**Problem Set 5**

Exercise 1

* Part 2

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

* Part 3
  + I think using OLS Regression is more suitable in understanding the factors that explain variability in wages because wage/logged wage is the response variable (LHS) of a regression model, and it is continuous. Logit regression is also used when the response variable has multiple outcomes, which isn't really applicable here.
* Part 4
  + Since we are using wages as the dependent variable on the left-hand side and we are trying to find the correlation between wages and other variables, I've created a model that checks the relationship between wage and all the continuous variables in the wage dataset, as well as two binary variables. I put those variables in an array and used the sm approach.
  + LHS: wage; RHS: education, experience, tenure, nonwhite, female
* Part 5

Table

Description automatically generated

* Part 6
  + Education, experience, and tenure all have positive correlations with wages, and they're all statistically significant because their p-values are all lower than 0.05. Variable female has a negative correlation with wage and it's also statistically significant because of its low p-value. The only nonsignificant variable is nonwhite.
* Part 7
  + Since the OLS regression yields an R-squared value of 0.364, this means that 36.4% of variation in the wage dataset can be explained by the variables educ, exper, tenure, nonwhite and female.
* Part 8
  + "wage": [150], "exper": [100], "tenure": [50], “educ": [100], "female": [0], "nonwhite": [0]})

Exercise 2

* Chart

  Description automatically generatedGraphical user interface

  Description automatically generatedPart 2

Chart

Description automatically generated

* Part 3
  + Logistic Regression is more suitable to analyze this problem because diabetes is a binary variable, meaning there are two outcomes, and each outcome has a probability between 0 and 1.
* Part 4
  + Since we are using diabetes as the dependent variable on the left-hand side and we are trying to find the correlation between diabetes and other variables, I've created a model that checks the relationship between diabetes and all the variables in the diabetes dataset. I put those variables in an array and used the sm approach.
  + LHS: diabetes, RHS: pregnant, glucose, pressure, triceps, insulin, mass, pedigree, age
* Part 5

Table

Description automatically generated

* Part 6
  + Glucose, mass and pedigree all have positive correlations with diabetes, and they're all statistically significant because their p-values are all lower than 0.05. The rest of the variables aren't statistically significant because their p-values are higher than 0.05.
* Part 7
  + The patient in the median of each of my independent variables has a 19.3% chance of having diabetes while the patient in the 75% tile has a 64.9% chance and the patient in the 25% tile has a 4.78% chance.

Code

"""This is the code file for both portions of problem set 5."""

import matplotlib.pyplot as plt

import seaborn as sns

import pandas as pd

import statsmodels.api as sm

# Exercise 1

def first\_exercise():

"""

This function loads and preps the dataset that contains information about

expected wage rates and its related variables. It also generates visuals

that describe the relationship between wage and years of education. It

also generates a model for wage data and estimates such model by producing

an OLS regression table.

"""

# Part 1

wage = pd.read\_csv("wage.csv").dropna()

# Part 2

sns.lmplot(x='educ', y='wage', data=wage)

fig1 = plt.figure(1)

fig1.savefig("1\_2\_1.png")

sns.lmplot(x='exper', y='wage', data=wage)

fig2 = plt.figure(2)

fig2.savefig("1\_2\_2.png")

sns.lmplot(x='tenure', y='wage', data=wage)

fig3 = plt.figure(3)

fig3.savefig("1\_2\_3.png")

# Part 5

lhs = wage['wage']

ind\_vars = ['educ', 'exper', 'tenure', 'nonwhite', 'female']

rhs = wage.loc[:, ind\_vars]

rhs = sm.add\_constant(rhs)

mod = sm.OLS(lhs, rhs)

res = mod.fit()

print(res.summary())

# Part 8

hypo = pd.DataFrame({"wage": [150], "exper": [100], "tenure": [50],

"educ": [100], "female": [0], "nonwhite": [0]})

print(res.predict(hypo))

# Exercise 2

def second\_exercise():

"""

This function loads and preps the dataset that contains information about

diabetes rates and its related variables. It also generates visuals

that describe the relationship between diabetes and variables like insulin

levels, pedigree and glucose. It also generates a model for diabetes data

and estimates such model by producing a logistic regression table.

"""

# Part 1

diabetes = pd.read\_csv("diabetes.csv").dropna()

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

# Part 2

sns.lmplot(x="insulin", y="diabetes", data=diabetes, logistic=True)

fig4 = plt.figure(4)

fig4.savefig("2\_2\_1.png")

sns.lmplot(x="pedigree", y="diabetes", data=diabetes, logistic=True)

fig5 = plt.figure(5)

fig5.savefig("2\_2\_2.png")

sns.lmplot(x="glucose", y="diabetes", data=diabetes, logistic=True)

fig6 = plt.figure(6)

fig6.savefig("2\_2\_3.png")

# Part 5

lhs = diabetes['diabetes']

ind\_vars = ['pregnant', 'glucose', 'pressure', 'triceps',

'insulin', 'mass', 'pedigree', 'age']

rhs = diabetes.loc[:, ind\_vars]

rhs = sm.add\_constant(rhs)

mod = sm.Logit(lhs, rhs)

res = mod.fit()

print(res.summary())

# Part 7

for percentiles in [.25, .5, .75]:

values = rhs.quantile(percentiles)

print(res.predict(values.to\_numpy()))