Tutorial topic 10, prediction and model quality

Willem Vervoort

Document generated on 2017-04-20

Objectives

- Interpret confidence intervals and prediction intervals;
- Calculate the bias and the RMSE for a validation set;
- Interpret model validation output to decide the best model.

EXERCISE 1 – Prediction and standard error of fit.

We will use the Corn dataset to predict y (CornP) for a specific pair of values of InorgP and OrgP, and we can simply substitute these into the fitted model:

$$\hat{y} = 66.465 + 1.290 * InorgP - 0.111 * OrgP$$

Here is a table of the data for which we would like a prediction

```
pander(newC[,2:3], caption="Organic and Inorganic Phosphorus data")
```

Table 1: Organic and Inorganic Phosphorus data

InorgP	OrgP
2.33	42
12.87	50
26.54	37
0.06	67
10.29	39

We can easily calculate the new values of CornP using the equation, but we cannot calculate the confidence intervals using the formula we learned in the lecture.

$$se_{\hat{y}} = \sqrt{s^2(1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{(n-1)*s_x^2})}$$

The predictions are:

```
newCP <- (66.645 + 1.290*newC$InorgP - 0.111*newC$OrgP)
pander(newCP)
```

64.99, *77.7*, *96.77*, *59.29* and *75.59*

A. Why can we not calculate the standard errors of the fit, using the equation given in the lecture?

Luckily R comes to the rescue and we can simply run:

```
mod <- lm(CornP~InorgP + OrgP, data=Corn)
newCornP <- predict(mod,newdata=newC, interval = "confidence", se.fit=T)
pander(newCornP$fit, caption="Prediction with confidence intervals")
```

Table 2: Prediction with confidence intervals

fit	lwr	upr
64.81	55.62	73.99
77.52	69.8	85.23
96.6	82.63	110.6
59.1	39.49	78.72
75.41	68.95	81.87

```
# and
newCornP_se <- predict(mod,newdata=newC, interval = "prediction")
pander(newCornP_se, caption="Prediction with prediction intervals")
```

Table 3: Prediction with prediction intervals

fit	lwr	upr
64.81	36.97	92.64
77.52	50.13	104.9
96.6	66.84	126.4
59.1	26.32	91.89
75.41	48.35	102.5

B. explain the difference between `interval = "confidence"` and `interval = "prediction"` (or the difference between the values in lwr and upr in the two tables).

We can identify the se of the fit for confidence interval from the r output:

```
pander(newCornP$se.fit, caption="se.fit for the confidence intervals")
```

1	2	3	4	5
4.282	3.597	6.513	9.144	3.013

C. What is the relationship between the se_fit of the prediction interval and the lwr and upr in the table?

EXERCISE 2 Calculate Bias and RMSE

Someone has gone out and actually measured CornP related to the data in the earlier table. Here are the results:

```
pander(newC, caption="observed data of CornP to match with Table 3" )
```

Table 5: observed data of CornP to match with Table 3

CornP	InorgP	OrgP
65.2	2.33	42
78.5	12.87	50
96.2	26.54	37
58.1	0.06	67
77.2	10.29	39

The bias in the prediction can be calculated as:

$$BIAS = \frac{\sqrt{\sum_{i-1}^{n}(y_i - \hat{y}_i)}}{n}$$

A. Calculate the bias in the prediction from Table 3 and Table 5. You can use R or excel for this, or do it by hand. Comment on wether the model over or under predicts.

The root mean square error of the prediction can be calculated as:

$$RMSE = \frac{\sqrt{\sum_{i-1}^{n}(y_i - \hat{y}_i)^2}}{n}$$

B. Calculate the RMSE in the prediction from the table 3 and table 5. You can use R or Excel to do this, or do it by hand. Comment on the accuracy of the model.

EXERCISE 3 Evaluate model quality using validation (an example exam question?)

To help you prepare, here is a question that could be similar to an exam question (no promises). This data consists of the measurements of Chlorophyll A (blue green algae). Blue green algae are a major drinking water quality concern, particularly in summer, however chlorophyll A is expensive and difficult to measure. We would like to find out what other water quality parameters can be used to predict chlorophyll A:

Response variable (y) Chla and transformed log10Chla: measure of chlorophyll A (mg/L)

Predictor variables:

1. maxPH: Maximum pH occurring on the measurement day, this is a measure of the alkalinity or acidity of the water (pH units);
2. Cl and transformed log10Cl: concentration of Chloride in the water on the measurement day (mg/L);
3. PO4 and transformed log10PO4: concentration of Phosphate in the water on the measurement day (mg/L).

Here is a snapshot of the data

	maxPH	Cl	log10Cl	PO4	log10PO4	Chla	log10Chla
3	8.1	40.02	5.33	187.1	2.272	15.6	1.193
5	8.06	55.35	10.42	97.58	1.989	10.5	1.021
20	7.83	88	1.944	586	2.768	16	1.204
41	8.3	54.14	1.734	326.9	2.514	11.84	1.073
70	7.4	13.5	1.13	104	2.017	21	1.322

The researchers, after going through variable selection have identified a model that includes maxPH, log10Cl and log10PO4 as the best model and now want to test this on a validation data set and assess the quality of the model.

Here is the summary of the final model and the residual plots

```
algae_mod <- lm(log10Chla ~ maxPH + log10Cl + log10PO4, data = chl_calib)
summary(algae_mod)
```

```
##
## Call:
## lm(formula = log10Chla ~ maxPH + log10Cl + log10PO4, data = chl_calib)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.11313 -0.25089 -0.02035  0.28834  1.09743
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.77485    0.60486  -6.241 4.47e-09 ***
## maxPH        0.38714    0.07562   5.120 9.51e-07 ***
## log10Cl      0.05237    0.03680   1.423    0.157
## log10PO4     0.61947    0.07075   8.756 4.62e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4042 on 146 degrees of freedom
## Multiple R-squared:  0.5018, Adjusted R-squared:  0.4916
## F-statistic: 49.02 on 3 and 146 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(algae_mod,which=c(1,2,5))
hist(rstandard(algae_mod))
```

```
par(mfrow=c(1,1))
```

A. Comment on the overall fit of the model

As the researchers have set aside part of the data for validation, we can check how well the model performs in validation. For this we first need to make a prediction and then calculate bias, RMSE and correlation and Lin's rho

Following the lecture material:

```
chl_predict <- predict(algae_mod, newdata = chl_valid)
# Accuracy (RMSE)
(RMSE_chl <- sqrt(mean((chl_valid$log10Chla - chl_predict)^2)))
```

```
## [1] 0.6375314
```

```
# Bias
(Bias_chl <- mean(chl_valid$log10Chla - chl_predict))
```

```
## [1] 0.5390444
```

B. Comment on the RMSE and bias and what this tells about the validation of the model and possibly the choice of the validation data set

The researchers now also calculated the correlation and Lin's concordance.

```
#Correlation
cor(chl_valid$log10Chla,chl_predict)
```

```
## [1] 0.3795479
```

```
#install.packages("epiR")
library(epiR)
```

```
## Warning: package 'epiR' was built under R version 3.3.3
```

```
## Loading required package: survival
```

```
## Loading required package: splines
```

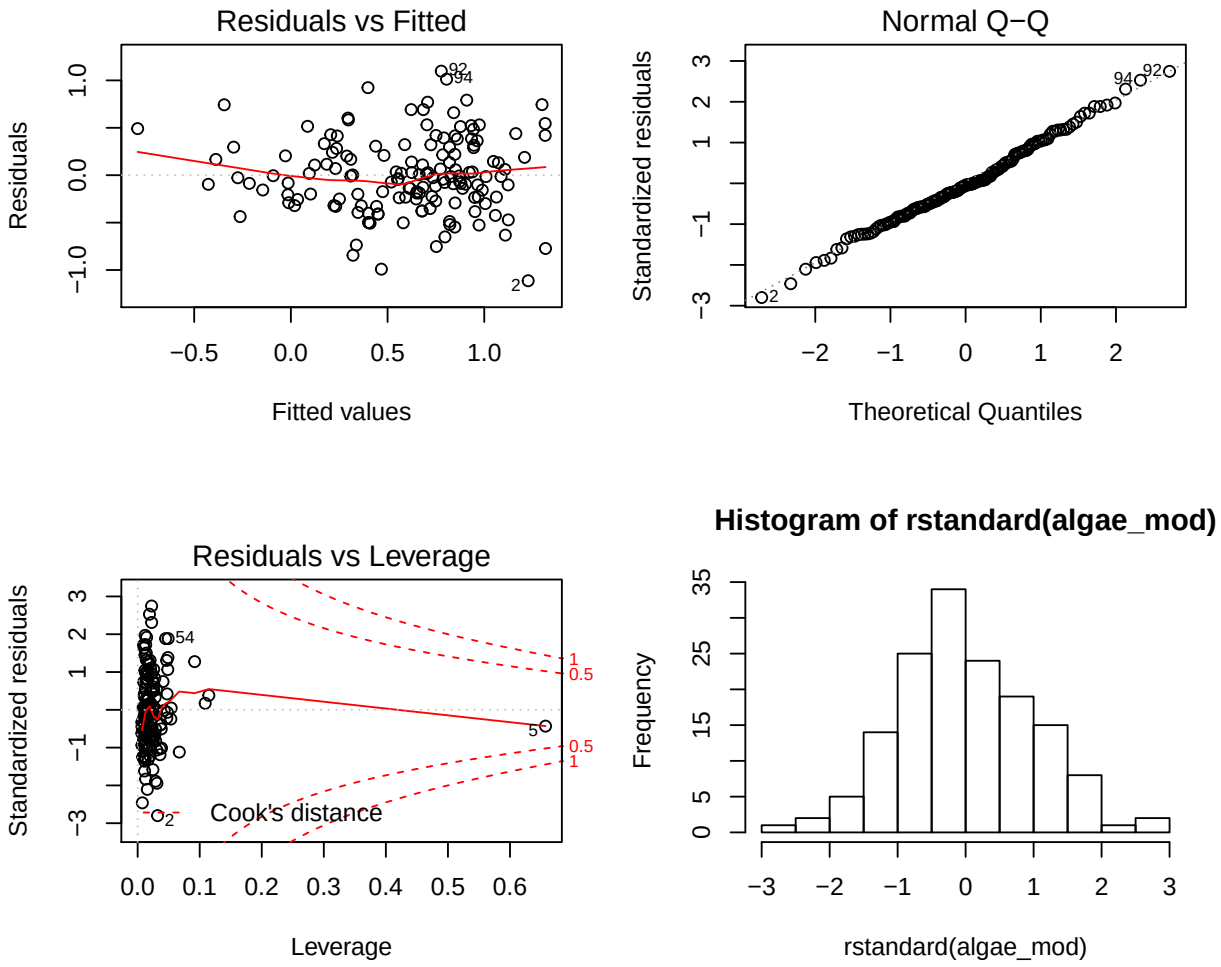

Figure 1: residual plots of the final algae prediction model

```
## Package epiR 0.9-82 is loaded

## Type help(epi.about) for summary information

##
```

```
valid.lcc<-epi.ccc(chl_valid$log10Chla, chl_predict)
valid.lcc$rho.c
```

```
##         est       lower     upper
## 1 0.1454675 0.006544931 0.2788813
```

C. Comment on the correlation and Lin's concordance in relation to the model output and explain if you think this model would be a useful model to make predictions of the occurence of Algae based on measurements of pH, Cl and PO4.

Clearly the correlation of the validation is poorer than for the calibration, but this initself is not surprising, this often happens. More concerning is the low value of Lin's concordance, which indicates the predictions are far from the 1:1 line and not in agreement. This suggest the model is not very good to predict future values of chla based on measurements of pH, Cl and PO4. Some more work identifying why the performance of the model in validation is poor is needed.

We can now plot observed against predicted.

```
# plot predicted versus observed
plot(chl_calib$log10Chla, predict(algae_mod),
     # colour = red, type = "16" and size is 20% larger
     pch = 16, col = "red", cex = 1.2,
     # add titles for axes and main
     xlab = "Observed dataset", ylab = "Predicted")
# insert a 1:1 line, dashed line, width = 2
abline(0, 1, lty = 2, lwd = 2)
# add the validation data
points(chl_valid$log10Chla, chl_predict,
       # colour = blue, type = "16" and size is 20% larger
       col = "blue", pch = 16, cex = 1.2)
# add a legend to the first plot
legend("topleft", c("calibration", "validation", "1 : 1 line"),
       pch = c(16, 16, NA), lty = c(NA, NA, 2), col = c("red", "blue", 1))
```

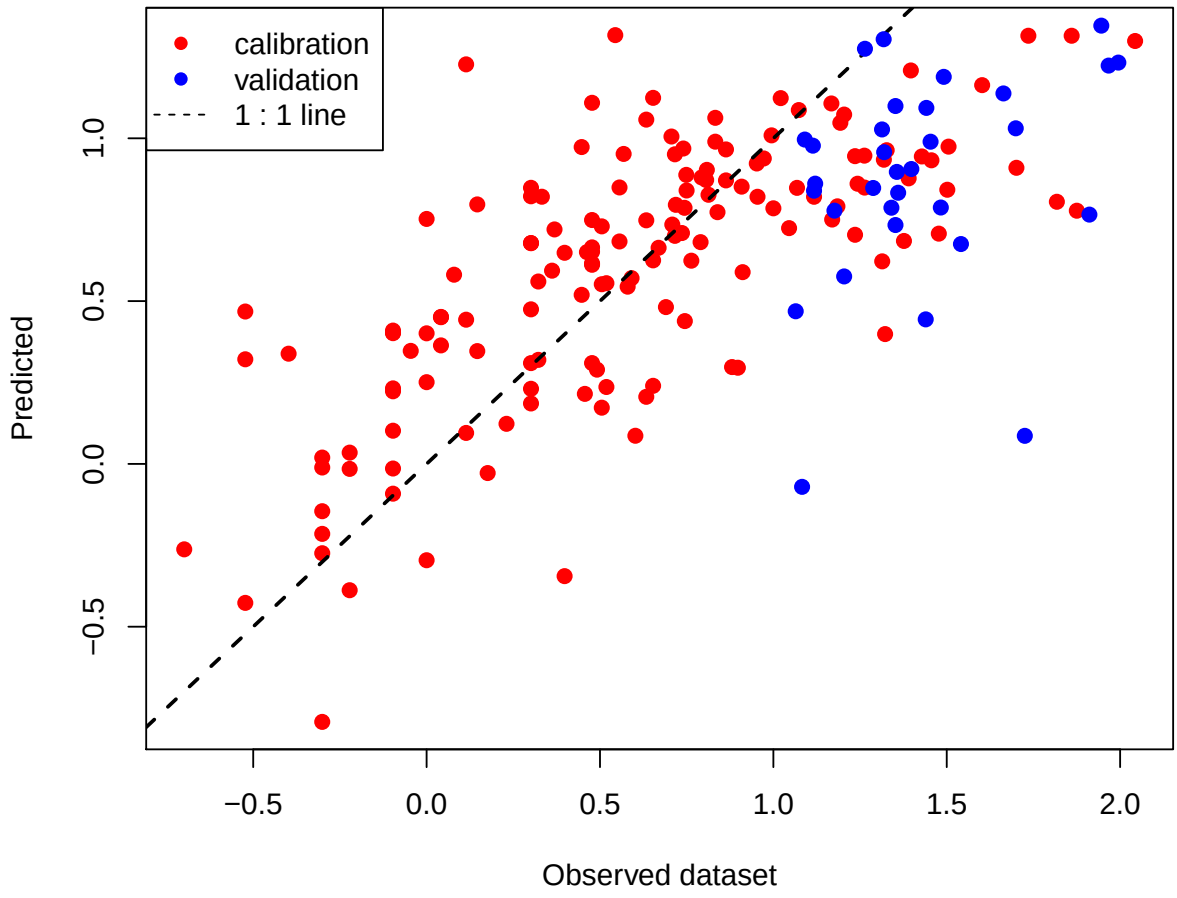

Figure 2: Plot of predicted versus observed for the algae model

d. Comment on whether this explains the results from b. and c.