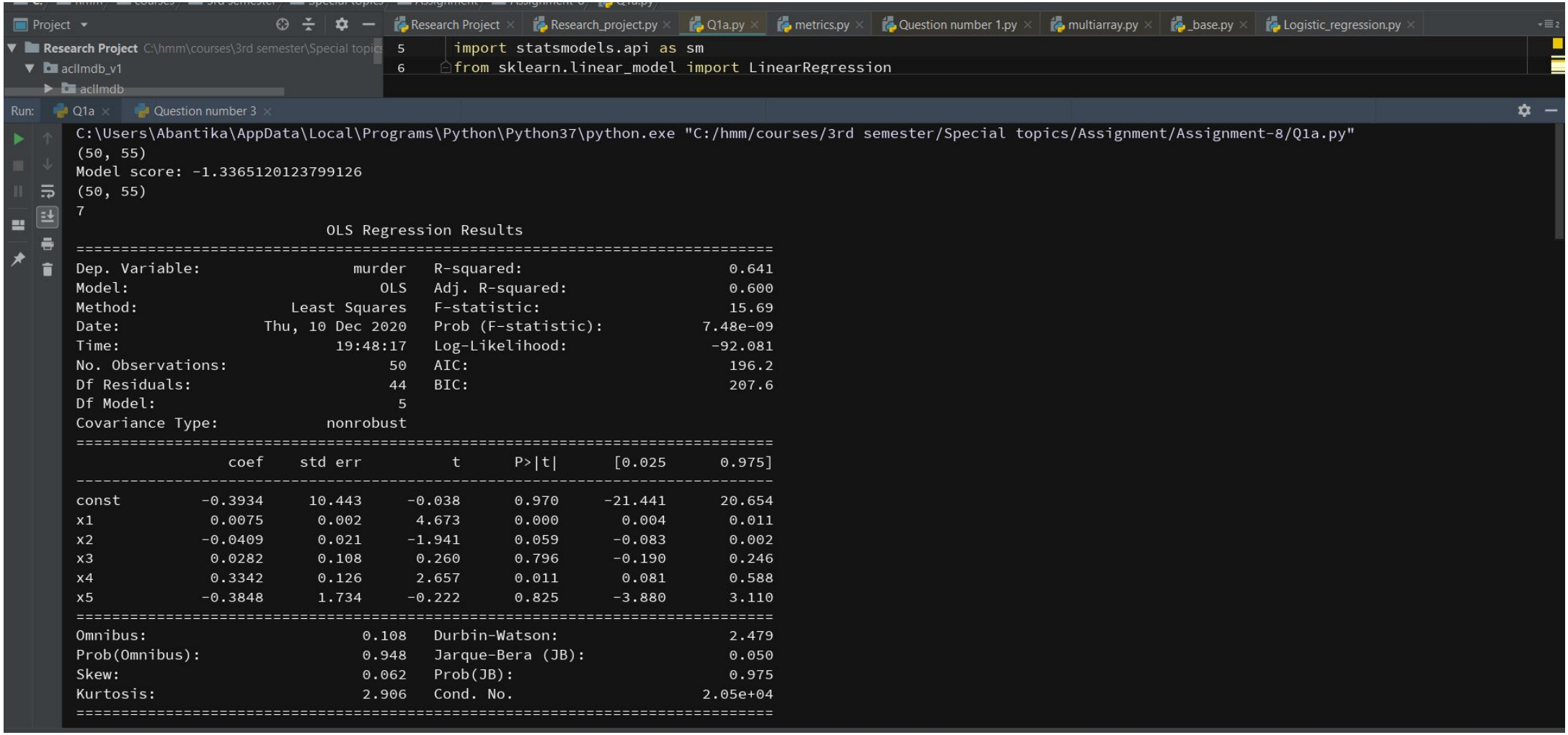
DFSC 5340.02 Assignment 8

Due: Thursday December 10th@11:59PM
Total Points: 110

1

1. (40 pts) Use software or write Python/Matlab/R script to analyze the Crime2 data file that is attached, excluding the observation for D.C. Let y = mur-der rate. For the five explanatory variables in that data file (excluding violent crime rate), with $\alpha = 0.10$ in tests,
   (a) Use backward elimination to select a model. Interpret the result.


Project
Research Project C:\hmm\courses\3rd semester\Special topics
aclImdb_v1
aclImdb
Research Project
Research_project.py
Q1a.py
metrics.py
Question number 1.py
multiarray.py
_base.py
Logistic_regression.py
5 import statsmodels.api as sm
6 from sklearn.linear_model import LinearRegression
Run:
Q1a
Question number 3
C:\Users\Abantika\AppData\Local\Programs\Python\Python37\python.exe "C:/hmm/courses/3rd semester/Special topics/Assignment/Assignment-8/Q1a.py"
(50, 55)
Model score: -1.3365120123799126
(50, 55)
7
OLS Regression Results
Dep. Variable: murder R-squared: 0.641
Model: OLS Adj. R-squared: 0.600
Method: Least Squares F-statistic: 15.69
Date: Thu, 10 Dec 2020 Prob (F-statistic): 7.48e-09
Time: 19:48:17 Log-Likelihood: -92.081
No. Observations: 50 AIC: 196.2
Df Residuals: 44 BIC: 207.6
Df Model: 5
Covariance Type: nonrobust
coef std err t P>|t| [0.025 0.975]
const -0.3934 10.443 -0.038 0.970 -21.441 20.654
x1 0.0075 0.002 4.673 0.000 0.004 0.011
x2 -0.0409 0.021 -1.941 0.059 -0.083 0.002
x3 0.0282 0.108 0.260 0.796 -0.190 0.246
x4 0.3342 0.126 2.657 0.011 0.081 0.588
x5 -0.3848 1.734 -0.222 0.825 -3.880 3.110
Omnibus: 0.108 Durbin-Watson: 2.479
Prob(Omnibus): 0.948 Jarque-Bera (JB): 0.050
Skew: 0.062 Prob(JB): 0.975
Kurtosis: 2.906 Cond. No. 2.05e+04

Command Prompt

```
Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.05e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
                                 OLS Regression Results
==============================================================================
Dep. Variable:                 murder   R-squared (uncentered):                   0.921
Model:                            OLS   Adj. R-squared (uncentered):              0.912
Method:                 Least Squares   F-statistic:                              105.1
Date:                Thu, 10 Dec 2020   Prob (F-statistic):                    1.18e-23
Time:                        18:45:54   Log-Likelihood:                         -92.082
No. Observations:                  50   AIC:                                      194.2
Df Residuals:                      45   BIC:                                      203.7
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1             0.0075      0.001      5.133      0.000       0.005       0.010
x2            -0.0410      0.021     -1.971      0.055      -0.083       0.001
x3             0.0242      0.023      1.066      0.292      -0.021       0.070
x4             0.3306      0.080      4.126      0.000       0.169       0.492
x5            -0.3707      1.675     -0.221      0.826      -3.744       3.002
==============================================================================
Omnibus:                        0.108   Durbin-Watson:                   2.478
Prob(Omnibus):                  0.947   Jarque-Bera (JB):                0.049
Skew:                           0.061   Prob(JB):                        0.976
Kurtosis:                       2.907   Cond. No.                     3.33e+03
==============================================================================

Notes:
[1] R² is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Command Prompt

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 murder   R-squared (uncentered):                   0.921
Model:                            OLS   Adj. R-squared (uncentered):              0.914
Method:                 Least Squares   F-statistic:                              134.1
Date:                Thu, 10 Dec 2020   Prob (F-statistic):                    9.75e-25
Time:                        18:45:54   Log-Likelihood:                         -92.109
No. Observations:                  50   AIC:                                      192.2
Df Residuals:                      46   BIC:                                      199.9
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1             0.0075      0.001      5.203      0.000       0.005       0.010
x2            -0.0404      0.020     -1.980      0.054      -0.081       0.001
x3             0.0235      0.022      1.056      0.296      -0.021       0.068
x4             0.3328      0.079      4.230      0.000       0.174       0.491
==============================================================================
Omnibus:                        0.147   Durbin-Watson:                   2.485
Prob(Omnibus):                  0.929   Jarque-Bera (JB):                0.098
Skew:                           0.094   Prob(JB):                        0.952
Kurtosis:                       2.892   Cond. No.                         158.
==============================================================================

Notes:
[1] R² is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.
                            OLS Regression Results
==============================================================================
Dep. Variable:                 murder   R-squared (uncentered):                   0.919
Model:                            OLS   Adj. R-squared (uncentered):              0.914
Method:                 Least Squares   F-statistic:                              178.0
```

Command Prompt

```
Date:                Thu, 10 Dec 2020   Prob (F-statistic):                1.16e-25
Time:                        18:45:54   Log-Likelihood:                    -92.708
No. Observations:                  50   AIC:                                 191.4
Df Residuals:                      47   BIC:                                 197.2
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1             0.0081      0.001      6.268      0.000       0.006       0.011
x2            -0.0205      0.008     -2.614      0.012      -0.036      -0.005
x3             0.3434      0.078      4.395      0.000       0.186       0.501
==============================================================================
Omnibus:                        0.568   Durbin-Watson:                   2.551
Prob(Omnibus):                  0.753   Jarque-Bera (JB):                0.635
Skew:                           0.233   Prob(JB):                        0.728
Kurtosis:                       2.703   Cond. No.                         155.
==============================================================================

Notes:
[1] R² is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.
                            OLS Regression Results
==============================================================================
Dep. Variable:                murder   R-squared (uncentered):                   0.921
Model:                            OLS   Adj. R-squared (uncentered):              0.914
Method:                 Least Squares   F-statistic:                              134.1
Date:                Thu, 10 Dec 2020   Prob (F-statistic):                    9.75e-25
Time:                        18:45:54   Log-Likelihood:                         -92.109
No. Observations:                  50   AIC:                                      192.2
Df Residuals:                      46   BIC:                                      199.9
Df Model:                           4
Covariance Type:            nonrobust
```

Command Prompt

```
Df Residuals:                       46   BIC:                                199.9
Df Model:                            4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1             0.0075      0.001      5.203      0.000       0.005       0.010
x2            -0.0404      0.020     -1.980      0.054      -0.081       0.001
x3             0.0235      0.022      1.056      0.296      -0.021       0.068
x4             0.3328      0.079      4.230      0.000       0.174       0.491
==============================================================================
Omnibus:                        0.147   Durbin-Watson:                   2.485
Prob(Omnibus):                  0.929   Jarque-Bera (JB):                0.098
Skew:                           0.094   Prob(JB):                        0.952
Kurtosis:                       2.892   Cond. No.                         158.
==============================================================================

Notes:
[1] R² is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.
Parameters:  x1        0.007288
x2       -0.042083
const    20.376305
x3       -0.174955
dtype: float64
R2:  0.5790527884801249
exog_names=  ['x1', 'x2', 'const', 'x3']
endog_names=  murder
```

(b) Use forward selection to select a model. Interpret the result.

b

```
Training dataset shape: (40, 55) (40,)
Testing dataset shape: (10, 55) (10,)
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
[Parallel(n_jobs=1)]: Done   1 out of   1 | elapsed:    2.5s remaining:    0.0s
[Parallel(n_jobs=1)]: Done  55 out of  55 | elapsed:    6.9s finished

[2020-12-10 18:34:45] Features: 1/3 -- score: 0.275[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
[Parallel(n_jobs=1)]: Done   1 out of   1 | elapsed:    0.0s remaining:    0.0s
[Parallel(n_jobs=1)]: Done  54 out of  54 | elapsed:    4.2s finished

[2020-12-10 18:34:50] Features: 2/3 -- score: 0.325[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
[Parallel(n_jobs=1)]: Done   1 out of   1 | elapsed:    0.0s remaining:    0.0s
[Parallel(n_jobs=1)]: Done  53 out of  53 | elapsed:    4.1s finished

[2020-12-10 18:34:54] Features: 3/3 -- score: 0.35features were selected:  [1, 2, 54]
```

(c) Compare results of the two selection procedures. How is it possible that a variable (percentage with a high school education) can be the first variable dropped in (a) yet the second added in (b)?

A feature selection procedure can keep out important and necessary variables. p-value provides a guide for making decisions about adding or dropping features. They are not actual true p-values for the tests conducted due to the sampling distribution of t or F statistics differing from the sampling distribution for a given chosen test. Any feature selection method should be used with caution.

(d) Now include the D.C. observation. Repeat (a) and (b) and compare

(d) Now include the D.C. observation. Repeat (a) and (b). and compare to results excluding D.C. What does this suggest about the influence outliers can have on automatic selection procedures?

OLS Regression Results

Dep. Variable:	murder	R-squared:	0.819
Model:	OLS	Adj. R-squared:	0.799
Method:	Least Squares	F-statistic:	40.67
Date:	Thu, 10 Dec 2020	Prob (F-statistic):	1.33e-15
Time:	18:26:06	Log-Likelihood:	-120.11
No. Observations:	51	AIC:	252.2
Df Residuals:	45	BIC:	263.8
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-53.4908	14.394	-3.716	0.001	-82.483	-24.499
x1	0.0164	0.002	7.667	0.000	0.012	0.021
x2	-0.0655	0.035	-1.876	0.067	-0.136	0.005
x3	0.5721	0.150	3.810	0.000	0.270	0.875
x4	0.8466	0.187	4.520	0.000	0.469	1.224
x5	-3.9522	2.818	-1.403	0.168	-9.628	1.723

Omnibus:	4.493	Durbin-Watson:	1.985
Prob(Omnibus):	0.106	Jarque-Bera (JB):	5.189
Skew:	0.048	Prob(JB):	0.0747
Kurtosis:	4.560	Cond. No.	1.89e+04

Command Prompt

```
strong multicollinearity or other numerical problems.
                            OLS Regression Results
==============================================================================
Dep. Variable:                 murder   R-squared:                       0.811
Model:                            OLS   Adj. R-squared:                  0.794
Method:                 Least Squares   F-statistic:                     49.31
Date:                Thu, 10 Dec 2020   Prob (F-statistic):           4.55e-16
Time:                        18:26:06   Log-Likelihood:                -121.20
No. Observations:                  51   AIC:                             252.4
Df Residuals:                      46   BIC:                             262.1
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        -51.3184     14.461     -3.549      0.001     -80.426     -22.211
x1             0.0161      0.002      7.497      0.000       0.012       0.020
x2            -0.0612      0.035     -1.741      0.088      -0.132       0.010
x3             0.5427      0.150      3.612      0.001       0.240       0.845
x4             0.8516      0.189      4.500      0.000       0.471       1.232
==============================================================================
Omnibus:                        5.419   Durbin-Watson:                   1.899
Prob(Omnibus):                  0.067   Jarque-Bera (JB):                6.709
Skew:                           0.203   Prob(JB):                       0.0349
Kurtosis:                       4.730   Cond. No.                     1.88e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.88e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
                            OLS Regression Results
==============================================================================
```

2. (30 pts) For a data set for 100 adults on y = height, x_1 = length of left leg, and x_2 = length of right leg, the model $E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2$ is fitted. Neither H_0: $\beta_1 = 0$ nor H_0: $\beta_2 = 0$ has a P-value below 0.05.
(a) Does this imply that length of leg is not a good predictor of height? Why?

x_1 = length of left leg
x_2 = length of right leg

$E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2$

Here, we can see x_1 and x_2 both are related with height(y). So, if we have one of the two variables in the model, then adding the other variable will not change anything. Individually each partial co-efficient may not be significant.

(b) Does this imply that H_0: $\beta_1 = \beta_2 = 0$ would not have a P-value below 0.05? Why? 6

$H_0: B_1 = B_2 = 0$ would probably have a small P-value.

The correlations between the explanatory variables and y are close to 0. Also, R^2 is close to 0. So, we can assume $H_0: B_1 = B_2 = 0$ also would have a small P-value.

7

(c) Suppose $r_{yx1} = 0.901$, $r_{yx2} = 0.902$, and $r_{x1x2} = 0.999$. Using forward selection and the potential predictors x_1 and x_2 with $\alpha = 0.05$ for tests, which model would you expect to be selected? Why?

$r_{yx_1x_2} =$

$\frac{r_{yx_1} - r_{yx_2} r_{x_1x_2}}{\sqrt{(1 - r^2_{yx_2})(1 - r^2_{x_1x_2})}}$

$= \frac{\cdot 98 \times 10^{-5}}{\cdot 0192}$

$= -\cdot 2964$

I expect to be selected x_2. Here, we can see x_2 and y have higher correlation. So in that case, we would select x_2 first. After that if we add x_1 in the model then it would not bring any change in the model. So, I would expect to be selected, model using with x_2.

$r_{yx_1} = \cdot 901$
$r_{yx_2} = \cdot 902$
$r_{x_1x_2} = 0 \cdot 999$

8

3. (40 pts) For white men in the United States, the following Table presents the number of deaths per thousand individuals of a fixed age within a period of a year.

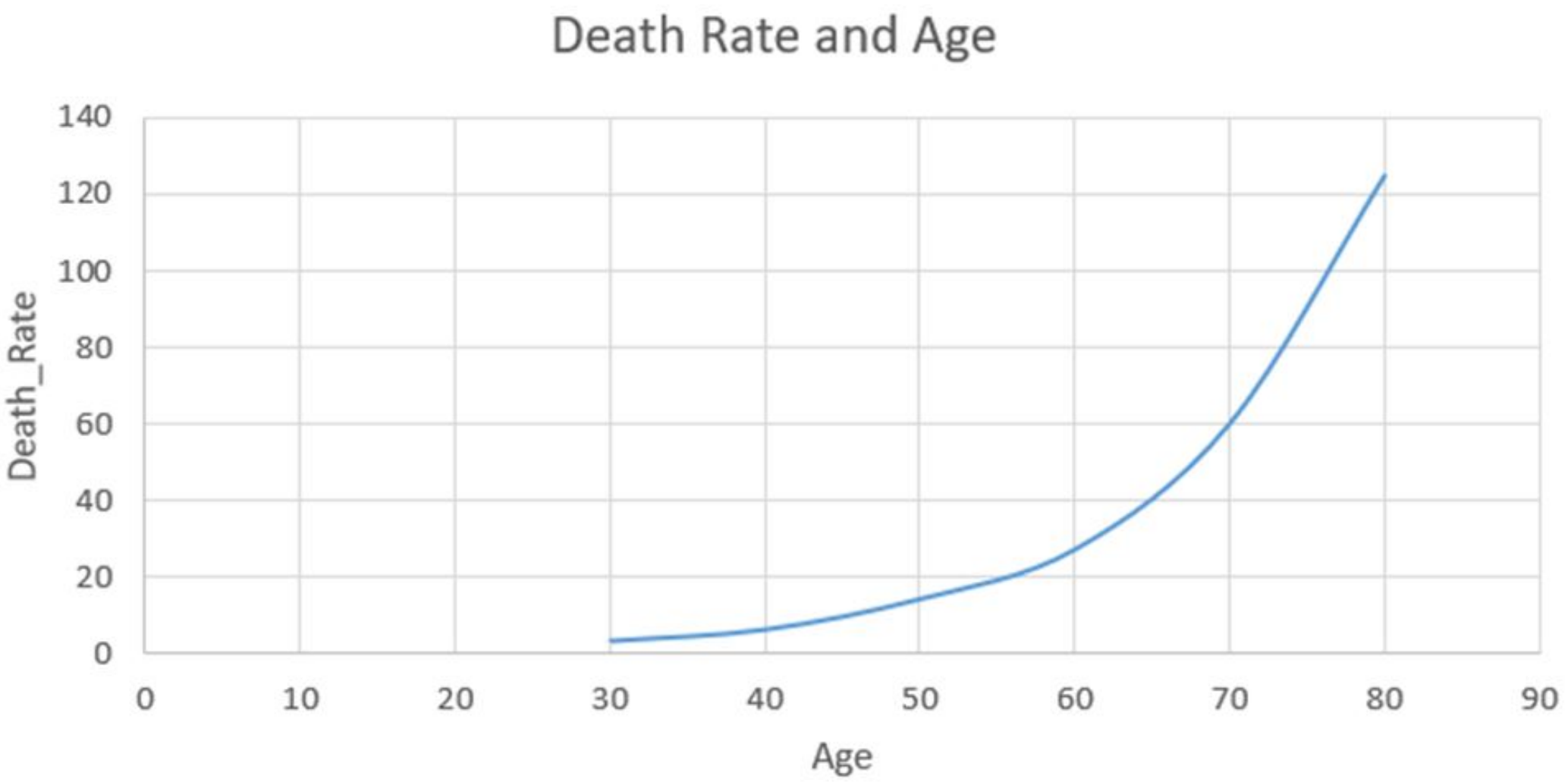
Age	Death Rate (Per Thousand)

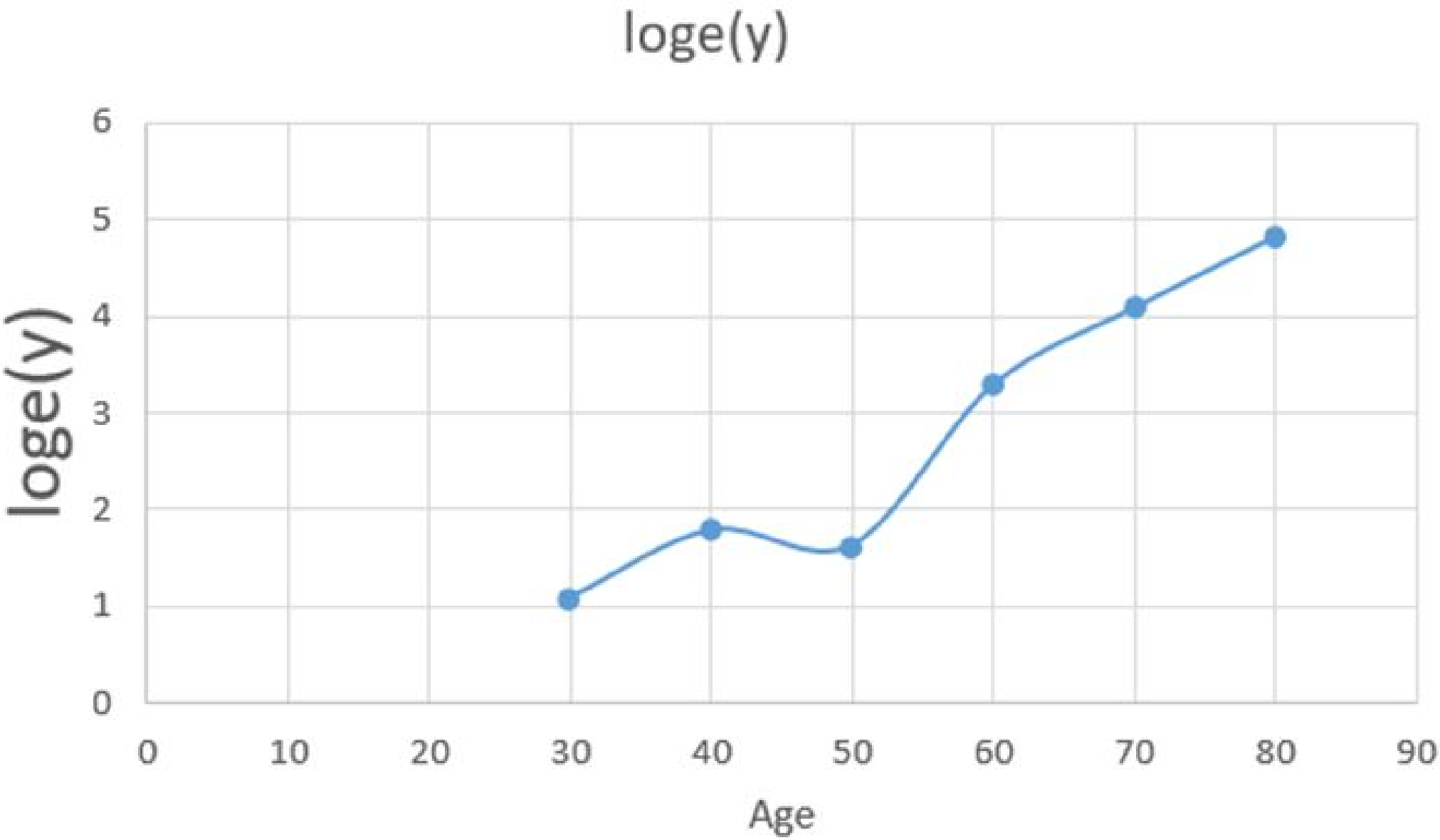
using with x_2.

3. (40 pts) For white men in the United States, the following Table presents the number of deaths per thousand individuals of a fixed age within a period of a year. 8

Age	Death Rate (Per Thousand)
30	3
40	6
50	14
60	27
70	60
80	125

(a) Plot x = age against y = death rate and against log y. What do these plots suggest about a good model for the relationship?


Death Rate and Age
Death_Rate
140
120
100
80
60
40
20
0
0
10
20
30
40
50
60
70
80
90
Age


loge(y)
loge(y)
6
5
4
3
2
1
0
0
10
20
30
40
50
60
70
80
90
Age

9

(b) Find the correlation between (i) x and y, (ii) x and log(y). What do these suggest about an appropriate model?

	x	y	log y
x	1·0	·89	·3936
y	·89	1·0	·89
log y	0·39	0·89	1·0

(i) x and y correlation = ·89

(ii) x and log(y) = ·3936

Here, we can see the correlation between x and log(y) is lower than x and y correlation.

An appropriate model could be exponential regression model.

10

(c) Using generalized linear models, find the prediction equation for the model $\log[F(y)] = \alpha + \beta x$.

$$\log e(\hat{u}) = -1·1452 + 0·0747x$$

```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:             Death_Rate   No. Observations:                    6
Model:                            GLM   Df Residuals:                        4
Model Family:                 Poisson   Df Model:                            1
Link Function:                    log   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -14.491
Date:                Thu, 10 Dec 2020   Deviance:                      0.10117
Time:                        17:27:25   Pearson chi2:                    0.101
No. Iterations:                     5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -1.1578      0.417     -2.778      0.005      -1.975      -0.341
Age            0.0748      0.006     13.038      0.000       0.064       0.086
==============================================================================
Coefficeients
Intercept   -1.157836
Age          0.074850
dtype: float64

p-Values
Intercept    5.467777e-03
Age          7.440952e-39
dtype: float64

Dependent variables
Death_Rate
```

(d) Find the prediction equation for $\hat{y}$. Interpret the parameter estimates.

$$\hat{y} = \hat{a}\hat{\beta}^{x}$$
$$= e^{-1.1453}\left(e^{0.0747}\right)^{x}$$
$$= .318\left(1.078\right)^{x}$$

We can see the predicted death rate at age $x+1$ equals 107.8% of the predicted death rate at age x.

The death rate increases by 7.8% for each addit year of age.