Regression and Bootstrapping

Prof Diamond

Isabella Buchanan

CS112 - Fall 2018

Code

All code can be found here.

Question 1

a) *The data generating equation is:*

$$happiness = 2 + 0.7 * puppy\_cuteness + rnorm(99,\ 3,\ 1)$$

b) *The regression results for the original 99 are:*

```
Call:
lm(formula = happiness ~ puppy_cuteness, data = data99)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.9640 -0.6981 -0.1159  0.5607  3.3650 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)     4.85771    0.24725   19.65   <2e-16 ***
puppy_cuteness  0.70789    0.04321   16.38   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9737 on 97 degrees of freedom
Multiple R-squared:  0.7345,	Adjusted R-squared:  0.7318 
F-statistic: 268.3 on 1 and 97 DF,  p-value: < 2.2e-16
```

c) *The regression results with the outlier are:*

```
Call:
lm(formula = happiness ~ puppy_cuteness, data = data100)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.8172 -1.4137  0.2347  1.2809  5.9588 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)     8.65134    0.28813  30.026   <2e-16 ***
puppy_cuteness -0.02724    0.03484  -0.782    0.436    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.054 on 98 degrees of freedom
Multiple R-squared:  0.006198,	Adjusted R-squared:  -0.003943 
F-statistic: 0.6112 on 1 and 98 DF,  p-value: 0.4362
```

d) Data visualisation:

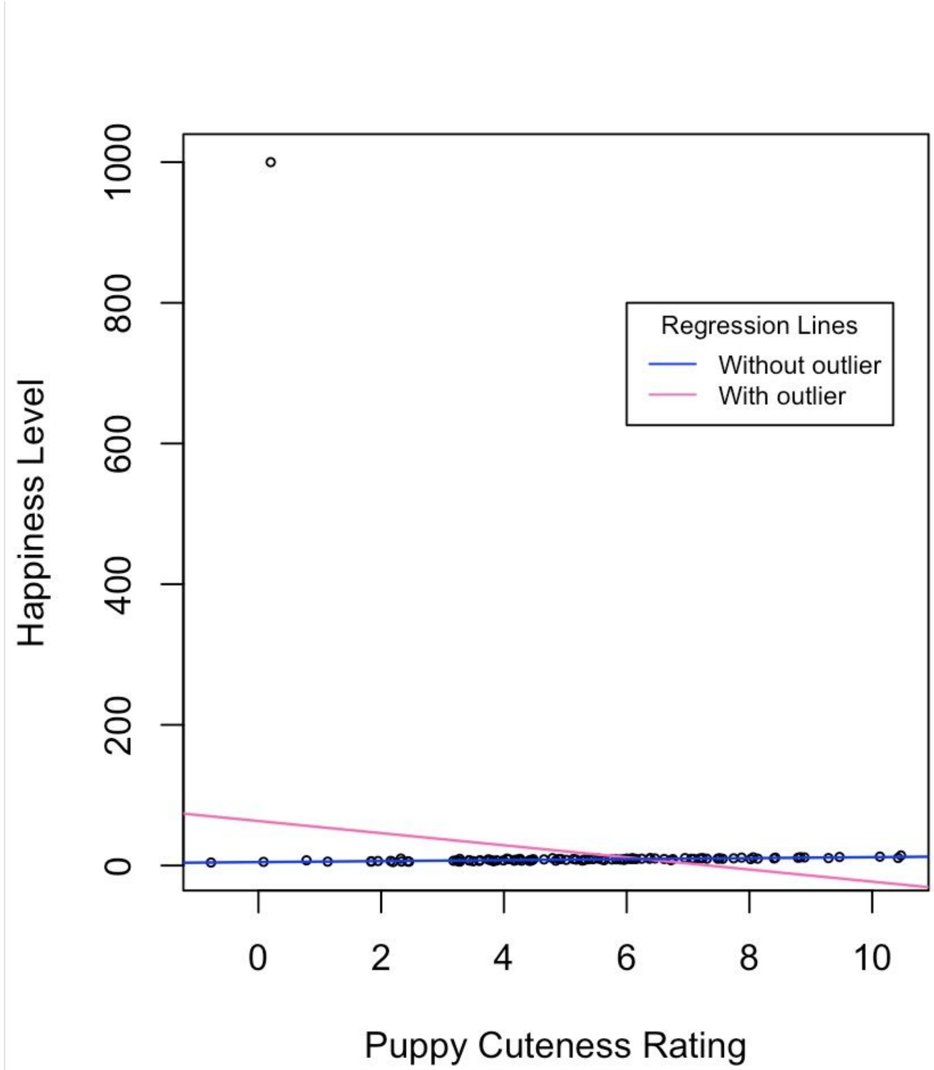

Figure 1. Visualisation demonstrating single variable regression predicting happiness level based on puppy cuteness rating on a dataset of 99 points (blue) and a dataset of 100 points (pink) including an outlier

e) 3 sentence caption:

Figure 1. The positively sloped blue line is fit on the data excluding the outlier, whereas the negatively sloped pink line is fit on the data including the outlier, an observation with an

unusual y value[1], which largely influences the model. This demonstrates the dangers of extrapolation using a model which does not fit the data well since using the pink model we would assume cuter puppies are correlated with lower happiness, which does not make logical sense, or fit the majority of the data.

[1] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: With applications in R*. New York: Springer. Retrieved November 7, 2016 from: Retrieved from http://www-bcf.usc.edu/~gareth/ISL/ISLR%20First%20Printing.pdf

Question 2

a)

Table 1. The 95% confidence intervals for varying ages with educ, re74, and re75 held at their medians and then 90% quantiles

Held at medians (educ = 10, re74 =0, re75 = 0)

Age	2.50%	97.50%
17	-6723.0492	15214.3593
18	-6664.253	15002.0859
19	-6635.2637	14884.2923
20	-6693.5304	14941.994
21	-7015.0774	14998.8827
22	-6774.7555	14795.3219
23	-6740.276	15046.1321
24	-7039.9606	15164.2034
25	-6611.3245	15056.9019
26	-6967.0963	14836.7661
27	-6714.2176	15135.2512
28	-6823.4306	15066.6595
29	-7006.4283	14739.7258
30	-6796.4602	15018.915
31	-7004.9264	15171.1567
32	-6692.3296	15257.3039
33	-6686.1506	14993.158
34	-6943.1415	15180.5011
35	-6787.8836	15047.3844
36	-6801.5579	15124.4176
37	-6672.8928	15137.3371
38	-6665.2064	15228.0586
39	-6920.2344	15454.8362
40	-6917.0704	15284.0151
41	-6840.7407	15378.5346
42	-6720.7531	15332.967
43	-6670.8133	15210.915
44	-6973.8173	15255.4792
45	-6894.531	15168.7507
46	-6935.2013	15325.2462
47	-6894.2467	15327.2245
48	-7158.596	15312.4742
49	-7035.6727	15270.1734
50	-7105.3437	15461.1622
51	-7099.879	15599.2524
52	-6767.6245	15245.2898
53	-6732.1175	15328.135
54	-7055.1565	15520.6398
55	-6899.1053	15617.5043

Held at 90% quantiles (educ = 12, re74 = 7628, re75 = 4493)

Age	2.50%	97.50%
17	-5341.255	17439.8656
18	-5348.4098	17459.5759
19	-5210.2126	17029.2348
20	-5199.5925	17255.0714
21	-5362.6383	17112.4014
22	-5029.2843	17406.3143
23	-5089.1027	17235.1364
24	-5008.2286	17013.776
25	-5142.0776	16930.6421
26	-4994.6473	17212.404
27	-5011.1668	17191.0095
28	-5263.0158	16923.1116
29	-5283.0741	17165.3086
30	-4931.3597	17071.0936
31	-4981.6996	17256.5888
32	-5221.728	17316.1904
33	-4888.6177	17232.1683
34	-4910.7618	17292.2585
35	-5419.1014	16979.3681
36	-5274.5539	17282.148
37	-5115.5065	17319.218
38	-5179.0735	17311.3107
39	-5200.9847	17218.7293
40	-5310.2984	17247.8926
41	-5681.1296	17301.1848
42	-5293.3466	17359.0143
43	-5662.8454	17467.257
44	-5529.9285	17585.151
45	-5608.1696	17383.3322
46	-5683.8342	17544.7498
47	-5939.4408	17665.4599
48	-5802.2686	18080.8675
49	-5914.3719	18164.7219
50	-5730.1032	17946.4703
51	-5694.8009	18302.4685
52	-6304.9698	18373.9399
53	-6547.6278	18098.3636
54	-6301.8664	18410.2803
55	-6208.6126	18666.3134

b)

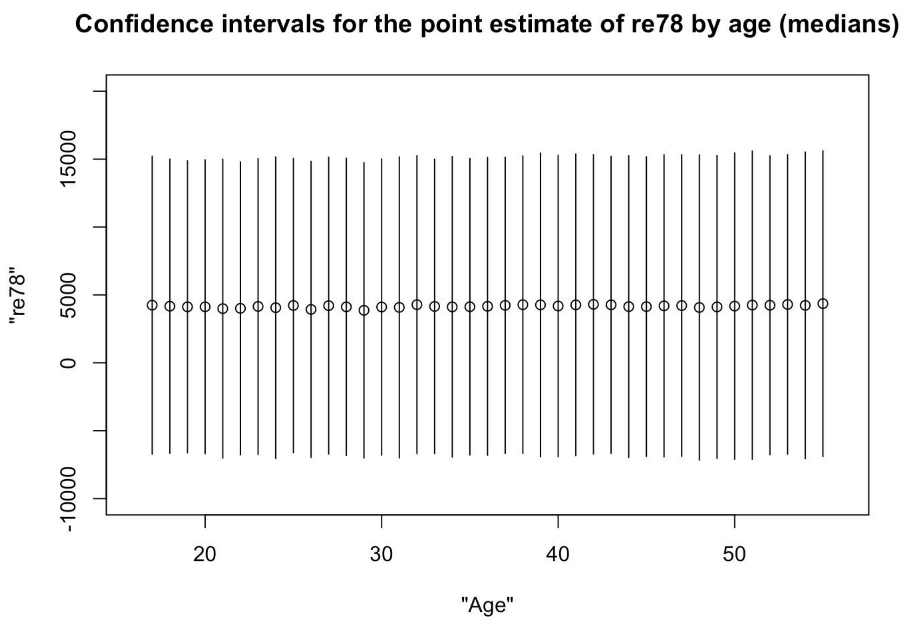

Figure 2. Scatter plot showing the 95% confidence intervals for estimates of re78 with educ, re74, and re75 held at their medians. The dots represent the mean of the confidence intervals estimated.

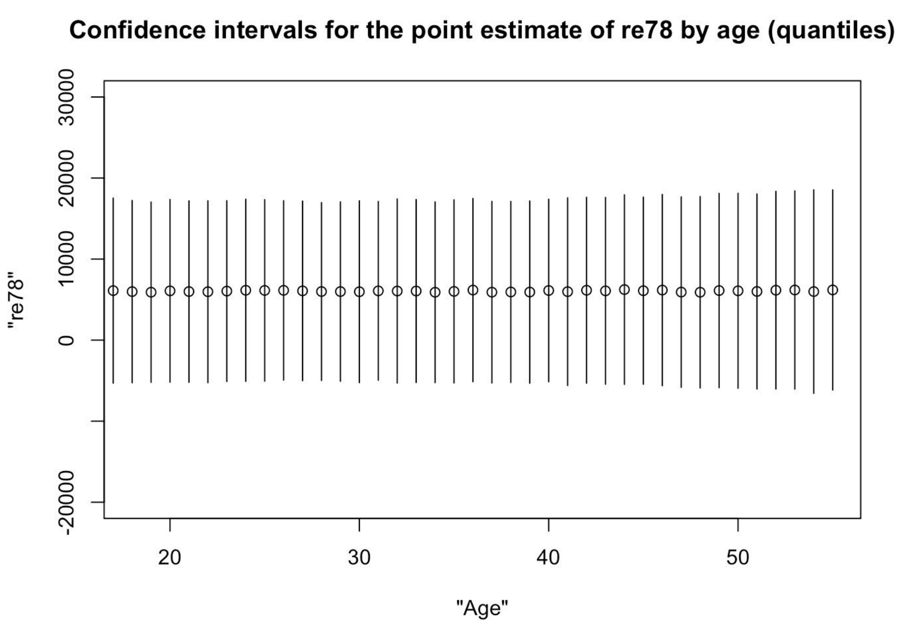

Figure 3. Scatter plot showing the 95% confidence intervals for estimates of re78 with educ, re74, and re75 held at their 90% quantiles. The dots represent the mean of the confidence intervals estimated.

Question 3

a)

Table 2. 95% confidence intervals for the coefficient of treatment using both bootstrapped and analytical methods

	2.5 %	97.5 %
Treatment for analytical	-40.52635	1813.134
Treatment for bootstrapped	-46.86176	1854.759

b)

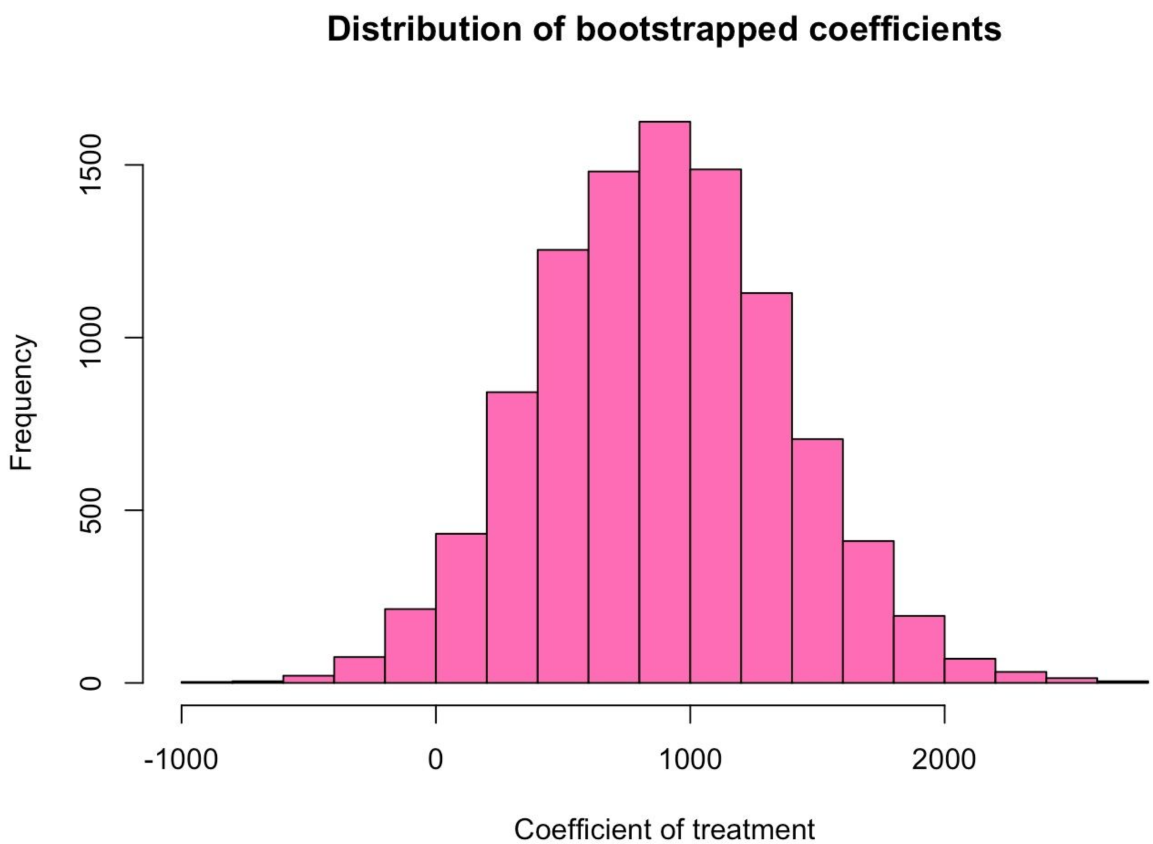

Figure 4. Distribution of the bootstrapped coefficients of treatment.

c)

The bootstrapped confidence intervals and analytical confidence intervals produce similar results. What is interesting is that the analytical confidence interval is smaller, and more precise than the bootstrapped confidence interval suggesting that for a linear regression, bootstrapping is not always necessary as R has powerful statistical software to compute error terms such as standard error.[2] The bootstrapped coefficients also produce a normal distribution due to resampling many times.

[2] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: With applications in R*. New York: Springer. Retrieved November 7, 2016 from: Retrieved from http://www-bcf.usc.edu/~gareth/ISL/ISLR%20First%20Printing.pdf

Question 4

My function returns:

```
my_r_squared <- function(actual_ys, y_hats) {
  TSS <- sum((actual_ys - mean(actual_ys))^2)
  RSS <- sum((actual_ys - y_hats)^2)
  return(1 - RSS/TSS)
}

my_r_squared(nsw_data$re78,predict(lin_fit1))
```

```
> my_r_squared(nsw_data$re78,predict(lin_fit1))
[1] 0.004871571
```

To confirm this:

```
lin_fit1 <- lm(re78~treat, data = nsw_data)
summary(lin_fit1)
```

```
Multiple R-squared:  0.004872,
```

Question 5

a)

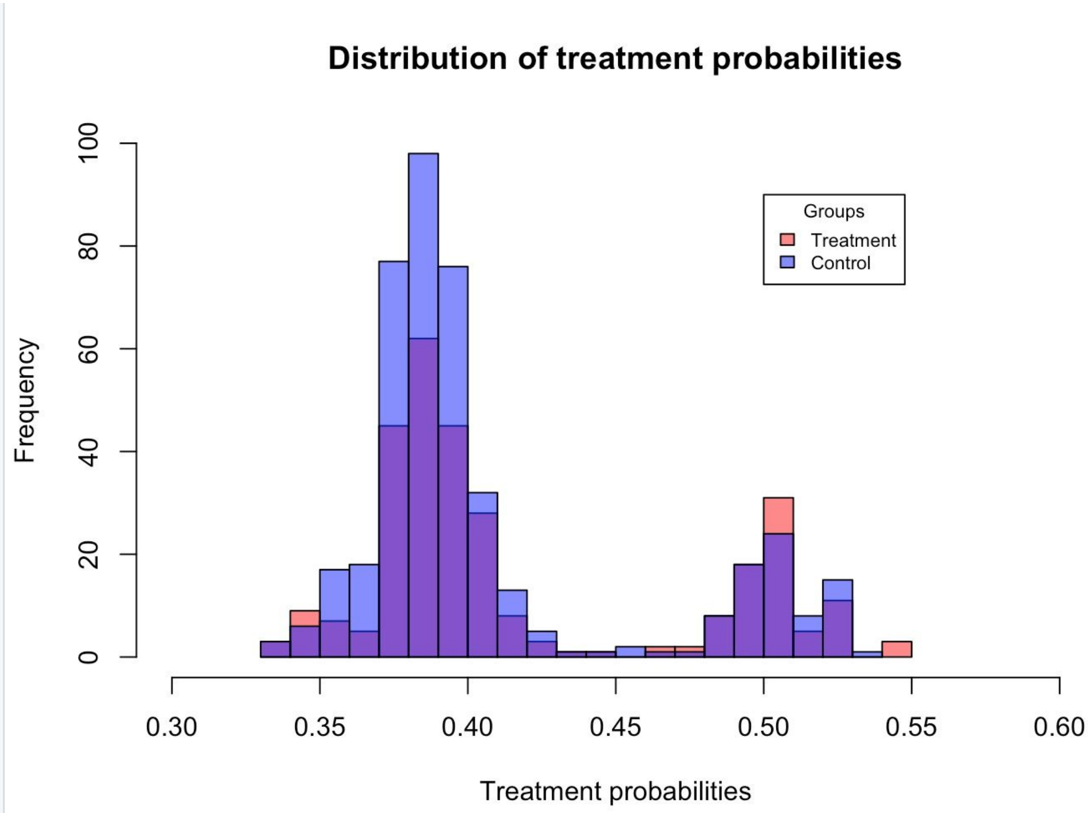

Figure 5. Histogram showing probabilities of being treated for treatment and control groups.

b)

The distributions largely overlap and have similar shapes, the control group has higher frequencies due its larger size. Intuitively, the histogram shows that most observations from the control group had low probabilities (below 0.5) of being in the treatment group. Counterintuitively, many observations in the treatment group also had low probabilities (below 0.5) of being in the treatment group suggesting our logistic model is not the best fit for our data.

Appendix

Code:

https://gist.github.com/bellabuchanan/8283a5b068e46fa9698d09fbffb47ead

References

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: With applications in R*. New York: Springer. Retrieved November 7, 2016 from: Retrieved from http://www-bcf.usc.edu/~gareth/ISL/ISLR%20First%20Printing.pdf