AN INTRODUCTION TO STATISTICAL LEANRING

Chapitre 1 - Introduction

L'apprentissage statistique se réfère à des outils pour comprendre les données.
On peut les classer en deux catégories : apprentissage supervisé et non supervisé
L'apprentissage supervisé a pour but de construire un modèle pour estimer, prédire un sortie basée sur des entrées.
L'apprentissage non supervisé quand à lui, il possède des sorties mais pas de sortie. On apprend les relations et structures des données

Chapitre 2 – apprentissage satistique

I Qu'est-ce que l'apprentissage statistique ?

De facon générale, on suppose observer une reponse quantitative et p differents predicteurs X1,X2,...,Xp.
On suppose que la relation entre Y et X=(X1,...,Xp) peut etre ecrite sous la forme
$Y = f(X) + \varepsilon$
où f est une fonction inconnue de X1,...,Xp et ε est une erreur aléatoire indépendante de X et de moyenne nulle.

L'apprentissage statistique consisite a s'approcher d'une estimation de f

Il y a deux raisons pour que l'on puisse désirer d'estimer f : prediction et inference

a) prediction

On peut prédire Y en utilisant la moyenne $\hat{Y} = \hat{f}(X)$
où $\hat{f}$ représente notre estimation de f et $\hat{Y}$ représente le résuklat de prédiction de Y
$\hat{f}$ est souvent considéré comme une boite noire

La préicsion de $\hat{Y}$ comme prediction de Y depend de deux quantités: erreur reductible et l'irreductible erreur. En général $\hat{f}$ ne sera pas un estimation parfaite de f, et son imprécision introduira quelques erreurs. Cette erreur est reductible car on peut potentiellement améliorer la précision de $\hat{f}$ en utilisant une technique statistique plus appropriée pour estimer f.
Cependant meme sil est possible de se rapprocher d'une estimation parfaite de f, notre prédiction aura encore des erreurs. Cela est du au fait que Y est aussi une fonction de ε qui ne peut etre prédite en utlisant X. Donc ε peut affecter notre prediction. Cela est connu sous le nom d'erreur irreductible.

Considérons une estimation de $\hat{f}$ et un jeu de pridcteurs X, qui implique la prediction $\hat{Y} = \hat{f}(X)$. On suppose que $\hat{f}$ et X sont fixés. On eput donc aisement montrer que
$E(Y - \hat{Y})^2 = E[f(X) + \varepsilon - \hat{f}(X)]^2$
$E(Y - \hat{Y})^2 = [f(X) - \hat{f}(X)]^2 + Var(\varepsilon)$
où la premier partie est reductible mais la seconde irreductible
Où $E(Y - \hat{Y})^2$ représente la moyenne ou la valeur attendue du carré de la différence entre la valeur prédite et la valeur actuelle de Y

Démonstration:

$$Var[X] = E[X^2] - E[X]^2$$
$$E[X^2] = Var[x] + E[X]^2$$
$$E[f] = f$$
$$y = f + \varepsilon$$
$$E[\varepsilon] = 0$$
$$E[y] = E[f + \varepsilon] = E[f] = f$$
$$Var[y] = E[(y - E[y])^2] = E[(y - f)^2]$$
$$E[(y - \hat{f})^2] = E[y^2 + \hat{f}^2 - 2y\hat{f}]$$
$$E[(y - \hat{f})^2] = E[y^2] + E[\hat{f}^2] - E[2y\hat{f}]$$
$$E[(y - \hat{f})^2] = Var[y] + E[y]^2 + Var[\hat{f}] + E[\hat{f}]^2 - 2fE[\hat{f}]$$
$$E[(y - \hat{f})^2] = Var[y] + Var[\hat{f}] + (f - E[\hat{f}])^2$$
$$E[(y - \hat{f})^2] = Var[y] + Var[\hat{f}] + E[(f - \hat{f})]^2$$
$$E[(y - \hat{f})^2] = Var[\varepsilon] + Var[\hat{f}] + Biais[\hat{f}]^2$$

Ce livre ce concentre sur les techniques d'estimations de f avec le but de minimiser l'erreur reductible.

b) inférence

On est souvent interessé à comprendre le fait que Y soit affecté par X1,..,Xp lorsuqil change.
Dans cette situation on veut estimer f mais notre objectif n'est pas necessairement de faire des predicions de Y. Au lieu de cela on veut comprendre la relation entre X et Y et plus scpeicifiquement comrpendre comment Y changes en fonction de X1,..,Xp. Maintenant $\hat{f}$ ne peut etre considéré comme une boite noire.
On s'interessera a differentes questions:
Quel prédicteur est associé a la reponse ?
Quel est la relation entre la reponse et chaque prédicteur ?
Est ce que la relation entre Y et chaque prédicteur puisse etre résumée en utilisant une equation linéaire, ou plus compliquée ?

Suivant la complexité du modele on aura un compromis a avoir (plus le modèle est compliqué, moins on peut effectuer d'infference) et inversement (plus le modele est simple plus on peut effectuer des inferences).
Le modèle linéaire par exemple est relativement simple et interretable mais la precision des predictions peut etre faible.

Tout au long de ce livre, on etudiera des modele lineaires et non-lineaire pour estimer f.
On supposera toujours que nous observons une jeu de n données differentes

Notre but est d'aplciquer les méthodes statistiques au jeu d'entrainement pour estimer la fonction inconnue f. En d'auters termes on veut trouver une fonction $\hat{f}$ comme $Y \approx \hat{f}(X)$ pour toute observation (X,Y).

La plupart des modeles d'apprentissage statistique pour cette tâche peuvent etre caractérisées comme parametrique ou non-parametrique.

a) les métodes paramétriques
Les méthodes paramétriques sont basés sur deux etapes.
1-/ premièrement, on fait une supposition de la forme et de la dimension de la fonction f.
Par exemple, une très simple supposition est que f est linéraire en X:

$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p$$

Ceci est un modèle linéraire que l'on etudiera plus dans le chapitre 3
Une fois les suppositions faites, au lieu d'avoir a estimer une fonction entière de dimension p, on peut se retouver simplement avec quelques coefficients (par exmple p+1 coefficient pour le modèle linéaire).
2-/Après que le modèle soit selectionné, on a besoin d'une procedure qui utilise les données d'entrainment pour adapter le modèle. Pour l'exemple du modèle linéaire, on utilise souvent les moindres carrés (a voir dans le chapitre 3) . Les moindres carrés est une des multitudes de méthodes pour ajuster le modèle linéaire.

L'approche du modele décrit ci-dessus est considéré comme paramétrique: on reduit le probleme de l'esitmation de f en estimant un certain nombre de parametres.
Le potnetiel désavantage d'une approche paramétrique est que le modèle choisit ne corresponde pas avec la vraie fonction inconnue f.
Si notre modèle est trop loin de la réalité nos estimation seront mauvaises.
On pourra essayer d'essayer des modèles flexible qui peuvent ajuster sur differents modèles possible de f.
Mais en général l'ajustement d'un modèle plus flexible requièrt d'estimer plus de paramètres. Ces modèles plus flexible peuvent amener à du surajustement.
Avec un trop petit nombre de données, l'ajustement linéaire peut s'avérer être le plus pratique et le plus proche.

b) les méthodes non-paramétriques
les méthodes non-paramétriques ne créent pas des supposition explicites sur la forme de la fonction f.
Au lieu de cela, elles recherchent une estimation de f qui devient aussi proche des données que possible sat être trop rugueux ou ondulé. Ce genre d'approche peut avoir un avatage majeure sur les approches paramétriques en évitant une supposition d'une fome de fonction de f particulière, elles ont le potentiel de précision plus large que possible des dimension de f.
Mais l''approche non paramétrique souffre d'un désavantage majeur: elle ne réduise pas le problème d'estimation de f à un petit nombre de paramètre . Un plus grand nombre de données est requise dans le but d'obtenir une estimation précise de f.
Pour obtenir d'obtenir un ajustement proche de f, on doit utiliser un plus petit niveau de souplesse.

Comme nous avons pu le voir, il y a des avantages et des inconvénients des deux cotés d'approche parramétrique et non-paramérique. On explorera les deux méthodes dans ce livre.

Comme nous verrons dans ce livre, nous examinerons que certaines sont moins flexibles ou plus restrictives. Par exemple, la regression linéaire est relativement inflexible, parce que elle peut générée uniquement des fonctions linéraires.

Voici un tableau de certaines techniques de l'interprétation possible en fonction de la flexibilité.

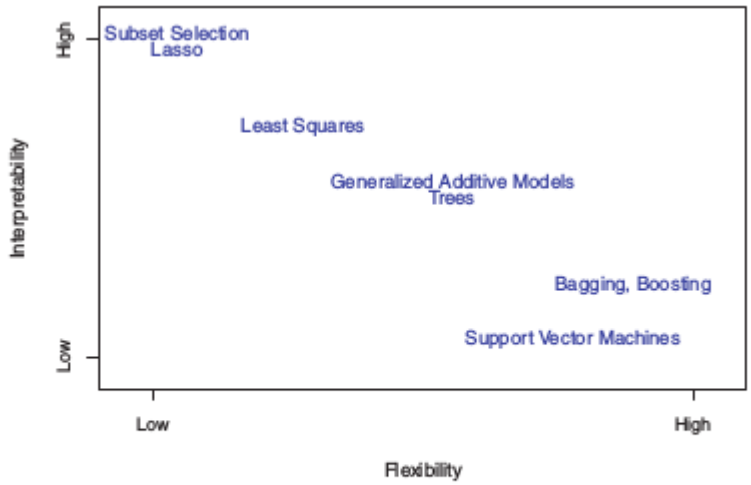


FIGURE 2.7. *A representation of the tradeoff between flexibility and interpretability, using different statistical learning methods. In general, as the flexibility of a method increases, its interpretability decreases.*

D'autres méthodes, comme thin-plate spline sont considérableement plus flexibles car elle génèrent un plus large champ de possibilités pour estimer f.
Pourquoi choisirerions nous d'utiliser des méthodes plus restricitves au lieu d'approches plus flexibles ?
Il y a plusieurs raisons qui nous ferais preferer un modèleplus restrictif. Si nous somme interessés par l'inférence, alors les modèles restrictifs sont plus interpretables. Par example quand inférer est le but ,le modèle linéaire peut être un bon choix puisqu'il serait plus facile de comprendre kes relations entre Y et X1,..,Xp . D'une autre manière, des approches plus flexibles, comme les splines et les methodes de boosting peuvent amener a des estimation compliquées de f qui serait difficile de comprendre.

On a donc établie que quand l'inférence est le but, il y a un avantage clair d'uliser des simples et relativement inflexibles methodes d'apprentissage statistiques. Dans certains cas nous sommes uniquement interssés par la prediction et l'interprétabilité du modèle prédictif n'est pas un enjeux.

La plupart des probleme d'apprentissage statistique tobe dans l'une des deux catégories: apprentissage supervisé ou non-supervisé. L'apprentissage no-supervisé décrit une situation plus difiicile qui pour chaque observation i=1,...,n on constate une mesure de vecteur xi mais non associé à une reponse yi. Il ne devient pas possible de faire des ajustements linéaire donc il n'y a pas de variable à prédire. Dans un crtain sens on voyage à l'aveugle. On peut tenter de comprendre les relations entre les variables ou entre les observations. L'une des techniques les plus utilisées et le cluster analysis, ou encore le clustering. Le but du clustering est de vérifier sur la bases des x1,...,x, si les observations créens des groupes distincts. On ne peut pas s'attendre à qu'une méthode de clustering assigne tout les points au bon groupe. Par exemple si nous avons p varaibles dans nos données alors p(p-1).2 différents nuages de points peuvent etre réalisés. Nous ne pouvons donc pas observer dans pluseurs dimensions autant de groupes. Donc des méthodes automatisées de clustering sont importants. Beaucoup de problèmes tombe directement dans une des deux cases apprentissage supervisé ou non-supervisé. Mais des fois ce n'est pas clair. Par exemple supposons que nous avons n observations. Pour m des observations où m < n, on a tout autant des mesures de prédictions et des reponses de mesures. Pour le reste de n-m observations, on a une mesure de prédiciton mais pas de reponse de mesure. On se réfère a ce genre de probleme un problème d'apprentissage semi-supervisé. Dans ce cas nous souhaitons utiliser une méthode d'apprentissage statistique qui peut prendre en compte les m observations qui pour chaue reponse de mesure soit possible autant que les n-m observations qui ne le sont pas.

Les problèmes de regression contre ceux de classification
Les variables peuvent être caractérisées soit quantitatives soit qualitative (aussi connues sous le nom de categoriques). Les variables quantitatives prennent des valeurs umériques. Les variables qualitatives prennent des valeurs des K différentes classes, ou catégories.

On se réfère a des problèmes de régressions pour des problème de réponse quantitatives et a des problèmes de classifications pour des problèmes de reponse qualitatives.
La distinction n'est pas toujours evidente. La régression linéaires aec les moindres carrées est utilisé pour des reponses quantitatives. Alors que la regression logisitque est typiquement utilisé pour des qualitatives. Qui est aussi souvent utilisé comme une méthode de classification. Mais quand il estime une probabilité de classe, il peut etre pensé comme une methode de regression.
Les k plus proches voisins et le boosting peuvent autant être utiliés pour des reponses quantitatives que qualitatives.
On a tendance a choisir un modele d'apprentissage statistique en fonction de la reponse qualitative ou quantitative. Plutot regression linéaire si quantitatif et regression logisitque quand quaitatif.
Il est souvent peut important de savoir si le prédicteur est qualitatif ou quantitatif.

II Evaluation de la précision du modèle

Pourquoi ne pas présenter les meilleurs méthodes que celle présentées? Pacequ'aucune mméthode domine les autres sur tout type de données. Choisir la meilleur approche peut s'avérer être l'un des plus grands challenge d'apprentissage statistique
Pour evaluer la performance d'un modèle d'apprentissage statistique sur un jeu de donné, nous avons besoin d'un outil de mesure pour connaitre la performance. Dans les cas de régression, l'outil de mesure le plus utilisé est le Mean Squared Error (MSE)

$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$ où $\hat{f}(x_i)$ est la prédiction que $\hat{f}$ donne une ième observation. La MSE sera petite si la prédiction sera proche de la vrai réponse et sera large si pour un des observations, la valeur prédite diffère significativement.

On effectue le MSE sur les données d'entrainement mais ce qui nous interesse le plus est sur les données de test. Donc on s'interesse aà la précision de la prédiction que nous obtenons quand nous appliquons notre methode à des données de test.

Supposons que nous faisons un ajustement de notre mméthode d'apprentissage statistique sur nos observations d'entrainement {(x1,y1),...,(xn,yn)} et que nous obtenons une estimations de $\hat{f}$. On peut calculer $\hat{f}(x_1),...,\hat{f}(x_n)$. S'ils sont approximativemet égaux à y1,...,yn alos notre MSE d'entrainement donnée par la formule plus haut est petite. Toutefois nous ne sommes plutot pas interessé par $\hat{f}(x_i) \approx y_i$, au lieu de cela on veut connaitre $\hat{f}(x_0)$ est approximativement égale à y0 où (x0,y0) est une observation non déjà fait par l'entrainement.
On veut choisir un modèle qui donne le plus petit test MSE contrairement au plus petit entrainement MSE. En d'autres mot si nous avons un grandn nombre de données de test, on pourrait calculer
$Ave(y_0 - \hat{f}(x_0))^2$ la moyenne du carré de l'erreur de prédiction pour ce jeu de test (x0,y0). Nous voudrions selectionner le modèle pour lequel la moyenne de cette quantité – la test MSE – aussi petit que possible.

Il n'y a pas de garanties que le jeu de test aient le minimum en se basant sur le minimum du jeu d'entrainement.

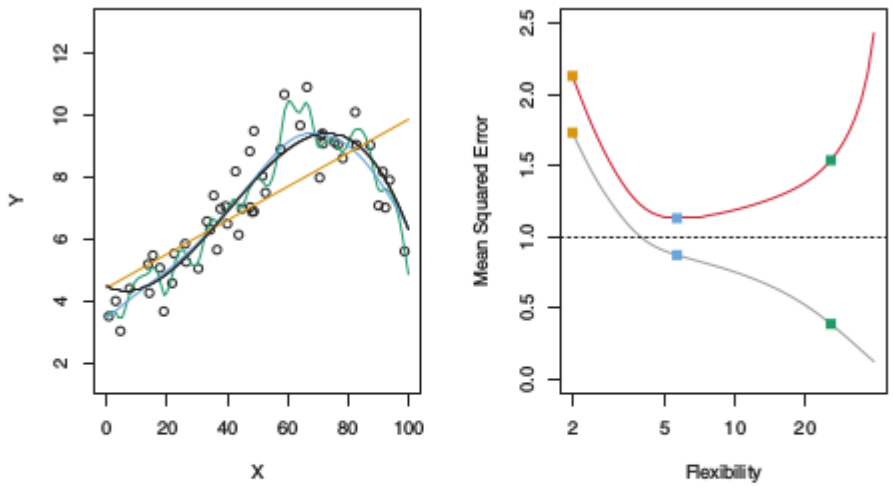


FIGURE 2.9. Left: *Data simulated from f, shown in black. Three estimates of f are shown: the linear regression line (orange curve), and two smoothing spline fits (blue and green curves).* Right: *Training MSE (grey curve), test MSE (red curve), and minimum possible test MSE over all methods (dashed line). Squares represent the training and test MSEs for the three fits shown in the left-hand panel.*

La ligne horizontale indique Var(ε)

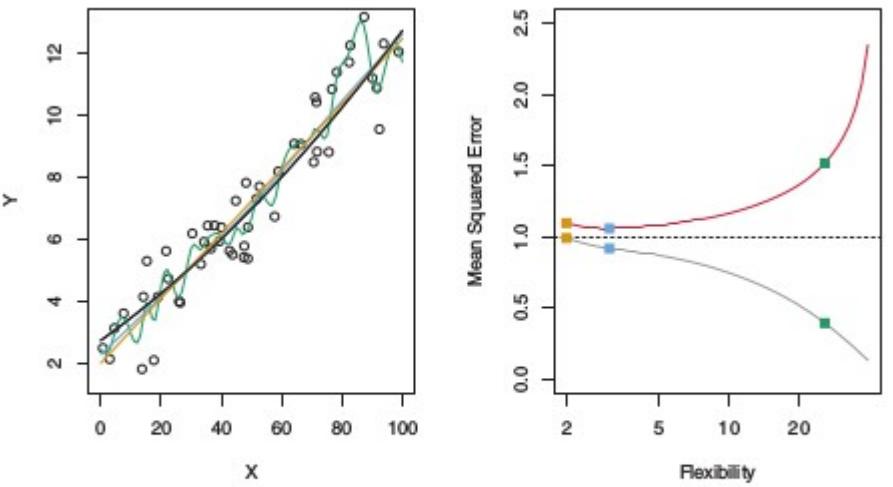


FIGURE 2.10. *Details are as in Figure 2.9, using a different true f that is much closer to linear. In this setting, linear regression provides a very good fit to the data.*

En pratique, on peut habituellemnt utiliser le MSE entrainement avec relaitvement d'aisance. Mais estimer le MSE test est considérablement plus copliqué car souvent il n'y a pas de jeu de test. Le nivea de flexibilité correspondant au modèle vec le test MSE peut considérablement varier en pratique (en fonction des données). Dans ce livre, nous allons etudier une variété d'approches qui peuvent être utilisées en pratique pour estimer ce point minimum.
Une technique importante est la cross-validation, qui set une méthode pour estimer le test MSE en utilisant uniquement le jeu d'entrainement.

Le dilemme biais-variance
Le test MSE pour un valeur x0 donnée peut toujours être décomposé en un seom de trois qantités fondamentales:
-la variane de $\hat{f}(x_0)$
-le carré du biais $\hat{f}(x_0)$
-et la varaince de l'erreur du terme ε

$$E(y_0-\hat{f}(x_0))^2=Var(\hat{f}(x_0))+[Bias(\hat{f}(x_0))]^2+Var(\varepsilon)$$

la notation $E(y_0-\hat{f}(x_0))^2$ est la valeur attendue du test MSE, et se réfère à la moyenne du test MSE que l'on obtiendrait si nous répetons l'estimation de f en utilisant un grand nombre de jeu d'entrainement et testé pour x0.
L'équation ci-dessus nous indique que pour minimiser l'erreur de test attendue, nous avons besoin de selectionner une modèle qui accomplie la plus petite variance et le plus petit biais. A noter que la variance et le biais au carré sont deux valeur strictement positives. Nous voyons donc que le test MSE ne peut jamais être en dessous de Var(ε) (se nommant l'erreur irreductible)

Que signifie le biais et la variance d'un modèle ?
La variance fait référence au montant que $\hat{f}$ changerait si nous changions de jeu de données. Alors que le jeu d'entrainement est utilisé pour faitre l'ajustement du modèle, différent jeu de données donnera different $\hat{f}$. Si un modèle à une haute variance alors de petits changements dans les données donneront de grands changements de $\hat{f}$. En général les modèles les plus flexibles ont une variance élevée.On peut voir sur le figure 2,9 que la courbe verte suit les observations de manière très proche. Il a donc une haute variance car changer le jeu de donnés causera l'estimation de $\hat{f}$ à changer considérablement. A l'inverse, la courbe orange est relativement inflexible et a une faible variance, car bouger n'importe quelle observation ne causera uniquement qu'un petit décalage de la position de la ligne.
Le biais fait référence à l'erreur introduite par l'approximation du reél problème, qui peut être extrêmement compliqué par rapport à un modèle simple. Par exemple une approche de regression linéaire sur un problème où la fonction réelle est plus compliqué que linéaire induira un haut biais. Généralement plus le modèle est flexible moins il a de biais.
De manière général, plus nous utilisons des modèles flexibles, plus la variance augmentera et plus le biais diminuera.
Le *biais* est l'erreur provenant d'hypothèses erronées dans l'algorithme d'apprentissage. Un biais élevé peut être lié à un algorithme qui manque de relations pertinentes entre les données en entrée et les sorties prévues (sous-apprentissage).
La *variance* est l'erreur due à la sensibilité aux petites fluctuations de l'échantillon d'apprentissage. Une variance élevée peut entraîner un surapprentissage, c'est-à-dire modéliser le bruit aléatoire des données d'apprentissage plutôt que les sorties prévues.
Donc au début le test MSE est trop sous-ajusté puis petit a petit en prenant un modèle plus flexible on voit le test MSE diminuer dû au biais qui diminue. Puis arrivé a un minimum, le test MSE réaugmente dû au sur-aprentissage qui est causé par l'augmentation de la variance.
VARIANCE élévée = surapprentissage
BIAIS élevé = sousappprentissage

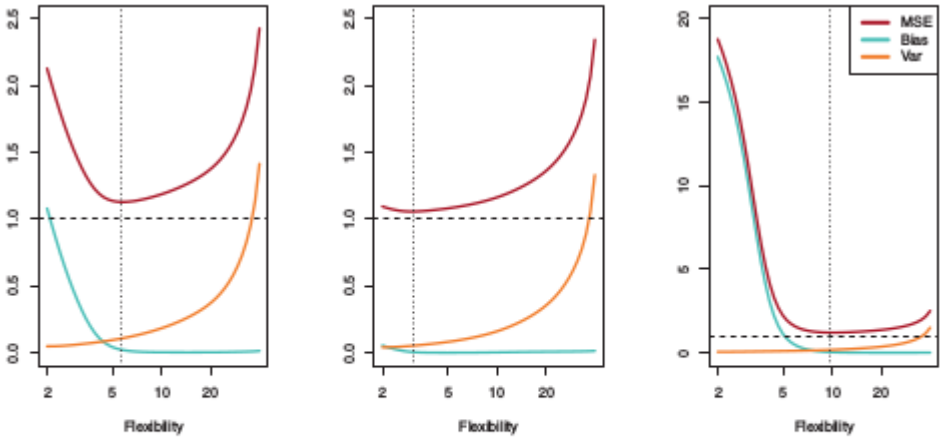


FIGURE 2.12. *Squared bias (blue curve), variance (orange curve), Var(ϵ) (dashed line), and test MSE (red curve) for the three data sets in Figures 2.9–2.11. The vertical dotted line indicates the flexibility level corresponding to the smallest test MSE.*

Les trois graphiques de la figure 2.12 illustrent l'équation du test MSE (voir formule séparée en plusieurs termes). On peut voir que le test MSE diffère des trois cas car le carré du biais et la variance change selon le jeu de données. Le deuxieme graphique montre que f est proche d'une fonction linéaire. Pour le troisième on est clairement dans du non-linéaire.

$$Biais[\hat{f}(x)] = E[\hat{f}(x) - f(x)]$$

$$Var[\hat{f}(x)] = E[(\hat{f}(x) - E[\hat{f}(x)])^2]$$

La relation entre le biais, la variance et le test MSE donné par l'équation et affiché par la fichuer 2.12 est connue sous le nom de dilemme biais-variance.
Il est plus facile d'obtenir des modèles avec très peu de biais mais haute variance ou l'inverse, des modèles avec haut biais et peu de variance.
Dans la vrai vie il est rare d'avoir la fonction f. Donc il est rare de pouvoir calculer le test MSE, la vaiance et le biais.
La validation croisée est un bon moyen pour estimer le test MSE avec les données d'entraînement.

On a parlé presque que de regression, maintenant de classification.
Supposons que nous recherchons à estimer f sur un base de jeu de données {(x1,y1),...,(xn,yn)} où y1,...yn sont qualitatives.
La technique la plus utilisée pour quantifier la precision de notre estimation de $\hat{f}$ est l'error rate qui est la proportion d'erreurs faites si nous appliquons notre estimation $\hat{f}$ aux données d'observations:

$\frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i)$ où $\hat{y}_i$ est la classe prédite pour la ième observation en utilsant $\hat{f}$

$I(y_i \neq \hat{y}_i)$ est une variable indicatrice qui est égale à 1 si $y_i \neq \hat{y}_i$ et 0 si $y_i = \hat{y}_i$.

Si $I(y_i \neq \hat{y}_i) = 0$ alors la ième observation a été classifiée correctement par notre modèle de classification. Sinon il est mal classifié. Donc cette équation calcule la fraction de classification incorrectes.
Cette équation est appelée trainning error rate car elle calcule a partir des données d'entrainment. Comme pour la régression, on est plus interessé par l'error rate résultant de données non d'entrainement.
Le test error rate est associé a un jeu de test de la forme (x0,y0) donné par

$Ave(I(y_0 \neq \hat{y}_0))$

Le classifieur de Bayes
Il est possible de montrer que l'équation plus haut est minimisé en moynne par un simple classifieur qui associe chaque observation à sa plus proche classe selon les prédictions. En d'autres mots nous souhaitons simplement assigner un test d'observation avec le predicteur x0 de la classe j pour lequel

$Pr(Y = j | X = x0)$ est grand. C'est une probabilité conditionnelle qui est la probabilité que Y=j selon un predicteur de x0. Ce très simple classifieur est appelé classifieur de Bayes.
Dans un problème à deux classes le classifieur de Bayes correspond à predire la classe 1 si

$Pr(Y = 1 | X = x0) > 0.5$ (2.10)

ou la classe 2 autrement.

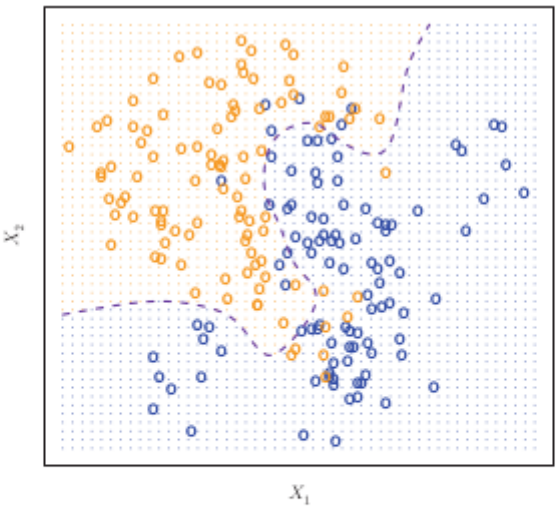


FIGURE 2.13. *A simulated data set consisting of* 100 *observations in each of two groups, indicated in blue and in orange. The purple dashed line represents the Bayes decision boundary. The orange background grid indicates the region in which a test observation will be assigned to the orange class, and the blue background grid indicates the region in which a test observation will be assigned to the blue class.*

La Figure 2.13 montre un exemple de dimension 2 (X1,X2) pour un problème à deux classes. Pour chaque valeur de X1 et X2 il y a differentes probabilités de reponse en étant orange ou bleue.La limite de décicison de Bayes est la ligne en pointillés.
Le classifieur de Bayes produit le plus petit test de taux d'erreur. Puisque le classifieur de Bayes choisira toujours la classe pour laquelle l'équation 2.10 est grand, le taux derreur à X=x0 sera

$1-max_j Pr(Y=j|X=x0)$.

En général l'ereur globale de Bayes est donnée par $1-E(max_j Pr(Y=j|X))$ (2.11)
où l'attente moyenne la probabilité la plus probable de X. L'erreur de Bayes est analogue a l'erreur irréductible discutée plus haut.

Les k-plus-proches voisins
En théorie nous voulons tojours prédire des reponses qualitatives en utilisant le classifieur de Bayes. Mais pour des données réelles, nous ne connaissons par la distribution conditionnelle de Y sachant X, et donc le classifieur de Bayes est impossible. Beaucoup d'approches tentent d'estimer la distribution conditionnelle de Y sachant X, et classifint une observation donnée à la classe dont l'estimation de la probabilité est la plus élevée. Un de ces modèles et le classifieur K plus proches voisins (KNN).
Etant donné un entier K et une observation x0, les KNN identifie les K points dans les données dentrainement qui sont les plus proches de x0, repréenté par No.

$$Pr(Y=j|X=x0)=\frac{1}{K}\sum_{i\in N_0} I(y_i=j)$$

En conclusion, KNN applique les règrle de Bayes et classifie l'observation x0 à tester à la classe ayant la plus grande probabilité.
Malgré le fait que c'est une approche très smple, KNN peut souvent produire des classificateur qui sont proche du classifieur de Bayes optimal.
Le choix de K est un grand effet sur le classifieur KNN obtenu.

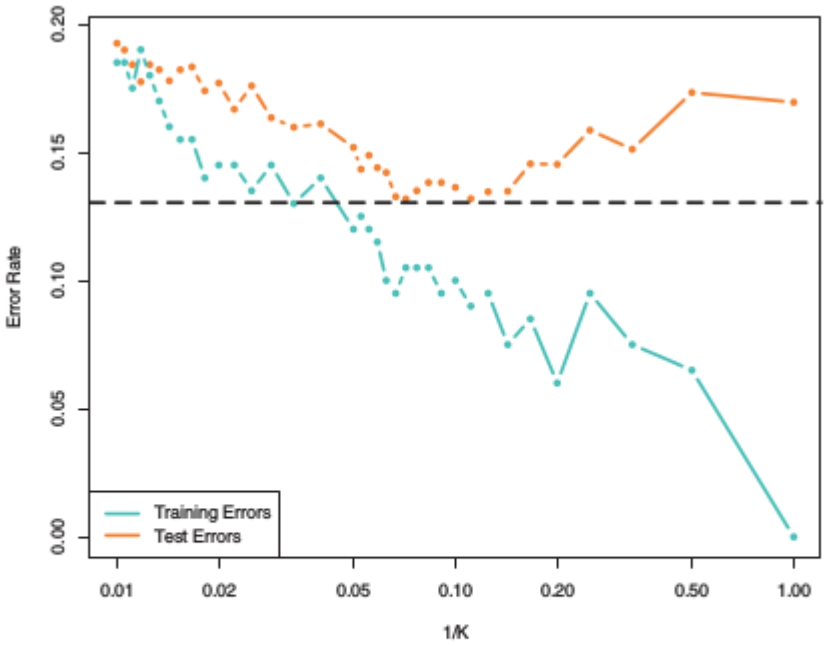


FIGURE 2.17. *The KNN training error rate (blue, 200 observations) and test error rate (orange, 5,000 observations) on the data from Figure 2.13, as the level of flexibility (assessed using $1/K$) increases, or equivalently as the number of neighbors K decreases. The black dashed line indicates the Bayes error rate. The jumpiness of the curves is due to the small size of the training data set.*

Pour K=1, la limite de décision est très flexible et et trouve des schémas dans le sdonnées qui ne correspondent pas a la limite de Bayes.
Quand K augmente , le modèle devient de moins en moins flexible et produit une limite de décision proche

du linéaire : petite variance mais grand biais
Quand 1/K augmente, ;le modèle devient plus flexible. Courbe en forme de U.
Dans les deux cas de classification et de régression, choisir le niveau de flexibilité est critique pour le succes de n'importe quel modèle statistique.
Le dilemme biais-variance et les resultat d'une courbe en U pour l'erreur de test, peut devenir une tache difficile.

Chapitre 3 – Régression linéaire

Supposons la relation linéaire $Y \approx \beta_0 + \beta_1 X$ (3.1)
Une fois que nous avons utilisé nos données d'entraînement pour estimer $\hat{\beta}_0$ et $\hat{\beta}_1$ nous pouvons faire des prédicitons $\hat{y} \approx \hat{\beta}_0 + \hat{\beta}_1 x$ où $\hat{y}$ indique la prédiction de Y sur la base de X=x.
Le chapeau ^ designe la valeur estimée pour un paramètre inconnu ou une prediction de la reponse.
Le plus commun est d'utiliser le critère des moindres carrés.
Posons $\hat{y}_i \approx \hat{\beta}_0 + \hat{\beta}_1 x_i$ la prédiction de Y vasée sur la ième valeur de X.
Donc $e_i = y_i - \hat{y}_i$ représente le ième résidu.
Nous définissons les résidues de sum de squares (RSS)

$RSS = e_1^2 + e_2^2 + ... + e_n^2$

$RSS = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + ... + (y_1 - \hat{\beta}_0 - \hat{\beta}_n x_n)^2$

L'approche par les moindres carrés choisit $\hat{\beta}_0$ et $\hat{\beta}_1$ pour minimiser le RSS .
En faisant quelques calculs, on peut montrer que les minimiseurs sont

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \quad (3.4.a)$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad (3.4.b)$$

où $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ et $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ qui sont les moyennes
les équations (3.4) definissents les estimations des coefficient des moindres carrés pour une régression linéaire

démonstration
On cherche RSS tel que

$$f(\hat{\beta}_0, \hat{\beta}_1) = min_{\hat{\beta}_0, \hat{\beta}_1} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

on atteint le minimum pour $\frac{\partial f(\hat{\beta}_0, \hat{\beta}_1)}{\partial \beta_k} = 0$

donc $-2\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i = 0$ et $-2\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$

donc $\sum_{i=1}^{n} (y_i - \hat{\beta}_1 x_i) x_i = n\bar{x}\hat{\beta}_0$ et $\bar{y} - \hat{\beta}_0 - \hat{\beta}_1 \bar{x} = 0$

donc $\sum_{i=1}^{n} (y_i x_i) - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = n\bar{x}(\bar{y} - \hat{\beta}_1 \bar{x})$ et $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

donc $\sum_{i=1}^{n} (y_i x_i) - \hat{\beta}_1 (\sum_{i=1}^{n} x_i^2 - n\bar{x}^2) = n\bar{x}\bar{y}$ et $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

donc $\sum_{i=1}^{n} (y_i x_i) - n\bar{x}\bar{y} - \hat{\beta}_1 (\sum_{i=1}^{n} x_i^2 - \bar{x}^2) = 0$ et $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

or $E[(X - E[X])^2] = E[X^2] - E[X]^2$ et $E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y]$

donc $\hat{\beta}_1 = \dfrac{\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$ et $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

On peut écrire $\hat{\beta}_1 = \dfrac{\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \dfrac{\sum_{i=1}^{n}(y_i-\bar{y})(x_i)}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \dfrac{\sum_{i=1}^{n}(y_i)(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$ (voir cela en développant)

On suppose que la relation entre X et Y prend la forme $Y = f(X) + \varepsilon$ pour une fonction f inconnue . Si f est approximé par une fonction linéaire, alors on peut écrire $Y = \beta_0 + \beta_1 X + \varepsilon$ (3.5)
On suppose que le terme d'erreur est indépendant de X.

Avec les coefficients de (3.4) on peut trace la ligne des moindres carrés.

Sur un grand jeux de données nous pouvons considérer que $\hat{\beta}_0 et \hat{\beta}_1$ seront égales exactement à $\beta_0 et \beta_1$
Par exemple pour verifier la précision de $\hat{\mu}$ étant l'estimation de la moyenne μ sur un jeu de données, on peut étudier l'erreur standard de $\hat{\mu}$ ecrite $SE(\hat{\mu})$

$Var(\hat{\mu}) = SE(\hat{\mu})^2 = \dfrac{\sigma^2}{n}$ (3.7) où σ est la déviation standard de chaque realisations de yi de Y.

Démonstration de $Var(\hat{\mu}) = \dfrac{\sigma^2}{n}$:

On a $E(\bar{\mu}) = E[\frac{1}{n}\sum_{i=1¿}^{n}\mu_i] = \frac{1}{n}\sum_{i=1}^{n}E[\mu_i] = \frac{n*\mu}{n} = \mu$ car on a $\bar{\mu} = \frac{1}{n}\sum_{i=1}^{n}\mu_i$

$Var(\hat{\mu}) = Var(\bar{y}) = E[(\bar{y} - E[\bar{y}])^2]$

$Var(\hat{\mu}) = E[(\frac{1}{n}\sum_{i=1}^{n}y_i - E[\frac{1}{n}\sum_{i=1}^{n}y_i])^2]$

$Var(\hat{\mu}) = \frac{1}{n^2}E[(\sum_{i=1}^{n}y_i - E[y_i])^2]$ on pose $a_i = y_i - E[y_i]$

$Var(\hat{\mu}) = \frac{1}{n^2}E[(\sum_{i=1}^{n}a_i)^2]$

$Var(\hat{\mu}) = \frac{1}{n^2}E[(\sum_{i=1}^{n}a_i)(\sum_{j=1}^{n}a_j)]$

$Var(\hat{\mu}) = \frac{1}{n^2}E[\sum_{i=1}^{n}a_i^2 + 2\sum_{1\le i<j\le n}a_i a_j]$

$Var(\hat{\mu}) = \frac{1}{n^2}(\sum_{i=1}^{n}E[a_i^2] + 2\sum_{1\le i<j\le n}E[a_i a_j])$

$Var(\hat{\mu}) = \frac{1}{n^2}(\sum_{i=1}^{n}E[a_i^2] + 2\sum_{1\le i<j\le n}E[a_i a_j])$ or $E[a_i] = E[y_i - E[y_i]] = E[y_i] - E[y_i] = 0$

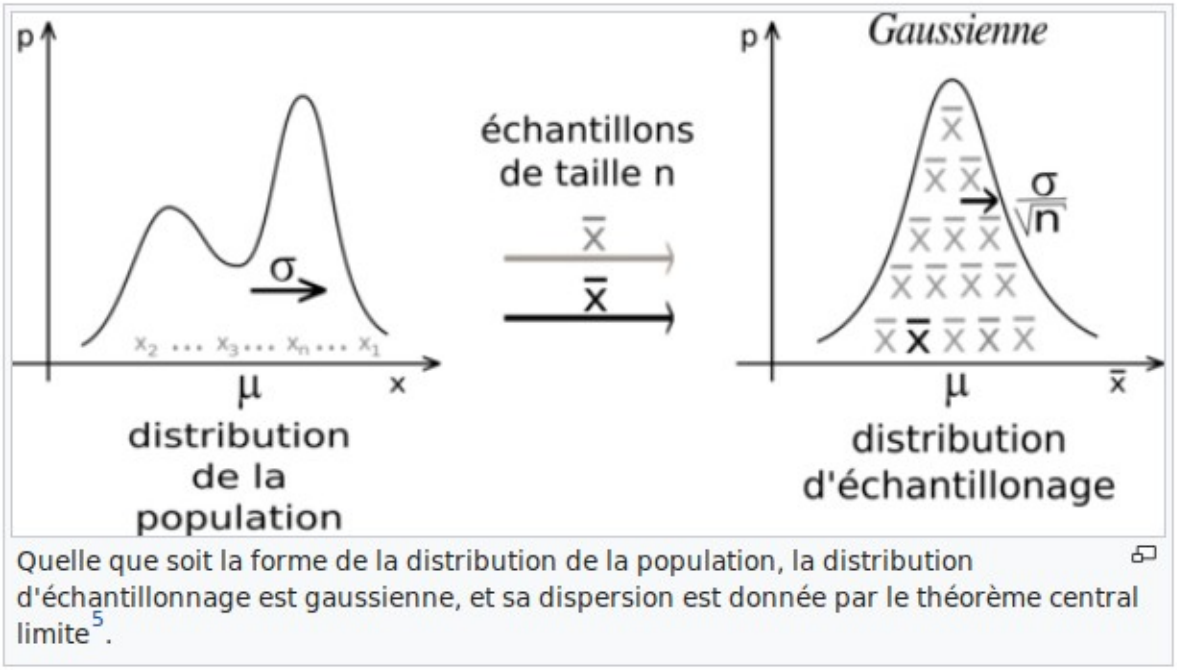
$$Var(\hat{\mu}) = \frac{1}{n^2}\sum_{i=1}^{n} E[(y_i - E[y_i])^2]$$

$$Var(\hat{\mu}) = \frac{1}{n} E[\frac{1}{n}\sum_{i=1}^{n} (y_i - E[y_i])^2]$$

$$Var(\hat{\mu}) = \frac{1}{n} E[\sigma^2]$$

$$Var(\hat{\mu}) = \frac{\sigma^2}{n}$$

On peut le voir graphiquement avec le théorème de la limite centrale



Quelle que soit la forme de la distribution de la population, la distribution d'échantillonnage est gaussienne, et sa dispersion est donnée par le théorème central limite[5].

$$cov(y_i, y_j) = \frac{1}{n^2}\sum_{i=1}^{n} (y_i - E[y_i])\sum_{j=1}^{n} (y_j - E[y_j]) = \frac{1}{n^2}\sum_{i=1}^{n} (\varepsilon_i - E[\varepsilon - i])\sum_{j=1}^{n} (\varepsilon_j - E[\varepsilon_j]) = cov(\varepsilon_i, \varepsilon_j)$$

$$cov(y_i, y_j) = 0 \; si \; j \neq i \; sinon \; var(\varepsilon_i, \varepsilon_i) = var(y_i, y_j) = \sigma^2$$

L'équation 3.7 nous dit comment la deviation diminue avec n. Plus on a d'observations, plus l'erreur standard devient petite. De manière similaire on peut désirer combient $\hat{\beta}_0 et \hat{\beta}_1$ sont proches des valeurs $\beta_0 et \beta_1$. Pour calculer l'erreur standard associé à $\hat{\beta}_0 et \hat{\beta}_1$ on utilise les formules suivantes

$$SE(\hat{\beta}_0)^2 = \sigma^2[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}] \quad \text{et} \quad SE(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

où $\sigma^2 = Var(\varepsilon)$

Demonstration:

$$\hat{\beta}_1=\frac{\sum_{i=1}^{n}(x_i-\bar{x})\,y_i}{\sum_{i=1}^{n}(x_i-\bar{x})^2}=\sum_{i=1}^{n}w_i Y_i \quad \text{avec} \quad w_i=\frac{x_i-\bar{x}}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

de même $\hat{\beta}_0=\bar{y}-\hat{\beta}_1\bar{x}=\sum_{i=1}^{n}(\frac{1}{n}-\bar{x}\,w_i)\,y_i$

on a $$\sum_{i=1}^{n}w_i=\sum_{i=1}^{n}\frac{x_i-\bar{x}}{\sum_{i=1}^{n}(x_i-\bar{x})^2}=0$$

on a $$\sum_{i=1}^{n}w_i x_i=\sum_{i=1}^{n}\frac{(x_i-\bar{x})\,x_i}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sum_{i=1}^{n}\frac{(x_i-\bar{x})(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}=1$$

on a $$\sum_{i=1}^{n}w_i^2=\sum_{i=1}^{n}\frac{(x_i-\bar{x})^2}{(\sum_{i=1}^{n}(x_i-\bar{x})^2)^2}=\frac{1}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

donc $$E[\hat{\beta}_1]=E[\sum_{i=1}^{n}w_i\,y_i]=E[\sum_{i=1}^{n}w_i(\beta_0+\hat{\beta}_1 x_i)]=\beta_1$$

$$E[\hat{\beta}_0]=E[\bar{y}-\hat{\beta}_1\bar{x}]=E[\beta_0\bar{y}+\beta_1\bar{x}+\bar{\varepsilon}]-\beta_1\bar{x}=\beta_0$$

$$Var(\hat{\beta}_0)=\sum_{i=1}^{n}(\frac{1}{n^2}-2\frac{\bar{x}\,w_i}{n}+\bar{x}^2 w_i^2)\sigma^2=(\frac{1}{n}-0+\bar{x}^2\sum_{i=1}^{n}w_i^2)\sigma^2=(\frac{1}{n}+\frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2})\sigma^2$$

$$Var(\hat{\beta}_1)=\sum_{i=1}^{n}w_i^2\sigma^2=\frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

(REFAIRE LES CALCULS A LA MAIN)

$$Cov(\hat{\beta}_1,\hat{\beta}_0)=Cov(\sum_{i=1}^{n}(\frac{1}{n}-\bar{x}\,w_i)\,y_i,\sum_{i=1}^{n}w_i\,y_i)=\sum_{i=1}^{n}\sum_{j=1}^{n}(\frac{1}{n}-\bar{x}\,w_i)\,w_j\,cov(y_i,y_j)$$

$$Cov(\hat{\beta}_1,\hat{\beta}_0)=\sum_{i=1}^{n}(\frac{1}{n}-\bar{x}\,w_i)w_i\sigma^2=\frac{-\bar{x}}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sigma^2$$

En général σ^2 n'est pas connu, mais il peut être estimé depuis les données.
L'estimation de σ est connu sous le nom de résidu d'erreur standard et est donné par la formule
$RSE=\sqrt{RSS/(n-2)}$ **(ZT.35)**

démonstration:

on appelle résidus les quantités $\hat{\varepsilon}_i = Y_i - \hat{Y}_i$ $\sum_{i=1}^{n} \hat{\varepsilon}_i = 0 \, car \, E[\hat{\varepsilon}_i] = 0$

estimer la variance σ^2 par $\frac{1}{n}\sum_{i=1}^{n} \hat{\varepsilon}_i^{\,2}$ mais c'est un estimateur biaisé on prendra donc

$\sigma^2 = \frac{1}{n-2}\sum_{i=1}^{n} \hat{\varepsilon}_i^{\,2}$ qui est un estimateur non biaisé

$$\hat{\varepsilon}_i = Y_i - \hat{Y}_i$$
$$\hat{\varepsilon}_i = \beta_0 + \beta_1 x_i + \varepsilon_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$
$$\hat{\varepsilon}_i = \bar{Y} - \beta_1 \bar{x} - \bar{\varepsilon} + \beta_1 x_i + \varepsilon_i - \bar{Y} + \hat{\beta}_1 \bar{x} - \hat{\beta}_1 x_i$$
$$\hat{\varepsilon}_i = (\varepsilon_i - \bar{\varepsilon}) + (\beta_1 - \hat{\beta}_1)(x_i - \bar{x})$$

donc $\hat{\varepsilon}_i^{\,2} = (\varepsilon_i - \bar{\varepsilon})^2 + 2(\varepsilon_i - \bar{\varepsilon})(\beta_1 - \hat{\beta}_1)(x_i - \bar{x}) + (\beta_1 - \hat{\beta}_1)^2 (x_i - \bar{x})^2$

or

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) y_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(\beta_0 + \beta_1 x_i + \varepsilon_i)}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \beta_1 + \frac{\sum_{i=1}^{n} (x_i - \bar{x})(\varepsilon_i - \bar{\varepsilon})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

donc $\sum_{i=1}^{n} \hat{\varepsilon}_i^{\,2} = \sum_{i=1}^{n} (\varepsilon_i - \bar{\varepsilon})^2 + \sum_{i=1}^{n} 2(\varepsilon_i - \bar{\varepsilon})(\beta_1 - \hat{\beta}_1)(x_i - \bar{x}) + \sum_{i=1}^{n} (\beta_1 - \hat{\beta}_1)^2 (x_i - \bar{x})^2$

$$\sum_{i=1}^{n} \hat{\varepsilon}_i^{\,2} = \sum_{i=1}^{n} (\varepsilon_i - \bar{\varepsilon})^2 + -2\sum_{i=1}^{n} (\beta_1 - \hat{\beta}_1)^2 (x_i - \bar{x})^2 + \sum_{i=1}^{n} (\beta_1 - \hat{\beta}_1)^2 (x_i - \bar{x})^2$$
$$\sum_{i=1}^{n} \hat{\varepsilon}_i^{\,2} = \sum_{i=1}^{n} (\varepsilon_i - \bar{\varepsilon})^2 - \sum_{i=1}^{n} (\beta_1 - \hat{\beta}_1)^2 (x_i - \bar{x})^2$$
$$E[\sum_{i=1}^{n} \hat{\varepsilon}_i^{\,2}] = E[\sum_{i=1}^{n} (\varepsilon_i - \bar{\varepsilon})^2] - E[\sum_{i=1}^{n} (\beta_1 - \hat{\beta}_1)^2 (x_i - \bar{x})^2]$$

or

$$E[\sum_{i=1}^{n} (\varepsilon_i - \bar{\varepsilon})^2] = E[-n\bar{\varepsilon}^2 + \sum_{i=1}^{n} (\varepsilon_i^2)] = \frac{-n}{n^2} E[(\sum_{i=1}^{n} \varepsilon_i)^2] + n\sigma^2 = n\sigma^2 - \frac{1}{n}\sum_{i=1}^{n} E[\varepsilon_i^2] - 2\sum_{1 \le i < j \le n} E[\varepsilon_i \varepsilon_j]$$

$$E[\sum_{i=1}^{n} (\varepsilon_i - \bar{\varepsilon})^2] = (n-1)\sigma^2 \quad \textbf{(A.64)}$$

et

$$E[\sum_{i=1}^{n} (\beta_1 - \hat{\beta}_1)^2 (x_i - \bar{x})^2] = E[(\beta_1 - \hat{\beta}_1)^2] E[\sum_{i=1}^{n} (x_i - \bar{x})^2] \quad \textbf{(APK.45)}$$

$$E[\sum_{i=1}^{n} (\beta_1 - \hat{\beta}_1)^2 (x_i - \bar{x})^2] = var(\hat{\beta}_1) E[\sum_{i=1}^{n} (x_i - \bar{x})^2] = \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} E[\sum_{i=1}^{n} (x_i - \bar{x})^2] = \sigma^2$$

donc $E[\sum_{i=1}^{n} \hat{\varepsilon}_i^{\,2}] = (n-2)\sigma^2$

Pour les régression linéaires, il y a approximativement 95% de chances que l'intervale $[\hat{\beta}_1 - 2.SE(\beta_1), \hat{\beta}_1 + 2.SE(\beta_1)]$ contiennent la vraie valeur de β_1 . Idem pour β_0 .

Les erreurs standards peuvent aussi être utilisées pour permettre des tests d'hypothèses sur les coefficients.
La plus connue est l'hypothèse nulle
H_0 : il n'y a pas de relation entre X et Y
contre l'hypothèse alternative
H_a : il y a une relation entre X et Y

Mathématiquement cela correspond à tester
H_0 : $\beta_1 = 0$
contre
H_0 : $\beta_1 \neq 0$

si $\beta_1 = 0$ alors le modèle $Y = \beta_0 + \beta_1 X + \varepsilon$ se réduit à $Y = \beta_0 + \varepsilon$ et X est non associé à Y. Pour tester l'hypothèse nulle, nous avons besoin de déterminer si $\hat{\beta}_1$ est suffisamment loin de zéro pour considérer que β_1 est non nulle. De combien ? Tout dépend de la précision de $\hat{\beta}_1$ qui dépend de $SE(\hat{\beta}_1)$.

En pratique on calcule une t-statistique $t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$ (3.14) qui mesure le nombre de déviation standard que $\hat{\beta}_1$ est loin de 0 .

S'il n'y a pas de relation entre Y et X alors on s'attendra que (3.14) aura une t-distribution de n-2 degrés de liberté.

La t-distribution a une forme en cloche et pour des valeurs de n plus grand approximativement que 30 il est quasi similaire à la distribution normale. En conséquence, il est simple de calculer la probabilité d'observer n'importe quel nombre egale a |t| ou plus grand en valeur absolue, en supposition de $\beta_1 = 0$. On appelle cette probabilité la p-valeur ($p = p(x|H_0)$). On interprete la p-valeur comme suit: une petite p-valeur indique qu'il est peut probable d'observer une association entre le prédicteur et la réponse, en l'abscence de n'importe quelle raison l'association entre le prédicteur et la reponse.
Si t est petit alors on est proche de la moyenne nulle, donc toute probabilité p(Y>|t|) doit etre grande pour ne pas invalider l'hypothèse nulle.
Par conséquent, si nous observons une petite p-valeur alors on peut inférer qu'il y a une association entre le prédicteur et la réponse. Nous rejetons l'hypothèse nulle, nous déclarons qu'il existe donc une relation entre X et Y- bein sur si la p-valeur est suffisamment petite. De manière général nousrejetons l'hypothèse nulle pour moins de 5 ou 1%. Quand n=30, cela correspond a une t-statistique autour de 2 et 2,75 respectivement.
Par exemple:

	Coefficient	Std. error	t-statistic	p-value
Intercept	7.0325	0.4578	15.36	< 0.0001
TV	0.0475	0.0027	17.67	< 0.0001

TABLE 3.1. *For the* Advertising *data, coefficients of the least squares model for the regression of number of units sold on TV advertising budget. An increase of* \$1,000 *in the TV advertising budget is associated with an increase in sales by around 50 units (Recall that the* sales *variable is in thousands of units, and the* TV *variable is in thousands of dollars).*

A noter que les coefficients sont très grands devant les erreurs standards. Donc la t-statistique est grande. Donc la p-valeur est très petite. Ce qui nous conduit à dire que la probabilité de voir ce genre de valeur si H_0 est vraie est virtuellement faux. Par conséquent on peut conclure que $\beta_1 \neq 0$ et $\beta_0 \neq 0$

Une fois que nous avons rejeté l'hypothèse nulle (3.12) en faveur de l'hypothèse alternative, iil est naturel de vouloir quantifier dans quelle mesure le modèle correspond aux données.
La qualité d'une régression linéaire est évalué par l'utilisation de deux valeurs: la résidual standard error (RSE) et le R^2-statistique.

L'erreur standard des résidus
un rappel du modèle $Y = \beta_0 + \beta_1 X + \varepsilon$ qui est associé avec chaque observation un terme d'erreur ε. A cause de la présence de ces terme d'erreur, meme si nous svaions la vai ligne de régressionn nous ne pourrions pas etre capable de prédire parfaitement Y a partir de X. La RSE est une estimation de la déviation standard de ε. C'est la moyenne des deviations des reponses provenatn de la vraie ligne de régression.

$$RSE = \sqrt{\frac{1}{n-2} RSS} = \sqrt{\frac{1}{n-2} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \quad (3.15) \text{ avec} \quad RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad (3.16)$$

Si $\hat{y}_i \approx y_i$ pour i=1,..,n alors (3.15) sera petit et nous pourrons conlure que le modele correspond bien aux données. Par contre , si $\hat{y}_i$ est très loin de yi pour une ou plusieurs observations, alors la RSE sera grnade. Ceci indique que le modèle ne correspond pas aux données.

R² statistique
La RSE fournit une mesure du manque d'ajustement du modèle (3.5) aux données. Mais il n'est jamais clair que la RSE est bonne ou non. La R² statistique fournit une alternative de mesure de l'ajustement. Elle prend la forme d'une proportion et prend ses valeurs entre 0 et 1 et est independante des dimensions de Y (contrairement a la RSE)

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS} \quad \text{où} \quad TSS = \sum (y_i - \bar{y})^2 \quad \text{TSS = Total Sum of Squares}$$

La TSS mesure la variance totale de la réponse Y, et peut être calculée avant la régression.
En revanche, la RSS mesure la quantité qui reste inexpliquée après avoir effectué la régression.
Par conséquent, TSS – RSS mesure la quantité de variabilité de la réponse qui est expliquée (ou supprimée) en effectuant la régression, et R² mesure la proportion de de la variabilité de Y expliquée par X.
Un R² proche de 1 indique indique qu'une grande partie de la variabilité de la réponse a été expliqué par la régression. Un nombre proche de 0 indique que le contraire. Cela peut se produire si le modèle n'est pas bon, ou que σ² est grand ou les deux.
La R² statistique à une interprétabilité meilleur que la RSE car elle est entre 0 et 1.
Quoi qu'il en soit il peut être interesant de savoir si R² est une bonne valeur et en général celaa dépendra de l'applicaiton.
La R² statistique est une mesure de la relation linéaire entre X et Y.

$$Cor(X,Y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}}$$ est aussi une mesure de la relation linéaire entre X et Y.

Cela suppose que nous pourrions être en mesure d'utiliser r=Cor(X,Y) au lieu de R² . En fait, on pourrait montrer que pour une simple régression linéaire $r^2 = R^2$. En d'autres mots, la correlation au carré et la R² statistique sont identiques.

Regression lineaire multiple

En général on a plusieurs paramètres pour la regression linéaire. Ce qu'on fait c'est qu'on ajoute des coefficients

$Y=\beta_0+\beta_1 X_1+\beta_2 X_2+...+\beta_p X_p+\varepsilon$ où X_j représente le jème predicteur et β_j ccc quantifie la'association entre cette variable et la réponse. On interprète β_j comme l'effet moyen sur Y pour une valeur croissante en X_j en gardant les autres prédicteur fixés.

Comme précedemment, les coefficients de regression $\beta_1, \beta_2, ..., \beta_p$ ne sont pas connus et nous devons les estimer. Nous avons donc les estimations $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p$ .On peut faire les predictions avec la formule

$\hat{y}=\hat{\beta}_0+\hat{\beta}_1 x_1+...+\hat{\beta}_p x_p$ (3.21)

Les parametres sont estimé avec encore les moindres carrés.

On choisit donc $\beta_1, \beta_2, ..., \beta_p$ pour minimiser la somme des residus au carré

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

$$RSS=\sum_{i=1}^{n}(y_i-\hat{\beta}_0-\hat{\beta}_1 x_1-...--\hat{\beta}_p x_m)^2 \quad (3.22)$$

Les valeurs $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p$ qui minimisent (3.22) sont les multiples coefficient de la régression linéaire simple.

Quand nous utilisons la regression linéaire multiple, nous sommes souvent interressés à répondre à ces questions.

1-/ est ce qu'au moins un des prédicteur $X_1,...,X_p$ est utile pour prédire la réponse ?

2-/ Est ce que tous les prédicteurs aident à expliquer Y, ou est ce qu'il y a seulement un sous ensemble utile pour predire ?

3-/ Comment le modèle s'ajuste bien par rapport aux données ?

4-/ Suivant in ensemble de valeur prédites, quelle valeur reponse devrions nous predire et commbien notre prédiction est précise ?

Premiere question : y a t il une relation entre la réponse et les prédicteurs ?

Pour la régression linéaire multiple avec p predicteurs, nous avons besoin de se demander si tous les coeffitiens sont nuls. Donc si $\beta_1=\beta_2=...=\beta_p=0$. Comme la regression linéaire simple, nous utilisons l'hypothèse de test pour repondre a la question. Nous testons l'hypothèse:

H_0: $\beta_1=\beta_2=...=\beta_p=0$ contre l'hypothèse H_a: au moins un β_j est non nul

Cette hypothèse est évalué avec la F-statistique:

$$F=\frac{(TSS-RSS)/p}{RSS/(n-p-1)}$$ (**pourquoi utilise t on RSS en bas**)

avec $RSS=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$ et $TSS=\sum(y_i-\bar{y})^2$

Si les suppositions d'un modèle linéaire sont corrects on peut montrer que $E[RSS/(n-p-1)]=\sigma^2$

demonstration:

$E[RSS]=E[||\hat{\varepsilon}||^2]=E[\hat{\varepsilon}'\hat{\varepsilon}]=E[tr(\hat{\varepsilon}'\hat{\varepsilon})]=E[tr(\hat{\varepsilon}\hat{\varepsilon}')]$ car tr(ABC)=tr(BCA)=tr(CAB)

$E[RSS]=tr(E[\hat{\varepsilon}\hat{\varepsilon}'])=tr(Var(\hat{\varepsilon}))$

or $Var(\hat{\varepsilon})=Var(Y-\hat{Y})$

or $\hat{Y}=X\hat{\beta}$

or $\hat{\beta}=(X^TX)^{-1}X^TY$ (point critique du lagrangien L)

donc $\hat{Y}=X(X^TX)^{-1}X^TY$

donc $Var(\hat{\varepsilon}) = Var((I_n - X(X^T X)^{-1} X^T) Y) = (I_n - X(X^T X)^{-1} X^T) Var(Y) = (I_n - X(X^T X)^{-1} X^T) \sigma^2$

donc $Var(\hat{\varepsilon}) = Var((I_n - X(X^T X)^{-1} X^T) Y) = (I_n - X(X^T X)^{-1} X^T) Var(Y) = (I_n - X(X^T X)^{-1} X^T) \sigma^2$

donc $E[RSS] = \sigma^2 [tr(I_n) - tr(X^T X (X^T X)^{-1})] = \sigma^2 [tr(I_n) - tr(I_{p+1})] = \sigma^2 (n - p - 1)$

donc $\sigma^2 = E[RSS/(n-p-1)]$ donc un estimateur sans biais est donné par $\hat{\sigma}^2 = RSS/(n-p-1)$

dans cette démontration on à fait l'hypothèse que l'on a effectivement un modèle linéaire

Si H_0 est vrai on a $E[(TSS - RSS)/p] = \sigma^2$

demonstration:

$$E[(TSS - RSS)/p] = E[\sum_{i=1}^{n} (y_i - \bar{y})^2 - \sum_{i=1}^{n} (y_i - \hat{y}_i)^2]/p$$

$$E[(TSS - RSS)/p] = 1/p(\sum_{i=1}^{n} E[(\varepsilon - \bar{\varepsilon})^2] - \sigma^2 (n - p - 1))$$ si on est dans l'hypothèse H_0

$$E[(TSS - RSS)/p] = 1/p((n-1)\sigma^2 - \sigma^2 (n - p - 1)) = \sigma^2$$ voir formule (A.64)

Par conséquent , lorsqu'il n'y a pas de relation entre Y et X nous nous attendons à que la F-statistique soit égale à 1.

Dans le cas où Ha est vrai alors $E[(TSS - RSS)/p] > \sigma^2$, donc on s'attend à que la F-statistique soit supérieure à 1.

De fàçon évidente des termes au carré en plus donneront $E[(TSS - RSS)/p] > \sigma^2$

De combien la largeur de la F-statistique doit être avant de rejeter H_0 et conclure qu'il y a une relation ? On voit bien que cela dépend des valeurs n et p. Quand n est grand, une F-statistique qui est un peu plus grand que 1 peut encore fournir l'évidence contre H_0.

Dans le cas contraire, une plus grande F-statistique est requise pour rejeter H_0 si n est petit.

Quand H0 est vrai et que les erreurs ε_i sont une distribution normale, la F-statistique suit une F-distribution. Pour des valeurs n et p données, n'importe quel logiciel peuut petre utilisé pour calculer la p-valeur associée avec la F-distribution. En se basant sur la p-valeur on peut determiner si oui ou non on rejete H_0. Des fois nous souhaitons tester un sous ensemble particulier de q coefficient nulles. Ce qui correspond à H0 : $\beta_{p-q+1} = \beta_{p-q+2} = ... = \beta_p = 0$. Dans ce cas on ajuste un second modèle qui utilise toutes les variables exceptés les derniers q. On suppose que la RSS de ce modèle est RSS_0 . Donc la F-statistique appropriée est

$$F = \frac{(RSS_0 - RSS)/q}{RSS/(n-p-1)}$$

TEXTE A REVOIR POUR COMPRENDRE

Given these individual p-values for each variable, why do we need to look at the overall F-statistic? After all, it seems likely that if any one of the p-values for the individual variables is very small, then at least one of the predictors is related to the response. However, this logic is flawed, especially when the number of predictors p is large.

For instance, consider an example in which p = 100 and H 0 : β 1 = β 2 = . . . = β p = 0 is true, so no variable is truly associated with the response. In this situation, about 5 % of the p-values associated with each variable (of the type shown in Table 3.4) will be below 0.05 by chance. In other words, we expect to see approximately five small p-values even in the absence of any true association between the predictors and the response. In fact, we are almost guaranteed that we will observe at least one p-value below 0.05 by chance! Hence, if we use the individual t-statistics and associated p-values in order to decide whether or not there is any association between the variables and the response, there is a very high chance that we will

incorrectly conclude that there is a relationship. However, the F-statistic does not suffer from this problem because it adjusts for the number of predictors. Hence, if H 0 is true, there is only a 5 % chance that the F-statistic will result in a p-value below 0.05, regardless of the number of predictors or the number of observations.

L'utilisation de la F-statistic pour tester n'importe quelle assiciation entre les predicteurs et la réponse fonctionne quand p est relativement petit, eet relativement plus petit que n.
Cependant il peut arriver que l'on ait un très grand nombre de variables. Si p>n alors il y a plus de coeffictients βj à estimer que do'bservation pour lequels on les estime. Dans ce cas, nous ne pouvons pas ajuster le modèle de régression linéaire en utilisant les moindres carrés, donc la F-statistique ne peut être utilisée et aucune autre technique que l'on ait pu voir de même.

Deuxième question : Decider des variables importantes
La première étape dans une régression multiple est de alculer la F-statistique et d'examiner les p-valeurs associées. Si nous concluons qu'il y a au moins un prédicteur qui est relié à la réponsen alors il est naturel de savoir lesquels ne sont pas bonnes! On pourrait regarder les p-valeurs individuelles mais comme nous avons vu, si p est grand il est probable de faire de mauvaises decouvertes.
Il est possible que tous les prédicteurs soient associéesà la réponsemais il est plus courrant que la éponse est reliée a un sous ensemble de prédicteurs. La tâche de déterminer lesquelles sont intéressantes s'appele la sélection des variables.
Idéalement on souhaiterait effectuer une selection de variable en essayant pleins de modèles différents, contenatnt chacun un ensemble de predicteur diffférents (exemple pour p=2 on a 4 possibilitées)
Comment determinons nous lequel des modèles est le meilleur ? Plusieurs statistiques peuvent etre utilisées pour juger de la qualité du modèle. Cela inclue Mallow's Cp, Akaike information criterion (AIC), Bayesian information criterion (BIC) et R^2. Ils seront dévellopés dans le chapitre 6.

Malheuresement il y a 2^p modèles qui contiennent un sousensemble de p variables. Cela signifie que meme pour un p modéré, essayer tous les sous-ensemble possibles de prédicteurs est impossible.
Donc si p est petit on peut considerer tous les modèles . Mais si p est grand nous aurons besoin une approche automatisé et efficace pour choisir un plus petit ensemeble de modèles à considérer.
Il existe 3 approches classique pour cette tâche:
- Forward selection: On commence par le modèle nulle (aucun prédicteurs). Puis nous ajustons p modèles de régression simple et nous ajoutons au modèle nul la variable qui nous donne le RSS le plus petit. Puis nous ajoutons a ce modèle la variable qui à le plus petit RSS pour le nouveau modèle a deux variables. On contiinue cette approche jusqu'à que nous soyons satisfait des résultats.
- Backward selection : On commence avec toutes les variables, et nos supprimons la variable avec la p-valeur la plus grande – qui est la variable qui est statisitquement la moins signifiante. Le nouveau modèle a p-1 varaible est construit et la variable avec la plus grande p-valeur est supprimée. On continue jusqu'à s'arreter au niveau d'une règle. On peut avoir comme regle de stopper quand le reste des variales ont une p-valeur sous un certain seuil.
- mixed selection: C'est une combinaison de foorward et backward selection. On commence avec aucune variable en faisant comme forward selection. O, continue d'en ajouter une par une. Si a un moment la p-valeur pour une des variables dans le modèle atteint un certain seuil, on supprime cette variable du modèle. On continue d'exécuter la forward et backward etapes jusqu'a que toutes les variables du modèles on suffisement une petite p-valeur, et que toute variable extérieure au modèle ait une grande p-valeur si on l'ajoute au modèle.

Backward selection ne peut pas etre utilisé si p>n alors que forward selection peut toujours etre utilisé.

Troisième question : l'ajustement du modèle

Deux des mesures numériques d''ajustement de modèle les plus connues sont RSE et R² . Ces quantités sont calculées et interprétées de la même manière que pour une regression linéaire simple.

Un R² proche de 1 indique que le modèle explique une grande partie de la variance dans les variables de réponse. Il faut un compromis entre la RSE et R² pour inclure les variables ou non dans le modèle.

On a $RSE=\sqrt{\frac{1}{n-p-1}RSS}$ pour la démonstration revoir **(ZT.35)** C'est à **(APK.45)** que ça change

$$E[\sum_{i=1}^{n}(\beta_1+...+\beta_p-\hat{\beta}_1-...\hat{\beta}_p)^2(x_i-\bar{x})^2]=E[(\beta_1-\hat{\beta}_1)^2+...+(\beta_p-\hat{\beta}_p)]E[\sum_{i=1}^{n}(x_i-\bar{x})^2]$$

$$E[\sum_{i=1}^{n}(\beta_1-\hat{\beta}_1)^2(x_i-\bar{x})^2]=var(\hat{\beta}_1+...+\hat{\beta}_p)\,E[\sum_{i=1}^{n}(x_i-\bar{x})^2]=\sigma^2\frac{p}{\sum_{i=1}^{n}(x_i-\bar{x})^2}E[\sum_{i=1}^{n}(x_i-\bar{x})^2]=\sigma^2 p$$

(il y a une erreur au niveau de la fraction, il faut calculer avec des xi1, ..., xip. mais en gros cest ca)

Donc avec un modèle de plusieurs variables peut avoir un RSE grand si l décroissance RSS est petit devant la croissance de p.

Quatrième question : les prédictions

Une fois les modèles de régression ajustés, il est simple d'appliquer $\hat{y}=\hat{\beta}_0+\hat{\beta}_1x_1+...+\hat{\beta}_px_p$ dans le but de prédire la reponse Y sur la base des valeurs des prédicteur X1,...,Xp.

Cependant il y a 3 sortes d'incertitudes associés à la prédiction

1-/ les estimation des coefficients $\hat{\beta}_0,\hat{\beta}_1,...,\hat{\beta}_p$ sont des estimations de $\beta_0,\beta_1,...,\beta_p$.

$\hat{y}=\hat{\beta}_0+\hat{\beta}_1X_1+...+\hat{\beta}_pX_p$ est juste une estimation de $f(X)=\beta_0+\beta_1X_1+...+\beta_pX_p$

L'imprécision des estimation des coefficients est dû a l'erreur reductilbe. On peut calculer l'intervalle de confiance dans le but de déterminer combien $\hat{Y}$ est proche de f(X).

2-/Il y a une source additionnelle potentiellement d'erreur reductible que l'on appelle biais du modèle. Quand nous utilisons un modèle linéaire nous estimons la meilleur approximation linéaire de la vraie surface. Toutefois ici nous ignorons la contradiction et opérons comme si le modèle linéaire était correct

3-/ Meme si nous connaissons f(X) – qui est que nous connaissons les vrais valeurs $\beta_0,\beta_1,...,\beta_p$ - la réponse ne peut être prédite parfaitement a cause de l'erreur aléatoire ε du modèle. Dans le chapitre 2 nous le réferions à l'erreur irréductible. Combien $\hat{Y}$ varie de Y ? Nous utilisons les intervalles de prédictions pour repondre à la question. Les intervalles de prédictions sont toujours plus large que l'intervalle de confiance, ar ils incorporent les deux erreur d'estimation de f(X) à savoir l'erreur irreductible et l'erreru reductible.

Intervalle de confiance exemple pour 95% des valeurs on a $[\hat{\beta}_1-2.SE(\beta_1),\hat{\beta}_1+2.SE(\beta_1)]$

intevalle de prédiciton plus large que l'intervalle de confiance

Les prédicteurs qualitatifs

Il peut arriver souvent que les prédicteurs soient des variables qualitatives

Les prédicteurs avec uniquement deux niveaux sont faciles a implementer dans le modèle de regression. On créé simplement une dummy variable (ou indicateur) qui prennents deux valeurs possibles .

On peut prendre xi= 1 si homme et 0 si femme

Ce qui nous donne $yi=\beta_0+\beta_1x_i+\varepsilon_i=\beta_0+\beta_1+\varepsilon_i$ si xi est 1 (donc un homme)

$yi=\beta_0+\beta_1x_i+\varepsilon_i=\beta_0+\varepsilon_i$ si xi est 0 (donc une femme)

Donc $\beta0$ est la moyenne des femmes et s $\beta0+\beta1$ est la moyenne des hommes et $\beta1$ l'écrat moyen entre les moyennes

Inverser les dummy variables (et donc 0 pout homme et 1 pour femme) affecte les variables et les moyennes des coefficients. Pour éviter le problème on peut utiliser des dummy variables de ce genre
on prend -1 si femme ou 1 si homme .
Ce qui nous donne $yi=\beta_0+\beta_1 x_i+\varepsilon_i=\beta_0+\beta_1+\varepsilon_i$ si xi est 1 (donc un homme)
$yi=\beta_0+\beta_1 x_i+\varepsilon_i=\beta_0-\beta_1+\varepsilon_i$ si xi est -1 (donc une femme)
La diffrence entre ces 3 techniques et que les coefficients `β0 et β1` ont diférentes significations.
`β0` peut être interpreté comme la moyenne de tous les hommes et femmes `β1` représente la part moyenne supplémentaire qu'on les femmes par rapport a `β0`.

On peut aussi se retrouver avec des prédicteurs qualitatifs ayant plus de 3 niveaux. Quand on est a plus de trois niveaux, une unique dummmy varaible ne suffit plus. Dans ce cas on créé plus de dummy variables.

Pour l'exemple de Caucasien , afro americain ou asiatique on utilise deux dummy variables
$yi=\beta_0+\beta_1 x_{i,1}+\beta_2 x_{i,2}+\varepsilon_i$
x_i,1 = 1 si la ième personne est asiatique sinon 0
x_i,2 = 1 si la ième personne est caucasienne sinon 0
Ce qui nous donne $yi=\beta_0+\beta_1+\varepsilon_i$ si la ième personne est asiatique
$yi=\beta_0+\beta_2+\varepsilon_i$ si la ième personne est Caucassienne
$yi=\beta_0+\varepsilon_i$ si la ième personne est afro americaine
Maintenant Beta_0 peut etre interprété comme la moyenne des afro americains,
Beta_1 comme la difference entre la moyenne entre asiatique et afro americain
Beta_2 comme la difference entre la moyenne entre Caucasien et afro americain
Conclusion : il y aura toujours un nombre n-1 de dummy variable pour n categories

REVOIR LA F-STATISTIQUE (ce que ca signifie)
Extensions du modèle linéaire
Le modèle de regression linéaire standard permet l'interpretation des reusltats et fonctionne assez bien à beaucoup de vrai problèmes. Toute fois, il suppose de très hautes hypothèses de restriciton qui sont souvent violée en pratique.
La deux suppositions les plus importntes sont la relation entre le predicteur et la reponse (elle sont additive et linéaire).
La supposition additive signifie que les effets de changement de Xj sur la reponse Y est independante des autres predicteurs.
L'hypothèse linéaire indique que la variation de la réponse Y due à une modification d'une unité de X j est constante, quelle que soit la valeur de X j.

Supprimer la supposition additive
Supposons le modèle de regression linéaire suivant $Y=\beta_0+\beta_1 X_1+\beta_2 X_2+\varepsilon$
suivant ce modèle si nous augmentons X1 par une unité alors Y agmentera en moyenne d'une unité Beta_1. A noter que laprésence de X2 n'altere pas l'etat. Alors qu'ne réalité il est possible que ca affecte X2.
Un des moyens pour etendre ce modèl est de permettre un effet d'interaction en incluant un troisième predicteur, appelé un terme d'interaction . Celui ci est construit en calculant le produit de X1 et X2.
Ce qui nous donne $Y=\beta_0+\beta_1 X_1+\beta_2 X_2+\beta_3 X_1 X_2+\varepsilon$
$Y=\beta_0+(\beta_1+\beta_3 X_2) X_1+\beta_2 X_2+\varepsilon$
$Y=\beta_0+\widetilde{\beta}_1 X_1+\beta_2 X_2+\varepsilon$ où $\widetilde{\beta}_1=\beta_1+\beta_3 X_2$
Quand $\widetilde{\beta}_1$ change avec X2, l'effet de X1 sur Y n est plus constant. Ajuster X2 changera l'impact de X1 sur Y.
La principe hierachique stipule que si nous incluons une interaction dans le modèle, nonus devons aussi inclure les princpaux effets et ce meme si les p-valeurs associées avec leur coefficient ne sont pas significatif.

En d'autres mot si l'interaction entee X1 et x2 semble importante, nous devons inclure X1 et x2 dans le modèle.

Pour le cas des variables qualitatives, ou quantitative et qualitatives on peut proceder de la meme manière. (Voir page 104 pour exemple en gros on a une dummy variable Beta_2 et une Beta_3 * X1 et nulles si non studient)

Des relations non-linéaires
Comme nous avons parlé précédemment, le modèle de régression linéaire suppose un relation linearire entre la reponse et les predicteurs. Mais dans certains cas la relatio nentre la reponse et les predicteurs peuvent etre non linéaires. On peut passer a une regression polynomiale.

Des problèmes potentiels
Quand nous ajustons une regression linéaire à un jeu de donné particulier, beaucoup de problemes se produisent. Les plus communs sont:
- non linearite de la relation reponse-predicteur
- correlation des termes d'erreur
- valeurs aberrantes
- for levier de points
- collinearité

En pratique, identifier et apprehender ces problemes est un etat de l'art.
1-/ non linearité du modele
La regression lineaire suppose qu'il y a une strict ligne reliant les predicteurs et la reponse. Si la relation est très loin de linéaire, alors toutes les conclusion que nous avons etablie sont mauvaises. La precision de prédiction du modèle peut etre nettement reduite.
Les residuals plot sont des outils graphiques pour identifier la non-linéarité. Soit un simple modèle de régression linéaire, nous pouvons afficher les résidus $e_i = y_i - \hat{y}_i$ contre les predicteurs. Dans le cas d'une regression multiple on a alors plusieurs predicteurs, au lieu de cela on affcihe les residus contre les valeurs predites $\hat{y}_i$.
Si les residual plots indique qu'il y a une association non linéaire dans les donnes, alors une approche simple est d'utiliser des transformations non-lineaire des predicteurs comme logX, sqrt(X) et X^2.

2-/ la correlation des termes d'erreur
Une supposition importante du modèle de regression lineaire est que les termes d'erreur $\varepsilon_1, \varepsilon_{2,..}, \varepsilon_n$ ne sont pas corrélés. Qu'est ce que ca signifie? Par exemple si l'erreur ne sont pas corrélé alors la fait que εi soit positive permet peu ou pas d'information sur le signe de $\varepsilon i+1$. L'erreur standard qui est calculée pour calculer les coefficients de regression ou les valeurs ajustées sont basées sur la supposition de termes d'erreurs non corrélés. Si en fait il y a corrlation parmis les terme d'erreurs, alors l'estimation des erreurs standard tenderont a sous estimer la vrai erreur standard. Les intervalles de confiance et de prédiction seraient donc trop petits qu'il ne devraient l'etre: cela causerait de conclusion errronées qu'un parametre soit significatif. En résumé, si les termes d'erreur sont corrélés, on aura un sentiment de confiance injustifié dans notre modèle

3-/la variance du terme d'erreur non-constante
Une supposition importante de la regression linéaire est que les termes d'erreur ont une variance constante, Var(εi)=σ²

Les erreurs standards, les intervalles de confiance, et les hypothèses de tests associées a un modèle linéaire dependent de cette supposition. Malheuresmeent, il est frequent que la varaince du terme d'erreur soit non constant. Par exmple la variance du terme d'erreur pourrait augmenter avec la valeur de la reponse. On peut identfier des variance non constantes des les erreurs, ou heteroscedacité par la présence d'un entonnoir dans les residuals plots. Par exemple voir figure 3.11 en utlisant log(Y)

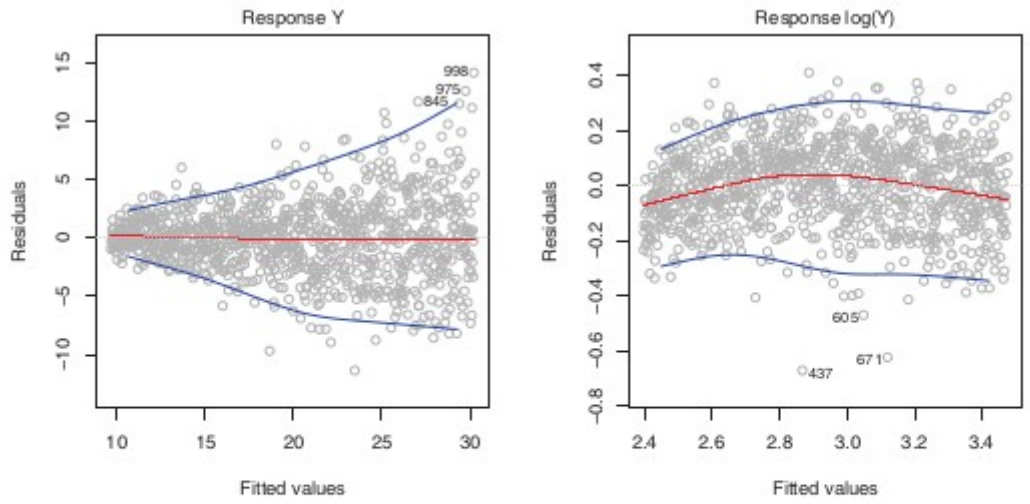


FIGURE 3.11. *Residual plots. In each plot, the red line is a smooth fit to the residuals, intended to make it easier to identify a trend. The blue lines track the outer quantiles of the residuals, and emphasize patterns.* Left: *The funnel shape indicates heteroscedasticity.* Right: *The response has been log transformed, and there is now no evidence of heteroscedasticity.*

avec
log(Y) les residus apparaissent avoir une variance constante, qui montre une relation non linéaire dans les données
Il arrive que l'on ait une bonne idée de la variance de chaque reponse. Parfois, nous avons une bonne idée de la variance de chaque réponse. Par exemple, la énième réponse pourrait être une moyenne de n i observations brutes. Si chacune de ces observations brutes n'est pas corrélée avec la variance σ², alors leur moyenne a la variance σ i ² = σ² / n i. Dans ce cas, une solution simple consiste à ajuster notre modèle par la méthode des moindres carrés pondérés, avec des pondérations proportionnelles à la valeur inverse.
variances - c'est-à-dire. w i = n i dans ce cas. La plupart des logiciels de régression linéaire permettent les poids d'observation.

4-/ Valeurs aberrantes
Un outliers est un point pour lequel yi est loin de la valeur prédite par le modèle. Elle peuvent apparaitre pour plusieurs raisons, comme un enregistrement incorrect durant l'observation de la collecte des donnees. Les residual plits peuvent etre utilisés pour identifier les outliers. En pratique il peut s'averer difficile de decider de combien les residus ont besoin pour etre considéré comme un outlier. Pour resoudre se probelme, au lieu d'afficher les residus, on peut afficher les residus studentisés, caluclé a partir de la division de chaque résidu par son erreur standard d'estimation. Les observation pour lesquels les residus sont plus grand que 3 en valeur absolue sont des possibles valeur aberante.
Si nous croyons qu'un outlier s'est produit dû a une erreur dans l'enregisrement des données, alors la solution est de simplement supprimer l'observation. Il faut quand meme faire attention que ce ne soit pas le modèle qui est mal choisit comme un predicteur manquant.

5-/ High leverage points
On vient de voir que les outliers sont des observations pour lesquelles la reponse yi est inhabituelle suivant un predicteur xi. En revhanche, des observations avec high leverage ont des valeur inhabituelles pour xi. Des poins de levier permettent de faire des regression plus précise. Elle ne sont pas aberante et doivent etre prisent en compte pour augmenter la qualtité de la regression. C'est pour cela qu'il est important d'identifier les observations de point de levier.

SUITE In a simple linear regression

Suite page 112
εβμσΠ
$\hat{f}$
ascertain : à verifier
assessing : evaluer, evaluation
assumption : supposition
despite : malgré
discrepancy : contradiction
for instance : par exemple
however: toutefois
in contrast : en revanche
lack of: manquer de
outliers: valeurs aberrantes
overall : global
perform : permettre / éxecuter
provided that : a condition que
rather : plutôt
scatter : distribution
scatterplot : nuage de points
scope: portée
seek : rechercher , tenter de
since : puisque
slope : pente
spread out : s'étendre
straightfoward : simple, honnête, juste
throughout : tout au long
whether : qu'il s'agisse,si
wider : plus large