MT5021 — Övningsprojekt A8 – Tillförsel av kväve till Östersjön

Mostafa Amini, Daniel Svedlund, Adam Carlén

2025-10-01

1 Sammanfattning

Vi analyserar sambanden mellan tillförsel av löst oorganiskt kväve (DIN) och demografiska, hydrologiska samt markanvändningsrelaterade faktorer i Östersjöns avrinningsområden. Data förbehandlas (enhetsbyten, loggar, andelar) och utforskas grafiskt. Linjära modeller med log-transformerad respons jämförs mot rå-respons och olika varianter av förklarande variabler. Variabelselektion görs med AIC, multikollinearitet kontrolleras med VIF och antaganden granskas med standarddiagnostik. Leave-One-Out CV med Duans smearing används för prediktion på kg-skalan. Resultaten visar att DIN främst ökar med befolkning, avrinning och antal svin, medan hög anslutning till reningsverk minskar DIN. Sverige har lägre grundnivå än övriga länder men kan inte förkasta nollhypotsen att sverige har samma lutningar som grannländer. Den bästa prediktionsmodellen är log-respons med log-arealer (stepAIC).

2 Inledning

Målet är att modellera och jämföra hur DIN (dissolved inorganic nitrogen) i tillrinnande vattendrag påverkas av avrinningsområdets storlek, befolkning, markanvändning, djurhållning, avrinning samt reningsanslutning, och att bedöma om sambanden skiljer mellan Sverige och övriga länder. Vi utvärderar modeller både för inferens och prediktion (LOOCV).

3 Data

Källa och struktur. Datasetet `baltic_DIN.csv` består av avrinningsområden med variabler $x_1, \ldots, x_{20}$ och respons y (DIN, kg). Vi omvandlar areor från ha till km², bildar loggar och andelar enligt uppgift. Datakontroller (andelar, täthet, arealsummor) utföll väl utan avvikelser av praktisk betydelse.

Länder och antal. Datasetet omfattar 7 länder och ett "MISSING" kategori:

ESTONIA (4), FINLAND (33), LATVIA (5), LITHUANIA (1), MISSING (22), POLAND (2), RUSSIA (2), SWEDEN (36)

Vi ser att Litaun har en observation och Polen och Ryssland har endast 2.

Variabler (urval).

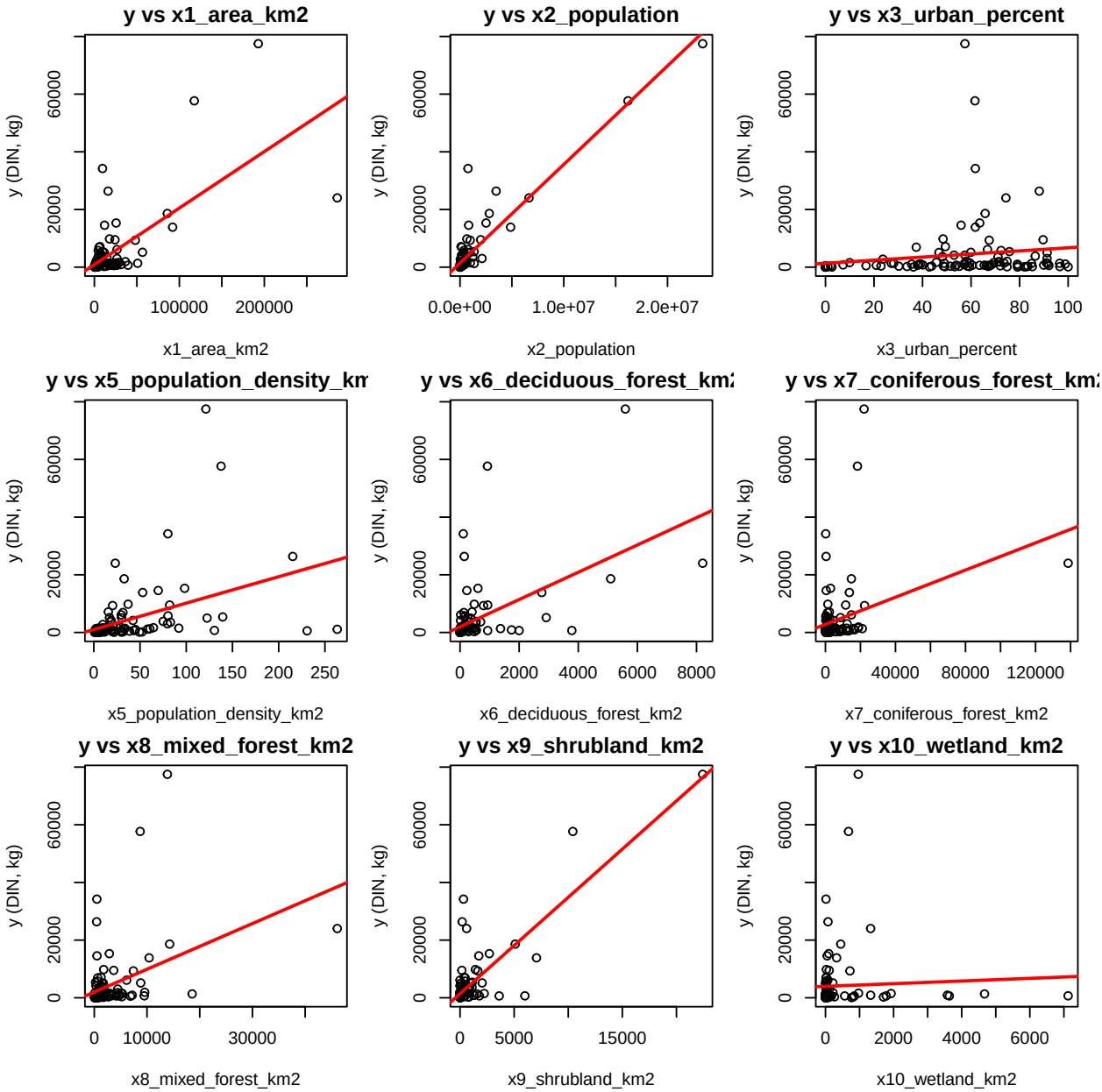
Variabel	Typ	Beskrivning
y	numerisk	DIN (kg)
$x1$	numerisk	Avrinningsområde (ha → km²)
$x2$	numerisk	Befolkning
$x3, x4$	numerisk	Andel stad/land (%)
$x5$	numerisk	Befolkningstäthet (inv/km²)
$x6$–$x15$	numerisk	Markanvändningsareor (ha → km²)
$x16$	numerisk	Avrinning (mm/år)

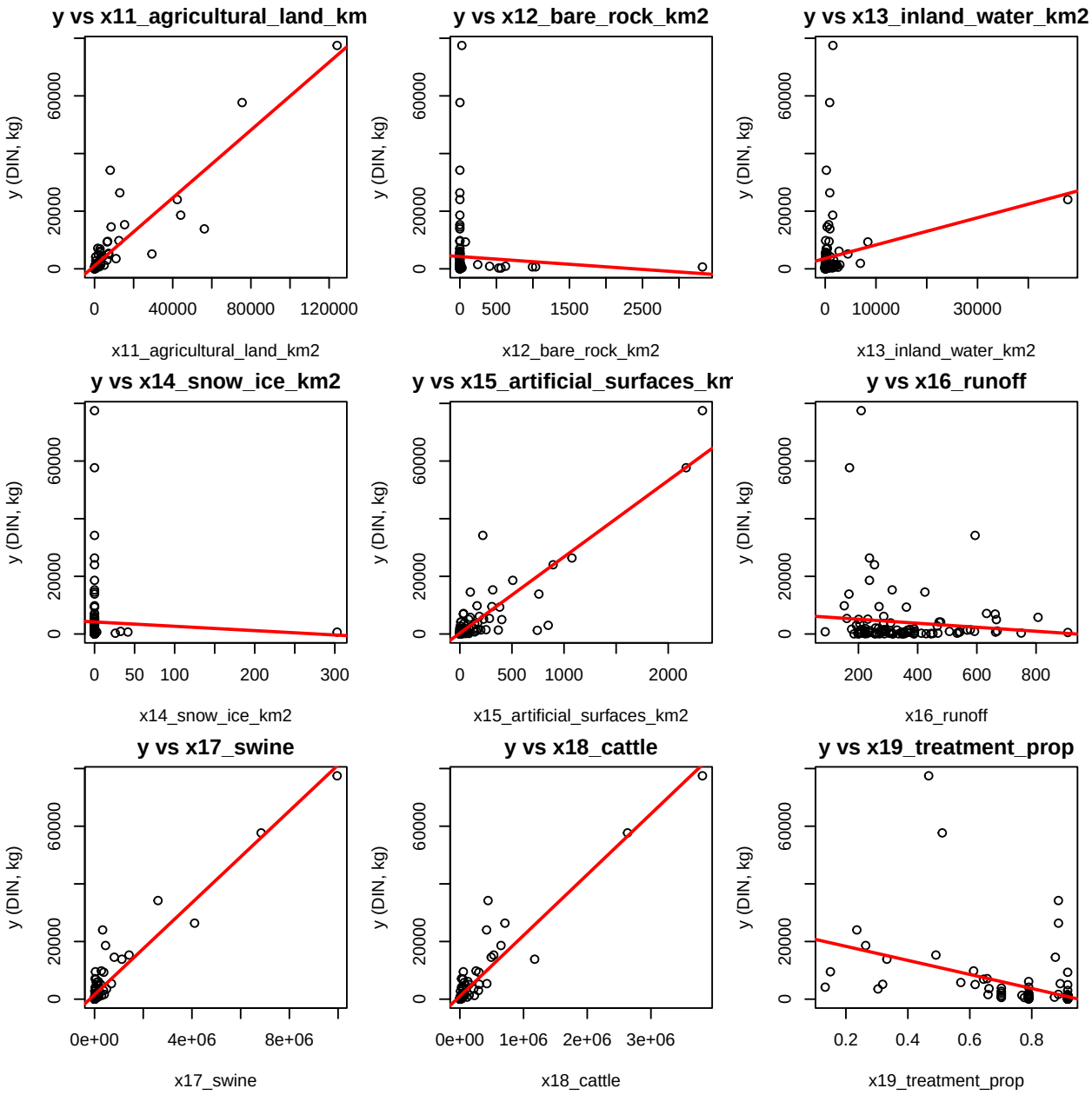
Variabel	Typ	Beskrivning
$x17, x18$	numerisk	Svin/Nöt (antal)
$x19$	numerisk	Anslutning reningsverk (%)
$x20$	char	Land

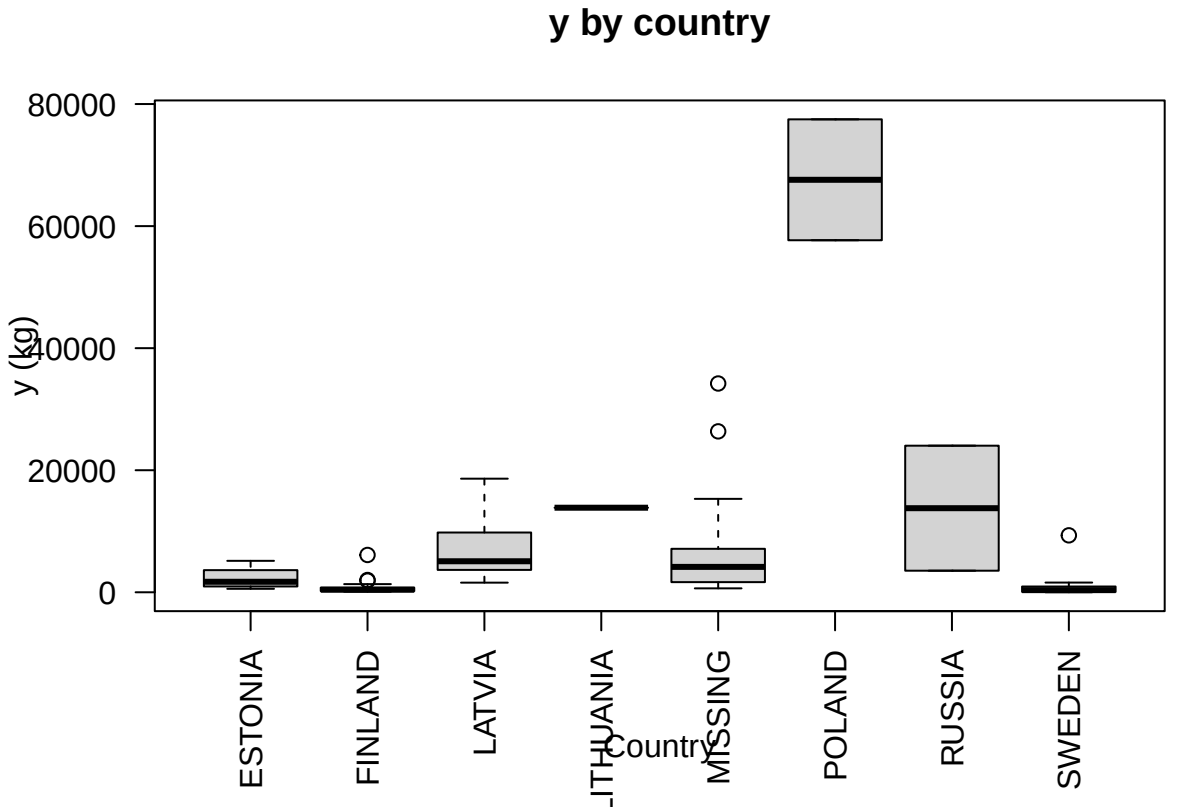
Vi omvandlade $x20$ till faktor för att utföra R funktioner.

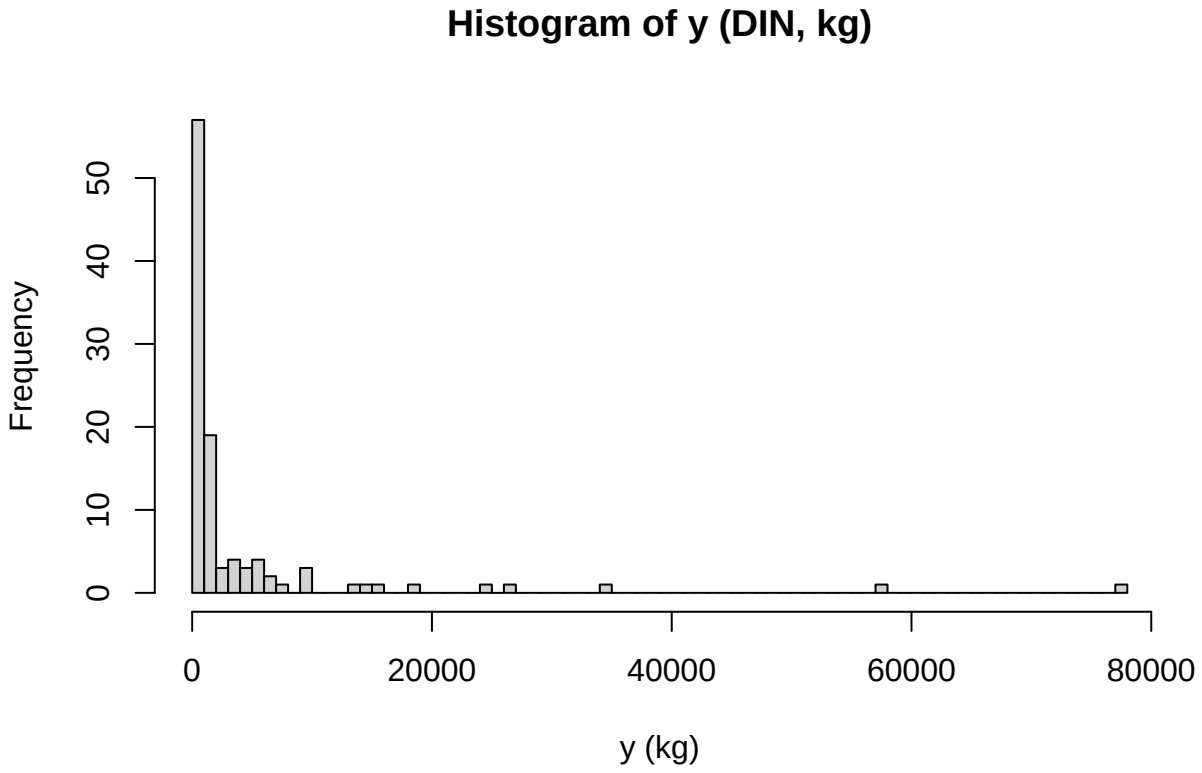
Data & Förberedelser

.




y vs x11_agricultural_land_km
y (DIN, kg)
0
20000
60000
0
40000
80000
120000
x11_agricultural_land_km2
y vs x12_bare_rock_km2
y (DIN, kg)
0
20000
60000
0
500
1500
2500
x12_bare_rock_km2
y vs x13_inland_water_km2
y (DIN, kg)
0
20000
60000
0
10000
30000
x13_inland_water_km2
y vs x14_snow_ice_km2
y (DIN, kg)
0
20000
60000
0
50
100
200
300
x14_snow_ice_km2
y vs x15_artificial_surfaces_km
y (DIN, kg)
0
20000
60000
0
500
1000
2000
x15_artificial_surfaces_km2
y vs x16_runoff
y (DIN, kg)
0
20000
60000
200
400
600
800
x16_runoff
y vs x17_swine
y (DIN, kg)
0
20000
60000
0e+00
4e+06
8e+06
x17_swine
y vs x18_cattle
y (DIN, kg)
0
20000
60000
0e+00
1e+06
2e+06
3e+06
x18_cattle
y vs x19_treatment_prop
y (DIN, kg)
0
20000
60000
0.2
0.4
0.6
0.8
x19_treatment_prop





Tolkning av scatterplot Spridningsdiagrammen (y mot x_1–x_{19}) visar tydligt positiva samband för x_1 (area), x_2 (befolkning), x_{11} (jordbruksmark), x_{15} (artificiella ytor), x_{17} (svin) och x_{18} (nöt); x_9 (buskmark) ser också positiv ut men effekten drivs av några få extrema värden, och x_{13} (inlandsvatten) är svagt–måttligt

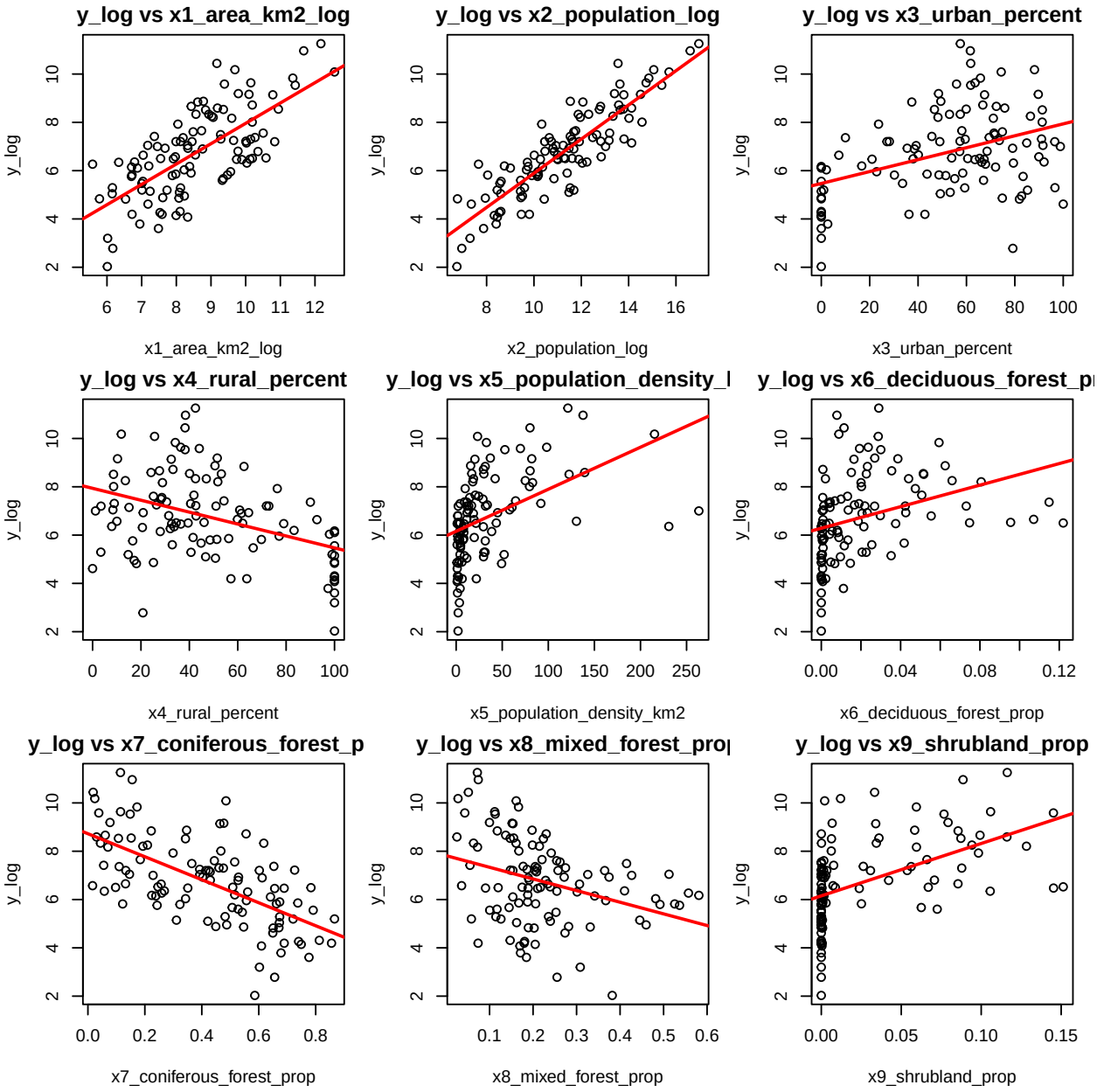
positiv. Sambanden är svaga eller oklara för x_3 (urban andel), x_5 (täthet), skogstyperna x_6–x_8 samt x_{10} (våtmark). Negativ lutning ses för x_{12} (häll/berg), x_{14} (snö/is) och x_{19} (reningsanslutning), medan x_{16} (avrinning) ligger nära noll eller svagt negativ i råa data. Variationen i y ökar med nivån och flera paneler påverkas av några mycket stora vattendrag (heteroskedasticitet och outliers). Boxplottar av y per land visar Polen högst, Ryssland och Lettland mellanhöga samt Sverige och Finland generellt låga nivåer, med stor inomlandsvariation och sned fördelning. Histogrammet visar att y är kraftigt högerskev med några extrema toppar, vilket motiverar en log-transform av y; dessutom ligger många punkter långt till vänster i plottarna, vilket ytterligare talar för log-transform för att få en mer symmetrisk fördelning.

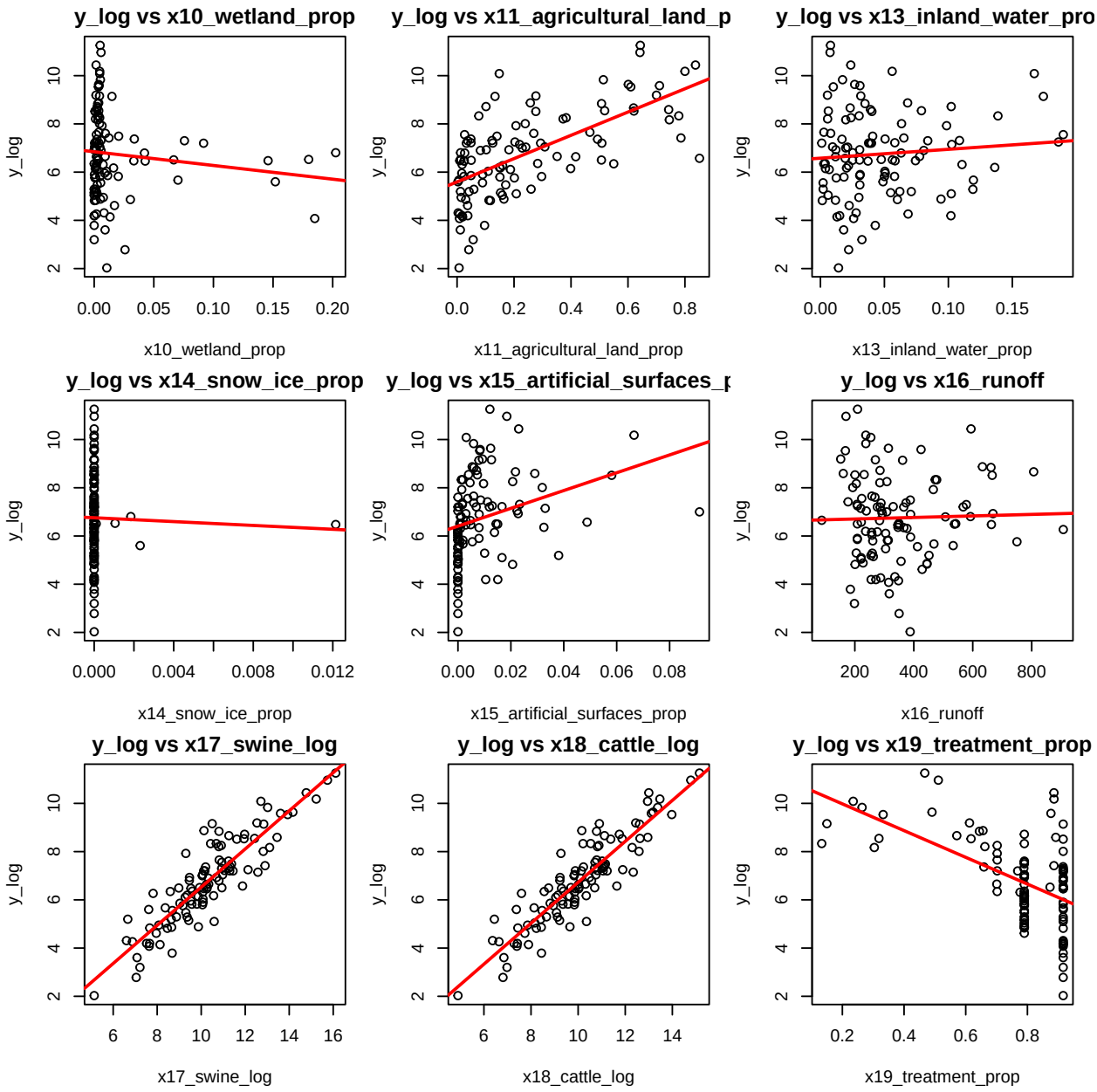
4 Statistisk modellering

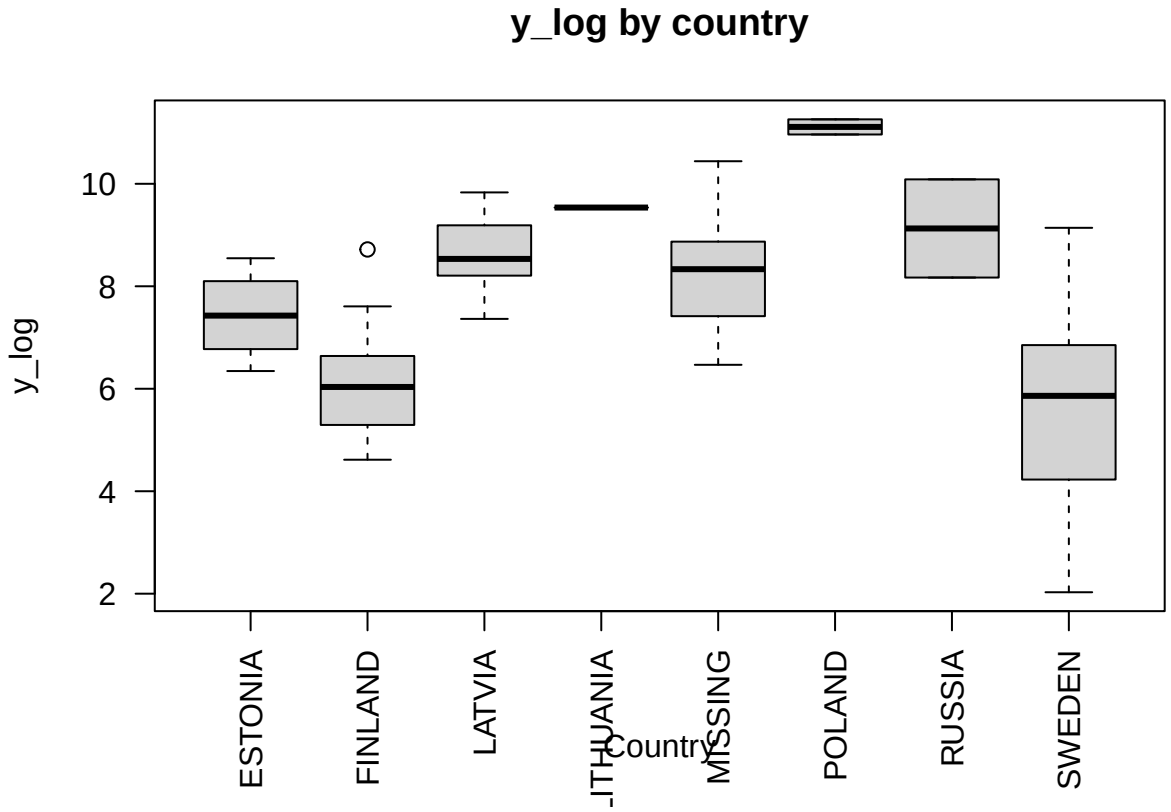
$x_3 + x_4 = 100$ (perfekt kollinearitet), x_5 är härledd ur x_2 och x_1, och summan av markandelsvariablerna x_6–x_{15} är x_1, vilket ger nära perfekt kollinearitet. Därför behöver vi utesluta minst en variabel (t.ex. x_{12}) och vi exkluderar x_4, x_5 och x_{12} när vi skapar linjära modeller. Vi börjar med att transformera responssvariabeln y samt x_1, x_2, x_{17} och x_{18}. För variablerna x_6–x_{15} använder vi proportionerna per observation, det vill säga värdet delat med x_1. Efter detta tog vi fram nya spridningsdiagram, histogram och boxplottar för att kontrollera effekterna.

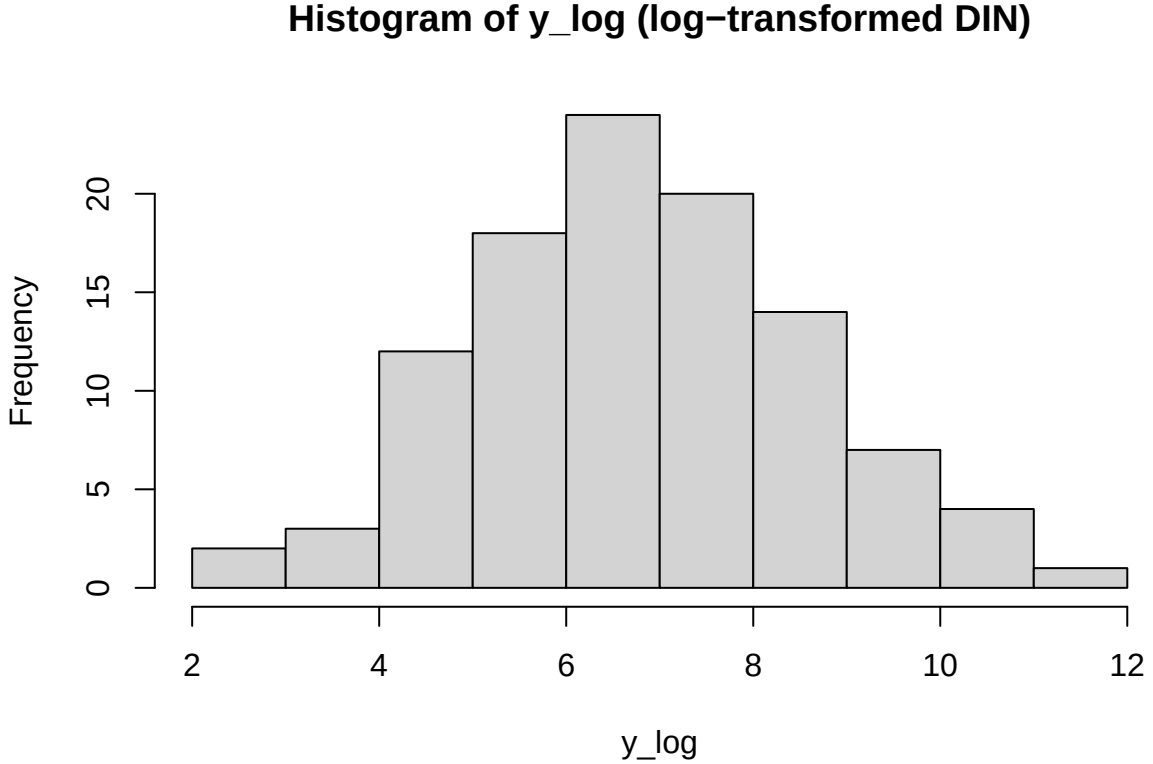
4.1 Grundläggande grafer

Vi ser positiva samband för area, befolkning, jordbruk, artificiella ytor, svin och nöt, och ett negativt samband för reningsanslutning. y är högerskev, därför använder vi en log-transform.


y_log vs x1_area_km2_log
y_log
x1_area_km2_log
y_log vs x2_population_log
y_log
x2_population_log
y_log vs x3_urban_percent
y_log
x3_urban_percent
y_log vs x4_rural_percent
y_log
x4_rural_percent
y_log vs x5_population_density_l
y_log
x5_population_density_km2
y_log vs x6_deciduous_forest_p
y_log
x6_deciduous_forest_prop
y_log vs x7_coniferous_forest_p
y_log
x7_coniferous_forest_prop
y_log vs x8_mixed_forest_pro
y_log
x8_mixed_forest_prop
y_log vs x9_shrubland_prop
y_log
x9_shrubland_prop


y_log vs x10_wetland_prop
y_log vs x11_agricultural_land_p
y_log vs x13_inland_water_pro
y_log vs x14_snow_ice_prop
y_log vs x15_artificial_surfaces_p
y_log vs x16_runoff
y_log vs x17_swine_log
y_log vs x18_cattle_log
y_log vs x19_treatment_prop
y_log
x10_wetland_prop
x11_agricultural_land_prop
x13_inland_water_prop
x14_snow_ice_prop
x15_artificial_surfaces_prop
x16_runoff
x17_swine_log
x18_cattle_log
x19_treatment_prop





Med $y_{\log}$ som respons ser vi starka positiva samband för x1_area_km2_log, x2_population_log, x5_population_density_km2, x11_agricultural_land_prop, x15_artificial_surfaces_prop, x17_swine_log och x18_cattle_log; även x9_shrubland_prop är positiv men delvis driven av några få stora fall. Sambanden är måttligt till svagt positiva för x3_urban_percent, x6_deciduous_forest_prop och

x13_inland_water_prop. Negativa samband ses för x4_rural_percent (spegel av x3), x7_coniferous_forest_prop, x8_mixed_forest_prop och x19_treatment_percent. Nära platta eller svagt negativa samband gäller för x10_wetland_prop, x14_snow_ice_prop samt x16_runoff, som ligger ungefär kring noll. Många av plotten ny visar tydliga linjära samband.

Boxplottar av den loggade responsen visar tydliga länderskillnader: POLAND har högst nivåer och liten spridning; RUSSIA och LITHUANIA ligger mycket högt (Litauen har en enstaka hög observation); LATVIA är också högt. ESTONIA och gruppen MISSING ligger mitt emellan, medan FINLAND och särskilt SWEDEN är lägst; Sverige har dessutom störst spridning bland de låga nivåerna.

Histogrammet av $y_{\log}$ är ungefär symmetriskt och klockformat, centrerat kring 7–8, vilket visar att log-transformen minskar skevhet, gör sambanden mer linjära och variansen jämnare.

Log-transformeringen gör därför att data passar bättre för linjär regression.

4.2 Samband i hela Östersjöområdet (log-respons)

Vi testade först alla variabler utan transformation (förutom x_4, x_5 och x_{12}). VIF-värdena var mycket höga för många variabler, vilket visar att variablerna är kollineära. Vi körde step-AIC i båda riktningar på denna modell, men VIF-värdena förblev höga.

Därefter valde vi att gå vidare med annat tillvägagångssätt: med log-transformerade responsvariabeln y och några förklaringsvariabler samt använde andelar för delområdena x_6–x_{15}. Vi testade alla variabler med dessa transformationer (fortfarande med x_4, x_5 och x_{12} exkluderade). Efter step-AIC i båda riktningar fick vi en reducerad modell med betydligt lägre VIF. Det justerade R^2 var 0.9705 för den icke-transformerade modellen, vilket var högre än för den transformerade modellen som hade $R^2 = 0.9348$.

Dessa variabler analyserade vi i vidare vår process: y_log ~ x1_area_km2_log + x2_population_log + x3_urban_percent + x6…x11 (proportioner, utom x12) + x13 + x14 + x15 + x16_runoff + x17_swine_log + x18_cattle_log + x19_treatment_percent + x20_country.

Sedan testade step-aic (båda riktningar) på modellen och fick vår reducerade förklaringsmodell.

I den fulla modellen ser vi starka och signifikanta samband för x2_population_log (+), x16_runoff (+) och x19_treatment_percent (–). x17_swine_log är svagt positiv. Övriga markanvändnings- och djurvariabler, inklusive x18_cattle_log, visar ingen tydlig effekt. På landsnivå ligger SWEDEN lägre än referenslandet Estland ($p < 0.01$). Modellens förklaringsgrad är hög med justerat $R^2 \approx 0.93$.

Stegvis AIC model (båda riktningar).

```
##
## Call:
## lm(formula = y_log ~ x2_population_log + x3_urban_percent + x6_deciduous_forest_prop +
##     x7_coniferous_forest_prop + x9_shrubland_prop + x16_runoff +
##     x17_swine_log + x19_treatment_prop + x20_country, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.05222 -0.33054  0.02721  0.29985  0.94501
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)               -1.622774   0.582625  -2.785  0.00653 **
## x2_population_log          0.331411   0.063599   5.211 1.20e-06 ***
## x3_urban_percent          -0.004195   0.002113  -1.985  0.05020 .
## x6_deciduous_forest_prop   3.747991   2.232128   1.679  0.09664 .
## x7_coniferous_forest_prop  0.725966   0.347321   2.090  0.03945 *
## x9_shrubland_prop          3.154264   1.632690   1.932  0.05655 .
## x16_runoff                 0.002809   0.000337   8.337 8.81e-13 ***
```

```
## x17_swine_log              0.473932   0.066013   7.179 2.02e-10 ***
## x19_treatment_prop        -1.212398   0.433287  -2.798  0.00630 **
## x20_countryFINLAND        -0.279977   0.279275  -1.003  0.31881
## x20_countryLATVIA          0.115383   0.317659   0.363  0.71729
## x20_countryLITHUANIA      -0.819364   0.546841  -1.498  0.13758
## x20_countryMISSING        -0.292969   0.270194  -1.084  0.28116
## x20_countryPOLAND         -0.557154   0.460560  -1.210  0.22959
## x20_countryRUSSIA         -0.696897   0.441688  -1.578  0.11816
## x20_countrySWEDEN         -0.889581   0.298397  -2.981  0.00370 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4605 on 89 degrees of freedom
## Multiple R-squared:  0.9442, Adjusted R-squared:  0.9348
## F-statistic: 100.4 on 15 and 89 DF,  p-value: < 2.2e-16
```

Med stegvis AIC kvarstår i huvudsak samma kärnsignaler och modellen förenklas: x2_population_log (+), x16_runoff (+), x17_swine_log (+) och x19_treatment_percent (–). Därutöver framträder svagare markanvändningssamband: x7_coniferous_forest_prop (+), x6_deciduous_forest_prop (positivt, på gränsen), x9_shrubland_prop (positivt, på gränsen) samt x3_urban_percent (negativt, på gränsen). SWEDEN förblir lägre, och anpassningen är i stort sett oförändrad (justerat $R^2 \approx 0.935$).

Sammanfattningsvis ökar DIN främst med befolkning, avrinning och antal svin, medan hög anslutning till reningsverk minskar DIN. Skogs- och buskandeler kan bidra svagt positivt. Sverige har en lägre basnivå än övriga länder givet övriga faktorer.

VIF-kontroll

Vi testade VIF på den modellen som reducerades av step-aic. Samtidigt som vi testade step-aic på den modellen innan reduceringen. I den fullständiga modellen ser vi extrem multikollinearitet bland markanvändningsvariablerna: VIF $>$ 2000 för x7_coniferous_forest_prop, x8_mixed_forest_prop och x11_agricultural_land_prop, samt VIF $\approx$ 80–90 för x17_swine_log och x18_cattle_log. Trots att x12_bare_rock_prop exkluderats för att undvika perfekt kollinearitet kvarstår mycket hög multikollinearitet mellan övriga markanvändningsproportioner. Orsaken är att proportionerna är starkt korrelerade med varandra—ökar en markanvändningstyp måste andra minska. De höga VIF-värdena innebär instabila koefficienter och opålitliga standardfel.

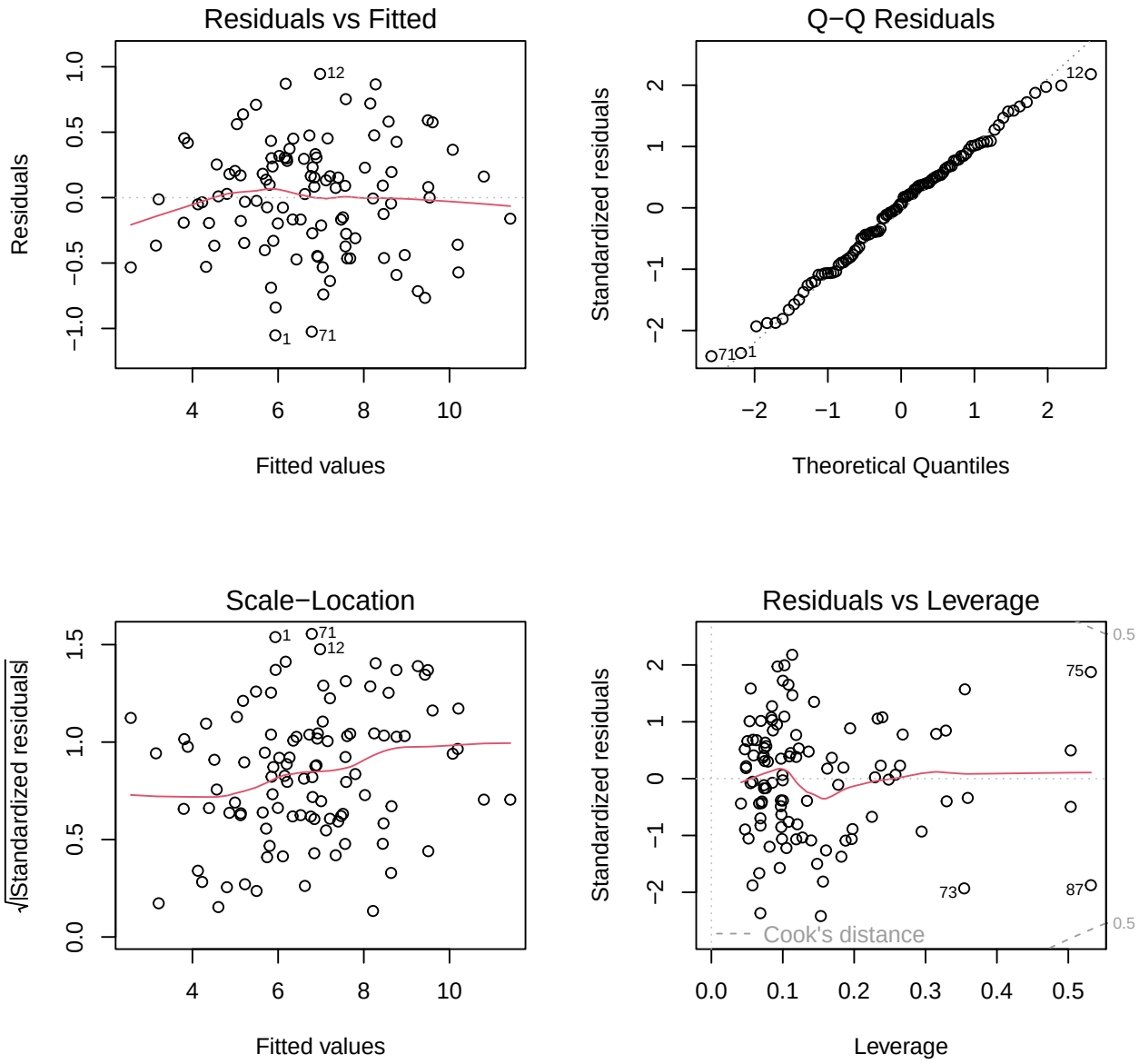
VIF-kontroll efter step-aic

```
##         x2_population_log          x3_urban_percent  x6_deciduous_forest_prop
##                  8.254936                  1.795104                  1.491762
## x7_coniferous_forest_prop         x9_shrubland_prop                x16_runoff
##                  2.810664                  1.856502                  1.180074
##             x17_swine_log        x19_treatment_prop
##                  7.893635                  1.504509
```

Tolkning av VIF-resultat

I StepAIC-modellen är multikollineariteten avsevärt reducerad. Alla VIF-värden är < 10; de högsta gäller x2_population_log (VIF = 8.3) och x17_swine_log (VIF = 7.9), medan majoriteten av VIF-värdena är < 3. StepAIC har därmed effektivt eliminerat de mest problematiska variablerna, vilket ger en mer stabil modell som är lämplig för inferens och tolkning av koefficienter. Vi ser inga anledning till att ta bort fler variabler.

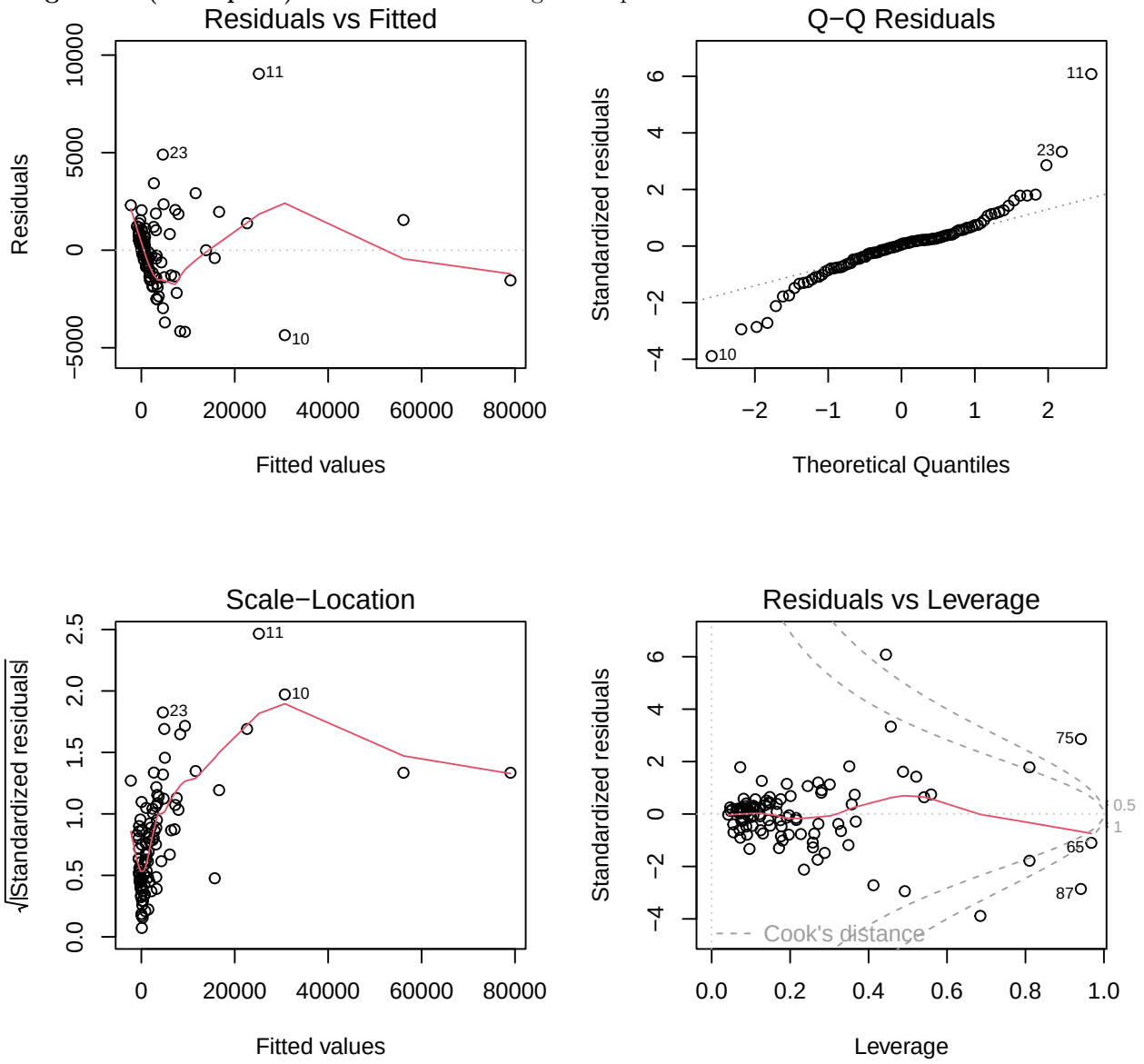
Diagnostik (transformerade modell efter stepa-ci) Vi testade med att ta fram diagnostikplottarna för den step-aic reducerade modell.



För step-AIC-modellen med log-y visar Residuals vs Fitted att punkterna ligger kring 0 utan tydlig kurvatur, vilket innebär att linearitet är okej. Q–Q-plottan ligger nära diagonalen med små svansavvikelser, så normaliteten är acceptabel. Scale–Location har en nästan plan loess-kurva, vilket tyder på högst mild heteroskedasticitet. I Residuals vs Leverage finns några punkter med något hög leverage, men Cook's D ligger väl under kritiska nivåer, vilket innebär att det inte finns några starkt inflytelserika observationer.

LM-antagandena är i stort uppfyllda och modellen är stabil och tolkbar för både inferens och prediktion.

Diagnostik (rå-respons) Vi testar modelldiagnostik på full modell med untransformerade variabler.



Diagnostik (rå-respons) Residuals vs Fitted visar tydlig kurvatur och en kraftig ”funnel”, vilket innebär bruten linearitet och stark heteroskedasticitet; några extrema punkter (t.ex. 11, 23, 10) sticker ut. Q–Q-plottan har klara avvikelser i båda svansarna, särskilt i den övre (11, 23), så normalantagandet håller dåligt. Scale–Location-plottan visar att variansen ökar med det anpassade värdet, vilket bekräftar tydlig heteroskedasticitet. I Residuals vs Leverage finns observationer med mycket hög leverage nära eller över Cook's $D \approx 1$ (punkterna längst till höger), vilket tyder på potentiellt starkt inflytande på koefficienterna.

Den linjära modellen på rå skala uppfyller alltså inte antagandena (nonlinearitet, icke-konstant varians och inflytelserika punkter). En log-respons är motiverad. Även om den otransformerade modellen har bättre justerat R^2 kan vi inte använda den eftersom antagandena inte håller. Den transformerade modellen uppfyller kraven för linjär regression bättre och är därför mer relevant.

Log-modellen uppfyller antagandena bättre än rå-modellen. Därför är vår slutliga förklaringsmodell den med transformerade variabler efter step-aic. Alltså med förklarande variabler x2_population_log, x3_urban_percent, x6_deciduous_forest_prop, x7_coniferous_forest_prop, x9_shrubland_prop, x16_runoff, x17_swine_log, x19_treatment_prop och länder.

4.3 Samma variabler i Sverige som i grannländer?

Vi skapade en kategorisk variabel ”Sweden group” som klassificerar varje observation som Sverige eller icke-Sverige. Därefter passade vi en linjär regressionsmodell som inkluderade alla variabler tillsammans med Sweden group, men exkluderade den ursprungliga landkategorin. För att få en mer riktad uppsättning variabler inför interaktionstester körde vi step-AIC på denna modell och genomförde sedan ett ANOVA-test för interaktioner på den utvalda delmängden.

De variabler som valdes (tillsammans med Sweden group) var: y_log, x1_area_km2_log, x2_population_log, x3_urban_percent, x9_shrubland_prop, x16_runoff, x18_cattle_log och x19_treatment_percent.

Test av interaktioner och huvudeffekt (Sweden vs Non-Sweden).

```
## y_log ~ x1_area_km2_log + x2_population_log + x3_urban_percent +
##     x9_shrubland_prop + x16_runoff + x18_cattle_log + x19_treatment_prop
```

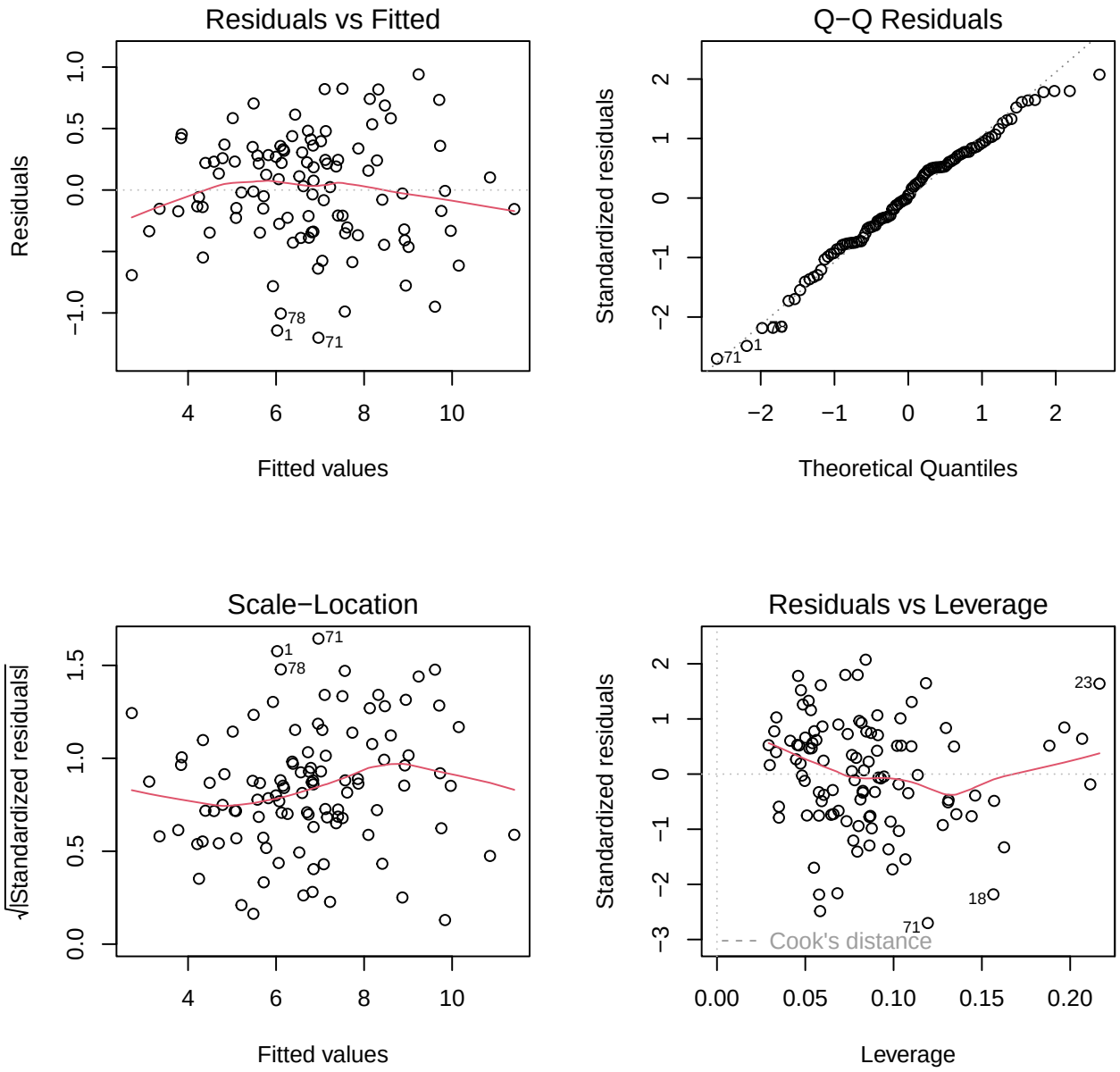
```
## Analysis of Variance Table
##
## Model 1: y_log ~ x1_area_km2_log + x2_population_log + x3_urban_percent +
##     x9_shrubland_prop + x16_runoff + x18_cattle_log + x19_treatment_prop +
##     sweden_group
## Model 2: y_log ~ (x1_area_km2_log + x2_population_log + x3_urban_percent +
##     x9_shrubland_prop + x16_runoff + x18_cattle_log + x19_treatment_prop) *
##     sweden_group
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     96 21.559
## 2     89 20.159  7    1.4003 0.8832 0.5232
```

$p = 0{,}523$, vilket ger inga bevis för att Sverige reagerar annorlunda på befolkning, avrinning, jordbruk m.m.

```
## Analysis of Variance Table
##
## Model 1: y_log ~ x1_area_km2_log + x2_population_log + x3_urban_percent +
##     x9_shrubland_prop + x16_runoff + x18_cattle_log + x19_treatment_prop
## Model 2: y_log ~ x1_area_km2_log + x2_population_log + x3_urban_percent +
##     x9_shrubland_prop + x16_runoff + x18_cattle_log + x19_treatment_prop +
##     sweden_group
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     97 25.151
## 2     96 21.559  1    3.5926 15.998 0.0001247 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Vi testar även huvudeffekten och får $p = 0{,}0001$, mycket signifikant. Sverige har därmed en annan grundnivå (intercept) för DIN än övriga länder även efter justering för de sju variablerna. RSS-värden: utan sweden_group $RSS = 25{,}15$, med sweden_group $RSS = 21{,}56$, en minskning på 3,59 (cirka 14 % mindre oförklarad variation).

Vi kan inte förkasta nollhypotesen att Sverige har samma lutningar som grannländerna—påverkan av variablerna är densamma—men Sverige har systematiskt lägre DIN-nivåer (annat intercept). Dessa slutsatser gäller under förutsättning att LM-antagandena är uppfyllda.



Diagnostik (vald log-modell) Residualerna är centrerade kring noll utan tydlig struktur; endast en svag trend anas, vilket innebär att linearitet en är okej. Q–Q-plotten ligger nära diagonalen med små svansavvikelser, så normalitetsantagandet är rimligt. Scale–Location visar en nästan platt loess; spridningen är något större runt mellannivåer (cirka 7–8 i anpassade värden) men ingen tydlig ”funnel”, vilket innebär högst mild heteroskedasticitet. I Residuals vs Leverage har de flesta punkter låg till måttlig leverage (mindre än cirka 0.20) och alla ligger inom Cook's $D < 0.5$, vilket innebär att det inte finns några klart inflytelserika observationer. Sammantaget är modellantagandena i stort uppfyllda, och små svans- eller variansskillnader bedöms inte påverka huvudslutsatserna.

Vi testade också med anova interaktioner med hela modellen utan att välja med step-aic. För den fulla modellen med alla prediktorer testade vi även ANOVA-interaktioner utan step-AIC och fick $p = 0{,}556$, vilket inte är signifikant. Slutsatsen blir densamma som för de utvalda variablerna: det finns inga bevis för att Sverige reagerar annorlunda på någon av de 16 prediktorerna. Diagnostikplottarna såg också bra ut för denna fulla modell. Vi testade dessutom med den variabelselektion som användes i avsnitt 4.2; inga låga p-värden framkom och vi kan inte förkasta nollhypotesen.

4.4 Prediktion (LOOCV) och Duans smearing

Vi utvärderar om DIN kan predikteras från förklaringsvariablerna genom att mäta modellernas prediktiva precision med Leave-One-Out Cross-Validation (LOOCV) via brute force: för varje observation i passar vi om modellen utan i, predikterar i på rätt skala, lagrar felet och upprepar detta för alla n. För log-responsmodeller backtransformerar vi prognoserna med Duan's smearing, eftersom log-transformen ger en systematisk bias vid baktransformering; därför används brute force i stället för hat-matrisen så att smearing-korrigeringen kan beräknas separat för varje träningsuppsättning.

Vi jämför tre modellfamiljer och deras stepAIC-varianter: (A) log-respons + log-arealer, (B) log-respons + arealandelar (proportioner), (C) rå respons + råa variabler. Urvalskriteriet är lägst LOOCV-RMSE på kg-skalan, och justerat R^2 rapporteras som komplement. Modellen med lägst LOOCV-RMSE används för slutlig diagnostik och redovisning av prediktioner. Litauen utesluts i LOOCV eftersom landet endast har en observation, vilket gör att landfaktorn saknas i vissa träningsfolds och modellen inte kan beräknas. För log-responsmodellerna (A, B, A_step, B_mapped) backtransformeras varje utelämnad prediktion enligt

$$\hat{y} = \exp(\hat{\eta}) \cdot \underbrace{\text{mean!}\big(\exp(e)\big)}_{\text{smearingfaktor}} - 1,$$

där $\hat{\eta}$ är den predikterade $\log(1+y)$ och e är residualerna från den aktuella träningspassningen.

Tre modellfamiljer.

```
##                         Model LOOCV_RMSE Relativt_prediktionsfel    Adj_R2
## 1  LOG-RESP + LOG-AREAS (A)   3626.017               0.3422803 0.9406082
## 2 LOG-RESP + PROP-AREAS (B)   9961.473               0.9403200 0.9322361
## 3    RAW-RESP + RAW-ALL (C)   4751.746               0.4485442 0.9699852
```

Tolkning: Modell A har lägst LOOCV-RMSE (3626) och ger bäst prediktion på kg-skalan, med justerat $R^2 \approx 0.941$. Modell C har högre LOOCV-RMSE (4752) men högst justerat $R^2 \approx 0.970$—den passar träningen bäst men predikterar sämre än A. Modell B är svagast (LOOCV-RMSE 9961; justerat $R^2 \approx 0.932$). Sammantaget är Modell A bäst för prediktion.

Vi testar nu att selektera en delmängd av variabler i alla tre fallen med step-AIC för att reducera modellerna till färre variabler och få enklare modeller.

LOOCV för step-modeller.

```
##                                     Model LOOCV_RMSE Relativt_prediktionsfel
## 1  A_step (LOG-RESP + LOG-AREAS, stepAIC)   3671.416               0.3465658
## 2 B_step (LOG-RESP + PROP-AREAS, stepAIC)   5613.371               0.5298780
## 3        C_step (RAW-RESP + RAW, stepAIC)   4653.785               0.4392972
##      Adj_R2
## 1 0.9428590
## 2 0.9347983
## 3 0.9704919
```

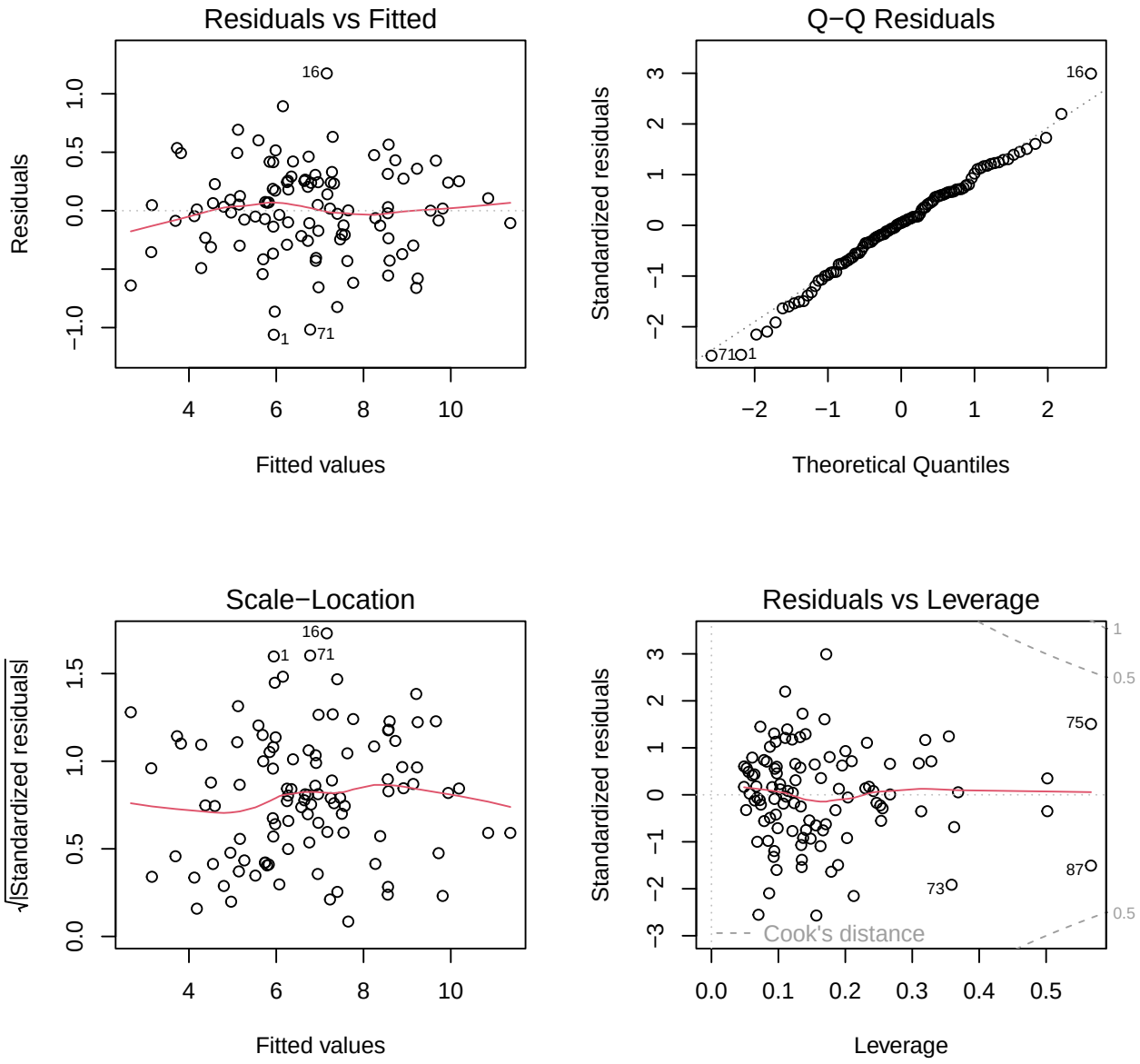
A_step (log y, log-areor) behåller x2_population_log (+), x16_runoff (+) och x17_swine_log (+); för markanvändning ingår x7 (+), x9 (+), x11 (+) och x8 (−), medan x3 är svagt negativ. Justerat $R^2_{\text{adj}} \approx 0.943$ och LOOCV-RMSE ≈ 3671 (bäst). B_step ger sämst prediktion med LOOCV-RMSE ≈ 5613. C_step (rå y, råa areor) har många areor (+); x1 (−) ("total vs delar") och x2 (−) efter kontroll; samt x16 (+), x17 (+), x18 (+) och x19 (−), med $R^2_{\text{adj}} \approx 0.970$ och LOOCV-RMSE ≈ 4654.

Vi väljer A_step som bästa prediktionsmodell: den använder log-transformerad respons y samt log-transformerade areor, befolkning och djur. Även om den fulla A-modellen utan selektion hade något lägre LOOCV-RMSE (3626) än A_step (3671) är skillnaden liten, och A_step är avsevärt enklare med färre variabler, vilket är att föredra.

Summary för slutliga prediktionsmodel

```
##
## Call:
## lm(formula = y_log ~ x2_population_log + x3_urban_percent + x7_coniferous_forest_km2_log +
##     x8_mixed_forest_km2_log + x9_shrubland_km2_log + x11_agricultural_land_km2_log +
##     x16_runoff + x17_swine_log + x18_cattle_log + x19_treatment_prop +
##     x20_country, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.0612 -0.2363  0.0180  0.2518  1.1741
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   0.2361757  0.6565208   0.360 0.719915
## x2_population_log             0.3318175  0.0666782   4.976 3.24e-06 ***
## x3_urban_percent             -0.0037869  0.0020355  -1.860 0.066195 .
## x7_coniferous_forest_km2_log  0.1653799  0.0685384   2.413 0.017923 *
## x8_mixed_forest_km2_log      -0.2205546  0.0837394  -2.634 0.009992 **
## x9_shrubland_km2_log          0.0517092  0.0256879   2.013 0.047209 *
## x11_agricultural_land_km2_log 1.0941674  0.3165246   3.457 0.000848 ***
## x16_runoff                    0.0028478  0.0003234   8.806 1.12e-13 ***
## x17_swine_log                 0.5888489  0.2129107   2.766 0.006935 **
## x18_cattle_log               -1.1720642  0.4540882  -2.581 0.011522 *
## x19_treatment_prop            0.9578997  0.6890732   1.390 0.168037
## x20_countryFINLAND            0.5651820  0.3425028   1.650 0.102518
## x20_countryLATVIA             0.3620599  0.3006057   1.204 0.231688
## x20_countryLITHUANIA         -0.1537447  0.5534108  -0.278 0.781815
## x20_countryMISSING            0.0608383  0.2790763   0.218 0.827940
## x20_countryPOLAND             0.0415309  0.4791640   0.087 0.931130
## x20_countryRUSSIA            -0.3458610  0.4374038  -0.791 0.431261
## x20_countrySWEDEN            -0.3455267  0.3179597  -1.087 0.280171
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4311 on 87 degrees of freedom
## Multiple R-squared:  0.9522, Adjusted R-squared:  0.9429
## F-statistic: 101.9 on 17 and 87 DF,  p-value: < 2.2e-16
```

4.5 Slutlig diagnostik (vald modell)



Residuals vs Fitted visar att residualerna ligger kring 0 med en mycket svag trend. Det finns enstaka utstickare (t.ex. 16 och 71) men inget systematiskt mönster. Q–Q-plotten ligger nära en rät linje med små svansavvikelser (höger: 16; vänster: 71 och 1), vilket gör normaliteten acceptabel. Scale–Location visar ungefär konstant spridning med högst mild heteroskedasticitet runt mellannivåer. I Residuals vs Leverage har de flesta punkter låg–måttlig leverage ($<\approx 0.3$); inga går över 0.5. Cook's $D < 0.5$, vilket innebär att inga observationer är starkt inflytelserika.

Sammanfattningsvis är LM-antagandena för A_step på $y_{\log}$ i stort uppfyllda, LOOCV_RMSEP-värdena är stabila och det finns inga inflytelserika observationer.

5 Resultat

Det går att påvisa statistisk signifikans för kvantitativa samband mellan våra förklarande varaibler och DIN. Den modell som etableras i slutet av del 4.2 uppfyller antaganden nödvändiga för linjär regression, har ett högt justerat R^2-värde, acceptabla värden vid VIF test, och har låga p-värden för några av förklaringsvari-

ablerna vilket visar på statistisk signifikans. Mer specifikt så påvisas ett positivt samband mellan DIN Befolkningen i området, avrinningen (mm/år , dvs liter per kvadratmeter och år) och antal svin inom området. Ett negativt samband etableras mellan DIN och andel av befolkningen som är ansluten till reningsverk.

Vidare för den valda modellen, vid genomfört anova test, så kan det inte påvisas att de förklarande variablerna skulle påverka DIN annorlunda i Sverige jämfört med reserande länder. Vi får därmed godta förklaringsmodellen som hyffsat gilltig för hela östersjö området.

Slutligen etableras i del 4.5 en prediktionsmodell som vi finner som den bästa bland de som testas med $RMSEP = 3671$ och ett gott relativt prediktionsfel ca 35 i balans med en hög förklaringsgrad justerad R^2 på ca 0.94.

6 Diskussion

Variabler att logaritmera(eller inte)

I den itterativa proccess som leder oss fram till den slutliga förklarringsmodellen i del 4.1-4.2 så prövas många saker. En del förklaringsvariabler logaritmeras medan areor transformeras till areaandelar. Med så många förklarande variabler blir det otänkbart att prova alla kombinationer. Vägledande blir de individuella plottarna av förklaringvariabler mot responsvariabeln, samt diagnostiska plottar. Om det inte ser ut att uppfylla modellantagenden eller om de plottarna är svårtolkade så provas en transformation.

Även om flera, inte alla, plottar som involverar en area som förklaringsvariabel, ser bättre ut efter en transformation till areaandel, så är det tänkbart att istället logaritmera dem, vilket inte prövas.

Outliers

I del 4.4-4.5 så förekommer för modell B och C några inlytelserika outliers. Det är tänkbart att prova att plocka bort dem och se om $RMSEP$-värdet förbättras utan en för stor förlust i justerad R^2

7 Gruppmedlemmarnas arbetsinsats

Vi alla började jobba på eget håll med egna modeller osv... för att sedan sammanstråla vid rapportskrivningen.

A Appendix: Grafer, R-kod, output om modeller