Diplomado en Ciencia de Datos UNAM

Modulo 3: Estadística para ciencia de datos

Dr. Roberto Bárcenas C.

Alumno: Ibarra Ramírez Sergio

Practca 1

El conjunto de datos 'Birthweight' contiene la información de 42 bebés al nacer. La pregunta de investigación es saber si existe una relación entre al peso al nacer y el tiempo de gestación. La variable dependiente es Peso al nacer (dada en libras) y la variable independiente para esta actividad es la edad gestacional del bebé al nacer (en semanas).

a) Realiza una descripción gráfica y de medidas estadísticas (descriptivas) de los datos.

Debemos antes que todo, importar los datos de Birthweight

In [2]:
```
import pandas as pd

data_Birthweight_path = "Birthweight.csv"


# Read the CSV file into a DataFrame
data_Birthweight = pd.read_csv(data_Birthweight_path)

# Print the DataFrame
print(data_Birthweight.head(7))
```

```
   ID  Gestation  Birthweight
0   1         44         4.55
1   2         40         4.32
2   3         41         4.10
3   4         44         4.07
4   5         42         3.94
5   6         38         3.93
6   7         40         3.77
```

Demos una "vista resumen " a nuestros datos

In [3]:
```
summary_Birthweight = data_Birthweight.describe()
print(summary_Birthweight)
```

```
              ID  Gestation  Birthweight
count  42.000000  42.000000    42.000000
mean   21.500000  39.190476     3.312857
std    12.267844   2.643336     0.603895
min     1.000000  33.000000     1.920000
25%    11.250000  38.000000     2.940000
50%    21.500000  39.500000     3.295000
75%    31.750000  41.000000     3.647500
max    42.000000  45.000000     4.570000
```

In [4]:
```
import pandas as pd

# Calculate variance
variance = data_Birthweight[['Gestation', 'Birthweight']].var()

# Calculate standard deviation
std_deviation = data_Birthweight[['Gestation', 'Birthweight']].std()

print("Variance:")
print(variance)

print("\nStandard Deviation:")
print(std_deviation)
```

```
Variance:
Gestation      6.987224
Birthweight    0.364689
dtype: float64

Standard Deviation:
Gestation      2.643336
Birthweight    0.603895
dtype: float64
```

Se observa que tenemos 42 datos de Gestation y Birthweight

- Para el caso de Gestation. La media es de 39.1 semanas, casi igual que la mediana (o percentil 59'0) que es de 39.5 semanas. Un valor minimo de 39 y maximo de 45 (Varianza de 6.9 semanas y sd de 2.6)
- Para el caso de Birthweight. La media es de 3.31 kg, casi igual que la mediana (o percentil 59'0) que es de 3.29 kg. Un valor minimo de 1.9 y maximo de 4.5 (Varianza de 0.36 semanas y sd de 0.6)

Utilicemos la libreria de seaborn para generar un gráfico que nos resuma

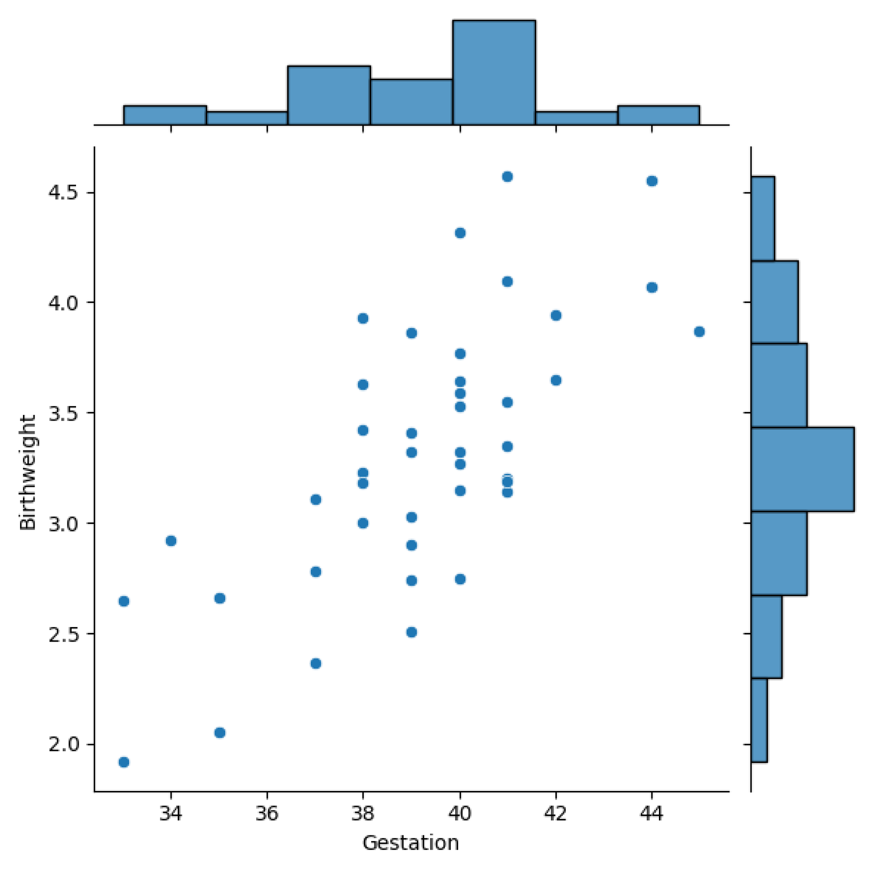
1. La relación que existe entre x = "Gestation" & y = "Birthweight"
2. La distribución individual de cada variable x & y

In [5]:
```
import seaborn as sns
import matplotlib.pyplot as plt

# Specify the DataFrame and variables
df = data_Birthweight
x = "Gestation"
y = "Birthweight"

# Create the scatter plot with histograms
sns.jointplot(x=x, y=y, data=df, kind="scatter")

# Show the plot
plt.show()
```



Se puede observar una relación "aproxiadamete lineal" positiva de Birthweight como function de Gestation. La variable de Birthweight parece estar "normalmente distribuida" mientras que la variable Gestation parece tener un sesgo a la derecha. Analisemos con mayor detalle la relación lineal entre ambas variables así como la normalidad de la variable dependiente

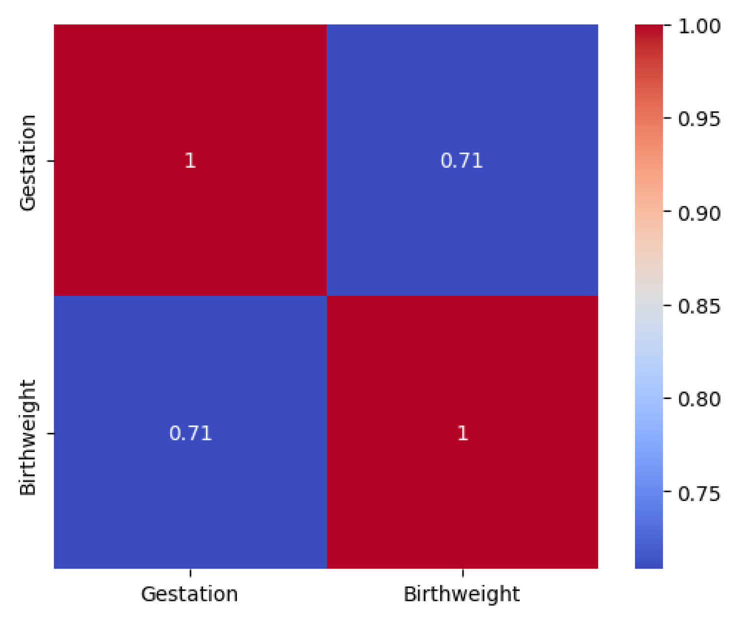
Para el análisis de la relación lineal entre ambas vaiable, determinamos su coeficiente de relacioón de Pearson

In [6]:
```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt



selected_columns = data_Birthweight.iloc[:, 1:3]  # Select columns 1 and 2 (indexing starts from 0)
cor_matrix_selected = selected_columns.corr()

sns.heatmap(cor_matrix_selected, annot=True, cmap='coolwarm', square=True)
plt.show()
```



Se observa que las variables tienen un coeficiente de relación de Person de 0.71. Lo que nos indicaría que el 71% de la varianza de la variable Birthweight puede ser explicada a tráves de su relación lineal con la variable Gestation

b) Realiza un análisis de regresión lineal y proporcionar estimadores puntuales de los parámetros

In [7]:
```
import statsmodels.api as sm

# Fit the linear regression model
X = data_Birthweight['Gestation']
y = data_Birthweight['Birthweight']

X = sm.add_constant(X)  # Add a constant term to the predictor variable
model = sm.GLM(y, X, family=sm.families.Gaussian(link=sm.families.links.identity()))
modelo_lineal_Birthweight = model.fit()

# Print the model summary
print(modelo_lineal_Birthweight.summary())
```

```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:            Birthweight   No. Observations:                   42
Model:                            GLM   Df Residuals:                       40
Model Family:                Gaussian   Df Model:                            1
Link Function:               identity   Scale:                         0.18627
Method:                          IRLS   Log-Likelihood:                -23.279
Date:                Sat, 20 May 2023   Deviance:                       7.4508
Time:                        13:11:53   Pearson chi2:                     7.45
No. Iterations:                     3   Pseudo R-squ. (CS):             0.6171
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -3.0289      1.002     -3.024      0.002      -4.992      -1.066
Gestation      0.1618      0.025      6.346      0.000       0.112       0.212
==============================================================================
```

La ecuación de relación lineal de Birthweight como función de Gestation quedaría:

$$y = -3.029 + 0.162 * x$$

Donde

y : Birthweigh [kg]

x : Gestation [Semanas]

c) Usando el error estándar, establece intervalos de confianza al 95% para los parámetros de la regresión.

In [8]:
```
# Get the confidence intervals
confidence_intervals = modelo_lineal_Birthweight.conf_int()
print(confidence_intervals)
```

```
                  0         1
const     -4.991878 -1.065895
Gestation  0.111841  0.211796
```

RESOLVAMOS LO MISMO DE MANERA ANALÍTICA (LOS VALORES PUEDEN DIFERIR UN POCO DEBIDO AL REDONDEO Y A LA BUSQUEDA EN LA TABLA t)

In [9]:
```
const = -3.028
std_err_for_const = 1.002
Gestation = 0.1618
std_err_for_Gestation = 0.025
```

Calculando los grados de libretad

In [10]:
```
n = 42 # Numero de datos que tenemos
p = 1 # Número de parámetros a determinar
```

In [11]:
```
gl = n - p -1
print(gl)
```

```
40
```

Utilizando la tabla de distribución t, el valor crítico para un nivel de confianza del 95 % con 40 grados de libertad es aproximadamente 2,021. Ahora, calculemos los intervalos de confianza para cada parámetro:

In [12]:
```
critical_value = 2.021
```

Para el termino constante u ordenada al origen

In [13]:
```
Lower_bound_constant = const - (critical_value*std_err_for_const)
print(Lower_bound_constant)


Upper_bound_constant = const + (critical_value*std_err_for_const)
print(Upper_bound_constant)
```

```
-5.053042
-1.002958
```

Para el termino del coeficiente de la variable independente Gestation

In [14]:
```
Lower_bound_gestation = Gestation - (critical_value*std_err_for_Gestation)
print(Lower_bound_gestation)


Upper_bound_gestation = Gestation + (critical_value*std_err_for_Gestation)
print(Upper_bound_gestation )
```

```
0.111275
0.21232499999999999
```

d) Realiza las pruebas de hipótesis para los parámetros y para determinar la significancia de la regresión

En realidad el método sm.GLM ya nos arrojá si los parámetros son estadísticamente significativos mediante la prueba t

In [15]:
```
# Print the model summary
print(modelo_lineal_Birthweight.summary())
```

```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:            Birthweight   No. Observations:                   42
Model:                            GLM   Df Residuals:                       40
Model Family:                Gaussian   Df Model:                            1
Link Function:               identity   Scale:                         0.18627
Method:                          IRLS   Log-Likelihood:                -23.279
Date:                Sat, 20 May 2023   Deviance:                       7.4508
Time:                        13:12:03   Pearson chi2:                     7.45
No. Iterations:                     3   Pseudo R-squ. (CS):             0.6171
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -3.0289      1.002     -3.024      0.002      -4.992      -1.066
Gestation      0.1618      0.025      6.346      0.000       0.112       0.212
==============================================================================
```

Y como se puede observar, los valores de p para ambas variables es < 0.05, lo que indica que son estadísticamente significativos Calculemos dicho valor de p y la signifcacia estadiastica de manera más explicita

In [17]:
```
import statsmodels.api as sm

# Realizar la prueba de hipótesis para los parámetros
t_const = modelo_lineal_Birthweight.params['const'] / modelo_lineal_Birthweight.bse['const']
t_Gestation = modelo_lineal_Birthweight.params['Gestation'] / modelo_lineal_Birthweight.bse['Gestation']

# Comparar con el valor crítico t
valor_critico_t = 2.01  # Obtener el valor crítico t correspondiente a un nivel de significancia del 95% y 40 grados de libertad
alpha = 0.05

if abs(t_const) > valor_critico_t:
    print("El parámetro constante es significativo")
else:
    print("El parámetro constante no es significativo")

if abs(t_Gestation) > valor_critico_t:
    print("El parámetro Gestation es significativo")
else:
    print("El parámetro Gestation no es significativo")
```

```
El parámetro constante es significativo
El parámetro Gestation es significativo
```

Para el caso de la REGRESION la significacia se determina Calculando el valor de F como función de los errores SSE y SSR

In [16]:
```
# Realizar la prueba F para la significancia de la regresión

SSR = modelo_lineal_Birthweight.deviance
SSE = modelo_lineal_Birthweight.deviance * modelo_lineal_Birthweight.df_resid

k = 2  # Número de coeficientes
n = 42  # Número de observaciones

valor_critico_F = 3.24  # Obtener el valor crítico F correspondiente a un nivel de significancia del 95% y 2 y 39 grados de libertad

F = (SSR / k) / (SSE / (n - k - 1))

if F > valor_critico_F:
    print("La regresión es significativa")
else:
    print("La regresión no es significativa")
```

```
La regresión no es significativa
```

Como podría quizá haberse intuido con un valor de R2 = 0.71 que la regresión NO ES ESTADÍSTICAMENTE SIGNIFICATIVA y NO SERÍA SUFICIENTE EXPLICAR LA VARIACIÓN TOTAL DEL BIRTHWEIGHT COMO FUNCIÓN DEL GESTATION

e) Con base en tu análisis, concluye sobre el contexto del problema y responde la pregunta de investigación

¿ Existe una relación entre al peso al nacer y el tiempo de gestación. La variable dependiente es Peso al nacer (dada en libras) y la variable independiente para esta actividad es la edad gestacional del bebé al nacer (en semanas).?

SI EXISTE UNA RELACIÓN LINEAL CON UN R2 = 0.71, PERO DICHA RELACIÓN NO ES ESTADÍSTICAMENTE SIGNIFICATIVA