Guide d'étude

Examen SRM: Statistics for Risk Modeling
Society of Actuaries (SOA)

Alec James van Rassel

Table des matières

Préliminaires

Rappel des bases de statistiques

Vidéos YouTube

- StatQuest: One or Two Tailed P-Values
- Khan Academy: P-values and significance tests

Notes sur les vidéos YouTube

StatQuest: One or Two Tailed P-Values

- **Test** new treatment ● vs old treatment ●
  - One-tailed : $\mathcal{H}_0$: The new treatment ● is *better* than the old treatment ●.
  - Two-tailed : $\mathcal{H}_0$: The new treatment ● is *better*, *worse* or *not significantly different* than the old treatment ●.
- **P-Hacking**
  - Deciding to test if the new treatment ● is *better* rather than *better*, *worse* or *not significantly different* than the old treatment ● *after* seeing the skewed distribution.
- **False Positive** *(recall the image of the normal distribution with the points)*
  - Usually, the two samples for ● and ● overlap (recall image where they're overlapping) but it can happen that they don't (recall image where they're spaced out) in which case the p-value is less than 0.05 and we have a ***false positive***.
- To decide which t test to use, ask ourselves **what we want to learn from the test** and then decide.

Khan Academy: P-values and significance tests

- Test de signification
  - i Déterminer la question
  - ii Établir les connus

1. Définir les hypothèses $\mathcal{H}_0$ et $\mathcal{H}_a$
2. Établir un niveau de signification α (souvent 0.05)
3. Obtenir un échantillon
4. Trouver la p-value
   - p-value = Pr{ statistique de test si $\mathcal{H}_0$ est vrai } $\neq$ Pr{ $\mathcal{H}_0$ est vrai sachant la statistique de test }
5. Décider
   - Si p-value $< \alpha$ alors on rejète l'hypothèse nulle
   - Si p-value $\geq \alpha$ alors on ne peut pas rejeter l'hypothèse nulle

Autres ressources

Liens

- Article sur la musique et le bruit ;
- Article sur l'analyse de séries chronologiques avec R ;

Sujets à l'étude

1 Basics of statistical learning (7.5% à 12.5%)

Information

Objective

Understand key concepts of statistical learning

Learning outcomes

1. Expliquer les différents types de problèmes, et différentes méthodes, de modélisation. Y compris :
   - Apprentissage supervisé vs non-supervisé
   - Régression vs classification
2. Expliquer les méthodes courantes pour évaluer la précision d'un modèle.
3. Utiliser les méthodes de base d'analyse exploratoire de données y compris la validation et vérification de données.

Related lessons ASM

1. Basics of Statistical Learning

Vidéos YouTube

- StatQuest: A Gentle Introduction to Machine Learning
- StatQuest: Machine Learning Fundamentals: Bias and Variance
- StatQuest: StatQuest: Quantiles and Percentiles, Clearly Explained!!!
- StatQuest: Quantile-Quantile plots (QQ plots), clearly explained
- StatQuest: Quantile Normalization

Résumés des chapitres

1. Basics of Statistical Learning

- Définitions de base
- Comparaisons
  - Prévision vs inférence
  - Paramétrique vs non-paramétrique
  - Flexibilité vs interprétabilité
  - Apprentissage supervisé vs non-supervisé
  - Régression vs classification
  - Biais vs variance
- Types de variables
  - Continue
  - Catégorique

    S'il y a une ordre logique, c'est *ordinal* sinon nominal.
  - Comptage
- Graphiques (scatter, box, qq)

Notes sur les vidéos YouTube

StatQuest: A Gentle Introduction to Machine Learning

- Le but de l'apprentissage machine est de faire des **prévisions**.
- Les *arbres de décisions* peuvent être utilisés pour la prévision ou la **classification** et la régression pour la prévision.
- Peut importe la méthode, l'important est de mesurer la ***précision*** *des prévisions*.
- On peut donc tester notre méthode avec du **testing data**.
- **Bias-Variance tradeoff** : Être bien ajusté, mais avoir des mauvaise prévisions.

StatQuest: Machine Learning Fundamentals: Bias and Variance

- **Biais** : Inhabilité pour une méthode de prévision de capturer la vraie relation. Pour exemple :
  - Essayer de fit une droite aux points du vidéo ; évidemment elle ne peut pas reproduire la vraie relation **VS** une squiggly line qui peut avec little to no biais.
  - Alors, la squiggly line sera mieux ajusté aux données d'entrainement mais pas aux testing data ; alias, elle sera **overfitted**.
- **Variance** : Variance de l'ajustement entre ensembles de données.
  - La squiggly line aura une variance très élevée mais pas la droite ; la droite est cohérente et aura toujours un **SSE** similaire.
- On cherche donc la meilleur combinaison des deux méthodes. Plusieurs méthodes existent pour la trouver dont la **régularisation**, le *boosting* et le *bagging*.

StatQuest: StatQuest: Quantiles and Percentiles, Clearly Explained!!!

Visualise the chart with the 15 points to which we add the lines.

1. The **median** is a quantile because it splits the data into 2 equally sized groups.
   - The (0.5 | 50%) quantile value is 4.5.
2. The (0.25 | 25%) quantile is such that (25% | a quarter) of the points are less than it.
3. **Technically**, quantiles are line which divide the data into equally sized groups.
4. **Technically**, percentiles are just quantiles which divide the data into 100 equally sized groups.
   - In practice, terminology is much more flexible ;
   - Even if there's less than 100%, we still call the (median | 50% quantile) the 50^{th} percentile.

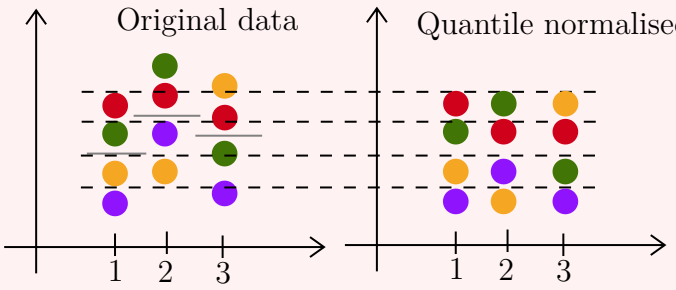
   - To calculate a quantile / percentile, just calculate how many values are less than it.
5. There are many other ways to calculate quantiles, R has 9 just by itself!
6. The lesson however, is that the smaller the sample the more variable the results.

StatQuest: Quantile-Quantile plots (QQ plots), clearly explained

1. Give each data point it's quantile.
2. Get a normal curve.
3. Add all the quantiles obtained to the curve.
   - That's to say the percentage and not the value.
4. Plot QQ graph.
   - If the data were normal, most data points would be on the line;
   - If the fit is not good, can try a different distribution;
   - Thus, can use the QQ plot to see which distribution fits the data.

- Could also use the QQ plot to compare 2 datasets.
- For example, if the second one had less data points we could use quartiles instead!
- We add the points at the intersection of the 2 datasets' quantiles and add a straight line.

StatQuest: Quantile Normalization



The idea for quantile normalisation ressembles the offset in linear regression. You account for differences in samples to adequately compare them.

To compensate for the differences in lightbulb intensity across samples, want to normalise the data points; you can see the difference in means across all three in gray.

1. Focus on the maximum value of each sample and take the mean.

   Extend this mean onto the new plot and that becomes the new value for all 3 samples.

2. Repeat for all the points.

   The end result is that the **values are all the same** but the **order of the colours is maintained** enabling an adequate comparison.

2 Linear Models (40% à 50%)

Information

Objective

Understand key concepts concerning Generalized Linear Models

Learning outcomes

1. Décrire et expliquer les composantes de la famille exponentielle et des fonctions de lien.
2. Estimer les paramètres par Least Squares et maximum de vraisemblance.
3. Interpréter les tests de vérification pour l'ajustement de modèle et la vérification des postulats graphiquement et numériquement.
4. Sélectionner un modèle approprié en prenant en compte :
   - Distributions et fonctions de lien
   - Transformations de variables, et leurs interactions
   - Statistique (du khi-carré) de Pearson
   - Tests t et F
   - AIC et BIC
   - Test du Rapport de Vraisemblance (TRV)
5. Interpréter les résultats du modèle dans le cadre de son utilisation pour résoudre aux problèmes d'affaires sous-jacents
6. Calculer et interpréter les valeurs prédites ainsi que les intervalles de confiance et de prévision.
7. Comprendre que d'autres méthodes peuvent différer du modèle OLS comme la régression Lasso, Ridge et KNN.

Related lessons ASM

2. Linear Regression : Estimating parameters
3. Linear Regression : Standard Error, R^2, and t statistic
4. Linear Regression : F statistic

5. Linear Regression : Validation
6. Resampling methods
7. Linear Regression : Subset Selection
8. Linear Regression : Shrinkage and Dimension Reduction

9. Linear Regression : Predictions
10. Interpreting Regression Results

11. Generalized Linear Models : Basics
12. Generalized Linear Models : Categorical Response
13. Generalized Linear Models : Count Response
14. Generalized Linear Models : Measures of Fit

Vidéos YouTube

- Khan Academy: Pearson's chi square test (goodness of fit)
- StatQuest: Fitting a line to data, aka least squares, aka linear regression
- StatQuest: R-squared explained
- StatQuest: Linear Models Pt.1 - Linear Regression
- StatQuest: Linear Regression in R
- StatQuest: Linear Models Pt.1.5 - Multiple Regression
- StatQuest: Multiple Regression in R
- StatQuest: Linear Models Pt.2 - t-tests and ANOVA
- StatQuest: Linear Models Pt.3 - Design Matrices
- StatQuest: Linear Models Pt.3 - Design Matrix Examples in R

- Phil Chan: Introduction to the Hat matrix in regression
- Phil Chan: Properties of the Hat matrix with proofs *(pas pertinent pour SRM, pas regarder dans ce cadre.)*
- Phil Chan: How is it the Hat matrix spans the column space of X? Really nice.
- Phil Chan: What does it mean to say the Hat matrix is an orthogonal projection? *(pas pertinent pour SRM, pas regarder dans ce cadre ; pas encore regardé moi-même mais je le note comme référence.)*
- Phil Chan: Should I look at raw, standardized, or studentized residuals? part 1 - what to look out for
- Phil Chan: Should I look at raw, standardized, or studentized residuals? part 2 - kinds of residuals
- Phil Chan: Should I look at raw, standardized, or studentized residuals? part 3
- Phil Chan: What's the difference between an outlier and a leverage point in regression?
- Phil Chan: Influential points - Cook's distance, DFFITS, DFBETAS
- jbstatistics: Leverage and Influential Points in Simple Linear Regression (*didn't watch the whole thing because time, but he breaks down the formula for leverage very well towards the end*)
- Ben Lambert: Variance Inflation Factors: testing for multicollinearity
- StatQuest: Machine Learning Fundamentals: Cross Validation
- Stephanie Glen: Adjusted R Squared
- Ben Lambert: Adjusted R squared
- ritvikmath: Vector Norms
- StatQuest: Regularization Part 1: Ridge Regression
- StatQuest: Regularization Part 2: Lasso Regression

- StatQuest: Logistic Regression

- StatQuest: Probability vs Likelihood
- StatQuest: Maximum Likelihood, clearly explained!!!

Pas regardé les 6 suivants mais je me les note comme référence au cas où que j'aurai le temps

- StatQuest: Maximum Likelihood For the Normal Distribution, step-by-step!
- StatQuest: Maximum Likelihood for the Exponential Distribution, Clearly Explained! V2.0
- StatQuest: Maximum Likelihood for the Binomial Distribution, Clearly Explained!!!
- StatQuest: Logistic Regression Details Pt1: Coefficients
- Logistic Regression Details Pt 2: Maximum Likelihood
- Logistic Regression Details Pt 3: R-squared and p-value

- StatQuest: Odds and Log(Odds), Clearly Explained!!!
- StatQuest: Odds Ratios and Log(Odds Ratios), Clearly Explained!!!

Régression linéaire

Résumés des chapitres

2. Linear Regression : Estimating parameters

1. Régression linéaire simple
   - Hypothèses pour les résidus ;
   - Différentes formulations des estimateurs. Particulièrement, d'exprimer $\hat{\beta}_1$ en fonction du coefficient de corrélation, la covariance et les variances échantillonnalles.
2. Régression linéaire multiple
   - Faire attention à la colinéarité ;

- Idée de ne pas avoir une colonne qui somme à un sinon intéraction avec la rangée de 1.

3. Généralisations de régression linéaire
   - Régression linéaire pondérée ;
   - Transformation (exposant, polynomiale, log, ...) ;
   - Famille de transformations Box-Cox.

Note sur les exercices : Généralement 2 types

1. Trouver estimations de $\hat{\beta}_1$ pour une régression linéaire simple ;
2. Trouver estimations de $\hat{\beta}_1$ pour une régression linéaire multiple.

Dans les deux cas c'est à partir de données qui incluent souvent le produit croisée des observations, ou la corrélation, etc.

3. Linear Regression : Standard Error R^2 and t statistic

1. Residual Standard Error of the Regression
   - Bien distinguer toutes les différentes façons de dire et d'écrire les équations pour les sommes des carrées.
2. R^2 : Coefficient de détermination
   - Savoir les différentes façon de l'écrire afin d'être en mesure de le calculer à partir d'une variété d'information donnée.
3. Statistique t
4. Graphique et algorithme de variable ajoutée et coefficient de corrélation partiel
   - Remember the algorithm for the added-variable plot (p.31)
   - Remember the formula for the partial correlation coefficient $\frac{t(b_1)}{\sqrt{t(b_1)^2+(n-p')}}$.

Note sur les exercices : Généralement de 3 à 4 types :

1. Trouver le R^2, MSE, etc. à partir d'information donnée ; donc savoir les différentes formulations pour ces équations ;
2. Trouver les statistiques t pour les paramètres ;
3. Trouver des intervalles de confiance pour les paramètres ;

4. Peut-être une sur le coefficient de corrélation mais c'est facile à trouver.

Donc il faut concrètement comprendre la distinction et les formules pour les SS, R^2 et variance des coefficients.

4. Linear Regression : F statistic

En gros le chapitre explique les deux différents tests F

1. Test de la validité gloable que l'on compare notre modèle à la moyenne.
   - › Voir les vidéos de StatQuest, il couvre les tests F en expliquant le reste de la régression linéaire.
   - › Savoir que pour un test du retrait d'une seule variable β_1, $F_{1,n-2} = (t_{n-2})^2$.
   - › Savoir la connection avec le R^2 en divisant par le SST.
2. Test pour le retrait de variables.
   - › Également lien avec le R^2.
   - › Faire attention au degrés de liberté si on donne les MSE pour les deux modèles ; pas le même pour les deux.

Note sur les exercices : Généralement de 2 types :

1. Trouver la statistique F pour le test global ;
2. Trouver la statistique F pour le test partiel.

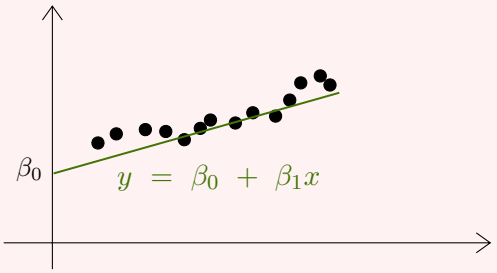
Dans les deux cas, souvent la question donne de l'information partielle et il faut savoir utiliser les différentes formulations pour l'équation. Pour exemple, avec R^2, avec le t^2, avec SSR ou SSE, etc.

Notes sur les vidéos YouTube

StatQuest: Fitting a line to data, aka least squares, aka linear regression

Recall the basic formula $y = ax + b$.

1. In Least Squares, we want to find the starting point on the line (b or β_0) and the slope (a or β_1) to minimise the distance of the points to the line.



2. The distance of the points to the line are the **residuals** ε.
3. When we plot the SSR for the different possible curve, we pick the one with the *smallest SSR*.
   - That is to say, we pick the one with the **Least Squares**.
   - That curve is the curve such that the derivative has a slope of 0 and it becomes the **Least Squares Estimator**.

StatQuest: R-squared explained

For comprehension, we define the measure of correlation by R rather than ρ.

1. The main idea to retain is that R^2 is a **metric of correlation**.
2. In fact, R^2 is literally the **correlation squared**, $R^2 = (R)^2$.
   - The advantage is that it's much easier to interpret and to calculate.
   - It's not obvious that $R = 0.7$ is twice as good as $R = 0.5$ but $R^2 = 0.7^2 = 0.5$ is clearly twice as good as $R^2 = 0.5^2 = 0.25$.
3. The same idea from the Least Squares video is used, we want the line that will minimise the SSR. Now however, we want to

quantify how good this new **line is** and we use the R^2.

4. $R^2 = \frac{\text{Var(mean)} - \text{Var(Least Squares line)}}{\text{Var(mean)}}$.
   - That is to say we divide the difference in the variation of the points to the line (SSR) by the variation of the points to the line that goes through the mean ;
   - Thereby, this has to be between 0 and 1 since the SSR for any line is $\leq$ SSR for the line going through the mean ;
   - Thus $0 \leq R^2 \leq 1$ and it is a percentage.
5. The interpretation of R^2 can be either :
   - There is 81% less variation around the line than the mean ;
   - The line explains 81% of the variation of the relationship between the two variables.
6. Thus, logically we want to maximise the R^2.
7. **Note** : R^2 doesn't indicate the direction of the relationship like R does.

StatQuest: Linear Models Pt.1 - Linear Regression

1. Reminder on R^2
   - Evaluates how well a line fits ;
   - SS(mean) = (data - mean)^2 : Sum of Squares around the mean ($\bar{y}$) ;
   - Var(mean) = $\frac{\text{(data} - \text{mean)}^2}{n}$: Variation of the data around the mean ($\bar{y}$) ;
   - SS(fit) = (data - fit)^2 : Sum of Squares around the fitted Least Squares line ;
   - Var(fit) = $\frac{\text{(data} - \text{fit)}^2}{n}$: Variation of the data around the best fit line ;
   - In general, Var(something) = $\frac{SS(\text{something})}{\text{number of things}}$;
   - Thus, $R^2 = \frac{SS(\text{mean}) - SS(\text{fit})}{SS(\text{mean})} \Leftrightarrow \frac{\text{Var(mean)} - \text{Var(fit)}}{\text{Var(mean)}}$.
2. Note on **adding variables**

- More parameters will always have a "***better***" fit because the Least Squares line will set useless variables's parameters to 0 ;
- Thereby, the SS(fit) will decrease causing the R^2 to increase ;
- This is not always a good thing. For example, if we add the result of a coin flip as a variable then it is possible that by chance heavier mice will get more heads. This would mean that a nonsensical variable would lead to a better R^2 ;
- This is why often times the adjusted R^2_a is reported instead of the R^2.

3. Note on the **number of data points**
   - If, for example, there were only 2 points then the line would have a perfect fine leading to a $R^2 = 1$.
   - However, this is true for any 2 points with a line connecting them.
   - Therefore, we want a measure of the significance of the R^2, we want a **p-value**.
4. To account for the 2 problems of the *number of parameters* and the *amount of data*, we define F.
   - First, we compare what the equations mean in text :

     $$R^2 = \frac{\text{variance in size that } \textbf{IS} \text{ explained by adding weight}}{\text{variance in size } \textbf{WITHOUT} \text{ weight taken into account}}$$
     $$F = \frac{\text{variance in size that } \textbf{IS} \text{ explained by adding weight}}{\text{variance in size } \textbf{NOT} \text{ explained by adding weight}}$$

     Thus, the denominator of F is the variance of the points to the line, or, **the error** in the prediction of mouse size.
   - More formally, we get :

     $$R^2 = \frac{SS(\text{mean}) - SS(\text{fit})}{SS(\text{mean})}$$
     $$F = \frac{(SS(\text{mean}) - SS(\text{fit})) \,/\, (p_{\text{fit}} - p_{\text{mean}})}{SS(\text{fit}) / (n - p_{\text{fit}})}$$

- $(p_{\text{fit}} - p_{\text{mean}})$ is the difference in parameters between the 2 lines; i.e., the number of parameters in addition to the y-intercept.
- $(n - p_{\text{fit}})$ is the number of observations minus the number of parameters.
- These 2 are the **degrees of freedom** for the F distribution.

5. Recall the **histogram of** the blocks of the F values for many random samples.
   - The p-value is obtained from the F distribution and is : $\frac{\text{number of blocks that are just as, if not more, extreme than the block for the fitted line}}{\text{the total number of blocks from all the simulations}}$;
   - In practice this is **approximated with a curve** where the **degrees of freedom** are the **shape parameters**.
   - We note that as the **number of points goes up**, the tail tapers off (light-tailed) and thus the **p-value gets smaller**.
   - This makes sense as the approximation gets closer to reality, the result is more credible and thereby more significant.

StatQuest: Linear Models Pt.1.5 - Multiple Regression

1. Same thing as simple linear regression but with more parameters.
2. R^2 is the same but p_{fit} will increase as p_{mean} stays equal to 1.
3. In both cases, this would be comparing the model (simple or multiple regression) to the mean but we can compare them to each other!
   - If want to know whether the addition of a second parameter, say tail length, is worth the time and effort it'll take to collect it's data we do a **partial F test** :

$$F = \frac{SS(\text{simple}) - SS(\text{multiple}) / (p_{\text{multiple}} - p_{\text{simple}})}{SS(\text{multiple}) / (n - p_{\text{multiple}})}$$

StatQuest: Linear Models Pt.2 - t-tests and ANOVA

1. The **goal** of a t-test is to **compare means** and see if there's a **significant difference** between them.
2. General steps to a t-test are the following (while visualising a t-test and simple linear regression side-by-side) :
   (a) Ignore x-axis and find the overall mean
   (b) Calculate the SS(mean) for both the t-test and regression
   (c) Fit a line (we care about the x-axis now). Recall that the t-test is two lines :
      › Use indicator variables represented by a **design matrix** which has 1's for the first group and 0's for the other, and vice-versa for all observations.
      › The equation becomes $y = \text{column } 1\mu_1 + \text{column } 2\mu_2$.
   (d) Calculate F now that the SS(mean) and SS(fit) is found.
      › Note that $p_{\text{mean}} = 1$ because there is only the intercept as an argument.
      › $p_{\text{fit}} = 2$ because of the 2 indicator variables for t-test and the 2 parameters for regression.
   (e) **ANOVA** is a generalization with more than just 2 groups' means to compare.
      › $p_{\text{mean}} = 1$ like before.
      › p_{fit} is the number of groups / the number of parameters ; 5 in the example.
      › Note that the design matrix can be rewritten with a column of 1's and is usally seen that way.

StatQuest: Linear Models Pt.3 - Design Matrices

1. Plus in one row at a time
2. Idea of 2 lines and predicted value will be on one or the other depending on the indicator variable.
3. Idea of comparing to the mean *or* a simpler model

Validation, sélection et qualité d'ajustement

Résumés des chapitres

5. Linear Regression : Validation

1. Validating model assumptions
   - › Matrice chapeau $\boldsymbol{H}$ et leverage h_{ii}
   - › Variance des résidus
   - › Types de résidus
     - – Résidus (raw) ;
     - – Résidus standardisés ;
     - – Résidus studentisés.
   - › Postulats appliqués à la variable réponse
     (a) Linéarité -> vérifié avec un QQplot ;
     (b) Normalité -> vérifé avec un graphique des résidus $\hat{\varepsilon}$ contre les prévisions $\hat{Y}$;
     (c) Homoscédasticité -> vérifiés avec un graphique des résidus contre chaque variable explicative qui ne devrait pas avoir de tendance (**pattern**) discernable (pour exemple, quadratique) ;
     (d) Indépendance des observations -> vérifiés avec la même graphique mais dans un ordre dont la corrélation est attendue (pour exemple, si données chronologiques selon le temps).
2. Outliers and influential points
   - › 2 types d'observations différents **(voir vidéos)** :
     (a) **Outliers** : espacés horizontalement et détectable par un **résidu élevé**.
     (b) **Leverage points** : espacés verticalement et détectable par un **high leverage** ; cependant, pas nécessairement mauvais.

- De plus, il y a des **influential points** (les mauvais leverage points).
- **Cook's distance** pour mesurer l'impact global sur les estimations des paramètres lorsque l'on retire chaque observation une à la fois.

3. Collinearity of explanatory variables ; VIF
   - Calcul du VIF_j et du $R^2_{(j)}$;
   - Relation avec l'écart-type estimé du paramètre $\hat{\beta}_j$.

Note sur les exercices :

1. Généralement peu d'exercices des examens.
   - Des exercices des examens, quelques choix multiples dont il faut dire quels affirmations sont vrai/faux de graphiques ;
   - Important est surtout de bien saisir les concepts et savoir les différentes formulations pour les équations (Cooks' Distance, VIF / s_{b_j}, ...).
2. Exercices de **résidus** aucune question d'examen, mais semble plus logique que ce soit englobé par une question plus générale.
3. Exercices de **influential points** :
   - une question d'examen ;
   - Faire des relations des différents formules pour les trouver à partir d'info partielle.
4. Exercices sur le **VIF**
   - Bien saisir le concept de faire une régression sur une autre variable explicative ;
   - Savoir la lien entre VIF et s_{b_j} ;
   - Savoir ce que représente s_{x_j} et bien saisir.

6. Resampling methods

1. Les méthodes classiques statistiques posent, et testent, des hypothèses.

- › Cependant, ces méthodes sont seulement valides pour des modèles de régression linéaire.

2. Les méthodes modernes utilisent la technologie pour directement évaluer la qualité des prévisions d'un modèle avec le MSE.
   - › Ces méthodes sont valides pour tous les modèles ;
   - › Le livre rappelle la balance entre la variance et le biais pour le MSE.
3. La problématique principale est que si l'on utilise toutes les données pour établir le modèle on ne peut pas le tester. Le chapitre couvre donc 2 méthodes de séparer les données pour les tester :
   (a) **Validation set approach** : On partitionne aléatoirement l'ensemble de données en 1 ensemble d'entrainement et un ensemble de validation.
      - › Habituellement, l'ensemble de validation comporte entre 25 et 35% des données ;
      - › Un désavantage est que le modèle est ajusté sur seulement une partie des données et donc la variabilité des résultats augmente et le MSE sera une sur-estimation du vrai MSE ;
      - › Un autre désavantage est que puisque la séparation est aléatoire, le MSE peut être très variable.

   (b) **Cross-validation** : Partitionne les données systématiquement et prends la moyenne des MSE des sous-ensembles. 2 méthodes sont abordées :
      i. **Leave One Out Cross Validation (LOOCV)** : Ajuste le modèle en excluant une observation à la fois pour toutes les observations et en prends la moyenne.
      ii. **k-fold Cross Validation** : Partitionne les données en k sous-ensembles et en prends la moyenne.
4. Le livre explique les pros/cons des différentes méthodes.

Note sur les exercices :

1. Savoir différents statistiques pour out of sample et la validation croisée ;
2. Savoir le lien entre k-fold et LOOOCV et le lien avec la variance/biais.

7. Linear Regression : Subset Selection

L'importance de la section est surtout de comprendre que puisque l'ajustement sera toujours *meilleur* avec plus de paramètres, nous devons établir des méthodes de réduire le nombre de paramètres.
S'il y a p paramètres, alors il y a 2^p possibilités de modèles. Il est donc important de comprendre qu'on ne peut pas tester autant de possibilités de modèles et donc qu'on doit établir des méthodes de sélection de modèle algorithmique.
Finalement, on veut comparer des différents modèles et c'est de là que vient les différentes statistiques de comparaison de modèles qui prennent en compte le nombre de paramètres p ainsi que l'ajustement du modèle SS(résidus).

1. Sélection de sous-ensembles
   - Les deux méthodes les mieux connues sont la sélection **backward** et la sélection **forward** ;
   - Il faut savoir non seulement leurs algorithmes, mais comprendre qu'ils font partie de l'approche **stepwise** ;
   - Faut savoir et comprendre le nombre de modèles possible maximal ;
   - Faut savoir et comprendre que ces méthodes ne garantie pas le meilleur modèle ;
   - Faut savoir la **mixed method** qui est en réalité assez semblable à faire des drop1 ;
   - Lire et comprendre les différents problèmes.
2. Sélection du meilleur modèle
   - La validation croisée est la plus précise mais nécessite beaucoup de computing power ;
   - La pénalité qu'on applique avec les diverses statistiques servent à estimer l'inflation du MSE en comparaison au vrai MSE ;
   - Lire et savoir les différentes formules pour l'AIC, BIC, C_p de Mallow, R_a^2 ;

- Comprendre que la pénalité du BIC est supérieur à celle de l'AIC et donc qu'il aura tendance à conserver moins de paramètres.s

Note sur les exercices : Généralement 5 types

1. Déterminer le nombre de possibilités de modèles selon les différents algorithmes stepwise ;
2. Sélectionner un modèle selon stepwise ;
3. Calculer / isoler le C_p de Mallows ;
4. Calculer / isoler le C_a^2.

8. Linear Regression : Shrinkage and Dimension Reduction

1. **Shrinkage methods** : On fait la sélection de variables soit en réduisant les coefficients ou en les retirant.
   - Régression **Ridge** : utile pour la multicolinéarité en réduisant des coefficients ;
   - Régression **Lasso** : utile pour la sélection de variables en posant des coefficients égale à 0 ;
   - Dans les deux cas, on trouve la balance entre l'augmentation du biais et la diminution de la variance où le MSE est minimisé ;
   - Nous avons soit le **tuning parameter** λ ou le **budget alloué s**.
2. **Méthodes de réduction de dimensions** : **Au lieu** de ***réduire*** le nombre de variables nécessaire pour le modèle, on ***crée*** des **nouvelles variables** qui sont des **combinaisons linéaires** des variables originales.

   Deux méthodes abordées :

   (a) **Principal Components Regression** : méthode *non-supervisée* (sans variable réponse) d'**identifier les variables** pour lesquelles les données sont **le plus variable** ;

   Note : Ceci est principalement couvert au **chapitre 17**, et on ne devrait pas vraiment s'attendre à le comprendre ici.

- Les ϕ_{ji} sont surnommés les **loadings** et les z_{i1} les **principal component scores** ;
- Les *scores* sont la distance entre les points et les *principal components* ;
- La **régression** de principal components (PCR) est **sur les principal components** ;
- Puisqu'ils sont des moyennes pondérées de toutes les variables, PCR ne **fait pas la sélection de variable** et est semblable à la régression ridge dans ce sens ;
- Le plus de composantes, le plus faible le biais et le plus élevé la variance.

(b) **Partial Least Squares** : méthode *supervisée* ;

- Puisque la variable réponse est prise en compte, la direction n'est pas aussi bien ajusté ;
- Les prédicteurs cependant seront mieux à expliquer la réponse ;
- Réduit le biais en comparaison au PCA, mais augmente la variance et donc n'est pas globalement supérieur au PCA.

3. **The curse of dimensionality** : Adding variables to a model with many variables will cause the model to deteriorate, unless the variables are truly related to the response.
   - Si le nombre de variable est large, particulièrement si $p \geq n$, l'ajustement sera parfait met les coefficients seront mal-définis ;
   - Donc, les statistiques vont indiquer un superbe modèle alors qu'en réalité il va très mal performer.

Note sur les exercices : Généralement soit :

1. Questions qualitatives
   - Les questions d'exam semblaient être +/- juste ça ;
   - Bien saisir le lien avec le SSE et les différentes méthodes de régression ;

- Savoir et comprendre le lien entre le budget alloué s et λ ;
- Savoir et comprendre la distinction entre PCA et PLS.

2. Questions quantitatives
   - Les questions les plus semblable à ce qui pourrait être dans l'exam je crois que ce serait comme 8.14 ou 8.3/8.8, le reste j'ai vraiment l'impression que c'est plus pour qu'on saisit les concepts.

Notes sur les vidéos YouTube

Phil Chan: Introduction to the Hat matrix in regression

It is essential to visualize the plane from the video.

1. $\mathbf{H}$ is a function of the explanatory variables ; $\mathbf{H} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$.
   - It is called the **hat** matrix ;
   - It is an **orthogonal projection** matrix that « *maps a vector into the column space spanned by* $\boldsymbol{X}$ ».
2. It is called the **Hat** matrix because it puts a *hat* on $\boldsymbol{Y}$; $\hat{\mathbf{Y}} = \mathbf{HY}$.
   - The proof for this consists of realizing we can isolate $\hat{\boldsymbol{\beta}}$ in the formula.
3. It is an **orthogonal** matrix because it picks the point on the plane (column space) closest to the real observation. By definition, that will be the point directly below the value. That is to say, the ***orthogonal*** point.
4. The project matrix $\mathbf{I} - \mathbf{H} = \mathbf{M}$ maps the plan orthogonal to the previous one. That is to say, the space between the observation and the original plane.
   - Doing this, we can visually see that $\mathbf{Y} = \mathbf{HY} + \mathbf{MY}$.

Phil Chan: How is it the Hat matrix spans the column space of X? Really nice.

This video explains why the hat matrix is what it is from a different angle which really helps to better understand.

1. $\boldsymbol{H}$ is a *projection matrix* of a vector (which is a function of $\boldsymbol{X}$) onto a column space.
   - Visualise the $H = (v_1\ v_2\ \ldots\ v_n)$ where there are bars above and below the v_is to exemplify *span*.
2. We want to get $\hat{\boldsymbol{\beta}}$ from the equation $\mathbf{Y} = \mathbf{X}\hat{\boldsymbol{\beta}}$ but we can't just inverse $\boldsymbol{X}$.
   - If $\boldsymbol{X}$ were square that would mean that there would be as many observations as parameters $(n = p)$;
   - If $\boldsymbol{Y}$ were « *on the same column space as* $\boldsymbol{X}$ » then $\boldsymbol{Y}$ would be on the same plane as $\boldsymbol{X}$ which is of course almost never the case;
   - This is why we have $\hat{\boldsymbol{Y}}$ *which* ***is*** on the column space of $\boldsymbol{X}$;
   - Finally, we recall from the previous video that that point will have to be *orthogonal* to $\boldsymbol{Y}$ and deduce that distance **must be** $\mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}$.
3. The next step is to realize the inner product of the observations to this vector must be zero and therefore that $\mathbf{X}^\top(\mathbf{Y} - \hat{\mathbf{Y}}) = 0$.
   - Finally we isolate this to find $\hat{\beta} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-\mathbf{1}}\mathbf{X}^\top\mathbf{Y}$ which means $\boldsymbol{H}$ must be $(\mathbf{X}^\top\mathbf{X})^{-\mathbf{1}}\mathbf{X}^\top$.

Phil Chan: Should I look at raw, standardized, or studentized residuals? part 1 - what to look out for

1. The (raw) residual $\hat{\varepsilon}_i = \text{observation}_i - \text{predicted}_i$.
2. Want the graphics to be random.
   - If there are far points that could suggest outliers;
   - If there is a pattern with the points (for example quadratic) that could suggest a non-linear relationship / collinearity;
   - If there is a pattern with the points and we plot it against

time that could suggest (auto)correlation ;

- If there is an irregular variance that could suggest heteroscedasticity.

3. Moral of the story, better with either the studentised or standardised residuals than the raw residuals.

Phil Chan: Should I look at raw, standardized, or studentized residuals? part 2 - kinds of residuals

There are 2 main problems that explain why we use the standardised or studentised residuals instead of the raw residuals :

1. The **heteroscedasticity** assumption ***doesn't*** **mean** the **variance** will be the **same for all (raw) residuals**.
   - If that were true, we'd have a horizontal line for the (raw) residuals which is never the case.
   - The reality is thus that some (raw) residuals' variance will be larger than others'.
2. The (raw) **residuals** are **measured** in the **same units as the observations**.
   - In case this doesn't make sense, recall that $\hat{\varepsilon} = \mathbf{Y} - \hat{\mathbf{Y}}$.
   - For example, they could range from -60000 to 340000 and thus it's hard to know whether that's good or bad.
   - It also makes it difficult to compare the (raw) residuals of different models.
3. Thus we transform them by reducing by the mean and standardising by the standard deviation ; that is to say, a **standard normal distribution**.
   - It is interesting to note that the 3 residuals (raw, standardised, and studentised) will all look similar but have a different y-axis.
   - With the transformation, this means that variance will be ≈ 1 because we've converted it into a standard normal distribution.

- The intuition for the general rule that we want the residuals to be between -3 and 3 is that 99.7% of observations will be within 3 standard deviations of the mean. That is to say, $3\sigma = 3(1)$.
- Therefore, **we can** still have some observations outside this range however it doesn't necessarily mean they're outliers. We should be concerned only if there are several.

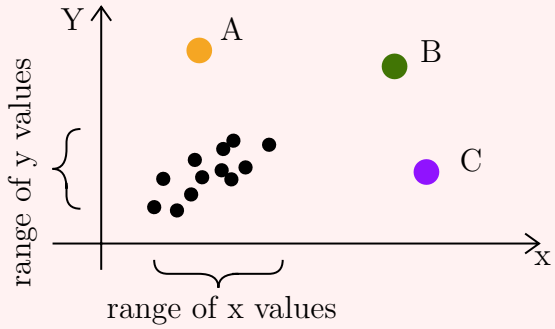
Phil Chan: Should I look at raw, standardized, or studentized residuals? part 3

Recall the formula $\text{Var}(\hat{\varepsilon}_i) = \sigma^2(1 - h_{ii})$ with $1/n \leq h_{ii} \leq 1$.
h_{ii} is the **leverage** of point i.

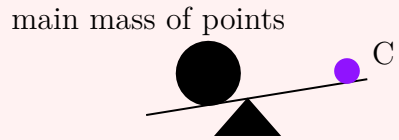
1. If h_{ii} were the same for all observations, then the variance would be constant but this is not realistic.
2. The problem in finding the variance is not with h_{ii}, which is a **function of the observations**, but rather in finding the **unknown** variance of the error term $\sigma^2 = \text{Var}(\varepsilon_i)$.
3. The difference between the standardised and studentised residuals is thus the estimation of σ^2.
   - For the standardised residuals, $\sigma^2 = s^2 = \frac{\text{SSE}}{n-p'}$;
   - For the studentised residuals, we estimate the model with the i^{th} observation excluded and obtain the variance of the error term $\sigma^2_{(i)}$ the same way **with the *new* model**.
4. It is important to understand that the variation of the residuals $\text{Var}(\hat{\varepsilon})$ can come from **either of** the 2 components of the formula.
   - Visually, it is hard to differentiate which one is causing the variance ;
   - For heteroscedasticity verification however, we're interested in the variance of the error term σ^2.

Phil Chan: What's the difference between an outlier and a leverage point in regression?

The graph below has a mass of points with an outlier (point A) and two leverage points (B and C) :



1. To understand the difference between **outliers** and **leverage points** it is important to compare them to the mass of points and ***not* the regression line**.
2. Outliers are points which are away from the main mass of points in the y direction (vertically).
   - When we *then* add the regression line, we see this means that **outliers have a large residual**.
3. Leverage points are away from the main mass of points in the x direction (horizontally).
   - There is a distinction to make between "*good*" and "*bad*" leverage points.
   - Visually, if we drew a line we can see that B is not too far out of the direction it would have whilst C is.
     - "*good*" leverage points won't "damage" the model, i.e. influence the line's direction, too much.
     - "*bad*" leverage points will. For example, C would drag the line down ; it would be an **influential point**.
   - To visualise, we can imagine a see-saw with C dragging down the main mass a bit :



Phil Chan: Influential points - Cook's distance, DFFITS, DFBETAS

1. Influential points affect both **output** and **conclusions** important ways.
2. The video covers 3 measures :
   - Cook's distance and DFFIT measure the **overall** change in paremeters when **each point in turn** is deleted.
   - DFBETA breaks it down to the individual change per parameter.
3. More generally, the level of inflation is proportional to the scaled residual (**outlyingness**) and it's **degree of leverage** in a multiplicative way.

$$\text{influence on parameters} = f(\text{leverage}) \times g(\text{outlyingness})$$

4. Finally, the important lesson is **don't just delete terms**, *reason it.*
   - The example has a model of prestige in function of income and education with a minister as an outlier.
   - If we just deleted the minister without thinking, we'd of missed an important insight into the model.

jbstatistics: Leverage and Influential Points in Simple Linear Regression

I didn't watch the whole video (because time) but it breaks down the formula for leverage very well towards the end.

1. The formula for leverage is $h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2}$.
   - Thus if the point x_i is far from the average, the denominator

wll be large causing the leverage to be high and vice-versa.

Ben Lambert: Variance Inflation Factors: testing for multicollinearity

1. The idea of the VIF is that we want to evaluate relationships between variables.
   - Usually, we would use a *correlation matrix* to compare the correlation of 2 variables and see if there's any relationship between them;
   - Usually, we would use a *scatterplot* to compare one variable to all of the others and see if there's any relationships between them;
   - The "*problem*" with both of these is that they're ***bivariate* methods only**;
   - Therefore, we want to generalize these and want to **explain one variable** as a **combination**, *or* ***linear combination***, **of the other variables**.
2. That being the idea for the **VIF**, we regress one of the explanatory variable x_j as a fonction of all the others.
3. We then find the R^2 of this new regression denoted $R^2_{(j)}$.
4. We then repeat this for all of the explanatory variables in the regression.
   - A **high value** of $R^2_{(j)}$ means it's likely there is **multicolinearity** with the ***linear* combination** of the other variables;
   - However, R^2 is hard to compare so we want to ***inflate* the differences** between the different values of $R^2_{(j)}$.
5. Therefore, we define $\text{VIF}_j = \frac{1}{1-R^2_{(j)}}$.
   - If $R^2_{(j)}$ is large, then so will the VIF_j;
   - As such, we want to minimise the VIF_j.

StatQuest: Machine Learning Fundamentals: Cross Validation

This video enables us to see the direct link between the holdout sample method, k-fold CV, and LOOCV. Also, visualise the four 25% blocks for the cross-validation.

1. If we used all the data for training, then there'd be no way to test the algorithm.
2. A *slightly* better idea would be to use 75% of the data for training and 25% for testing.
   - › The downside is that the 25% we choose is arbitrary, why one block and not another ?
3. An *even better* idea is thus to adjust the model with 3 blocks and test one the 4th one 4 times.
   - › This is **k-fold cross validation** with $k = 4$.
4. If we set $k = n$, we adjust the model and test the data for every observation. This is n-fold cross-validation, or rather **Leave One Out Cross-Validation *(LOOCV)***.
5. A **tuning parameter** is a parameter that's not estimated but guessed (like lambda in Ridge/Lasso regression). We can therefore use k-fold cross validation to help find the best value for it.

StatQuest: Regularization Part 1: Ridge Regression

1. J'ai écouté ce vidéo et le prochain pendant la session et ils sont incroyable pour bien expliquer. Cependant, pas le temps en ce moment pour les regarder encore.

StatQuest: Regularization Part 2: Lasso Regression

1. J'ai écouté ce vidéo et le dernier pendant la session et ils sont incroyable pour bien expliquer. Cependant, pas le temps en ce moment pour les regarder encore.

Prévisions et interprétations

Résumés des chapitres

9. Linear Regression : Predictions

Parameter Risk	Process Risk
intervalle de confiance	intervalle de prévision
pour la valeur moyenne	pour la valeur prédite
$\mathrm{E}[Y^*\|x^*] = \beta_0 + \beta_1 x^*$	$Y^* = \beta_0 + \beta_1 x^*$
$\hat{y}^* \pm t_{1-\frac{\alpha}{2},n-2} s \sqrt{\frac{1}{n} + \frac{(x^*-\bar{x})^2}{\sum_{i=1}^n (x_i-\bar{x})^2}}$	$\hat{y}^* \pm t_{1-\frac{\alpha}{2},n-2} s \sqrt{1 + \frac{1}{n} + \frac{(x^*-\bar{x})^2}{\sum_{i=1}^n (x_i-\bar{x})^2}}$

En régression linéaire multiple, $\widehat{\mathrm{Var}}(y^*) = s^2(1 + (\mathbf{x}^*)^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}^*)$

Note sur les exercices :

1. Les questions d'examen semble être plus axée sur utiliser les relations pour arriver à un intervalle de prévision ;
2. Ne semble pas que le multivarié serait sur l'examen donc difficile de juger si apprendre ses formules vaut la peine.

10. Interpreting Regression Results

1. Statistical significance.
   - Statistical significance $\neq$ practical significance. L'estimation du paramètre pourrait être trops faible pour avoir un vrai impact en dollars et sous ;
   - $s_{b_j} = \frac{s}{\sqrt{n-1}} \frac{\sqrt{VIF_j}}{s_{x_j}}$ est donc influencé par 4 facteurs, une augmentation de la signifiance (décroissance de s_{b_j}) peut être en raison de :
     (a) Une baisse de s qui peut être réduit en mesurant plus précisément y ;
     (b) Une baisse du VIF_j qui être réduit en utilisant des variables explicatives moins colinéaires ;
     (c) Une augmentation de $\sqrt{n-1}$ en ayant plus d'observations ;

(d) Une augmentation de s_{x_j} avec des variables plus écartées.

2. Utilités des modèles de régression.

   Déterminer si le prix d'une maison est raisonnable avec un modèle de régression linéaire simple.

3. Sélection de variables.

   Positifs du overfitting :

   › Prévisions seront sans biais. Exclure une variable pertinente peut mener à un biais.

   Négatifs du overfitting :

   › Les modèles plus simple sont plus faciles à interpréter ;

   › Les modèles plus simples vont mieux performer avec des données externes ;

   › Les variables inutiles peuvent mener à de la colinéarité ;

4. Data collection

   › **Sampling frame error** : mauvaise population est utilisée pour choisir un échantillon. (pour exemple, ceux qui achète des bonds pensent déjà vivre plus longtemps donc ce n'est pas un échantillon représentatif de la population) ;

   › Les variables explicatives pourrait être **censurées**. Des données censurées de façon significative peut mener à un biais dans l'estimation des coefficients ;

   › Les variables explicatives pourrait être **tronquées**. Ceci est un problème plus sévère que des données censurées puisqu'elles ne sont pas observées du tout ;

   › L'**exclusion** de variables explicatives peut mener à un **biais** ;

   De plus, ceci peut mener à l'inclusion de variables endogènes (dépendantes sur d'autres variables) comme l'exemple de séries chronos.

   › Il peut avoir un problème de **données manquantes**.

Modèles linéaires généralisés

Résumés des chapitres

11. Basics

1. Famille exponentielle linéaire
   - Paramètres de dispersion et d'**échelle** ;
   - Savoir Tweedie ;
2. Fonction de lien
   - Modéliser *une fonction* de la moyenne - prédicteur linéaire ;
   - L'estimation du GLM est sans biais lorsque la fonction de lien canonique est utilisée ;
3. Estimation
   - Matrice d'information Fisher ;
   - Les dérivées partielles sont les *scores* ;
4. Sur-dispersion
   - Estimation du paramètre de dispersion via la statistique du khi-carré de Pearson ;

Note sur les exercices :

1. Trouver des fonctions d'une distribution t.q. $\text{Var}(Y)$, $\text{V}(\mu)$, $b(\theta)$, etc. ;
2. Déterminer si une distribution fait partie de la famille exponentielle (domaine, etc.) ;
3. **Trouver la prévision de l'espérance ou la variance** à partir de données ;

   Vraiment beaucoup de ce dernier type pour MAS-I / S et c'est **vraiment facile**, donc probable que ce serait dans SRM ;
4. Questions à choix multiple.

12. Categorical Response

1. Binomial (binary) response
   - η est le **systematic component** ;
   - Dans un **logistic model**, $\eta = \ln\left(\frac{\pi}{1-\pi}\right)$;
   - Idée du *treshold interpretation* avec $y^* = \eta + \varepsilon$ que l'on isole ;

   On devrait s'attendre à peu, ou aucune, questions sur ces 2 sujets (réponse ordinale et nominale) et l'auteur suggère qu'on **peut même les sauter en entier si on est pressé**.
2. Nominal response
   - Généralisation du binary response avec plus de catégories, $c > 2$;
   - Peut calculer le **relative odds** de la catégorie j à la catégorie de base c.
3. Ordinal response
   - Les catégories ont un ordre ;
   - Idée du odds ratio qui est maintenant cumulatif et peut soit avoir un coefficient différent pour chaque catégorie ou pas (sauf l'intercepte qui sera toujours différent).

Note sur les exercices : Personnellement j'en ai arraché pas mal dans ce chapitre avec la réponse nominale / ordinale. Ceci dit, peu probable que ce sera dans l'exam donc c'est ça.

1. Réponse binaire généralement :
   - Exercices qu'il faut trouver le odds ratio (5, 9 à 12, 17) ;
   - Exercices qu'il faut trouver une prob / une différence de probs (4, 6 à 8, 13 à 16).
2. Réponse nominale généralement :
   - Exercices qu'il faut trouver le odds ratio (19 à 20) ;
   - Exercices qu'il faut trouver une prob / une différence de probs (21 à 24).
3. Réponse ordinale généralement :

- Exercices qu'il faut trouver le odds ratio / cumulative odds ratio / relative odds ratio (26, 27, 32) ;
- Exercices qu'il faut trouver une prob / une différence de probs (25, 27, 29 à 31, 33 à 35) ;
- J'ai éprouvé beaucoup de difficulté à bien comprendre comment isoler la probabilité et/ou le odds ratio pour le lien logit ; mais, peu probable que ce soit dans l'examen donc à votre guise.

13. Count Response

1. Poisson response
2. Sur-dispersion et modèles avec binomiale négative ;

   Noter la formule pour estimer le paramètre de dispersion et le lien avec celle de chapitre 11.
3. Other count models

   Noter que les formules pour la l'espérance et la variance sont du même format pour les trois.

   (a) **Zero-inflated** models ;

$$\Pr(Y = j) = \begin{cases} \pi + (1-\pi)h(0) & j = 0 \\ (1-\pi)h(j) & j > 0 \end{cases}$$
$$\mathrm{E}[Y_i] = (1-\pi_i)\mu_i$$
$$\mathrm{Var}(Y_i) = (1-\pi_i)\mu_i + \pi_i(1-\pi_i)\mu_i^2$$

   (b) **Hurdle** models : Rationale is that the response is the result of a two-step process :
   - The decision to make the count greater than 0 (the hurdle) ;
   - Determining the non-zero count.

$$\Pr(Y=j) = \begin{cases} \pi & j=0 \\ kh(j) & j>0 \end{cases}$$

$$\mathrm{E}[Y_i] = k\mu_i$$

$$\mathrm{Var}(Y_i) = k\mu_i + k(1-k)\mu_i^2$$

De plus, on trouve :

$$k<1 \Rightarrow 1-k>0 \therefore \mathrm{Var} > \mathrm{E}$$

$$k>1 \Rightarrow 1-k<0 \therefore \mathrm{Var} < \mathrm{E}$$

où $k = \frac{1-\pi}{1-h(0)}$

(c) **Heterogeneity** models ;

Lorsque $\{Y_i|\alpha_i\} \sim$ Poisson, on a :

$$\mathrm{E}[Y_i] = \mu_i$$

$$\mathrm{Var}(Y_i) = \mu_i + \mathrm{Var}(\mathrm{e}^{\alpha_i})\mu_i^2$$

(d) **Latent** models ;

Note sur les exercices :

1. Poisson
   - Une bonne question un peu plus tricky serait une comme le numéro 5 ;
   - De plus, le numéro 9 est une bonne pratique pour question que je crois pourrait être dans l'examen ;
   - Aussi, des questions classiques comme trouver la variance/espérance/offset/... ce à quoi les questions d'examen semble plus se rapprocher.
2. Autres
   - Ça revient pas mal toujours à soit trouver une prob (hurdle et poisson gonflée à zéro) ou trouver la variance / espérance (tous) ;
   - Pour apprendre les formules des variances / espérance, ça devient facile en voyant la pattern commune aux trois ! ;
   - Dans tous les cas, il n'avait pas de questions d'exam sur ces 4 modèles et la feuille de formule de Coaching n'inclut pas les formules pour.

14. Measures of Fit

Note importante : Lorsqu'on test si un modèle est une simplification adéquate d'un autre, on teste (pour exemple) :

$$\mathcal{H}_0 : \beta_1 = \beta_2 = 0 \qquad \mathcal{H}_1 : \beta_1 \neq 0 \cup \beta_2 \neq 0$$

Mais ceci *n'est* ***pas*** *un test bilatéral* ! En réalité, on test :

$$\mathcal{H}_0 : p = 0 \qquad \mathcal{H}_1 : p > 0$$

Ce qui est un test unilatéral et donc on veut que la statistique du TRV soit $> \chi^2_{q,1-\alpha}$. Le seuil n'est pas divisé par deux.

1. **Pearson chi-square** : Pour évaluer la qualité d'ajustement d'un modèle.
   - Si les données sont groupées on utilise la première définition avec le $X^2 = \sum_{i=1}^{n} \frac{(n_i - n\hat{p}_i)^2}{n\hat{p}_i}$;
   - Si les donnés sont par observations (indépendantes) on utilise $X^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\phi \mathrm{V}(\hat{\mu}_i)}$.
2. **Likelihood Ratio Test *(LRT)***
   - On peut voir le test du rapport de vraisemblance comme l'équivalent du test F pour les GLMs ;

     On évalue si un modèle, ou un ensemble de ses paramètres, est significatif s ;
   - Statistique : $LRT = -2(\tilde{\ell} - \hat{\ell}) \approx \chi^2_q$;

     q est le nombre de contraintes (paramètres à retirer)

     $\hat{\ell}$ est la log-vraisemblance du modèle sans contraintes **(complet)** ;

     $\tilde{\ell}$ est la log-vraisemblance du modèle avec contraintes **(réduit)** ;
3. **Deviance** : Pour évaluer la qualité d'ajustement d'un modèle ;

   $D = 2\phi \left(\ell(\boldsymbol{b}_{\text{saturé}}) - \ell(\boldsymbol{b})\right) \approx \chi^2_{n-p'}$.
   - On peut voir la déviance comme la *déviance*, ou **l'écart**, entre les observations et les prévisions.

Si ceci semble familier, c'est puisque c'est l'interprétation du SSE en régression linéaire simple ;

Donc, *(**je** pose que)* on peut voir la déviance comme étant l'équivalent aux GLMs **de la famille exponentielle** du SSE ;

- La loi normale exemplifie ceci avec une déviance = SSE ;
- Par la suite, les déviances des distributions Bernoulli et Poisson deviennent logique avec ceci en tête :

$$D = -2\left(\sum_{y_i=1}\ln(\hat{y}_i)\sum_{y_i=0}\ln(1-\hat{y}_i)\right)$$
$$D = 2\sum_{i=1}^{n}\left[y_i\ln\left(\frac{y_i}{\hat{y}_i}\right) - (y_i - \hat{y}_i)\right]$$

  De plus, lorsqu'un lien log et que le coefficient β_0 sont utilisés, et que la somme des résidus est nulle ($\sum(y_i - \hat{y}_i) = 0$) on obtient $\ln(1) = 0$. Ce faisant, la deuxième partie de la formule de la déviance pour la Poisson peut être omise ;
- Plus formellement, le **modèle saturé** est le modèle avec le *meilleur ajustement possible*. C'est-à-dire qu'il y a un paramètre par observation ;
- La **scaled deviance** est la statistique comparant le modèle saturé au modèle proposé ;
- La **déviance** est ré-obtenue en multipliant le paramètre d'échelle ϕ avec la scaled deviance ;
- On généraliser également le TRV, $LRT = -2(\tilde{\ell} - \hat{\ell}) = (\tilde{D} - \hat{D}) \approx \chi^2_q$;
  - Donc, avec la déviance on appel le modèle complet / sans contraintes le *modèle saturé* puisqu'on compare notre modèle à celui de base au lieu de comparer 2 modèles ;
  - Pareillement, lorsqu'on récrit le TRV avec la déviance on retourne à la notation réduit vs complet (avec vs sans contraintes) ;

- Finalement, on note que la log-vraisemblance du modèle complet (saturé) devrait être supérieure à celle du modèle réduite ;

4. Penalized loglikelihood tests

$$AIC = -2\ell(\hat{\boldsymbol{\beta}}) + 2p' \qquad BIC = -2\ell(\hat{\boldsymbol{\beta}}) + \ln(n)p'$$

   - On rappel que $\hat{\boldsymbol{\beta}}$ est le vecteur des $p' = p + 1$ coefficients et donc la pénalité est de p' ;
   - Cependant, s'il a également un paramètre de dispersion ϕ à estimer alors la pénalité est $p + 2$;

5. Max-scaled R^2 and pseudo-R^2

$$R^2_{ms} = \frac{1 - \mathrm{e}^{2\left(\ell_0 - \ell(\hat{\beta})\right)/n}}{1 - \mathrm{e}^{2\ell_0/n}} \qquad R^2_{pse.} = \frac{\ell_0 - \ell(\hat{\boldsymbol{\beta}})}{\ell_0 - \hat{\ell}}$$

   - Le « *problème* » avec le max-scaled R^2_{ms} est qu'il sera uniquement 1 pour un modèle parfait puisqu'avec un modèle parfait, $\ell(\hat{\boldsymbol{\beta}}) = 0$ et donc le ratio est de 1 ;
   - Dans le cas de régression linéaire simple, le pseudo $R^2_{pse.}$ se réduit à uniquement R^2 ;

6. Résidus
   - Utilités :
     - Identifier les covariantes et/ou les pattern ;
     - Identifier les outliers ;
     - Illustrer l'hétéroscédasticité / tendances temporelles ;
     - Individuellement illustrer l'impact d'une variable explicative sur le modèle ;
   - **Pearson**
   - **Deviance**
   - *Anscombe*

Note sur les exercices :

1. **Pearson chi-square** (3 -> 1 past/sample exam questions)

- Bien distinguer le cas de données groupées du cas de données individuelles ;
- Noter que la fonction de variance est celle pour la famille exponentielle dans la 2e définition ;

2. **LRT et deviance** (12 -> 5 past/sample exam questions)
   - Beaucoup d'anciennes questions d'examens donc à savoir ;
   - Très utile de connaître les formules de la déviance pour les 3 distributions ainsi que la simplification pour la Poisson sinon faut recalculer à chaque fois ;
   - Bien saisir les seuils et les tests d'hypothèse pour la khi-carré ;
3. **AIC** et **BIC** (12 -> 6 past exam questions)
   - Ces questions sont plutôt difficiles et ont apparu systématiquement dans les examens de la S/MAS-I. Ce faisant, il me semble logique qu'elles apparaissent dans SRM ;
   - Je conseille de noter ce type d'exercices comme des exercices à pratiquer puisqu'il faut vraiment connaître les liens entre le TRV/AIC/BIC comme sa poche ;
   - Également, faut comprendre l'impact et le raisonnement sous-jacente à la pénalité des paramètres ;
4. Max-scaled R^2 and pseudo-R^2 : savoir les formules. (4)
5. Résidus : savoir les formules. (4)

Notes sur les vidéos YouTube

StatQuest: Logistic Regression

1. Logistic regression predicts whether something is T/F, instead of predicting something continuous like size.
2. Instead of fitting a line, we fit a "s" shaped logistic function.
   - The curve therefore goes from 0 to 1 and **predicts a probability** that a mouse is obese based on its weight ;

- If we weighed a heavy mouse, a high probability it's obese ; if weighed a light mouse high probability it's not obese.

3. Logistic regression usually **used for classification**.

   For example, if the probability that a mouse is obese is $> 50\%$ then we'll **classify** it as obese.

4. We can make simple models.

   For example, Obesity in function of Weight.

5. We can have continuous, categorical, etc. types of variables.
   - We test if a variable's effect on the prediction is significantly different from 0 but can't directly compare 2 models like in linear regression (Wald's test).

6. Logistic regression **fits the line differently** than linear regression
   - In linear regression, we fit with least squares where the sum of the residuals is minimised ;
   - In logistic regression we don't have the same concept of a "residual" and thereby can't use least squares nor calculate R^2 but uses maximum likelihood instead ;
   - Calculate likelihood of observing each data point and multiply them together ;
   - Then shift the line and repeat until the curve with the maximum likelihood is found.

StatQuest: Probability vs Likelihood

1. Probability is what we're used to : Prob(mouse weighs 34 grams | mean = 34, standard deviation = 2.5) = **Pr(data | distribution)**.

   We find the most likely observations given some parameters for the curve.

   It's the area under a fixed distribution.

2. Likelihood : L(mean = 34, standard deviation = 2.5 | mouse weighs 34 grams) = **L(distribution | data)**.

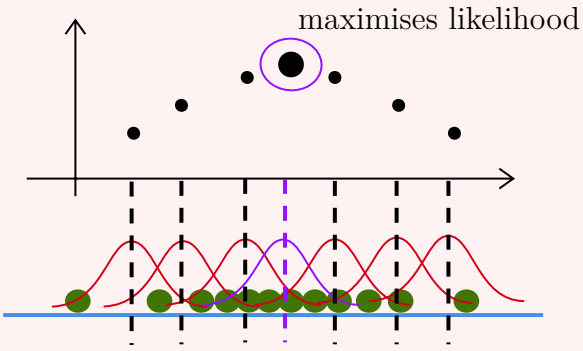
We find the most likely curve given some oboservation(s).

Likelihoods are the y-axis value for fixed data points with distributions which can be moved.

StatQuest: Maximum Likelihood, clearly explained!!!

1. Depending on where we plot the normal density curve, the probability of observing our data points can go up or down.
2. We plot the curve around the mean which we then shift around.
3. We shift the location of the mean, around which the curve is centered, until we find the location that maximises the **likelihood** of observing the weights we measured.

   This is the **maximum likelihood estimate for the mean** (of the distribution, not data but for a normal those are equal).
4. The MLE for the standard deviation is the same procedure.



3 Time Series Models (12.5% à 17.5%)

Information

Objective

Understand key concepts concerning regression-based time series models.

Learning outcomes

1. Définir et expliquer les concepts, et composantes, des processus de séries chronologiques stochastiques. Incluant :
   - Les marches aléatoires
   - Stationnarité
   - Auto-corrélation
2. Décrire des modèles de séries chronologiques précis dont :
   - Exponential smoothing
   - Autoregressive
   - Autoregressive conditionally heteroskedastic models
3. Calculer et interpréter les valeurs prédites et leurs intervalles de confiance.

Related lessons ASM

19. Time Series : Basics
20. Time Series : Autoregressive Models
21. Time Series : Forecasting Models

Vidéos YouTube

- Dr. Daniel Soper: Visualizing Random Walks in Three Dimensions
- Ben Lambert: Time series vs cross sectional data

- Ben Lambert: Time series Gauss Markov conditions
- ritvikmath: Time Series Talk : Autocorrelation and Partial Autocorrelation
- ritvikmath: Time Series Talk: Stationarity
- ritvikmath: Time Series Talk : White Noise
- ritvikmath: Time Series Talk : Autoregressive Model

Résumés des chapitres

19. Time Series : Basics

Time series -> Séries chronologiques

1. **Introduction**

   Séries chronologiques

   - Série d'observations $y_1, y_2, \ldots, y_T$ sur des intervalles de temps consécutives ;
   - L'analyse de séries chronologiques consiste à essayer de déchiffrer des patterns dans les séries pour en faire des prévisions ;

     Ces patterns peuvent prendre plusieurs formes y inclut de relier les observations y_t aux observations précédentes de la série ou, au temps lui-même ;

   Données

   - **Longitudinales** : Données d'un processus variant avec le temps ;
   - **Transversales (cross-sectional)** : l'inverse, elles ne sont pas organisées chronologiquement ;
   - Pour bien saisir la distinction, on peut visualiser des données **transversales comme une image** et des **longitudinales comme une vidéo** ;
   - La différence entre des données chronologiques vs transversales est bien expliquée dans le vidéo de Ben Lambert ;

Modèle de

- **causalité** : Sans variables explicatives reliées au temps ;

  Négatif : les modèles de causalité dépistent la **corrélation** et *pas* la causalité ;

  Négatif : les modèles de causalité doivent connaître les valeurs des variables indépendantes pour faire des prévisions sur la variable dépendante ;

- **série chronologique** : L'inverse.

  Positif : Les modèles de séries chronologiques peuvent trouver des relations causales avec le temps ;

3 composantes aux séries chronologiques :

(a) **Trend** T_t : Long term pattern des données ;

(b) **Seasonality** S_t : Cyclical pattern des données ;

(c) **Random patterns** ε_t ;

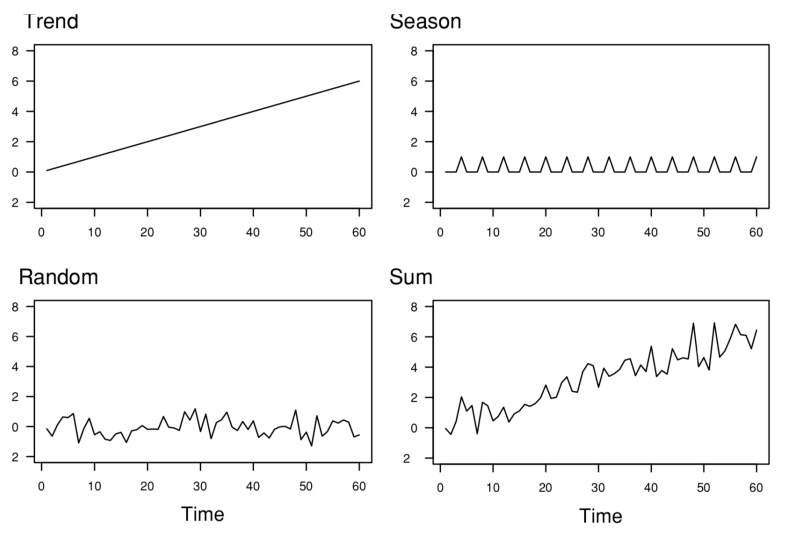
Tendance

(a) Additive : $y_t = T_t + S_t + \varepsilon_t$

(b) Multiplicative : $y_t = T_t \times S_t + \varepsilon_t$

Ajustement et évaluation :

- Lorsqu'il y une tendance, ou un cycle, **saisonnale** on peut ajuster le modèle avec une variable binaire ;
- Lorsqu'il y a une transition (**Regime change**) on peut ajuster le modèle avec une variable binaire égale à 0 avant le point fixé et 1 après ;
- Un scatter plot du modèle contre le temps, i.e. un **time series plot**, permet d'évaluer la série chronologique. Par exemple :



Problèmes :

› On surnomme parfois les prévisions de séries chronologiques comme étant des prévisions « **naïves** ».

Les prévisions des modèles de série chronologique prennent en compte uniquement les observations passées pour faire des prévisions ;

Elles ne sont ***ni* ajustés** *ni* **évaluées** pour des **relations causales**.

Ce faisant, elles ne sont pas très représentatives de la réalité et sont ***naïves*** ;

› De plus, le modèle de régression va allouer le **plus de poids aux observations** avec des variables explicatives **anormalement élevées** ;

On peut visualiser que si le temps était une variable explicative, alors les données le plus récentes et les plus anciennes seraient les plus éloigné du temps moyen ;

Par exemple, si le temps va de 1 à 10 alors le temps moyen est de 5.5 et les points au début et à la fin sont les plus éloignés ;

Le modèle va donc attribuer **plus de poids** aux observations **récentes et anciennes** qu'aux autres, ce qui n'est pas idéal ;

Ce problème est adressé avec du **smoothing** au chapitre 21 ;

Incertitude des prévisions :

- Avec des données transversales, le plus éloigné un point est de la masse principale, le moins nous sommes certain des prévisions ;

  Ce faisant, nous sommes majoritairement intéressés aux prévisions sur des points près de la masse principale ;
- Pareillement, avec des données chronologiques, le plus loin dans le futur une prévision, le moins nous en sommes certains ;

  Ce faisant, nous sommes majoritairement intéressés aux prévisions d'évènements pas trop loin dans le temps ;

2. **Moyenne** et **variance**

   Stationnarité

   - Si la **moyenne** *(au temps t)* $\mathrm{E}[y_t] = \mu(t)$ ne varie pas dans le temps, la série est **stationnary in the mean** ;

     Ce faisant, on peut l'estimer avec les valeurs observées ;
   - Pareillement, si la **variance** *(au temps t)* $\mathrm{Var}(y_t) = \sigma^2(t)$ ne varie pas dans le temps, la série est **stationnary in the variance** ;

     Ce faisant, on peut l'estimer avec la variance échantillonnale ;
   - Cependant, lorsque les données sont corrélées (ce à quoi on s'attend) alors la variance échantillonnale va **sous-estimer** la vraie variance ;
   - L'**autocorrelation, ou *serial correlation*)** est mesuré par la corrélation échantillonnale ;

     À noter que la sous-estimation de la variance devient de moins en moins importante plus la taille de la série augmente ;

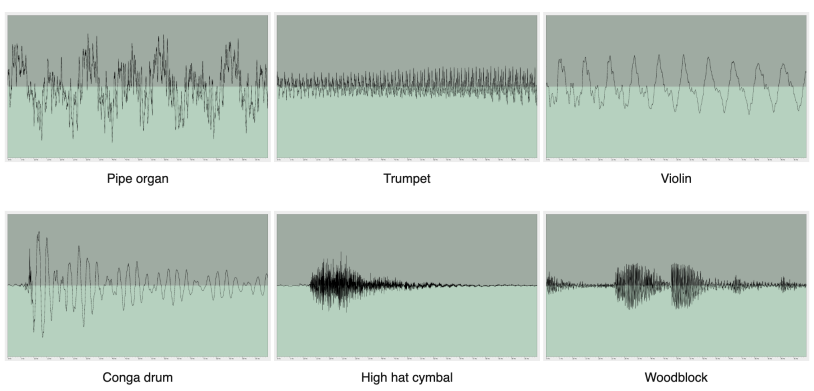
- Le distance k des observations d'une série chronologique au temps t avec la même série au temps $t + k$ est le **lag** ;
- Si la série a une espérance stationnaire et que la variance et corrélation sont fonction seulement du *lag*, et non du temps lui-même, alors la série est dite ***(weakly)* stationary (par défaut)** ;
- Si aucun des moments plus élevés varient avec le temps, alors la série est ***strongly* stationary** ;

3. **White noise** : Une série chronologique **white noise** w_t est caractérisé par :
   - L'indépendance des termes ;

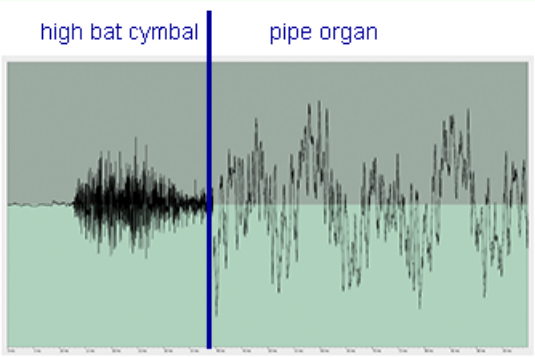
     Ce faisant, la corrélation r_k à tous **lag** supérieurs à 0 est de 0 (r_0 est toujours égale à 1) ;
   - Une espérance constante μ_w ;

     Habituellement, elle est nulle ;
   - Une variance constante σ_w^2 ;
   - Habituellement, on suppose également la normalité ;

   Réduction de série alias **time series *filtering*** :
   - Nous cherchons toujours à réduire toute série chronologique à une série white noise en **identifiant des patterns** ;
   - Pour visualiser ce que le time series filtering représente, on observe ces graphiques d'ondes sonores d'instruments :



Et si nous avions une combinaison de ces fréquences :



- Cette idée d'observer des tendances dans des graphiques d'ondes sonores est exactement la même idée que le *time series filtering* ;

  On cherche à identifier des tendances, des cycles, des changements de régime, etc. dans une série qui peuvent sembler d'être aléatoires ;

- Dans le même ordre d'idée, lorsqu'un musicien joue son instrument on imagine que ce n'est pas parfait et qu'il y a un certain niveau d'erreur !

  Ceci est donc le **white noise**, l'erreur aléatoire de la fréquence observée à la fréquence attendue ;

  Alors, pour néanmoins reconnaître une fréquence d'instrument imparfait, on alloue une variabilité (ou déviance) à la série « parfaite » ;

- La partie inexplicable par les patterns qui demeure est dite d'être **irréductible** et est donc traitée comme du ***white noise*** ;

4. **Marches aléatoires** : Une **marche aléatoire** est :
   - Une série chronologique ***non*-stationnaire** ;
   - Ayant comme valeur le niveau initial y_0 en plus de l'accumulation du *white noise* w_t ;
   - L'expression est donc de la forme : $y_t = y_{t-1} + w_t, t \geq 1$.

Notes sur l'*espérance* et la *variance* :

- Si $\mu_c = 0$ la série est néanmoins non-stationnaire puisque la variance croît avec le temps ;
- Si $\mu_w \neq 0$ la série est une marche aléatoire **avec drift** où μ_w est le drift ;
- On déduit de plus que la différence entre marches aléatoires $y_t - y_{t-1}$ est du white noise ;

Pour distinguer et comprendre la distinction entre une tendance dans le temps (à gauche) et une marche aléatoire (à droite) on compare les deux :

$$y_t = y_0 + tk + \varepsilon_t \qquad \text{vs} \qquad y_t = y_0 + t\mu_w + \sum_{j=1}^{t} \varepsilon_j$$

D'où on peut voir le lien entre l'erreur d'un modèle causal et d'un modèle chronologique. Malgré une espérance identique, si $k = \mu_w$ la variance de la série chrono va croître avec le temps ;

Exemples théoriques de **filtering** :

- Différencier une marche aléatoire ;
- Prendre le log de la série. Ceci peut stabiliser la variance ;
- Finalement, prendre le log de la différence des séries qui correspond *approximativement* à des changements proportionnels ;

5. **Control charts** : Graphiques pour observer l'évolution d'un processus avec le temps.
   - Les données sont placées en ordre chronologique ;
   - Le graphique comporte une ligne centrale et deux lignes pour les bornes supérieures et inférieures (typiquement Q1 et Q3) ;
   - L'axe des y sert à évaluer la moyenne du processus et l'axe des x permet d'observer l'étendu ;
   - Ce faisant, l'utilité est un peu semblable à celle d'un box-plot :
     - On évalue si un processus est relativement stable en observant leur distribution relative au temps ;

- En revanche, un boxplot permet vérifier que les données n'ont pas trop de points aberrants et qu'elles ne sont pas trop répandues ;
- On peut donc évaluer si un processus est stable (en contrôle) ou instable (hors de contrôle) ;

Xbar charts : Calculer la moyenne des séries de k observations.

› Pour exemple avec $k = 5$, on calcule la moyenne des observations 1 à 5, 6 à 10, 11 à 15, etc. ;

› Puisque la variance des moyennes sera plus faible que la variance de la série, des tendances inhabituelles devraient être plus apparentes ;

R charts : Calculer l'étendu des séries de k observations.

› L'étendu est la valeur maximale - la valeur minimale ;

› Ceci donne une idée de la dispersion des données pour évaluer la volatilité et évaluer des patterns ;

Exemple de xbar plot :

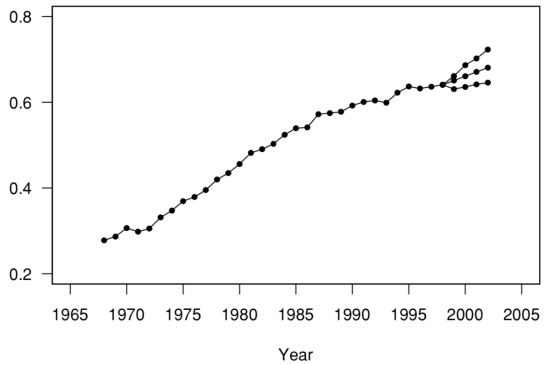


Figure 7.8 Time series plot of labor force participation rates with forecast values for 1999– 2002. The middle series represent the point forecasts. The upper and lower series represent the upper and lower 95% forecast intervals. Data for 1968– 98 represent actual values.

6. **Evaluating forecasts** *(évaluer les prévisions)*

   On utilise le **out-of-sample validation**

   › Les données sont partitionnées jusqu'au point fixe T_1 où $T_1 < T$;

   › Ces données sont utilisés pour ajuster le modèle et le reste pour le tester ;

   › Avec les données de test on calcule le résidu $e_t = \hat{y}_t - y_t$;

Par la suite on évalue la qualité du modèle avec ces statistiques *(voir formules)* :

- Le **Mean Error** (ME) et **Mean Percentage Error** (MPE) détectent des **trend patterns** ;
- **Cependant**, ils ne **détectent pas** des problèmes lorsque les *résidus sont positifs et négatifs* avec une *moyenne faible* alors que les **autres mesures oui** ;
- De plus, MPE ne peut pas être utilisé lorsque la série contient des 0s et peut être incohérente avec des termes négatifs ;
- Le **Mean Square Error** (MSE) et **Mean Absolute Error** (MAE) adressent le problème de résidus nuls avec ME et le **Mean Absolute Percentage Error** (MAPE) pour le MPE ;
- Cependant, le MAPE a les même problèmes que le MPE avec les termes nuls ;

Note sur les exercices :

1. Systématiquement les examens de la CAS ont une question où il faut calculer le sample correlation pour un tableau de données dont fort probable ce sera dans SRM (1 à 6) ;
2. Question qu'il faut trouver la prévision ou l'écart-type (d'ailleurs, 2 des sample questions de SRM portent sur ceci) ;
3. Question qualitative sur les propriétés (stationnarité, prévisions, homogénéité, etc.) ;
4. Question comme 17 à 20 dont il faut calculer une des statistiques pour un random walk ou du white noise ;
5. Avec un peu de réflexion la logique derrière les 5 formules devient évidente et il n'est pas nécessaire de les mémoriser ;

Formules chapitre 19

$$s^2 = \frac{\sum_{t=1}^{n}(y_t - \bar{y})^2}{n-1} \qquad r_k = \frac{\sum_{t=k+1}^{T}(y_{t-k} - \bar{y})(y_t - \bar{y})}{\sum_{t=1}^{T}(y_t - \bar{y})^2}$$

$$r_0 = 1$$

$$ME = \frac{1}{T - T_1}\sum_{t=T_1}^{T} e_t \qquad MPE = \frac{100}{T - T_1}\sum_{t=T_1}^{T} \frac{e_t}{y_t}$$

$$MSE = \frac{1}{T - T_1}\sum_{t=T_1}^{T} e_t^2$$

$$MAE = \frac{1}{T - T_1}\sum_{t=T_1}^{T} \left|e_t\right| \qquad MAPE = \frac{100}{T - T_1}\sum_{t=T_1}^{T} \left|\frac{e_t}{y_t}\right|$$

Une marche aléatoire est définie par $y_t = y_{t-1} + w_t$ lorsque $t \geq 1$. Ce faisant, on peut définir la série de white noise $w_t = y_t - y_{t-1}$.
Pour une série white noise w_t : Pour une marche aléatoire y_t :

$$\mathrm{Var}(w_t) = \sigma_w^2 \qquad \mathrm{Var}(y_t) = t\sigma_w^2$$
$$\mathrm{E}[w_t] = \mu_w \qquad \mathrm{E}[y_t] = y_0 + \underbrace{t\mu_w}_{\text{drift si } \mu_w \neq 0}$$

Measure	White noise	Random walk
Estimated standard error	$se_{\hat{y}_{T+l}} = s_y\sqrt{1 + \frac{1}{T}}$	$se_{\hat{y}_{T+l}} = s_w\sqrt{l}$
l-period lookahead forecast	$\hat{y}_{T+l} = \bar{y}\ \forall l$	$\hat{y}_{T+l} = y_T + l\bar{w}$
Forecast interval of y_{T+l}	$\bar{y} \pm t_{T-1,1-\frac{\alpha}{2}} s_y\sqrt{1 + \frac{1}{T}}$	
Approximate 95% forecast interval of y_{T+l}	$\bar{y} \pm 2s_y$	$\bar{y} + l\bar{w} \pm 2s_w\sqrt{l}$

où $\bar{w}$ est la moyenne échantillonnale de w_t.

20. Time Series : Autoregressive Models

Pas encore fait, notes pour me donner une idée du chapitre

1. Intro
   - Un **autoregressive model** d'ordre 1, AR(1), est une série chronologique dont chaque terme peut être exprimé en fonction du précédent :

     $$y_t = \beta_0 + \beta_1 y_{t-1} + \varepsilon_t$$

   - Pour exemple, si $\beta_1 = 0$ alors c'est une série white noise et si $\beta_1 = 1$ c'est une marche aléatoire ;
   - On rappel qu'une série white noise est stationnaire, donc on veut que $0 < |\beta_1| < 1$ puisque lorsque $|\beta_1| < 1$ la série est stationnaire ;
   - Voir la formule de la **vraie auto-corrélation** au lag k ;
   - Coefficients ;
2. Forecasting with AR(1) series

Note sur les exercices :

1.

Formules chapitre 20

La **vraie auto-corrélation** au lag k pour un processus AR(1) stationnaire :

$$\rho_k = \beta_1^k$$

21. Time Series : Forecasting Models

1. Moving average smoothing
   - Average k consecutive terms and create a new time series
   - Moving average smoothing is weighted least square with weights of 1 on the most recent k periods and 0 on earlier periods ;

2. Exponential smoothing
    - Moving average weights are constant for the k previous periods and then drop off to 0 ;
    - To have a smoother pattern of weights, use exponential smoothing to put weights that're proportionnal ;
    - Impact of different weights ;
3. Seasonal models
    - Fixed effects ;
    - Autoregressive seasonal models ;
    - Seasonal exponential smoothing ;
4. Unit root tests
5. ARCH and GARCH models

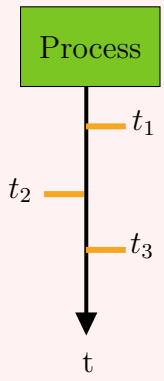
Note sur les exercices :

1.

Notes sur les vidéos YouTube

Ben Lambert: Time series vs cross sectional data

Le vidéo est utile pour commencer à conceptualiser la différence en approche des données auxquelles nous sommes habituées (*cross-sectional data*) des données de séries chronologiques.
Pour ce, et les prochains vidéos de Ben Lambert, $u_t = \varepsilon_t$.



- With cross-sectional data, we can think of the data as a population from which we take random samples ;
- With time series data however, it is not so simple ;
- Time series data is actually a **process** that evolves with time ;
- Thereby, when we see observations of the process, we're really sampling the process at different time (for example, $t_1, t_2, t_3, \dots$) ;
- The « *population* » is actually all the observations for all the possible periods of in time but that is quite untangible ;
- Thus, we have more specific conditions than normal cross-sectional data with regards to independance, etc.

Ben Lambert: Time series Gauss Markov conditions

Rappel : $u_t = \varepsilon_t$.
On a appris en modèle à propos des 4 hypothèses pour un LM et que la 4ème n'était pas *vraiment* nécessaire. En réalité, les 3 premières sont le théorème **Gauss-Markov** qui prouve qu'un estimateur est le **Best Linear Unbiased Estimator *(BLUE)***.
Il est important de faire cette distinction pour les séries chronologiques puisque les conditions sont légèrement différentes d'une façon importante. Entres autres, puisque les données ne sont pas indépendantes nous avons plusieurs conditions pour la corrélation / colinéarité. Ce vidéo explique le tout.

The first three prove the estimator is unbiased.

1. Linearity : $Y_t = \alpha + \beta_1 x_{1t} + \beta_2 x_{2t} + \varepsilon_t$.

   Thus we have $p = 2$.

2. Nul expectation

$$\mathrm{E}[\varepsilon_t | x_{pk}] = 0 \qquad \text{vs} \qquad \mathrm{E}[\varepsilon_i | x_{pi}] = 0$$

- In linear regression, we only need the i^{th} observation's expected value to be 0 for all p parameters ;
- In time series however, we need the expected value of the observation over **all possible time intervals** k to be 0.

3. No perfect colinearity (i.e. no multicolinearity)
4. Homoscedasticity

   The same idea holds for the variance.

$$\text{Var}(\varepsilon_t|x_{pk}) = \sigma^2 \qquad \text{vs} \qquad \text{Var}(\varepsilon_i|x_{pi}) = \sigma^2$$

5. No serial correlation : $\text{cov}(\varepsilon_t, \varepsilon_s|x_{pk}) = 0$

   The same idea holds again for the correlation.

With these conditions, we prove the OLS is BLUE.

4 Principal Component Analysis (2.5% à 7.5%)

Information

Objective

UUnderstand key concepts concerning Principal Components Analysis

Learning outcomes

1. Définir « principal components »
2. Interpréter les résultats d'une analyse par composantes principales en prenant en compte :
   - « Loading Factors »
   - Proportion de la variance expliquée
3. Expliquer les applications de l'analyse par composantes principales

Related lessons ASM

15. K-Nearest neighbors
17. Principal Component Analysis

15. K-Nearest neighbors

Pas encore fait, notes pour me donner une idée du chapitre

1. The Bayes classifier
   - Error rate with and without the Bayes classifier ;
   - It is the best classifier ;
   - Bayes decision boundaries ;
2. KNN classifier
   - Caveat with Bayes is we don't know the probabilities

$\Pr(Y = j|x_0)$;

› Estimate them with the KNN classifier ;

3. KNN regression

› KNN is non-parametric ;

› In KNN regression K is selected and the value of the response at any point is the average of the values at the K nearest observations ;

› Comparision to other methods ;

› Complexity of interpretation vs parametric methods ;

Note sur les exercices :

1.

17. Principal Component Analysis

Pas encore fait, notes pour me donner une idée du chapitre

1. Loadings and scores

› PCA's an unsupervised method for visualising data ;

› Creates principal components that summarize correlated variables ;

These are linear combinations of the existing variables ;

› Loadings ϕ_{ji} : the p weights on the variables X_j in the expression for Z_i ;

› Scores z_{ki} : p coordinates of the observations in the Z coordinate system ;

Score i for observation k is the distance of point k from 0, in the Z_i direction ;

2. Biplots

›

3. Approximation and scaling

› Another interpretation of principal components is that they are the best linear approximations of the observations ;

› Scale matters in PCA unlike linear regression ;

4. Proportion of variance explained (PVE)
    - Cross-validation isn't possible for unsupervised methods so how to determine the number of principal components to use ?
    - Can look at the PVE by the principal components

Note sur les exercices :

1.

Vidéos YouTube

- StatQuest: K-nearest neighbors, Clearly Explained
- StatQuest: PCA main ideas in only 5 minutes!!!
- StatQuest: Principal Component Analysis (PCA) clearly explained (2015)
- StatQuest: Principal Component Analysis (PCA) (step by step)
- StatQuest: PCA - Practical Tips
- StatQuest: PCA in R

5 Decision Trees (10% à 15%)

Information

Objective

Understand key concepts concerning decision tree models

Learning outcomes

1. Expliquer l'utilité et les applications des arbres de décisions.
2. Expliquer et interpréter les arbres de décisions en considérant les arbres de régression et le « recursive binary splitting ».
3. Expliquer et interpréter le « bagging », « boosting » et les forêts aléatoires.
4. Expliquer et interpréter les arbres de classification, leur construction, le « Gini Index » et « entropy ».
5. Comparer les arbres de décisions aux modèles linéaires.
6. Interpréter les résultats d'une « decision tree analysis ».

Related lessons ASM

16. Decision Trees

16. Decision Trees

Pas encore fait, notes pour me donner une idée du chapitre

1. Building decision trees
   - › Non-parametric alternatives to regression ;
   - › Similarity to KNN wrt/ the split of predictors into regions and assignment of an average value, ... ;
   - › **nodes** ;
     - – leave / terminal ;
     - – Intermediate ;

- Types of variables can be categorical, count, or continuous ;

  Note : a cut point must be selected for continuous variables ;
- Every split's binary ;
- Optimal **continuous** (*regression*) trees minimize the MSE ;
  - Because too many possibilities, use recursive binary splitting ;
  - Select binary split to minimise MSE ;
  - **greedy** because, much like stepwise, doesn't optimise future MSE only current split ;
  - Algo stops when number of observations below a fixed number *(e.g. 5)* ;
- Balance with the size of the tree ;

  More splits = more flexibility => more variance and less bias ;
- To optimize the size of a the tree, we *prune* it (similar to the Lasso) ;
  - *cost complexity pruning* oro *weakest link pruning* ;
  - Tuning parameter α ;
  - Cost of tree is α per terminal node ;
  - For each value of α we prune the tree to minimize <> and select optimal tree with cross-validation ;
  - higher α => smaller tree
- To optimise a **classification** tree, we minimize the classification error rate instead of the MSE ;
- The classification error rate is not sufficiently sensitive to grow the tree so we use the **Gini index** or the **cross-entropy** instead ;
- Note on measure to use for pruning ;
- Residual mean deviance ;
- Advantages over linear models
- Shortcomings ;

2. Bagging, random forests, boosting

   Questions on this would be qualitative rather than quantitative because they necessitate heavy computing ;

   (a) Bagging

   Form of boostrapping ;

   Bootstrapping is not on the syllabus, but it's needed to understand bagging ;

   (b) Random forests

   Bagged trees may be correlated and random forests corrects this I think ;

   (c) Boosting

   Only discussed for regression (continuous) setting ;

   Parameters ;

   Algorithm ;

Note sur les exercices :

1.

Vidéos YouTube

- StatQuest: Decision Trees
- StatQuest: Decision Trees, Part 2 - Feature Selection and Missing Data
- StatQuest: Regression Trees, Clearly Explained!!!
- StatQuest: How to Prune Regression Trees, Clearly Explained!!!

6 Cluster Analysis (10% à 15%)

Information

Objective

Understand key concepts concerning cluster analysis

Learning outcomes

1. Expliquer les utilités du « clustering ».
2. Expliquer le « K-means clustering ».
3. Expliquer le « hierarchical clustering ».
4. Expliquer les méthodes pour décider le nombre de « clusters »
5. Comparer le « hierarchical » contre le « K-means clustering ».

Related lessons ASM

18. Cluster Analysis

18. Cluster Analysis

Pas encore fait, notes pour me donner une idée du chapitre

Unsupervised learning method ;

Groups observations into homogeneous clusters, groups of similar observations ;

Contrast to PCA ;

Examples of application ;

1. K-means clustering
   - › Decide the number of clusters in advance ;
   - › Clusters are exhaustive and mutually exhaustive ;

     Each observation will belong to one and none belongs to more than one ;

- Clusters selected to minimise the dissimilarities between points within the clusters ;
- Centroid of clusters ;
- Algorithm ;

2. Hierarchical clustering
   - Doesn't specify the number of clusters ;
   - Results in bigger clusters containing smaller clusterings containing smaller clusterings .... ;
   - **Bottom-up** clustering ;
   - **agglomerative** clustering ;
   - Dendrogram ;
   - **Linkage** : dissimilarity between clusters
     (a) Complete linkage
     (b) Single linkage
     (c) Average linkage
     (d) Centroid linkage
3. Issues with clustering
   - Assumption of hierarchy with hierarchical clustering ;
   - Decisions that need to be made ;
   - Difficulty in validating clusters ;
   - Lack of robustness ;

Note sur les exercices :

1.

Vidéos YouTube

- StatQuest: Hierarchical Clustering
- StatQuest: K-means clustering