Project

**Contents**

1. [Question 1](#_bookmark1) 2
2. [Question 2](#_bookmark2) 8
3. [Question 3](#_bookmark3) 13
4. [Question 4](#_bookmark4) 18
5. [Question 5](#_bookmark5) 21
6. [Question 6](#_bookmark6) 24

# Question 1

The file **stockdata.csv** uploaded to the blackboard is a dataset that contains the price of a stock in the last 100 days as the response and the following as predictors:

* + **vol**: Volatility of the stock.
  + **cap.to.gdp**: The ratio of the market cap to GDP.
  + **q.ratio**: The ratio of market cap to net worth.
  + **gaap**: Shiller Cape Index.
  + **avg.allocation**: Average investor equity allocation of the stock.

Fit a model to explain price in terms of the predictors. Perform regression diagnostics to answer the following questions. Display any plots that are relevant and explain your reasoning. Suggest possible improvements if there are any.

Table 1: Sample of Stocks Data

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| X | cap.to.gdp | q.ratio | gaap | trailing.pe | avg.allocation | price | vol |
| 1 | 0.06936 | 0.483 | 0.973 | 0.5996 | 0.001411 | 1.143 | 0.92 |
| 2 | 0.8178 | 0.8411 | 0.2173 | 0.8306 | 0.2766 | 1.162 | 0.22 |
| 3 | 0.9426 | 0.4551 | 0.5205 | 0.7397 | 0.4798 | 1.165 | 0.02 |
| 4 | 0.2694 | 0.8605 | 0.8279 | 0.5644 | 0.5103 | 1.165 | 0.37 |
| 5 | 0.1693 | 0.6761 | 0.9641 | 0.3643 | 0.9031 | 1.166 | 0.93 |
| 6 | 0.0339 | 0.7276 | 0.6319 | 0.9902 | 0.1554 | 1.153 | 0.42 |

## 1.1 (a)

Fit a model to explain **price** in terms of the predictors. Which variables are important, can any of the variables be removed? Please use F-test to justify.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | t value | Pr(>|t|) |  |
| **(Intercept)** | 1.109 | 0.001272 | 871.5 | 1.006e-183 | \* \* \* |
| **cap.to.gdp** | 0.0209 | 0.001053 | 19.84 | 3.488e-35 | \* \* \* |
| **q.ratio** | 0.01811 | 0.001041 | 17.39 | 5.53e-31 | \* \* \* |
| **gaap** | 0.01633 | 0.0009298 | 17.56 | 2.803e-31 | \* \* \* |
| **trailing.pe** | 0.01438 | 0.000975 | 14.75 | 4.435e-26 | \* \* \* |
| **avg.allocation** | 0.02259 | 0.0009978 | 22.64 | 1.319e-39 | \* \* \* |
| **vol** | -0.0005667 | 0.0009918 | -0.5714 | 0.5691 |  |

Table 3: Model Summary

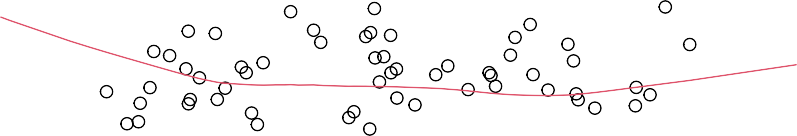
Observations Residual Std. Error *R*2 Adjusted *R*2

100 0.002758 0.9535 0.9505

## 1.2 (b)

Check the constant variance assumption of the errors.

Scale−Location



59

42

85

Standardized residuals

1.0

1.5

1.13 1.14 1.15 1.16 1.17 1.18

0.0

0.5

Fitted values

A clear curve pattern can be observed in the residuals. Therefore, the assumption of constant variance of the errors is violated.

## 1.3 (c)

Check the independentness of the errors assumption.

##### Hypothesis:

*H*0 : There is no correlation among the residuals.

*H*1 : The residuals are auto-correlated.

lag Autocorrelation D-W Statistic p-value 1 0.07666919 1.839674 0.404

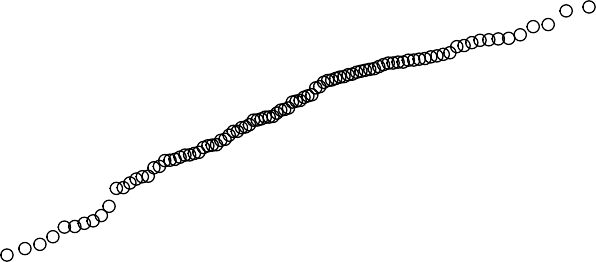
Alternative hypothesis: rho != 0

Since, p-value ≮ 0.05, therefore we do not reject *H*0 and conclude that there is no correlation among the residuals.

## 1.4 (d)

Check the normality assumption.

Normal Q−Q



59

42 85

Standardized residuals

1

2

3

−2 −1 0 1 2

−2

−1

0

Theoretical Quantiles

##### Hypothesis:

*H*0 : The residuals are normally distributed.

*H*1 : The residuals are not normally distributed.

Shapiro-Wilk normality test

data: residuals(stocks\_model) W = 0.97252, p-value = 0.0346

For the Shapiro Wilk Test, since the p-value < 0.05, we reject *H*0 and conclude that the residuals are not normally distributed. The same can be observed from the QQ-Plot of the residuals. The points deviate from the normal line at tails.

## 1.5 (e)

Is non-linearity a problem?

0.0 0.6

cap.to.gdp

0.0

1.0

0.0

1.0



X



0 40 100

0.0

1.0

0.0

1.0



q.ratio











0.0 0.6

1.13

0.0

1.0

0.0 0.6







gaap











trailing.pe







0.0 0.6

0.0 0.6











avg.allocation





price

1.13 1.17

0.0 0.6









vol



0.0

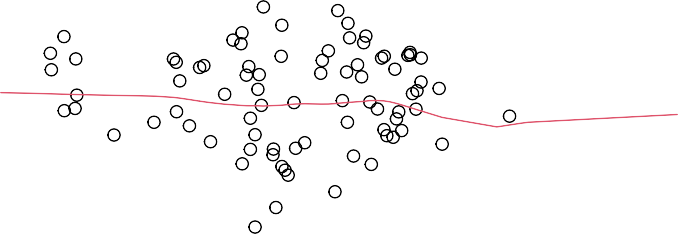
1.0

0 80

## 1.6 (f)

Check for outliers, compute and plot Cook’s Distance

Residuals vs Leverage



36

Cook's distance

85

9

Standardized residuals

0

1

2

3

0.00 0.02 0.04 0.06 0.08 0.10 0.12 0.14

−3

−2

−1

Leverage

The observations 9, 36 and 85 are potential outliers in the data.

Cook's distance

9

85

36

Cook's distance

0.08

0 20 40 60 80 100

0.00

0.04

Obs. number

## 1.7 (g)

Check for influential points.

|  |  |  |
| --- | --- | --- |
| price | cap.to.gdp q.ratio gaap trailing.pe avg.allocation | vol |
| 1 1.127683 | 0.02287774 0.1466545 0.5426289 0.6547472 0.1711072 | 0.09 |
| 2 1.160614 | 0.94108754 0.8894984 0.3154241 0.1179111 0.6639421 | 0.86 |
| 3 1.177698 | 0.71841339 0.8152898 0.9223454 0.9728031 0.7486021 | 0.28 |
| .fitted | .resid .std.resid .hat .sigma .cooksd |  |
| 1 1.133936 | -0.006253064 -2.403777 0.1103318 0.002685365 0.10236769 |  |
| 2 1.165848 | -0.005233825 -2.005918 0.1049586 0.002712238 0.06740666 |  |
| 3 1.184290 | -0.006591951 -2.502884 0.0880370 0.002677869 0.08639160 |  |

## 1.8 (h)

The return at time *t* is defined as: *r*(*t*) = *p*(*t*+1) *−* 1

*p*(*t*)

where p is the price data for day *t*. Are the returns normally distributed? Please justify your answer using QQ-Plots and Normality Tests.

0

5

10

## Histogram of Returns

Density

15

20

25

−1.02 −1.00 −0.98 −0.96

Returns

##### Hypothesis:

*H*0 : The returns are normally distributed.

*H*1 : The returns are not normally distributed.

Shapiro-Wilk normality test data: returns

W = 0.98581, p-value = 0.3695

Since the p-value ≮ 0.05, we do not reject *H*0 and conclude that the returns are normally distributed. Also, the histogram of returns follows normal distribution.

# Question 2

Repeat the same from (a) to (h) on the **cheddar** dataset (except part (i)) form the book by fitting a model with **taste** as the response and the other three variables as predictors. Answer the questions posed in the first question.

Table 4: Sample of Cheddar Data

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | taste | Acetic | H2S | Lactic |
| 12.3 | 4.543 | 3.135 | 0.86 |
| 20.9 | 5.159 | 5.043 | 1.53 |
| 39 | 5.366 | 5.438 | 1.57 |
| 47.9 | 5.759 | 7.496 | 1.81 |
| 5.6 | 4.663 | 3.807 | 0.99 |
| 25.9 | 5.697 | 7.601 | 1.09 |
| **2.1 (a)** |  |  |  |  |

Fit a model to explain **taste** in terms of the predictors. Which variables are important, can any of the variables be removed? Please use F-test to justify.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | t value | Pr(>|t|) |
| **(Intercept)** | -28.88 | 19.74 | -1.463 | 0.1554 |
| **Acetic** | 0.3277 | 4.46 | 0.07349 | 0.942 |

**H2S** 3.912 1.248 3.133 0.004247 \* \*

**Lactic** 19.67 8.629 2.28 0.03108 \*

Table 6: Model Summary

Observations Residual Std. Error *R*2 Adjusted *R*2

30 10.13 0.6518 0.6116

## 2.2 (b)

Check the constant variance assumption of the errors.

Residuals vs Fitted



15

12

8

Residuals

10

20

30

0 10 20 30 40 50

−20

−10

0

Fitted values

There is no pattern in the Residuals vs Fitted Values plot. Therefore, the assumption of constant variance of the errors is not violated.

## 2.3 (c)

Check the independentness of the errors assumption.

##### Hypothesis:

*H*0 : There is no correlation among the residuals.

*H*1 : The residuals are auto-correlated.

lag Autocorrelation D-W Statistic p-value 1 0.1692325 1.57513 0.182

Alternative hypothesis: rho != 0

Since, p-value ≮ 0.05, therefore we do not reject *H*0 and conclude that there is no correlation among the residuals.

## 2.4 (d)

Check the normality assumption.

Normal Q−Q



15

12

8

Standardized residuals

1

2

3

−2 −1 0 1 2

−1

0

Theoretical Quantiles

##### Hypothesis:

*H*0 : The residuals are normally distributed.

*H*1 : The residuals are not normally distributed.

Shapiro-Wilk normality test data: residuals(taste\_model)

W = 0.98021, p-value = 0.8312

For the Shapiro Wilk Test, since the p-value ≮ 0.05, we do not reject *H*0 and conclude that the residuals are normally distributed. The same can be observed from the QQ-Plot of the residuals.

## 2.5 (e)

Is non-linearity a problem?

4.5 5.0 5.5 6.0 6.5 1.0 1.4 1.8

taste

Acetic

4.5

5.5

6.5

0 20 40

H2S

0 10 30 50 4 6 8 10

Lactic

1.0

1.6

4 6 8 10

There is no issue of nonlinearity in the data. Clear positive linear relationships can be observed. Also, the point to note here is the problem of multi-collinearity in the data as the predictors are also exhibit linear relationships.

## 2.6 (f)

Check for outliers, compute and plot Cook’s Distance

Residuals vs Leverage

3

0.5



15

12

30

Cook's distance

Standardized residuals

0

1

2

0.00

−2

0.05 0.10 0.15 0.20 0.25

0.5

−1

Leverage

Cook's distance

15

12

30

Cook's distance

0.10

0.15

0 5 10 15 20 25 30

0.00

0.05

Obs. number

## 2.7 (g)

Check for influential points.

# Question 3

The problem is to discover relation between US new house construction starts data (**HOUST**) and macro economic indicators: **GDP**, **CPI**, Population (**POP**). Please download the relevant data from **house.zip** from blackboard. The description for this data can be found in [https://fred.stlouisfed.org/.](https://fred.stlouisfed.org/)

Table 7: Sample of HOUST Data

|  |  |
| --- | --- |
| DATE | HOUST |
| 1975-10-01 | 296.6 |
| 1976-01-01 | 280.8 |
| 1976-04-01 | 439.3 |
| 1976-07-01 | 434.3 |
| 1976-10-01 | 382.9 |
| 1977-01-01 | 367.4 |

Table 8: Sample of GDP Data

|  |  |
| --- | --- |
| DATE | GDP |
| 1976-01-01 | 58.6 |
| 1976-04-01 | 32.4 |
| 1976-07-01 | 33.6 |
| 1976-10-01 | 47.9 |
| 1977-01-01 | 54.1 |
| 1977-04-01 | 67.7 |

Table 9: Sample of CPI Data

|  |  |
| --- | --- |
| DATE | CPI |
| 1976-01-01 | 0.633 |
| 1976-04-01 | 0.5 |
| 1976-07-01 | 0.9 |
| 1976-10-01 | 0.833 |
| 1977-01-01 | 1.067 |
| 1977-04-01 | 1.033 |

Table 10: Sample of POP Data

|  |  |
| --- | --- |
| DATE | POP |
| 1976-01-01 | 462 |
| 1976-04-01 | 562 |
| 1976-07-01 | 579 |
| 1976-10-01 | 510 |
| 1977-01-01 | 529 |
| 1977-04-01 | 617 |

## 3.1 (a)

Data preparation: combine all data into an R dataframe object, and construct dummy or factor variable for 4 quarters. First model is *HOUST ~ GDP + CPI + QUARTER*.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | t value | Pr(>|t|) |  |
| **(Intercept)** | 271.1 | 20.91 | 12.96 | 1.917e-26 | \* \* \* |
| **GDP CPI** | 0.2208  1.847 | 0.1207  9.83 | 1.829  0.1879 | 0.06933 |  |
| **QUARTER2** | 105.3 | 23.54 | 4.475 | 1.477e-05 | \* \* \* |
| **QUARTER3** | 88.29 | 23.45 | 3.764 | 0.0002373 | \* \* \* |
| **QUARTER4** | 30.5 | 23.42 | 1.302 | 0.1948 |  |

0.8512

Table 12: Model Summary

Observations Residual Std. Error *R*2 Adjusted *R*2

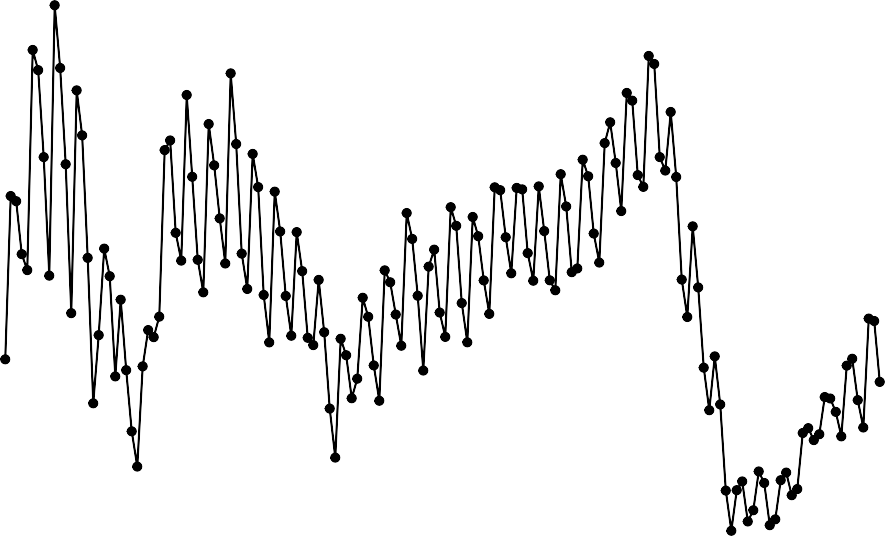
160 104.4 0.1782 0.1515

## 3.2 (b)

Do you think the data needs some cleaning? If so, clean the data.

#### Quarterly Trends in Number of New House Constructions

600



500

400

HOUST

300

200

100

1980

1990

DATE

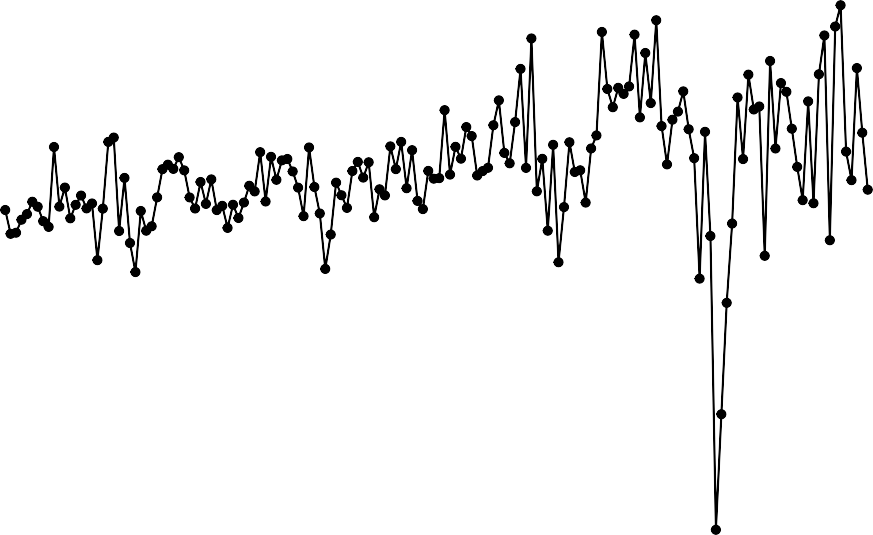
2000

2010

Data Souce: Federal Reserve Bank of St. Loius

#### Quarterly Trends in GDP

300



200

100

0

GDP

−100

−200

−300

1980

1990

DATE

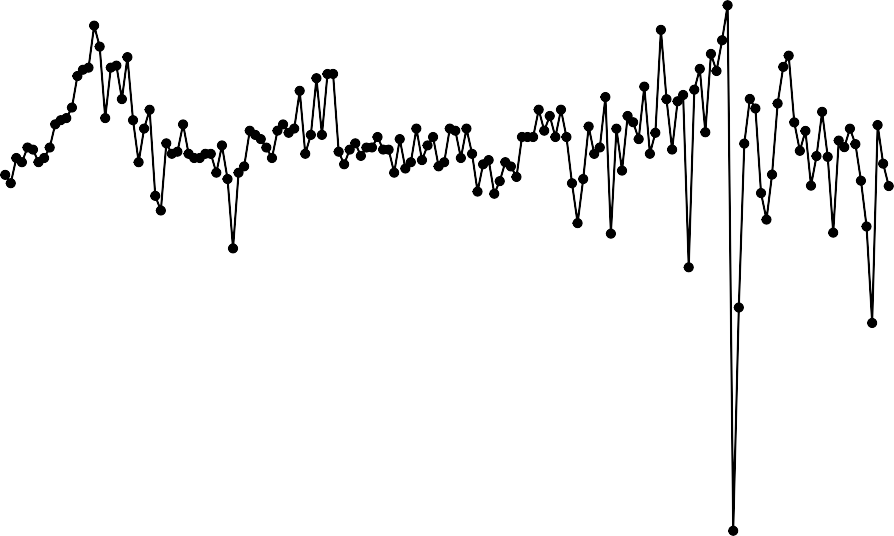
2000

2010

Data Souce: Federal Reserve Bank of St. Loius

#### Quarterly Trends in CPI

2



0

CPI

−2

−4

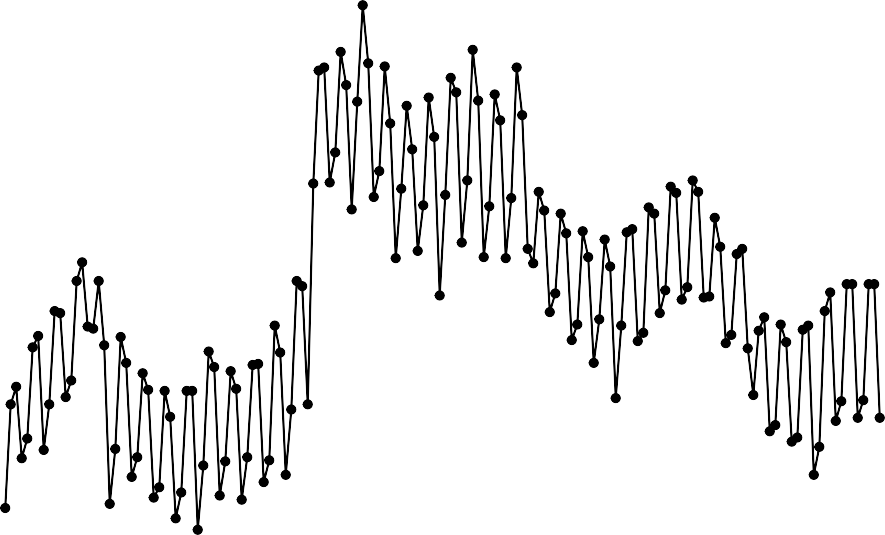
1980 1990 2000 2010

DATE

Data Souce: Federal Reserve Bank of St. Loius

#### Quarterly Trends in Population

900



800

700

POP

600

500

1980 1990 2000 2010

DATE

Data Souce: Federal Reserve Bank of St. Loius

## 3.3 (c)

Is there a seasonal effect you observe in the data? Show necessary steps and explanation. This is an open-ended question and you are free to use any tool that you find appropriate.

## 3.4 (d)

Do a pair-wise comparison for different quarters. Which quarter do you think is the best one to buy a house? Show necessary steps and explanation. Use any statistical test or tool that you think is appropriate, this is an open-ended question and there is no one way of answering this question.

#### Boxplot

600

500

400

HOUST

300

200

100

1 2 3 4

QUARTER

Data Souce: Federal Reserve Bank of St. Loius

## 3.5 (e)

Add population to the first model, do the steps (b) and (c) again.

# Question 4

Read the **train.csv** and **test.csv** files in R which contains training and test data containing information on ten thousand customers. The aim here is to predict which customer will default on their credit card debt. These datasets contains the following information/variables:

* + **default** : A factor with levels **No** and **Yes** indicating whether the customer defaulted on their debt.
  + **student** : A factor with levels **No** and **Yes** indicating whether the customer is a student.
  + **balance** : The average balance that the customer has remaining on their credit card after making their monthly payment.
  + **income** : Income of customer.

Table 13: Sample of Train Data

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| X | default | student | balance | income |
| 1 | No | No | 729.5 | 44362 |
| 3 | No | No | 1074 | 31767 |
| 6 | No | Yes | 919.6 | 7492 |
| 7 | No | No | 825.5 | 24905 |
| 9 | No | No | 1161 | 37469 |
| 10 | No | No | 0 | 29275 |

Table 14: Sample of Test Data

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| X | default | student | balance | income |
| 2 | No | Yes | 817.2 | 12106 |
| 4 | No | No | 529.3 | 35704 |
| 5 | No | No | 785.7 | 38463 |
| 8 | No | Yes | 808.7 | 17600 |
| 11 | No | Yes | 0 | 21871 |
| 13 | No | No | 237 | 28252 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | z value | Pr(>|z|) |  |
| **(Intercept)** | -11.73 | 0.5768 | -20.34 | 5.668e-92 | \* \* \* |
| **balance** | 0.005787 | 0.0003028 | 19.12 | 1.883e-81 | \* \* \* |
| **income** | 1.719e-05 | 6.466e-06 | 2.658 | 0.007854 | \* \* |

(Dispersion parameter for binomial family taken to be 1 )

Null deviance: 1723.0 on 6046 degrees of freedom Residual deviance: 901.6 on 6044 degrees of freedom

## 4.1 (b)

[1] 907.6183

## 4.2 (c)

##### Interpretation:

|  |  |  |  |
| --- | --- | --- | --- |
| (Intercept) | | balance | income |
| -11.7318 | | 0.0058 | 0.0000 |
| [1] | -11.7318 | | |
| [1] | 0.0058 | | |
| [1] | 0 | | |

* + If all the other factors are considered to be null or zero, the average.

•

## 4.3 (d)

y

x No Yes No 3805 10

Yes 99 39

[1] 97.2426

## 4.4 (e)

(Intercept) balance income

-1.173178e+01 5.787483e-03 1.718741e-05

[1] 0.6296417

## 4.5 (f)

2000

0.0 0.2 0.4 0.6 0.8 1.0



train$balance

0

500 1000

train$student

[1] 0.2024676

## 4.6 (g)

Estimate Std. Error z value Pr(>|z|)

**(Intercept)** -10.91 0.6481 -16.83 1.473e-63 \* \* \*

**balance** 0.005907 0.0003102 19.04 7.896e-81 \* \* \*

**income** -5.013e-06 1.079e-05 -0.4647 0.6421

**student** -0.8095 0.3133 -2.584 0.009778 \* \*

(Dispersion parameter for binomial family taken to be 1 )

Null deviance: 1723 on 6046 degrees of freedom Residual deviance: 895 on 6043 degrees of freedom

## 4.7 (h)

From the coefficient of the variable **student** in the model, we can conclude that it is more likely for a student to default compared to a non-student for different values of income level.

# Question 5

These days, there are a lot of discussions about what should the healthcare system look like in US. For a scientific discussion, one should need to have a model of demand in the healthcare system. In this question, we will work on the dataset **dvisit** which is about modeling the demand for doctor visits in terms of explanatory variables such as **age**, **income**, **existence** of health insurance and others. To load this dataset, we type in the commands:

which downloads the library **faraway**, loads this library and then pulls up the dataset **dvisits** from this li- brary. The information about this dataset from the Faraway package can be found at the following document: <https://cran.r-project.org/web/packages/faraway/faraway.pdf>

The following exercise is about fitting a model to data and checking diagnostics of it, making sure that our model is right.

##### Hints:

In class, we provide solutions in R to a similar problem but for a different dataset. I will also give many hints in the class for doing the homework, we will go over the homework questions together.

## 5.1 (a)

Using the **dvisits** dataset, fit a model with the **hospdays** as the response and other variables as potential predictors. Make sure that variables in your model are significant. Note that there is no single perfect model for this dataset, you can do your best for the fit. We can accept any model as long as your variables are reasonably significant and you can justify the variables in your model in words about why/how they should be predictive. We will accept all the models as long as they do not have any serious flaws in them, so feel free to be creative and do not be afraid about playing with variables. Perform regression disgnostics on this model to answer the following questions. Display any plots that are relevant.

Table 19: Sample of Dvisits Data (continued below)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| sex | age | agesq | income | levyplus | freepoor | freerepa | illness |
| 1 | 0.19 | 0.0361 | 0.55 | 1 | 0 | 0 | 1 |
| 1 | 0.19 | 0.0361 | 0.45 | 1 | 0 | 0 | 1 |
| 0 | 0.19 | 0.0361 | 0.9 | 0 | 0 | 0 | 3 |
| 0 | 0.19 | 0.0361 | 0.15 | 0 | 0 | 0 | 1 |
| 0 | 0.19 | 0.0361 | 0.45 | 0 | 0 | 0 | 2 |
| 1 | 0.19 | 0.0361 | 0.35 | 0 | 0 | 0 | 5 |

Table 20: Table continues below

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| actdays | hscore | chcond1 | chcond2 | doctorco | nondocco | hospadmi |
| 4 | 1 | 0 | 0 | 1 | 0 | 0 |
| 2 | 1 | 0 | 0 | 1 | 0 | 0 |
| 0 | 0 | 0 | 0 | 1 | 0 | 1 |
| 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| 5 | 1 | 1 | 0 | 1 | 0 | 0 |
| 1 | 9 | 1 | 0 | 1 | 0 | 0 |

|  |  |  |  |
| --- | --- | --- | --- |
| hospdays | medicine | prescrib | nonpresc |
| 0 | 1 | 1 | 0 |
| 0 | 2 | 1 | 1 |
| 4 | 2 | 1 | 1 |
| 0 | 0 | 0 | 0 |
| 0 | 3 | 1 | 2 |
| 0 | 1 | 1 | 0 |

First, we’ll try to fit the model including all the variables as predictors.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | t value | Pr(>|t|) |  |
| **(Intercept)** | 1.435 | 0.5317 | 2.699 | 0.006969 | \* \* |
| **sex** | -0.3851 | 0.1606 | -2.398 | 0.01654 | \* |
| **age** | -8.624 | 3.019 | -2.856 | 0.004303 | \* \* |
| **agesq** | 10.75 | 3.384 | 3.178 | 0.001493 | \* \* |
| **income** | -0.1431 | 0.2432 | -0.5885 | 0.5562 |  |
| **levyplus freepoor** | -0.003811  -0.2492 | 0.183  0.3855 | -0.02083  -0.6464 | 0.9834 |  |
| **freerepa** | 0.5935 | 0.2808 | 2.113 | 0.03462 | \* |
| **illness** | -0.0226 | 0.06429 | -0.3515 | 0.7252 |  |
| **actdays** | 0.08281 | 0.02946 | 2.811 | 0.004957 | \* \* |
| **hscore chcond1** | 0.04221  0.2113 | 0.03829  0.1761 | 1.102  1.2 | 0.2704 |  |
| **chcond2** | 0.8815 | 0.27 | 3.265 | 0.001102 | \* \* |
| **doctorco** | -0.1747 | 0.1045 | -1.672 | 0.09457 |  |
| **nondocco** | 0.4807 | 0.07876 | 6.103 | 1.118e-09 | \* \* \* |
| **hospadmi** | 5.531 | 0.1536 | 36 | 1.69e-253 | \* \* \* |
| **medicine** | -0.05945 | 0.1053 | -0.5647 | 0.5723 |  |
| **prescrib** | 0.1525 | 0.1184 | 1.289 | 0.1976 |  |

0.518

0.2302

Table 23: Model Summary

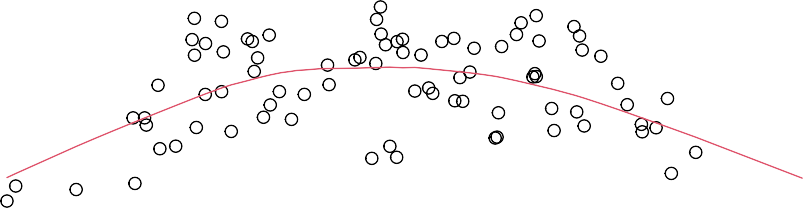
Observations Residual Std. Error *R*2 Adjusted *R*2

5190 5.243 0.2686 0.2662

## 5.2 (b)

Why is your model a good/reasonable model? Check the constant variance assumption for the errors.

Residuals vs Fitted



59

42

85

Residuals

0.000

0.005

1.13 1.14 1.15 1.16 1.17 1.18

−0.005

Fitted values

## 5.3 (c)

Check the normality assumption.

## 5.4 (d)

Are the errors correlated?

## 5.5 (e)

Check for leverage points, outliers, influential points.

## 5.6 (f)

Check the structure of the relationship between the predictors and the response.

Our PhD students (and Masters students who are interested in doing academic research) can also check out the following research article about this dataset to get more information about economics of healthvare and potential research topics in this direction. However, this material is completely optional, not required for this class (but provided for students interested in research).

# Question 6

The following data provides the COVID-19 cases per state since January: [https://covidtracking.com/api/](https://covidtracking.com/api/v1/states/daily.csv) [v1/states/daily.csv](https://covidtracking.com/api/v1/states/daily.csv)

The purpose is to predict “the total number of cases in US per day” with linear regression. Please use the data till the end of September for training and the rest for testing. Perform diagnostics on your model and show that your model is a good model. This is a harder question (such as the optional homework 3), so a “perfect model” may not exist; the purpose is to do “our best”.