Actividad: Regresión Lineal 2

Tania Sayuri Guizado Hernandez A01640092

Regresión lineal múltiple

Utiliza un modelo de regresión lineal múltiple para predecir el radio del tumor. Las variables regresoras de tu modelo deben de ser todas las variables de la base de datos.

Importamos las librerías necesarias para el desarrollo de la actividad.

In [1]:
```
# Cargamos las librerias necesarias para la actividad
import pandas as pd
import numpy as np
import statsmodels.api as sm
from sklearn.model_selection import train_test_split
import statsmodels.formula.api as smf
from sklearn.preprocessing import StandardScaler
from sklearn.preprocessing import LabelEncoder
import seaborn as sns
import matplotlib.pyplot as plt
from scipy.stats import t
import scipy.stats as stats
df = pd.read_csv('breast_cancer.csv')
df.drop(['id','diagnosis','concave points_mean','concave points_se','concave points_worst'],ax

class color:
    PURPLE = '\033[95m'
    CYAN = '\033[96m'
    DARKCYAN = '\033[36m'
    BLUE = '\033[94m'
    GREEN = '\033[92m'
    YELLOW = '\033[93m'
    RED = '\033[91m'
    BOLD = '\033[1m'
    UNDERLINE = '\033[4m'
    END = '\033[0m'
```

1.- Base de datos completa. No se observan valores faltantes. En caso de haberlos se realiza imputación simple.

Confirmamos que la base de datos este completa y no contenga valores nulos. Ya que de lo contrario será necesario realizar una imputación de datos.

In [2]:
```
df.isnull().sum()
```

Out[2]:
```
radius_mean               0
texture_mean              0
perimeter_mean            0
area_mean                 0
smoothness_mean           0
compactness_mean          0
concavity_mean            0
symmetry_mean             0
fractal_dimension_mean    0
radius_se                 0
texture_se                0
perimeter_se              0
area_se                   0
smoothness_se             0
compactness_se            0
concavity_se              0
symmetry_se               0
fractal_dimension_se      0
radius_worst              0
texture_worst             0
perimeter_worst           0
area_worst                0
smoothness_worst          0
compactness_worst         0
concavity_worst           0
symmetry_worst            0
fractal_dimension_worst   0
dtype: int64
```

2.-Mostrar que las variables regresoras son independientes. En caso de no serlo realizar el procedimiento correspondiente.

Guardamos la correlación y se verifica que no exista alta ni baja correlación entre las variables.

In [3]:
```
# Guardar la correlacion
correlacion= df.corr()
# Verificamos la alta correlacion
alta_corr=np.where((correlacion>0.95)&(correlacion<1))
alta_corr
```

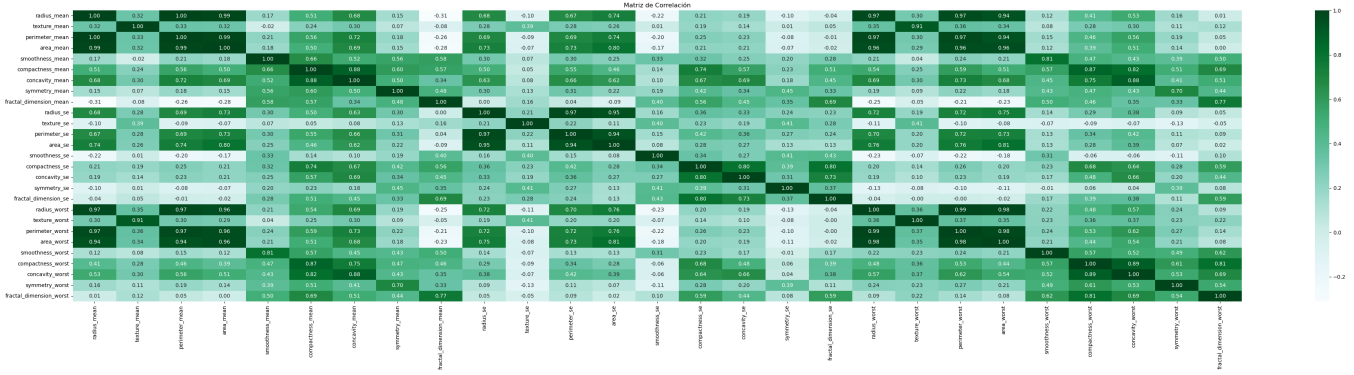
Out[3]:
```
(array([ 0,  0,  0,  0,  2,  2,  2,  2,  3,  3,  3,  3,  3,  9,  9, 11, 12,
        18, 18, 18, 18, 18, 20, 20, 20, 20, 20, 21, 21, 21], dtype=int64),
 array([ 2,  3, 18, 20,  0,  3, 18, 20,  0,  2, 18, 20, 21, 11, 12,  9,  9,
         0,  2,  3, 20, 21,  0,  2,  3, 18, 21,  3, 18, 20], dtype=int64))
```

In [4]:
```
# Verificamos la baja correlacion
baja_corr = np.where((correlacion < -0.95) & (correlacion > -1))
baja_corr
```

Out[4]:
```
(array([], dtype=int64), array([], dtype=int64))
```

En la matriz de correlación se aprecia mejor entre que variables se esta presentando tanto bajas como altas correlaciones.

In [5]:
```
# Ploteamos la matriz de correlacion
plt.figure(figsize=(50, 10))
sns.heatmap(correlacion, annot=True, cmap="BuGn", fmt=".2f", linewidths=0)
plt.title("Matriz de Correlación")
plt.show()
```



Como los datos mostraron alta correlación fue necesario hacer una estandarización de los datos

In [6]:
```
# Estandarizacion de los datos
scaler = StandardScaler()
df_estandar=scaler.fit_transform(df)
df_estandar=pd.DataFrame(df_estandar,columns=df.columns)
columns_names = df.columns.values
```

In [7]:
```
#Entrenamiento y prueba del modelo
entrenamiento, prueba = train_test_split(df_estandar, test_size=0.2, random_state=42)

#Modelo OLS
modelo = smf.ols(formula='radius_mean~texture_mean+perimeter_mean+area_mean+smoothness_mean+co
modelo = modelo.fit()
print(modelo.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:            radius_mean   R-squared:                       1.000
Model:                            OLS   Adj. R-squared:                  1.000
Method:                 Least Squares   F-statistic:                 6.611e+04
Date:                Sat, 16 Sep 2023   Prob (F-statistic):               0.00
Time:                        00:08:51   Log-Likelihood:                 1240.8
No. Observations:                 455   AIC:                            -2428.
Df Residuals:                     428   BIC:                            -2316.
Df Model:                          26
Covariance Type:            nonrobust
============================================================================================
                               coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------
Intercept                    0.0005      0.001      0.630      0.529      -0.001       0.002
texture_mean                -0.0016      0.003     -0.598      0.550      -0.007       0.004
perimeter_mean               0.9492      0.018     54.007      0.000       0.915       0.984
area_mean                    0.0715      0.013      5.299      0.000       0.045       0.098
smoothness_mean              0.0067      0.002      3.253      0.001       0.003       0.011
compactness_mean            -0.0565      0.005    -11.860      0.000      -0.066      -0.047
concavity_mean              -0.0363      0.004     -8.830      0.000      -0.044      -0.028
symmetry_mean                0.0038      0.002      2.443      0.015       0.001       0.007
fractal_dimension_mean       0.0072      0.003      2.382      0.018       0.001       0.013
radius_se                    0.0045      0.006      0.694      0.488      -0.008       0.017
texture_se               -9.373e-05      0.002     -0.058      0.953      -0.003       0.003
perimeter_se                -0.0163      0.006     -2.742      0.006      -0.028      -0.005
area_se                      0.0006      0.004      0.129      0.897      -0.008       0.009
smoothness_se                0.0014      0.001      0.958      0.338      -0.001       0.004
compactness_se              -0.0018      0.003     -0.662      0.508      -0.007       0.004
concavity_se                 0.0144      0.002      6.440      0.000       0.010       0.019
symmetry_se                  0.0044      0.002      2.462      0.014       0.001       0.008
fractal_dimension_se        -0.0032      0.002     -1.415      0.158      -0.008       0.001
radius_worst                 0.2323      0.018     12.784      0.000       0.197       0.268
texture_worst                0.0002      0.003      0.059      0.953      -0.006       0.007
perimeter_worst             -0.1139      0.015     -7.626      0.000      -0.143      -0.085
area_worst                  -0.0840      0.013     -6.369      0.000      -0.110      -0.058
smoothness_worst            -0.0049      0.002     -2.064      0.040      -0.010      -0.000
compactness_worst            0.0157      0.005      3.477      0.001       0.007       0.025
concavity_worst              0.0010      0.004      0.268      0.788      -0.007       0.009
symmetry_worst              -0.0048      0.002     -2.069      0.039      -0.009      -0.000
fractal_dimension_worst     -0.0035      0.003     -1.050      0.294      -0.010       0.003
==============================================================================
Omnibus:                       46.518   Durbin-Watson:                   2.076
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              200.782
Skew:                           0.306   Prob(JB):                     2.52e-44
Kurtosis:                       6.196   Cond. No.                         120.
==============================================================================

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

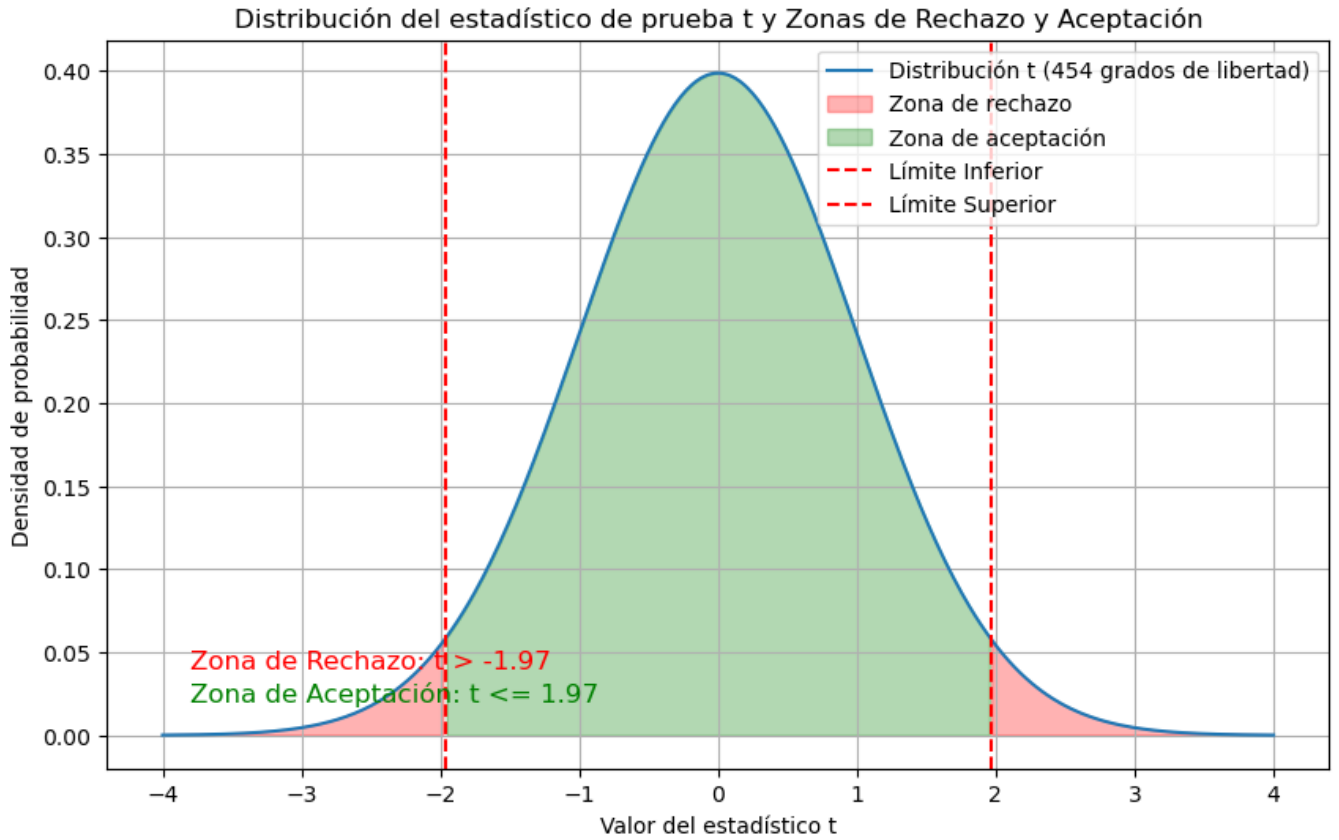
Haciendo un análisis del anterior modelo, el valor del estadístico F es extremadamente alto y la probabilidad asociada cercana a cero indican que el modelo en su conjunto es estadísticamente significativo. Esto sugiere que al menos una de las variables independientes incluidas en el modelo tiene un impacto significativo en la variable dependiente (radius_mean). Aparte el valor R^2 es 1, o que indica que el modelo se ajusta perfectamente a los datos de entrenamiento.

3.-Hipótesis nula de los coeficientes de regresión. Estadístico de prueba, distribución del estadístico de prueba.

Para un 95% de confianza realiza un diagrama en donde se muestre la distribución del estadístico de prueba, la zona de aceptación y la zona de rechazo.

In [8]:
```
# Calcular los estadísticos t
nivel_de_confianza = 0.95
alpha2 = (1 - nivel_de_confianza) / 2
grados_de_libertad = len(entrenamiento) - 1
valor_critico_t = t.ppf(1 - alpha2, df=grados_de_libertad)

limite_inferior = -valor_critico_t
limite_superior = valor_critico_t
```

In [9]:
```
# Rango de valores para el estadístico t
rango_t = np.linspace(-4, 4, 400)
densidad_t = t.pdf(rango_t, df=grados_de_libertad)  # Distribución t-Student

# Crear el gráfico
plt.figure(figsize=(10, 6))
plt.plot(rango_t, densidad_t, label=f'Distribución t ({grados_de_libertad} grados de libertad
plt.fill_between(rango_t, 0, densidad_t, where=(rango_t < limite_inferior) | (rango_t > limite
plt.fill_between(rango_t, 0, densidad_t, where=(rango_t >= limite_inferior) & (rango_t <= lim

# Agrega líneas verticales para el estadístico de prueba y los límites
plt.axvline(limite_inferior, color='red', linestyle='--', label='Límite Inferior')
plt.axvline(limite_superior, color='red', linestyle='--', label='Límite Superior')

# Etiquetas y leyenda
plt.title('Distribución del estadístico de prueba t y Zonas de Rechazo y Aceptación')
plt.xlabel('Valor del estadístico t')
plt.ylabel('Densidad de probabilidad')
plt.legend()
plt.grid()

# Etiquetas en el gráfico
plt.text(-3.8, 0.04, f'Zona de Rechazo: t > {limite_inferior:.2f}', fontsize=12, color='red')
plt.text(-3.8, 0.02, f'Zona de Aceptación: t <= {limite_superior:.2f}', fontsize=12, color='g

# Mostrar el gráfico
plt.show()
```

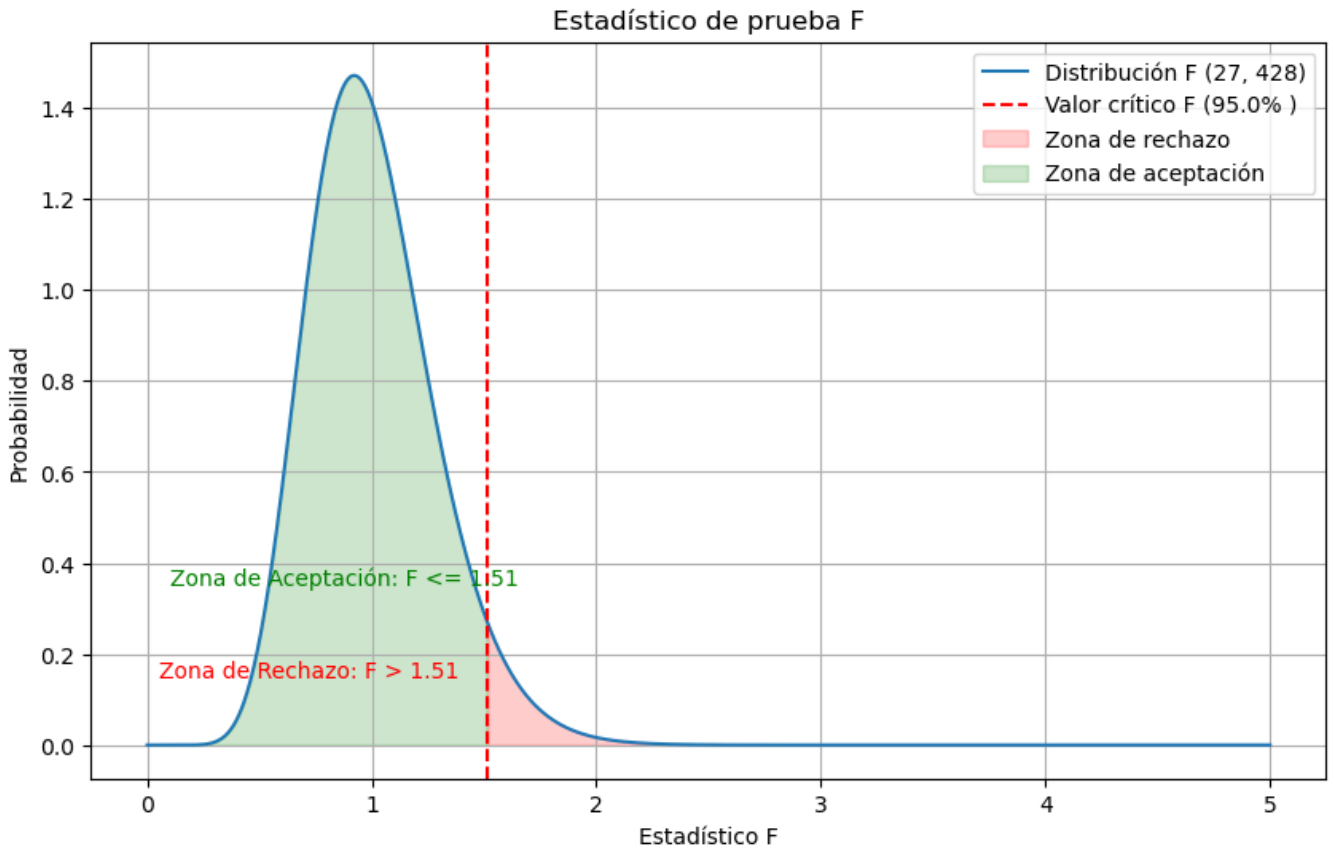


Para este punto la prueba de hipótesis la realizamos con el estadístico de prueba y distribución t-Student. Tomando en consideración que teniamos que tener un nivel de confianza del 95% el calculo resultando del valor crítico fue de aproximadamente 1.9652 y este valor sería el que define la zona de rechazo y aceptación(-1.97,1.97) en la distribución t. En esta misma, con 454 grados de libertad, se asemeja una distribución normal estándar debido a la muestra tan grande de la base de datos de breast_cancer.

Sobre el gráfico que la curva asemeja una campana, los valores críticos de t se representan por los límites inferior y superior. Los valores del estadístico de prueba que se sitúan en la zona de rechazo (en rojo) muestran pruebas suficientes para refutar la hipótesis nula, mientras que los que se sitúan dentro de la zona de aceptación (en verde) no muestran pruebas significativas.

4.-Hipótesis nula de la significancia del modelo (prueba F-Fisher). Menciona que distribución tiene el estadístico de prueba con qué número de grados de libertad. Para un 95% de confianza realiza un diagrama en donde se muestre la distribución del estadístico de prueba, la zona de aceptación y la zona de rechazo.

In [10]:
```
from scipy.stats import f
coeficientes = modelo.params
# Grados de libertad del modelo y del error
df_model = len(coeficientes)
df_error = len(entrenamiento) - len(coeficientes)

# Nivel de confianza
nivel_de_confianza = 0.95

# Valor crítico F para el nivel de confianza y grados de libertad
valor_critico_F = f.ppf(nivel_de_confianza, dfn=df_model, dfd=df_error)

rango_F = np.linspace(0, 5, 1000)
densidad_F = f.pdf(rango_F, dfn=df_model, dfd=df_error)

plt.figure(figsize=(10, 6))
plt.plot(rango_F, densidad_F, label=f'Distribución F ({df_model}, {df_error})')
plt.axvline(x=valor_critico_F, color='red', linestyle='--', label=f'Valor crítico F ({nivel_d
plt.fill_between(rango_F, densidad_F, where=((rango_F > valor_critico_F)), color='red', alpha=
plt.fill_between(rango_F, densidad_F, where=((rango_F <= valor_critico_F)), color='green', al

plt.title('Estadístico de prueba F')
plt.xlabel('Estadístico F')
plt.ylabel('Probabilidad')
plt.legend()
plt.grid()

plt.text(0.05, 0.15, f'Zona de Rechazo: F > {valor_critico_F:.2f}', fontsize=10, color='red')
plt.text(0.1, 0.35, f'Zona de Aceptación: F <= {valor_critico_F:.2f}', fontsize=10, color='gr

plt.show()
```



En este caso para el estadísitico de prueba F y su distribución hubo 27 grados de libertad en el modelo y 428 en el error. Considerando el solicitado de nivel de confiaza del 95% el valor crítico F fue de aproximadamente 1.512. En el gráfico solo se muestra la cola derecha, que suele ser típica en este estadístico y sus valores del mismo mayores al valor crítico deberán caer en la zona de rechazo (inidcaran que al menos una vriable independiente es significativa) mientras que los valores menores caen en la zona de aceptación (afirmando que las variables no son significativas).

5.- Realiza un modelo de regresión hacia atrás (backward). Explica el criterio para ir eliminando variables del modelo.

Explicando un poco mejor el modelo de regresión backward, en pocas palabras, es una técnica utilizada para simplificar un modelo de regresión múltiple eliminando gradualmente las variables predictoras que tienen un impacto menos significativo en la predicción de la variable dependiente. Su significancia se evalúa si su p-valor es mayor a 0.05.

En cada paso, **se eliminará la variable independiente** *menos significativa*, y el modelo se ajustará nuevamente sin esa variable.

In [11]:

```
# Tomamos las variables para X y Y en nuestro modelo de regresion backward
df_estandar['intercept'] = 1
X = df_estandar[[ 'texture_mean', 'perimeter_mean', 'area_mean', 'smoothness_mean',
    'compactness_mean', 'concavity_mean', 'symmetry_mean', 'fractal_dimension_mean',
    'radius_se', 'perimeter_se', 'area_se', 'smoothness_se',
    'compactness_se', 'concavity_se', 'symmetry_se', 'fractal_dimension_se',
    'radius_worst', 'texture_worst', 'perimeter_worst', 'area_worst',
    'smoothness_worst', 'compactness_worst', 'concavity_worst', 'symmetry_worst',
    'fractal_dimension_worst']]
Y = df_estandar['radius_mean']

modelo2 = sm.OLS(Y, X).fit()

while any(modelo2.pvalues > 0.05):
    variable_menos_significativa = modelo2.pvalues.idxmax()
    X = X.drop(variable_menos_significativa, axis=1)
    modelo2 = sm.OLS(Y, X).fit()
    print(color.BOLD + '\n--- La variable eliminada fue: '+ color.BLUE + variable_menos_signi
    print('\n')
    print(modelo2.summary())
```

--- La variable eliminada fue: texture_worst

OLS Regression Results

Dep. Variable:	radius_mean	R-squared (uncentered):	1.000
Model:	OLS	Adj. R-squared (uncentered):	1.000
Method:	Least Squares	F-statistic:	8.357e+04
Date:	Sat, 16 Sep 2023	Prob (F-statistic):	0.00
Time:	00:08:52	Log-Likelihood:	1528.7
No. Observations:	569	AIC:	-3009.
Df Residuals:	545	BIC:	-2905.
Df Model:	24		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
texture_mean	-0.0012	0.001	-1.443	0.150	-0.003	0.000
perimeter_mean	0.9379	0.016	60.464	0.000	0.907	0.968
area_mean	0.0797	0.012	6.675	0.000	0.056	0.103
smoothness_mean	0.0088	0.002	4.746	0.000	0.005	0.012
compactness_mean	-0.0570	0.004	-13.277	0.000	-0.065	-0.049
concavity_mean	-0.0341	0.004	-8.922	0.000	-0.042	-0.027
symmetry_mean	0.0029	0.001	2.020	0.044	7.87e-05	0.006
fractal_dimension_mean	0.0037	0.003	1.333	0.183	-0.002	0.009
radius_se	0.0093	0.006	1.573	0.116	-0.002	0.021
perimeter_se	-0.0191	0.005	-3.506	0.000	-0.030	-0.008
area_se	-0.0013	0.004	-0.308	0.759	-0.009	0.007
smoothness_se	0.0025	0.001	1.818	0.070	-0.000	0.005
compactness_se	-0.0015	0.003	-0.570	0.569	-0.007	0.004
concavity_se	0.0139	0.002	6.578	0.000	0.010	0.018
symmetry_se	0.0005	0.001	0.346	0.730	-0.002	0.003
fractal_dimension_se	-0.0018	0.002	-0.837	0.403	-0.006	0.002
radius_worst	0.2181	0.017	12.823	0.000	0.185	0.251
perimeter_worst	-0.0985	0.013	-7.379	0.000	-0.125	-0.072
area_worst	-0.0839	0.012	-7.208	0.000	-0.107	-0.061
smoothness_worst	-0.0073	0.002	-3.417	0.001	-0.012	-0.003
compactness_worst	0.0158	0.004	3.783	0.000	0.008	0.024
concavity_worst	-0.0004	0.004	-0.106	0.916	-0.008	0.007
symmetry_worst	-0.0027	0.002	-1.304	0.193	-0.007	0.001
fractal_dimension_worst	-0.0025	0.003	-0.804	0.422	-0.008	0.004

Omnibus:	76.188	Durbin-Watson:	1.916
Prob(Omnibus):	0.000	Jarque-Bera (JB):	691.023
Skew:	-0.135	Prob(JB):	8.84e-151
Kurtosis:	8.392	Cond. No.	116.

Notes:
[1] R² is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.

--- La variable eliminada fue: concavity_worst

OLS Regression Results

Dep. Variable:	radius_mean	R-squared (uncentered):	1.000
Model:	OLS	Adj. R-squared (uncentered):	1.000
Method:	Least Squares	F-statistic:	8.737e+04
Date:	Sat, 16 Sep 2023	Prob (F-statistic):	0.00
Time:	00:08:52	Log-Likelihood:	1528.7
No. Observations:	569	AIC:	-3011.
Df Residuals:	546	BIC:	-2911.
Df Model:	23		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
texture_mean	-0.0012	0.001	-1.444	0.149	-0.003	0.000
perimeter_mean	0.9378	0.015	60.544	0.000	0.907	0.968
area_mean	0.0797	0.012	6.709	0.000	0.056	0.103
smoothness_mean	0.0088	0.002	4.759	0.000	0.005	0.012
compactness_mean	-0.0569	0.004	-13.736	0.000	-0.065	-0.049
concavity_mean	-0.0343	0.003	-11.349	0.000	-0.040	-0.028
symmetry_mean	0.0029	0.001	2.020	0.044	7.93e-05	0.006
fractal_dimension_mean	0.0037	0.003	1.350	0.177	-0.002	0.009
radius_se	0.0094	0.006	1.583	0.114	-0.002	0.021
perimeter_se	-0.0192	0.005	-3.512	0.000	-0.030	-0.008
area_se	-0.0013	0.004	-0.306	0.759	-0.009	0.007
smoothness_se	0.0025	0.001	1.843	0.066	-0.000	0.005
compactness_se	-0.0015	0.003	-0.563	0.574	-0.007	0.004
concavity_se	0.0138	0.002	7.480	0.000	0.010	0.017
symmetry_se	0.0005	0.001	0.341	0.733	-0.002	0.003
fractal_dimension_se	-0.0018	0.002	-0.835	0.404	-0.006	0.002
radius_worst	0.2180	0.017	12.834	0.000	0.185	0.251
perimeter_worst	-0.0985	0.013	-7.386	0.000	-0.125	-0.072
area_worst	-0.0840	0.012	-7.218	0.000	-0.107	-0.061
smoothness_worst	-0.0074	0.002	-3.460	0.001	-0.012	-0.003
compactness_worst	0.0156	0.004	4.166	0.000	0.008	0.023
symmetry_worst	-0.0027	0.002	-1.302	0.194	-0.007	0.001
fractal_dimension_worst	-0.0025	0.003	-0.851	0.395	-0.008	0.003

Omnibus:	76.132	Durbin-Watson:	1.917
Prob(Omnibus):	0.000	Jarque-Bera (JB):	689.166
Skew:	-0.136	Prob(JB):	2.24e-150
Kurtosis:	8.385	Cond. No.	112.

Notes:
[1] R² is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.

--- La variable eliminada fue: area_se

OLS Regression Results

Dep. Variable:	radius_mean	R-squared (uncentered):	1.000
Model:	OLS	Adj. R-squared (uncentered):	1.000
Method:	Least Squares	F-statistic:	9.149e+04
Date:	Sat, 16 Sep 2023	Prob (F-statistic):	0.00
Time:	00:08:52	Log-Likelihood:	1528.6
No. Observations:	569	AIC:	-3013.
Df Residuals:	547	BIC:	-2918.
Df Model:	22		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
texture_mean	-0.0011	0.001	-1.426	0.155	-0.003	0.000
perimeter_mean	0.9363	0.015	64.087	0.000	0.908	0.965
area_mean	0.0803	0.012	6.853	0.000	0.057	0.103
smoothness_mean	0.0088	0.002	4.768	0.000	0.005	0.012
compactness_mean	-0.0568	0.004	-13.760	0.000	-0.065	-0.049
concavity_mean	-0.0343	0.003	-11.354	0.000	-0.040	-0.028
symmetry_mean	0.0029	0.001	2.119	0.035	0.000	0.006
fractal_dimension_mean	0.0038	0.003	1.395	0.164	-0.002	0.009
radius_se	0.0086	0.005	1.619	0.106	-0.002	0.019
perimeter_se	-0.0195	0.005	-3.678	0.000	-0.030	-0.009
smoothness_se	0.0025	0.001	1.888	0.060	-0.000	0.005
compactness_se	-0.0015	0.003	-0.565	0.572	-0.007	0.004
concavity_se	0.0138	0.002	7.482	0.000	0.010	0.017
symmetry_se	0.0006	0.001	0.425	0.671	-0.002	0.003
fractal_dimension_se	-0.0017	0.002	-0.816	0.415	-0.006	0.002

```
radius_worst                0.2201      0.016     14.187      0.000       0.190       0.251
perimeter_worst            -0.0977      0.013     -7.474      0.000      -0.123      -0.072
area_worst                 -0.0860      0.010     -9.016      0.000      -0.105      -0.067
smoothness_worst           -0.0074      0.002     -3.516      0.000      -0.012      -0.003
compactness_worst           0.0155      0.004      4.161      0.000       0.008       0.023
symmetry_worst             -0.0028      0.002     -1.415      0.158      -0.007       0.001
fractal_dimension_worst    -0.0026      0.003     -0.884      0.377      -0.008       0.003
==============================================================================
Omnibus:                       75.876   Durbin-Watson:                   1.917
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              685.394
Skew:                          -0.132   Prob(JB):                    1.47e-149
Kurtosis:                       8.370   Cond. No.                         99.0
==============================================================================

Notes:
[1] R² is computed without centering (uncentered) since the model does not contain a constan
t.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

--- La variable eliminada fue: symmetry_se

```
                                 OLS Regression Results
==============================================================================================
Dep. Variable:            radius_mean   R-squared (uncentered):                   1.000
Model:                            OLS   Adj. R-squared (uncentered):              1.000
Method:                 Least Squares   F-statistic:                          9.599e+04
Date:                Sat, 16 Sep 2023   Prob (F-statistic):                        0.00
Time:                        00:08:52   Log-Likelihood:                         1528.5
No. Observations:                 569   AIC:                                    -3015.
Df Residuals:                     548   BIC:                                    -2924.
Df Model:                          21
Covariance Type:            nonrobust
==============================================================================================
                               coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------------
texture_mean                -0.0011      0.001     -1.393      0.164      -0.003       0.000
perimeter_mean               0.9369      0.015     64.490      0.000       0.908       0.965
area_mean                    0.0800      0.012      6.845      0.000       0.057       0.103
smoothness_mean              0.0089      0.002      4.952      0.000       0.005       0.012
compactness_mean            -0.0568      0.004    -13.773      0.000      -0.065      -0.049
concavity_mean              -0.0342      0.003    -11.356      0.000      -0.040      -0.028
symmetry_mean                0.0028      0.001      2.084      0.038       0.000       0.005
fractal_dimension_mean       0.0039      0.003      1.451      0.147      -0.001       0.009
radius_se                    0.0086      0.005      1.617      0.106      -0.002       0.019
perimeter_se                -0.0193      0.005     -3.656      0.000      -0.030      -0.009
smoothness_se                0.0027      0.001      2.183      0.029       0.000       0.005
compactness_se              -0.0014      0.003     -0.515      0.607      -0.007       0.004
concavity_se                 0.0138      0.002      7.479      0.000       0.010       0.017
fractal_dimension_se        -0.0017      0.002     -0.792      0.429      -0.006       0.002
radius_worst                 0.2199      0.015     14.192      0.000       0.189       0.250
perimeter_worst             -0.0983      0.013     -7.568      0.000      -0.124      -0.073
area_worst                  -0.0857      0.010     -9.015      0.000      -0.104      -0.067
smoothness_worst            -0.0077      0.002     -3.897      0.000      -0.012      -0.004
compactness_worst            0.0154      0.004      4.144      0.000       0.008       0.023
symmetry_worst              -0.0022      0.001     -1.604      0.109      -0.005       0.000
fractal_dimension_worst     -0.0028      0.003     -0.969      0.333      -0.009       0.003
==============================================================================
Omnibus:                       73.545   Durbin-Watson:                   1.919
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              641.568
Skew:                          -0.109   Prob(JB):                    4.85e-140
Kurtosis:                       8.197   Cond. No.                         98.6
==============================================================================

Notes:
[1] R² is computed without centering (uncentered) since the model does not contain a constan
t.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

--- La variable eliminada fue: compactness_se

OLS Regression Results

Dep. Variable:	radius_mean	R-squared (uncentered):	1.000
Model:	OLS	Adj. R-squared (uncentered):	1.000
Method:	Least Squares	F-statistic:	1.009e+05
Date:	Sat, 16 Sep 2023	Prob (F-statistic):	0.00
Time:	00:08:52	Log-Likelihood:	1528.4
No. Observations:	569	AIC:	-3017.
Df Residuals:	549	BIC:	-2930.
Df Model:	20		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
texture_mean	-0.0011	0.001	-1.392	0.165	-0.003	0.000
perimeter_mean	0.9364	0.014	64.618	0.000	0.908	0.965
area_mean	0.0804	0.012	6.890	0.000	0.057	0.103
smoothness_mean	0.0090	0.002	4.969	0.000	0.005	0.012
compactness_mean	-0.0573	0.004	-14.181	0.000	-0.065	-0.049
concavity_mean	-0.0339	0.003	-11.533	0.000	-0.040	-0.028
symmetry_mean	0.0028	0.001	2.094	0.037	0.000	0.005
fractal_dimension_mean	0.0040	0.003	1.487	0.138	-0.001	0.009
radius_se	0.0090	0.005	1.718	0.086	-0.001	0.019
perimeter_se	-0.0198	0.005	-3.832	0.000	-0.030	-0.010
smoothness_se	0.0025	0.001	2.139	0.033	0.000	0.005
concavity_se	0.0134	0.002	7.912	0.000	0.010	0.017
fractal_dimension_se	-0.0023	0.002	-1.329	0.184	-0.006	0.001
radius_worst	0.2197	0.015	14.192	0.000	0.189	0.250
perimeter_worst	-0.0973	0.013	-7.582	0.000	-0.123	-0.072
area_worst	-0.0862	0.009	-9.116	0.000	-0.105	-0.068
smoothness_worst	-0.0075	0.002	-3.873	0.000	-0.011	-0.004
compactness_worst	0.0144	0.003	4.589	0.000	0.008	0.021
symmetry_worst	-0.0022	0.001	-1.608	0.108	-0.005	0.000
fractal_dimension_worst	-0.0024	0.003	-0.852	0.395	-0.008	0.003

Omnibus:	72.564	Durbin-Watson:	1.922
Prob(Omnibus):	0.000	Jarque-Bera (JB):	622.855
Skew:	-0.101	Prob(JB):	5.61e-136
Kurtosis:	8.122	Cond. No.	96.4

Notes:
[1] R^2 is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.

--- La variable eliminada fue: fractal_dimension_worst

OLS Regression Results

Dep. Variable:	radius_mean	R-squared (uncentered):	1.000
Model:	OLS	Adj. R-squared (uncentered):	1.000
Method:	Least Squares	F-statistic:	1.063e+05
Date:	Sat, 16 Sep 2023	Prob (F-statistic):	0.00
Time:	00:08:52	Log-Likelihood:	1528.0
No. Observations:	569	AIC:	-3018.
Df Residuals:	550	BIC:	-2935.
Df Model:	19		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
texture_mean	-0.0011	0.001	-1.408	0.160	-0.003	0.000
perimeter_mean	0.9366	0.014	64.650	0.000	0.908	0.965
area_mean	0.0802	0.012	6.880	0.000	0.057	0.103

```
smoothness_mean             0.0092      0.002      5.144      0.000       0.006       0.013
compactness_mean           -0.0560      0.004    -14.924      0.000      -0.063      -0.049
concavity_mean             -0.0343      0.003    -11.829      0.000      -0.040      -0.029
symmetry_mean               0.0028      0.001      2.082      0.038       0.000       0.005
fractal_dimension_mean      0.0027      0.002      1.219      0.223      -0.002       0.007
radius_se                   0.0092      0.005      1.761      0.079      -0.001       0.019
perimeter_se               -0.0199      0.005     -3.861      0.000      -0.030      -0.010
smoothness_se               0.0027      0.001      2.431      0.015       0.001       0.005
concavity_se                0.0138      0.002      8.525      0.000       0.011       0.017
fractal_dimension_se       -0.0030      0.002     -1.912      0.056      -0.006    8.08e-05
radius_worst                0.2180      0.015     14.204      0.000       0.188       0.248
perimeter_worst            -0.0961      0.013     -7.535      0.000      -0.121      -0.071
area_worst                 -0.0858      0.009     -9.086      0.000      -0.104      -0.067
smoothness_worst           -0.0080      0.002     -4.310      0.000      -0.012      -0.004
compactness_worst           0.0125      0.002      5.654      0.000       0.008       0.017
symmetry_worst             -0.0022      0.001     -1.584      0.114      -0.005       0.001
==============================================================================
Omnibus:                       72.913   Durbin-Watson:                   1.922
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              628.194
Skew:                          -0.107   Prob(JB):                    3.88e-137
Kurtosis:                       8.143   Cond. No.                         95.0
==============================================================================

Notes:
[1] R² is computed without centering (uncentered) since the model does not contain a constan
t.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

--- La variable eliminada fue: fractal_dimension_mean

```
                                 OLS Regression Results
================================================================================================
Dep. Variable:            radius_mean   R-squared (uncentered):                   1.000
Model:                            OLS   Adj. R-squared (uncentered):              1.000
Method:                 Least Squares   F-statistic:                          1.121e+05
Date:                Sat, 16 Sep 2023   Prob (F-statistic):                        0.00
Time:                        00:08:52   Log-Likelihood:                         1527.2
No. Observations:                 569   AIC:                                    -3018.
Df Residuals:                     551   BIC:                                    -2940.
Df Model:                          18
Covariance Type:            nonrobust
================================================================================================
                           coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
texture_mean            -0.0012      0.001     -1.475      0.141      -0.003       0.000
perimeter_mean           0.9307      0.014     68.092      0.000       0.904       0.958
area_mean                0.0837      0.011      7.401      0.000       0.061       0.106
smoothness_mean          0.0095      0.002      5.395      0.000       0.006       0.013
compactness_mean        -0.0535      0.003    -16.974      0.000      -0.060      -0.047
concavity_mean          -0.0339      0.003    -11.761      0.000      -0.040      -0.028
symmetry_mean            0.0027      0.001      2.033      0.043    9.12e-05       0.005
radius_se                0.0079      0.005      1.544      0.123      -0.002       0.018
perimeter_se            -0.0189      0.005     -3.708      0.000      -0.029      -0.009
smoothness_se            0.0026      0.001      2.277      0.023       0.000       0.005
concavity_se             0.0135      0.002      8.439      0.000       0.010       0.017
fractal_dimension_se    -0.0019      0.001     -1.485      0.138      -0.004       0.001
radius_worst             0.2211      0.015     14.609      0.000       0.191       0.251
perimeter_worst         -0.0981      0.013     -7.743      0.000      -0.123      -0.073
area_worst              -0.0865      0.009     -9.181      0.000      -0.105      -0.068
smoothness_worst        -0.0078      0.002     -4.222      0.000      -0.011      -0.004
compactness_worst        0.0120      0.002      5.521      0.000       0.008       0.016
symmetry_worst          -0.0021      0.001     -1.515      0.130      -0.005       0.001
==============================================================================
Omnibus:                       71.500   Durbin-Watson:                   1.914
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              603.404
Skew:                          -0.090   Prob(JB):                    9.39e-132
Kurtosis:                       8.042   Cond. No.                         91.3
==============================================================================
```

Notes:
[1] R^2 is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.

--- La variable eliminada fue: texture_mean

OLS Regression Results

Dep. Variable:	radius_mean	R-squared (uncentered):	1.000
Model:	OLS	Adj. R-squared (uncentered):	1.000
Method:	Least Squares	F-statistic:	1.184e+05
Date:	Sat, 16 Sep 2023	Prob (F-statistic):	0.00
Time:	00:08:52	Log-Likelihood:	1526.1
No. Observations:	569	AIC:	-3018.
Df Residuals:	552	BIC:	-2944.
Df Model:	17		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
perimeter_mean	0.9307	0.014	68.019	0.000	0.904	0.958
area_mean	0.0842	0.011	7.442	0.000	0.062	0.106
smoothness_mean	0.0099	0.002	5.702	0.000	0.007	0.013
compactness_mean	-0.0533	0.003	-16.909	0.000	-0.060	-0.047
concavity_mean	-0.0344	0.003	-11.952	0.000	-0.040	-0.029
symmetry_mean	0.0026	0.001	1.978	0.048	1.85e-05	0.005
radius_se	0.0081	0.005	1.589	0.113	-0.002	0.018
perimeter_se	-0.0192	0.005	-3.781	0.000	-0.029	-0.009
smoothness_se	0.0023	0.001	2.084	0.038	0.000	0.004
concavity_se	0.0136	0.002	8.550	0.000	0.010	0.017
fractal_dimension_se	-0.0019	0.001	-1.449	0.148	-0.004	0.001
radius_worst	0.2194	0.015	14.523	0.000	0.190	0.249
perimeter_worst	-0.0972	0.013	-7.675	0.000	-0.122	-0.072
area_worst	-0.0863	0.009	-9.145	0.000	-0.105	-0.068
smoothness_worst	-0.0079	0.002	-4.246	0.000	-0.012	-0.004
compactness_worst	0.0116	0.002	5.382	0.000	0.007	0.016
symmetry_worst	-0.0020	0.001	-1.437	0.151	-0.005	0.001

Omnibus:	73.605	Durbin-Watson:	1.916
Prob(Omnibus):	0.000	Jarque-Bera (JB):	651.590
Skew:	-0.090	Prob(JB):	3.23e-142
Kurtosis:	8.239	Cond. No.	90.6

Notes:
[1] R^2 is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.

--- La variable eliminada fue: symmetry_worst

OLS Regression Results

Dep. Variable:	radius_mean	R-squared (uncentered):	1.000
Model:	OLS	Adj. R-squared (uncentered):	1.000
Method:	Least Squares	F-statistic:	1.256e+05
Date:	Sat, 16 Sep 2023	Prob (F-statistic):	0.00
Time:	00:08:52	Log-Likelihood:	1525.0
No. Observations:	569	AIC:	-3018.
Df Residuals:	553	BIC:	-2949.
Df Model:	16		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]

```
perimeter_mean           0.9347      0.013     69.692      0.000       0.908       0.961
area_mean                0.0814      0.011      7.297      0.000       0.059       0.103
smoothness_mean          0.0102      0.002      5.853      0.000       0.007       0.014
compactness_mean        -0.0531      0.003    -16.847      0.000      -0.059      -0.047
concavity_mean          -0.0343      0.003    -11.926      0.000      -0.040      -0.029
symmetry_mean            0.0013      0.001      1.361      0.174      -0.001       0.003
radius_se                0.0087      0.005      1.701      0.090      -0.001       0.019
perimeter_se            -0.0195      0.005     -3.824      0.000      -0.029      -0.009
smoothness_se            0.0025      0.001      2.305      0.022       0.000       0.005
concavity_se             0.0135      0.002      8.502      0.000       0.010       0.017
fractal_dimension_se    -0.0017      0.001     -1.308      0.191      -0.004       0.001
radius_worst             0.2154      0.015     14.494      0.000       0.186       0.245
perimeter_worst         -0.0970      0.013     -7.654      0.000      -0.122      -0.072
area_worst              -0.0836      0.009     -9.029      0.000      -0.102      -0.065
smoothness_worst        -0.0082      0.002     -4.466      0.000      -0.012      -0.005
compactness_worst        0.0108      0.002      5.182      0.000       0.007       0.015
==============================================================================
Omnibus:                       77.721   Durbin-Watson:                   1.911
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              718.228
Skew:                          -0.153   Prob(JB):                    1.09e-156
Kurtosis:                       8.496   Cond. No.                         87.6
==============================================================================

Notes:
[1] R² is computed without centering (uncentered) since the model does not contain a constan
t.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

--- La variable eliminada fue: fractal_dimension_se

```
                                 OLS Regression Results
================================================================================================
Dep. Variable:            radius_mean   R-squared (uncentered):                   1.000
Model:                            OLS   Adj. R-squared (uncentered):              1.000
Method:                 Least Squares   F-statistic:                          1.338e+05
Date:                Sat, 16 Sep 2023   Prob (F-statistic):                        0.00
Time:                        00:08:52   Log-Likelihood:                         1524.2
No. Observations:                 569   AIC:                                    -3018.
Df Residuals:                     554   BIC:                                    -2953.
Df Model:                          15
Covariance Type:            nonrobust
================================================================================================
                           coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
perimeter_mean           0.9356      0.013     69.794      0.000       0.909       0.962
area_mean                0.0808      0.011      7.243      0.000       0.059       0.103
smoothness_mean          0.0102      0.002      5.859      0.000       0.007       0.014
compactness_mean        -0.0544      0.003    -18.193      0.000      -0.060      -0.049
concavity_mean          -0.0336      0.003    -11.898      0.000      -0.039      -0.028
symmetry_mean            0.0014      0.001      1.456      0.146      -0.000       0.003
radius_se                0.0078      0.005      1.539      0.124      -0.002       0.018
perimeter_se            -0.0187      0.005     -3.691      0.000      -0.029      -0.009
smoothness_se            0.0022      0.001      2.057      0.040    9.93e-05       0.004
concavity_se             0.0125      0.001      9.088      0.000       0.010       0.015
radius_worst             0.2152      0.015     14.478      0.000       0.186       0.244
perimeter_worst         -0.0967      0.013     -7.623      0.000      -0.122      -0.072
area_worst              -0.0837      0.009     -9.031      0.000      -0.102      -0.065
smoothness_worst        -0.0079      0.002     -4.324      0.000      -0.012      -0.004
compactness_worst        0.0107      0.002      5.130      0.000       0.007       0.015
==============================================================================
Omnibus:                       74.569   Durbin-Watson:                   1.911
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              666.014
Skew:                          -0.109   Prob(JB):                    2.38e-145
Kurtosis:                       8.296   Cond. No.                         87.1
==============================================================================

Notes:
[1] R² is computed without centering (uncentered) since the model does not contain a constan
```

t.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.

--- La variable eliminada fue: symmetry_mean

OLS Regression Results

Dep. Variable:	radius_mean	R-squared (uncentered):	1.000
Model:	OLS	Adj. R-squared (uncentered):	1.000
Method:	Least Squares	F-statistic:	1.431e+05
Date:	Sat, 16 Sep 2023	Prob (F-statistic):	0.00
Time:	00:08:52	Log-Likelihood:	1523.1
No. Observations:	569	AIC:	-3018.
Df Residuals:	555	BIC:	-2957.
Df Model:	14		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
perimeter_mean	0.9342	0.013	69.796	0.000	0.908	0.960
area_mean	0.0813	0.011	7.290	0.000	0.059	0.103
smoothness_mean	0.0106	0.002	6.144	0.000	0.007	0.014
compactness_mean	-0.0536	0.003	-18.211	0.000	-0.059	-0.048
concavity_mean	-0.0332	0.003	-11.809	0.000	-0.039	-0.028
radius_se	0.0085	0.005	1.676	0.094	-0.001	0.018
perimeter_se	-0.0192	0.005	-3.800	0.000	-0.029	-0.009
smoothness_se	0.0022	0.001	2.027	0.043	6.73e-05	0.004
concavity_se	0.0124	0.001	9.009	0.000	0.010	0.015
radius_worst	0.2157	0.015	14.494	0.000	0.186	0.245
perimeter_worst	-0.0959	0.013	-7.560	0.000	-0.121	-0.071
area_worst	-0.0845	0.009	-9.133	0.000	-0.103	-0.066
smoothness_worst	-0.0081	0.002	-4.435	0.000	-0.012	-0.005
compactness_worst	0.0106	0.002	5.075	0.000	0.006	0.015

Omnibus:	72.226	Durbin-Watson:	1.916
Prob(Omnibus):	0.000	Jarque-Bera (JB):	622.372
Skew:	-0.082	Prob(JB):	7.14e-136
Kurtosis:	8.121	Cond. No.	86.0

Notes:
[1] R² is computed without centering (uncentered) since the model does not contain a constan
t.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.

--- La variable eliminada fue: radius_se

OLS Regression Results

Dep. Variable:	radius_mean	R-squared (uncentered):	1.000
Model:	OLS	Adj. R-squared (uncentered):	1.000
Method:	Least Squares	F-statistic:	1.536e+05
Date:	Sat, 16 Sep 2023	Prob (F-statistic):	0.00
Time:	00:08:52	Log-Likelihood:	1521.6
No. Observations:	569	AIC:	-3017.
Df Residuals:	556	BIC:	-2961.
Df Model:	13		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
perimeter_mean	0.9314	0.013	70.006	0.000	0.905	0.958
area_mean	0.0821	0.011	7.347	0.000	0.060	0.104
smoothness_mean	0.0110	0.002	6.459	0.000	0.008	0.014
compactness_mean	-0.0538	0.003	-18.234	0.000	-0.060	-0.048
concavity_mean	-0.0324	0.003	-11.681	0.000	-0.038	-0.027
perimeter_se	-0.0112	0.002	-7.032	0.000	-0.014	-0.008

```
smoothness_se          0.0025      0.001      2.362      0.019       0.000       0.005
concavity_se           0.0123      0.001      8.932      0.000       0.010       0.015
radius_worst           0.2313      0.012     19.972      0.000       0.209       0.254
perimeter_worst       -0.1092      0.010    -11.032      0.000      -0.129      -0.090
area_worst            -0.0850      0.009     -9.174      0.000      -0.103      -0.067
smoothness_worst      -0.0085      0.002     -4.698      0.000      -0.012      -0.005
compactness_worst      0.0106      0.002      5.072      0.000       0.006       0.015
==============================================================================
Omnibus:                       75.002   Durbin-Watson:                   1.910
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              675.381
Skew:                          -0.111   Prob(JB):                    2.20e-147
Kurtosis:                       8.333   Cond. No.                         79.8
==============================================================================

Notes:
[1] R² is computed without centering (uncentered) since the model does not contain a constan
t.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

6.-Comparación entre datos reales y predicción. Análisis de los resultados.

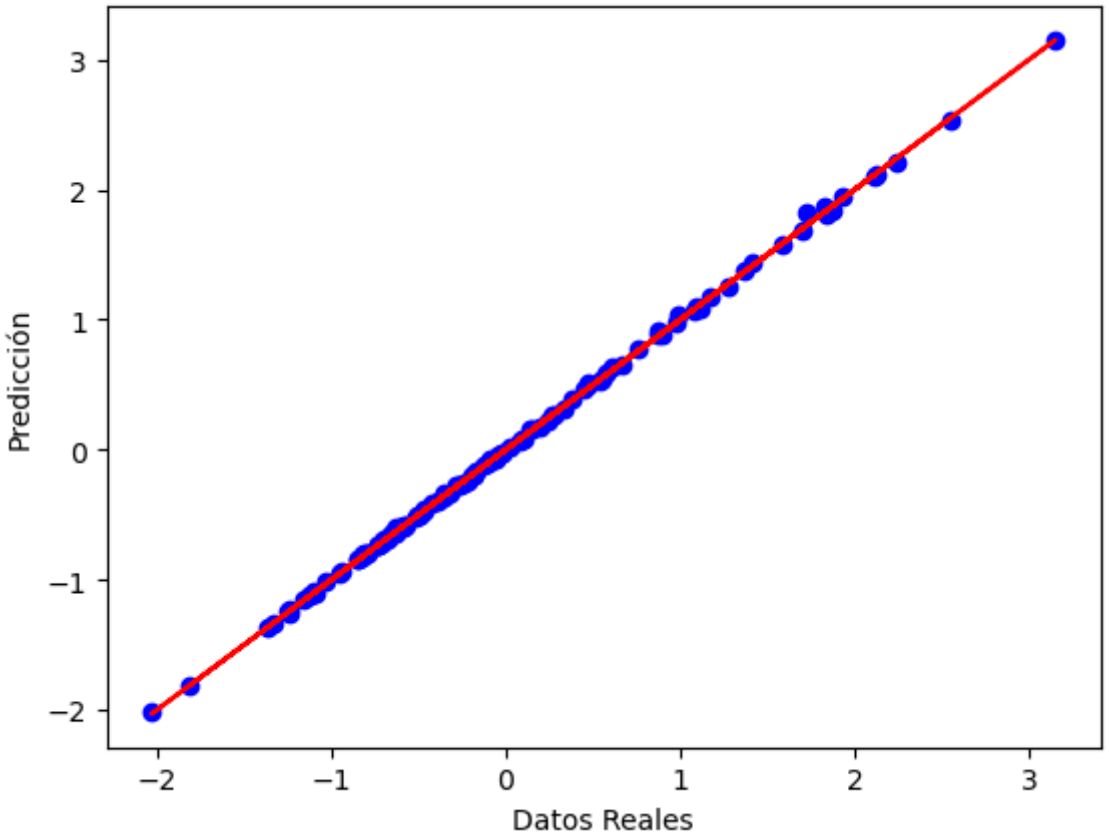
El modelo de regresión múltiple muestra un R-squared perfecto, lo que indico que el modelo se ajusta perfectamente a los datos, asi mismo se puede determinar que los coeficientes de las variables independientes son todos significativos.

A continuación este análisis encaja en veracidad, al ver como la gráfica de valores de predicción y valores reales se alinea con la línea roja porque esto significa que el modelo de regresión está haciendo predicciones muy precisas y que los valores predichos son prácticamente idénticos a los valores reales. También se realiza el grafico e histograma de los residuos.

In [12]:
```
y_aprox=modelo2.params[0]*prueba['perimeter_mean']+modelo2.params[1]*prueba['area_mean']+model
tabla=pd.DataFrame({'Real': prueba['radius_mean'], 'Prediccion': y_aprox, 'Errores': prueba['
```

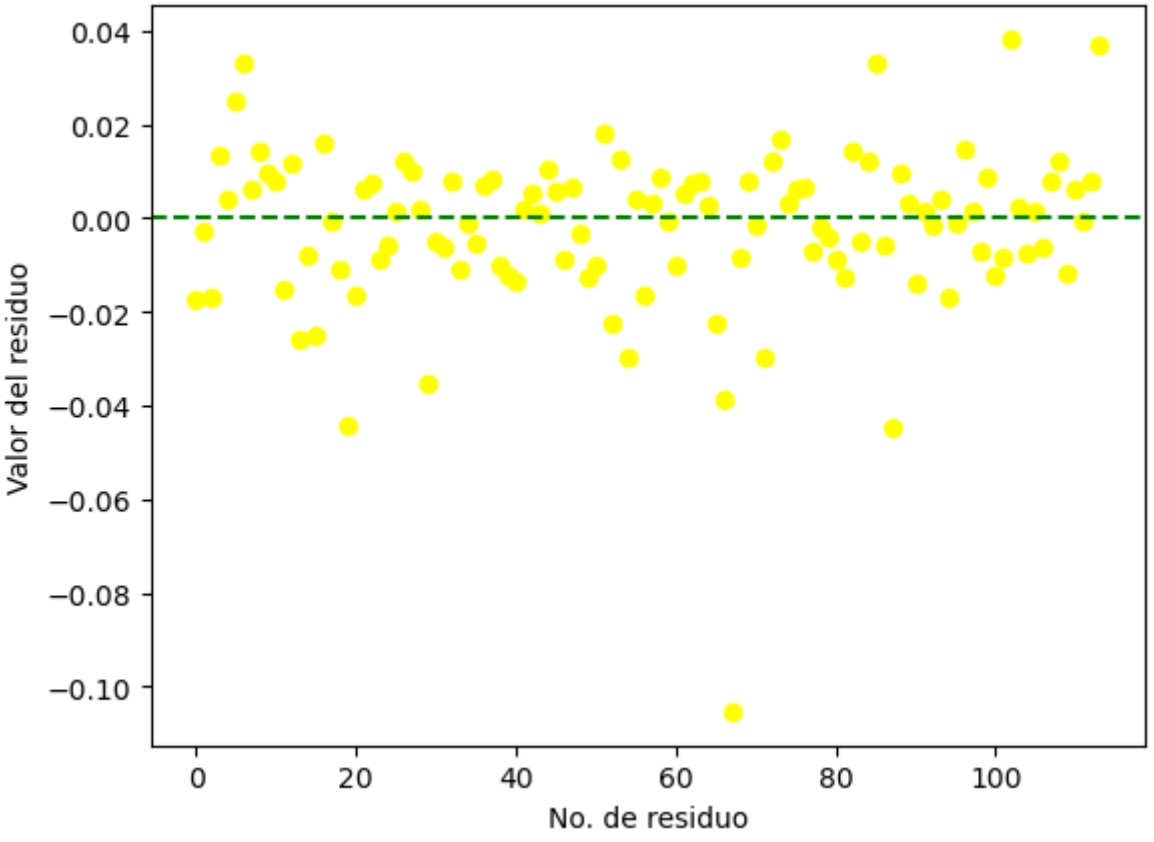
In [13]:
```
plt.scatter(prueba['radius_mean'], y_aprox, color='blue')
plt.plot(prueba['radius_mean'],prueba['radius_mean'], color='red')
plt.xlabel("Datos Reales")
plt.ylabel("Predicción")
```

Out[13]:
```
Text(0, 0.5, 'Predicción')
```



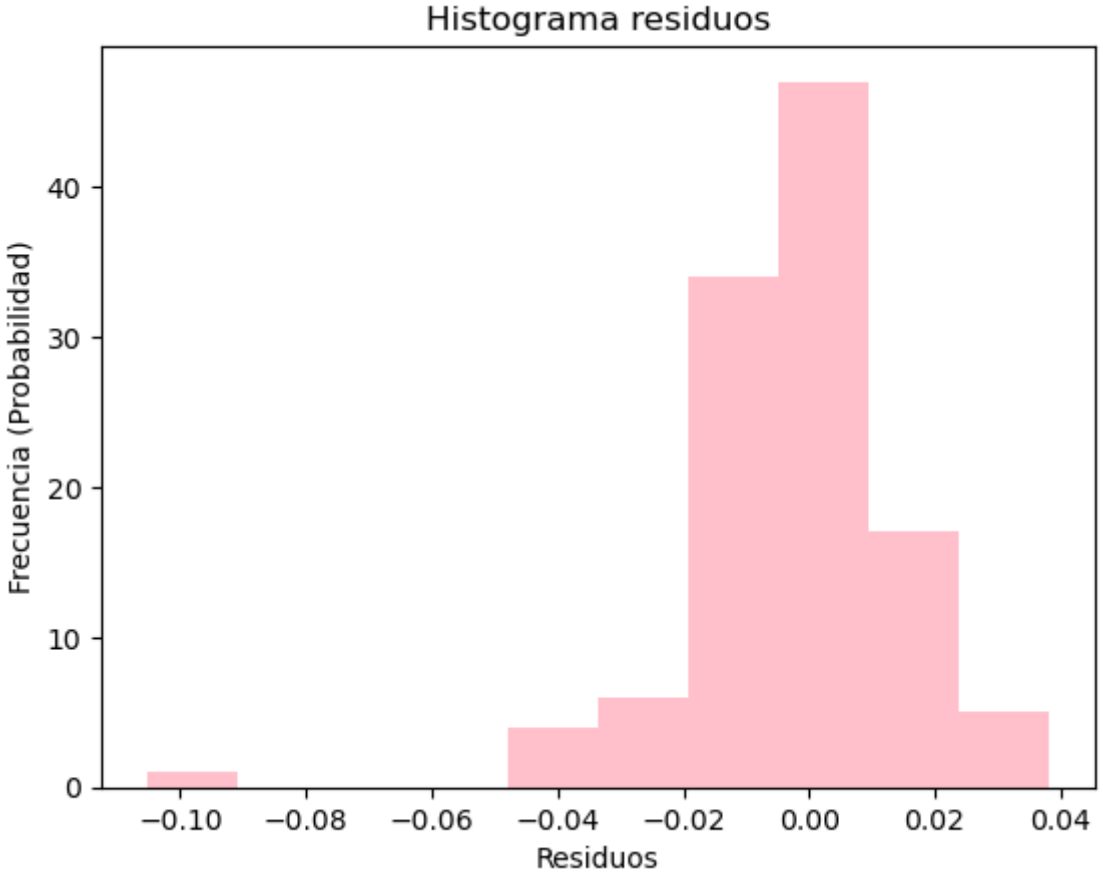
In [14]:
```
plt.scatter(range(tabla.shape[0]),tabla['Errores'], color='yellow')
plt.axhline(y=0, linestyle='--', color='green')
plt.xlabel("No. de residuo")
plt.ylabel("Valor del residuo")
```

Out[14]:
```
Text(0, 0.5, 'Valor del residuo')
```



In [15]:
```
plt.hist(x=tabla['Errores'], color='pink')
plt.title('Histograma residuos')
plt.xlabel("Residuos")
plt.ylabel("Frecuencia (Probabilidad)")
```

Out[15]:
```
Text(0, 0.5, 'Frecuencia (Probabilidad)')
```


Histograma residuos
Frecuencia (Probabilidad)
0
10
20
30
40
−0.10
−0.08
−0.06
−0.04
−0.02
0.00
0.02
0.04
Residuos