Problem Set 5

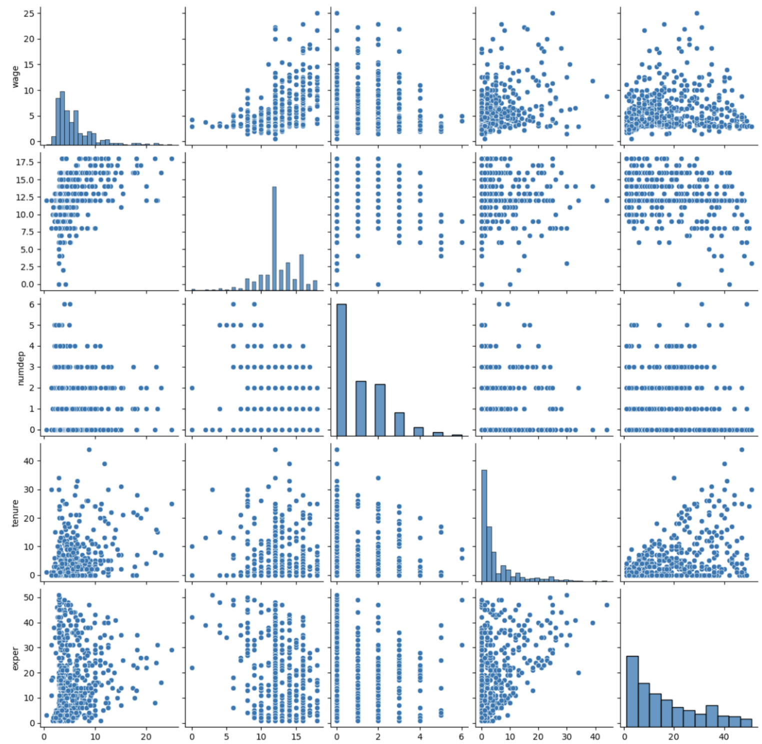
Justin Chen

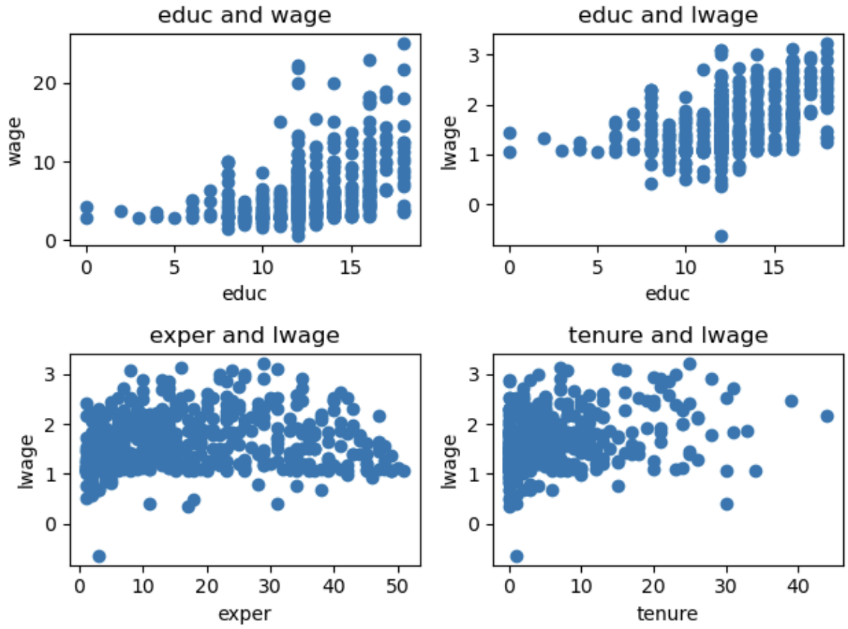
Exercise 1: Wage

1.1

Data is checked for null values. The code is attached in the end.

1.2





1.3

I think OLS Regression is more suitable for understanding what factors explain variability in wage. The graphs of education against wages, education against lwages, experience or tenure against lwages do show a linear relationship.

1.4

$$wage = \beta_0 + \beta_1 educ_i + \beta_2 exper_i + \beta_3 tenure_i + \beta_4 married_i + \beta_5 female_i + \epsilon_i$$

1.5

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                   wage   R-squared:                       0.416
Model:                            OLS   Adj. R-squared:                  0.408
Method:                 Least Squares   F-statistic:                     52.70
Date:                Fri, 18 Nov 2022   Prob (F-statistic):           1.20e-56
Time:                        10:38:28   Log-Likelihood:                -1291.6
No. Observations:                 526   AIC:                             2599.
Df Residuals:                     518   BIC:                             2633.
Df Model:                           7
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -0.3094      0.741     -0.418      0.676      -1.765       1.146
educ           0.3908      0.056      7.016      0.000       0.281       0.500
exper          0.0101      0.012      0.864      0.388      -0.013       0.033
tenure         0.1354      0.020      6.646      0.000       0.095       0.175
married        0.6104      0.276      2.215      0.027       0.069       1.152
female        -1.5728      0.260     -6.059      0.000      -2.083      -1.063
profocc        1.7488      0.304      5.748      0.000       1.151       2.347
west           1.0253      0.331      3.097      0.002       0.375       1.676
==============================================================================
Omnibus:                      174.906   Durbin-Watson:                   1.848
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              672.881
Skew:                           1.485   Prob(JB):                    7.69e-147
Kurtosis:                       7.678   Cond. No.                         151.
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

1.6

An additional year of education increases the wage by 0.3908; an additional year of potential experience increases the wage by 0.01; an addition year with current employer increases the wage by 0.135; if married, the wage increases by 0.61; if the person is female, the wage decreases by 1.573; if the person works in professional occupation, the wage increases by 1.749 and if the person lives in west region, the wage increases by 1.025

The regression shows that the coefficients for educ, tenure and whether the person is female are statistically significant at $\alpha = 0.05$

1.7

The model explains 41.6% of the variation of wages for different individuals.

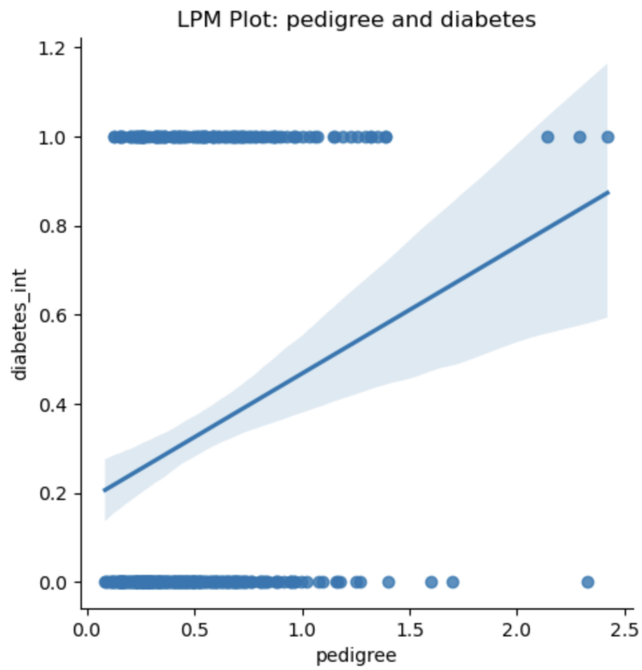
1.8

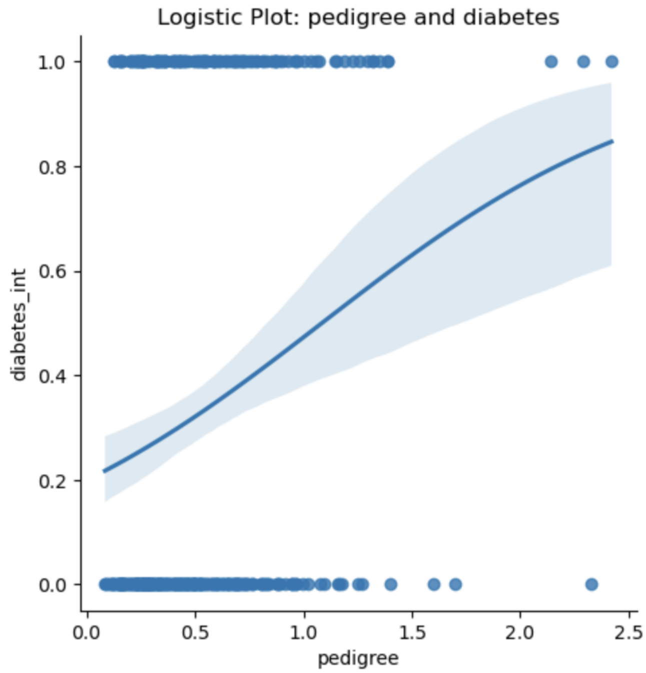
A person with 281 years of education, 270 years of potential experience, 255 years with current employer, married, male, works in professional occupation and lives in western region can have hourly wages of $150.

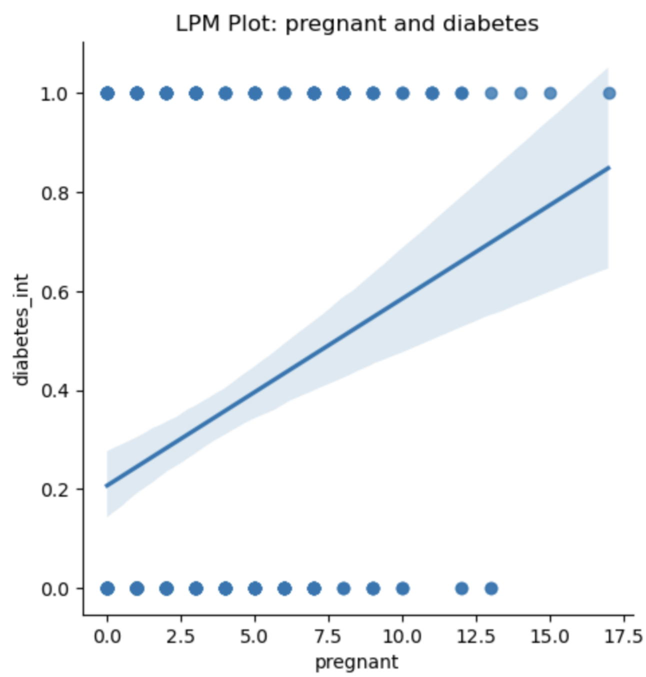
2.1

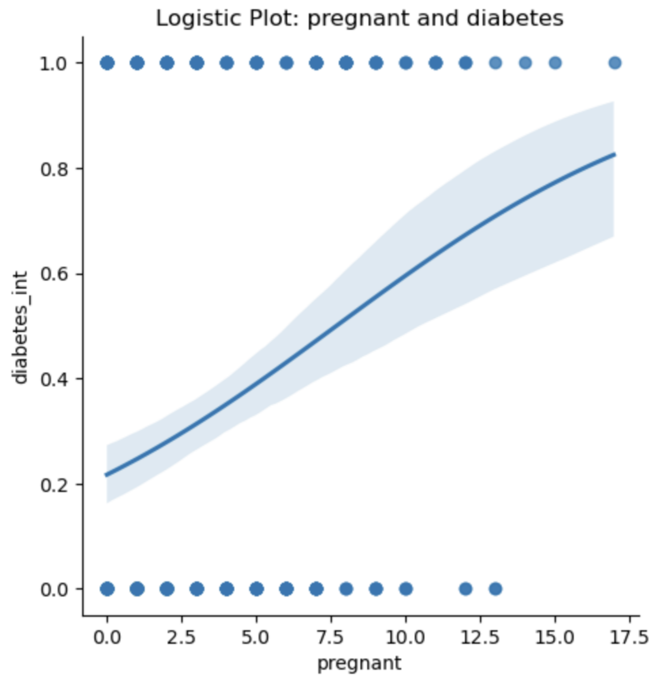
Data is checked for null values. The code is attached in the end.

2.2

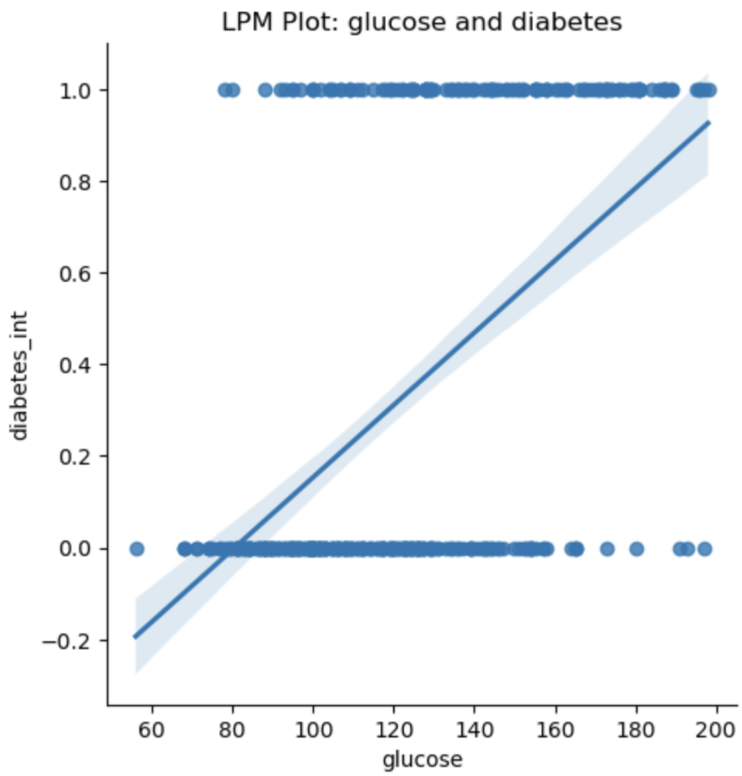



Logistic Plot: pedigree and diabetes
diabetes_int
pedigree
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.5
1.0
1.5
2.0
2.5

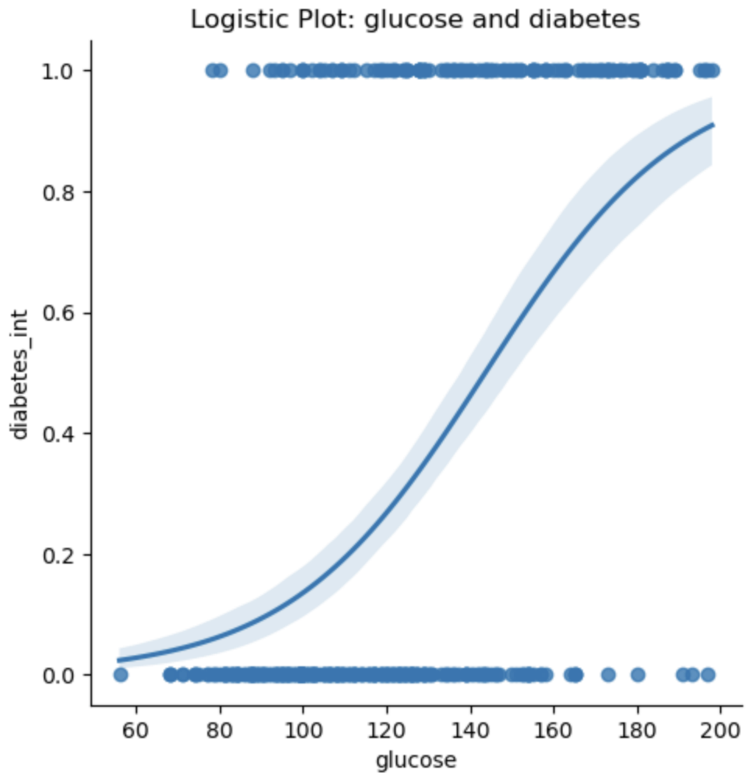

LPM Plot: pregnant and diabetes
diabetes_int
pregnant
0.0
0.2
0.4
0.6
0.8
1.0
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5

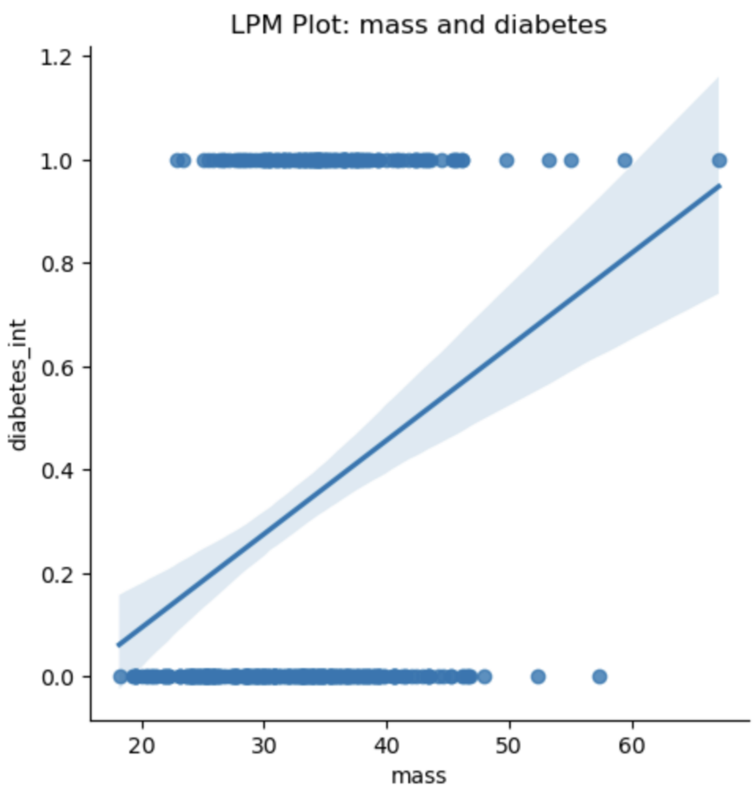


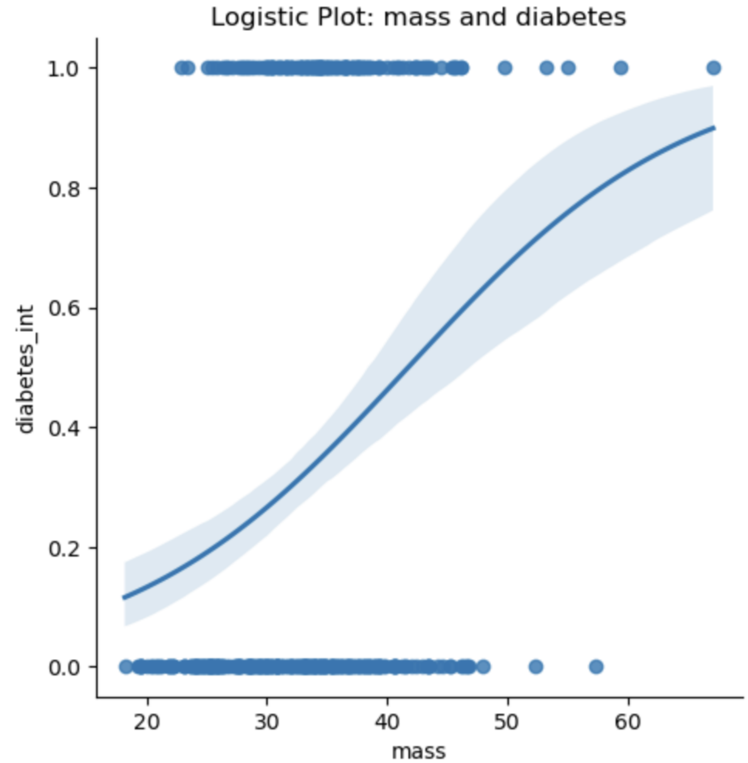
Logistic Plot: pregnant and diabetes
1.0
0.8
0.6
0.4
0.2
0.0
diabetes_int
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
pregnant



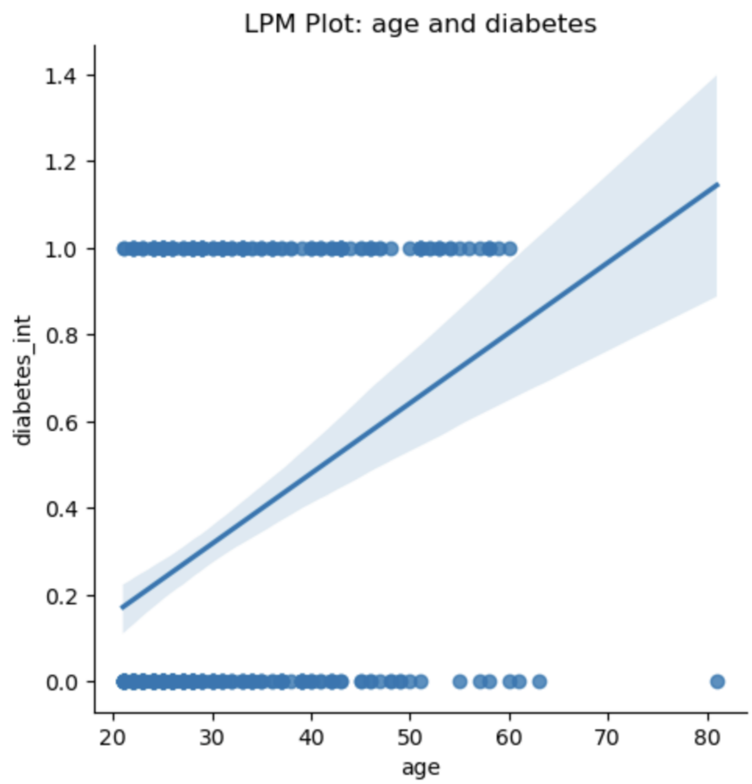
LPM Plot: glucose and diabetes
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
diabetes_int
60
80
100
120
140
160
180
200
glucose


Logistic Plot: glucose and diabetes
diabetes_int
1.0
0.8
0.6
0.4
0.2
0.0
60
80
100
120
140
160
180
200
glucose

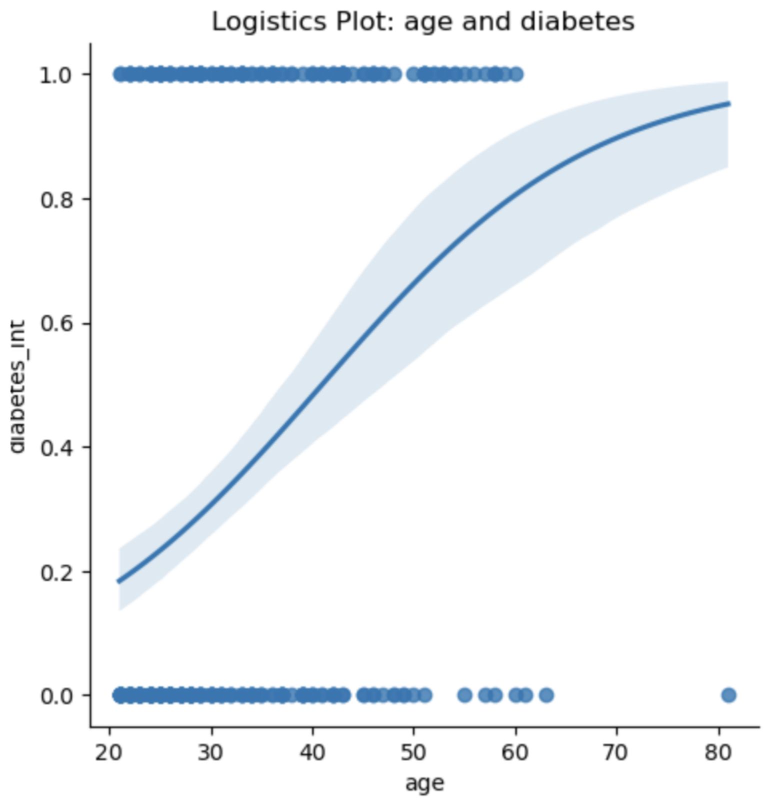

LPM Plot: mass and diabetes
diabetes_int
1.2
1.0
0.8
0.6
0.4
0.2
0.0
20
30
40
50
60
mass



Logistic Plot: mass and diabetes
1.0
0.8
0.6
0.4
0.2
0.0
diabetes_int
20
30
40
50
60
mass



LPM Plot: age and diabetes
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
diabetes_int
20
30
40
50
60
70
80
age



2.3

A Logistic Regression is more suitable. The dependent variable is a 0 or 1 binary outcome. The regression should be predicting the probability.

2.4

$$pr(diabetes = 1|X, \beta) = \frac{e^{\beta_0 + \beta_1 pregnant + \beta_2 glucose + \beta_3 mass + \beta_4 pedigree + \beta_5 age}}{1 + e^{\beta_0 + \beta_1 pregnant + \beta_2 glucose + \beta_3 mass + \beta_4 pedigree + \beta_5 age}}$$

2.5

```
Optimization terminated successfully.
         Current function value: 0.439905
         Iterations 7
                           Logit Regression Results
==============================================================================
Dep. Variable:           diabetes_int   No. Observations:                  392
Model:                          Logit   Df Residuals:                      386
Method:                           MLE   Df Model:                            5
Date:                Fri, 18 Nov 2022   Pseudo R-squ.:                  0.3076
Time:                        11:44:37   Log-Likelihood:                -172.44
converged:                       True   LL-Null:                       -249.05
Covariance Type:            nonrobust   LLR p-value:                 2.764e-31
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -9.9921      1.087     -9.193      0.000     -12.122      -7.862
pregnant       0.0840      0.055      1.526      0.127      -0.024       0.192
glucose        0.0365      0.005      7.324      0.000       0.027       0.046
mass           0.0781      0.021      3.792      0.000       0.038       0.119
pedigree       1.1509      0.424      2.713      0.007       0.319       1.982
age            0.0344      0.018      1.929      0.054      -0.001       0.069
==============================================================================
```

2.6

- When other variables are 0, the odds ratio of having diabetes is $e^{-9.9921} = 4.576000965045239e-05$
- A unit increase of pregnant times increases the log odds of having diabetes by 0.084;
- A unit increase of plasma glucose concentration increases the log odds of having diabetes by 0.0365;
- A unit increase of body mass index increases the log odds of having diabetes by 0.0781;
- A unit increase of pedigree increases the log odds of having diabetes by 1.1509;
- A unit increase of age increases the log odds of having diavetes by 0.0344

2.7

Probability of the patient with median value of the variables to have diabetes is 0.1906

Probability of the patient with 75th percentile value of the variables to have diabetes is 0.638

Probability of the patient with 25th percentile value of the variables to have diabetes is 0.048

	pregnant	glucose	mass	pedigree	age
count	392.000000	392.000000	392.000000	392.000000	392.000000
mean	3.301020	122.627551	33.086224	0.523046	30.864796
std	3.211424	30.860781	7.027659	0.345488	10.200777
min	0.000000	56.000000	18.200000	0.085000	21.000000
25%	1.000000	99.000000	28.400000	0.269750	23.000000
50%	2.000000	119.000000	33.200000	0.449500	27.000000
75%	5.000000	143.000000	37.100000	0.687000	36.000000
max	17.000000	198.000000	67.100000	2.420000	81.000000

```
hypothetical_prediction
```

```
array([0.19055135, 0.63824141, 0.04838718])
```

```
"""
Justin Chen
Problem Set 5
"""

import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import statsmodels.api as sm
import statsmodels.formula.api as smf
from statsmodels.discrete.discrete_model import Logit

# 1.1
w_df = pd.read_csv('wage.csv')
print(w_df.isnull().sum())
```

```
# 1.2
sns.pairplot(w_df[['wage', 'educ', 'numdep', 'tenure', 'exper']])

fig = plt.figure()
ax1 = fig.add_subplot(2, 2, 1)
ax2 = fig.add_subplot(2, 2, 2)
ax3 = fig.add_subplot(2, 2, 3)
ax4 = fig.add_subplot(2, 2, 4)

ax1.scatter(w_df['educ'], w_df['wage'])
ax1.set_title('educ and wage')
ax1.set_xlabel('educ')
ax1.set_ylabel('wage')

ax2.scatter(w_df['educ'], w_df['lwage'])
ax2.set_title('educ and lwage')
ax2.set_xlabel('educ')
ax2.set_ylabel('lwage')

ax3.scatter(w_df['exper'], w_df['lwage'])
ax3.set_title('exper and lwage')
ax3.set_xlabel('exper')
ax3.set_ylabel('lwage')

ax4.scatter(w_df['tenure'], w_df['lwage'])
ax4.set_title('tenure and lwage')
ax4.set_xlabel('tenure')
ax4.set_ylabel('lwage')

plt.show()
plt.tight_layout()

# 1.5
mod = smf.ols(formula='wage ~ educ + exper + tenure + married + female + profocc + west',
                data=w_df)
res = mod.fit()
print(res.summary())

# 1.8
hypothetical = pd.DataFrame({"educ": [281], "exper": [270], "tenure": [255],
"married": [1], "female": [0], "profocc": [1], "west": [1]})
hypothetical_prediction = res.predict(hypothetical)
print(hypothetical_prediction)

# 2.1
d_df = pd.read_csv('diabetes.csv')
print(d_df.isnull().sum())

diabetes_diag = {'neg': 0, 'pos':1}
d_df['diabetes_int'] = d_df['diabetes'].map(diabetes_diag)

# 2.2
sns.lmplot(x='pedigree', y='diabetes_int', data=d_df).set(
```

```
    title='LPM Plot: pedigree and diabetes')
sns.lmplot(x='pedigree', y='diabetes_int', data=d_df, logistic= True).set(
    title='Logistic Plot: pedigree and diabetes')

sns.lmplot(x='pregnant', y='diabetes_int', data=d_df).set(
    title='LPM Plot: pregnant and diabetes')
sns.lmplot(x='pregnant', y='diabetes_int', data=d_df, logistic= True).set(
    title='Logistic Plot: pregnant and diabetes')

sns.lmplot(x='glucose', y='diabetes_int', data=d_df).set(
    title='LPM Plot: glucose and diabetes')
sns.lmplot(x='glucose', y='diabetes_int', data=d_df, logistic= True).set(
    title='Logistic Plot: glucose and diabetes')

sns.lmplot(x='mass', y='diabetes_int', data=d_df).set(
    title='LPM Plot: mass and diabetes')
sns.lmplot(x='mass', y='diabetes_int', data=d_df, logistic= True).set(
    title='Logistic Plot: mass and diabetes')

sns.lmplot(x='age', y='diabetes_int', data=d_df).set(
    title='LPM Plot: age and diabetes')
sns.lmplot(x='age', y='diabetes_int', data=d_df, logistic= True).set(
    title='Logistics Plot: age and diabetes')
plt.show()

# 2.5
dfy = d_df['diabetes_int']
dfx = sm.add_constant(d_df[['pregnant', 'glucose', 'mass', 'pedigree', 'age']])
mod = Logit(dfy, dfx)
res = mod.fit()
print(res.summary())

# 2.7
d_df[['pregnant', 'glucose', 'mass', 'pedigree', 'age']].describe()

hypothetical = [[1, 2, 119, 33.2, 0.4495, 27],
[1, 5, 143, 37.1, 0.687, 36], [1, 1, 99, 28.4, 0.26975, 23]]
hypothetical_prediction = res.predict(hypothetical)
print(hypothetical_prediction)
```