10b (obs=10);

proc freq data=part10b; tables prediction_grade;

Best Model, Revisited with all Dummy coded Variables

If a dummy coded variable included, add the rest from that parameter

Report the model, this is the final model, report happily

Conclusion / Reflection

What were the challenges within this data set?

What are the recommendations for improving prediction accuracy

Procedures

```
title 'Assignment 5';
libname mydata '/scs/crb519/PREDICT_410/SAS_Data/' access=readonly;

* create a temporary variable (data source is read only);
data ames;
  set mydata.ames_housing_data;

ods graphics on;

proc sort data=ames;
  by OverallQual;

proc means data=ames;
  var SalePrice;
  by OverallQual;

proc reg data=ames;
  model SalePrice = OverallQual;

data ames_dummy_oc;
  set ames;
  keep SalePrice OverallQual oc_1 oc_2 oc_3 oc_4 oc_5 oc_6 oc_7 oc_8 oc_9 oc_10;
  if OverallQual in (1 2 3 4 5 6 7 8 9 10) then do;
    oc_1 = (OverallQual eq 1);
    oc_2 = (OverallQual eq 2);
    oc_3 = (OverallQual eq 3);
    oc_4 = (OverallQual eq 4);
    oc_5 = (OverallQual eq 5);
```

```
    oc_6 = (OverallQual eq 6);
    oc_7 = (OverallQual eq 7);
    oc_8 = (OverallQual eq 8);
    oc_9 = (OverallQual eq 9);
    oc_10 = (OverallQual eq 10);
    end;

proc freq data=ames_dummy_oc;
  tables OverallQual oc_1 oc_2 oc_3 oc_4 oc_5 oc_6 oc_7 oc_8 oc_9 oc_10;

proc reg data=ames_dummy_oc;
  model saleprice = oc_1 oc_2 oc_3 oc_4 oc_5 oc_6 oc_7 oc_8 oc_9;

run;
```

References

[1]Wikipedia, “Likert scale — wikipedia, the free encyclopedia.” 2015 [Online]. Available: http://en.wikipedia.org/w/index.php?title=Likert_scale&oldid=653173542