ANALYSE ET EVALUATION D'UN MODELE

Daniel Wallach
INRA Environnement & Agronomie
UMR ARCHE

1

Une partie du texte des commentaires est en anglais. Je m'en excuse. Nous sommes en train de publier un livre sur la modélisation et il était plus facile par endroit d'utiliser directement le texte du livre. (J'en profite pour faire de la pub. Le livre s'appelle « WORKING WITH DYNAMIC CROP MODELS. Evaluating, analyzing, parameterizing and using them », éditeurs Daniel Wallach, David Makowski et Jim Jones, maison d'édition Elsevier, sortie vers mars 2006.)

Plan

Analyse de sensibilité

Evaluation

- Comparaison modèle-observations
- Qualité de prédiction
- Quel niveau de complexité pour un modèle?

2

La première partie, l'analyse de sensibilité, est assez spécifique aux modèles dynamiques. En général des modèles statiques sont suffisamment simples pour que l'équivalent d'une analyse de sensibilité se fait en regardant les équations.

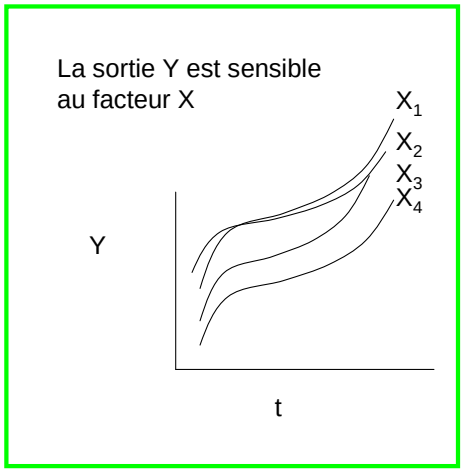
La partie évaluation s'appliquent en général aux modèles statiques ou dynamiques.

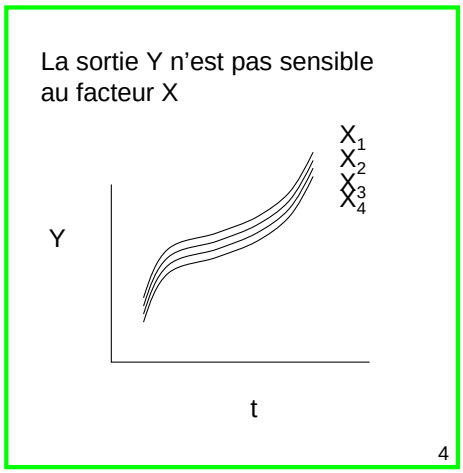
Analyse de sensibilité

Analyse de sensibilité

C'est quoi?

- Analyser comment les sorties varient en fonction des facteurs d'entrée





Par exemple, Y pourrait être la biomasse et X pourrait être un paramètre qui détermine l'effet de stress hydrique sur l'augmentation de la biomasse. On étudie différentes valeurs de X et on résout les équations pour connaître les conséquences sur Y.

Une question importante est de savoir comment choisir les valeurs de X. Il y a deux grandes familles de choix.

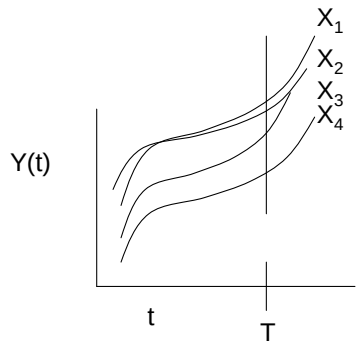
1. Arbitraire – on fait varier chaque facteur de -20%, _10%, +10%, +20%.
2. Il y a un niveau d'incertitude attaché au facteur. Par exemple, la densité de semis est connu à 5% près. On fait varier X alors dans cette plage.

On pourrait aussi faire varier plusieurs facteurs en même temps, pour étudier des interactions. (Le résultat de faire varier X1 et X2 n'est pas nécessairement la somme des effets individuels de X1 et de X2).

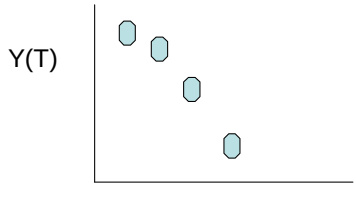
Analyse de sensibilité

Autre présentation

La sortie Y(t) en fonction de t pour plusieurs X



La sortie Y(T) en fonction de X



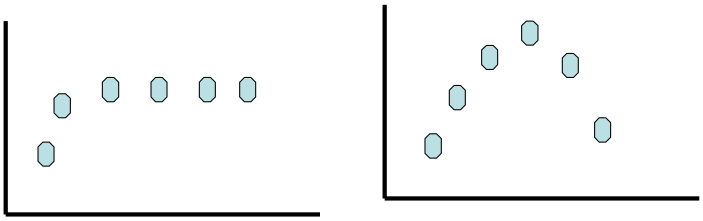
5

Il y a une deuxième façon de présenter les résultats qui est souvent utile. Au lieu de présenter Y(t) en fonction de t pour plusieurs valeurs du facteur X, on présente Y(T) en fonction de X, où T est un temps qui nous intéresse particulièrement (par exemple, récolte).

Analyse de sensibilité

Pourquoi (1) ?

- Explorer comment les sorties varient en fonction des entrées
- Vérifier le comportement général du modèle



6

Comme on a dit, le comportement du modèle après intégration en fonction des variables d'entrée est implicite dans les équations dynamiques, mais c'est trop complexe pour être vu sans une étude informatique spécifique. Cette étude s'appelle l'analyse de sensibilité.

Considèrons par exemple l'effet de la densité de semis sur le rendement. Quand la densité augmente, cela accélère la mise en place du système foliaire, donc la captation du rayonnement augmente ce qui a tendance à augmenter le rendement. Mais s'il y a des problèmes de stress hydrique, ils seront exacerbés par une augmentation de la densité à cause de l'augmentation de la transpiration, et cela aura tendance à réduire le rendement. L'effet final sur le rendement n'est pas évident. Une analyse de sensibilité avec densité comme facteur pourra le dire.

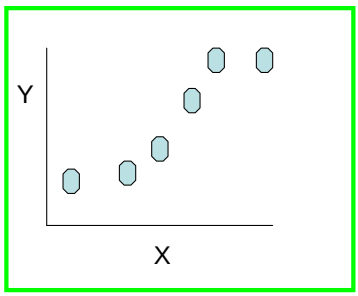
Dans d'autres cas, on connaît le comportement général du système et on veut vérifier que le modèle a le même comportement. Un exemple serait l'effet de la quantité d'irrigation sur le rendement du maïs. On s'attend à une réponse comme celle de gauche, le rendement augmentant avec une plus grande disponibilité en eau. Si par contre on trouve une réponse comme celle de droite, on cherchera l'erreur dans le modèle.

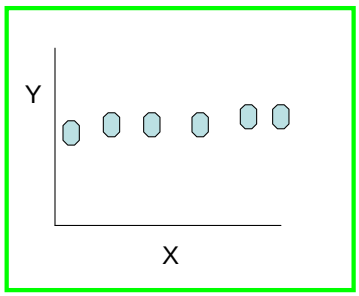
Digression – un domaine où l'analyse de sensibilité est très utilisée concerne les centrales nucléaires. Par exemple, les ingénieurs de ce domaine utilisent des modèles dynamiques pour calculer la pression dans les tuyaux, mais des équations, les valeurs des paramètres et les valeurs des variables d'entrée sont incertaines. Il est important pour eux (pour nous tous d'ailleurs) d'une part de s'assurer que les tuyaux tiendront malgré ces incertitudes, d'autre part d'identifier les sources principales d'incertitude.

Analyse de sensibilité

Pourquoi (2)?

- Identifier facteurs (paramètres ou entrées) importants
- Identifier facteurs dont l'estimation est cruciale
- Explorer possibilités de simplification





7

Certains facteurs auront une grande importance sur la sortie en question (graphique de gauche), d'autres n'auront que très peu d'effet.

Si la valeur d'un paramètre ou variable d'entrée n'a pas d'effet sur la sortie, il n'est pas important de la connaître.

On pourrait aller plus loin. On pourrait carrément l'éliminer du modèle et ainsi simplifier le modèle.

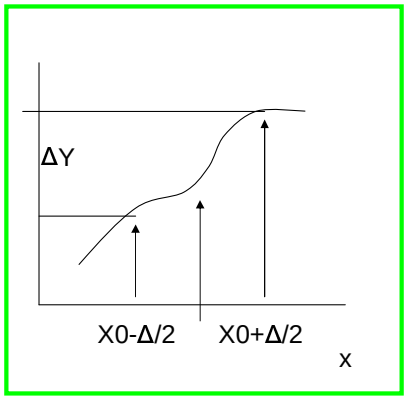
Par exemple, supposons que nous ayons un modèle de culture avec une partie qui décrit la dynamique du phosphore dans le sol, l'absorption de P par la plante et l'effet de P sur la croissance et développement de la culture. Il y aura plusieurs paramètres alors liés à cette partie du modèle. Il y aura aussi au moins une variable d'entrée, la quantité de phosphore initiale.

Supposons que Y (p.e. rendement) est insensible à tous ces paramètres et variables d'entrée. On pourrait alors simplifier le modèle en éliminant toute cette partie. (Bien sur, il faudrait vérifier que le manque de réponse est général et non pas lié à un seul site-année).

Analyse de sensibilité

Indice de sensibilité

- ΔY/ ΔX



8

Jusqu'ici on a regardé comment une variable de sortie Y varie quand un facteur X varie dans un intervalle. Faire ça pour plusieurs variables de sortie et de facteurs peut être très long et difficile à analyser. Il est alors important de définir un indice pour résumer l'importance d'un facteur. Comment quantifier l'importance d'un facteur pour une sortie? C'est-à-dire, comment définir un « indice de sensibilité »?

Un indice très utilisé est calculé en regardant Y quand le facteur X est déplacé à une valeur X-ΔX/2 puis à une valeur X+ΔX/2 (X étant sa valeur nominale, c'est-à-dire notre meilleure estimation). Soit ΔY le changement correspondant en Y. Alors le deuxième indice est ΔY/ ΔX. Il y a aussi des variantes, par exemple (ΔY/Y)/(ΔX/X) qui mesure le changement relatif de Y par rapport au changement relatif de X.

Evaluation

L'analyse d'incertitude concerne uniquement le modèle. Elle aide à mieux comprendre le fonctionnement du modèle. Mais elle ne concerne pas directement la relation entre modèle et monde réel. Le modèle pourrait être complètement faux, l'analyse de sensibilité ne le révèlerait pas.

C'est l'activité d'évaluation qui est concernée par la relation entre le modèle et le monde réel.

Evaluation. Pourquoi?

- Nécessaire pour démarrer
  - Définir objectif, critère de qualité
- Nécessaire pour améliorer le modèle
- Nécessaire pour l'utilisateur

10

Model evaluation is important for several reasons. Firstly, the simple fact of deciding to evaluate a model obliges one to answer some basic questions, including what is the objective of the model, what is the range of conditions where the model will be used, what level of quality will be acceptable.

Secondly, model improvement is impossible without evaluation. In the absence of a measure of model quality, how can one decide whether improvement is called for, and how can one know if a modified model is an improvement? As we shall see, evaluation can provide not only an overall indication of quality but can also quantify the errors resulting from different causes. Then further efforts can be focused on reducing the major errors.

Finally, evaluation is important for potential users of a model. The user needs information about the quality of the model in order to decide how much credence to give to model results.

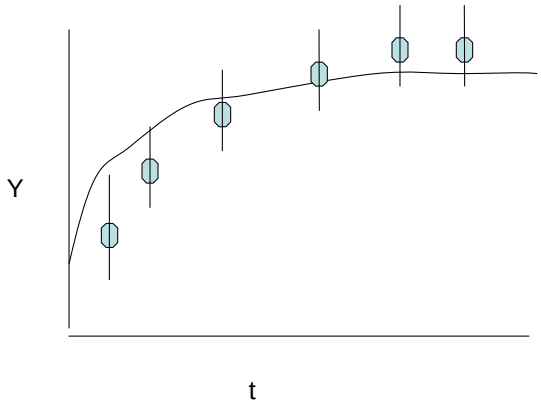
Comparaison entre valeurs calculées par le modèle et valeurs observées

Un premier aspect du travail d'évaluation est la comparaison entre valeurs observées et valeurs calculées. On verra par la suite que l'on doit faire attention aux conclusions que l'on en tire. Mais d'abord, on parlera des approches qui sont utilisées pour faire cette comparaison.

Il existe un très grand nombre de façons pour comparer valeurs observées et valeurs calculées. Beaucoup sont intéressantes, et apportent des informations utiles. Néanmoins, on n'a pas le temps ici de les présenter toutes. On ne présentera que quelques approches, qui sont parmi les plus utilisées.

Comparaison modèle-observations

Graphique avec valeurs calculées et observations (plusieurs réponses, une situation)



12

C’est dans le cas où on a des observations dans le temps sur une variable (par exemple, poids vif d’une brebis).

Comparaison modèle-observations

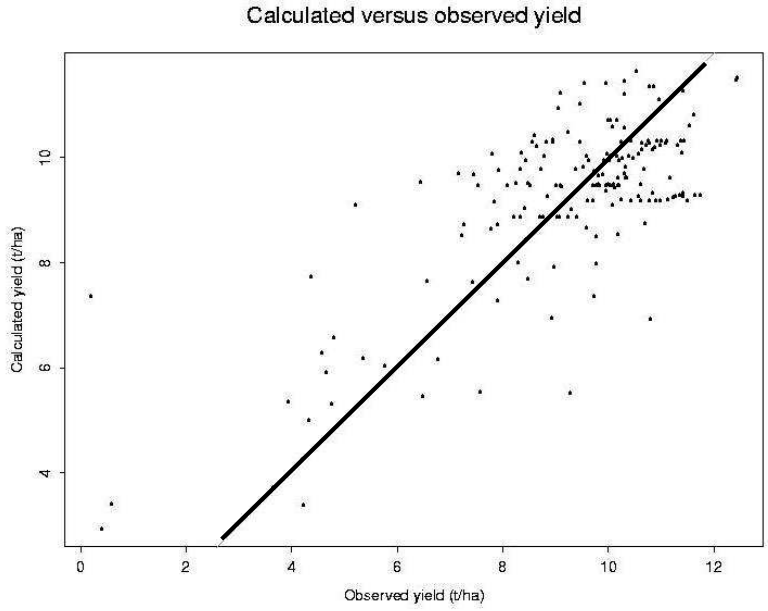
Graphique avec valeurs calculées et observations (une réponse, plusieurs situations)

- Valeurs calculées contre valeurs observées

13

C’est probablement la présentation la plus universelle pour comparer valeurs calculées et valeurs observées.

Comparaison modèle-observations



14

On a, pour chaque situation, une valeur observe (en abscisse) et une valeur calculé (en ordonnée). La ligne qui passe par 0 et qui a une pente de 1, représente l'équation (valeur calculée)=(valeur observée). Si le modèle était parfait, tous les points tombera exactement sur cette ligne. C'est donc la distance de la ligne qui mesure l'erreur du modèle.

Ici, les points viennent d'un modèle pour le maïs. Il s'agit des résultats d'essais sur l'irrigation du maïs, l'objectif du modèle étant de raisonner les stratégies d'irrigation. On a l'impression que globalement les valeurs calculées augmentent quand les valeurs observées augmentent, mais il y a pas mal de variabilité.

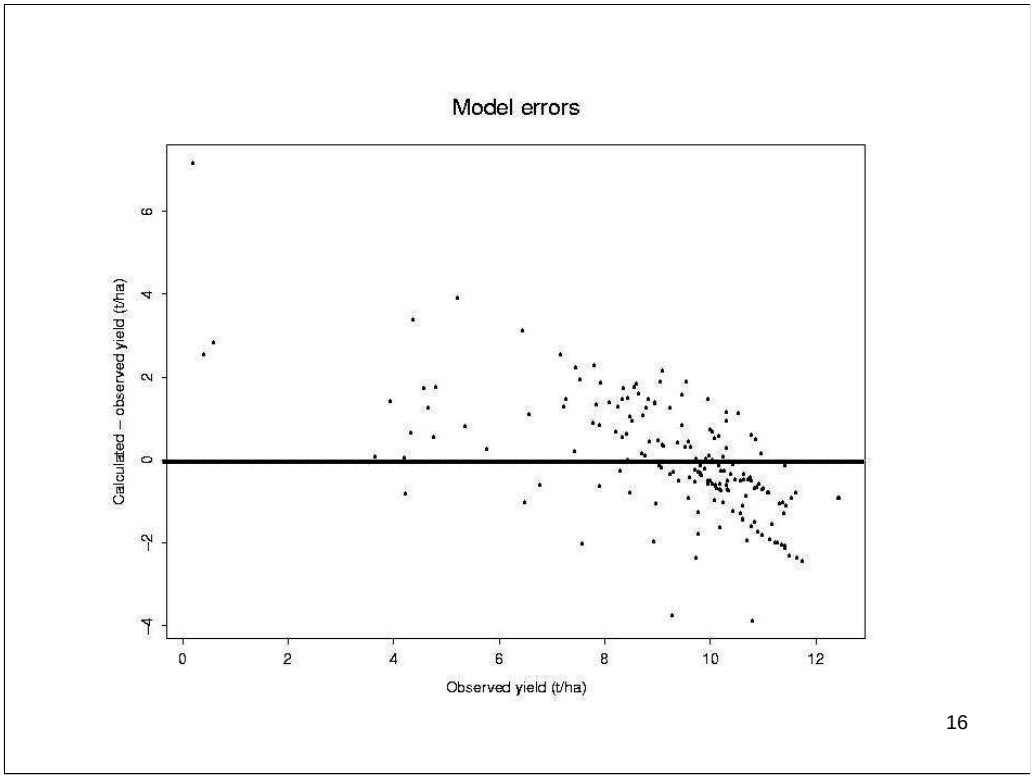
Comparaison modèle-observations

Graphique avec valeurs calculées et observations (une réponse, plusieurs situations)

- Valeurs des résidus contre valeurs observées.

15

Un résidu est la différence entre la valeur calculée et la valeur observée (c'est-à-dire, c'est l'erreur du modèle).



Il s'agit ici des mêmes données que sur le graphique précédent, seulement la présentation est différente.

En statistique, ce type de graphique est la façon classique de présenter la qualité du modèle. C'est maintenant la distance à la ligne horizontale Y=0 qui mesure l'erreur du modèle.

Ce type de graphique permet de tester en même temps les deux modèles; le modèle pour l'espérance et le modèle pour l'erreur. Le modèle pour l'espérance est bien si les erreurs sont faibles partout. Ici il semble que certaines des eerreurs sont assez appréciables. Le modèle statistique est souvent que l'espérance de l'erreur aléatoire est zero et que la variance est constante. Ici, il semble ici que le modèle sur-estime le rendement en cas de stress important (petit rendement) et sous-estime pour des grands rendements. C'est à dire, l'espérance de l'erreur aléatoire est positive pour petit Y et négative pour grand Y.

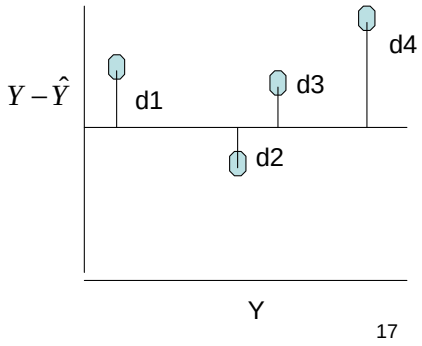
Ce type de graphique est moins utilisé que le précédent pour des modèles, qui est très dommage. Les deux peuvent être utiles et complémentaires.

Comparaison modèle-observations

Mesure numérique (1)

- L'erreur quadratique moyenne (Mean Squared Error MSE)

$$MSE = (1/N) \sum_{i=1}^{N} (Y_i - \hat{Y}_i)^2$$



17

Les graphiques sont très utiles pour voir d'un coup d'œil la distance entre valeurs observées et valeurs calculées par le modèle. Pourtant, on a besoin aussi de valeurs numériques comme mesures simples de cette distance. .

MSE est probablement le critère quantitatif le plus utilisé pour comparer valeurs calculées et valeurs observées.

Pour chaque observation on prend la différence entre valeur calculée et valeur observée (c'est l'erreur du modèle), on prend le carré, et enfin on prend la moyenne des carrés sur toutes les observations.

On peut montrer que MSE est une somme de trois contributions différentes. La première est le biais du modèle (l'erreur moyenne) au carré. Ce terme pourrait être grand, par exemple, si le modèle sur-estime systématiquement Par exemple, si le modèle ne prend pas en comptes des maladies qui sont en effet importantes, ça pourrait être le cas. Le deuxième terme est la différence entre la variabilité des résultats observées et la variabilité des résultats calculés. Si par exemple le modèle est bon en moyenne mais n'est pas suffisamment sensible aux stress, alors les résultats calculées seront peu variables avec différents niveaux de stress comparés aux résultats observés. Le dernier terme enfin dépend en détail de la corrélation entre valeurs observées et calculées et est plus difficile à interpréter que les autres.

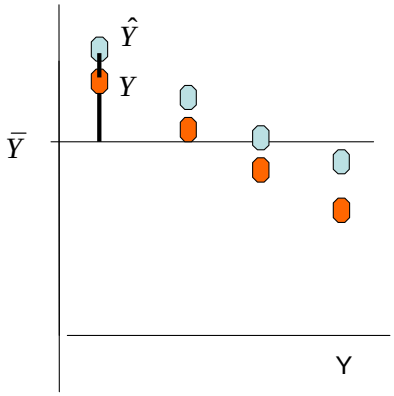
Il est facile de calculer, à partir d'un jeu de données qui comporte valeurs observées et valeurs calculées, ces trois termes. On peut alors avoir une première idée sur le type d'erreur qui prédomine, et cela peut aider à améliorer le modèle.

Comparaison modèle-observations

Mesure numérique (2)

- L'efficience

$$EF = 1 - \frac{\sum_{i=1}^{N} \left(Y_i - \hat{Y}_i \right)^2}{\sum_{i=1}^{N} \left(Y_i - \bar{Y} \right)^2}$$



$$\hat{Y}_i = Y_i \Rightarrow EF = 1 \qquad \hat{Y}_i = \bar{Y} \Rightarrow EF = 0$$

18

L'efficience est un deuxième critère très utilisé pour mesurer l'accord entre valeurs mesurées et valeurs calculées.

Yi est la valeur calculée pour situation i. Yi chapeau est la valeur correspondante calculée par le modèle. Y barre représente la moyenne des Yi.

Si le modèle est parfait, les valeurs calculées égalent exactement les valeurs mesurées et EF=1. C'est la valeur maximale.

Si les valeurs calculées égalent la moyenne des valeurs mesurées, le numérateur = le dénominateur et EF=0. C'est-à-dire, si un modèle n'explique pas plus de la variabilité des mesures que la moyenne des mesures, son efficience est 0. L'efficience peut être encore pire (négative).

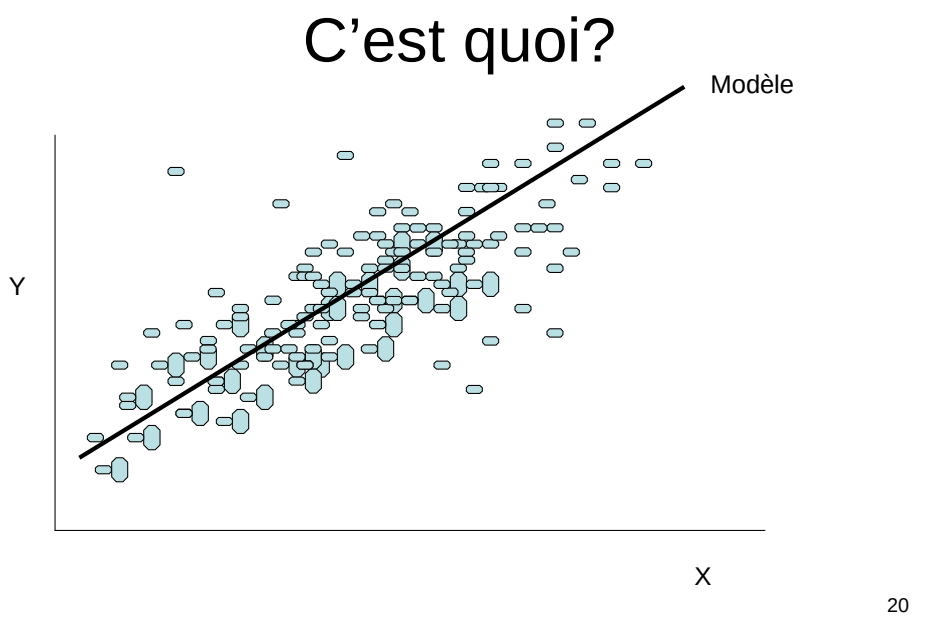
Ce type de mesure a l'avantage d'être comparable entre différents modèles et données. Les valeurs 0 et 1 ont toujours le même sens.

Évaluation. Erreur de prédiction

In general, however, our real interest is not in how well the model reproduces data that has already been measured, but rather in how well it can predict new results. The assumption, often implicit, underlying the use of past measurements is that the agreement of the model with those data can inform us about how the model will perform in the future. However, that assumption is not always founded.

On va d’abord définir un critère de qualité de prédiction. Ensuite on expliquera comment on peut estimer la qualité de prédiction. Enfin, on verra que cela aide à comprendre comment faire un modèle pour qu’il soit un bon prédicteur.

Erreur de prédiction



Avant de définir un critère, il faut comprendre ce que l’on entend par la qualité ou dans l’autre sens l’erreur de prédiction?

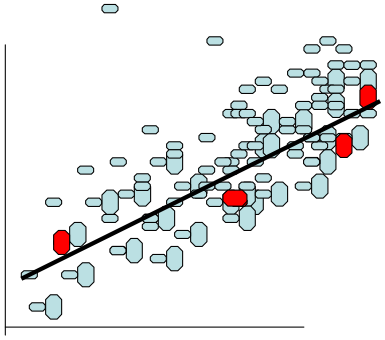
D’abord, il faut définir la gamme de situations pour lesquelles on veut faire des prédictions. On appelle ça la population cible. L’erreur de prédiction prend en compte l’erreur pour toutes ces situations. Par exemple, tous les champs de maïs en France. C’est une population infinie parce qu’il y a un nombre infini de climats possibles, sans parler des décisions de gestions et des sols.

Erreur de prédiction

Critère de qualité de prédiction

- MSEP=erreur quadratique moyenne de prédiction (Mean squared error of prediction)

$$MSEP = E\left[\left(Y - \hat{Y}\right)\right]^2$$



Rappel $MSE = (1/N)\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2$

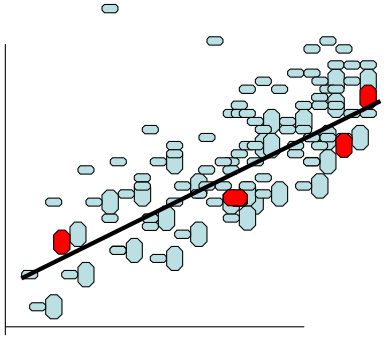
21

Le critère communément utilisé en statistiques pour évaluer la qualité de prédiction est l'erreur quadratique moyenne de prédiction (Mean Squared Error of Prediction = MSEP). Cela ressemble beaucoup à MSE. La différence est que MSE comporte une somme sur les observations, tandis que MSEP comporte une espérance. Cela veut dire qu'il s'agit d'une moyenne sur toutes les situations où on voudrait appliquer le modèle.

Erreur de prédiction

Estimation de MSEP

- MSEP ressemble à MSE. Peut-on alors utiliser MSE pour estimer MSEP?
  - Si les mesures sont un échantillon de la distribution cible
  - Si les données n'ont pas été utilisés pour estimer les paramètres du modèle.



22

MSEP concerne toutes les situations où on voudrait utiliser le modèle. On ne peut pas en général observer toutes ces situations. C'est-à-dire, en général on ne peut pas mesurer directement MSEP. On doit l'estimer à partir des quelques situations où on a des mesures.

On a vu que MSEP et MSE se ressemblent beaucoup. Il est alors tentant d'utiliser MSE pour estimer MSEP. C'est-à-dire, on utiliserait la moyenne des erreurs quadratiques mesurées pour estimer la moyenne des erreurs quadratiques sur toute la population cible.

Sous deux conditions, c'est exactement cela que l'on fait. Mais les conditions sont très importantes. Si on ne remplit pas les conditions, on peut se tromper de beaucoup en utilisant MSE comme estimateur de MSEP.

La première condition concerne la population cible. Pour que MSE soit un estimateur raisonnable de MSEP, il faut que les situations mesurées soient un échantillon représentative des situations de la population cible. C'est évident mais important. Supposons que les mesures ont été faites dans des essais de tournesol avec protection phytosanitaire quasi-complète, et que la population cible concerne des champs d'exploitants avec une gamme de niveaux de protection. MSE ne mesure alors que la qualité du modèle avec protection poussé, et ne peut pas fournir d'informations sur la qualité de prédiction en présence de maladies. MSE ne sera pas un estimateur raisonnable de MSEP.

La deuxième condition concerne l'utilisation des données mesurées pour la mise au point du modèle. Si le modèle a été ajusté aux données du jeu de données, alors MSE n'est pas un estimateur raisonnable de MSEP. Il est relativement facile de voir pourquoi. Dans ce cas, les paramètres du modèle ont été choisi pour reproduire spécifiquement les données utilisées pour calculer MSE. MSEP par contre concerne l'erreur du modèle pour toutes les situations. Cette erreur sera en général plus grande que l'erreur entre modèle et les valeurs auxquelles le modèle a été ajusté. Les 2 transparents suivants illustrent cet effet.

Erreur de prédiction

Exemple de la différence entre MSE et MSEP

- 8 observations (Y,x1,x2,x3,x4,x5)

$$Y = \theta^{T0} + \theta^{T1} x_1 + \theta^{T2} x_2 + \theta^{T3} x_3 + \theta^{T4} x_4 + \theta^{T5} x_5 + \varepsilon \quad \varepsilon \sim N(0, \sigma^2)$$

- 5 modèles linéaires

$$\hat{Y}_{2\ paramètres} = \theta^{(0)} + \theta^{(1)} x_1$$

$$\hat{Y}_{3\ paramètres} = \theta^{(0)} + \theta^{(1)} x_1 + \theta^{(2)} x_2$$

$$\hat{Y}_{4\ paramètres} = \theta^{(0)} + \theta^{(1)} x_1 + \theta^{(2)} x_2 + \theta^{(3)} x_3$$

$$\hat{Y}_{5\ paramètres} = \theta^{(0)} + \theta^{(1)} x_1 + \theta^{(2)} x_2 + \theta^{(3)} x_3 + \theta^{(4)} x_4$$

$$\hat{Y}_{6\ paramètres} = \theta^{(0)} + \theta^{(1)} x_1 + \theta^{(2)} x_2 + \theta^{(3)} x_3 + \theta^{(4)} x_4 + \theta^{(5)} x_5$$

23

Ce n'est pas un modèle dynamique mais un modèle linéaire, mais le principe est exactement le même. Cette discussion s'applique aussi bien aux modèles dynamiques qu'aux modèles statiques.

On a un jeu de 8 situations (8 données), avec valeurs observées, variables explicatives et valeurs calculées par le modèle. Le « vrai » modèle, que l'on a utilisé pour générer des données, est un modèle linéaire avec un terme aléatoire d'espérance 0.

On teste 5 modèles différents. Le premier est de la forme $\theta_0 + \theta_1 X_1$, le deuxième a la forme $\theta_0 + \theta_1 X_1 + \theta_2 X_2$ etc. Chaque modèle rajoute une variable explicative par rapport au modèle précédent. On ajustera les paramètres de chaque modèle aux 8 données.

Une fois que l'on a fait l'ajustement, on peut calculer MSE pour chaque modèle. Pour cet exemple artificiel où on connaît le vrai modèle (c'est-à-dire l'équation qui définit toute la population cible) on peut aussi calculer MSEP. Les résultats sont donnés sur la transparente suivante.

Erreur de prédiction

Modèle	Paramètres à estimer Valeurs estimées par moindres carrées	Λ	Δ	$MSEP(\hat{\theta})$	MSE
2 params	$\theta^{(0)}, \theta^{(1)}$ 2.535, 8.275	4.04	0.36	4.40	4.61
3 params	$\theta^{(0)}, \theta^{(1)}, \theta^{(2)}$ 2.121, 8.005, 2.065	0.04	0.02	0.06	0.01
4 params	$\theta^{(0)}, \theta^{(1)}, \theta^{(2)}, \theta^{(3)}$ 2.046, 7.971, 2.085, 0.091	0.04	0.01	0.05	0.01
5 params	$\theta^{(0)}, \theta^{(1)}, \theta^{(2)}, \theta^{(3)}, \theta^{(4)}$ 1.906, 7.906, 2.036, 0.169, 0.156	0.04	0.05	0.09	0.004
6 params	$\theta^{(0)}, \theta^{(1)}, \theta^{(2)}, \theta^{(3)}, \theta^{(4)}, \theta^{(5)}$ 1.641, 7.735, 1.967, 0.237, 0.230, -0.174	0.04	0.35	0.39	0.0003

24

Pour le prmieer modèle on a une seule variable explicative et on ajuste 2 paramètres, dans le deuxième cas 2 variables explicatives (et 3 paramètres), jusqu'au dernier modèle qui a 5 variables explicatives et 6 paramètres.

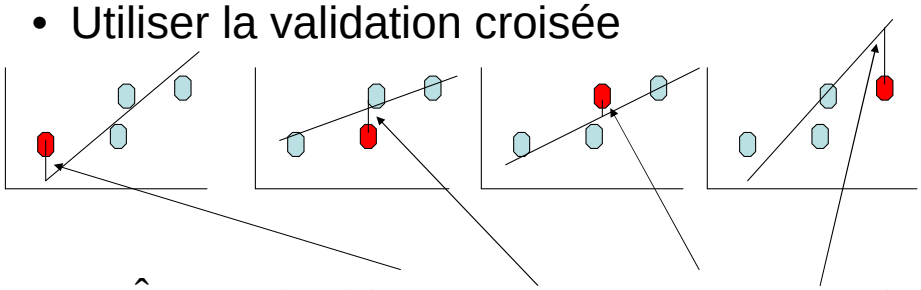
Regardons d'abord la différence entre MSE et MSEP. Pour le modèle le plus simple ils ne sont pas très différents, puis les différences augmentent. Pour modèles 2 et 3 il y a un facteur de 5 ou 6 de différence, puis un facteur de 22 pour le modèle 4 et enfin, pour le dernier modèle, avec 6 paramètres, un facteur de 1300! C'est-à-dire, plus on ajuste de paramètres au modèle, plus MSE sous-estime MSEP, et cette sous-estimation peut être énorme.

Ensuite, regardons le comportement général de MSE et de MSEP. MSE diminue systématiquement quand on ajuste plus de paramètres. Si on base le choix de modèle sur MSE, on choisira toujours le modèle le plus complexe. MSEP par contre diminue au début puis augmente, avec un minimum pour le modèle 3. Ce comportement est typique. L'erreur de prédiction a en général un minimum pour un complexité intermédiaire. Ce tableau illustre par ailleurs la notion de « surparamétrisation ». Les modèles 4 et 5 sont surparamétrés. L'ajustement de paramètres supplémentaires, par rapport au modèle 3, a fait augmenter l'erreur de prédiction.

Erreur de prédiction

Comment estimer MSEP si MSE n'est pas un bon estimateur?

- Diviser le jeu de données en deux parties
  - Estimer les paramètres sur une partie, calculer MSEP sur la deuxième partie
- Utiliser la validation croisée



$$\hat{MSEP} = (1/4)(MSE_1 + MSE_2 + MSE_3 + MSE_4)$$ 25

On estime les paramètres sur toutes les données sauf une et on calcule MSE=MSEP pour celle-là.

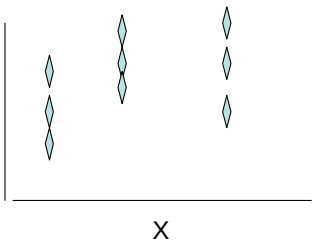
On laisse tomber tour à tour chaque donnée

A la fin, l'estimateur de MSEP=moyenne des MSE

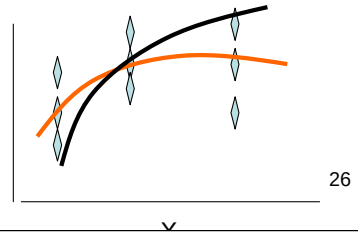
Niveau de complexité

$$MSEP = \Lambda + \Delta$$

$$\Lambda = E_X\left[\text{var}\ (Y|X)\right]$$
$$= \text{population variance}$$



$$\Delta = E_X\{\,[E_Y(Y|X) - \hat{Y}(X)]^2\}$$
$$= \text{squared bias}$$



Une anlayse plus approfondie de MSEP peut aider à comprendre l'effet de complexité sur la qualité de prédiction d'un modèle.

MSEP can be written as the sum of two terms.

The population variance, Λ, depends on how much *Y* varies for fixed values of the explanatory variables in the model. When *X* is fixed *Y* still varies, within the target population, because not all the variables that affect *Y* are included in the model. That variability is then averaged over *X*. Note that Λ does not involve the model. The form of the model is irrelevant here. It is only the choice of the explanatory variables that is important. If the explanatory variables in the model do not explain most of the variability in *Y*, then the remaining variability in *Y* for fixed *X* is large and Λ is large. Consider for example a model which does not include initial soil mineral nitrogen. If *Y* (for example yield) for the target population is strongly affected by initial soil nitrogen, then Λ will large. We see that the choice of explanatory variables is a major decision as far as prediction accuracy is concerned. That choice sets a minimum value for mean squared prediction error. Even if the model is the best possible, the mean squared error of prediction cannot be less than the population variance Λ.

The squared bias term, Δ, does depend on the form of the model. Once the choice of explanatory variables in the model is made, then the best model (minimum value of *MSEP* is the model that predicts a value equal to at each value of *X*. The bias measures the distance between this best prediction and the model prediction, averaged over the target distribution of *X* values. The bias may be due to errors in the form of the model or to errors in the parameter values.

Niveau de complexité

Model	Parameters in the model Least squares parameter values	Λ	Δ	$MSEP(\hat{\theta})$	*MSE*
$f_1(X;\theta)$	$\theta^{(0)},\theta^{(1)}$ 2.535, 8.275	4.04	0.36	4.40	4.61
$f_2(X;\theta)$	$\theta^{(0)},\theta^{(1)},\theta^{(2)}$ 2.121, 8.005, 2.065	0.04	0.02	0.06	0.01
$f_3(X;\theta)$	$\theta^{(0)},\theta^{(1)},\theta^{(2)},\theta^{(3)}$ 2.046, 7.971, 2.085, 0.091	0.04	0.01	0.05	0.01
$f_4(X;\theta)$	$\theta^{(0)},\theta^{(1)},\theta^{(2)},\theta^{(3)},\theta^{(4)}$ 1.906, 7.906, 2.036, 0.169, 0.156	0.04	0.05	0.09	0.004
$f_5(X;\theta)$	$\theta^{(0)},\theta^{(1)},\theta^{(2)},\theta^{(3)},\theta^{(4)},\theta^{(5)}$ 1.641, 7.735, 1.967, 0.237, 0.230, -0.174	0.04	0.35	0.39	0.0003

27

La somme des deux termes Λ et Δ = MSEP. En rajoutant de la complexité au modèle (en ajoutant des variables explicatives supplémentaires), on fait diminuer Λ. C'est forcément le cas. La variabilité inexpliqué par les variables explicatives du modèle diminue forcément si on rajoute des variables explicatives supplémentaires. Au pire une nouvelle variable explicative n'est pas lié à Y et dans ce cas Λ reste inchangé. Par contre, Δ augmente en moyenne. Ce terme provient de deux sources d'erreur; erreurs dans la orme des équations et erreurs dans les valeurs des paramètres. Supposons que la forme du modèle est correcte. Dans ce cas, ce terme résulte uniquement des erreurs dans les valeurs des paramètres. Chaque paramètre supplémentaire doit être estimé et apporte donc une erreur d'estimation supplémentaire.

The above decomposition of *MSEP* into two terms can help to understand the consequences of choosing different levels of detail for a model. Adding more detail in general involves including additional explanatory variables. This has two opposing consequences. On the one hand, adding additional explanatory variables will reduce (or at worst leave unchanged) the unexplained variability in *Y* once *X* is fixed. That is, Λ will decrease or at worst remain unchanged. On the other hand, there will in general be additional equations and parameters to estimate in conjunction with the additional explanatory variables. This will in general lead to an increase in the squared bias term Δ.

Suppose that one has a preliminary model and wants to decide whether or not to add additional explanatory variables. There is a better chance that the additional explanatory variables will reduce *MSEP* if

•they play an important role in determining the variability in *Y* in the target population, so that adding them to the model reduces Λ by a substantial amount.

•the associated equations and parameters can be well estimated from the available data, so that the additional detail does not cause a substantial increase in Δ.

FIN

28