

## Kunskapskontroll

Denna kunskapskontroll består av två delar.

- 1) Datainsamling av riktig data som kommer göras i **grupp (3-5 personer)**.
  - Vi delar in grupperna på lektionen. Är du inte på lektionen så kontakta en klasskamrat och gå med i en grupp. Hjälp gärna åt. Kontakta Antonio om du behöver support.
  - Gruppdiskussionerna kommer utgöra kursens muntliga examinationsdel.
- 2) Individuellt arbete där en regressionsanalys på den insamlade datan kommer göras samt utnyttjande av extern data från statistiska centralbyrån (SCB). I vanlig ordning så kommer en rapport att skrivas.

På Omniway skall en GitHub länk, som innehåller följande, lämnas in:

1. Datan som samlats in.
2. En rapport som följer den vanliga mallen "rapport\_mall" (samma som användes i föregående kurs).
3. R koden.

Det är väldigt spännande att arbeta med riktig data som inte har "facit". Nu behöver ni själva argumentera för varför det ni gör är rimligt. Jag ser fram emot att följa era arbeten och se slutresultaten.

Lycka till.  
Antonio

## Betygskriterier

Vi har kommit längre in i utbildningen vilket medför att de som satsar på VG har betydligt större utmaning än i början av utbildningen. Det är spännande, kul och lärorikt så även om det är krävande så är det bra och kul att utmana sig själv.

Nedan ser du betygskriterierna.

Examination kommer att ske genom:

1 inlämningsuppgift (IG/G/VG)

1 gruppuppgift vilken ska redovisas skriftligen och muntligen (IG/G)

### Betygsskala

Följande betygsskala tillämpas:

VG = Väl Godkänd, G = Godkänd, IG = Icke Godkänd

Läranderesultat	Inlämningsuppgift (G/VG)	Gruppuppgift (G)
1	x	
2	x	x
3	x	

### Principer för betygssättning

För betyget Godkänd ska den studerande

- Ha nått samtliga läranderesultat för kursen

För betyget Väl Godkänd ska den studerande:

- Uppnått kraven för betyget Godkänd
- I en skriftlig rapport lösa ett problem genom att implementera metoder och modeller från regressionsanalys på ett fördjupat sätt med hög säkerhet
- Redogöra för och kritiskt diskutera modellval, modellanpassning och modellutvärdering


## Del 1: Datainsamling, görs i grupp

1. Samtliga i gruppen skall kolla på följande video (3 timmar lång) för att lära sig Excel. Excel är ett verktyg som alla bör kunna då det frekvent används i arbetslivet:

<https://www.youtube.com/watch?v=4UMLFC1SoHM&list=PLgzaMbMPEHEX2aR9-EXfD6psvezSMcHJ6&index=1&t=3s>

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 gruppens uppgift är att samla in data om bilar och lagra den i Excel. Excel är gratis tillgängligt för alla på skolan, har du inte tillgång till det så kontakta din utbildningsledare.

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:



Inlagd: idag 13:31  
Uppsala ([hitta.se](#))

Spara

Mazda 3 Cosmo Sedan 2.0 e-SKYACTIV-X M 186hk

**289 900 kr** ~~299 000 kr~~

3 049 kr/mån hos Mazda Finans  
[Beräkna din månadskostnad](#)

Säljes av:  
Uppsala Bilgalleri AB  
Företag

Skicka meddelande

Visa telefonnummer Köp online hos DNB

**Fakta**

Bränsle Bensin	Växellåda Automat	Miltal 1 358	Modellår 2021
Biltyp Sedan	Drivning Tväxeldrivna	Hästkrafter 187 Hk	Färg Svart (Svart Metal...
Motorstorlek 1998 cc	Datum i trafik 2021-11-12	Märke MAZDA	Modell MAZDA MAZDA3

Visa mindre fakta

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 ni 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).
- Vilken typ av data skall ni samla in?
- Hur skall ni samla in datan på ett konsistent sätt i gruppen?
- Kan man göra några kontroller så datan är "rimlig"?
- **Hur mycket** data skall ni 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?
3. Vad var bra i grupparbetet och vad kan utvecklas?
4. Vad är dina styrkor och utvecklingsmöjligheter när du arbetar i grupp?
5. Finns det något du hade gjort annorlunda? Vad i sådana fall?

## Del 2: Regressionsmodellering, görs individuellt

I denna delen skall du (1) besvara teoretiska frågor, (2) använda extern data från statistiska centralbyrån och (3) använda regressionsmodellering för att modellera den insamlade datan. De som satsar på VG skall även göra (4) API, som är en modifikation av (2), se nedan.

### (1) Teoretiska frågor

Besvara följande teoretiska 7 frågor:

1. Kolla på följande video: [https://www.youtube.com/watch?v=X9\\_ISJ0YpGw&t=290s](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:  
$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + \varepsilon.$$
  
Hur tolkas beta parametrarna?
5. Din kollega Hassan 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  $k = 1, 2, \dots, p$ :
    - (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  $\mathcal{M}_0, \dots, \mathcal{M}_p$  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.

## (2) Extern data

Extern data är i många fall väldigt värdefullt. Nu kommer vi använda extern data från statistiska centralbyrån (SCB).

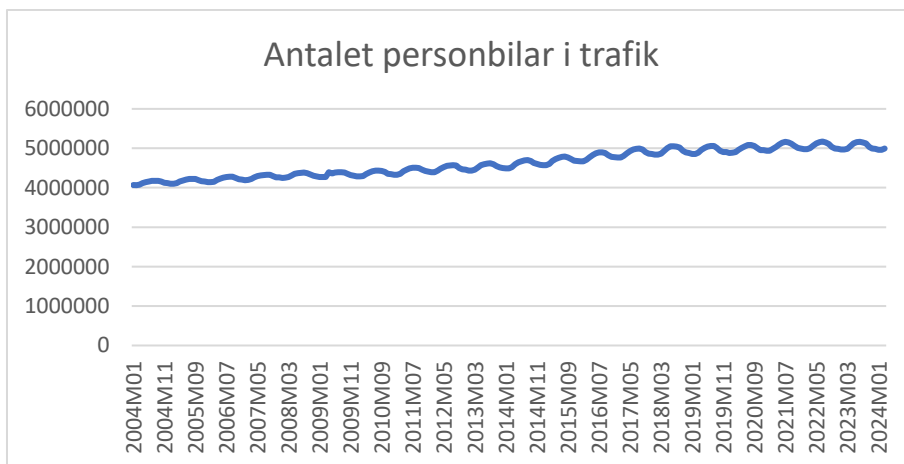
Använd data som du kan "väva in i din rapport", här ser du potentiellt väldigt intressant användning av extern data: <https://www.statistikdatabasen.scb.se/pxweb/sv/ssd/>

-	Transporter och kommunikationer
-	<b>Fordonsstatistik</b>
-	Fordonsstatistik
	Nyregistrerade personbilar efter län och kommun samt drivmedel. Månad 2006M01-2024M03 [2024-04-03]
	Fordon enligt bilregistret efter fordonsslag och bestånd. Månad 1975M01-2024M03 [2024-04-03]
	Fordon i trafik efter län och kommun samt fordonsslag. År 2002-2023 [2024-02-15]
	Personbilar i trafik efter län och kommun samt ägande. År 2002-2023 [2024-02-15]
+	Ekonomiska indikatorer

Själva datan ses här:

	2004M01	2004M02	2004M03	2004M04	2004M05	2004M06	2004M07	2004M08	2004M09	2004M10	2004M11
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	4 174 889

Och jag valde att ladda ned den till Excel (CSV fil) och visualisera:



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

*"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 kan extern data 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.

**De som satsar på VG kommer samla in extern data men istället för att göra det manuellt så kommer ett API användas, se (4) nedan.**

### (3) Regressionsmodellering

Gör en komplett regressionsmodellering på den insamlade datan. Notera, de som satsar på VG behöver göra denna delen på ett *"fördjupat sätt med hög säkerhet"* samt *"Redogöra för och kritiskt diskutera modellval, modellanpassning och modellutvärdering"*.

Rent praktiskt så innebär det t.ex. att potentiella problem såsom outliers, icke-linjaritet, icke-normalitet måste undersökas och anpassa slutsatserna så att du både tänker och skriver med precision.

### (4) VG del- API

För att uppnå VG så behöver du sammanfattningsvis ha hög kvalitet i allting som görs. Det innebär att om du t.ex. inte besvarar de teoretiska frågorna från (1) på ett bra sätt så är VG kriterierna inte uppfyllda.

I (2) så istället för att ladda ned data manuellt så skall du nyttja API:et <https://www.scb.se/vara-tjanster/oppna-data/api-for-statistikdatabasen/>. Det finns instruktioner för hur det görs på hemsidan.

## For R utvecklare

En av användarna av Statistikdatabasen har utvecklat en modul med exempelkod för R-utvecklare. Koden finns tillgänglig på GitHub.

[Hur du hämtar data från Statistikdatabasen i R \(Github\)](#)