MATH 424: Homework Chapter 7: Some Regression Pitfalls

Due on Tuesday, November 7, 2017

Kafai 11:10am

Jonathan Dombrowski

Contents

Q 6

Characteristics of sea ice melt ponds. Surface albedo is defined as the ratio of solar energy directed upward from a surface over energy inci- dent upon the surface. Surface albedo is a critical climatological parameter of sea ice. The National Snow and Ice Data Center (NSIDC) collects data on the albedo, depth, and physical characteristics of ice melt ponds in the Canadian Arctic, includ- ing ice type (classified as first-year ice, multiyear ice, or landfast ice). Data for 504 ice melt ponds located in the Barrow Strait in the Canadian Arctic are saved in the PONDICE file. Environmental engineers want to model the broadband surface albedo level, y, of the ice as a function of pond depth, x 1 (meters), and ice type, represented by the dummy variables x 2 = 1 if first-year ice, 0 if not and x 3 = 1 if multiyear ice, 0 if not. Ultimately, the engineers will use the model to predict the surface albedo level of an ice melt pond. Access the data in the PONDICE file and identify the experi- mental region for the engineers. What advice do you give them about the use of the prediction equation?

a

The problem was a bit unclear when associating which type of ice went with each level of x2. Based on the assumption that the number in x2 corresponds with the order the variables were mentioned in the text, the experimental regions in the data are as follows:
1st year icetype = 1: depth between (0.02, 0.36)
Landfast icetype = 2: depth between (0.07, 0.64)
Multiyear icetype = 3: depth between (0.00, 0.86)

b

The advice that I can give is that when using the prediction equation $y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3$ is only to use it to predict values within the intervals listed above; any values outside the corresonding are not accurate. The VIF scores do not show anything to be concerned about, so as long as the prediction values stay within the intervals, there should not be any complications.

The VIF scores are appended

```
variables               VIF
PONDICE.broadbandalb    1.143472
PONDICE.depth           1.344388
PONDICE.icetype         NA
```

Q 19

Cooling method for gas turbines. Refer to the Journal of Engineering for Gas Turbines and Power (January 2005) study of a high-pressure inlet fogging method for a gas turbine engine, Exercise 6.10 (p. 343). Recall that a number of independent variables were used to predict the heat rate (kilojoules per kilowatt per hour) for each in a sample of 67 gas turbines augmented with high-pressure inlet fogging. For this exercise, consider a first-order model for heat rate as a function of the quantitative independent variables cycle speed (revolutions per minute), cycle pres- sure ratio, inlet temperature (C), exhaust gas temperature (C), air mass flow rate (kilograms per second), and horsepower (Hp units). Theoret- ically, the heat rate should increase as cycle speed increases. In contrast, theory states that the heat rate will decrease as any of the other independent variables increase. The model was fit to the data in the GASTURBINE file with the results shown in the accompanying MINITAB printout. Do you detect any signs of multicollinearity? If so, how should the model be modified?

a

Based on the minitab output, there are multiple variables that look like they need to be dropped. Any values that exceed the VIF value 10 have a very high likelihood for multicolinearity. The high VIF values coupled with the fact that the beta estimates defy the theory of the science behind the data is grounds to suggest multi-collinearity.
Cutting the highest VIF scoring variable until all VIF scores were below the threshold for proceeding with the regression. Below are the VIF scores after trimming and the correlation chart respectively.

Variables <fctr>	VIF <dbl>
rpm	1.963139
cpratio	1.409922
exhtemp	1.947429
power	3.189915

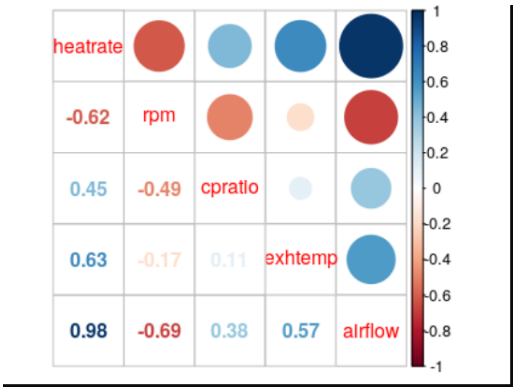
4 rows

After elininating terms based on VIF, until the VIF score of all terms is below the threshold, the model that is created after running through this screening process is shown below. All of the terms are statistically significant and are viable based off of the VIF scores.
Now that multiple correlation coefficients are taked care, the pairwise coefficients need to be taken care of.

```
lm(formula = heatrate ~ rpm + cpratio + exhtemp + airflow)
Residuals:
   Min     1Q Median     3Q    Max
-26850  -9180  -1995  10580  33549

Coefficients:
              Estimate Std. Error t value Pr(>t)
(Intercept) -1.730e+05  2.615e+04  -6.618 9.79e-09 ***
rpm          1.637e+00  3.848e-01   4.255 7.20e-05 ***
cpratio      2.885e+03  4.704e+02   6.133 6.62e-08 ***
exhtemp      2.015e+02  5.123e+01   3.932 0.000215 ***
airflow      3.993e+02  1.356e+01  29.451  < 2e-16 ***
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1          1
```

```
Residual standard error: 14100 on 62 degrees of freedom
Multiple R-squared:  0.9791,    Adjusted R-squared:  0.9777
F-statistic: 725.6 on 4 and 62 DF,  p-value: < 2.2e-16
```



After this correlation, we can see that airflow should be removed from the model due to the incredibly high correlation with our y (heatrate). A method of variable selection should then be applied to decide) After removing Airflow, we now have a valid set of data to work with. The next step would be a variable screening method to proceed with the actual regression. At this point we can apply the stepwise selection method to decide how to progress now that we can confirm, not just assume the variables are not collinear. The stepwise application results as follows :

```
Stepwise Selection Method

Candidate Terms:

1 . rpm
2 . cpratio
3 . exhtemp
```

Stepwise Selection Summary

Step	Variable	Added/ Removed	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	exhtemp	addition	0.398	0.389	57.9680	1696.2335	73857.7898
2	rpm	addition	0.665	0.654	6.3870	1659.0336	55554.1029
3	cpratio	addition	0.687	0.672	4.0000	1656.5232	54139.9156

The selected model after screening is predicting heatrate based off of rpm, cpratio, and exhtemp. The model is as follows $E(y) = -5.337 \times 10^5 - 5.943 x_{rpm} + 3.776 \times 10^3 x_{cpratio} + 1.147 \times 10^3 x_{exhtemp}$
In summary, yes, we detected signs of multicollinearity, we handled this by removing the collinear terms, then removing variables based on their pairwise correlation, then finally running the dataset through a variable screening method; in this case stepwise in order to refine the prediction model.

Q 21

Multicollinearity in real estate data. D. Hamilton illustrated the multicollinearity problem with an example using the data shown in the accompany- ing table. The values of x 1 , x 2 , and y in the table at right represent appraised land value, appraised improvements value, and sale price, respectively, of a randomly selected residential property. (All measurements are in thousands of dollars.)
(a) Calculate the coefficient of correlation between y and x 1 . Is there evidence of a linear rela- tionship between sale price and appraised land value?
(b) Calculate the coefficient of correlation between y and x 2 . Is there evidence of a linear rela- tionship between sale price and appraised improvements?
(c) Based on the results in parts a and b, do you think the model E(y) = 0 + 1 x 1 + 2 x 2 will be useful for predicting sale price?
(d) Use a statistical computer software package to fit the model in part c, and conduct a test of model adequacy. In particular, note the value of R 2 . Does the result agree with your answer to part c?
(e) Calculate the coefficient of correlation between x 1 and x 2 . What does the result imply?
(f) Many researchers avoid the problems of mul- ticollinearity by always omitting all but one of the redundant variables from the model. Would you recommend this strategy for this example? Explain. (Hamilton notes that, in this case, such a strategy can amount to throwing out the baby with the bathwater.)

a

With 95% confidence, we can say that the true value for the correlation coefficient is between (-0.51, 0.51), with the estimate being 0.00249. We fail to reject the notion that the correlation is zero with a t-test.
$H_0 : R^2 = 0$
$H_a : R^2 \neq 0$
p-value = 0.993
We fail to reject the null hypothesis and state that R = 0. There is not evidence of a linear relationship between sale price and appraised land value.

b

With 95% confidence, we can say that the true value for the correlation coefficient is between (-0.10, 0.77), with the estimate being 0.434. We fail to reject the notion that the correlation is zero with a t-test.
$H_0 : R^2 = 0$
$H_a : R^2 \neq 0$
p-value = 0.106
We fail to reject the null hypothesis and state that R = 0. There is not evidence of a linear relationship between improvement price and appraised land value.

c

Basing a conclusion solely off of the information from parts a and b, The model $y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$ seems to be a good candidate for regression for predicting appraised value. The correlation between the subset of pairwise variables and the y does not indicate anything of multicolinearity, so the model seems fine to proceed to testing. However there doesn't seem to be any useful correlation between the y and the respective x's, therefore the resultant regression model most likely will not be useful in predicting appraisal value.

Q 21 continued on next page...

d

Fitting the model $y = \hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2$ to the dataset, the results form an accurate model. The readout is as follows.

```
Call:
lm(formula = y ~ x1 + x2)

Residuals:
     Min       1Q   Median       3Q      Max
-0.13632 -0.09452 -0.02279  0.08629  0.16325

Coefficients:
             Estimate Std. Error t value Pr(>t)
(Intercept) -45.154136   0.611418  -73.85   <2e-16 ***
x1            3.097008   0.012274  252.31   <2e-16 ***
x2            1.031859   0.003684  280.08   <2e-16 ***
---
Signif. codes:
0    ***    0.001    **    0.01    *    0.05    .    0.1          1

Residual standard error: 0.1072 on 12 degrees of freedom
Multiple R-squared:  0.9998,    Adjusted R-squared:  0.9998
F-statistic: 3.922e+04 on 2 and 12 DF,  p-value: < 2.2e-16
```

The correlation between the model and the dataset is unreasonably high. High enough to prompt questioning of the data's independence. The value confirms the notion in part c that the model was a good candidate for a regression.

e

The results from calculating the R^2 value between the x1 and x2 is as follows :

```
data:  x1 and x2
t = -7.4348, df = 13, p-value = 4.94e-06
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.9665398 -0.7188453
sample estimates:
       cor
-0.8997765
```

The value -0.89977 is concerning, but not *technically* within our threshold of 0.9 for rejection or pairwise correlation. This close of a value lends itself to suggest that x1 and x2 are correlated, which would mean looking to remove one of them from the model. However attempting that only led to p-values for the betas that did not reject the null hypothesis that they were equal to zero; which would render the model useless.

f

Removing x1 results in a model that is not statistically significant.

```
Call:
```

```
lm(formula = y ~ x1)

Residuals:
     Min       1Q   Median       3Q      Max
-11.6910  -6.7912  -0.0326   6.4412  11.2993

Coefficients:
             Estimate Std. Error t value Pr(>t)
(Intercept) 1.199e+02  1.267e+01   9.463  3.4e-07 ***
x1          3.747e-03  4.161e-01   0.009    0.993
---
Signif. codes:
0    ***    0.001    **    0.01    *    0.05    .    0.1         1

Residual standard error: 8.324 on 13 degrees of freedom
Multiple R-squared:  6.24e-06,  Adjusted R-squared:  -0.07692
F-statistic: 8.112e-05 on 1 and 13 DF,  p-value: 0.993
```

Removing x1 results in a model that is not statistically significant.

```
Call:
lm(formula = y ~ x2)

Residuals:
     Min       1Q   Median       3Q      Max
-10.8999  -6.3345   0.0023   6.1458  10.4033

Coefficients:
            Estimate Std. Error t value Pr(>t)
(Intercept) 106.3194     8.1094  13.111 7.18e-09 ***
x2            0.1955     0.1125   1.737    0.106
---
Signif. codes:
0    ***    0.001    **    0.01    *    0.05    .    0.1         1

Residual standard error: 7.499 on 13 degrees of freedom
Multiple R-squared:  0.1884,    Adjusted R-squared:  0.126
F-statistic: 3.018 on 1 and 13 DF,  p-value: 0.106
```

Both of which are unusable regressions. In this case, avoiding the problems of multicollinearity by omitting all but one of the redundant variables from the model, destroys the model. Therefore we can't drop x_1 or x_2. I would not reccomend this strategy foor the situation as it results in a dead end where the two variable model results in a very significant model.