Greyson Newton

1. (10 pts) Assume a response variable, y, is linear correlated with three main effects, x_1, x_2, x_3. To investigate the relationships between response variable and main effects, an engineer conducts experiments, obtains the data shown in Figure 1 and fits a linear regression model. But the engineer misses a main effect (x_2) when fitting the model, leading to the following regression model:

$$y = \beta_0 + \beta_1 x_1 + \beta_3 x_3 + \epsilon,\ \epsilon \sim (0, \sigma^2)$$

a) Estimate the coefficients (β_0, β_1, β_3) in the fitted model.

[44.81, 4.96, 1.3116066323613498]

b) Is the fitted model biased? If yes, please estimate the bias on (β_0, β_1, β_2, β_3); If no, please explain why.

The least squared method is mathematically unbiased.

c) Give the 95% confidence intervals for (β_0, β_1, β_2, β_3) in the full model.

11.857061177815892

d) Compare the R^2 and adjusted R^2 of the full model and the model fitted by engineer. What conclusion you could draw?

The full model's R2 and adjusted R2 values are closer to 1 compared to the r2 numbers of the engineer's model. This implies that the full model is a better fit to the relationships in the data.

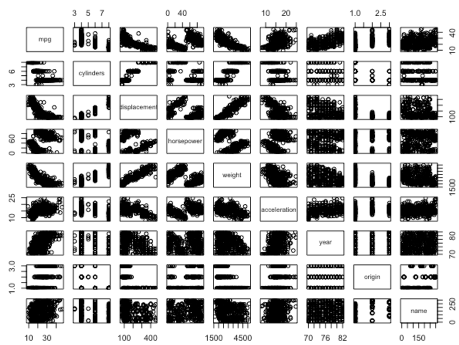
x_1	x_2	x_3	y
-1	-1	-1	32
1	-1	-1	46
-1	1	-1	57
1	1	-1	65
-1	-1	1	36
1	-1	1	48
-1	1	1	57
1	1	1	68
0	1	0	50
0	0	1	44
1	1	0	53
1	0	1	56

Table 1: Dataset of Problem 1.

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 12 & 2 & 2 & 2 \\ 2 & 10 & 1 & 1 \\ 2 & 1 & 10 & 0 \\ 2 & 1 & 0 & 10 \end{bmatrix},\ \mathbf{X}'\mathbf{y} = \begin{bmatrix} 612 \\ 154 \\ 188 \\ 109 \end{bmatrix} \text{ where } \mathbf{X} = [\mathbf{1}, \boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3]$$

2. (10 pts) Problem 3.9 on Page 122 of textbook.

a) > pairs (Auto)



b) > cor(Auto[1:8])

```
                    mpg  cylinders displacement horsepower     weight acceleration       year
mpg           1.0000000 -0.7762599   -0.8044430  0.4228227 -0.8317389    0.4222974  0.5814695
cylinders    -0.7762599  1.0000000    0.9509199 -0.5466585  0.8970169   -0.5040606 -0.3467172
displacement -0.8044430  0.9509199    1.0000000 -0.4820705  0.9331044   -0.5441618 -0.3698041
horsepower    0.4228227 -0.5466585   -0.4820705  1.0000000 -0.4821507    0.2662877  0.1274167
weight       -0.8317389  0.8970169    0.9331044 -0.4821507  1.0000000   -0.4195023 -0.3079004
acceleration  0.4222974 -0.5040606   -0.5441618  0.2662877 -0.4195023    1.0000000  0.2829009
year          0.5814695 -0.3467172   -0.3698041  0.1274167 -0.3079004    0.2829009  1.0000000
origin        0.5636979 -0.5649716   -0.6106643  0.2973734 -0.5812652    0.2100836  0.1843141
                 origin
mpg           0.5636979
cylinders    -0.5649716
displacement -0.6106643
horsepower    0.2973734
weight       -0.5812652
acceleration  0.2100836
year          0.1843141
origin        1.0000000
```

c)

i)

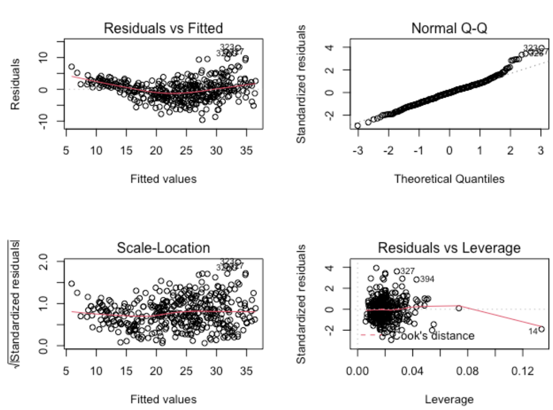
> fit = lm(mpg~.-name,data=Auto)
> summary(fit)

```
Call:
lm(formula = mpg ~ . - name, data = Auto)

Residuals:
   Min     1Q Median     3Q    Max
-9.629 -2.034 -0.046  1.801 13.010

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2.128e+01  4.259e+00  -4.998 8.78e-07 ***
cylinders    -2.927e-01  3.382e-01  -0.865   0.3874
displacement  1.603e-02  7.284e-03   2.201   0.0283 *
horsepower    7.942e-03  6.809e-03   1.166   0.2442
weight       -6.870e-03  5.799e-04 -11.846  < 2e-16 ***
acceleration  1.539e-01  7.750e-02   1.986   0.0477 *
year          7.734e-01  4.939e-02  15.661  < 2e-16 ***
origin        1.346e+00  2.691e-01   5.004 8.52e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.331 on 389 degrees of freedom
Multiple R-squared:  0.822, Adjusted R-squared:  0.8188
F-statistic: 256.7 on 7 and 389 DF,  p-value: < 2.2e-16
```

ii) **The p-value implies that all variables are statistically significant aside from "*horsepower*", "*cylinder*", and "*acceleration*".**

iii) **The coefficient of "*year*" maps a unit change in the variable "year" to a .7507 change in the isolated variable being studied.**

d)
> par(mfrow=c(2,2))
> plot(fit)



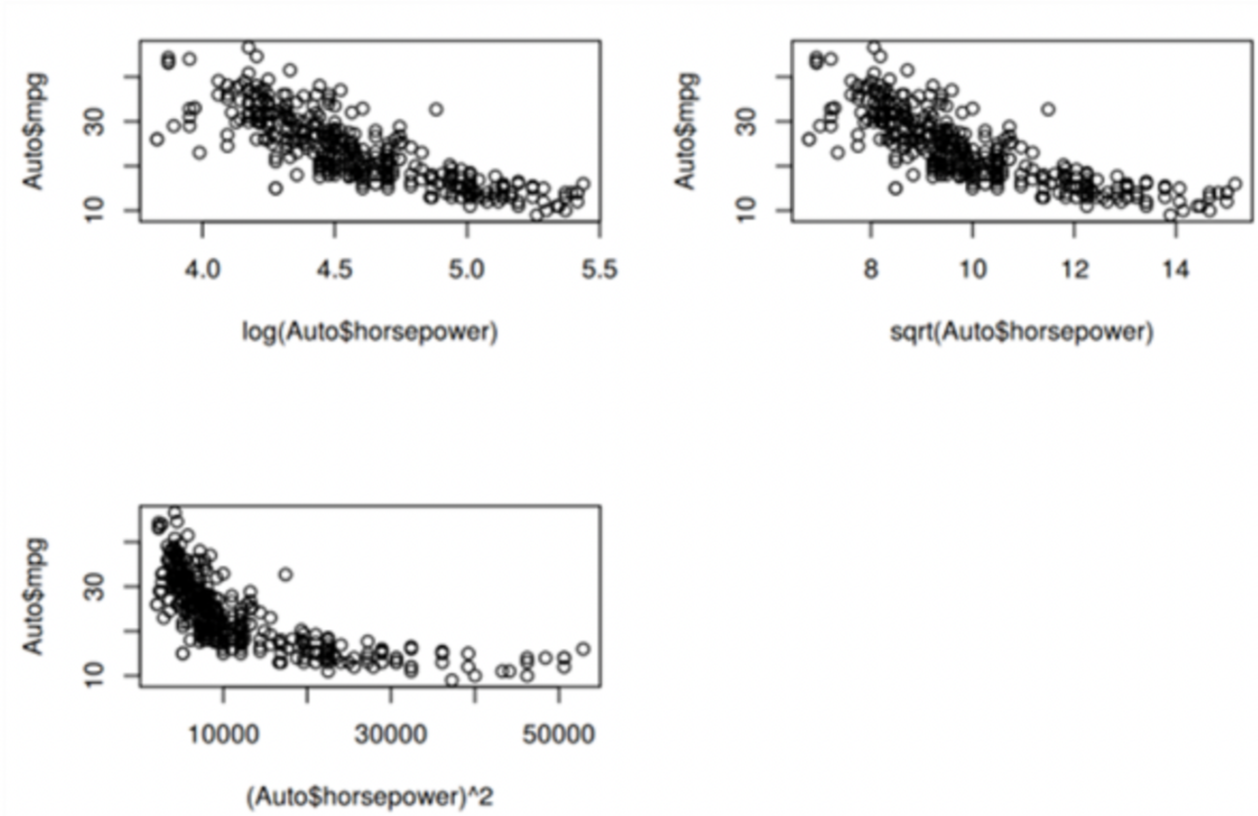
e)

```
> fit = lm(mpg ~ cylinders*displacement+displacement*weight,data=Auto[1:8])
> summary(fit)

Call:
lm(formula = mpg ~ cylinders * displacement + displacement *
    weight, data = Auto[1:8])

Residuals:
     Min       1Q   Median       3Q      Max
-13.3564  -2.4882  -0.3635   1.8469  17.8176

Coefficients:
                        Estimate Std. Error
(Intercept)            5.285e+01  2.233e+00
cylinders              7.580e-01  7.645e-01
displacement          -7.514e-02  1.669e-02
weight                -9.931e-03  1.323e-03
cylinders:displacement -2.893e-03  3.424e-03
displacement:weight    2.147e-05  4.996e-06
                       t value Pr(>|t|)
(Intercept)             23.673  < 2e-16 ***
cylinders                0.992    0.322
displacement            -4.502 8.90e-06 ***
weight                  -7.505 4.19e-13 ***
cylinders:displacement  -0.845    0.399
displacement:weight      4.298 2.18e-05 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.115 on 391 degrees of freedom
Multiple R-squared:  0.7269,  Adjusted R-squared:  0.7234
F-statistic: 208.2 on 5 and 391 DF,  p-value: < 2.2e-16
```

The p-value implies no statistical significance between the variables, "*mpg*" and "*cylinders*".

f)

```
> par(mfrow=c(2,2))
> plot(log(Auto$horsepower),Auto$mpg)
> plot(sqrt(Auto$horsepower),Auto$mpg)
> plot((Auto$horsepower)^2,Auto$mpg)
```



It seems as though the logged “*Horsepower”* has a very true Linear relationship when plotted against “*mpg*”, implying an exponential trend between the two variables .

3. (10 pts) *Piecewise-polynomial fitting*: Assume predictor $f(x)$ is a piecewise- polynomial function with the following form:

$$f(x) = \begin{cases} \theta_1 + \theta_2 x + \theta_3 x^2 & x \le a \\ \theta_4 + \theta_5 x + \theta_6 x^2 & x > a \end{cases}$$

Where a is given.

a) Please give two equality constraints to guarantee the predictor is continuous and has the first order derivation at a (smoothness).

$\boldsymbol{\theta_1 + \theta_2 x + \theta_3 x^2 = \theta_4 + \theta_5 x + \theta_6 x^2}$

AND

$\boldsymbol{\theta_2 + 2\theta_3 x = \theta_5 + 2\theta_6 x}$

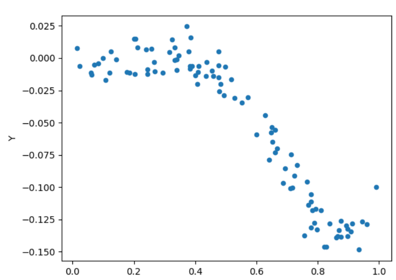
b) Please write the constrained least square problem to estimate the parameters $(\theta_1, \ldots, \theta_6)$ under the two equality constraints in (a). $D = \{(x_1, y_1), \ldots, (x_N, y_N)\}$

$\mathbf{SUM}_{x<a}(\mathbf{y\_i} - \theta_1 + \theta_2 x\_i + \theta_3 x\_i^2)^2 + \mathbf{SUM}_{x>a}(\mathbf{y\_i} - \theta_4 + \theta_5 x\_i + \theta_6 x\_i^2)^2$

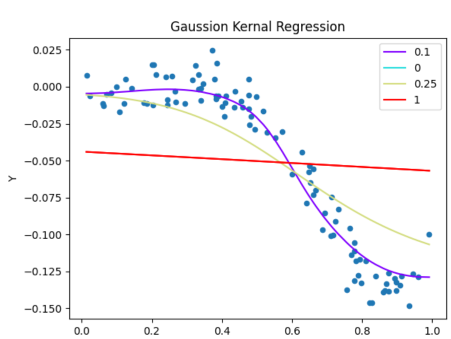
The above expression is constrained by smoothness/continuous equality constraints from part a

4. (10 pts) Download the datasets nonlintrain.txt and nonlintest.txt from Blackboard and answer the following questions.

a) Plot the 100 x points versus these 100 y points in training data to get an idea of the trend.



b) Fit a kernel regression on the training data, with 3 different values of the bandwidth parameter: 0.01, 0.25, and 1. You should use the Gaussian kernel. For each bandwidth value, plot the estimated regression function from kernel regression over top of the training points.

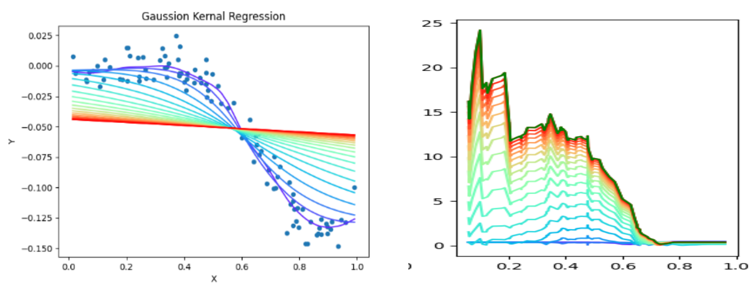


c) By inspection, what happens to the kernel regression fit as we drive the bandwidth parameter down to 0?

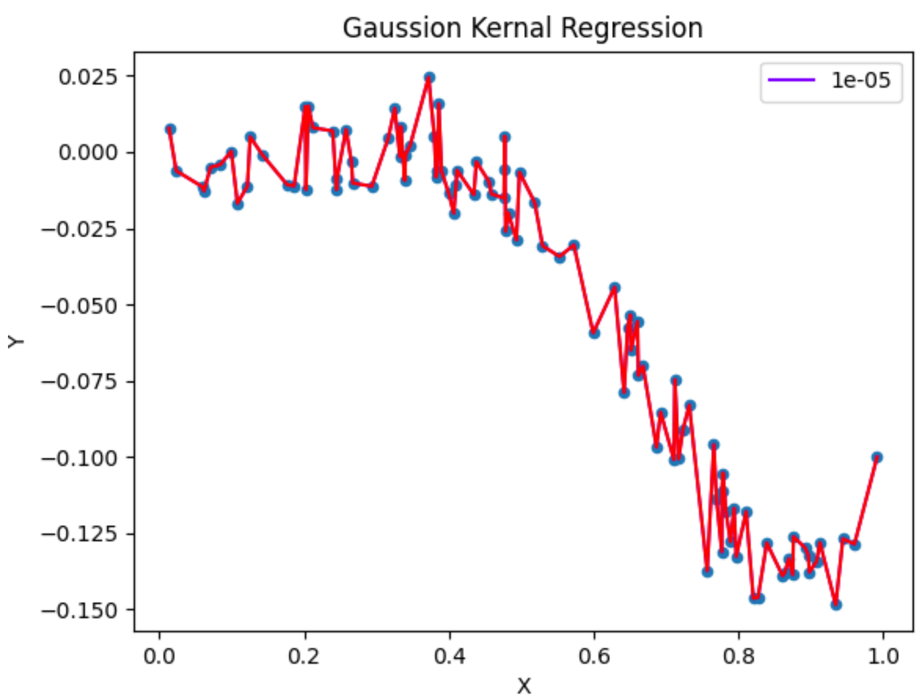
Fit converges to trends in data.

d) Investigate the predictive performance on test set. For a set of 20 bandwidth values, equally spaced between 0.01 and 1, fit a kernel regression to the training points and predict the regression function at the test x points. Evaluate its test error, measured in terms of squared error loss to the test y points. Hence, you will have a curve of 20 test errors; plot this test error curve as a function of the underlying bandwidth values.

Loss functions for Bandwidths



e) According to this test error curve, what is the optimal bandwidth value? What is its associated test error? Plot the kernel regression fit, over top of the training points, at this optimal bandwidth value. Looking at the plot, does your eye agree that this is really the best bandwidth value? Why or why not?



0 and 0… Yes because it most accurately reflects the trend data. Lowering bandwidth seems to minimize Gaussian Kernal Regression residuals. However overfitting could be in an issue. Look at a bandwidth of 0.000001 and a 0-valued loss function: