Kunskapskontroll

I denna kunskapskontroll kommer vi arbeta med riktig data och tillämpa regressionsmodellering. Kunskapskontrollen består av:

1. Lära dig grunderna i Excel.
2. Besvara teoretiska frågor.
3. Hantera extern data från SCB.
4. Data från blocket.
5. Regressionsmodellering på datan från blocket.
6. Rapportskrivande.
7. Självutvärdering

På Omniway ska ni lämna in:

1. En rapport som följer den vanliga mallen ”rapport_mall” (samma som användes i föregående kurs).
2. R koden.

Rent generellt är jag mycket nöjd över den utvecklingen ni gjort. Detta är utbildningens mest krävande kunskapskontroll hitintills. Jag ser fram emot att dels se er inställning under arbetet (det kommer vara krävande) dels se vad ni lyckas åstadkomma.

Lycka till.
Antonio Prgomet

1. Lära dig grunderna i Excel

Som data scientist förväntas du kunna Excel. Kolla på kapitel 1-6 samt kapitel 8 i denna videon för att lära dig grunderna i Excel:
https://www.youtube.com/watch?v=4UMLFC1SoHM&list=PLgzaMbMPEHEx2aR9-EXfD6psvezSMcHJ6&index=1&t=3s

I denna kunskapskontroll kommer datan sparas i Excel och därför behöver ni kunna grunderna.

2. Teoretiska frågor

Besvara följande teoretiska 7 frågor:

1. Kolla på följande video: https://www.youtube.com/watch?v=X9_ISJ0YpGw&t=290s , beskriv kortfattat vad en Quantile-Quantile (QQ) plot är.
2. Din kollega Karin frågar dig följande: *”Jag har hört att i Maskininlärning så är fokus på prediktioner medan man i statistisk regressionsanalys kan göra såväl prediktioner som statistisk inferens. Vad menas med det, kan du ge några exempel?”* Vad svarar du Karin?
3. Vad är skillnaden på ”konfidensintervall” och ”prediktionsintervall” för predikterade värden?
4. Den multipla linjära regressionsmodellen kan skrivas som:
   $Y = \beta_0 + \beta_1 x_1 + \beta_1 x_2 + \ldots + \beta_p x_p + \varepsilon$.
   Hur tolkas beta parametrarna?

5. Din kollega Nils frågar dig följande: *”Stämmer det att man i statistisk regressionsmodellering inte behöver använda träning, validering och test set om man nyttjar mått såsom BIC? Vad är logiken bakom detta?”* Vad svarar du Hassan?
6. Förklara algoritmen nedan för ”Best subset selection”

Algorithm 6.1 *Best subset selection*

1. Let $\mathcal{M}_0$ denote the *null model*, which contains no predictors. This model simply predicts the sample mean for each observation.
2. For $k = 1, 2, \ldots p$:
   (a) Fit all $\binom{p}{k}$ models that contain exactly k predictors.
   (b) Pick the best among these $\binom{p}{k}$ models, and call it $\mathcal{M}_k$. Here *best* is defined as having the smallest RSS, or equivalently largest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using using the prediction error on a validation set, C_p (AIC), BIC, or adjusted R^2. Or use the cross-validation method.

7. Ett citat från statistikern George Box är: *“All models are wrong, some are useful.”* Förklara vad som menas med det citatet.

3. Extern data

De som satsar på betyget godkänt ska använda sig av extern data genom en manuell process enligt nedan.

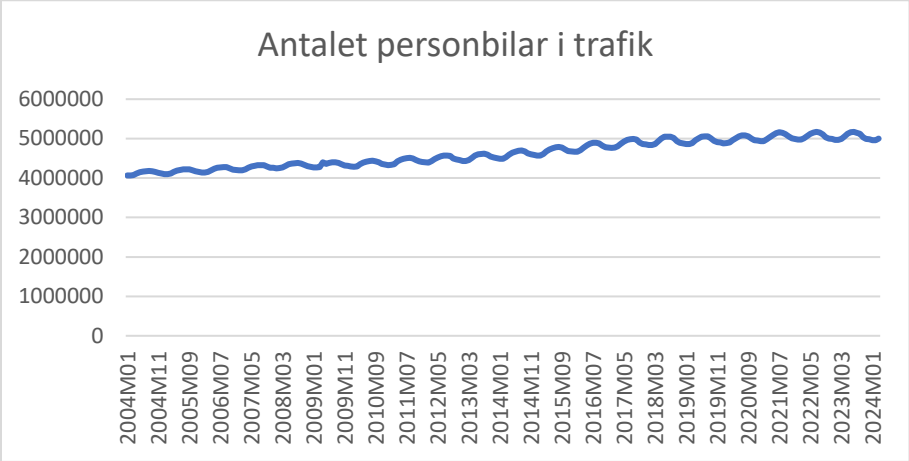
Extern data är i många fall väldigt värdefullt. Antonio använde t.ex. extern data på jobbet kopplat till räntor och liknande för diverse ekonomiska analyser. Det fina är att det är gratis tillgängligt för företag och privatpersoner. Nu kommer vi använda extern data från statistiska centralbyrån (SCB).

Använd data som du kan ”väva in i din rapport”, här ser du potentiellt väldigt intressant användning av extern data: https://www.statistikdatabasen.scb.se/pxweb/sv/ssd/

Själva datan ses här:

	2004M01	2004M02	2004M03	2004M04	2004M05	2004M06	2004M07	2004M08	2004M09	2004M10	20
Personbilar											
I trafik	4 065 919	4 065 292	4 078 304	4 114 364	4 144 593	4 161 752	4 172 697	4 177 187	4 174 889	4 159 816	4 1

Och jag valde att ladda ned den till Excel (CSV fil) och visualisera (tryck på ”verktyg → spara resultat som” för att ladda ned datan):



Genom att nyttja extern data så kan man skapa en väldigt övertygande argumentation i t.ex. ett företag eller i skolan. Exempelvis hade man kunnat skriva enligt nedan:

”Antalet personbilar i trafik ökar vilket medför att automatisk prissättning via regressionsmodellering är en värdeskapande innovation. Genom att dessutom kunna tillföra statistisk inferens så kan kunderna bättre förstå vad som påverkar prissättningen.”

Om man t.ex. hade skrivit om miljö så hade extern data kunnat användas för att exempelvis skriva följande:

”På 20 år så har antalet personbilar i Sverige ökat med 1 000 000 stycken. Detta är något som har en direkt påverkan på miljön … ”

sammanfattningsvis, extern data via t.ex. SCB kan vara väldigt användbart.

I rapporten så ska den externa datan användas i t.ex. inledning delen för att motivera varför arbetet är intressant, eller i analys delen när man kanske reflekterar kring värdet av att skapa modeller för prissättning. Exakt hur ni gör är mindre intressant eftersom jag är mest intresserad av att ni faktiskt lär er att det finns extern data och kan använda det.

De som satsar på betyget väl godkänt ska hämta den externa datan genom att använda sig av API. De som lyssnade på företagen från ledningsgruppen minns att detta var bra att kunna för att vara väl förberedd när man börjar arbeta inom IT-branschen.

Så istället för att ladda ned data manuellt så skall du nyttja API:et: https://www.scb.se/vara-tjanster/oppna-data/pxwebapi/api-for-statistikdatabasen/ .

Det finns instruktioner för hur det görs på hemsidan ovan, se rubriken *"För R utvecklare"*.

4. Data från blocket

För att kunna skapa en regressionsmodell behövs data.

- **De som satsar på betyget godkänt** har ett färdigt dataset, *"data_insamling_volvo_blocket"*, som tidigare studenter har samlat in. Notera att ni behöver undersöka datan för att dels förstå den dels hantera eventuella problem eller felaktigheter som kan finnas. Det finns här inget facit utan ni måste undersöka. Det är precis så det ofta är i verkligheten.
- **De som satsar på betyget väl godkänt** ska samla in egen data från blocket. Om ni samlar in egen data så är det en *stark rekommendation* att ni gör det i grupper. Då kan ni ha ett möte i gruppen och därefter slutföra datainsamlingen effektivt på någon timme. Ni ansvarar själva för att koordinera er. Se *"appendix – datainsamling"* längst ned i dokumentet för saker att tänka på vid datainsamling.

5. Regressionsmodellering

Skapa en regressionsmodell för datan på blocket. Målet är att du ska skapa en prediktiv modell som kan förutsäga pris på bilar. Du ska även analysera modellen (t.ex. i vilken grad de teoretiska antaganden är uppfyllda och hur det påverkar tolkningen) och tolka resultaten (exempelvis vilka variabler som är signifikanta, hur starka effekterna är, konfidensintervall, hypotesprövning, med mera).

6. Rapportskrivande

Skriv en rapport där du använder den vanliga mallen "rapport_mall".

7. Självutvärdering

I slutet av rapporten ska du inkludera en självutvärdering som innehåller:

1. Vad tycker du har varit roligast i kunskapskontrollen?`
2. Hur har du hanterat utmaningar? Vilka lärdomar tar du med dig till framtida kurser?
3. Vilket betyg anser du att du ska ha och varför?
4. Något du vill lyfta till Antonio?

Betygskriterier

Exakta betygskriterier finns i kursplanen. De som satsar på VG behöver genomgående visa på hög säkerhet i det som görs. Skriv koncist. Särskilt kommer Antonio kolla på följande:

1. Att ni samlat in extern data från SCB via API.
2. Att ni samlat in data manuellt från blocket och gjort det på ett bra sätt.
3. Att ni kan tolka era regressionsmodeller. I rapporten bör ni ha en rubrik som heter ”Undersökning av teoretiska antaganden” där ni analyserar nedanstående ”potentiella problem” från lektionsanteckningarna från YouTube videon. Syftet är att jag vill stärka er förmåga i att inte ”låsa er” utan vänja er vid att *argumentera* och *anpassa slutsatser* snarare än som bebisar be om facit från någon stackars lärare eller bara säga ”detta går inte”. Det är så seniora människor inom data science branschen arbetar eftersom det finns inget facit. Därför behöver man argumentera för sina val. Jag ser att vissa i utbildningen redan lyckats komma till denna nivån och det är ett beteende vi vill förstärka.

Notera att satsar man på VG är ”time-management” av central betydelse, precis som på arbetsplatser i verkligheten.

Potentiella Problem - Regressionsmodellering

- När vi skapar regressionsmodeller så kan det uppstå problem ifall antaganden bryts. Dessa är:

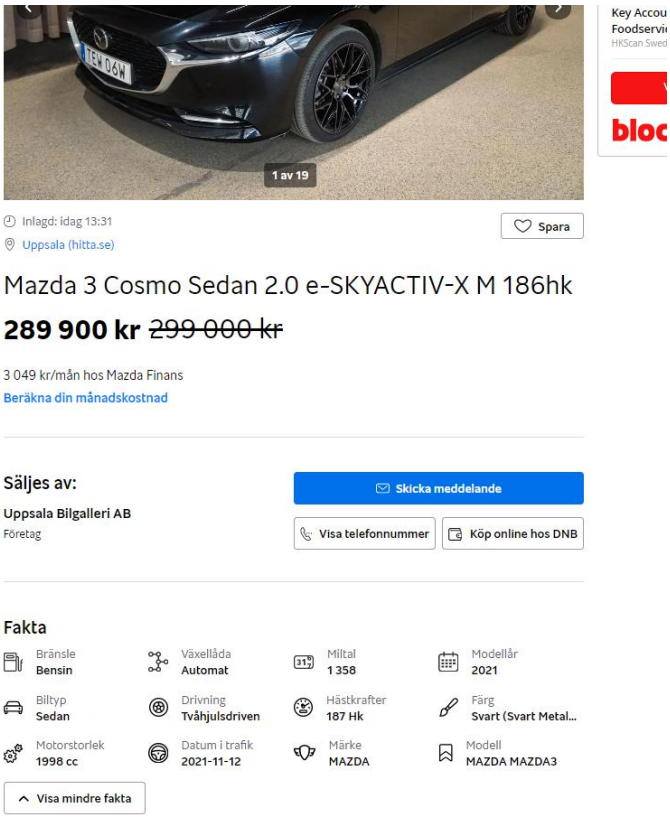
1. Icke-linjärt förhållande mellan den beroende variabeln och de oberoende variablerna.
2. Korellerade residualer - Ej oberoende residualer.
3. Icke-konstant varians på residualerna (Heteroskedasticitet).
4. Ej normalfördelade residualer.
5. Outliers.
6. "High Leverage" punkter.
7. Kollinearitet/Multikollinearitet.

- En modell är en förenkling av verkligheten och vi tror i praktiken aldrig att en modell och dess antaganden alltid är helt uppfyllda.
- Därför skall man inte bli lamslagen när antaganden inte uppfylls.
- Ytterst är vi intresserade av att ha en modell som är **användbar**.

Appendix- Datainsamling

Blocket (https://www.blocket.se/) är en sida där säljare och köpare möts för att kunna göra affärer. Ett vanligt förekommande objekt är bilar och uppgiften är att samla in data om bilar och lagra den i Excel. Excel är gratis tillgängligt för alla på skolan. *Det rekommenderas starkt att ni samarbetar i grupper för att samla in data.*

Exempel på hur en annons kan se ut ser du nedan, det framgår t.ex. vilken typ av bränsle, växellåda, miltal, modell med mera bilen har:



Det finns flertalet saker att tänka på och som ni i gruppen behöver diskutera igenom innan datainsamlingen, några exempel är:

- Ni kommer göra en modell, vad är syftet med modellen och vilken data behövs för det?
- Vilken typ av fordon vill vi modellera? Exempelvis kan det vara problematiskt om hälften är exklusiva bilar såsom Ferrari och andra hälften vanliga bilar såsom Mazda.
- Säkerställ att datan ni samlar in går att läsa in i R och att det blir som ni tänker er. Gör alltså en ”Proof of Concept” (POC). Det är trist om man efter all datainsamling inser att det inte funkar.
- Vilken typ av data skall vi samla in?
- Hur skall vi samla in datan på ett konsistent sett i gruppen?
- Kan man göra några kontroller så datan är ”rimlig”?
- **Hur mycket** data skall vi samla in?

När gruppen är klar med datainsamlingen så skall du besvara följande frågor kortfattat (ha ett kapitel i rapportens metod del som t.ex. heter ”Datainsamling”):

1. Vem du har arbetat i grupp med?
2. Hur har ni i gruppen arbetat tillsammans med datainsamlingen?
3. Några lärdommar i processen att samla in data manuellt?