Datorlaboration 2

Statistik och dataanalys 2 (ST1201), VT2023

Introduktion

Alla datorlaborationer ska genomföras som *Quarto-notebooks*. En stor fördel med notebook-formatet är att det låter er skapa ert egna kursmaterial genom att kombinera kod med text som beskriver vad koden gör. Att skriva sitt egna kursmaterial, alltså att med egna ord tvingas förklara hur saker fungerar, är ett av dom bästa sätten att lära sig.

Dom första tre laborationerna är utformade så att dom matchar mot dom tre delarna i inlämningsuppgiften. När du är klar med dagens laboration är du alltså redo att börja med del 2 på inlämningsuppgiften.

- Det är helt OK att ni sammarbetar under labben, men skriv din egna labbrapport! Det är viktigt att faktiskt skriva koden själv.
- Om du fastnar, testa att se om du kan hitta en lösning i dokumentationen. För att se dokumentationen för en viss funktion skriver du helt enkelt ett frågetecken följt av funktionens namn i konsolen, exempelvis `?lm`. Funkar inte det, testa google, och funkar inte det, frågga labbansvarig. Vi finns där för att svar på dina frågor, med det är viktigt att du tränar på att lösa problem själv.

Innehåll

I den här laboration kommer du lära dig att

- använda ett F-test för att jämföra två modeller,
- ställa upp en VIF-beräkning "manuellt",
- testa om antagandet om homoskedasticitet är uppfyllt med White's test.

Innan du går vidare, skapa ett tomt quarto-dokument på samma sätt som du gjorde under första labben. Alltså, skapa ett nytt quarto-dokument, radera allt utom preamble, ändra `title` till något passande, och lägg sedan till en kodchunk (med rätt chunkinställningar) som du kan ha alla dina `librar()`-anrop i i början.

Under labben kommer du jobba med datasetet `bike` som finns i paketet `sda1`.

Del 1 - F-test för jämförelse mellan modeller

- ☐ `bike` innehåller en variable `dteday` som vi inte kommer behöva i den här labben. Använd `select()` från `dplyr` för att välja ut alla kolumner *utom* `dteday` och spara dess i en ny `data.frame` eller `tibble` med namnet `datb`. (När vi väljer "alla kolumner utom xyz" så säger vi att vi väljer ut *komplementet* till xyz. Skriv `?select` i konsolen för att ta reda på du man kan välja ut komplementet till `dteday`.) Skriv över `datb` med datasetet som inte innehåller `dteday`.
- ☐ Använd `head()` (behövs ej om du använder ett `tibble`) för att undersöka vilka variabler som finns i datasetet.

I den här uppgiften ska du använda ett F-test för att undersöka om det är en bra att inkludera årstiderna i följande modell:

$$\text{nRides} = \beta_0 + \beta_1 \text{temp} + \beta_2 \text{hum} + \beta_3 \text{spring} + \beta_4 \text{summer} + \beta_5 \text{fall} + \varepsilon$$

För att göra detta ska du

- ☐ Skatta två regressionsmodeller med `lm()`. En modell skall motsvara modellen ovan (din **U**n**R**estricted model), och en ska motsvara din **R**estricted modell. Du behöver alltså själv lista ut vilken din **R**estricted model ska vara!
- ☐ Använd `reg_summary()` dina två modeller och identifiera dom delar du behöver för att göra ett F-test.
- ☐ Beräkna värdet på teststatistikan.

Slutligen ska du beräkna det kritiska värdet med R, genom att använda funktionen `qf()`. För att beräkna det kritiska värdet på signifikansnivå alpha, för en fördelning med a respektive b frihetsgrader skriver vi

```
qf(1 - alpha, df1 = a, df2 = b)
```

- ☐ Beräkna det kritiska värdet med hjälp av `qf()` för $\alpha = 0.1$. Vad blir din slutsats?

Del 2 - Multikollinearitet och variance inflation factor (VIF)

I den här deluppgiften kommer du behöva funktionen `correlate()` från paketet `corrr`. Antagligen är det inte installerat på datorn du arbetar på, så du kommer behöva börja med att installera det med `install.packages("corrr")`.

- ☐ Använd det du har lärt dig om `dplyr` till skapa en kedja av pipes som gör följande med ditt dataset `datb`: väljer ut kolumnerna `temp`, `hum`, och `windspeed`; använder funktionen `correlate()` för att ta fram en tabell som visar parvisa korrelationer (detta kallas en korrelationsmatrix). (`correlate()` behöver inga argument, den kommer ta datsetet som du skickat vidare med din pipe och skapa en korrelationsmatris.)
- ☐ Eftersom att en korrelationsmatris är *symmetrisk* (alltså triangeln ovanför diagonalen som går från topp vänster til botten höger och triangeln under diagonale är spegelbilder av varandra) så är det vanligt att bara skriva ut halva matrisen. Funktionen `shave()` från `corrr` gör just detta. Bygg ut din kedja från frågan innan med hjälp av `shave()`.
- ☐ Diagonel i matrisen ovan är satt till `NA`. Om du istället vill ha 1 på diagonalen (eftersom att korrelationen av en variabel med sig själv är 1), kan du lägga till agumentet `diagonal = 1` i `correlate()`. Testa!

När du gått igenom alla steg ovan borde du ha följande korrelationsmatris

```
Correlation computed with
* Method: 'pearson'
* Missing treated using: 'pairwise.complete.obs'

# A tibble: 3 x 4
  term        temp    hum windspeed
  <chr>      <dbl>  <dbl>     <dbl>
1 temp       1     NA            NA
2 hum        0.127  1            NA
3 windspeed -0.158 -0.248         1
```

I multipel regression vill vi helst ha förklarande variabler som är starkt korrelerade med utfallsvariablen, men som inte är starkt korrelerade med *varandra*. När vi har förklarande variabler som är korrelerade kallar vi det *multikollinearitet*. Multikollinearitet är ett problem eftersom att det gör det svårt att identifiera vilken förklarande variabel som bidrar till att förklara utfallsvariabeln.

Om vi har en stor modell med många förklarande variabler och funderar på att lägga till ytterligare en förklarande variable bör vi tänka på multikollinearitet. För att undersöka multikollinearitet kan vi beräkna VIF (variance inflation factor) för den nya förklarande variabeln.

Om VIF är väldigt hög betyder det att den nya förklarande variabeln inte bidrar med så mycket nytt. Om vi kan “beräkna” vad den nya variabeln har för värde utifrån dom variabler som redan är med i modellen så kommer den såklart inte att kunna bidra med någon ny information om utfallsvariabeln.

Antag att du har formulerat följande modell

$$\text{nRides} = \beta_0 + \beta_1\text{temp} + \beta_2\text{windspeed} + \beta_3\text{vår} + \beta_4\text{sommar} + \beta_5\text{höst} + \varepsilon$$

Du funderar på att lägga till variabeln `hum` (luftfuktighet) men oroar dig för multikollinearitet, så du börjar med att beräkna VIF för `hum`.

- ☐ Hur ser regressionsmodellen du behöver använda för att beräkna VIF för `hum` ut? Skatta den modellen och ta fram dess förklaringsgrad med `reg_summary()`.
- ☐ Beräkna VIF med hjälp av förklaringsgraden. Verkar det vara en bra idé att inkludera `hum` i din modell?

Del 3 - Test av homoskedasticitet med White's test

Ett av dom tre antagandena vi ofta gör om feltermernas fördelning är *homoskedasticitet*, alltså att feltermernas varians är konstant. Detta betyder att variansen till exempel inte får bero på dom förklarande variablerna.

Detta antagande är ofta inte uppfyllt. Tänkt dig till exempel om vi vill undersöka sambandet mellan vikt och ålder, och att dina observationer innehåller både spädbarn och fullvuxna människor. Fysikens lagar gör det *omöjligt* för ett spädbarnen att variera lika mycket i vikt som en fullvuxen människa. Det kommer alltså att finnas ett *samband* mellan vår förklarande variabel (ålder) och feltermernas varians. När det finns ett sånt samband säger vi att vi har *heteroskedastiska feltermer*.

Ett sätt att upptäcka om vi har heteroskedasticitet är att plotta residualerna mot dom olika förklarande variablerna. Om vi ser något mönster i data, exempelvis ett trattformat utseende, så är det ett tecken på att vi har problem med heteroskedasticitet.

Ett annat sätt att upptäcka heteroskedasticitet är genom att använda ett test. I den här uppgiften ska du använda Whites test för homoskedasticitet.

Whites test fungerar så att du specificerar en regressionsmodell för dom kvadrerade residualerna, med följande som förklarande variabler:

- dina förklarande variabler,
- kvadraterna av dina förklarande variabler,
- interaktioner mellan dina förklarnde variabler.

Anledningen till att vi har dom kvadrerade residualerna som utfallsvariabel är för att det finns ett samband mellan dom kvadrerade residualerna och feltermernas varians (vikitigt att förstå detta!). Om vi försöker göra en modell som kan förklara dom kvadrerade residualerna med våra förklarade variabler så betyder det att feltermernas varians antagligen har ett samband med någon av dom förklarande variablerna!

I fallet när vi har två förklarande variabler, i exemplet nedan X_1 och X_2, så ska vi alltså använda följande regression:

$$e^2 = \tilde{\beta}_0 + \tilde{\beta}_1 X_1 + \tilde{\beta}_2 X_2 + \tilde{\beta}_3 X_1^2 + \tilde{\beta}_4 X_2^2 + \tilde{\beta}_5 X_1 \cdot X_2 + \varepsilon$$

Du ska i den här uppgiften utgå ifrån en modell som endast har `temp` och `windspeed` som förklarande variabler och utföra Whites test för att se om det finns några tecken på heteroskedasticitet.

- ☐ Använd `mutate()` för att lägga till `temp2` (alltså `temp` upphöjt till 2) och `windspeed2` (`windspeed` upphöjt till 2) till ditt dataset.

- ☐ Skatta en regressionsmodell med `nRides` som beroende variabel och `temp` och `windspeed` som förklarande variabler. Spara din modell som `reg_modell`.
- ☐ Använd `mutate()` för att lägga till en ny kolumn till ditt dataset som innehåller dom kvadrerade residualerna från regressionsmodellen. Du kommer åt resdi-ualerna genom `reg_modell$residuals`, och kan lägga till genom `mutate(e2 = reg_modell$residuals^2)`.
- ☐ Skatta en regressionsmodell enligt ovan, som alltså innehåller `temp`, `temp2`, `windspeed`, `windspeed2`, och `temp:windspeed`, och spara som `reg_white`.

White's test har nollhypotesen att feltermerna är homoskedastiska, vilket betyder att

$$H_0 : \tilde{\beta}_1 = \tilde{\beta}_2 = \tilde{\beta}_3 = \tilde{\beta}_4 = \tilde{\beta}_5 = 0$$

Testvariabeln för Whites test är nR^2, där R^2 är förklaringsgraden i modellen som har e^2 som utfallsvariabel, och n är antal observationer i datasetet.

- ☐ Beräkna nR^2.

För stora stickprov är nR^2 χ^2-fördelad med k frihetsgrader, där k är antalet restriktioner under nollhypotesen. För att beräkna det kritiska värdet kan du använda `qchisq()`. Det första argumentet ska vara $1 - \alpha$, så exempelvis 0.95 för 5% signifikansnivå, och det andra argumentet ska vara antalet frihetsgrader.

- ☐ Beräkna det kritiska värdet på 1% signifikansnivå. Förkastar du nollhypotesen?
- ☐ Använd `reg_residuals()` på din orginalmodell. Verkar resultatet från Whites test rimligt? Varför/varför inte?