CS112 – Regression and Bootstrapping

Armin Hamp

18 October 2018

Link to code

https://gist.github.com/hamparmin/9f97b2e0327e54574719d49fa5e52f44

Question 1

(a)

```
#creating sample n=99
x <- rnorm(99,50,12)
y <- 4*x+3+rnorm(x,50,14)
set1 <- data.frame(x,y)
```

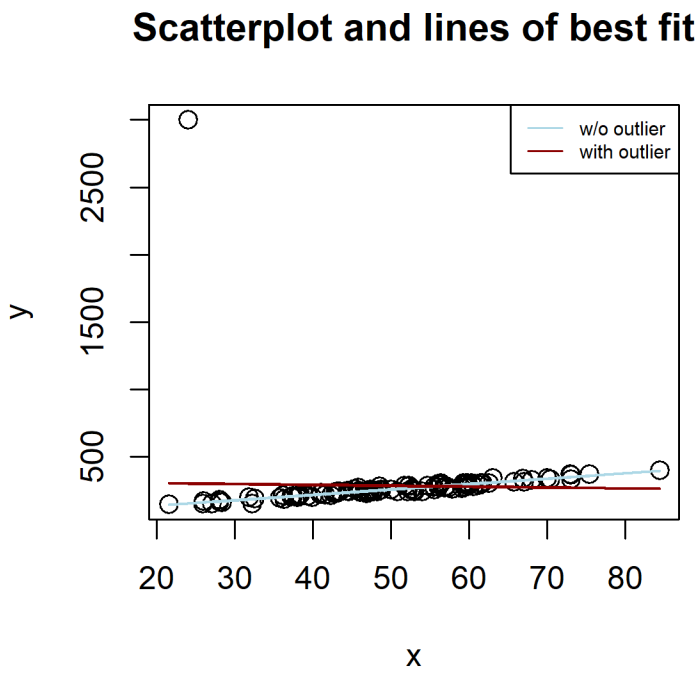
(b) Model without outlier

```
Call:
lm(formula = y ~ x, data = set1)

Residuals:
    Min      1Q  Median      3Q     Max
-34.209  -7.896   0.081  10.080  34.944

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  53.9682     5.9453   9.077 1.32e-14 ***
x             4.0305     0.1151  35.010  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.12 on 97 degrees of freedom
Multiple R-squared:  0.9267,    Adjusted R-squared:  0.9259
F-statistic:  1226 on 1 and 97 DF,  p-value: < 2.2e-16
```

(c) Model with outlier

```
Call:
lm(formula = y ~ x, data = set2)

Residuals:
    Min      1Q  Median      3Q     Max
-158.26  -61.62  -28.72   10.98 2699.27

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 316.6669   115.1051   2.751  0.00708 **
x            -0.6641     2.2377  -0.297  0.76728
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 280.5 on 98 degrees of freedom
Multiple R-squared:  0.0008978, Adjusted R-squared:  -0.009297
F-statistic: 0.08807 on 1 and 98 DF,  p-value: 0.7673
```

(d)



(e)

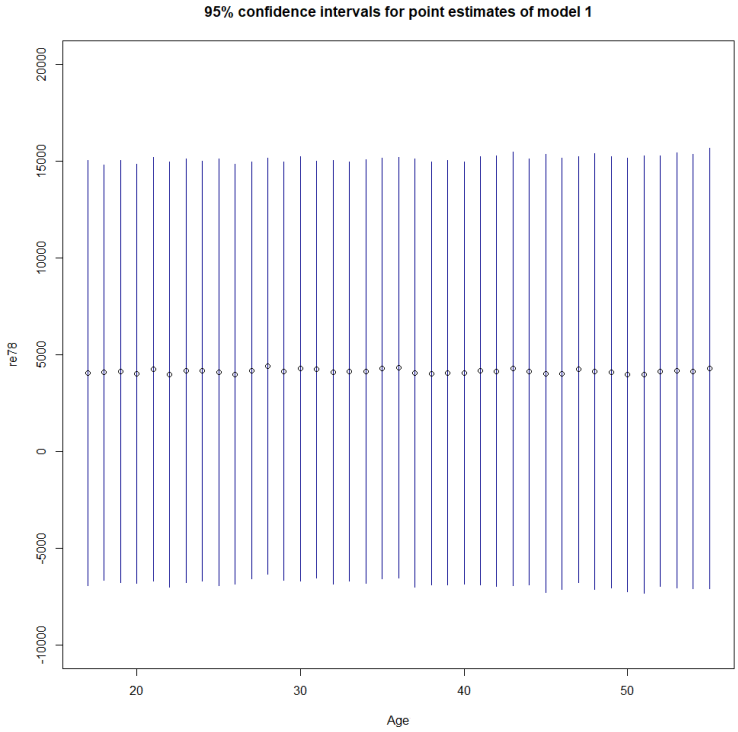
Outliers are observations with an unusual y_i value, that can have a significant effect on our regression. This is especially true if the y-value for our outlier is different enough to influence the least squares fitting of the regression. As we can see above, a single outlier (x=24, y=3000) was able to change our otherwise precise regression coefficient from positive to negative.

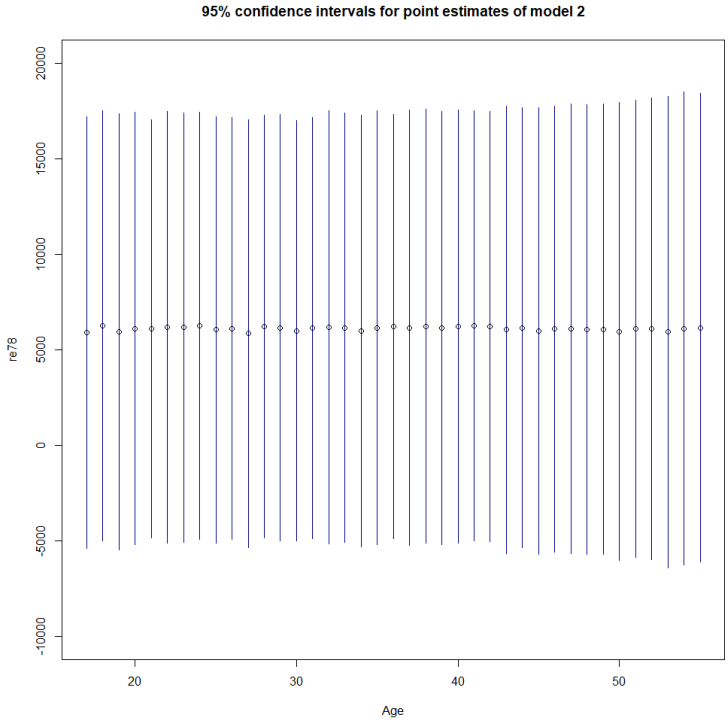
Question 2

(a) See table on next page:

MODEL 1			MODEL 2		
Age	***2.50%***	***97.50%***	***Age***	***2.50%***	***97.50%***
17	-6928	15020	*17*	-5255	17357
18	-6637	14798	*18*	-5343	17347
19	-6785	15049	*19*	-4956	17509
20	-6795	14829	*20*	-5156	17040
21	-6708	15171	*21*	-4998	17144
22	-7012	14941	*22*	-4841	17365
23	-6757	15092	*23*	-5368	17211
24	-6695	15001	*24*	-5194	17319
25	-6943	15108	*25*	-5166	17123
26	-6860	14832	*26*	-4966	17029
27	-6583	14952	*27*	-5042	17065
28	-6325	15129	*28*	-5069	17117
29	-6656	14954	*29*	-5014	17275
30	-6673	15245	*30*	-5120	16972
31	-6519	14992	*31*	-5187	17393
32	-6849	15025	*32*	-5254	17245
33	-6687	14959	*33*	-4716	17392
34	-6818	15067	*34*	-4888	17417
35	-6566	15161	*35*	-4959	17214
36	-6524	15200	*36*	-4935	17467
37	-6993	15125	*37*	-5262	17453
38	-6900	14940	*38*	-5297	17683
39	-6893	15014	*39*	-5221	17240
40	-6861	14940	*40*	-5171	17220
41	-6897	15224	*41*	-5205	17264
42	-6981	15274	*42*	-5366	17496
43	-6915	15481	*43*	-5573	17487
44	-6880	15112	*44*	-5703	17751
45	-7287	15341	*45*	-5579	17697
46	-7134	15149	*46*	-5749	17725
47	-6759	15244	*47*	-5872	18041
48	-7111	15395	*48*	-6024	17988
49	-7038	15225	*49*	-6146	18092
50	-7250	15159	*50*	-6004	18099
51	-7318	15248	*51*	-6294	18205
52	-6985	15258	*52*	-6023	18082
53	-7065	15426	*53*	-6647	18499
54	-7069	15358	*54*	-5956	18247
55	-7098	15647	*55*	-6543	18653

(b)


95% confidence intervals for point estimates of model 1
re78
20000
15000
10000
5000
0
-5000
-10000
20
30
40
50
Age

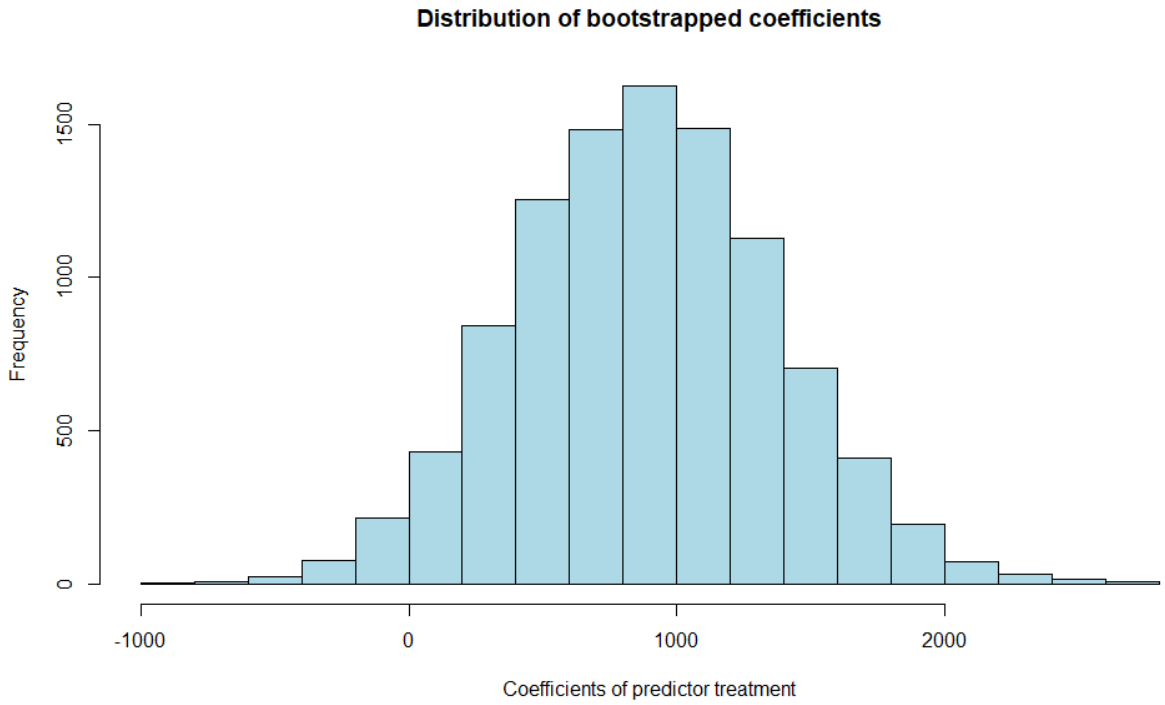

95% confidence intervals for point estimates of model 2
re78
20000
15000
10000
5000
0
-5000
-10000
20
30
40
50
Age

Question 3

(a)

```
                                  2.5%    97.5%
confint with bootstrapping   -46.86176 1854.759
confint based on single coef -40.52635 1813.134
```

(b)



(c)

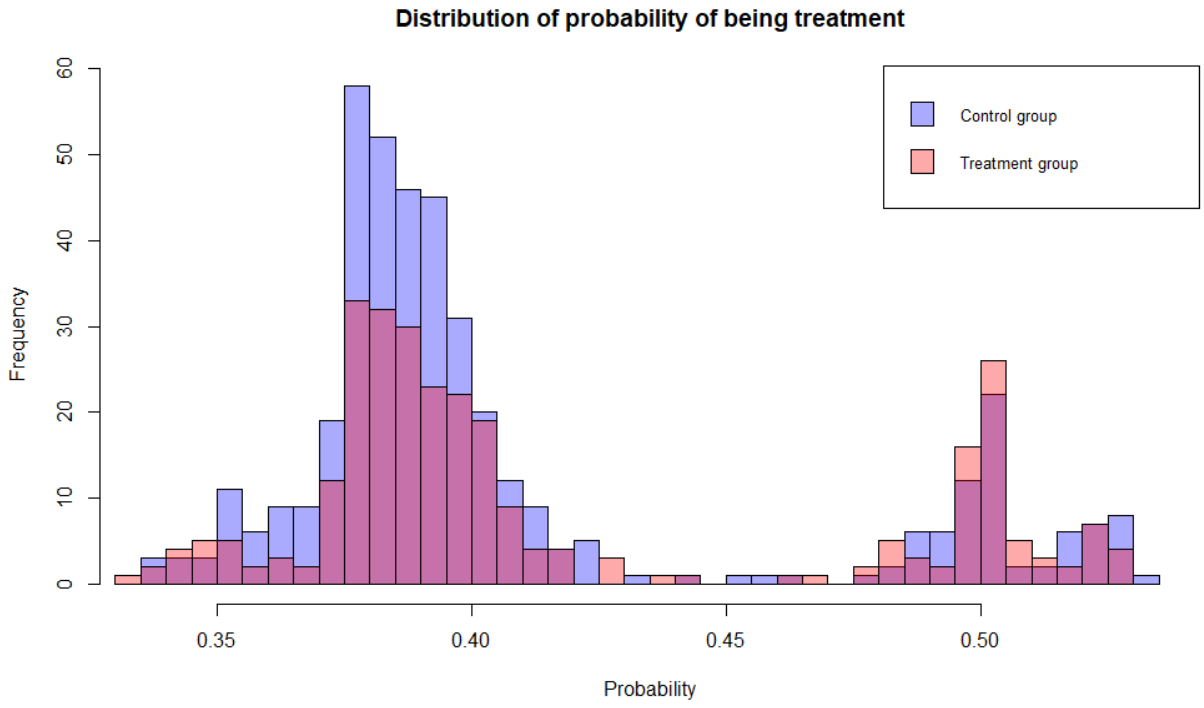
What is interesting is that the bootstrapped confidence interval for the coefficients is about the same (a little larger) than the analytically obtained interval, using the *confint*() function. This is surprising, as during previous exercises we saw that there are cases where we can obtain a more precise confidence interval with bootstrapping. This result is probably due to *treatment* being a poor predictor of *re78*.

Question 4

```
#calculating r^2 by the equation R2=1-rss/tss
rsq <- function(ypredicted, yactual) {
  tss <- sum((yactual - mean(yactual))^2)
  rss <- sum((yactual - ypredicted)^2)
  return(1-(rss/tss))
}

#testing rsq on nsw dataset
rsq(predict(fit_nsw),nsw$re78)

Output
[1] 0.004871571

> summary(fit_nsw)
...
Multiple R-squared:  0.004872
```

Question 5

(a)



(b)

The histogram shows that most observations from the control set received low (between 0.35 and 0.42) probabilities, which seems intuitive at first, if our model was trained for treatment. But, we can also see that most of the observations from the treatment group were also assigned similar probabilities, which tells us that our predictors are not great for establishing a clear threshold between control and treated. However, this is not necessarily bad, as it indicates that the assignment of treatment was not dependent on any other variables measured, which is good practice for experiments with the intent of establishing causal relationships.