Kunskapskontroll

R-programmering

Kamila Nigmatullina

EC Utbildning

202404

Abstract

A goal of this report was to create a model for predicting prices for electric cars. The model was created using “backwards selection” which helped to find out significant characteristics influencing the price. The variables were chosen: millage, model-year, type of the car, horses, model.

The model was successfully chosen and evaluated on the test-data.

Innehållsförteckning

1 Inledning

I nytiden när marknad av begagnade personliga bilar utvecklas mer och mer är det viktigt att inte stå stilla och förbättra tjänster och verktyg kring det. Det hade varit en stor konkurrensfördel för webbplatformer som ger möjlighet att sälja och köpa personliga bilar att ha en automatisk prissättning tjänst. Då kan platformen rekommendera prissättningen till kunden utan att hen behöver bläddra och undersöka marknaden själv. Den här tjänsten skulle spara massa tid för kunder och locka mer användare till platförmen.

Särskilt ny marknaden i detta område är marknaden av elbilar. Därför kan det bli svårt för människor att bestämma pris vid försäljning av sin begagnad bil.

Med tanke på detta bestämde vi att skapa en Linjer-regressionsmodell som skulle prediktera pris på begagnade elbilar.

Syftet med denna rapport är att skapa en fungerande modell sim ska hjälpa att prediktera pris på begagnade elbilar, baserad på flera karakteristiker.

För att uppfylla syftet så kommer följande frågeställning(ar) att besvaras:

1. Går det att samla in relevant data själv från existerande platformer? (Blocket.se)
2. Går det att hitta linjära samband mellan olika karakteristiker av bilar och pris?
3. Går det att skapa en fungerande modell som ska hjälpa att prediktera priser på begagnade elbilar?
4. Vilka variabler är signifikanta och vilka är mindre viktiga vid prissättningen?

2 Teori

2.1 Modell

Multilinjär modell används för att prediktera sambandet mellan en beroende variabel och flera oberoende variabler.

2.2 Dataset

Datan var insamlad från Blocket.se i en excell tabell.

Datan består av variablerna:

Miltal, Modellår, Biltyp, Drivning, Hästkrafter, Färg, Märke, Modell, Län, Pris

3 Metod

3.1 Datainsamling

Datan var samlad in från Blocket.se i en excell fil (2404152328_Cars) vid ett grupparbete.

Den var laddad i R-studio och rensad med hjälp av kod och sparad i en ny fil (carclean.csv).

	X	Miltal	Modellår	Biltyp	Drivning	Hästkrafter	Färg	Märke	Modell	Län	Pris
1	1	1491	2022	halvkombi	2	204	grå	hyundai	kona	kronoberg	349800
2	2	9800	2019	sedan	2	120	svart	hyundai	ioniq	västragötaland	144000
3	3	6729	2020	suv	4	408	svart	audi	etron	skåne	528900
4	4	1270	2022	suv	2	136	röd	hyundai	kona	uppsala	259900
5	5	11748	2020	halvkombi	2	136	blå	hyundai	ioniq	örebro	209800
6	6	4355	2020	suv	4	408	svart	audi	etron	skåne	548900
7	7	6400	2022	suv	2	204	blå	hyundai	kona	örebro	319800
8	8	7400	2020	suv	2	204	grå	hyundai	kona	örebro	274800
9	9	4998	2020	suv	2	204	grå	hyundai	kona	örebro	289800
10	10	3525	2020	suv	2	204	röd	hyundai	kona	örebro	289800

Tabell 1: Uppladdad data.

3.2 Preprocessing

För att atbeta med datan och skapa modeller behöver den förberedas. Den första kolumnen var raderad eftersom det fanns bara ordningstal som störde. Variabler av klass ”character” (Biltyp, Färt, Modell, Märke) konvertarades till faktor så att modellen hanterar inte varje värde separat. Variabeln ”län” var raderad pga det visade sig framöver att den inte var signifikant.

	Miltal	Modellår	Biltyp	Drivning	Hästkrafter	Färg	Märke	Modell	Pris
1	1491	2022	halvkombi	2	204	grå	hyundai	kona	349800
2	9800	2019	sedan	2	120	svart	hyundai	ioniq	144000
3	6729	2020	suv	4	408	svart	audi	etron	528900
4	1270	2022	suv	2	136	röd	hyundai	kona	259900
5	11748	2020	halvkombi	2	136	blå	hyundai	ioniq	209800
6	4355	2020	suv	4	408	svart	audi	etron	548900
7	6400	2022	suv	2	204	blå	hyundai	kona	319800
8	7400	2020	suv	2	204	grå	hyundai	kona	274800
9	4998	2020	suv	2	204	grå	hyundai	kona	289800
10	3525	2020	suv	2	204	röd	hyundai	kona	289800

Tabell 2: Data efter preprocessing.

3.3 Splittring datan till ”train” och ”test”

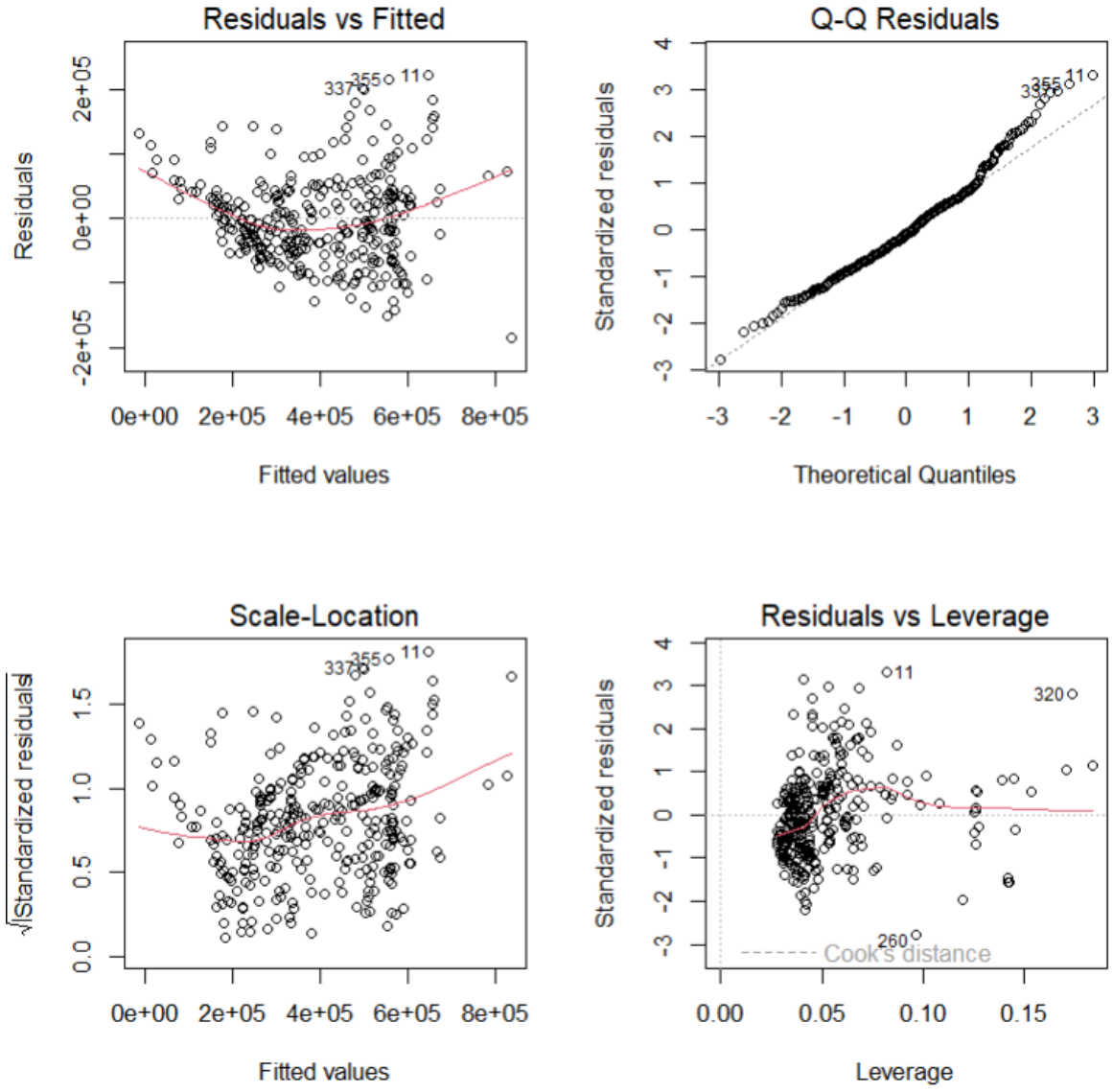
Dataset var fördelat till tränings- och testdata, respektive 80% och 20%.

3.4 Modeller

Tre multilinjära modeller skapades:

1. Modell 1: variablerna valdes teoretiskt enligt logik.

```
Residual standard error: 70030 on 310 degrees of freedom
Multiple R-squared:  0.8394,    Adjusted R-squared:  0.8311
F-statistic: 101.2 on 16 and 310 DF,  p-value: < 2.2e-16
```



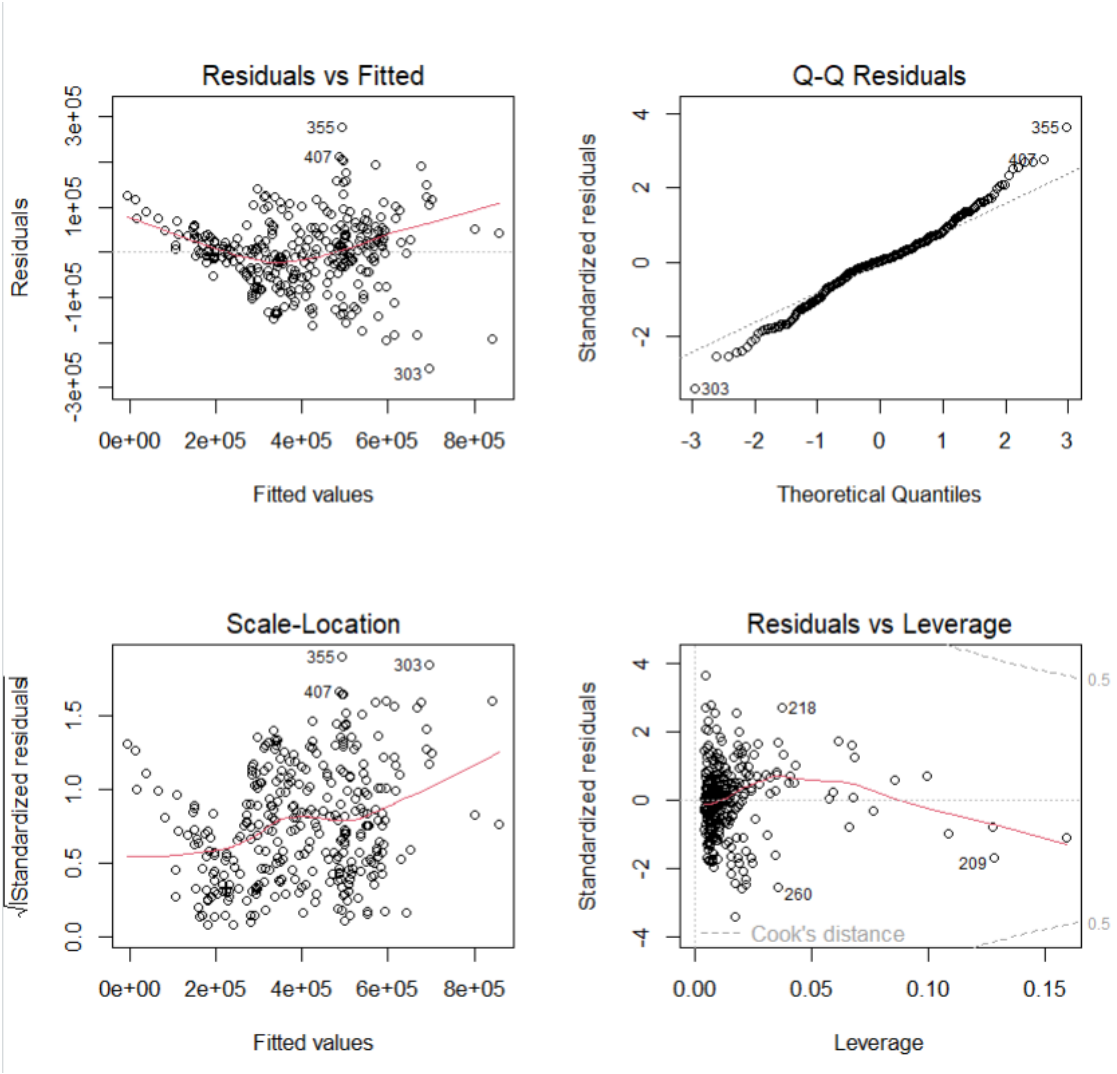
Figur 1: Sammanfattning av resultaten för Model 1.

2. Modell 2: alla variablerna användes. Det visade sig att ”län” var oviktig, därför blev det raderad. Signifikanta variablerna blev Miltal, Modellår, Hästkrafter och kombinerade effekten av Modelår och Miltal på Pris.

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -1.075e+08  9.001e+06 -11.945  < 2e-16 ***
Miltal           6.787e+03  1.086e+03   6.250 1.30e-09 ***
Modellår         5.329e+04  4.453e+03  11.967  < 2e-16 ***
Hästkrafter      9.568e+02  3.866e+01  24.749  < 2e-16 ***
Miltal:Modellår -3.364e+00  5.376e-01  -6.257 1.25e-09 ***
```

Tabell 4: Illustrering av signifikanta variabler

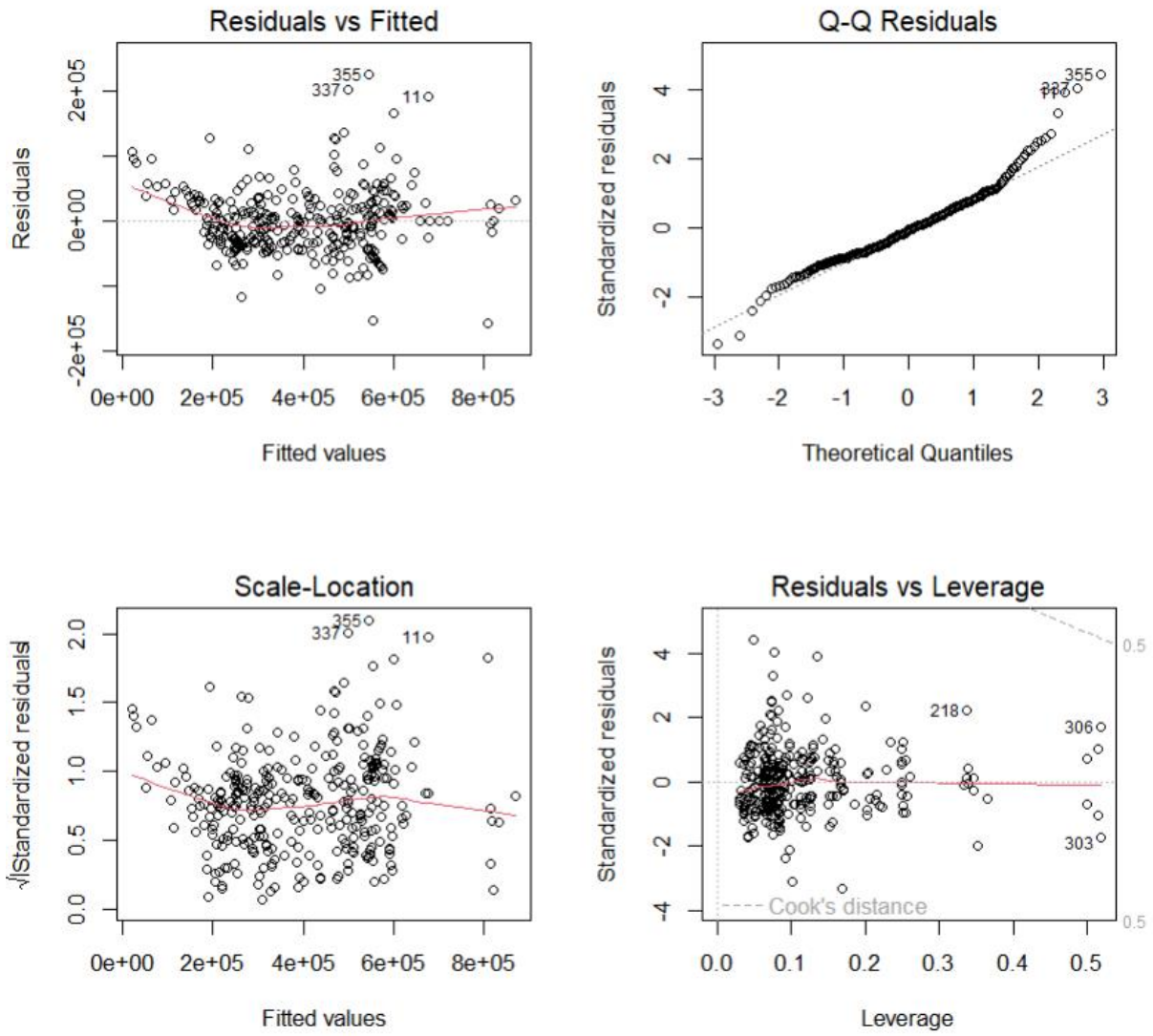
```
Residual standard error: 76370 on 322 degrees of freedom
Multiple R-squared:  0.8016,    Adjusted R-squared:  0.7991
F-statistic: 325.2 on 4 and 322 DF,  p-value: < 2.2e-16
```



Figur 2: Sammanfattning av resultaten för Model 2.

3. Modell 3: används en automatisk funktion ”backward stepwise selection” som tar bort stegvist variablerna som inte förbättrar modellen.

```
Residual standard error: 52140 on 284 degrees of freedom
Multiple R-squared:  0.9184,    Adjusted R-squared:  0.9063
F-statistic: 76.12 on 42 and 284 DF,  p-value: < 2.2e-16
```



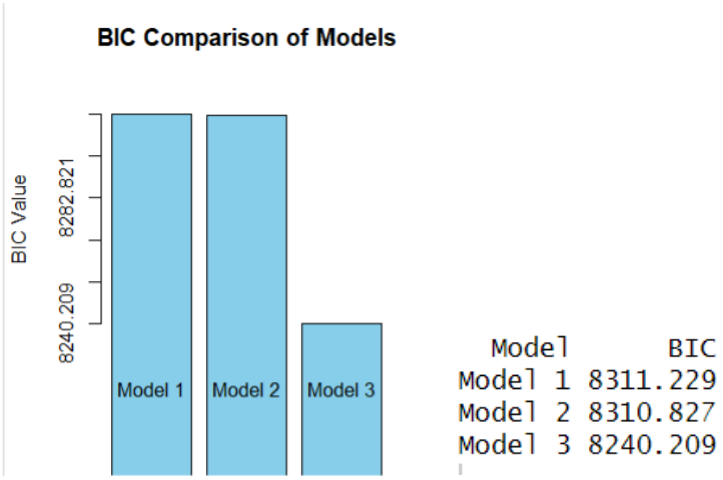
Figur 3: Sammanfattning av resultaten för Model 3.

4 Resultat och Diskussion

Tre olika modeller skapades. Vid utvärdering visade det sig att Modell 3 är bäst att prediktera pris, enligt coefficenter:

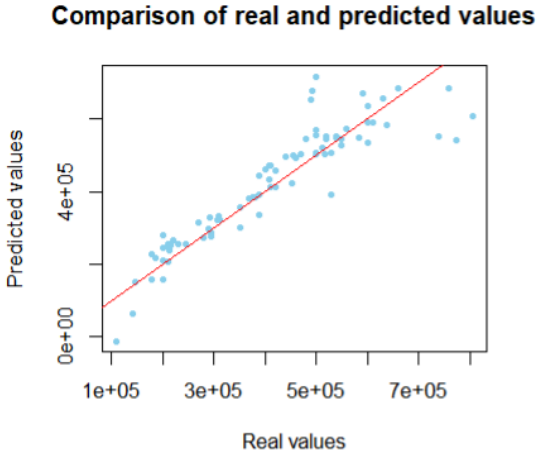
RSE för olika modeller	
Modell 1	70030
Modell 2	76370
Modell 3	52140

Tabell 3: Root Mean Squared Error (RMSE) för de tre valda modellerna.



Figur 4: BIC för för de tre valda modellerna.

Tre olika modeller skapades. Vid utvärdering visade det sig att Modell 3 är bäst för att prediktera priset på elbilar.



Figur 5: Jämförelse av verklig och predikterad data.

5 Slutsatser

Syftet med rapporten var uppfylld och frågorna besvarade:

1. Går det att samla in relevant data själv från existerande platformer? (Blocket.se)
   Data var successivt insamlad från Blocket.se.

2. Går det att hitta linjära samband mellan olika karakteristiker av bilar och pris?
   Det gick att hitta linjära samband mellan olika karakteristiker och pris. Flera karakteristiker dock visade sig som ej signifikanta för prediktion.

3. Går det att skapa en fungerande modell som ska hjälpa att prediktera priser på begagnade elbilar?
   Modell som skulle hjälpa att prediktera priser på begagnade elbilar var skapad och visade en bra prestanda. Den måste dock finjusteras pga outliers och avvikelser.

4. Vilka variabler är signifikanta och vilka är mindre viktiga vid prissättningen?
   Variablerna som visade sig signifikanta för prissättningen är Miltal, Modellår, Biltyp, Hästkrafter, Modell.

6 Teoretiska frågor

1. Kolla på följande video: https://www.youtube.com/watch?v=X9_ISJ0YpGw&t=290s , beskriv kortfattat vad en Quantile-Quantile (QQ) plot är.

En QQ plot är en plot som visualiserar jämförelse mellan faktiska och förväntade fördelningen (normalfördelning). Om prickorna ligger nära den räta linjen då betyder det att datan följer normalfördelningen. Om det är tvärtom då betyder det att datan inte är normalfördelad.

2. Din kollega Karin frågar dig följande: ”Jag har hört att i Maskininlärning så är fokus på prediktioner medan man i statistisk regressionsanalys kan göra såväl prediktioner som statistisk inferens. Vad menas med det, kan du ge några exempel?” Vad svarar du Karin?

I maskininlärning söker man mönster och gör prediktioner på det. I en statistisk regressionsanalys kan man gå på detaljer och dra slutsatser på hur olika faktorer och deras sambandet kan påverka varandra.

Exempel: vid maskininlärning kan man prediktera priser på bilar. Vid statistisk regressionsanalys kan man dra slutsatser på vilka karakteristiker av bilen kan påverka pris.

3. Vad är skillnaden på ”konfidensintervall” och ”prediktionsintervall” för predikterade värden?

Konfidensinterval var den sanna värde ligger (standard med säkerhet 95%), medan prediktionsintervall visar var framtida observationer kan ligga.

4. Den multipla linjära regressionsmodellen kan skrivas som:

$Y = \beta 0 + \beta 1x1 + \beta 1x2 + \ldots + \beta pxp + \varepsilon$. **Hur tolkas beta parametrarna?**

Beta variablerna är koefficenter. $\beta 0$ exempelvis intercept, $\beta 1$ är lutningen av den räta linjen.

5. Din kollega Hassan frågar dig följande: ”Stämmer det att man i statistisk regressionsmodellering inte behöver använda träning, validering och test set om man nyttjar mått såsom BIC? Vad är logiken bakom detta?” Vad svarar du Hassan?

Det stämmer. BIC tar hänsyn till både modellens passning och komplexitet. Därför blir det möjligt att jämvöra och välja den bästa modellen utan att behöva separera datan till train och test. Det är dock bättre att fördela datan ändå som jag läst.

6. Förklara algoritmen nedan för ”Best subset selection”

Best subset selection väljer den bästa kombinationen av variabler för en regressionsmidell. Den testar olika kombinationer och väljer den som har lägst residual sum of squares (RSS). Den passar inte till för stor mängd av variabler eftersom det blir för tungt att pröva miljoner kombinationer.

7. Ett citat från statistikern George Box är: “All models are wrong, some are useful.”

Förklara vad som menas med det citatet.

Det menas att man ska inte lita 100% på modellen. Den kan hjälpa, men aldrig blir 100% perfekt.

Appendix A

Källkoder för R-Studio:

https://github.com/xkotyplasticx/R

Konfidensintervaller för den valda Modell 3 (på testdata):

	fit	lwr	upr
1	356648.36	320891.27	392405.46
3	507992.19	487098.35	528886.03
6	526786.11	505067.12	548505.10
15	410813.27	376654.35	444972.19
24	442880.74	402787.09	482974.40
29	216777.72	190297.74	243257.71
39	548044.96	524041.25	572048.68
43	279913.89	254096.28	305731.50
45	244216.53	216673.07	271759.98
47	274047.08	249376.89	298717.26
50	267427.25	246774.51	288079.99
51	-11180.61	-49457.19	27095.97
63	255742.40	236258.63	275226.18
74	226129.06	207606.96	244651.16
76	156761.25	137260.47	176262.03
81	252409.53	233082.88	271736.19
83	211604.68	190402.39	232806.98
104	316273.85	289518.82	343028.87
109	589245.96	563476.99	615014.92
112	590037.61	564264.54	615810.69
121	497563.56	470483.49	524643.63
137	533987.21	502712.18	565262.25
147	151100.01	108783.31	193416.71
150	500863.96	473646.85	528081.07
152	301440.18	255390.48	347489.88
153	385359.16	359258.95	411459.38
159	379896.73	353644.98	406148.49
160	238565.29	195907.74	281222.84
168	156821.38	109310.61	204332.14
178	456341.95	429308.07	483375.83
180	321231.59	275083.12	367380.06
190	543685.80	481715.05	605656.56
195	493739.74	466864.89	520614.58
196	470172.32	441520.49	498824.15
200	470230.06	443886.05	496574.07
202	717402.87	662600.09	772205.66
207	413208.83	384633.62	441784.04
208	335257.14	302832.91	367681.37
226	503035.64	475544.22	530527.06
229	654690.44	597215.93	712164.96
230	675970.17	618306.78	733633.55
232	384482.14	354480.23	414484.06
242	424603.82	393082.34	456125.29
243	521096.38	490513.36	551679.40
245	543972.73	513351.13	574594.33
251	504210.35	473167.28	535253.42
255	391126.03	361267.80	420984.26
262	65263.24	16739.46	113787.01
264	390715.90	359071.54	422360.26
272	582286.38	551098.98	613473.78
280	333156.30	296831.70	369480.90
284	247010.68	212556.48	281464.88
293	285941.45	250966.50	320916.40
295	278199.05	242686.92	313711.18
298	297190.89	262559.60	331822.17
300	321099.37	284563.67	357635.08
311	434232.69	374222.57	494242.81

```
318 683521.97 580367.30 786676.64
321 556916.89 533601.25 580232.52
328 505013.06 476812.71 533213.41
330 502875.59 474644.52 531106.66
334 460559.18 431417.11 489701.26
336 546505.40 517324.02 575686.78
339 552886.15 524009.05 581763.25
342 554304.42 531157.00 577451.84
344 548091.13 525117.04 571065.22
351 506358.88 478165.78 534551.97
353 570199.18 540492.19 599906.17
354 605987.94 577535.71 634440.16
358 548510.71 525521.09 571500.33
359 550504.47 527526.63 573482.31
360 552483.61 529428.66 575538.56
363 541400.42 518458.36 564342.48
381 255639.86 229395.71 281884.02
383 329733.88 299063.15 360404.62
387 637229.73 534426.58 740032.88
389 671714.55 568954.72 774474.38
390 684696.94 611996.18 757397.70
391 571631.04 498052.35 645209.74
392 254104.42 227879.77 280329.08
393 206229.73 176627.74 235831.73
401 655231.49 582328.35 728134.63
```

Prediktionsintervaller för den valda Modell 3 (på testdata):

```
          fit        lwr      upr
1   356648.36  247964.31 465332.4
3   507992.19  403253.44 612730.9
6   526786.11  421879.63 631692.6
15  410813.27  302644.49 518982.1
24  442880.74  332693.83 553067.7
29  216777.72  110783.18 322772.3
39  548044.96  442641.77 653448.2
43  279913.89  174082.88 385744.9
45  244216.53  137951.31 350481.7
47  274047.08  168490.12 379604.0
50  267427.25  162736.33 372118.2
51  -11180.61 -120719.42  98358.2
63  255742.40  151275.80 360209.0
74  226129.06  121837.53 330420.6
76  156761.25   52291.47 261231.0
81  252409.53  147972.12 356846.9
83  211604.68  106803.96 316405.4
104 316273.85  210210.26 422337.4
109 589245.96  483426.80 695065.1
112 590037.61  484217.46 695857.8
121 497563.56  391417.51 603709.6
137 533987.21  426694.24 641280.2
147 151100.01   40084.89 262115.1
150 500863.96  394682.86 607045.1
152 301440.18  188949.17 413931.2
153 385359.16  279458.85 491259.5
159 379896.73  273958.98 485834.5
160 238565.29  127419.79 349710.8
168 156821.38   43724.39 269918.4
178 456341.95  350207.67 562476.2
180 321231.59  208700.10 433763.1
190 543685.80  423794.07 663577.5
195 493739.74  387645.86 599833.6
196 470172.32  363614.44 576730.2
200 470230.06  364269.40 576190.7
202 717402.87  601054.26 833751.5
207 413208.83  306671.53 519746.1
208 335257.14  227623.58 442890.7
```

226	503035.64	396783.90	609287.4
229	654690.44	537059.78	772321.1
230	675970.17	558247.10	793693.2
232	384482.14	277553.34	491410.9
242	424603.82	317238.75	531968.9
243	521096.38	414003.08	628189.7
245	543972.73	436868.41	651077.1
251	504210.35	396984.77	611435.9
255	391126.03	284237.45	498014.6
262	65263.24	-48263.02	178789.5
264	390715.90	283314.69	498117.1
272	582286.38	475018.92	689553.8
280	333156.30	224284.22	442028.4
284	247010.68	138748.29	355273.1
293	285941.45	177512.20	394370.7
295	278199.05	169595.34	386802.8
298	297190.89	188872.00	405509.8
300	321099.37	212156.67	430042.1
311	434232.69	315342.53	553122.8
318	683521.97	538007.23	829036.7
321	556916.89	451668.26	662165.5
328	505013.06	398575.69	611450.4
330	502875.59	396430.08	609321.1
334	460559.18	353868.44	567249.9
336	546505.40	439803.91	653206.9
339	552886.15	446267.48	659504.8
342	554304.42	449092.93	659515.9
344	548091.13	442917.64	653264.6
351	506358.88	399923.43	612794.3
353	570199.18	463352.75	677045.6
354	605987.94	499483.56	712492.3
358	548510.71	443333.82	653687.6
359	550504.47	445330.16	655678.8
360	552483.61	447292.43	657674.8
363	541400.42	436233.93	646566.9
381	255639.86	149703.99	361575.7
383	329733.88	222615.50	436852.3
387	637229.73	491963.97	782495.5
389	671714.55	526479.44	816949.7
390	684696.94	558923.11	810470.8
391	571631.04	445347.72	697914.4
392	254104.42	148173.38	360035.5
393	206229.73	99412.45	313047.0
401	655231.49	529340.57	781122.4

Källförteckning

Bruno Scalia C. F. Leite (2022). Hämtat från https://towardsdatascience.com/multiple-linear-regression-theory-and-applications-677ec2cd04ac

Elena Kosourova (2022). Hämtat från https://www.dataquest.io/blog/apply-functions-in-r-sapply-lapply-tapply/

Information om bilar för datainsamling. Hämtat från https://www.blocket.se/

The R Manuals. Hämtat från https://cran.r-project.org/manuals.html

Whuber (2017). Hämtat från https://stats.stackexchange.com/questions/11551/is-there-a-good-browser-viewer-to-see-an-r-dataset-rda-file