Fahri Ulkat 090220756

Question 1

a. Ommited variable bias: If the model excludes an important variable that is related to the independent variables, the estimates may be misleading and lead to incorrect results.

b. Partial effect: When other regressors are constant, the effect on Y of a change in relevant independent variable X.

c. R2 and Adjusted R2:

R^2 shows how much the independent variables explain the variance in the dependent variable. It has a value between 0 and 1.

Adjusted R^2: Since when a new variable added, an increase in R^2 is inevitable we need a better measure. Adjusted R^2 measures more accurately by taking into account the number of variables in the model and this measre does not have to increase when we add anothre regressor because of penalization.

d. Perfect multicollinearity: An independent variable is a linear function of one or more other variables. This causes the model to fail to give accurate results.

e.Dummy variable trap: Including all dummy variables for a category would lead to perfect multicollinearity, making the model invalid.

f. Imperfect multicollinearity: It occurs when there is a high but imperfect linear relationship between independent variables. This can increase the magnitude of standard errors and make estimation difficult.

Question 2

Following questions refer to the estimation results in Table 1 below, computed using data for 2015 from the Current Population Survey. The dataset consists of information on 7178 full-time full-year workers. The highest educational achievement for each worker was a bachelor's degree. Let AHE denote average hourly earnings, college denote a binary variable that equals 1 if worker has a bachelor's degree, equals 0 otherwise; age denote age in years. Also, there are four regional dummies, and the regional dummies, northeast, midwest, south and west.

In [4]:
```
from IPython.display import Image, display

display(Image("Screenshot_6.png"))
```

Table 1: Dependent variable: AHE

	(1)	(2)	(3)
college	10.47	10.44	10.42
female	-4.69	-4.56	-4.57
age		0.61	0.61
northeast		0.74	
midwest		-1.54	
south		-0.44	
constant	18.15	0.11	0.33
Observations	7178	7178	7178
R^2	0.165	0.182	0.185

a. Do workers with college degrees earn more on average than workers with no college degree? How much more? Do men earn more than women on average? How much more?

Workers with college degrees earn **$10.47** more on average than workers with no college. The coefficient for female is -4.69 which means women earn **$4.69** less than men.

b. Why is the regressor West omitted from the regression? What would happen if it was included?

If we include "west", **dummy variable trap** will occur and it will cause perfect multicollinearity.

c. Juanita is a 28-year-old female college graduate from the South. Jennifer is a 28-year-old female college graduate from the Midwest. Calculate the expected difference in earnings between Juanita and Jennifer.

Since they are both 28 years old and females, the difference will occur just because of region difference which is **-0.44 - (-1.54) = $1.10**

Question 3

In [5]:
```
display(Image("Screenshot_1.png"))
```

Question 3

In this exercise, we will work with the data file `birthweight_smoking.xlsx`, which contains data for a random sample of babies born in Pennsylvania in 1989. The data includes the baby's birth weight together with various characteristic of the mother, including whether she smoked during the pregnancy. The data set includes the following variables:

- `birthweight`: birth weight of infant (in grams)
- `smoker`: indicator equal to one if the mother smoked during pregnancy and zero, otherwise.
- `age`: age
- `educ`: years of educational attainment (more than 16 years coded as 17)
- `unmarried`: indicator = 1 if mother is unmarried
- `alcohol`: indicator=1 if mother drank alcohol during pregnancy
- `drinks`: number of drinks per week
- `tripre1`: indicator=1 if 1st prenatal care visit in 1st trimester
- `tripre2`: indicator=1 if 1st prenatal care visit in 2nd trimester
- `tripre3`: indicator=1 if 1st prenatal care visit in 3rd trimester
- `tripre0`: indicator=1 if no prenatal visits
- `nprevist`: total number of prenatal visits

Use either R or Pyhton to answer the following quesitons.

a. Use the stargazer package to generate a table of descriptive statistics.

b. Run a regression between `birthweight` and `smoker`. Show your estimation output. Interpret the estimated coefficient on `smoker`.

c. Regress `birthweight` on `smoker`, `alcohol`, `nprevist` and `educ`. Show your estimation output. Interpret the estimated coefficient on `smoker`. Compare your result with part (b).

d. Consider regression in part (c). Interpret the estimated coefficients on `nprevist` and `educ`.

e. Consider the regression in part (c). Interpret R^2 and $\bar{R}^2$. Why they are so similar?

f. Jane smoked during her pregnancy, did not drink alcohol, had 8 prenatal care visits and has 12 years of educational attainment. Use the regression in part (c) to predict the birth weight of Jane's child.

g. An alternative way to control for prenatal visits is to use the binary variables `tripre0` through `tripre3`. Regress `birthweight` on `smoker`, `alcohol`, `tripre1`, `tripre2`, `tripre3` and `tripre0`. Note that R will not estimate the coefficient associated with `tripre0` in order to avoid the dummy variable trap. Show your result and interpret the estimated coefficient on `tripre1`.

In [12]:
```
import pandas as pd
from stargazer.stargazer import Stargazer
```

In [7]:
```
data = pd.read_csv("birthweight_smoking.csv", sep=";")
```

a.

In [18]:
```
data.describe()
```

Out[18]:

	nprevist	alcohol	tripre1	tripre2	tripre3	tripre0	birthweight	smoker	unm
count	3000.000000	3000.000000	3000.000000	3000.000000	3000.000000	3000.000000	3000.000000	3000.000000	3000.0
mean	10.991667	0.019333	0.804000	0.153000	0.033000	0.010000	3382.933667	0.194000	0.2
std	3.672069	0.137717	0.397035	0.360048	0.178666	0.099515	592.162889	0.395495	0.4
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	425.000000	0.000000	0.0
25%	9.000000	0.000000	1.000000	0.000000	0.000000	0.000000	3062.000000	0.000000	0.0
50%	12.000000	0.000000	1.000000	0.000000	0.000000	0.000000	3420.000000	0.000000	0.0
75%	13.000000	0.000000	1.000000	0.000000	0.000000	0.000000	3750.000000	0.000000	0.0
max	35.000000	1.000000	1.000000	1.000000	1.000000	1.000000	5755.000000	1.000000	1.0

b.

In [19]:
```
import statsmodels.formula.api as smf

model = smf.ols('birthweight ~ smoker', data=data).fit()
print(model.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:            birthweight   R-squared:                       0.029
Model:                            OLS   Adj. R-squared:                  0.028
Method:                 Least Squares   F-statistic:                     88.28
Date:                Mon, 09 Dec 2024   Prob (F-statistic):           1.09e-20
Time:                        00:39:42   Log-Likelihood:                -23364.
No. Observations:                3000   AIC:                         4.673e+04
Df Residuals:                    2998   BIC:                         4.674e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept   3432.0600     11.871    289.115      0.000    3408.784    3455.336
smoker      -253.2284     26.951     -9.396      0.000    -306.074    -200.383
==============================================================================
Omnibus:                      473.891   Durbin-Watson:                   1.973
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              1247.472
Skew:                          -0.858   Prob(JB):                     1.30e-271
Kurtosis:                       5.652   Cond. No.                         2.64
==============================================================================

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

c.

In [20]:
```
model = smf.ols('birthweight ~ smoker + alcohol + nprevist + educ', data=data).fit()
print(model.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:            birthweight   R-squared:                       0.074
Model:                            OLS   Adj. R-squared:                  0.072
Method:                 Least Squares   F-statistic:                     59.53
Date:                Mon, 09 Dec 2024   Prob (F-statistic):           1.95e-48
Time:                        00:40:20   Log-Likelihood:                -23293.
No. Observations:                3000   AIC:                         4.660e+04
Df Residuals:                    2995   BIC:                         4.663e+04
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept   2954.6037     68.958     42.846      0.000    2819.394    3089.813
smoker      -207.9322     27.337     -7.606      0.000    -261.532    -154.332
alcohol     -35.4913     76.277     -0.465      0.642    -185.052     114.069
nprevist     33.1679      2.909     11.403      0.000      27.465      38.871
educ          8.1184      5.039      1.611      0.107      -1.762      17.999
==============================================================================
Omnibus:                      373.688   Durbin-Watson:                   1.971
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              871.742
Skew:                          -0.727   Prob(JB):                     5.05e-190
Kurtosis:                       5.205   Cond. No.                         127.
==============================================================================

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

The new coefficient of smoke is **-207.93** which is less than SLR coefficient. It means if the mother smokes birthweight decreases by -207.93

d. "nprevist" coefficient is **33.16** which means each previsit increases birth weight by 33.16. "educ" coefficient is **8.11** which means each additional year of education increases birth weight by 8.11.

e. R2 and Adjusted R2 is around 0.07 which means these variables explains the variance of dependent variables by %7. Since number of data is large enough R2 and Adjusted R2 are so similar.

f.

In [32]:
```
jane = {
    'smoker' : 1,
    'alcohol' : 0,
    'nprevist' : 12,
    'educ': 12
}

janes_child = model.predict(jane)
print(f"Jane's child birthweight = {janes_child.iloc[0]}")
```

```
Jane's child birthweight = 3242.107345036429?
```

g.

In [35]:
```
## to avoid dummy variable trap ( will ignore tripre0
    model = smf.ols('birthweight ~ smoker + alcohol + tripre1 + tripre2 + tripre3', data=data).fit()
    print(model.summary())
```

```
                            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:                Mon, 09 Dec 2024   Prob (F-statistic):           5.20e-29
Time:                        01:06:56   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]
------------------------------------------------------------------------------
Intercept   2756.5806    105.686     25.838      0.000    2547.396    2965.766
smoker      -238.8476     27.165     -8.424      0.000    -292.111    -175.584
alcohol     -15.1000     77.541     -0.195      0.846    -167.138    136.938
tripre1     697.9687    105.876      6.531      0.000     488.411    907.526
tripre2     597.1315    109.421      5.457      0.000     382.584    811.679
tripre3     561.0135    120.876      4.641      0.000     324.004    798.022
==============================================================================
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.                         27.0
==============================================================================

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

If the first prenatal care visit happened in 1st trimester the birthweight will increase by **697.9687**