

Pracownia Chemii Teoretycznej A/A' – objaśnienie do
 ćwiczenia 7 (QSAR)
 wersja 2.0 (14.12.2011)

Aby zgromadzić wszystkie potrzebne do zrobienia opisu dane w jednym pliku tekstowym, należy w pliku `zad1jiao_new.sas` (zakładam, że już zostały wprowadzone zmiany, o których była mowa na zajęciach), po wywołaniu procedury `pls` zakończonej `run;quit;` umieścić nową linię:

```
proc print data=predpls;run;
```

`predpls` zawiera tabelę, w której znajduje się m.in. przewidywana wielkość `pSw`. Do tabeli `predpls` można się zazwyczaj dostać tylko poprzez `Explorer` → `Libraries` → `Work` → `Predpls`. Dzięki procedurze `print` wszystkie wyniki z procedury `pls` lądują w `Output`, który jest łatwo zapamiętać jako plik tekstowy i przeanalizować w domu.

Uwaga: musimy wykonać procedurę `pls` w sumie 4 razy, dla wartości `nfac` równej 4,3,2,1.

Wyszukiwanie odpowiednich wielkości z `Output`:

1. Model REG: szukamy **drugiego** przebiegu procedury REG, czyli

```
The REG Procedure
Model: MODEL1
Dependent Variable: pSw
```

i dalej

```
Stepwise Selection: Step 1
```

- (a) następnie poszukujemy tabeli zatytułowanej:

```
Analysis of Variance
```

Znajdujemy w niej parametr `F`, p. kolumna `F Value`, oraz `Pr`, a także parametr `Corrected Total` i odchylenie standardowe modelu `Root MSE` (porównaj je z typową wielkością `pSw`!)

- (b) R^2 znajdujemy pod spodem, p. `R-Square`
- (c) W tabeli `Parameter Estimates` znajdujemy

- i. jakie deskryptory zostały użyte w modelu REG – kolumna **Variable**, przy czym **Intercept** oznacza wyraz wolny,
 - ii. jakie są współczynniki regresji liniowej – kolumna **Parameter Estimate**,
 - iii. oraz jakie są odchylenia standardowe dla tych współczynników – kolumna **Standard Error**
 - iv. wielkości VIF, czyli **Variance Inflation** – ostatnia kolumna
- (d) Znajdujemy tabelę **Output Statistics**, a w niej przewidywaną wartość **pSw** dla badanej cząsteczki. Pod tabelą znajduje się dodatkowo

Predicted Residual SS (PRESS)

Używając **Corrected Total** oraz **PRESS**, obliczamy Q^2 .

2. Model PLS

- (a) Szukamy **The PLS Procedure**, jeśli dołączyliśmy drukowanie tabel do głównego **Outputu**, to znajdziemy tabelę, w której ostatnia kolumna będzie nosić nazwę **pred_pls**. W tej kolumnie znajdziemy wartość **pSw** dla wybranej cząsteczki.
- (b) W sumie w **Outputcie** powinny być cztery wywołania procedury PLS, rozróżnialne po **Number of Factors** (równemu 4 lub 3 lub 2 lub 1).
- (c) ww. procedurę trzeba więc powtórzyć pozostałych **nfac**.

Model ma własności predykcyjne, jeśli

1. $R^2 > 0.8$
2. $Q^2 > 0.7$
3. F duże
4. $VIF < 10$