**HT 2020 Inlämningsuppgift i Regressionsanalys, dagtid**

|  |
| --- |
|  |

Övningslärare: Maria Anna Di Lucca

Övningsgrupp: Raum 2

Inlämningsdatum: fredag 27 november kl. 17:00

Gruppmedlemmar (3 personer per grupp):

|  |  |  |  |
| --- | --- | --- | --- |
| **För- och efternamn** | **Pnr** | **Mobilnr** | **E–mail** |
| Leo Söderberg | 980114–0717 | 0793101331 | leo.soderberg@gmail.com |
| Michaela Machacny | 950322–6947 | 0700596078 | michaela247@hotmail.com |
| Moa Dahlin | 840521–3904 | 0702703207 | moadahlin@gmail.com |
| Sara Standert | 790330–0205 | 0701088429 | sara.standert@gmail.com |

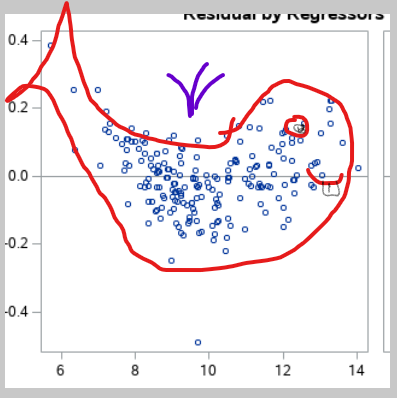
*Övningslärarens anteckningar*:

**Resultat efter första rättningen**:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| □ | Godkänt | □ | Komplettering | Lärarens signatur: ……………………… |

*OBS! Vid eventuell komplettering skall ursprungsuppgiften med försättsblad bifogas*

Kommentarer (vid komplettering):



/ Ha en trevlig läsning hälsningar residual valen Rubert

**1. Inledning**

Beroendevariabel: Voterate.

Förklaringsvariabler (version 2): Population, Education, Houses, Longitude, Area.

**2. Deskriptiv statistik**

Tabell 1: Beskrivande tabell över variabler

|  |  |  |
| --- | --- | --- |
| **Variabel** | **Beskrivning** | **Typ** |
| **Y = Voterate** | Andel röster i området | Kvantitativ, kontinuerlig |
| **X1 = Population** | Antal individer som är 18 år eller äldre (basvärdet) i området | Kvantitativ, kontinuerlig |
| **X2= Education** | Antal individer som innehar utbildning på minst high school-nivå i området | Kvantitativ, kontinuerlig |
| **X3= Houses** | Antalet egenägda house units i området | Kvantitativ, kontinuerlig |
| **X4 = Longitude** | Hur långt väster- österut området ligger | Kvantitativ, kontinuerlig |
| **X5= Area (dummy)** | Intervjuade bosatta i centrala staden (1) eller utanför (0) i området | Kategori, dikotom |

Tabell 2: Beskrivande mått över samtliga variabler

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Medelvärde** | **Standardavvikelse** | **Median** | **Minimum** | **Maximum** |
| **Y = Voterate** | 0,5540621 | 0,1275801 | 0,55146 | 0,04703 | 1,10526 |
| **X1= Population** | 9,92967 | 1,64313 | 9,6479 | 5,717 | 14,0272 |
| **X2= Education** | 9,24666 | 1,71023 | 8,85065 | 5,451 | 13,4166 |
| **X3= Houses** | 8,86886 | 1,62363 | 8,65545 | 4,5951 | 12,7145 |
| **X4= Longitude** | -0,9236974 | 0,1362628 | -0,84596 | -1,23118 | -0,71988 |
| **X5= Area (dummy)** | – | – | – | – | – |

|  |  |
| --- | --- |
| **Area** | **Antal observationer** |
| Centrala delar | 92 |
| Utanför centrala delar | 108 |

**Kommentar:**

Voterate: Vi ser att andelen Voterate är minimum 4,70 procent och maximum 111 procent. En andel kan inte vara högre än 100 procent och detta tyder därmed på något fel.

Population: Området med den minsta befolkningsmängden har 5,72 enheter och området med störst befolkningsmängd har 14 enheter. I genomsnitt bor det 9,93 enheter i områdena.

Education: I genomsnitt har områdena 9,25 enheter med en utbildningsnivå motsvarande high school eller högre. Området med den lägsta utbildningsnivån har ett genomsnitt på 5,45 enheter och området med störst utbildningsnivå har 13,42 enheter.

Houses: Området med minst antal egenägda bostäder har i genomsnitt 4,60 enheter medan området med flest har 12,72 enheter. I medeltal har områdena 8,87 enheter och medianen är 8,66

Longitude: Samtliga värden för områdena är negativa, förutom standardavvikelsen som alltid är positiv. Det området med lägst värde har -1,23 och det område med högst värde har -0,72. Medelvärdet är -0,92 och medianen är -0,85.

Area: Då denna variabel är en dummyvariabel går det inte att räkna på medelvärde, median, standardavvikelse et cetera. Det som går att beskriva med dummyvariabeln är att om den innehar värdet 0 bor de intervjuade utanför den centrala staden och om den innehar värdet 1 bor de intervjuade i den centrala staden. 92 centrala, 108 utanför.

Allmänt: Vi ser att medelvärdet och medianen är snarlika för alla förklaringsvariabler, vilket tyder på en normalfördelad distribution av dem.

Tabell 3: Pearsons korrelationsmatris

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Pearson Correlation Coefficients, N = 200**  **Prob > |r| under H0: Rho=0** | | | | | | |
|  | **Voterate** | **Population** | **Education** | **Houses** | **Longitude** | **Area** |
| **Voterate** | 1.00000 | -0.45070  <.0001 | -0.35623  <.0001 | -0.40039  <.0001 | -0.42456  <.0001 | -0.29138  <.0001 |
| **Population** | -0.45070  <.0001 | 1.00000 | 0.99007  <.0001 | 0.99092  <.0001 | 0.25928  0.0002 | 0.82983  <.0001 |
| **Education** | -0.35623  <.0001 | 0.99007  <.0001 | 1.00000 | 0.98314  <.0001 | 0.18058  0.0105 | 0.83987  <.0001 |
| **Houses** | -0.40039  <.0001 | 0.99092  <.0001 | 0.98314  <.0001 | 1.00000 | 0.26336  0.0002 | 0.82993  <.0001 |
| **Longitude** | -0.42456  <.0001 | 0.25928  0.0002 | 0.18058  0.0105 | 0.26336  0.0002 | 1.00000 | 0.15329  0.0302 |
| **Area** | -0.29138  <.0001 | 0.82983  <.0001 | 0.83987  <.0001 | 0.82993  <.0001 | 0.15329  0.0302 | 1.00000 |

Grön = starka samband (± 0,80 – 1,00), Röd = svaga samband (± 0,00 – 0,20)

Det finns ett starkt positivt samband mellan Population och Education, Population och Houses, Population och Area, Education och Area, Houses och Education, Houses och Area.

Det finns ett svagt positivt samband mellan Education och Longitude samt Longitude och Area.

Samtliga korrelationer för Voterate är negativa (-0,29 till -0,46).

|  |  |
| --- | --- |
| Figur 1a | Figur 1b |
| Figur 1c | Figur 1d |
| Figur 1e |  |

Figur 1a) Om vi skulle anpassa en linjär regressionslinje till scatterplotten skulle det behövas två linjer för att passa punkterna. En linje med negativt lutning mellan värdena 6 och 10 samt en linje mellan värdena 10 och 14 där lutningen är relativt plan. Detta antyder att de troligtvis bör vara en icke-linjär linje.

Figur 1b) Om vi skulle anpassa en linjär regressionslinje till scatterplotten skulle det behövas två linjer för att passa punkterna. En linje med negativt lutning mellan värdena 6 och 10 samt en linje mellan värdena 8 och 14 där lutningen är relativt plan. Detta antyder att de troligtvis bör vara en icke-linjär linje.

Figur 1c) Om vi skulle anpassa en linjär regressionslinje till scatterplotten skulle det behövas två linjer för att passa punkterna. En linje med negativt lutning mellan värdena 4 och 10 samt en linje mellan värdena 10 och 13 där lutningen är relativt plan. Detta antyder att det troligtvis bör vara en icke-linjär linje.

Figur 1d) Scatterplotten visar att punkterna ligger i kluster. Detta antyder att det troligtvis föreligger en annan variabels påverkan på relationen.

Figur 1e) Då det är en kvalitativ variabel fungerar det inte med scatterplot. Istället presenterar vi resultatet i en boxplot. Vi kan i den utläsa att distributionen är större utanför den centrala staden jämfört med den centrala staden. Vi kan även se att vi har två punkter i boxplotten över “outside city” som ligger mellan 1,5 och 3 gånger kvartilavståndet från kvartil ett och tre (Nyqvist, 2017). Dessa är markerade med cirkel.

|  |  |
| --- | --- |
| Figur 2.1 a | Figur 2.1 b |
| Figur 2.2a | Figur 2.2b |
| Figur 2.3a | Figur 2.3b |
| Figur 2.4a | Figur 2.4b |
| Figur 2.5a | Figur 2.5b |
| Figur 2.6a | Figur 2.6b |

För samtliga plottar (figurer 2.1a – 2.6b) går det att se att det inte spelar någon större roll vilken som är den beroende respektive oberoende variabeln. Då sambandet i plotten blir “detsamma” men spegelvänt och vridet.

Det kan utläsas i plottarna för Population och Education (figur 2.1 a och b) att det finns ett starkt positivt samband då punkterna ligger relativt nära varandra och formar en linje uppåt. Variansen tycks inte vara konstant eftersom punkterna ligger glesare för de låga värdena jämfört med de högre.

I plottarna för Houses och Population (figur 2.2 a och b) går det att utläsa att det finns ett starkt positivt samband då punkterna ligger relativt nära varandra och formar en linje uppåt, med en avvikande punkt. Variansen är relativt jämn och punkterna ligger centrerade längs en linje.

Det kan utläsas i plottarna för Population och Longitude (figur 2.3 a och b) att det troligen finns andra variabler som påverkar då värdena återfinns i kluster i varierande storlek, varav de två mittersta är störst. Det verkar även finnas en stor varians då klustren ligger i linjer och inte är formade i cirklar.

Plottarna för Education och Houses (figur 2.4 a och b) visar att det finns ett starkt positivt samband. Variansen tycks inte vara konstant då den avviker mer för de låga värdena jämfört med de höga. Det finns även en punkt som avviker. De här plotterna är relativt lika de för Houses och Population.

Plottarna för Education och Longitude (figur 2.5 a och b) samt Houses och Longitude (figur 2.6 a och b) visar att det troligen finns andra variabler som påverkar dessa då punkterna ligger i kluster. Plottarna är nästintill identiska med de för Population och Longitude.

**Boxplotts med Area och övriga oberoende variabler:**

|  |  |
| --- | --- |
| Figur 3a | Figur 3b |
| Figur 3c | Figur 3d |

Boxplotten för Education (figur 3a) visar att det är ett mindre antal som har högre utbildning i “outside city” jämfört med “central city”. Boxplotten för “outside” city har även ett värde som ligger utanför 1,5 gånger kvartilavståndet från undre kvartilen. Det går att se att den övre kvartilen har en större spridning jämfört med den undre för “central city”.

Boxplotten för Population (figur 3b) visar att det är färre invånare “outside city” än i “central city”. Boxplotten för “outside city” har även två värden som ligger utanför 1,5 gånger kvartilavståndet från den undre kvartilen (Nyqvist, 2017). Det går att se att den övre kvartilen har en större spridning än den undre för “central city”.

Boxplotten för Longitude (figur 3c) visar att det är högre spridning för de lägre värdena i både “outside city” och “inside city”. Det går att se att den nedre kvartilen har en större spridning jämfört med den övre för “outside city”.

Boxplotten för Houses (figur 3d) visar att antalet “house units” är betydligt färre “outside city” än i “central city”. Denna plott är väldigt lik plotten över Population. Boxplotten för “outside city” har tre värden som ligger utanför 1,5 gånger kvartilavståndet från undre kvartilen.

Medelvärdet och medianen är relativt lika för alla förutom Longitude, för denna går det att se att medianen och medelvärdet skiljer sig markant.

**3. Val av modell**

**ENKEL REGRESSION**

Tabell 4 a: Samtliga möjliga regressionsmodeller med en oberoende variabel

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **R-Square** | **Adj. R-Sq.** | **F Value** | **Pr > F** |
| **Voterate= a+bX1** | 0,2031 | 0,1991 | 50,47 | <0,0001 |
| **Voterate=a+bX2** | 0,1269 | 0,1225 | 28,78 | <0,0001 |
| **Voterate= a+bX3** | 0,1603 | 0,1561 | 37,80 | <0,0001 |
| **Voterate=a+bX4** | 0,1803 | 0,1761 | 43,43 | <0,0001 |
| **Voterate= a+bX5** | 0,0849 | 0,0803 | 18,37 | <0,0001 |

Tabell 4 b: Samtliga möjliga regressionsmodeller med en oberoende variabel

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Parameter Estimates** | | | | | | | | | |
| **Model** | **Variable** | **Label** | **DF** | **Parameter**  **Estimate** | **Standard**  **Error** | **t- Value** | **Pr > |t|** | **95 % Confidence Limits** | |
| **1** | **Intercept** | **Voterate** | 1 | 0.90154 | 0.04957 | 18.19 | <.0001 | 0.80379 | 0.99930 |
|  | **X1** | **Population** | 1 | -0.03499 | 0.00493 | -7.1 | <.0001 | -0.04471 | -0.02528 |
| **2** | **Intercept** | **Voterate** | 1 | 0.79978 | 0.04658 | 17.17 | <.0001 | 0.70793 | 0.89164 |
|  | **X2** | **Education** | 1 | -0.02657 | 0.00495 | -5.36 | <.0001 | -0.03634 | -0.01681 |
| **3** | **Intercept** | **Voterate** | 1 | 0.83309 | 0.04613 | 18.06 | <.0001 | 0.74211 | 0.92406 |
|  | **X3** | **Houses** | 1 | -0.03146 | 0.00512 | -6.15 | <.0001 | -0.04155 | -0.02137 |
| **4** | **Intercept** | **Voterate** | 1 | 0.18688 | 0.05625 | 3.32 | 0.0011 | 0.07596 | 0.29780 |
|  | **X4** | **Longitude** | 1 | -0.39751 | 0.06024 | -6.60 | <.0001 | -0.51631 | -0.27871 |
| **5** | **Intercept** | **Voterate** | 1 | 0.58829 | 0.01177 | 49.97 | <.0001 | 0.56507 | 0.61150 |
|  | **X5** | **Area** | 1 | -0.07440 | 0.01736 | -4.29 | <.0001 | -0.10863 | -0.04017 |

Utifrån tabellerna är Population den bästa förklaringsvariabeln då den har högst R²-värde och korrelationsvärde.

Modell: Voterate = 0,90154 – 0,03499X

Korrelationsmatrisvärde Population/Voterate: – 0,45070, p-värde: <0,0001

R²: 0,2031

**Modellkontroll**

Vi kan utläsa att både korrelationskoefficienten och koefficienten är negativa, vilket antyder att koefficienten förmodligen är rätt modellerad. Korrelationen visar på vilken riktning sambandet har, och koefficienten visar på vilken riktning variabeln har. Dessa skall ha samma tecken.

Modell: Voterate = 0,90154 – 0,03499X

Interceptet tolkas ej. Men koefficienten går att tolka att för varje extra enhet Population kommer Voterate att minska med 0,03499

Kritiskt F-värde från tabell: 3,84 (v1=1 v2=198 α=0,05)

Observerat F-värde: 50,47, p-värde: <0,0001

Då observerat värde är större än kritiska samt att p-värdet är extremt lågt, går det att utläsa att modellen i sin helhet är signifikant.

Kritiskt t-värde från tabell: ±1,972 (v=199 α=0,05)

Observerat t-värde: –7,1, p-värde: <0,0001.

Då observerat värde är större än kritiska samt att p-värdet är extremt lågt, går det att utläsa att variabeln är signifikant.

Vi gör följande antaganden: feltermerna bildar en normalfördelning med samma varians och är oberoende av varandra (Andersson, Jorner & Ågren, 2007). Detta analyseras i bilderna nedan.

|  |  |
| --- | --- |
| (Figur 4a)  Det går att utläsa att prickarna inte ligger med samma avstånd från nollinjen samt att de bildar ett mönster och inte ligger “random”. Därmed antas de ha olika varians.  De tenderar att vara positiva värden för punkterna för låga och höga värden, medan de i mitten tenderar att ha negativa värden. | (Figur 4b)  Plotten används för att analysera misstankar om outliers ifall de ligger utanför intervallet -2 till 2. Då det finns punkter utanför intervallet -2 till 2, antyder det att det förmodligen förekommer outliers bland de observerade värdena. |
| (Figur 4c)  Punkterna ligger längs med linjen, vilket antyder att det till stor del är en normalfördelning. Dock finns det några punkter som har ett avstånd från linjen och därmed antas det observerade värdena inte vara helt normalfördelade. | (Figur 4d)  Det går att utläsa att staplarna till vänster är mer som en jämn trappa jämfört med de på den högra sidan. vilket antyder att kurvan till stor del är normalfördelad men aningen skev åt vänster. |

|  |
| --- |
| (Figur 4e)   Det går att utläsa ett visst mönster bland punkterna vilket antyder att modellen förmodligen inte ska vara linjär. |

**Vald modell med en oberoende variabel – utan outliers**  
Vi tog vår valda “bästa modell” och eliminerade outliers och fick då följande modell:  
Antal observerad data efter eliminering: 195

Modell: Voterate = 0,83808 – 0,02872X

Interceptet tolkas ej. Men tolkningen av koefficienten blir att för varje extra enhet Population kommer Voterate att minska med 0,02872

**Modellkontroll**  
Kritiskt F-värde från tabell: 3,84 (v1=1 v2=198 α=0,05)

Observerat F-värde: 41,62, p-värde: <0,0001

Då observerat värde är större än kritiska samt att p-värdet är extremt lågt, går det att utläsa att modellen i sin helhet är signifikant. Kritiskt t-värde från tabell: ±1,972 (v=199 α=0,05)

Observerat t-värde: –6,45, p-värde: <0,0001.

Då observerat värde är större än kritiska samt att p-värdet är extremt lågt, går det att utläsa att variabeln är signifikant.

R² och Adjusted R² är 0.1774 respektive 0.1731. Detta är lägre värden än modellen med outliers. Men då residualplottarna nedan visar bättre resultat, antas denna modell vara bättre.

|  |  |
| --- | --- |
| (Figur 5a)  Punkterna ser mer slumpmässiga ut än i modellen med outliers. Därmed antas de ha en mer konstant varians. | (Figur 5b)  Punkterna utanför gränsvärdena 2 och -2 befinner sig närmare gränserna eftersom outliers har eliminerats och antas därför inte vara av betydelse. |
| (Figur 5c)  Punkterna ligger längs med linjen, vilket antyder att det till stor del är en normalfördelning. Dock finns det några punkter som har ett avstånd från linjen och övriga punkter. Därmed antas det observerade värdena inte vara helt normalfördelade. | (Figur 5d)  Precis som i figuren med outliers går det att utläsa att staplarna till höger är mer ojämna än de till vänster. Det antyder att kurvan till stor del är normalfördelad men aningen skev åt vänster. |

**MULTIPEL REGRESSION**

Tabell 5a: Samtliga möjliga regressionsmodeller med två oberoende variabler

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **R-Square** | **Adj. R-Sq.** | **F Value** | **Pr > F** |
| **Voterate= a+b1X1+b2X2** | 0.6130 | 0.6091 | 156.02 | <.0001 |
| **Voterate= a+b1X1+b2X3** | 0.3213 | 0.3144 | 46.62 | <.0001 |
| **Voterate= a+b1X1+b2X4** | 0.3046 | 0.2976 | 43.15 | <.0001 |
| **Voterate= a+b1X1+b2X5** | 0.2251 | 0.2172 | 28.61 | <.0001 |
| **Voterate= a+b2X2+b3X3** | 0.2022 | 0.1941 | 24.96 | <.0001 |
| **Voterate= a+b2X2+b4X4** | 0.2610 | 0.2535 | 34.80 | <.0001 |
| **Voterate= a+b2X2+b5X5** | 0.1271 | 0.1182 | 14.34 | <.0001 |
| **Voterate= a+b3X3+b4X4** | 0.2697 | 0.2623 | 36.38 | <.0001 |
| **Voterate= a+b3X3+b5X5** | 0.1657 | 0.1572 | 19.56 | <.0001 |
| **Voterate= a+b4X4+b5X5** | 0.2327 | 0.2249 | 29.87 | <.0001 |

Tabell 5 b: Samtliga möjliga regressionsmodeller med två oberoende variabler ·

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Parameter Estimates** | | | | | | | | | |
| **Model** | **Variable** | **Label** | **DF** | **Parameter**  **Estimate** | **Standard**  **Error** | | **t- Value** | **Pr > |t|** | **Variance**  **Inflation** |
| **1** | **Intercept** | Voterate | 1 | 1.23651 | 0.04168 | 29.67 | | <.0001 | 0 |
|  | **X1** | Population | 1 | -0.38511 | 0.02448 | -15.73 | | <.0001 | 50.60846 |
|  | **X2** | Education | 1 | 0.33975 | 0.02352 | 14.44 | | <.0001 | 50.60846 |
| **2** | **Intercept** | Voterate | 1 | 1.07305 | 0.05442 | 19.72 | | <.0001 | 0 |
|  | **X1** | Population | 1 | -0.23166 | 0.03389 | -6.83 | | <.0001 | 55.30723 |
|  | **X3** | Houses | 1 | 0.20085 | 0.03430 | 5.86 | | <.0001 | 55.30723 |
| **3** | **Intercept** | Voterate | 1 | 0.55030 | 0.08028 | 6.85 | | <.0001 | 0 |
|  | **X1** | Population | 1 | -0.02835 | 0.00478 | -5.94 | | <.0001 | 1.07207 |
|  | **X4** | Longitude | 1 | -0.30886 | 0.05760 | -5.36 | | <.0001 | 1.07207 |
| **4** | **Intercept** | Voterate | 1 | 1.04014 | 0.07648 | 13.60 | | <.0001 | 0 |
|  | **X1** | Population | 1 | -0.05209 | 0.00873 | -5.97 | | <.0001 | 3.21145 |
|  | **X5** | Area | 1 | 0.06775 | 0.02870 | 2.36 | | 0.0192 | 3.21145 |
| **5** | **Intercept** | Voterate | 1 | 0.82789 | 0.04511 | 18.35 | | <.0001 | 0 |
|  | **X2** | Education | 1 | 0.08346 | 0.02596 | 3.21 | | 0.0015 | 29.90812 |
|  | **X3** | Houses | 1 | -0.11789 | 0.02735 | -4.31 | | <.0001 | 29.90812 |
| **6** | **Intercept** | Voterate | 1 | 0.43135 | 0.07511 | 5.74 | | <.0001 | 0 |
|  | **X2** | Education | 1 | -0.02156 | 0.00465 | -4.64 | | <.0001 | 1.03371 |
|  | **X4** | Longitude | 1 | -0.34865 | 0.05830 | -5.98 | | <.0001 | 1.03371 |
| **7** | **Intercept** | Voterate | 1 | 0.81202 | 0.07341 | 11.06 | | <.0001 | 0 |
|  | **X2** | Education | 1 | -0.02823 | 0.00915 | -3.09 | | 0.0023 | 3.39427 |
|  | **X5** | Area | 1 | 0.00677 | 0.03131 | 0.22 | | 0.8291 | 3.39427 |
| **8** | **Intercept** | Voterate | 1 | 0.47360 | 0.07898 | 6.00 | | <.0001 | 0 |
|  | **X3** | Houses | 1 | -0.02437 | 0.00496 | -4.91 | | <.0001 | 1.07453 |
|  | **X4** | Longitude | 1 | -0.32105 | 0.05909 | -5.43 | | <.0001 | 1.07453 |
| **9** | **Intercept** | Voterate | 1 | 0.89368 | 0.07083 | 12.62 | | <.0001 | 0 |
|  | **X3** | Houses | 1 | -0.04003 | 0.00917 | -4.37 | | <.0001 | 3.21317 |
|  | **X5** | Area | 1 | 0.03357 | 0.02979 | 1.13 | | 0.2611 | 3.21317 |
| **10** | **Intercept** | Voterate | 1 | 0.24483 | 0.05679 | 4.31 | | <.0001 | 0 |
|  | **X4** | Longitude | 1 | -0.36425 | 0.05913 | -6.16 | | <.0001 | 1.02406 |
|  | **X5** | Area | 1 | -0.05917 | 0.01613 | -3.67 | | 0.0003 | 1.02406 |

Vald modell: 3. Voterate = 0,55030 – 0,02835X1 – 0,30886X4

Motivering: Modell 1 har högst R²-värden men då den har ett VIF-värde på 50,61 föreligger förmodligen en besvärande multikollinearitet (Andersson, Jorner & Ågren, 2007). Detta kan bidra till osäkra parameterskattningar och därav valdes denna modell bort. Näst högsta R²-värdet är modell 2 men även den har ett högt VIF-värde (55,31), och valdes därmed också bort.

Den modell som hade högst R²-värde och ett godkänt VIF på mindre än 10 var modell 3, varav den valdes.

**Modellkontroll**

F-värdet och t-värdena för modellen visar på att både modellen i sin helhet och koefficienterna är signifikanta. Då kritiska t-värdet är ±1,972 (v=198 α=0,05) och det kritiska F-värdet (v1 =2, v2=197 α=0,05) är 3,00

R²-värdet på 0.3046 tolkas som att modellen kan förklara cirka 30 procent av den beroende variabeln Voterate. Även Adjusted R² ligger runt 30 procent vilket tyder på en hög trovärdighet då resultatet inte förändras vid hänsyn taget till antalet variabler, det vill säga det tyder på att inga onödiga variabler lagts till.

Tolkning av modellens koefficienter: Interceptet (0,5503) tolkas inte (Andersson, Jorner & Ågren, 2007) då en tolkning av interceptet hade inneburit att om de två oberoende variablerna var noll, hade Voterate varit 0,5503. Detta är inte logiskt då noll population och noll longitude inte borde generera någon Voterate alls.

Om koefficienten Population ökar en enhet kommer Voterate att minska med 0,02835 givet att Longitude är konstant.

Om koefficienten Longitude ökar en enhet kommer Voterate att minska med 0,30886 givet att Population är konstant.

Vi kan även se i korrelationsmatrisen att både Population och Longitude har negativa värden (–0,4507 respektive –0,42456), detta antyder att det finns ett negativt samband med de oberoende variablerna och den beroende, vilket konfirmeras med koefficienterna. Longitude och Population har värdet 0,25928. Detta antyder att det inte föreligger något starkt samband mellan de oberoende koefficienterna.

Tabell 6: Utdrag ur tabell 5 för vald modell

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Modell** | **R-Square** | **Adj. R-Sq.** | **F Value** | **Pr > F** |
| **Voterate= a+b1X1+b2X4** | 0.3046 | 0.2976 | 43.15 | <.0001 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **Label** | **DF** | **Parameter**  **Estimate** | **Standard Error** | **t- Value** | **Pr > |t|** | **Variance**  **Inflation** |
| **Intercept** | Voterate | 1 | 0.55030 | 0.08028 | 6.85 | <.0001 | 0 |
| **X1** | Population | 1 | -0.02835 | 0.00478 | -5.94 | <.0001 | 1.07207 |
| **X4** | Longitude | 1 | -0.30886 | 0.05760 | -5.36 | <.0001 | 1.07207 |

|  |  |
| --- | --- |
| (Figur 6a)  Det går att utläsa att prickarna inte ligger med samma avstånd från nollinjen, samt att de bildar ett mönster och inte ligger “random”. Det tyder på att de inte har samma varians. Det går att se att för låga och höga förväntade värden är residuerna positiva och för de förväntade värdena i mitten är majoriteten negativa. Ett par punkter sticker iväg mer från de övriga. | (Figur 6b)  Plotten används för att identifiera eventuella outliers, som ligger utanför intervallet -2 till 2. Det går att utläsa i plotten att det finns punkter utanför intervallet -2 till 2, vilket ger misstankar om att det kan finnas outliers i den observerade datan som kan behövas korrigeras. |
| (Figur 6c)  Punkterna ligger längs med linjen, vilket antyder att det till stor del föreligger normalfördelning. Det finns två observerade värden som avviker från linjen och därmed är de observerade värdena inte helt normalfördelade. | (Figur 6d)  Det går att utläsa att staplarna till höger är mer som en jämn trappa jämfört med den vänstra sidan, vilket antyder att kurvan till stor del är normalfördelad men med svans till vänster. |

|  |
| --- |
| (Figur 6e och 6f)  I den vänstra plotten (figur 6e) syns ett mönster vilket antyder att modellen inte är optimal. Det har även tidigare analys vittnat om.  I den högra plotten (figur 6f) har residualerna bildat ett mönster av kluster, vilket även var fallet i scatterplotterna med Longitude. Detta beror förmodligen på att en annan variabel påverkar (troligtvis latitude då även det är en geografisk variabel). Då klustrena inte är av samma distribution antas även att variansen inte är detsamma |

**Vald modell med två oberoende variabler – utan outliers**

Vi sökte efter jackknifes och genomförde en rensning av outliers. Efter borttagningen återstod 190 observationer det vill säga tio, som utgör fem procent, antags vara outliers som eliminerades.

Därmed skapades modellen: Voterate = 0,50805 – 0,02627X1 – 0.32750X4.

**Modellkontroll**

Kritiska F-värdet är: 3,00 (v1 =2, v2=187 α=0,05)   
Kritiska t-värdet är: ±1,972 (v=190 α=0,05)  
F-värdet är: 54,60 och därmed är modellen i helhet signifikant.  
Koefficienternas t-värden är: –6,38 och –6,76 och därmed är båda koefficienterna signifikanta.   
VIF för båda koefficienterna är: 1,04572 och därmed antas ingen besvärande multikollinearitet finnas.  
  
Modellens koefficienter har inte ändrats extremt jämfört med innan, Tolkningen av Population blir nu istället att en extra enhet minskar Voterate med 0,02627 (tidigare 0,02835) givet att Longitude är konstant och Voterate minskar med 0,32750 (tidigare 0,30886) då en enhet extra Longitude adderas, givet att Population är konstant.

Denna nya modell har ett R² värde på 0,3687 och R²-Adjusted på 0,3619. Jämfört med modellen inklusive outliers (tidigare 0,3046 respektive 0,2976) har förklaringsgraden ökat och därmed antas det att modellen förklarar bättre, vilket antas rimligt då outliers kan bidra till en felanpassad regressionslinje, då man behövt ta hänsyn till outliers.

|  |  |
| --- | --- |
| (Figur 7a) Inget synbart mönster. Vi kan därför anta att residualernas varianser är likafördelade. | (Figur 7b)  Det går att utläsa från plotten att det finns några värden utanför 2 och -2, men då dessa ligger nära gränsen förväntas de inte påverka resultatet särskilt. |
| (Figur 7c) Det går att se att punkterna följer linjen relativt jämnt och därav antas residualerna vara normalfördelade. Dock syns det mindre svansar i ändarna, men dessa är acceptabla. | (Figur 7d) Det går nu att utläsa att histogrammet ser relativt normalfördelad ut. Några staplar går upp och ner, men detta tros bero på att vårt urval är cirka 200. Om urvalsstorleken ökar torde dessa jämnas ut. |

|  |
| --- |
| (Figur 7e och 7f)  Population: Inget synbart mönster. Vi kan därför anta att residualernas varianser är likafördelade.  Longitude: Det går att se att residualerna bildar kluster, därav antas en annan variabel påverka. Dessutom går det att se att klustrena inte ser lika ut, och det går att anta att variansen inte är likafördelad. |

**Vald modell med två oberoende variabler – utan intercept**

Vi skapade en modell med samma oberoende variabler som valdes ut som den “bästa modellen” med två oberoende variabler, men tog bort interceptet ur regressionsmodellen.

Därmed skapades modellen: Voterate = – 0,00339X1 – 0,63095X4.

**Modellkontroll**  
F-värdet är: 2194,93 och därmed är modellen i sin helhet signifikant.  
Koefficientens t-värden är: –0,99 och –17,07 och därmed är enbart den sista koefficienten (Longitude) signifikant.   
  
VIF för båda koefficienterna är 16,91519 och därmed antas att det finns multikollinearitet. Detta kan bidra till att koefficientens värde blir missvisande, vilket är troligt i detta fall då koefficientens värde har förändrats extremt.  
  
Då det finns misstanke om att koefficienterna är påverkade av multikollinearitet, väljer vi att inte tolka dessa.   
  
Denna nya modell har ett R² värde på 0,9568 och R²-Adjusted på 0,9564. Jämfört med modellen inklusive outliers och intercept (tidigare 0,3046 respektive 0,2976) har förklaringsgraden ökat och regressionslinjen förklarar nu nästintill hela modellen. Men då vi misstänker att det blivit något missvisande på grund av det höga VIF-värdet, samt då en av de oberoende variablerna inte längre är signifikant, antar vi, trots de högre R²-värdena, att modellen i sin helhet inte förklarar den beroende variabeln bättre.

|  |  |
| --- | --- |
| (Figur 8a)  Tydliga “klustermönster” med två kluster i mitten och mindre i sidorna. En skillnad från figur 6a där ett särskilt mönster inte går att utläsa. Två punkter avviker fortfarande, precis som i figur 6a, från de övriga. | (Figur 8b)  I plotten går att utläsa ett större antal outliers än i figur 6b. Deras placering är längre ifrån gränsen, -2 och 2, än de i den valda figuren.  I plotten går att se att det finns fler punkter utanför intervallet -2 till 2. Samt att punkterna nu är samlade i kluster och inte lika utspridda. |
| (Figur 8c)  Det går att se att punkterna följer linjen relativt jämnt. Dock syns det mindre svansar i ändarna. Jämfört med den tidigare modellen, figur 6c, finns det fler punkter som avviker från linjen, och därav anses inte residualerna vara fullt normalfördelade. | (Figur 8d)  Det går att utläsa att histogrammet ser ungefär ut som i figur 6d, med undantaget att för en del staplar har höjden minskat. Det går även att utläsa att fördelningen över x-axeln har förändrats. Histogrammet anses inte vara fullt normalfördelad. |

|  |
| --- |
| (Figur 8e och 8f)  Population: Liksom i 6e finns inget synbart mönster men punkterna är i detta fall snedfördelade åt höger för de negativa residualerna. Vi kan därför anta att residualernas varianser inte är likafördelade i samma grad som i 6e.  Longitude: Det går att se att residualerna bildar kluster, därav antas en annan variabel påverka. Dessutom går det att se att klustrena inte ser likadana ut, och det går därmed inte att anta att variansen är likafördelad. |

**Modell med fem oberoende variabler – utan outliers**

Vi testade att konstruera en modell med alla givna oberoende variabler, utan outliers, då vi ville testa samtliga variabler som vi samlat in data på. Outliers togs bort då vi vet att dessa kan påverka resultatet på så vis att regressionslinjen blir skev.

**Modellkontroll**

Longitude har t-värde –0,04 och Area har t-värde –1,45 och är därmed är ingen av dem signifikanta. De övriga variablerna är signifikanta.

Population, Education och Houses har alla höga VIF-värden vilket tyder på besvärande multikollinearitet.

Modellens F-värde är dock 108,51 vilket betyder att den som helhet är signifikant.  
R² samt Adjusted R²-värdena är 0,7488 respektive 0,7419, vilket antyder att modellen är bra passad till datan, men detta kan ha blivit påverkat av de höga värdena på VIF.

I figur 9a och 9b syns en avvikande punkt, vilket tros vara dummyvariabeln Area. I övrigt syns inget tydligt mönster i figurerna och få värden finns utanför –2 och 2. I figur 9c och 9d går det läsa att den observerade datan är normalfördelad. Trots att residualfigurerna visar bra värden, visar VIF-värdena och de inte signifikanta variablerna att modellen förmodligen inte är bra.   
  
Då modellen tydligt påvisade att flera variabler hade höga VIF-värden, och därmed multikollinearitet, valdes det inte att undersöka om modellen blivit bättre utan intercept.

|  |  |
| --- | --- |
| (Figur 9a) | (Figur 9b) |
| (Figur 9c) | (Figur 9d) |

I och med ovanstående analys påvisade att variablerna Population- Education- Houses troligen medgav multikollinearitet i modellen var vi intresserade att se variabeln Area borde adderas till modellen. Vi testade först modellen med intercept och fann att alla koefficienter förutom Arean, med p-värdet 0,1789, var signifikanta.  
  
Därför testades modellen även utan intercept då vi vet att det kan påverka koefficienterna. I och med det blev Arean signifikant, med inte Population samt fick både Population och Longitud fick höga VIF-värden. Då det i modellen med intercept endast var Arean som inte var signifikant, antas det att Arean förmodligen inte är en relevant variabel.

F-värdet för modellen med intercept är 39,87, vilket betyder att modellen i sin helhet är signifikant. Både R²-värdet och Adjusted R²-värde (0,3940 respektive 0,3841) är bättre än i modellen utan Area vilket antyder att variabeln har relevans för modellen, men då den inte är signifikant kan det behövas fler observationer för att säkerställa variabelns relevans.

Residualerna bildar ett “v-mönster” (se figur a och b) och därför antas variansen inte vara konstant. I figurerna c och d går det att se att observationerna är relativt normalfördelade, men att histogrammet inte visar ett fint “trappmönster”.

|  |  |
| --- | --- |
| (Figur 10a) | (Figur 10b) |
| (Figur 10c) | (Figur 10d) |

Då vi såg att Arean eventuellt såg ut att vara en bra variabel testade vi att logaritmera modellen, och fick följande data:  
F-värde för modellen är 34,22, vilket visar på att modellen i sin helhet är signifikant. Dock är variabeln Area fortfarande inte signifikant.  
R²- och Adjusted R²-värdena för modellen är 0,3581 respektive 0,3477, vilket är högre än modellen utan Area, men lägre än den ologgade modellen. Trots loggningen verkar variabeln Area inte ändras till signifikant och vi antar därför att det inte är modellstrukturen som påverkar.

Residualplottarna visar att residualerna inte kan antas ha likafördelad varians. Plottarna visar inte heller, precis som innan, på en normalfördelning.

|  |  |
| --- | --- |
| (Figur 11a) | (Figur 11b) |
| (Figur 11c) | (Figur 11d) |

**Vald modell av alla testade modeller**

Modell “Två variabler utan outliers”: Voterate = 0,50805 – 0,02627X1 – 0.32750X4 valdes. Detta eftersom F-värdet, koefficienternas t-värden är signifikanta och VIF-värdena är låga och därmed tycks det inte föreligga någon besvärande multikollinearitet.   
  
Modellen har ett R²-värde på 0.3687 och R²-Adjusted på 0,3619 vilket är högre än för samma modell inklusive outliers. I de olika testade modellerna förekom R²-värden som var högre, men då vissa koefficienter inte var signifikanta alternativt att det förekom höga VIF-värden valdes dessa bort. Modellen valdes även då residualanalysen för denna visade på att residualerna hade en konstant varians och är relativt normalfördelad.

**4. Avslutning**

I modellen med enbart en förklaringsvariabel gick det att utläsa att variabeln Population gav en förklaringsgrad på cirka 20 procent, vilket var det högsta värdet med avseende på de givna variabler.   
  
Även vid de olika modellerna med två förklaringsvariabler går det att utläsa att kombinationerna med högst förklaringsgrad innehåller variabeln Population, vilket ger ett starkt intryck av att det är en viktig variabel. Dock syns även en hög multikollinearitet med samtliga kombinationerna av variablerna Population-Education-Houses. Det är inte föga förvånande givet korrelationsmatrisen där samtliga kombinationer har en korrelation med varandra runt 0,98–0,99 vilket påvisar ett väldigt starkt positivt samband.  
  
I tabell 2 över beskrivande mått går det att utläsa att den beroende variabeln högst troligen innefattar outliers, då det maximala värdet var 1,10526. Då variabeln tolkas som andelen röster i området kan detta värde inte vara över 1, eftersom en andel inte kan vara mer än 100 procent!  
  
I och med misstanken om outliers testades att ta bort dessa både i den valda enkla och den multipla modellen. I båda fallen resulterade det i bättre residualplottar. Från att ha påvisat ett mönster i scatterplotten över residualerna, till att nu inte påvisa något mönster. Även normalfördelningen förbättrades från en svagt skev kurva till en mer normalfördelad kurva.  
  
För att kontrollera att interceptet inte förvred datan testades även modellen utan intercept, detta medgav dock att variabeln Longitude inte längre var signifikant och därmed slopades denna modell.

Vi drog slutsatsen att det fanns påverkande outliers i datan och vid fortsatt analys valdes att direkt eliminera dessa. För att undersöka om modellen möjligtvis borde bestå av alla givna variabler, trots att vi redan hade en misstanke om att så inte borde vara fallet, skapades en modell med samtliga variabler. Residualplottarna visade på gott resultat, dock med en avvikande punkt som tros bero på Dummyvariabeln Area, och därmed inte helt avgörande för vår analys. Som misstänkt visade variablerna Population-Education-Houses höga värden på VIF, samt slutade variablerna Longitude och Area att vara signifikanta. Därmed, trots den höga förklaringsgraden, valde vi att inte gå vidare med denna modell då misstanke om multikollinearitet förelåg.

Då de tre variablerna Education-Population-Houses tydligt visade på att de inte borde vara i samma modell undersöktes ifall enbart Area skulle adderas. Detta testades genom att lägga till enbart den, logaritmera den beroende variabeln samt att skapa en modell utan intercept. Samtliga analyser påvisade ett högt p-värde för variabeln Area och därmed uteslöts även denna.  
  
Slutsatsen drogs därmed att den bästa modellen, utifrån den givna data, var den med två förklaringsvariabler: Population och Longitude med korrigering av outliers.

Den valda modellen har en förklaringsgrad på 37 procent, vilket kan anses vara lågt, men med de variabler vi har att tillgå anser vi att modellen är det bästa alternativet. Alternativen med högre förklaringsgrad har, som tidigare nämnts, lägre validitet och därmed anses de inte vara tillförlitliga.

Residualplottarna visar inte på perfekt normalfördelning, men detta tros bero på att urvalet består av enbart 190 observationer. Vi anar att histogrammet över normalfördelningen kommer att jämnas ut om ett större urval undersöks.

För de enskilda residualplottarna, figur 7 f, går det att utläsa att variansen för Longitude inte är likafördelad, vilket förmodligen beror på att variabeln är en geografisk variabel (tillhörande det geografiska koordinatsystemet) och därmed är beroende av Latitude.

Den valda modellen har låga VIF-värden (1,04572) och F-värde på 54,60. Då kritiska F-värdet är 3,00 tolkas modellen vara signifikant. Vid en analys av de enskilda variablerna går det att utläsa att samtliga koefficienter är signifikanta.

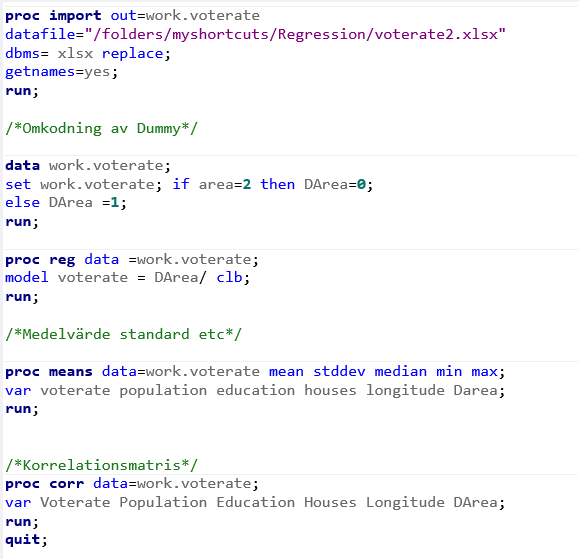
Då vi misstänker att modellen med variabel Longitude är beroende av variabeln Latitude hade det varit intressant att även testa en modell innehållande variabeln Latitude för att se om förklaringsgraden ökar samt om residualplottarna blir bättre.

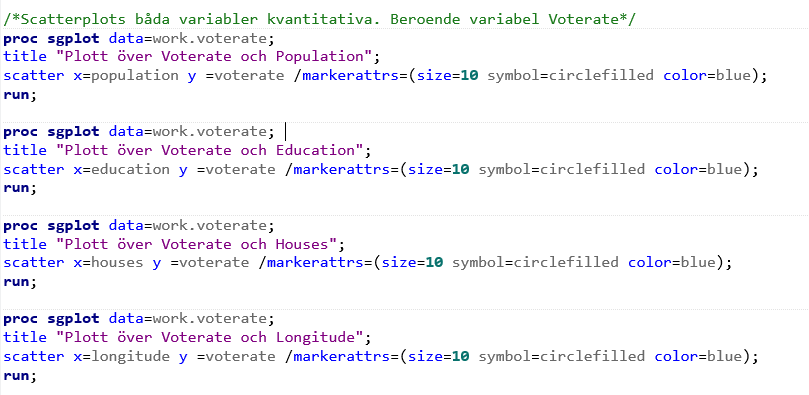
**5. Referensförteckning**

Andersson, G. Jorner, U. och Ågren, A. (2007). Regressions- och tidsserieanalys. 3:8 uppl., Lund: Studentlitteratur.

Nyqvist, H. (2017). Statistikens Grunder. 1:2 uppl., Lund: Studentlitteratur.

**6. Bilagor (med SAS-kod obligatorisk):  
  
Sas kod:**

****

****

