F9 - Multipel linjär regression

Statistisk översiktskurs

Ulf Högnäs

Statistiska institutionen
Stockholms universitet

April 9, 2025

Contents

Dummyvariabler

Loans

Enkel linjär regression

```
library(openintro)
loans <- loans_full_schema
View(loans)
table(loans$public_record_bankrupt)
# Ny variabel, 1 = minst en konkurs, 0 = inga konkurser
loans$bankrupcy <- as.numeric(loans$public_record_bankrupt>0)
# Skatta en linjär modell
model_bankrupt <- lm(interest_rate ~ bankrupcy, data = loans)
summary(model_bankrupt)
```

Dummyvariabel

Bankrupcy är ett exempel på en **dummyvariabel**

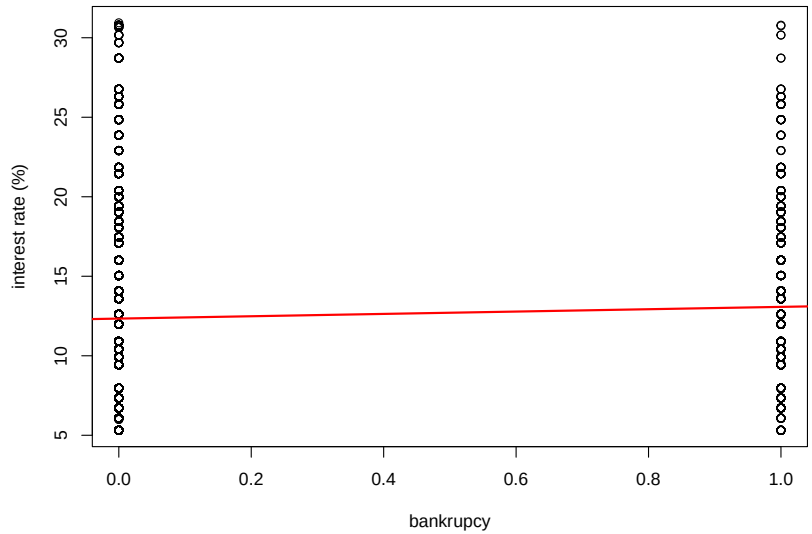
1	minst en konkurs
0	inga konkurser

```
Call:
lm(formula = interest_rate ~ bankrupcy, data = loans)

Residuals:
    Min      1Q  Median      3Q     Max
-7.7648 -3.6448 -0.4548  2.7120 18.6020

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  12.3380     0.0533 231.490  < 2e-16 ***
bankrupcy     0.7368     0.1529   4.819 1.47e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.996 on 9998 degrees of freedom
Multiple R-squared:  0.002317,  Adjusted R-squared:  0.002217
F-statistic: 23.22 on 1 and 9998 DF,  p-value: 1.467e-06
```



Flera “levels” med dummyvariabler

- Många kategoriska variabler har fler än två kategorier
- Exempel: `verified_income`

```
Not Verified Source Verified        Verified
        3594            4116            2290
```

- `verified_income` har tre “levels” i R
- Vi skattar en modell ...

```
table(loans$verified_income)
model_verified <- lm(interest_rate ~ verified_income, data = loans)
summary(model_verified)
```

Dropped level

```
> model_verified <- lm(interest_rate ~ verified_income, data = loans)
> summary(model_verified)

Call:
lm(formula = interest_rate ~ verified_income, data = loans)

Residuals:
    Min      1Q  Median      3Q     Max
-9.0437 -3.7495 -0.6795  2.5345 19.6905

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                     11.09946    0.08091  137.18   <2e-16 ***
verified_incomeSource Verified   1.41602    0.11074   12.79   <2e-16 ***
verified_incomeVerified          3.25429    0.12970   25.09   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.851 on 9997 degrees of freedom
Multiple R-squared:  0.05945,   Adjusted R-squared:  0.05926
F-statistic: 315.9 on 2 and 9997 DF,  p-value: < 2.2e-16
```

Vad hände med kategorin “not verified”?

Base Category - den osynliga kategorin

- Tre alterativ för varje observation
  1. `verified_incomeSource Verified` $= 1$
     `verified_incomeVerified` $= 0$
  2. `verified_incomeSource Verified` $= 0$
     `verified_incomeVerified` $= 1$
  3. `verified_incomeSource Verified` $= 0$
     `verified_incomeVerified` $= 0$
- Alternativ 3 betyder ingen av dessa två, d.v.s. “not verified”
- Denna kategori har blivit **base category**
- I denna modell är den skattade räntan för “not verified” lika med interceptet, dvs. 11.1%[†]
- Varför är räntan lägst för denna grupp?

[†]Beräkna den skattade räntan för resterande grupper. Lösning på s. 139 i boken.

Multipel linjär regression

Multipel linjär regression

- Vi antar att det finns ett linjärt samband mellan responsvaribeln och flera förklaringsvariabler
- Vi antar också att det finns variation ε som inte föklaras av modellen
- $y = \beta_0 + \beta_1 \cdot x_1 + \cdots + \beta_k \cdot x_k + \varepsilon$
- Från den skattade modellen kan vi få prediktioner

$$\hat{y} = b_0 + b_1 \cdot x_1 + \cdots + b_k \cdot x_k$$

Roligt att veta

- Istället för att anpassa en linje i 2D så anpassar vi ett hyperplane i ett rum med högre dimensioner
- man kan inte rita dimensioner högre en tre, men matten funkar ändå

Exempel: utlåningsräntor

```
Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                     1.893839   0.210245   9.008  < 2e-16 ***
verified_incomeSource Verified  0.997468   0.099187  10.056  < 2e-16 ***
verified_incomeVerified         2.563168   0.117184  21.873  < 2e-16 ***
debt_to_income                  0.021832   0.002937   7.434 1.14e-13 ***
credit_util                     4.896988   0.161887  30.249  < 2e-16 ***
bankrupcy                       0.391105   0.132286   2.957  0.00312 **
term                            0.153388   0.003944  38.889  < 2e-16 ***
credit_checks                   0.228321   0.018242  12.516  < 2e-16 ***
issue_monthJan-2018             0.045510   0.108034   0.421  0.67358
issue_monthMar-2018            -0.041624   0.106545  -0.391  0.69605
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.3 on 9964 degrees of freedom
  (24 observations deleted due to missingness)
Multiple R-squared:  0.2604,    Adjusted R-squared:  0.2597
F-statistic: 389.7 on 9 and 9964 DF,  p-value: < 2.2e-16
```

En modell med flera förklaringsvariabler

Att tolka koefficienter, ceteris paribus

Vad betyder det att koefficienten för bankrupcy är 0.39?

- Jämför två kunder, Madison och Harper
  - Madion har aldrig gått igenom en konkurs
  - Harper har gått igenom (minst) en konkurs
  - Allt annat i modellen är lika för Madison och Harper
  - **Ceteris paribus** ("allt annat hålls konstant")
- I så fall är den skattade räntan 0.39 procentenheter högre för Harper än för Madison

Uppgift

Tolka koefficienten för `verified_incomeSource Verified` (den är 0.997).

Att välja modell

Förklaringsgrad (R^2)

I föreläsning 8 ingick en intuitiv beskrivning av R^2. Här är den matematiska definitionen.

R^2

Coefficient of Determination (sv. förklaringsgrad)

$$
\begin{aligned}
R^2 &= 1 - \frac{\text{Variationen i residualerna}}{\text{Variationen i utfallet}} \\
&= 1 - \frac{\mathrm{Var}(e)}{\mathrm{Var}(y)}
\end{aligned}
$$

R^2 mäter hur stor andel av variationen i utfallet som förklaras av modellen.

Justerad förklaringsgrad (R^2_{adj})

R^2_{adj} - Adjusted Coefficient of Determination

Den justerade förklaringsgraden tar hänsyn till antalet variabler i modellen och beräknas som

$$\begin{aligned} R^2_{\text{adj}} &= 1 - \frac{s^2_{\text{residualer}}/(n-k-1)}{s^2_{\text{utfall}}/(n-1)} \\ &= 1 - \frac{s^2_{\text{residualer}}}{s^2_{\text{utfall}}} \cdot \frac{n-1}{n-k-1} \end{aligned}$$

- n är antalet observationer som används för att anpassa modellen.
- k är antalet förklarande variabler i modellen.

Vad händer om antalet förklarande variabler ökar, utan att residualernas storlek minskas nämnvärt?

Backward elimination

Den förra modellen hade $R^2 = 0.2604$ respektive $R^2_{adj} = 0.2597$

```
Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                     1.896283   0.198294   9.563  < 2e-16 ***
verified_incomeSource Verified 0.996255   0.099139  10.049  < 2e-16 ***
verified_incomeVerified        2.560756   0.117131  21.862  < 2e-16 ***
debt_to_income                 0.021852   0.002937   7.441 1.08e-13 ***
credit_util                    4.896475   0.161845  30.254  < 2e-16 ***
bankrupcy                      0.391645   0.132272   2.961  0.00307 **
term                           0.153356   0.003943  38.893  < 2e-16 ***
credit_checks                  0.228409   0.018241  12.522  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.3 on 9966 degrees of freedom
  (24 observations deleted due to missingness)
Multiple R-squared:  0.2603,    Adjusted R-squared:  0.2598
```

Vi tog bort `issue_month`. Vad hände med R^2 respektive R^2_{adj}?

Backward elimination

Ett sätt att välja vilka förklaringsvariabler som ska vara med är
Backward elimination

1. Skatta en stor modell med många eller alla förklaringsvariabler
2. Ta bort förklaringsvariabeln med sämst p-värde och skatta modellen igen
3. Om R^2_{adj} ökade, upprepa steg 2
4. Om R^2_{adj} minskade, lägg tillbaka variabeln och prova att ta bort en annan variabel
5. Du är klar när du inte kan förbättra R^2_{adj}

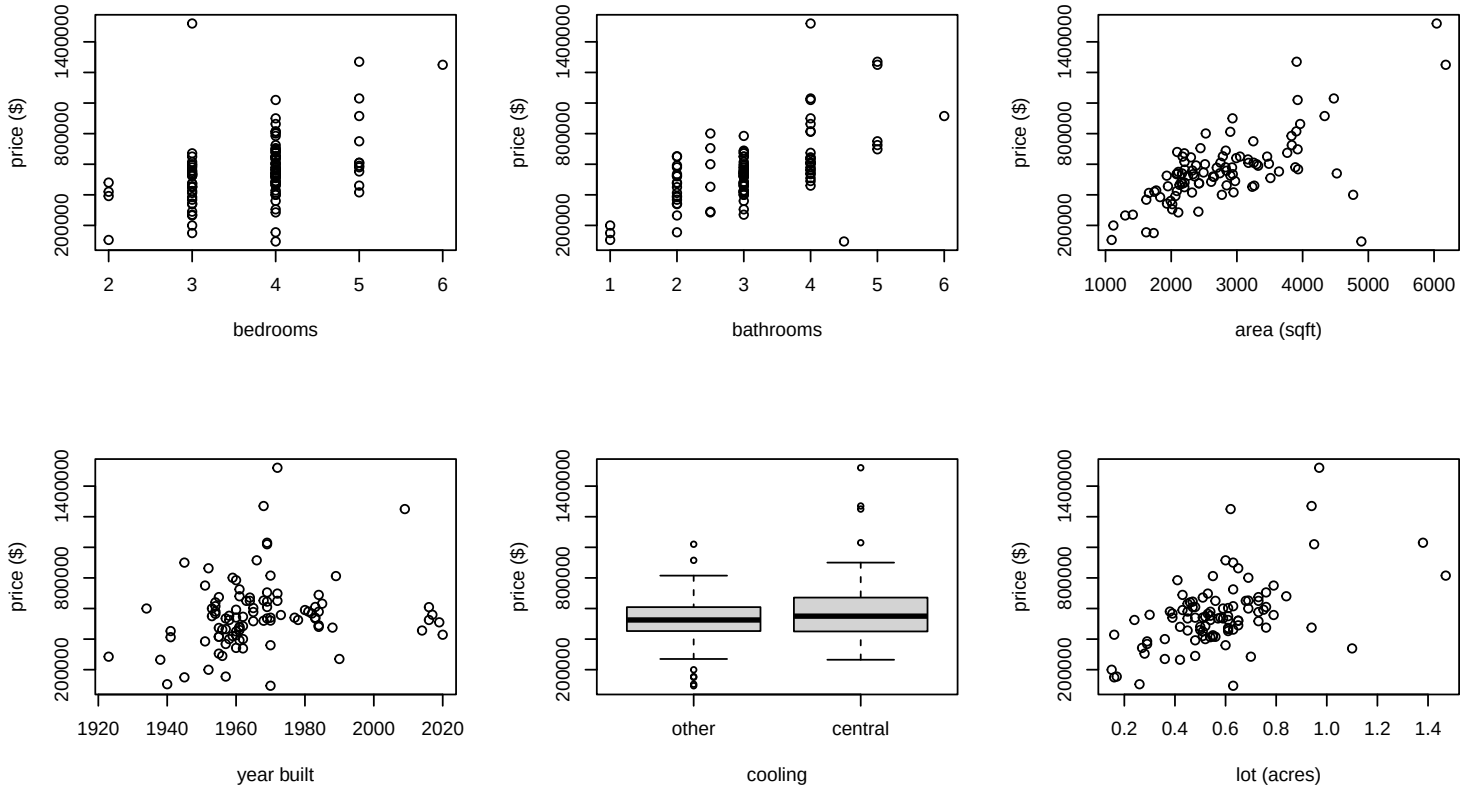
Case study: Houses for sale

Duke Forest

Duke Forest, Durham, NC. Zillow.com.

Det första huset i duke_forest från openintro. Zillow.com.

Korrelationsplottar



Några förklaringsvariabler vs pris

Första modellen

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -2.911e+06  1.788e+06  -1.628  0.10703
bed            -1.369e+04  2.593e+04  -0.528  0.59874
bath            4.108e+04  2.466e+04   1.666  0.09928 .
area            1.021e+02  2.313e+01   4.416 2.79e-05 ***
year_built      1.459e+03  9.140e+02   1.597  0.11385
coolingcentral  8.407e+04  3.034e+04   2.771  0.00679 **
lot             3.561e+05  7.594e+04   4.690 9.70e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 145100 on 90 degrees of freedom
```

Första modellen, från boken

Table 10.4: Summary of least squares fit for price on multiple predictor variables.

term	estimate	std.error	statistic	p.value
`(Intercept)`	-2,910,715	1,787,934	-1.63	0.107
`area`	102	23	4.42	<0.0001
`bed`	-13,692	25,928	-0.53	0.5987
`bath`	41,076	24,662	1.67	0.0993
`year_built`	1,459	914	1.60	0.1139
`coolingcentral`	84,065	30,338	2.77	0.0068
`lot`	356,141	75,940	4.69	<0.0001
Adjusted R-sq = 0.5896				
df = 90				

Från s. 172

Uppgifter

Uppgift

1. Tolka koefficienten för `bed`. Varför tror ni att den är negativ?
2. Vilken variabel bör vi prova att ta bort i vår backward elimination?

Observation 18.

Observation 83.

Andra modellen, vi eliminerade bed

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -2.953e+06  1.779e+06  -1.660  0.10043
bath            3.623e+04  2.280e+04   1.589  0.11552
area            9.906e+01  2.230e+01   4.443 2.49e-05 ***
year_built      1.466e+03  9.102e+02   1.611  0.11068
coolingcentral  8.386e+04  3.022e+04   2.775  0.00669 **
lot             3.571e+05  7.562e+04   4.723 8.41e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 144600 on 91 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.6141,    Adjusted R-squared:  0.5929
```

Förbättrades modellen när vi eliminerade bed?

Multikollinearitet

Multikollinearitet

Multikollinaritet uppstår när de förklarande variablerna är starkt korrelerade med varandra. När variablerna i modellen är alltför inbördes korrelerade kan det bli svårt att tolka regressionskoefficienterna.

Är det den arean eller antalet sovrum som avgör priset? Modellen kan inte hålla isär effekterna.

Tredje modellen, vi eliminerade `bath`

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -3.706e+06  1.729e+06  -2.143  0.03473 *
area            1.209e+02  1.772e+01   6.820 9.39e-10 ***
year_built      1.870e+03  8.813e+02   2.122  0.03653 *
coolingcentral  9.221e+04  3.000e+04   3.074  0.00278 **
lot             3.693e+05  7.585e+04   4.868 4.65e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 145800 on 92 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.6034,    Adjusted R-squared:  0.5861
```

Förbättrades modellen när vi eliminerade `bath`?

Winner: Andra modellen

Uppgift

Beräkna det skattade värdet för ett hus som

- har area 1661 sq ft
- har tre sovrum[a]
- har två badrum
- är byggt 1941
- inte har central cooling
- har en 0.54 acre lot

[a] Vad ska vi göra med denna information?

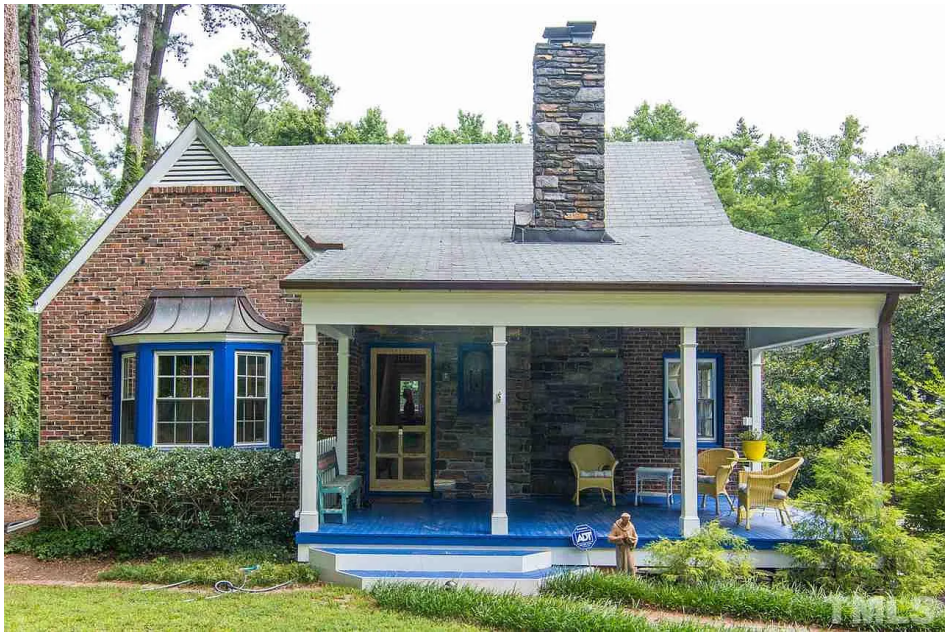
term	estimate
(Intercept)	-2,952,641
area	99
bath	36,228
year_built	1,466
coolingcentral	83,856
lot	357,119
Adjusted R-sq = 0.5929	
df = 91	

Koefficienter för den andra modellen

Beräkna Residualen för observation 69

Uppgift

Förklaringsvariablernas värden på förra bilden är hämtade från rad 69, *2618 Pickett Rd, Durham, NC 27705*. Husets försäljningspris var $\$412,500$. Vad blir residualen för denna observation?



Observation 69.

Modellens antaganden

Kort om modellens antaganden

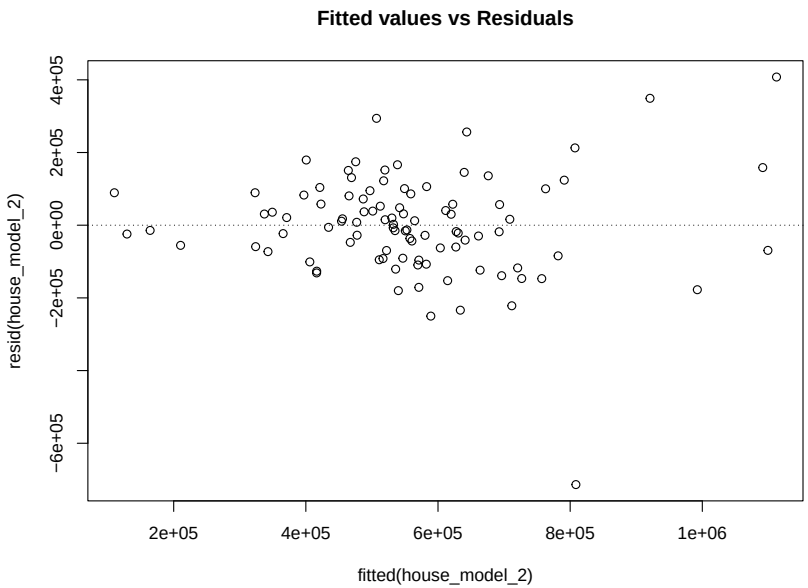
Regression vilar på några antaganden som måste stämma någorlunda för att modellen ska fungera. Se även bild 15 från föreläsning 8.

Antaganden för linjär regression

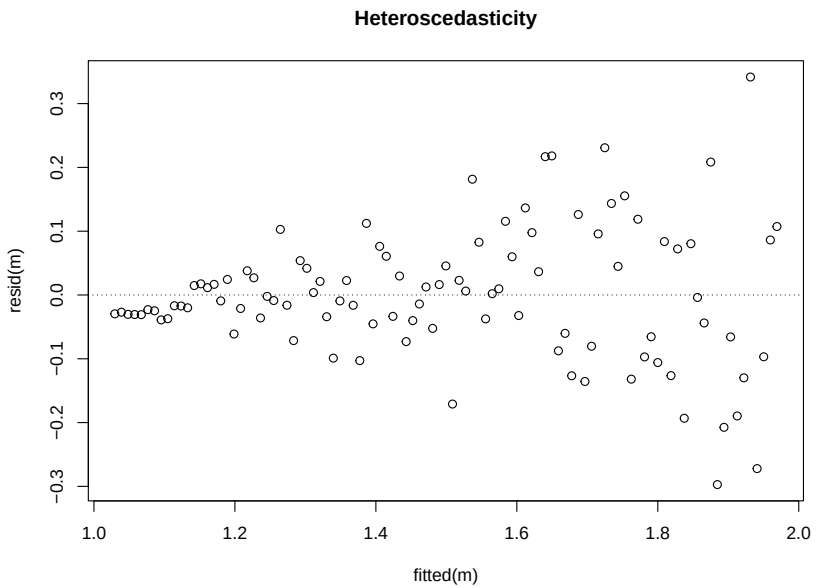
1. **Linjära samband** Sambandet mellan varje förklarande variabel och responsvariabeln är linjärt.
2. **Oberoende observationer** Observationerna är oberoende av varandra.
3. **Homoskedasticitet** Variansen i residualerna är konstant
4. **Normalfördelade residualer** Residualerna är normalfördelade.
5. **Ingen multikollinaritet** De förklarande variablerna är inte starkt korrelerade med varandra.

Exempel. Antagande 3: Homoskedasticitet

För att undersöka om homoskedasticietet är ett rimligt antagande kan man plotta skattade värden (x-axel) mot residualerna (y-axel)

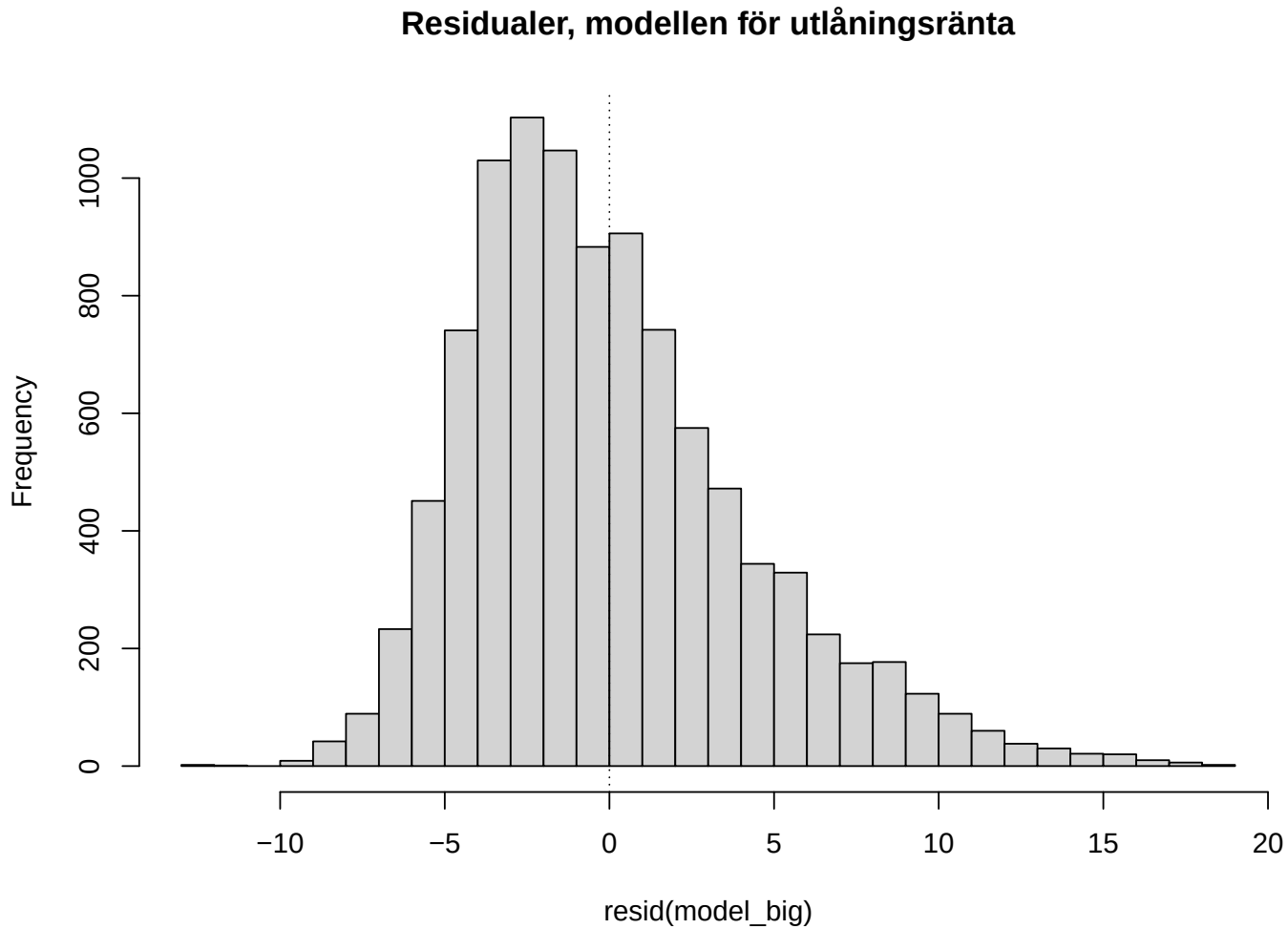


OK...? Inte det värsta jag har sett



Katastrof

Exempel. Antagande 4: Normalfördelade residualer

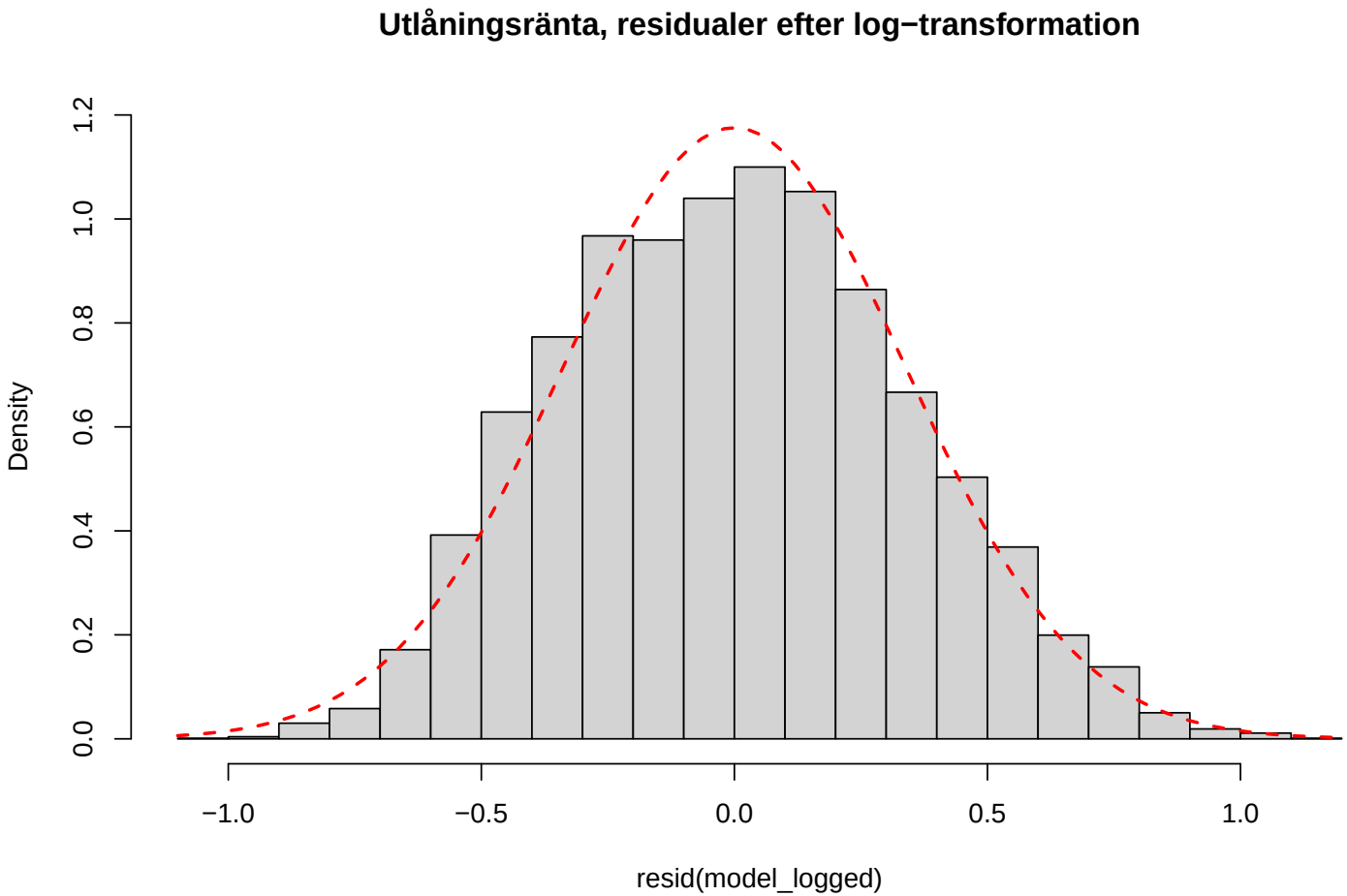


Är antagandet om normalfördelade residualer ungefär uppfyllt?

Transformationer

- Om modellens antaganden inte alls är uppfyllda kan vi inte lita på prediktioner och inferens
- Vi kan hitta resultat som egentligen är felaktiga
- Lösningen kan vara en **transformation** av någon variabel
- En vanlig lösning är att ta logaritmen av varje y-värde (responsvariabeln) och sedan skatta modellen igen
  1. Transfomera responsvariabeln med `log()`
  2. Skatta modellen
  3. Stoppa in varje prediktion i `exp()` för att få värden som går att förstå

Residualer efter en framgångsrik transformation



Detta ser bra ut