Birthweight Smoking Analysis

Initial Setup

In [13]:
```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
import statsmodels.api as sm

# Converted to csv since pandas wasn't playing nice
smoking_df = pd.read_csv("birthweight_smoking-1.csv")
# Reorder according to the documentation order
smoking_df = smoking_df[["birthweight", "smoker", "age", "educ", "unmarried"
display(smoking_df)
```

	birthweight	smoker	age	educ	unmarried	alcohol	drinks	tripre1	tripre2	trip
0	4253	1	27	12	1	0	0	1	0	
1	3459	0	24	16	0	0	0	0	1	
2	2920	1	23	11	0	0	0	1	0	
3	2600	0	28	17	0	0	0	1	0	
4	3742	0	27	13	0	0	0	1	0	
...	...	...	...	...	...	...	...	...	...	
2995	2520	0	42	12	0	0	0	0	1	
2996	3062	0	27	17	0	0	0	1	0	
2997	3799	0	28	12	0	0	0	0	1	
2998	2070	1	21	11	0	0	0	1	0	
2999	2948	0	23	14	0	0	0	1	0	

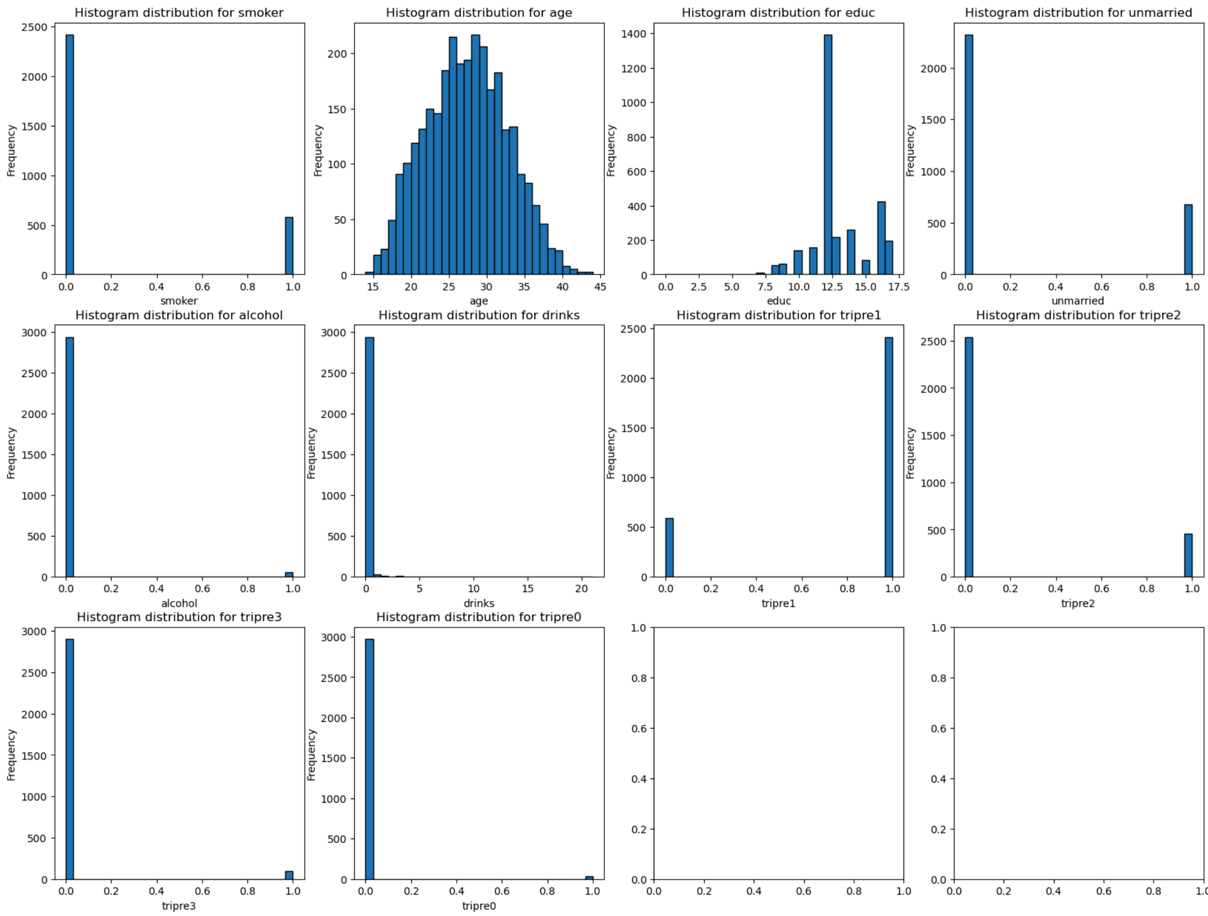
3000 rows × 12 columns

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

In [28]:
```
# Summary statistics for all numerical columns
print(smoking_df.describe())

# Create histograms for relevant variables
fig, axes = plt.subplots(3, 4, figsize=(20, 15))
axes = axes.flatten()

predictor_vars = list(smoking_df.columns)[1:11]
for i, predictor_var in enumerate(predictor_vars):
    axes[i].hist(smoking_df[predictor_var], bins=30, edgecolor="black")
```

```python
    axes[i].set_title(f"Histogram distribution for {predictor_var}")
    axes[i].set_xlabel(predictor_var)
    axes[i].set_ylabel("Frequency")

plt.show()
```

```
       birthweight       smoker          age         educ    unmarried  \
count  3000.000000  3000.000000  3000.000000  3000.000000  3000.000000   
mean   3382.933667     0.194000    26.889000    12.907000     0.226667   
std     592.162889     0.395495     5.362487     2.166699     0.418745   
min     425.000000     0.000000    14.000000     0.000000     0.000000   
25%    3062.000000     0.000000    23.000000    12.000000     0.000000   
50%    3420.000000     0.000000    27.000000    12.000000     0.000000   
75%    3750.000000     0.000000    31.000000    14.000000     0.000000   
max    5755.000000     1.000000    44.000000    17.000000     1.000000   

           alcohol       drinks      tripre1      tripre2      tripre3  \
count  3000.000000  3000.000000  3000.000000  3000.000000  3000.000000   
mean      0.019333     0.058333     0.804000     0.153000     0.033000   
std       0.137717     0.687814     0.397035     0.360048     0.178666   
min       0.000000     0.000000     0.000000     0.000000     0.000000   
25%       0.000000     0.000000     1.000000     0.000000     0.000000   
50%       0.000000     0.000000     1.000000     0.000000     0.000000   
75%       0.000000     0.000000     1.000000     0.000000     0.000000   
max       1.000000    21.000000     1.000000     1.000000     1.000000   

           tripre0     nprevist  
count  3000.000000  3000.000000  
mean      0.010000    10.991667  
std       0.099515     3.672069  
min       0.000000     0.000000  
25%       0.000000     9.000000  
50%       0.000000    12.000000  
75%       0.000000    13.000000  
max       1.000000    35.000000  
```



A. What is the average value of birthweight for mothers who smoke? For mothers who don’t smoke?

In [14]:
```
smoker_avg_birthweight = smoking_df[smoking_df["smoker"] == 1]["birthweight"
non_smoker_avg_birthweight = smoking_df[smoking_df["smoker"] == 0]["birthwei

print(f"{smoker_avg_birthweight=}")
print(f"{non_smoker_avg_birthweight=}")
```

```
smoker_avg_birthweight=3178.831615120275
non_smoker_avg_birthweight=3432.0599669148055
```

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. Two commands, model<-lm(response~predictor), followed by summary(model) will give you least squares result for an individual predictor (e.g., smoker) against the response. This will give you a rough idea of what might be important. Try for regressions 2 through 11.

In [ ]:
```
predictors = list(smoking_df.columns)[1:11]

fig, axes = plt.subplots(3, 4, figsize=(20, 25))
axes = axes.flatten()

correlations = {}
regression_results = {}
```

```
for i, predictor in enumerate(predictors):
    # Calculate correlation
    correlation = smoking_df[predictor].corr(smoking_df["birthweight"])
    correlations[predictor] = correlation

    # Perform simple linear regression
    intercept = sm.add_constant(smoking_df[predictor])
    model = sm.OLS(smoking_df["birthweight"], intercept).fit()
    regression_results[predictor] = model

    # Plot predictor vs birthweight
    axes[i].scatter(smoking_df[predictor], smoking_df["birthweight"])

    # Add regression line
    x_range = np.linspace(smoking_df[predictor].min(), smoking_df[predictor]
    y_pred = model.params[0] + model.params[1] * x_range
    axes[i].plot(x_range, y_pred, "r--", linewidth=2)

    axes[i].set_xlabel(predictor)
    axes[i].set_ylabel('birthweight')
    axes[i].set_title(f'{predictor} vs birthweight\nCorr: {correlation:.3f},

    print(f"{predictor} {model.summary()=}")

plt.show()
```

```
smoker model.summary()=<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
==
Dep. Variable:            birthweight   R-squared:                       0.0
29
Model:                            OLS   Adj. R-squared:                  0.0
28
Method:                 Least Squares   F-statistic:                     88.
28
Date:                Sun, 10 Aug 2025   Prob (F-statistic):           1.09e-
20
Time:                        05:23:37   Log-Likelihood:                -2336
4.
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      0.97
5]
------------------------------------------------------------------------------
--
const       3432.0600     11.871    289.115      0.000    3408.784    3455.3
36
smoker      -253.2284     26.951     -9.396      0.000    -306.074    -200.3
83
==============================================================================
==
Omnibus:                      473.891   Durbin-Watson:                   1.9
73
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1247.4
72
Skew:                          -0.858   Prob(JB):                    1.30e-2
71
Kurtosis:                       5.652   Cond. No.                         2.
64
==============================================================================
==

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is corre
ctly specified.
"""
age model.summary()=<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
==
Dep. Variable:            birthweight   R-squared:                       0.0
06
Model:                            OLS   Adj. R-squared:                  0.0
```

```
06
Method:                 Least Squares   F-statistic:                     19.
35
Date:                Sun, 10 Aug 2025   Prob (F-statistic):           1.13e-
05
Time:                        05:23:37   Log-Likelihood:                -2339
8.
No. Observations:                3000   AIC:                         4.680e+
04
Df Residuals:                    2998   BIC:                         4.681e+
04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
==
                 coef    std err          t      P>|t|      [0.025      0.97
5]
------------------------------------------------------------------------------
--
const       3145.1747     55.119     57.061      0.000    3037.099    3253.2
50
age            8.8422      2.010      4.398      0.000       4.901      12.7
84
==============================================================================
==
Omnibus:                      459.571   Durbin-Watson:                   1.9
77
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1196.2
54
Skew:                          -0.838   Prob(JB):                    1.73e-2
60
Kurtosis:                       5.600   Cond. No.                         14
0.
==============================================================================
==

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is corre
ctly specified.
"""
educ model.summary()=<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
==
Dep. Variable:            birthweight   R-squared:                       0.0
11
Model:                            OLS   Adj. R-squared:                  0.0
11
Method:                 Least Squares   F-statistic:                     33.
56
Date:                Sun, 10 Aug 2025   Prob (F-statistic):           7.65e-
09
Time:                        05:23:37   Log-Likelihood:                -2339
1.
No. Observations:                3000   AIC:                         4.679e+
```

```
04
Df Residuals:                    2998   BIC:                         4.680e+
04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
==
                 coef    std err          t      P>|t|      [0.025      0.97
5]
------------------------------------------------------------------------------
--
const       3011.8137     64.963     46.362      0.000    2884.437    3139.1
91
educ          28.7534      4.964      5.793      0.000      19.021      38.4
86
==============================================================================
==
Omnibus:                      441.177   Durbin-Watson:                   1.9
73
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1139.8
87
Skew:                          -0.809   Prob(JB):                    3.00e-2
48
Kurtosis:                       5.549   Cond. No.                          7
9.5
==============================================================================
==

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is corre
ctly specified.
"""
unmarried model.summary()=<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
==
Dep. Variable:            birthweight   R-squared:                       0.0
41
Model:                            OLS   Adj. R-squared:                  0.0
41
Method:                 Least Squares   F-statistic:                      12
9.2
Date:                Sun, 10 Aug 2025   Prob (F-statistic):            2.46e-
29
Time:                        05:23:37   Log-Likelihood:                -2334
4.
No. Observations:                3000   AIC:                         4.669e+
04
Df Residuals:                    2998   BIC:                         4.670e+
04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
==
                 coef    std err          t      P>|t|      [0.025      0.97
```

```
5]
------------------------------------------------------------------------------
--
const       3448.0780     12.040    286.396      0.000    3424.471    3471.6
85
unmarried   -287.4015     25.288    -11.365      0.000    -336.985    -237.8
18
==============================================================================
==
Omnibus:                      426.713   Durbin-Watson:                   1.9
75
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1112.4
13
Skew:                          -0.782   Prob(JB):                    2.77e-2
42
Kurtosis:                       5.540   Cond. No.                         2.
54
==============================================================================
==

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is corre
ctly specified.
"""
alcohol model.summary()=<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
==
Dep. Variable:            birthweight   R-squared:                       0.0
01
Model:                            OLS   Adj. R-squared:                  0.0
01
Method:                 Least Squares   F-statistic:                     3.3
98
Date:                Sun, 10 Aug 2025   Prob (F-statistic):             0.06
54
Time:                        05:23:37   Log-Likelihood:                -2340
6.
No. Observations:                3000   AIC:                         4.682e+
04
Df Residuals:                    2998   BIC:                         4.683e+
04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
==
                 coef    std err          t      P>|t|      [0.025      0.97
5]
------------------------------------------------------------------------------
--
const       3385.7308     10.913    310.246      0.000    3364.333    3407.1
29
alcohol     -144.6791     78.486     -1.843      0.065    -298.571       9.2
13
==============================================================================
```

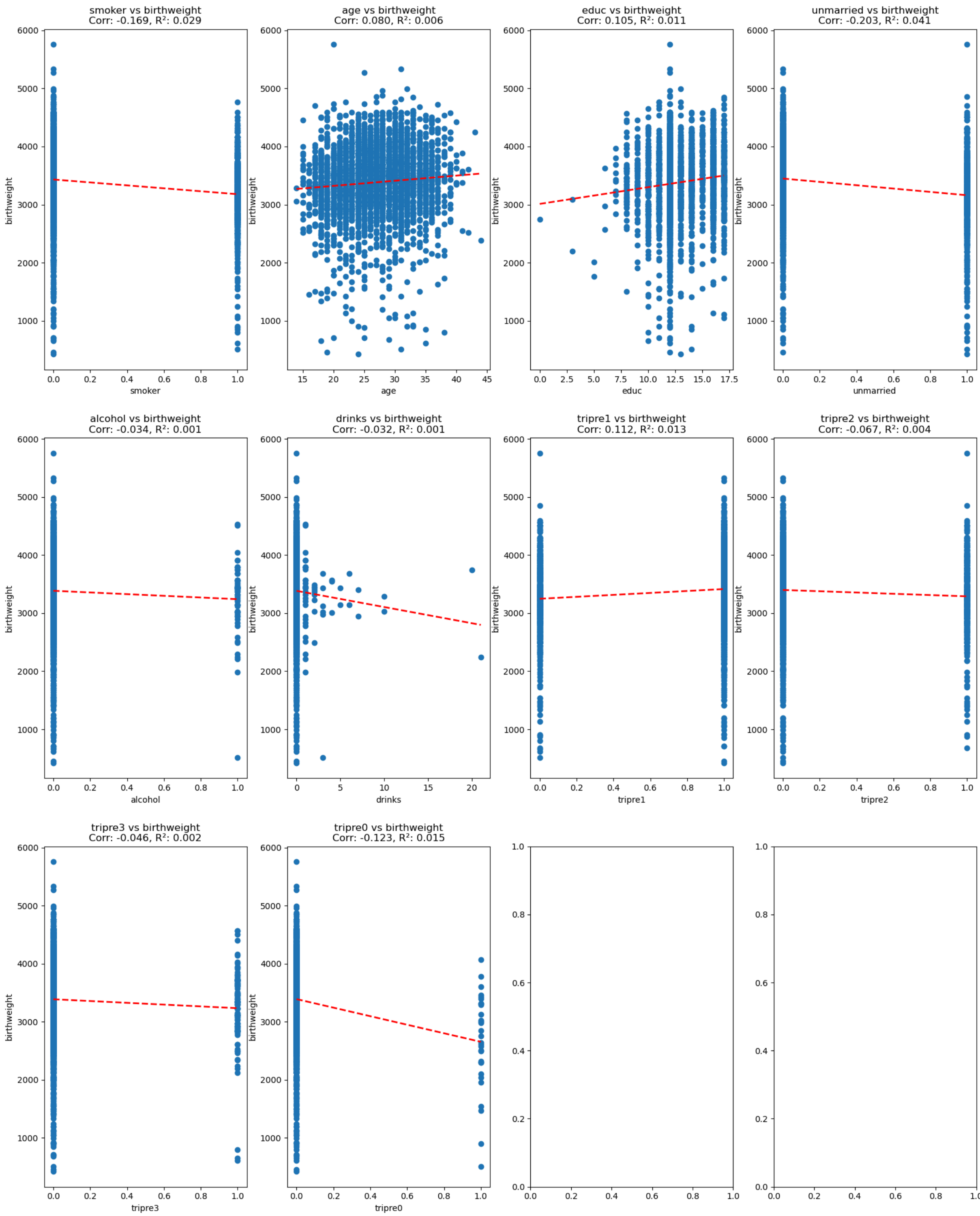
```
==
Omnibus:                      448.589   Durbin-Watson:                   1.9
84
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1145.1
53
Skew:                          -0.826   Prob(JB):                   2.15e-2
49
Kurtosis:                       5.537   Cond. No.                         7.
27
============================================================================
==

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is corre
ctly specified.
"""
drinks model.summary()=<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
============================================================================
==
Dep. Variable:            birthweight   R-squared:                       0.0
01
Model:                            OLS   Adj. R-squared:                  0.0
01
Method:                 Least Squares   F-statistic:                     3.1
52
Date:                Sun, 10 Aug 2025   Prob (F-statistic):             0.07
59
Time:                        05:23:37   Log-Likelihood:                -2340
6.
No. Observations:                3000   AIC:                       4.682e+
04
Df Residuals:                    2998   BIC:                       4.683e+
04
Df Model:                           1
Covariance Type:            nonrobust
============================================================================
==
                 coef    std err          t      P>|t|      [0.025      0.97
5]
----------------------------------------------------------------------------
--
const       3384.5613     10.846    312.048      0.000    3363.294    3405.8
28
drinks       -27.9022     15.715     -1.775      0.076     -58.716       2.9
12
============================================================================
==
Omnibus:                      449.492   Durbin-Watson:                   1.9
83
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1150.3
03
Skew:                          -0.826   Prob(JB):                   1.64e-2
50
Kurtosis:                       5.544   Cond. No.                         1.
```

```
46
==============================================================================
==

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is corre
ctly specified.
"""
tripre1 model.summary()=<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
==
Dep. Variable:            birthweight   R-squared:                       0.0
13
Model:                            OLS   Adj. R-squared:                  0.0
12
Method:                 Least Squares   F-statistic:                     38.
34
Date:                Sun, 10 Aug 2025   Prob (F-statistic):           6.74e-
10
Time:                        05:23:37   Log-Likelihood:                -2338
9.
No. Observations:                3000   AIC:                         4.678e+
04
Df Residuals:                    2998   BIC:                         4.679e+
04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
==
                 coef    std err          t      P>|t|      [0.025      0.97
5]
------------------------------------------------------------------------------
--
const       3248.1786     24.270    133.837      0.000    3200.592    3295.7
66
tripre1      167.6058     27.067      6.192      0.000     114.534     220.6
77
==============================================================================
==
Omnibus:                      429.414   Durbin-Watson:                   1.9
82
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1092.1
92
Skew:                          -0.795   Prob(JB):                    6.81e-2
38
Kurtosis:                       5.492   Cond. No.                         4.
31
==============================================================================
==

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

```
tripre2 model.summary()=<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results                            
=============================================================================
==
Dep. Variable:            birthweight   R-squared:                       0.0
04
Model:                            OLS   Adj. R-squared:                  0.0
04
Method:                 Least Squares   F-statistic:                     13.
40
Date:                Sun, 10 Aug 2025   Prob (F-statistic):           0.0002
56
Time:                        05:23:37   Log-Likelihood:                -2340
1.
No. Observations:                3000   AIC:                         4.681e+
04
Df Residuals:                    2998   BIC:                         4.682e+
04
Df Model:                           1                                         
Covariance Type:            nonrobust                                         
=============================================================================
==
                 coef    std err          t      P>|t|      [0.025      0.97
5]
-----------------------------------------------------------------------------
--
const       3399.7214     11.723    290.002      0.000    3376.735    3422.7
08
tripre2     -109.7235     29.971     -3.661      0.000    -168.489     -50.9
58
=============================================================================
==
Omnibus:                      447.961   Durbin-Watson:                   1.9
81
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1161.0
36
Skew:                          -0.820   Prob(JB):                    7.66e-2
53
Kurtosis:                       5.569   Cond. No.                         2.
85
=============================================================================
==

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is corre
ctly specified.
"""
tripre3 model.summary()=<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results                            
=============================================================================
==
Dep. Variable:            birthweight   R-squared:                       0.0
02
Model:                            OLS   Adj. R-squared:                  0.0
```

```
02
Method:                 Least Squares   F-statistic:                     6.4
95
Date:                Sun, 10 Aug 2025   Prob (F-statistic):             0.01
09
Time:                        05:23:37   Log-Likelihood:                -2340
4.
No. Observations:                3000   AIC:                        4.681e+
04
Df Residuals:                    2998   BIC:                        4.682e+
04
Df Model:                           1
Covariance Type:            nonrobust
============================================================================
==
                 coef    std err          t      P>|t|      [0.025      0.97
5]
----------------------------------------------------------------------------
--
const       3388.0190     10.984    308.444      0.000    3366.482    3409.5
56
tripre3     -154.0998     60.466     -2.549      0.011    -272.659     -35.5
40
============================================================================
==
Omnibus:                      443.461   Durbin-Watson:                   1.9
86
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1123.6
46
Skew:                          -0.819   Prob(JB):                    1.01e-2
44
Kurtosis:                       5.511   Cond. No.                         5.
60
============================================================================
==

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is corre
ctly specified.
"""
tripre0 model.summary()=<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
============================================================================
==
Dep. Variable:            birthweight   R-squared:                       0.0
15
Model:                            OLS   Adj. R-squared:                  0.0
15
Method:                 Least Squares   F-statistic:                     46.
43
Date:                Sun, 10 Aug 2025   Prob (F-statistic):           1.14e-
11
Time:                        05:23:37   Log-Likelihood:                -2338
5.
No. Observations:                3000   AIC:                        4.677e+
```

```
04
Df Residuals:                    2998   BIC:                         4.679e+
04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
==
                 coef    std err          t      P>|t|      [0.025      0.97
5]
------------------------------------------------------------------------------
--
const       3390.2825     10.784    314.368      0.000    3369.137    3411.4
28
tripre0     -734.8825    107.844     -6.814      0.000    -946.339    -523.4
26
==============================================================================
==
Omnibus:                      425.987   Durbin-Watson:                   1.9
88
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1072.7
38
Skew:                          -0.792   Prob(JB):                    1.14e-2
33
Kurtosis:                       5.464   Cond. No.                          1
0.1
==============================================================================
==

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



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

Looking at the regression of birthweight on smoker, there seems to be a negative correlation between maternal smoking and infant birthweight, that is, smoking seems to have a negative effect and reduce the 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?)

The regression above likely does not solely accurately capture the impact of smoking on birthweight, as some of the key assumptions of the linear regression model include normality and independence, both of which are likely correlated to other factors that cause low birthweight due to lifestyle choices or tightly coupled backgrounds.

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?

In [49]:
```
x_multi = sm.add_constant(smoking_df[["smoker", "alcohol", "nprevist"]])
multi_model = sm.OLS(smoking_df["birthweight"], x_multi).fit()
print(f"{multi_model.summary()=}")

pct_change = abs((-253.2284 - (-217.5801)) / -253.228)
print(f"{pct_change=}")
```

```
multi_model.summary()=<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
============================================================================
==
Dep. Variable:            birthweight   R-squared:                       0.0
73
Model:                            OLS   Adj. R-squared:                  0.0
72
Method:                 Least Squares   F-statistic:                     78.
47
Date:                Sun, 10 Aug 2025   Prob (F-statistic):           7.31e-
49
Time:                        05:56:49   Log-Likelihood:                -2329
4.
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      0.97
5]
----------------------------------------------------------------------------
--
const       3051.2486     34.016     89.701      0.000    2984.552    3117.9
46
smoker      -217.5801     26.680     -8.155      0.000    -269.892    -165.2
68
alcohol      -30.4913     76.234     -0.400      0.689    -179.968     118.9
85
nprevist      34.0699      2.855     11.933      0.000      28.472      39.6
68
============================================================================
==
Omnibus:                      374.095   Durbin-Watson:                   1.9
74
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              869.2
20
Skew:                          -0.729   Prob(JB):                    1.78e-1
89
Kurtosis:                       5.197   Cond. No.                          8
5.2
============================================================================
==

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is corre
ctly specified.
"""
pct_change=0.1407755066580315
```

The exclusion of these variables could lead to a biased regression other variables like alcohol and nprevist could also have an effect on birthweight, so smoking could appear to be more harmful than it actually is on its own. The coefficient seems to have changed by ~14%, which is a relatively large amount to have shifted just due to two variables, so the impact is indeed somewhat substantially different.

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.

In [52]:
```
smoker_jane = 1
alcohol_jane = 0
nprevist_jane = 8

intercept, smoker_coeff, alcohol_coeff, nprevisit_coeff = multi_model.params

birthweight_jane = intercept + smoker_coeff * smoker_jane + alcohol_coeff *

print(f"{birthweight_jane=}")
```

```
birthweight_jane=3106.227800368578
```

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.

In [57]:
```
x_tri_multi = sm.add_constant(smoking_df[["smoker", "alcohol", "tripre0", "t
x_tri_multi_model = sm.OLS(smoking_df["birthweight"], x_tri_multi).fit()
print(f"{x_tri_multi_model.summary()=}")
```

```
x_tri_multi_model.summary()=<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
==
Dep. Variable:            birthweight   R-squared:                       0.0
46
Model:                            OLS   Adj. R-squared:                  0.0
45
Method:                 Least Squares   F-statistic:                     29.
18
Date:                Sun, 10 Aug 2025   Prob (F-statistic):           5.20e-
29
Time:                        06:04:13   Log-Likelihood:                -2333
6.
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      0.97
5]
------------------------------------------------------------------------------
--
const       3454.5493     12.650    273.077      0.000    3429.745    3479.3
54
smoker      -228.8476     27.165     -8.424      0.000    -282.111    -175.5
84
alcohol      -15.1000     77.541     -0.195      0.846    -167.138     136.9
38
tripre0     -697.9687    106.876     -6.531      0.000    -907.526    -488.4
11
tripre2     -100.8373     29.619     -3.404      0.001    -158.913     -42.7
62
tripre3     -136.9553     59.581     -2.299      0.022    -253.780     -20.1
31
==============================================================================
==
Omnibus:                      443.968   Durbin-Watson:                   1.9
76
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1157.6
34
Skew:                          -0.811   Prob(JB):                    4.20e-2
52
Kurtosis:                       5.575   Cond. No.                          1
0.5
==============================================================================
==

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

A. Why is tripre1 excluded from the model? What happens if you include it in the regression?

In [59]:
```
x_tri_full_multi = sm.add_constant(smoking_df[["smoker", "alcohol", "tripre0
x_tri_full_multi = sm.OLS(smoking_df["birthweight"], x_tri_multi).fit()
print(f"{x_tri_full_multi.summary()=}")
```

```
x_tri_full_multi.summary()=<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
============================================================================
==
Dep. Variable:          birthweight   R-squared:                       0.0
46
Model:                          OLS   Adj. R-squared:                  0.0
45
Method:               Least Squares   F-statistic:                     29.
18
Date:              Sun, 10 Aug 2025   Prob (F-statistic):           5.20e-
29
Time:                      06:04:44   Log-Likelihood:                -2333
6.
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      0.97
5]
----------------------------------------------------------------------------
--
const       2576.4872     25.835     99.727      0.000    2525.830    2627.1
44
smoker      -228.8476     27.165     -8.424      0.000    -282.111    -175.5
84
alcohol      -15.1000     77.541     -0.195      0.846    -167.138     136.9
38
tripre0      180.0934     85.926      2.096      0.036      11.614     348.5
72
tripre1      878.0621     26.598     33.012      0.000     825.909     930.2
15
tripre2      777.2249     32.505     23.911      0.000     713.490     840.9
60
tripre3      741.1069     51.552     14.376      0.000     640.027     842.1
87
============================================================================
==
Omnibus:                    443.968   Durbin-Watson:                   1.9
76
Prob(Omnibus):                0.000   Jarque-Bera (JB):             1157.6
34
Skew:                        -0.811   Prob(JB):                     4.20e-2
52
Kurtosis:                     5.575   Cond. No.                     1.42e+
16
============================================================================
==

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

```
ctly specified.
[2] The smallest eigenvalue is 2.63e-29. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.
"""
```

It seems that including `tripre1` drastically affects the coefficients of the tripre variables by turning all of them to positive whereas they were previously all negative. It is likely that the results become meaningless because these tripre variables add up perfectly to 1, and with including these all we cannot do an accurate regression since we cannot find unique solutions for the coefficients of the regression.

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

This coefficient measures the difference in average birthweight of babies to mothers who did not have any prenatal care visits compared to mothers who had their prenatal care visit in the first trimester. The ~-698 value indicates that the average birthweight of babies with no prenatal care visits is approximately 698 grams less than those born to mothers with their first prenatal care visit in the first trimester.

C. Interpret the value of the estimated coefficients on tripre2 and tripre3.

- The coefficient of tripre2, ~-101, means that babies born to mothers with their first prenatal care visit in the 2nd trimester had birthweights of 101 grams less than babies born to mothers with their first prenatal care visit in the first trimester, with everything else in the model held constant.
- The coefficient of tripre3, ~-137, means that babies born to mothers with their first prenatal care visit in the 2nd trimester had birthweights of 137 grams less than babies born to mothers with their first prenatal care visit in the first trimester, with everything else in the model held constant.

D. Does the regression in (3) explain a larger fraction of the variance in birthweight than the regression in (2c)? (Hint: consider R2 .)

The regression in (3) has R^2 value of 0.046 while the regression in part (2C) has R^2 value of 0.073. We can therefore determine that with the smaller R^2 value, it actually explains a **lower fraction** of the variance.

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

In [ ]:
```
x_unmarried_multi = sm.add_constant(smoking_df[["smoker", "alcohol", "nprevi
x_unmarried_multi_model = sm.OLS(smoking_df["birthweight"], x_unmarried_mult
print(f"{x_unmarried_multi_model.summary()=}")
```

```
x_unmarried_multi_model.summary()=<class 'statsmodels.iolib.summary.Summar
y'>
"""
                            OLS Regression Results
==============================================================================
==
Dep. Variable:            birthweight   R-squared:                       0.0
89
Model:                            OLS   Adj. R-squared:                  0.0
87
Method:                 Least Squares   F-statistic:                     72.
79
Date:                Sun, 10 Aug 2025   Prob (F-statistic):           6.12e-
59
Time:                        19:07:08   Log-Likelihood:                -2326
8.
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      0.97
5]
------------------------------------------------------------------------------
--
const       3134.4000     35.656     87.907      0.000    3064.487    3204.3
13
smoker      -175.3769     27.099     -6.472      0.000    -228.511    -122.2
43
alcohol      -21.0835     75.607     -0.279      0.780    -169.331     127.1
64
nprevist      29.6025      2.898     10.213      0.000      23.920      35.2
86
unmarried   -187.1332     26.007     -7.195      0.000    -238.128    -136.1
39
==============================================================================
==
Omnibus:                      369.861   Durbin-Watson:                   1.9
67
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              880.8
70
Skew:                          -0.714   Prob(JB):                     5.27e-1
92
Kurtosis:                       5.238   Cond. No.                          8
5.2
==============================================================================
==

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

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?

The coefficient on smoker in this regression is ~-175, which indicates a 175 gram decrease in birthweight for mothers who smoke relative to those who do not in this model. This is lower than the coefficients from (2A) and (2C), which are ~-253 and ~-218 respectively, which indicates a much larger decrease in birthweight (253 and 218 grams respectively) attributed to the fact of whether the mother of the child is a smoker or not.

B. Interpret differences in estimated effects.

The smaller absolute value of the smoking coefficient in this regression indicates that the regression with fewer variables is less accurate, and controlling for more variables leads to a model that is closer to modeling the actual effect of smoking on birthweight, with the previously simpler regressions in question 2 overestimating its effect and possibly having some correlation that is obscured with other variables. With more variables, we can further isolate and decorrelate these variables, so the true value is likely closer to the regression we just did, with a smaller effect on birthweight truly attributed solely to smoking.

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

The unmarried coefficient is ~-187 with an extremely small p-value (that displays as 0.000 when rounded to the thousandths), which in turn is below our threshold of considering it to be statistically significant, so the effect is indeed statistically significant. The effect seems to be that a mother being unmarried is associated with an approximately 187 gram decrease in child birthweight, while holding other variables in the model constant. The magnitude seems to be fairly large -- it is larger than the magnitude of all other variables within the model, including smoking, which we previously found to be a large factor in the other models.

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).

It makes more sense to disagree with the family advocacy group. In this model, the coefficient indicates that being unmarried while controlling for the other variables not present in the model such as age and education, while the coefficient itself reflects the impact of being unmarried while combined with other factors like the behaviors of smoking or drinking during the pregnancy, so it cannot be taken just at face value as an impactful factor in isolation.

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

All the other factors that are left in the data set should honestly be included -- they are all relevant to the attributes of the mother and it would be more impactful to include them so that we can also better isolate and understand the impacts of each factor individually. So we should add age and education, which we haven't yet used, to the regression.

A. Try adding in some of these additional variables. Share your findings and conclusions.

In [12]:

```
x_age_multi = sm.add_constant(smoking_df[["smoker", "alcohol", "nprevist", "
x_age_multi_model = sm.OLS(smoking_df["birthweight"], x_age_multi).fit()
print(f"{x_age_multi_model.summary()=}")

x_educ_multi = sm.add_constant(smoking_df[["smoker", "alcohol", "nprevist",
x_educ_multi_model = sm.OLS(smoking_df["birthweight"], x_educ_multi).fit()
print(f"{x_educ_multi_model.summary()=}")

x_both_multi = sm.add_constant(smoking_df[["smoker", "alcohol", "nprevist",
x_both_multi_model = sm.OLS(smoking_df["birthweight"], x_both_multi).fit()
print(f"{x_both_multi_model.summary()=}")
```

```
x_age_multi_model.summary()=<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
============================================================================
==
Dep. Variable:            birthweight   R-squared:                       0.0
89
Model:                            OLS   Adj. R-squared:                  0.0
87
Method:                 Least Squares   F-statistic:                     58.
51
Date:                Sun, 10 Aug 2025   Prob (F-statistic):           2.88e-
58
Time:                        20:08:12   Log-Likelihood:                -2326
8.
No. Observations:                3000   AIC:                         4.655e+
04
Df Residuals:                    2994   BIC:                         4.658e+
04
Df Model:                           5
Covariance Type:            nonrobust
============================================================================
==
                 coef    std err          t      P>|t|      [0.025      0.97
5]
----------------------------------------------------------------------------
--
const       3201.4887     68.100     47.011      0.000    3067.960    3335.0
17
smoker      -177.1410     27.140     -6.527      0.000    -230.356    -123.9
26
alcohol      -14.6826     75.806     -0.194      0.846    -163.319     133.9
54
nprevist      29.7909      2.903     10.263      0.000      24.099      35.4
83
unmarried   -199.4890     28.116     -7.095      0.000    -254.617    -144.3
61
age           -2.4597      2.127     -1.156      0.248      -6.631       1.7
11
============================================================================
==
Omnibus:                      366.192   Durbin-Watson:                   1.9
69
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              872.2
65
Skew:                          -0.708   Prob(JB):                    3.89e-1
90
Kurtosis:                       5.231   Cond. No.                         21
8.
============================================================================
==

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

```
x_educ_multi_model.summary()=<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
==
Dep. Variable:            birthweight   R-squared:                       0.0
89
Model:                            OLS   Adj. R-squared:                  0.0
87
Method:                 Least Squares   F-statistic:                     58.
24
Date:                Sun, 10 Aug 2025   Prob (F-statistic):           5.23e-
58
Time:                        20:08:12   Log-Likelihood:                -2326
8.
No. Observations:                3000   AIC:                         4.655e+
04
Df Residuals:                    2994   BIC:                         4.658e+
04
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
==
                 coef    std err          t      P>|t|      [0.025      0.97
5]
------------------------------------------------------------------------------
--
const       3157.9138     74.290     42.508      0.000    3012.248    3303.5
79
smoker      -177.0028     27.475     -6.442      0.000    -230.874    -123.1
31
alcohol      -19.7940     75.703     -0.261      0.794    -168.229     128.6
41
nprevist      29.7471      2.926     10.165      0.000      24.009      35.4
85
unmarried   -189.8066     27.046     -7.018      0.000    -242.837    -136.7
76
educ          -1.8754      5.198     -0.361      0.718     -12.068       8.3
17
==============================================================================
==
Omnibus:                      369.874   Durbin-Watson:                   1.9
68
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              880.5
15
Skew:                          -0.714   Prob(JB):                    6.29e-1
92
Kurtosis:                       5.237   Cond. No.                         12
8.
==============================================================================
==

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

```
x_both_multi_model.summary()=<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
=============================================================================
==
Dep. Variable:            birthweight   R-squared:                       0.0
89
Model:                            OLS   Adj. R-squared:                  0.0
87
Method:                 Least Squares   F-statistic:                     48.
74
Date:                Sun, 10 Aug 2025   Prob (F-statistic):           2.32e-
57
Time:                        20:08:12   Log-Likelihood:                -2326
8.
No. Observations:                3000   AIC:                         4.655e+
04
Df Residuals:                    2993   BIC:                         4.659e+
04
Df Model:                           6
Covariance Type:            nonrobust
=============================================================================
==
                 coef    std err          t      P>|t|      [0.025      0.97
5]
-----------------------------------------------------------------------------
--
const       3199.4264     83.337     38.392      0.000    3036.023    3362.8
30
smoker      -176.9589     27.474     -6.441      0.000    -230.828    -123.0
89
alcohol      -14.7583     75.839     -0.195      0.846    -163.460     133.9
43
nprevist      29.7751      2.926     10.175      0.000      24.037      35.5
13
unmarried   -199.3195     28.396     -7.019      0.000    -254.997    -143.6
42
age           -2.4935      2.269     -1.099      0.272      -6.942       1.9
55
educ           0.2380      5.542      0.043      0.966     -10.629      11.1
05
=============================================================================
==
Omnibus:                      366.140   Durbin-Watson:                   1.9
69
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              872.1
90
Skew:                          -0.707   Prob(JB):                    4.04e-1
90
Kurtosis:                       5.231   Cond. No.                         26
6.
=============================================================================
==

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

```
ctly specified.
"""
```

We added in age, education, and then both to the model.

Overall, we find that adding age has a very little effect on the birthweight with single-digit coefficients in all contexts, albeit with a relatively high p-value. Adding in education also did not have a statistically significant effect, even smaller than that of age when it comes to coefficient. Both of these had little impact on the model's R^2 value as well, so it's safe to say they did not contribute much. The effect of the other factors stayed relatively constant, though.

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.

- (i) For California in 1989, there is some generalizability and validity. The year is the same, and the country is the same, but the socioeconomic structure of the state, sociocultural norms around education, marriage, smoking, and alcohol, and even other confounding factors such as climate. Still, there is some validity to the data.
- (ii) For Illinois in 2015, there is limited generalizability and validity. The country is still the same, and the state is geographically similar, but there are a lot of differences in the 26 years that have passed when it comes to medical advice and cultural norms around relevant factors.
- (iii) For South Korea in 2014, there is little if any generalizability and validity. The country and demographics and culture are completely different, and completely different medical systems and norms.

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)

Across the different regressions that we ran, we found that with more controls, the regression with more variables had smoking at significantly different coefficients and impact on birthweight relative to the other regressions that were ran with fewer variables in parts (2A) and (2C) (and even those two regression models had a pretty nontrivial difference in coefficient). This shows what we hoped to determine, which is that isolating the effect that smoking has on birthweight is difficult to isolate and can vary wildly due to other confounding variables that may be interlinked for other reasons. As such, the earlier models with higher absolute value coefficients for maternal smoking's effect seemed to miss confounding variables. We do see that maternal smoking still has some meaningful effect on child birthweight both statistically and magnitude-wise. However, it is still important to not that this is in the context of several other variables that were considered in the model as well.