***Greyson Newton***

1. (10 pts) Assume a response variable, y, is linear correlated with three main effects, 𝑥1, 𝑥2, 𝑥3. 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 (𝑥2) when fitting the model, leading to the following regression model:

y = 𝛽0 + 𝛽1𝑥1 + 𝛽3𝑥3 + 𝜖, 𝜖~(0, 𝜎2)

* 1. Estimate the coefficients (𝛽0, 𝛽1, 𝛽3) in the fitted model.

**[44.81, 4.96, 1.3116066323613498]**

* 1. Is the fitted model biased? If yes, please estimate the bias on (𝛽0, 𝛽1, 𝛽2, 𝛽3); If no, please explain why.

**The least squared method is mathematically unbiased.**

* 1. Give the 95% confidence intervals for (𝛽0, 𝛽1, 𝛽2, 𝛽3) in the full model.

**11.857061177815892**

* 1. Compare the 𝑅2 and adjusted 𝑅2 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.**

|  |  |  |  |
| --- | --- | --- | --- |
| 𝑥1 | 𝑥2 | 𝑥3 | 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.

12 2

𝐗′𝐗 = [ 2 10

|  |  |  |  |
| --- | --- | --- | --- |
| 2 | 1 1 | 0 0 | 188 |
| 2 | 1 0 10 | | 109 |

2 2 612

1 1 ], 𝐗′𝐲 = [ 154 ] where 𝐗 = [𝟏, 𝒙1, 𝒙2, 𝒙3]

1. (10 pts) Problem 3.9 on Page 122 of textbook.

a) > pairs (Auto)

A picture containing text

Description automatically generated

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)

Diagram

Description automatically generated

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)**

**Diagram

Description automatically generated**

**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 .**

1. (10 pts) *Piecewise-polynomial fitting*: Assume predictor 𝑓(𝑥) is a piecewise- polynomial function with the following form:

Where 𝑎 is given.

𝑓(𝑥) = {

𝜃1 + 𝜃2𝑥 + 𝜃3𝑥2 𝑥 ≤ 𝑎

𝜃4 + 𝜃5 𝑥 + 𝜃6 𝑥2 𝑥 > 𝑎

* 1. Please give two equality constraints to guarantee the predictor is continuous and has the first order derivation at 𝑎 (smoothness).  
     **𝜃1 + 𝜃2𝑥 + 𝜃3𝑥2 = 𝜃4 + 𝜃5 𝑥 + 𝜃6 𝑥2**

**AND**

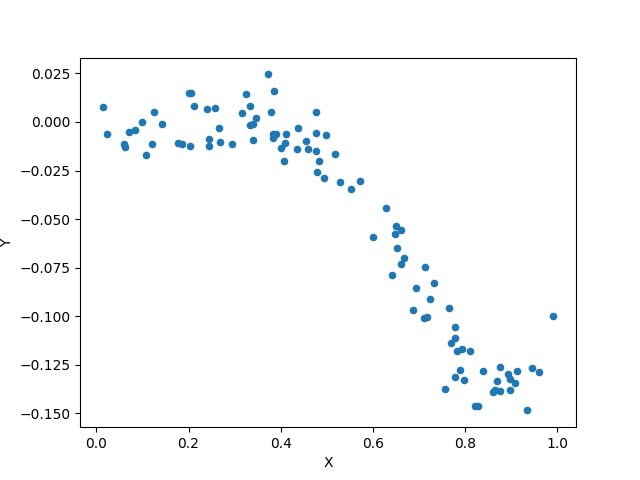
**𝜃2 +2 𝜃3𝑥 = 𝜃5 +2 𝜃6 𝑥**

* 1. Please write the constrained least square problem to estimate the parameters (𝜃1, … , 𝜃6) under the two equality constraints in (a). 𝐷 = {(𝑥1, 𝑦1), … , (𝑥𝑁, 𝑦𝑁)}

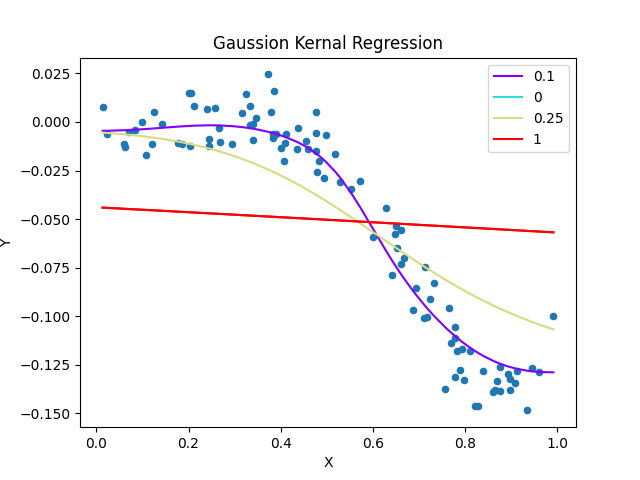
**SUMx<a(y\_i - 𝜃1 + 𝜃2𝑥\_i + 𝜃3𝑥\_i2)2 +SUMx>a(y\_i - 𝜃4 + 𝜃5𝑥\_i + 𝜃6𝑥\_i2)2**

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

1. (10 pts) Download the datasets nonlintrain.txt and nonlintest.txt from Blackboard and answer the following questions.
   1. Plot the 100 x points versus these 100 y points in training data to get an idea of the trend.



* 1. 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.

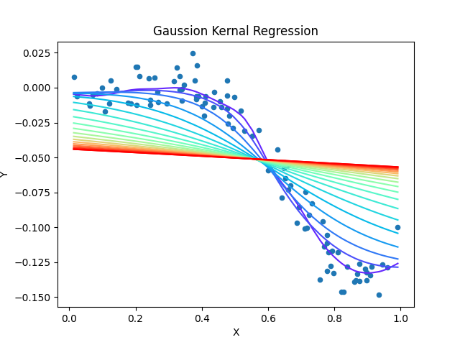


* 1. By inspection, what happens to the kernel regression fit as we drive the bandwidth parameter down to 0?

**Fit converges to trends in data.**

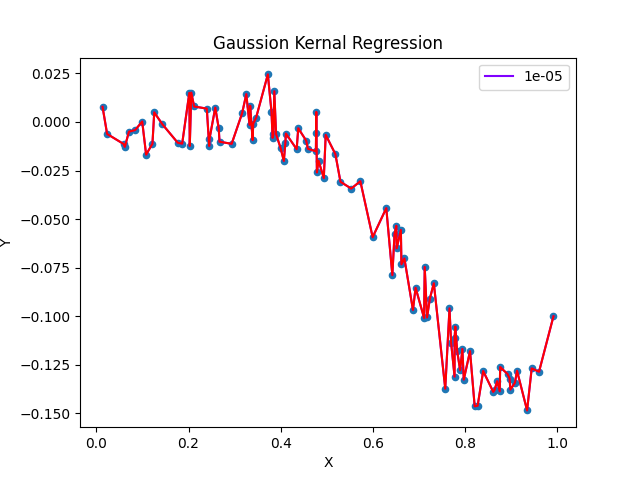
* 1. Investigate the predictive performance on test set. For a set of 20 bandwidth values, equally spaced between 0.01 and 1, fit a kernel regression to the training points and predict the regression function at the test x points. Evaluate its test error, measured in terms of squared error loss to the 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**

Chart, histogram

Description automatically generated

* 1. 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:**