Assignment 3 - PDF

Devon Park - dap2189

2025-11-08

Question 1

The data below show plasma inorganic phosphate measurements obtained from 13 controls ('group' = 0) and 20 obese patients ('group' =1) taken initially ('initial') and 3 hours after an oral glucose challenge ('final'). The aim is to find if there is a difference in the final phosphate measurements between the two groups after adjusting for any initial difference.

Table 1: First few rows of Q1 dataset

group	initial	final
0	4.3	2.5
0	3.7	3.2
0	4.0	3.1
0	3.6	3.9
0	4.1	3.4
0	3.8	3.6

(a) Write down a simple ANCOVA model for data shown, using the control group as a reference.

```
##
## Call:
## lm(formula = final ~ group + initial, data = Q1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.0570 -0.4166  0.1031  0.4629  0.8495
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.8888     0.6373   1.395   0.1733
## group        -0.2340     0.2154  -1.086   0.2860
## initial       0.6005     0.1508   3.983   0.0004 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5755 on 30 degrees of freedom
## Multiple R-squared:  0.3462, Adjusted R-squared:  0.3026
## F-statistic: 7.942 on 2 and 30 DF,  p-value: 0.001706
```

Parallel slopes Model:

$final = \beta_0 + \beta_1 * group + \beta_2 * initial + \epsilon$

Model with coefficients:

$final = 0.889 - 0.234x_{group} + 0.601x_{initial}$

Note: group = 0 for the control (reference group) and group = 1 for obese patients.

(b) Test for any difference in the outcome between the two groups. Looking at our regression model, we see the coefficient for variable group is -0.234 with a p-value of 0.286. This means that the adjusted difference in the final phosphate readings for obese participants compared to control participants is -0.234. However, the p-value is greater than 0.05 which means this difference is not statistically significant at the 5% significance level.

(c) Does the initial phosphate value contribute significantly to the final phosphate value? Looking at our regression model, we see the coefficient for variable initial is 0.601 with a p-value of 0.004. This means that for every one unit increase in initial phosphate value, the final phosphate value increases by roughly 0.6 units, after adjusting for group. That the p-value is less than 0.05 suggests the association between initial and final phosphate scores is statistically significant.

(d) Repeat the ANCOVA without assuming parallel slopes for the two groups The interaction term tests whether the relation between initial and final phosphate values differ by group. If slopes are parallel $\beta_3 = 0$.

```
##
## Call:
## lm(formula = final ~ group + initial + group:initial, data = Q1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.0575 -0.4316  0.1190  0.4544  0.8917
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     0.5380     1.5830   0.340    0.736
## group           0.1900     1.7593   0.108    0.915
## initial         0.6862     0.3848   1.783    0.085 .
## group:initial  -0.1019     0.4195  -0.243    0.810
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5847 on 29 degrees of freedom
## Multiple R-squared:  0.3475, Adjusted R-squared:   0.28
## F-statistic: 5.148 on 3 and 29 DF,  p-value: 0.005628
```

```
## Analysis of Variance Table
##
## Model 1: final ~ group + initial
## Model 2: final ~ group + initial + group:initial
##   Res.Df    RSS Df Sum of Sq     F Pr(>F)
## 1     30 9.9359
## 2     29 9.9157  1  0.020172 0.059 0.8098
```

Non-parallel slopes Model:

$final = \beta_0 + \beta_1 * group + \beta_2 * initial + \beta_3 * (group * initial) + \epsilon$

Model with coefficients:

$final = 0.538 + 0.190x_{group} + 0.686x_{initial} - 0.102x_{group*initial}$

Note: group = 0 for the control patients (reference group) and group = 1 for obese patients.

The interaction term in model2 is $\beta_3 = -0.102$ with a p-value of 0.810. That the p-value is >0.05 suggests that initial and final phosphate values are not statistically significantly different depending on group. In other words, the association between initial and final phosphate scores is roughly similar across both groups.

Additionally, I ran `anova(m1, m2)` to compare m1 (parallel slopes model) with m2 (non-parallel slopes model). From this test, I get an F test statistic of 0.059 with a p-value of 0.810 suggesting that adding the interaction term does not significantly improve the model. In other words, we fail to reject the null hypothesis $H_0 : \beta_3 = 0$.

Question 2

The US Navy attempts to develop equations for estimation of manpower needs for manning installations such as Bachelor Officers Quarters (BOQ). Regression equations are developed from data taken by measurement teams. The data in the attached excel file were collected from 25 BOQ sites.

Variables of Dataset:

X_1 = average daily occupancy

X_2 = monthly average number of check-ins

X_3 = weekly hours of service desk operation

X_4 = square feet of common use area

X_5 = Number of building wings

X_6 = operational berthing capacity

X_7 = number of rooms

Y = monthly man-hours

(a) Fit the regression model

Table 2: First few rows of Q2 dataset

Site	x1	x2	x3	x4	x5	x6	x7	y
1	2.0	4.00	4.0	1.26	1	6	6	180.23
2	3.0	1.58	40.0	1.25	1	5	5	182.61
3	16.6	23.78	40.0	1.00	1	13	13	164.38
4	7.0	2.37	168.0	1.00	1	7	8	284.55
5	5.3	1.67	42.5	7.79	3	25	25	199.92
6	16.5	8.25	168.0	1.12	2	19	19	267.38

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7, data = Q2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -869.39 -180.75  -29.81  184.22  826.55
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 135.0380   237.7999   0.568  0.57755
## x1           -1.2841     0.8046  -1.596  0.12895
## x2            1.8040     0.5163   3.494  0.00278 **
```

```
## x3            0.6686     1.8463   0.362  0.72170
## x4          -21.4363    10.1701  -2.108  0.05020 .
## x5            5.6224    14.7452   0.381  0.70770
## x6          -14.4896     4.2196  -3.434  0.00317 **
## x7           29.3349     6.3643   4.609  0.00025 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 455.1 on 17 degrees of freedom
## Multiple R-squared:  0.9613, Adjusted R-squared:  0.9453
## F-statistic: 60.27 on 7 and 17 DF,  p-value: 9.199e-11
```

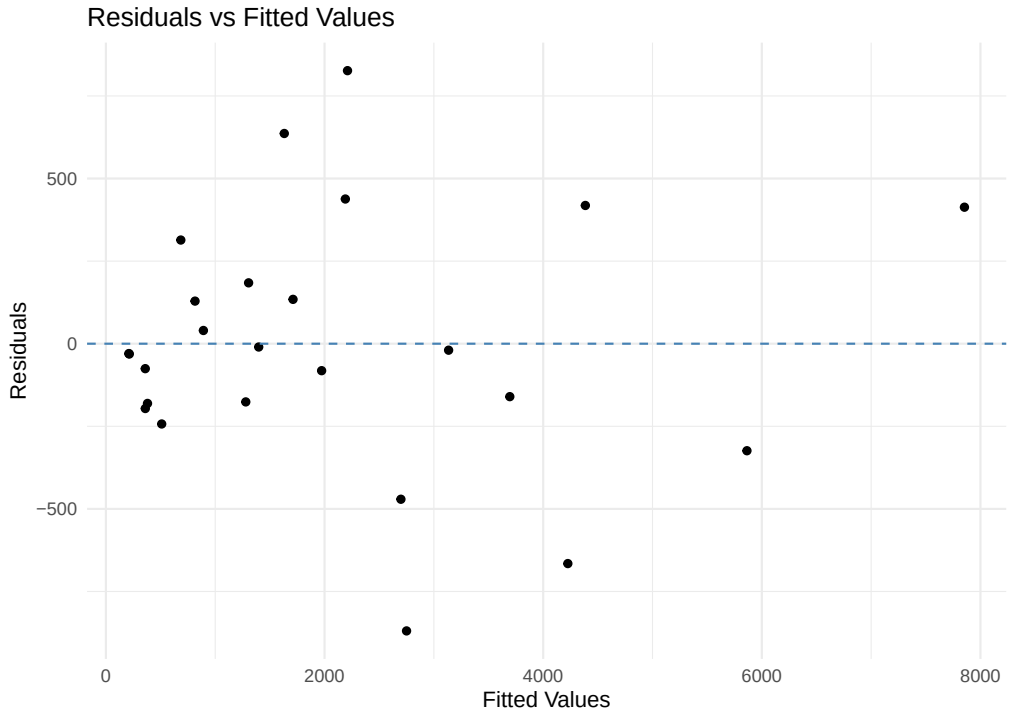
From our regression model I get the following result: Regression Model Form:

$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \epsilon$

Regresssion Model:

$Y = 135.038 - 1.284x_1 + 1.804x_2 + 0.669x_3 - 21.436x_4 + 5.622x_5 - 14.490x_6 + 29.335x_7$

(b) Perform a residual plot and make any necessary transformations.



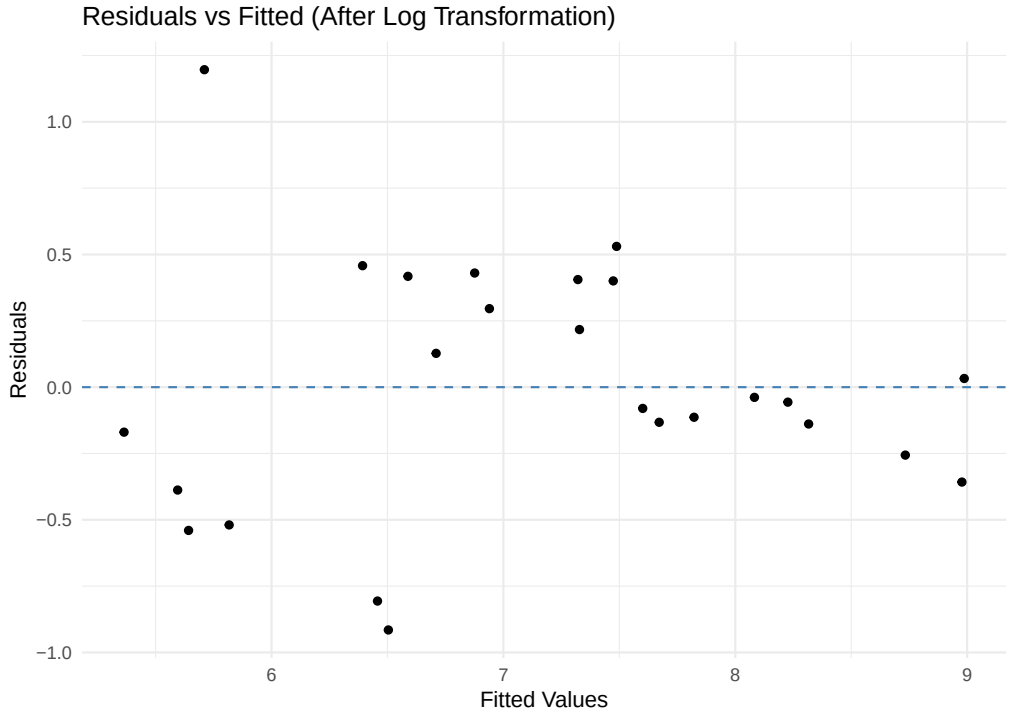
Based on the residual plot, we see a funneling effect suggesting heteroscedasticity (residuals spreading as our fitted values increase). This suggests that we should transform our data using a natural log.

(c) Refit a linear regression based on your transformation. Refit the model with the log transformation:

```
##
## Call:
## lm(formula = log_y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7, data = Q2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -0.9152 -0.2561 -0.0565  0.4004  1.1965
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.290e+00  2.901e-01  18.235 1.35e-12 ***
## x1          -9.815e-05  9.816e-04  -0.100  0.92152
## x2           5.637e-04  6.298e-04   0.895  0.38324
## x3           6.612e-03  2.252e-03   2.936  0.00924 **
## x4           1.857e-02  1.241e-02   1.497  0.15270
## x5          -8.452e-03  1.799e-02  -0.470  0.64443
## x6          -4.957e-03  5.148e-03  -0.963  0.34905
## x7           1.001e-02  7.764e-03   1.290  0.21438
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5552 on 17 degrees of freedom
## Multiple R-squared:  0.8407, Adjusted R-squared:  0.775
## F-statistic: 12.81 on 7 and 17 DF,  p-value: 1.138e-05
```

As an added step, we can check the residuals again: - After the log transformation, we see that our residuals form a cloud like shape suggesting a more even spread.



(d) Compute the hat diagonals, Standardized residuals, Cook's distance, and Studentized residuals for the model in (c)

Site	y	residuals	fitted	log_y	residuals_log	fitted_log	hat	std_resid	stud_resid	cooks
1	180.23	-29.815	210.045	5.194	-0.170	5.364	0.257	-0.354	-0.345	0.005
2	182.61	-31.225	213.835	5.207	-0.388	5.595	0.161	-0.762	-0.753	0.014
3	164.38	-196.161	360.541	5.102	-0.540	5.642	0.161	-1.062	-1.066	0.027
4	284.55	-75.543	360.093	5.651	-0.806	6.457	0.163	-1.587	-1.668	0.061
5	199.92	-180.752	380.672	5.298	-0.519	5.817	0.147	-1.013	-1.014	0.022

Site	y	residuals	fitted	log_y	residuals_log	fitted_log	hat	std_resid	stud_resid	cooks
6	267.38	-242.980	510.360	5.589	-0.915	6.504	0.159	-1.797	-1.937	0.076
7	999.09	313.842	685.248	6.907	1.196	5.710	0.183	2.384	2.834	0.159
8	1103.24	-176.187	1279.427	7.006	0.418	6.588	0.359	0.940	0.936	0.062
9	944.21	128.966	815.244	6.850	0.458	6.393	0.281	0.972	0.971	0.046
10	931.84	40.022	891.818	6.837	0.127	6.710	0.130	0.246	0.239	0.001
11	2268.06	636.320	1631.740	7.727	0.405	7.321	0.124	0.780	0.771	0.011
12	1489.50	184.222	1305.278	7.306	0.430	6.876	0.202	0.867	0.861	0.024
13	1891.70	-81.657	1973.357	7.545	0.217	7.328	0.080	0.408	0.397	0.002
14	1387.82	-9.990	1397.810	7.235	0.296	6.940	0.097	0.561	0.549	0.004
15	3559.92	-665.544	4225.464	8.177	-0.139	8.316	0.558	-0.376	-0.366	0.022
16	3115.29	-19.585	3134.875	8.044	-0.038	8.083	0.402	-0.090	-0.087	0.001
17	2227.76	-470.820	2698.580	7.709	-0.113	7.822	0.368	-0.257	-0.250	0.005
18	4804.24	418.470	4385.770	8.477	-0.256	8.733	0.446	-0.620	-0.608	0.039
19	2628.32	438.007	2190.313	7.874	0.400	7.474	0.087	0.755	0.745	0.007
20	1880.84	-869.388	2750.228	7.539	-0.132	7.672	0.366	-0.299	-0.291	0.006
21	3036.63	826.554	2210.076	8.019	0.530	7.488	0.070	0.991	0.990	0.009
22	5539.98	-323.936	5863.916	8.620	-0.358	8.977	0.785	-1.391	-1.433	0.884
23	3534.49	-160.268	3694.758	8.170	-0.056	8.227	0.988	-0.947	-0.944	9.606
24	8266.77	413.218	7853.552	9.020	0.033	8.987	0.876	0.167	0.162	0.025
25	1845.89	134.232	1711.658	7.521	-0.080	7.601	0.547	-0.214	-0.208	0.007

Note: Removed columns x1:x7 for the above table to make the tables fit on the page

(e) Identify any influential points and discuss the nature of the influence. We can use cooks distance to look at how much the fitted model changes when a certain observation is removed. We consider a point influential if D (cook's distance) is greater than 1.

Site	y	residuals	fitted	log_y	residuals_log	fitted_log	hat	std_resid	stud_resid	cooks	influential
1	180.23	-29.815	210.045	5.194	-0.170	5.364	0.257	-0.354	-0.345	0.005	FALSE
2	182.61	-31.225	213.835	5.207	-0.388	5.595	0.161	-0.762	-0.753	0.014	FALSE
3	164.38	-196.161	360.541	5.102	-0.540	5.642	0.161	-1.062	-1.066	0.027	FALSE
4	284.55	-75.543	360.093	5.651	-0.806	6.457	0.163	-1.587	-1.668	0.061	FALSE
5	199.92	-180.752	380.672	5.298	-0.519	5.817	0.147	-1.013	-1.014	0.022	FALSE
6	267.38	-242.980	510.360	5.589	-0.915	6.504	0.159	-1.797	-1.937	0.076	FALSE
7	999.09	313.842	685.248	6.907	1.196	5.710	0.183	2.384	2.834	0.159	FALSE
8	1103.24	-176.187	1279.427	7.006	0.418	6.588	0.359	0.940	0.936	0.062	FALSE
9	944.21	128.966	815.244	6.850	0.458	6.393	0.281	0.972	0.971	0.046	FALSE
10	931.84	40.022	891.818	6.837	0.127	6.710	0.130	0.246	0.239	0.001	FALSE
11	2268.06	636.320	1631.740	7.727	0.405	7.321	0.124	0.780	0.771	0.011	FALSE
12	1489.50	184.222	1305.278	7.306	0.430	6.876	0.202	0.867	0.861	0.024	FALSE
13	1891.70	-81.657	1973.357	7.545	0.217	7.328	0.080	0.408	0.397	0.002	FALSE
14	1387.82	-9.990	1397.810	7.235	0.296	6.940	0.097	0.561	0.549	0.004	FALSE
15	3559.92	-665.544	4225.464	8.177	-0.139	8.316	0.558	-0.376	-0.366	0.022	FALSE
16	3115.29	-19.585	3134.875	8.044	-0.038	8.083	0.402	-0.090	-0.087	0.001	FALSE
17	2227.76	-470.820	2698.580	7.709	-0.113	7.822	0.368	-0.257	-0.250	0.005	FALSE

Site	y	residuals	fitted	log_y	residuals_log	fitted_log	hat	std_resid	stud_resid	cooks	influential
18	4804.24	418.470	4385.770	8.477	-0.256	8.733	0.446	-0.620	-0.608	0.039	FALSE
19	2628.32	438.007	2190.313	7.874	0.400	7.474	0.087	0.755	0.745	0.007	FALSE
20	1880.84	-869.388	2750.228	7.539	-0.132	7.672	0.366	-0.299	-0.291	0.006	FALSE
21	3036.63	826.554	2210.076	8.019	0.530	7.488	0.070	0.991	0.990	0.009	FALSE
22	5539.98	-323.936	5863.916	8.620	-0.358	8.977	0.785	-1.391	-1.433	0.884	FALSE
23	3534.49	-160.268	3694.758	8.170	-0.056	8.227	0.988	-0.947	-0.944	9.606	TRUE
24	8266.77	413.218	7853.552	9.020	0.033	8.987	0.876	0.167	0.162	0.025	FALSE
25	1845.89	134.232	1711.658	7.521	-0.080	7.601	0.547	-0.214	-0.208	0.007	FALSE

By filtering our table, we can see how there is only 1 influential point:

Site	y	residuals	fitted	log_y	residuals_log	fitted_log	hat	std_resid	stud_resid	cooks	influential
23	3534.49	-160.268	3694.758	8.17	-0.056	8.227	0.988	-0.947	-0.944	9.606	TRUE

Note: Removed columns x2:x7 for the above table to make the tables fit on the page

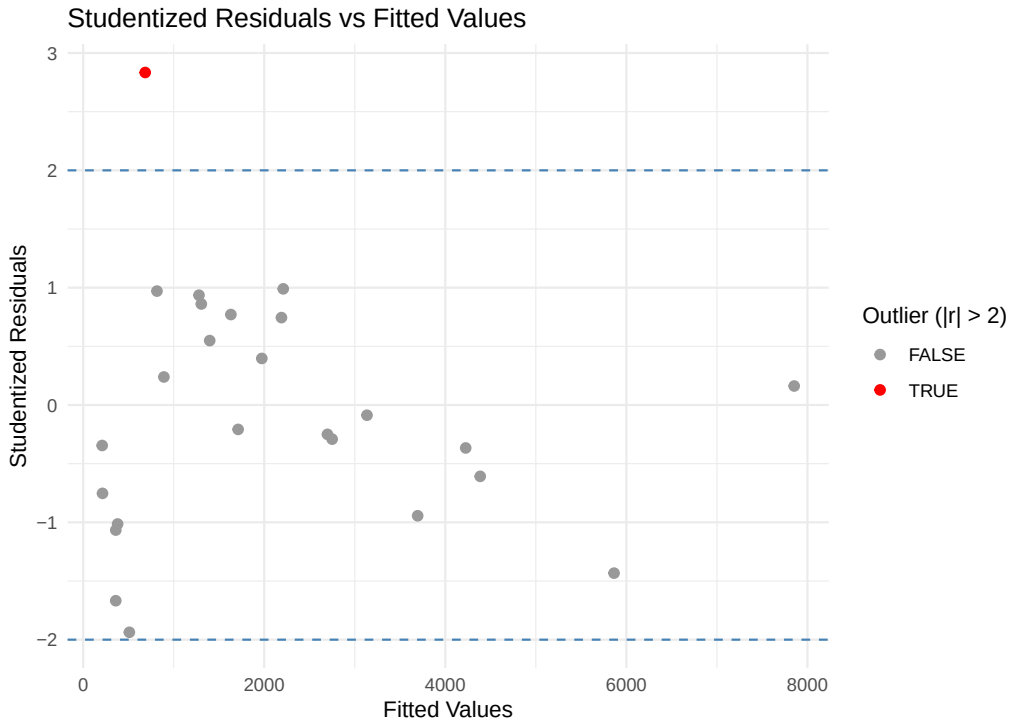
Points that are high leverage (large hat diagonal) tell us that the observation's value on the x axis is distant from the mean. With our influential point we see a large hat diagonol value of 0.988. This indicates that it was distant on the x-axis compared to the other points. As it is an influential point, so we will not remove it from our data.

(f) Identify any outliers and take any remedial action necessary. We look a the studentized residuals (standardized residuals) to find outliers ,r. If $|r| > 2$, we classify this point as an outlier.

Site	y	residuals	fitted	log_y	residuals_log	fitted_log	hat	std_resid	stud_resid	cooks	influential
7	999.09	313.842	685.248	6.907	1.196	5.71	0.183	2.384	2.834	0.159	FALSE

Note: Removed columns x1:x7 for the above table to make the tables fit on the page

We can see outliers graphically below:



The point at site 7 is found as an outlier because the absolute value of the studentized residuals is greater than 2. So, we remove this point from our dataset and refit the model.

```
##
## Call:
## lm(formula = log_y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7, data = Q2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.87110 -0.21372 -0.02625  0.33315  0.60769
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.993e+00  2.655e-01  18.805 2.47e-12 ***
## x1          -1.292e-05  8.261e-04  -0.016 0.987719
## x2           7.033e-04  5.320e-04   1.322 0.204759
## x3           8.177e-03  1.973e-03   4.144 0.000763 ***
## x4           2.397e-02  1.061e-02   2.260 0.038104 *
## x5          -7.540e-03  1.513e-02  -0.498 0.625078
## x6          -4.019e-03  4.342e-03  -0.926 0.368398
## x7           7.999e-03  6.568e-03   1.218 0.240948
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.467 on 16 degrees of freedom
## Multiple R-squared:  0.8937, Adjusted R-squared:  0.8472
## F-statistic: 19.22 on 7 and 16 DF,  p-value: 1.127e-06
```

After remediation, here is the final model:

$$ln(Y) = 4.993 - 0.000013X_1 + 0.000703X_2 + 0.008177X_3 + 0.023974X_4 - 0.007540X_5 - 0.004019X_6 + 0.007999X_7$$