Student Name: **Luke Nguyen**
Student ID: **D5850A**

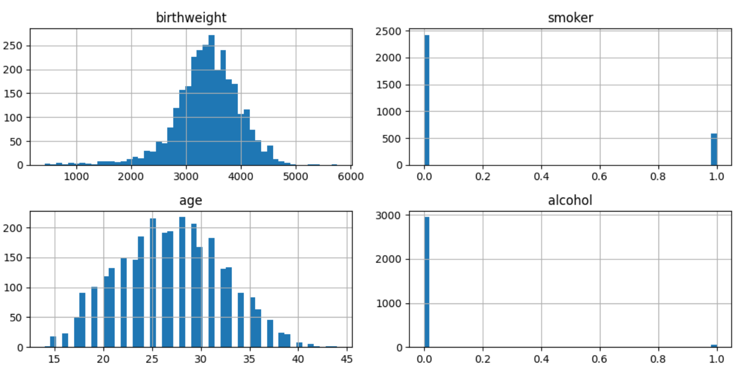
Statistical Methods and Data Analysis (EN.625.603)

Project 2

Question 1
Get to know your data. Make histograms and summary statistics of your data to get a sense of distributions.

(a) What is the average value of birthweight for mothers who smoke? For mothers who don't smoke?

Solution




```
       birthweight       smoker          age      alcohol
count  3000.000000  3000.000000  3000.000000  3000.000000
mean   3382.933667     0.194000    26.889000     0.019333
std     592.162889     0.395495     5.362487     0.137717
min     425.000000     0.000000    14.000000     0.000000
25%    3062.000000     0.000000    23.000000     0.000000
50%    3420.000000     0.000000    27.000000     0.000000
75%    3750.000000     0.000000    31.000000     0.000000
max    5755.000000     1.000000    44.000000     1.000000
smoker
0    3432.059967
1    3178.831615
```

From Python and Pandas, we can see that the average value of birthweight for mothers who smoke is 3178.83 and for mothers who don't smoke is 3432.06.

Question 2
Consider associations. Plot each predictor (variables 2 through 11 in the pdf data description) against the response (birthweight). You could also do a quick line fit or get its correlation. Correlation is with "cor()". A line fit can be achieved using the linear model function. Try for regressions 2 through 11.

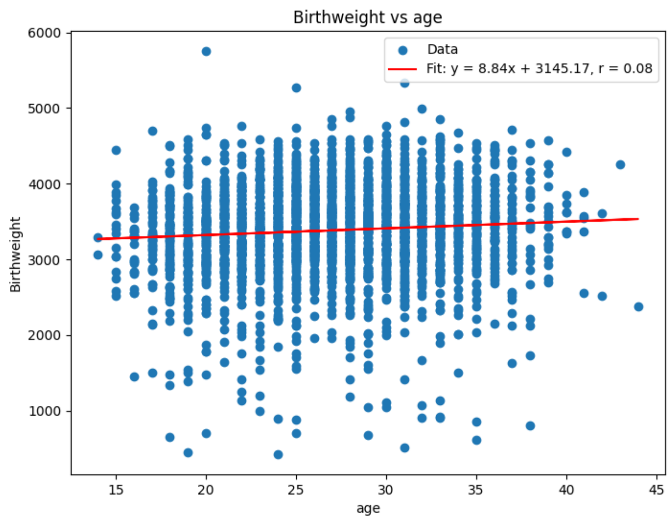
(a) What does a regression of birthweight on the binary variable smoker suggest about the relationship between maternal smoking and infant birthweight?

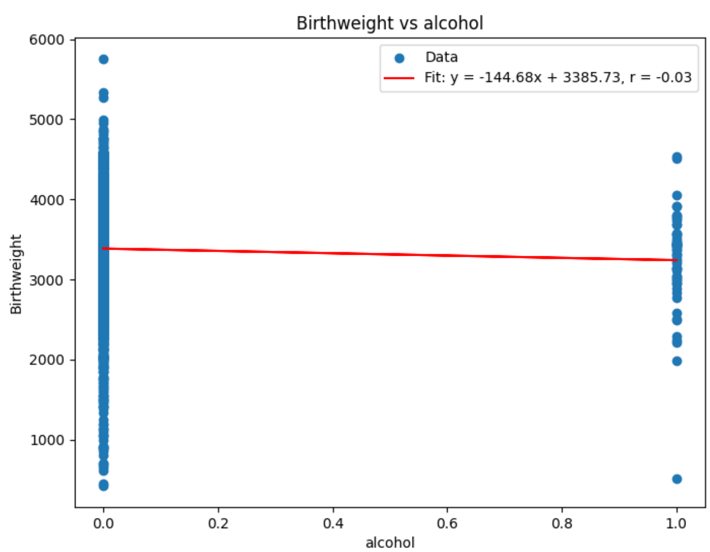
(b) Do you think the regression above accurately captures the impact of smoking on birthweight? (Consider the assumptions of the linear regression model and whether they are met. Hint: do you think smoking is uncorrelated with other factors that cause low birthweight?)

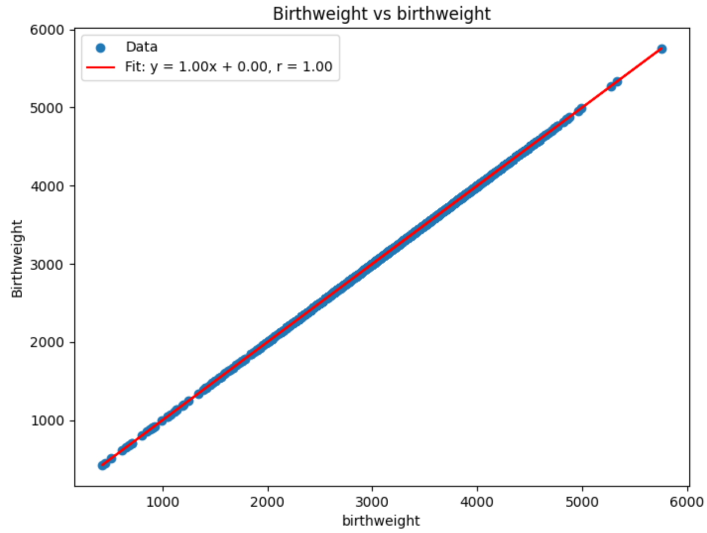
(c) Regress birthweight on smoker, alcohol, and nprevist. Explain why the exclusion of these variables could lead to a biased regression coefficient in (a) above. Is the estimated effect of smoking on birthweight substantially different from the regression in (a) above?

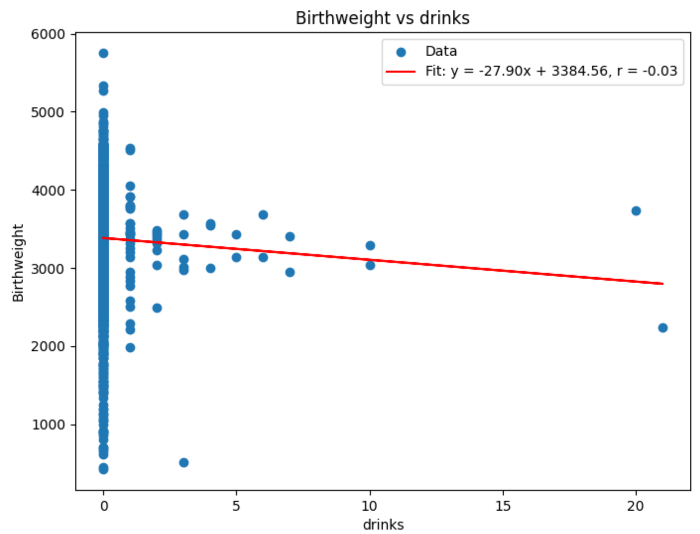
(d) Jane smoked during her pregnancy, did not drink alcohol, and had 8 prenatal care visits. Use the regression in (c) to predict the birthweight of Jane's infant.

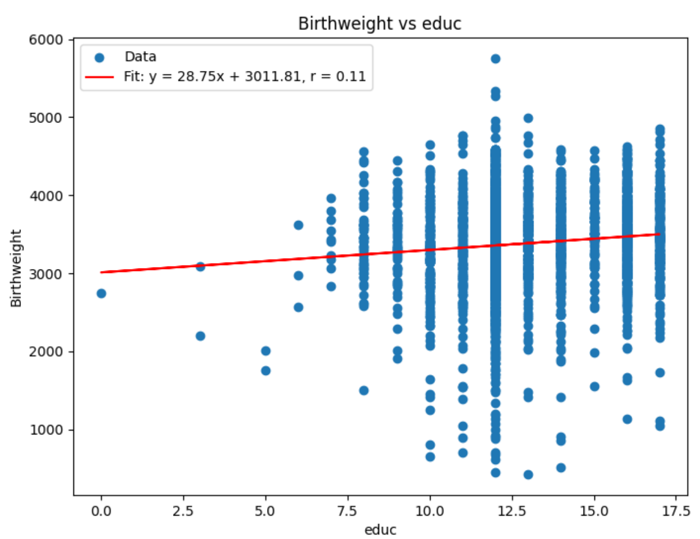
Solution
All plots and regression lines and fitting model for each predictor against birthweight are shown below.

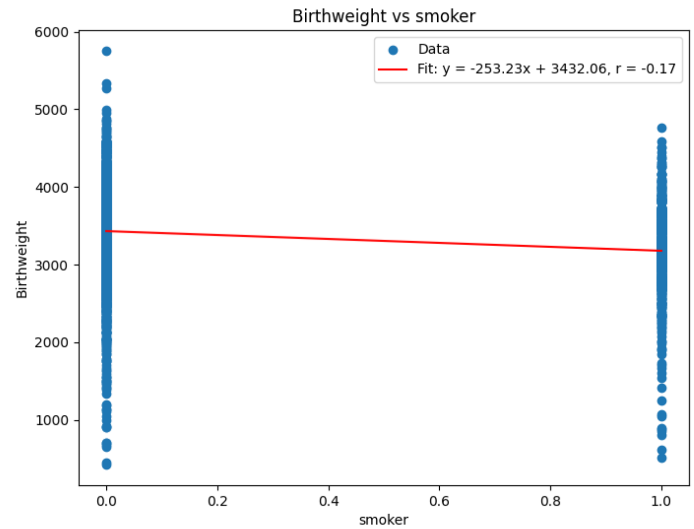


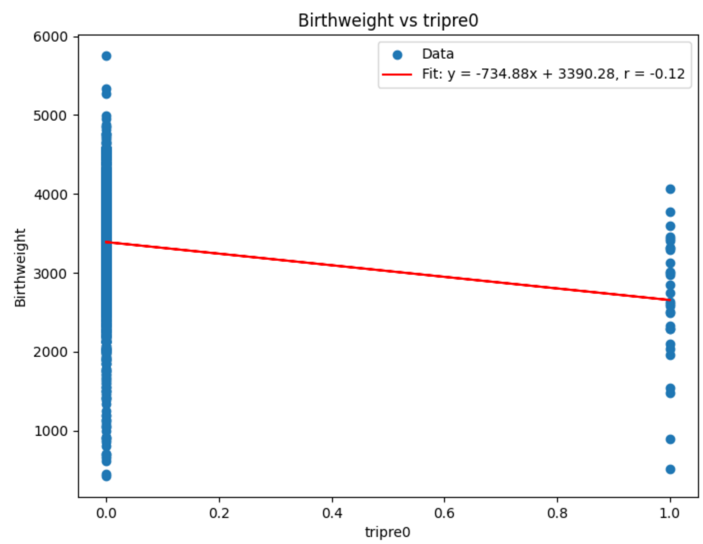


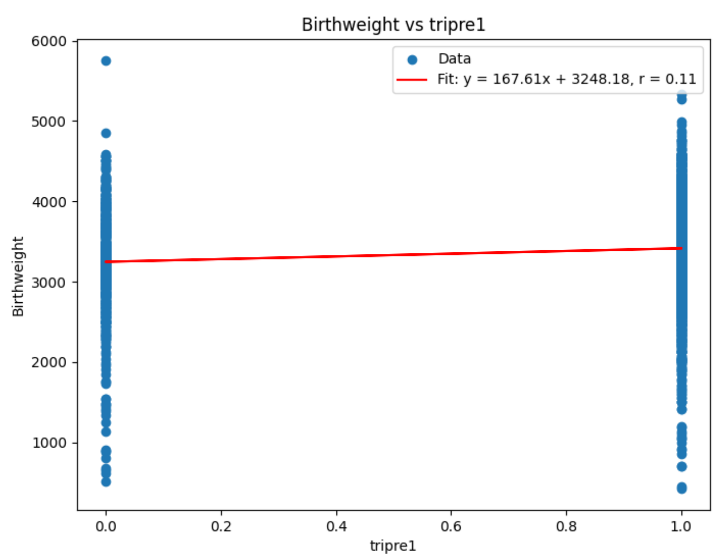

Birthweight vs birthweight
Data
Fit: y = 1.00x + 0.00, r = 1.00
Birthweight
birthweight
1000
2000
3000
4000
5000
6000

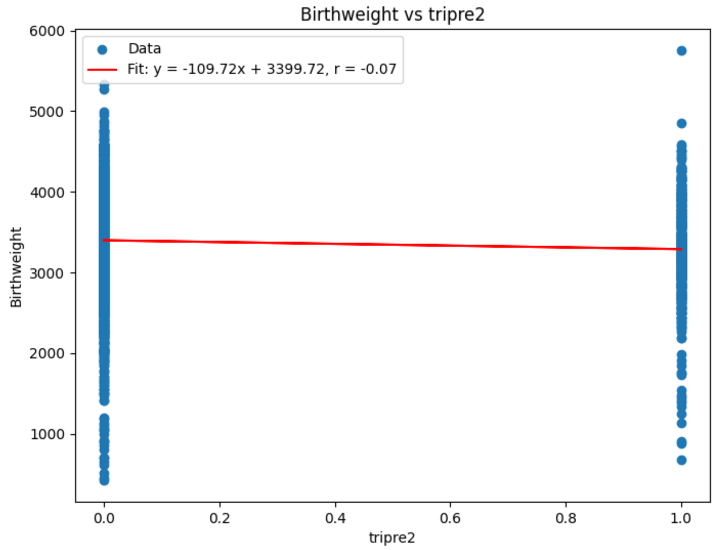

Birthweight vs drinks
Data
Fit: y = -27.90x + 3384.56, r = -0.03
Birthweight
drinks
0
5
10
15
20
1000
2000
3000
4000
5000
6000

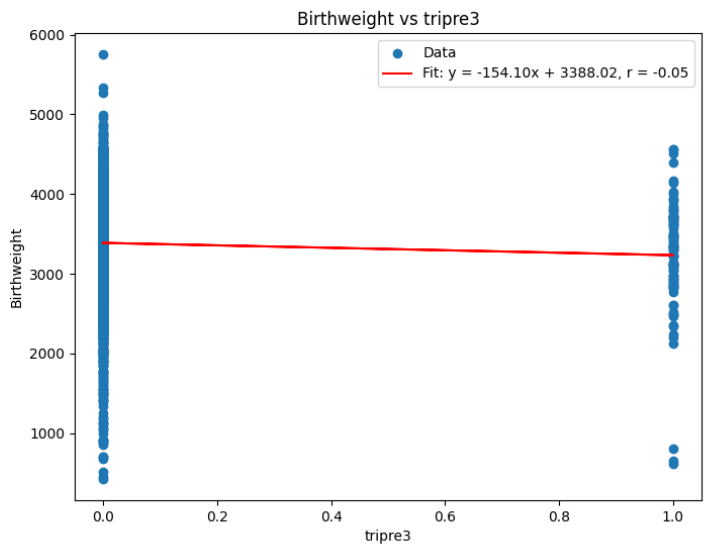

Birthweight vs educ
Data
Fit: y = 28.75x + 3011.81, r = 0.11
Birthweight
educ
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
1000
2000
3000
4000
5000
6000

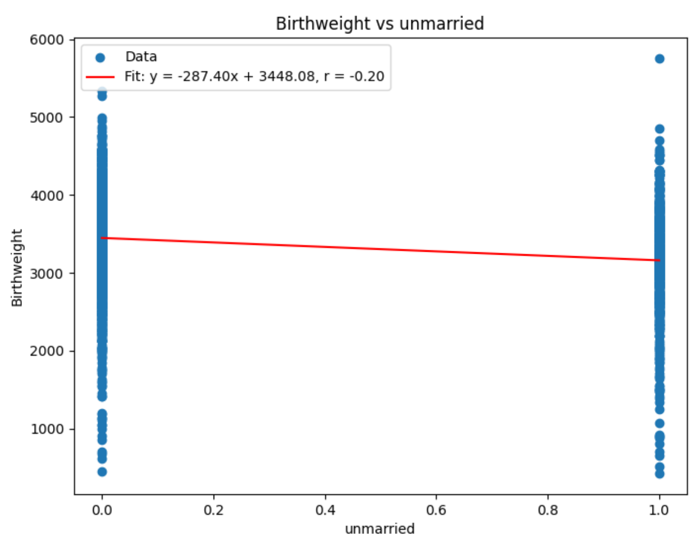

Birthweight vs smoker
Data
Fit: y = -253.23x + 3432.06, r = -0.17
Birthweight
smoker
0.0
0.2
0.4
0.6
0.8
1.0
1000
2000
3000
4000
5000
6000


Birthweight vs tripre0
Data
Fit: y = -734.88x + 3390.28, r = -0.12
Birthweight
tripre0
0.0
0.2
0.4
0.6
0.8
1.0
1000
2000
3000
4000
5000
6000


Birthweight vs tripre1
Data
Fit: y = 167.61x + 3248.18, r = 0.11
Birthweight
tripre1
0.0
0.2
0.4
0.6
0.8
1.0
1000
2000
3000
4000
5000
6000


Birthweight vs tripre2
Data
Fit: y = -109.72x + 3399.72, r = -0.07
Birthweight
tripre2
6000
5000
4000
3000
2000
1000
0.0
0.2
0.4
0.6
0.8
1.0


Birthweight vs tripre3
Data
Fit: y = -154.10x + 3388.02, r = -0.05
Birthweight
tripre3
6000
5000
4000
3000
2000
1000
0.0
0.2
0.4
0.6
0.8
1.0

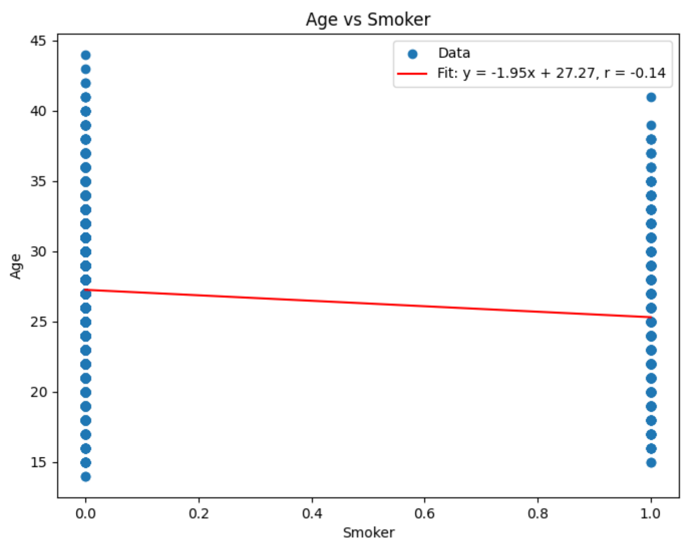

Birthweight vs unmarried
Data
Fit: y = -287.40x + 3448.08, r = -0.20
Birthweight
unmarried
6000
5000
4000
3000
2000
1000
0.0
0.2
0.4
0.6
0.8
1.0

(a) The fitting model for the regression of birthweight on the binary variable smoker is

$$\text{birthweight} = 3369.58 - 254.47 \times \text{smoker}$$

The regression suggests that the birthweight of infants whose mothers smoke is 254.47 grams less than the birthweight of infants whose mothers don't smoke.

(b) It might not be accurate because the assumptions of the linear regression model might not be met. For example, smoking might be correlated with other factors that cause low birthweight, one example is age of mother. The younger the mother is, the more likely she is to smoke and the more likely her infant is to have low birthweight. Plot and regression line and fitting model of age of mother against birthweight are shown below



(c) Regressing birthweight on smoker, alcohol, and nprevist, we have

$$\text{birthweight} = 3051.2486 - 217.5801 \times \text{smoker} - 30.4913 \times \text{alcohol} + 34.0699 \times \text{nprevist}$$

The exclusion of these variables could lead to a biased regression coefficient in (a) above because

i. The estimated effect of smoking on birthweight is -254.47 grams if regress birthweight against smoker only

ii. The estimated effect of smoking on birthweight is -217.58 grams if regress birthweight against smoker, alcohol, and nprevist

And the estimated effect of smoking on birthweight is substantially different.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:            birthweight   R-squared:                       0.073
Model:                            OLS   Adj. R-squared:                  0.072
Method:                 Least Squares   F-statistic:                     78.47
Date:                Wed, 16 Aug 2023   Prob (F-statistic):           7.31e-49
Time:                        04:35:45   Log-Likelihood:                -23294.
No. Observations:                3000   AIC:                         4.660e+04
Df Residuals:                    2996   BIC:                         4.662e+04
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       3051.2486     34.016     89.701      0.000    2984.552    3117.946
smoker      -217.5801     26.680     -8.155      0.000    -269.892    -165.268
alcohol      -30.4913     76.234     -0.400      0.689    -179.968     118.985
nprevist      34.0699      2.855     11.933      0.000      28.472      39.668
==============================================================================
Omnibus:                      374.095   Durbin-Watson:                   1.974
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              869.220
Skew:                          -0.729   Prob(JB):                    1.78e-189
Kurtosis:                       5.197   Cond. No.                         85.2
==============================================================================

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

(d) Jane smoked, did not drink alcohol and had 8 prenatal care visits. Her infant's birthweight is

$$\begin{aligned} \text{birthweight} &= 3051.2486 - 217.5801 \times 1 - 30.4913 \times 0 + 34.0699 \times 8 \\ &= 3106.2277 \end{aligned}$$

Question 3

An alternative way to control for prenatal visits is to use binary variables tripre0 through tripre3. Regress birthweight on smoker, alcohol, tripre0, tripre2, and tripre3.

(a) Why is tripre1 excluded from the model? What happens if you include it in the regression?

(b) The estimated coefficient on tripre0 is large and negative. What does this coefficient measure? Interpret its value.

(c) Interpret the value of the estimated coefficients on tripre2 and tripre3.

(d) Does the regression in (3) explain a larger fraction of the variance in birthweight than the regression in (2c)? (Hint: consider $\bar{R}^2$.)

Solution

The fitting model for the regression of birthweight on smoker, alcohol, tripre0, tripre2, is as follows

```
                            OLS Regression Results
==============================================================================
Dep. Variable:            birthweight   R-squared:                       0.046
Model:                            OLS   Adj. R-squared:                  0.045
Method:                 Least Squares   F-statistic:                     29.18
Date:                Wed, 16 Aug 2023   Prob (F-statistic):           5.20e-29
Time:                        05:02:33   Log-Likelihood:                -23336.
No. Observations:                3000   AIC:                         4.668e+04
Df Residuals:                    2994   BIC:                         4.672e+04
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       3454.5493     12.650    273.077      0.000    3429.745    3479.354
smoker      -228.8476     27.165     -8.424      0.000    -282.111    -175.584
alcohol      -15.1000     77.541     -0.195      0.846    -167.138     136.938
tripre0     -697.9687    106.876     -6.531      0.000    -907.526    -488.411
tripre2     -100.8373     29.619     -3.404      0.001    -158.913     -42.762
tripre3     -136.9553     59.581     -2.299      0.022    -253.780     -20.131
==============================================================================
Omnibus:                      443.968   Durbin-Watson:                   1.976
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1157.634
Skew:                          -0.811   Prob(JB):                    4.20e-252
Kurtosis:                       5.575   Cond. No.                         10.5
==============================================================================

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

(a) tripre1 is excluded from the model because it is a linear combination of tripre0, tripre2, and tripre3. If we include it in the regression, the regression would be perfect multicollinearity overparameterized on tripre0, tripre1, tripre2, and tripre3.

	tripre0	tripre1	tripre2	tripre3
count	3000.000000	3000.000000	3000.000000	3000.000000
mean	0.010000	0.804000	0.153000	0.033000
std	0.099515	0.397035	0.360048	0.178666
min	0.000000	0.000000	0.000000	0.000000
25%	0.000000	1.000000	0.000000	0.000000
50%	0.000000	1.000000	0.000000	0.000000
75%	0.000000	1.000000	0.000000	0.000000
max	1.000000	1.000000	1.000000	1.000000

As showing in the figure above

$$\overline{tripre1} = 1 - \overline{tripre0} - \overline{tripre2} - \overline{tripre3}$$

(b) The estimated coefficient on tripre0 is large and negative. This coefficient measures the difference in birthweight between infants whose mothers had no prenatal visits and infants whose mothers had one or more prenatal visit. The difference is 698 grams which is very substantial. This suggests that prenatal visits has strong positive linear relationship with infant's birthweight.
By common sense, having no prenatal visits could mean that the pregnant mother is not aware of the importance of prenatal visits, or she is not able to afford, both of which could be strong signs of not having enough resources to support the infant's growth.

(c) The estimated coefficients on tripre2, tripre3 are -100.84 and -136.96 respectively. These coefficients measure the difference in birthweight between infants whose mothers had first prenatal visit in the second/third trimester and infants whose mothers had first prenatal visit in the first trimester or none. It suggests that the earlier the mother has her first prenatal visit, the more likely her infant is to have higher birthweight.

(d) R^2 in (2c) is 0.073, while R^2 in (3) is 0.046. The regression in (2c) explains a larger fraction of the variance in birthweight than the regression in (3).

Question 4
Consider adding an additional regressor: Regress birthweight on smoker, alcohol, nprevist, and unmarried.

(a) Compare the coefficient on smoker in this regression to the coefficients on smoker in regressions (2a) and (2c). What is the estimated effect of smoking on birthweight in each regression?

(b) Interpret differences in estimated effects.

(c) Interpret the estimated effect of marital status on birthweight. Is the coefficient on unmarried statistically significant? Is the magnitude of the coefficient large?

(d) A family advocacy group notes that the large coefficient suggests that public policies that encourage marriage will lead, on average, to healthier babies. Do you agree? (Hint: consider some of the various factors that unmarried may be controlling for and how this affects the interpretation of this coefficient).

Solution

The fitting model for the regression of birthweight on smoker, alcohol, nprevist, and unmarried is as follows

```
                            OLS Regression Results
==============================================================================
Dep. Variable:            birthweight   R-squared:                       0.089
Model:                            OLS   Adj. R-squared:                  0.087
Method:                 Least Squares   F-statistic:                     72.79
Date:                Wed, 16 Aug 2023   Prob (F-statistic):           6.12e-59
Time:                        05:20:20   Log-Likelihood:                -23268.
No. Observations:                3000   AIC:                         4.655e+04
Df Residuals:                    2995   BIC:                         4.658e+04
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       3134.4000     35.656     87.907      0.000    3064.487    3204.313
smoker      -175.3769     27.099     -6.472      0.000    -228.511    -122.243
alcohol      -21.0835     75.607     -0.279      0.780    -169.331     127.164
nprevist      29.6025      2.898     10.213      0.000      23.920      35.286
unmarried   -187.1332     26.007     -7.195      0.000    -238.128    -136.139
==============================================================================
Omnibus:                      369.861   Durbin-Watson:                   1.967
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              880.870
Skew:                          -0.714   Prob(JB):                    5.27e-192
Kurtosis:                       5.238   Cond. No.                         85.2
==============================================================================

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

(a) Smoker coefficient in this regression is -175.38, while smoker coefficient in (2a) is -254.47 and smoker coefficient in (2c) is -217.58. The estimated effect

  i. In (2a) is lowering birthweight by 254.47 grams
  ii. In (2c) is lowering birthweight by 217.58 grams
  iii. In this regression is lowering birthweight by 175.38 grams

(b) Smoker coefficient in this regression is less significant than smoker coefficient in (2a) and (2c). This suggests that the estimated effect of smoking on birthweight is less significant in this regression than in (2a) and (2c), this is because the effect of smoking on birthweight is confounded by unmarried in this regression. A possible explanation is that unmarried mothers are more likely to smoke than married mothers, and unmarried mothers are more likely to have lower birthweight infants than married mothers. Therefore, the estimated effect of smoking on birthweight is less significant in this regression than in (2a) and (2c).

(c) The coefficient on unmarried is statistically significant because p-value is less than 0.05. The magnitude of the coefficient is -187 which is very large because it would put the infant's birthweight at 0.31 standard deviations below the mean, if the birthweight is normally distributed.

(d) Although the coefficient of unmarried is large and its effects on lowering birthweight is statistically significant. We can't conclude that public policies that encourage marriage will lead, on average, to healthier babies. Because, unmarried could have strong positive linear relationship with other regressors, such as smoking, alcohol, and nprevist which are all negatively correlated with birthweight.

Question 5
Consider the other coefficients in this data set. Which do you think should be included in the regression?

(a) Try adding in some of these additional variables. Share your findings and conclusions.

(b) The data set includes babies born in Pennsylvania in 1989. Discuss the external validity of your analysis for: (i) California in 1989, (ii) Illinois in 2015, (iii) South Korea in 2014.

(c) Overall, explain your conclusions on how maternal smoking impacts birthweight (hint: the regressions you're running should be helping you see that isolating the causal effect of smoking on birthweight is difficult because there are a lot of other confounding variables).

Solution
Considering that age and education were not considered in the previous regressions, I am adding them in a new regression.
The fitting model for the regression of birthweight on age and education is as follows

```
                            OLS Regression Results
==============================================================================
Dep. Variable:            birthweight   R-squared:                       0.012
Model:                            OLS   Adj. R-squared:                  0.012
Method:                 Least Squares   F-statistic:                     18.88
Date:                Wed, 16 Aug 2023   Prob (F-statistic):           7.11e-09
Time:                        05:46:59   Log-Likelihood:                -23389.
No. Observations:                3000   AIC:                         4.678e+04
Df Residuals:                    2997   BIC:                         4.680e+04
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       2953.9689     70.840     41.699      0.000    2815.069    3092.869
age            4.5724      2.239      2.042      0.041       0.182       8.963
educ          23.7095      5.542      4.278      0.000      12.843      34.576
==============================================================================
Omnibus:                      447.814   Durbin-Watson:                   1.971
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1163.955
Skew:                          -0.819   Prob(JB):                    1.78e-253
Kurtosis:                       5.575   Cond. No.                         200.
==============================================================================

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

(a) The p-value of age is 0.041 which indicates that age has statistically significant effect on birthweight. For each year older the mother is, the infant's birthweight increases by 4.58 grams
The p-value of education is 0.000 which indicates that education has statistically significant effect on birthweight. For each year of education the mother has, the infant's birthweight increases by 23.71

(b)
   i. Applying the analysis of this dataset Pennsylvania in 1989 to California in 1989 could be valid because the data is collected from the same year and the same country. Thus, the culture and the environment are similar, and the effect of unmarried, smoker, alcohol, and nprevist on birthweight should be similar.
   ii. Applying the analysis of this dataset Pennsylvania in 1989 to Illinois in 2015 will not be valid because the data is collected by two points in time and the difference is almost 30 years. Thus, the culture and the environment have changed a lot.
   iii. Applying the analysis of this dataset Pennsylvania in 1989 to South Korea in 2014 will be absolutely invalid, the culture and the environment are totally different. Smoker, alcohol, age could even have positive linear relationship with birthweight in South Korea.

(c) Maternal smoking has negative linear relationship with birthweight. However, isolating the causal effect of smoking on birthweight is difficult because there are a lot of other confounding variables. For example, unmarried, alcohol, age, and education are all confounding variables. The estimated effect of smoking on birthweight is less significant in the regression of birthweight on smoker, alcohol, nprevist, and unmarried than in the regression of birthweight on smoker, alcohol, and nprevist. This is because unmarried is confounding the effect of smoking on birthweight. A possible explanation is that unmarried mothers are more likely to smoke than married mothers, and unmarried mothers are more likely to have lower birthweight infants than married mothers. Therefore, the estimated effect of smoking on birthweight is less significant in the regression of birthweight on smoker, alcohol, nprevist, and unmarried than in the regression of birthweight on smoker, alcohol, and nprevist.

```
# All code used to generate the answers are below
    import pandas as pd
    import numpy as np
    import statsmodels.api as sm
    from scipy import stats
    import scipy.stats
    import matplotlib.pyplot as plt
    import pandas as pd
    import matplotlib.pyplot as plt
    import numpy as np
    from scipy.stats import linregress


    class WeightSmokingDataframe:
        df = None
        def __init__(self):
            file_path = 'weight_smoking.xlsx'
            file_sheet_name = 'Data'
            self.df = pd.read_excel(file_path, sheet_name=file_sheet_name)


    def part_1_solution():
        weightSmoking = WeightSmokingDataframe()
        print(weightSmoking.df[['birthweight', 'smoker', 'age', 'alcohol']].describe(
            include='all'))
        weightSmoking.df[['birthweight', 'smoker', 'age', 'alcohol']].hist(
            bins=50, figsize=(10, 5))
        average_birthweights = weightSmoking.df.groupby('smoker')[
            'birthweight'].mean()
        print(average_birthweights)

        plt.tight_layout()
        plt.show()


    def part_2_solution():
        weightSmoking = WeightSmokingDataframe()
        predictor_columns = weightSmoking.df.columns[1:12]

        for column in predictor_columns:
            slope, intercept, r_value, p_value, std_err = linregress(
                weightSmoking.df[column], weightSmoking.df['birthweight'])
            line = slope * weightSmoking.df[column] + intercept
            plt.figure(figsize=(8, 6))
            plt.scatter(weightSmoking.df[column],
                        weightSmoking.df['birthweight'], label='Data')
            plt.plot(weightSmoking.df[column], line, color='red',
                     label='Fit: y = {:.2f}x + {:.2f}, r = {:.2f}'.format(slope,
                                                   intercept, r_value))
```

```python
        plt.xlabel(column)
        plt.ylabel('Birthweight')
        plt.title('Birthweight vs ' + column)
        plt.legend()
        plt.show()


def scatter_plot(x_col, y_col):
    df = WeightSmokingDataframe().df
    slope, intercept, r_value, p_value, std_err = linregress(
        df[x_col], df[y_col])
    line = slope * df[x_col] + intercept
    plt.figure(figsize=(8, 6))
    plt.scatter(df[x_col], df[y_col], label='Data')
    plt.plot(df[x_col], line, color='red',
             label='Fit: y = {:.2f}x + {:.2f}, r = {:.2f}'.format(slope, intercept, r_value))
    plt.xlabel(x_col)
    plt.ylabel(y_col)
    plt.title(f'{y_col} vs {x_col}')
    plt.legend()
    plt.show()


def part_2_solution_abcde():
    scatter_plot("smoker", "birthweight")
    scatter_plot("alcohol", "birthweight")
    scatter_plot("nprevist", "birthweight")


def part_2_solution_c():
    df = WeightSmokingDataframe().df
    df = sm.add_constant(df)
    model_c = sm.OLS(df['birthweight'],
                     df[['const', 'smoker', 'alcohol', 'nprevist']])
    results_c = model_c.fit()
    print(results_c.summary())


def part_3_solution():
    df = WeightSmokingDataframe().df
    df = sm.add_constant(df)
    model_c = sm.OLS(df['birthweight'],
                     df[['const', 'smoker', 'alcohol', 'tripre0', 'tripre2', 'tripre3']])
    results_c = model_c.fit()
    print()
    print(results_c.summary())


def part_3a_solution():
    weightSmoking = WeightSmokingDataframe()
    print()
    print(weightSmoking.df[['tripre0', 'tripre1', 'tripre2', 'tripre3']].describe(
        include='all'))


def part_4_solution():
    df = WeightSmokingDataframe().df
    df = sm.add_constant(df)
    model_c = sm.OLS(df['birthweight'],
```

```
                                  df[['const', 'smoker', 'alcohol', 'nprevist', 'unmarried']])
        results_c = model_c.fit()
        print()
        print(results_c.summary())


    def part_5_solution():
        df = WeightSmokingDataframe().df
        df = sm.add_constant(df)
                                  df[['const', 'age', 'educ']])
        results_c = model_c.fit()
        print()
        print(results_c.summary())


    if __name__ == '__main__':
        part_5_solution()
```