Birth Weight Analysis

1. Data Preparation and Initial Model

Column	Data Type	Description
id	int64	Identification code
low	int64	Birth weight < 2500g (binary: 0=no, 1=yes)
age	int64	Age of mother
lwt	int64	Weight at last menstrual period
race	object	Race
smoke	int64	Smoked during pregnancy (binary: 0=no, 1=yes)
ptl	int64	Premature labor history (count)
ht	int64	Has history of hypertension (binary: 0=no, 1=yes)
ui	int64	Presence of uterine irritability (binary: 0=no, 1=yes)
ftv	int64	Number of visits to physician during 1st trimester
bwt	int64	Birth weight (grams)
race2	int64	Second race variable
race3	int64	Third race variable

The analysis begins with data preparation, including the creation of race segmentation variables (race_white, race_black, race_other) and the dropping of original race-related columns. The initial full model is then fitted using all available predictors.

1.1 Initial Model Summary

Dep. Variable:	bwt	**R-squared:**	0.668
Model:	OLS	**Adj. R-squared:**	0.649
Method:	Least Squares	**F-statistic:**	35.74
Date:	Sat, 28 Sep 2024	**Prob (F-statistic):**	1.60e-37
Time:	22:23:28	**Log-Likelihood:**	-1409.4
No. Observations:	189	**AIC:**	2841.
Df Residuals:	178	**BIC:**	2877.
Df Model:	10		

Covariance Type:	nonrobust					
	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	2517.0669	152.368	16.520	0.000	2216.386	2817.748
low	-1118.6446	74.150	-15.086	0.000	-1264.972	-972.318
age	-7.8539	6.396	-1.228	0.221	-20.475	4.768
lwt	1.5075	1.168	1.290	0.199	-0.798	3.813
smoke	-172.7168	71.714	-2.408	0.017	-314.235	-31.199
ptl	81.2508	68.277	1.190	0.236	-53.485	215.987
ht	-182.5864	137.108	-1.332	0.185	-453.152	87.979
ui	-339.3759	92.982	-3.650	0.000	-522.865	-155.887
ftv	-6.6709	30.871	-0.216	0.829	-67.592	54.250
race_white	984.5147	71.912	13.691	0.000	842.605	1126.424
race_black	743.2896	88.831	8.367	0.000	567.993	918.586
race_other	789.2626	61.619	12.809	0.000	667.665	910.860

Omnibus:	3.672	**Durbin-Watson:**	0.499
Prob(Omnibus):	0.159	**Jarque-Bera (JB):**	3.316
Skew:	0.245	**Prob(JB):**	0.190
Kurtosis:	3.426	**Cond. No.**	6.94e+17

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 7.19e-30. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.

- R-squared: 0.668 (66.8% of variance in birth weight explained)
- Adjusted R-squared: 0.649
- F-statistic: 35.74 (p-value: 1.60e-37)

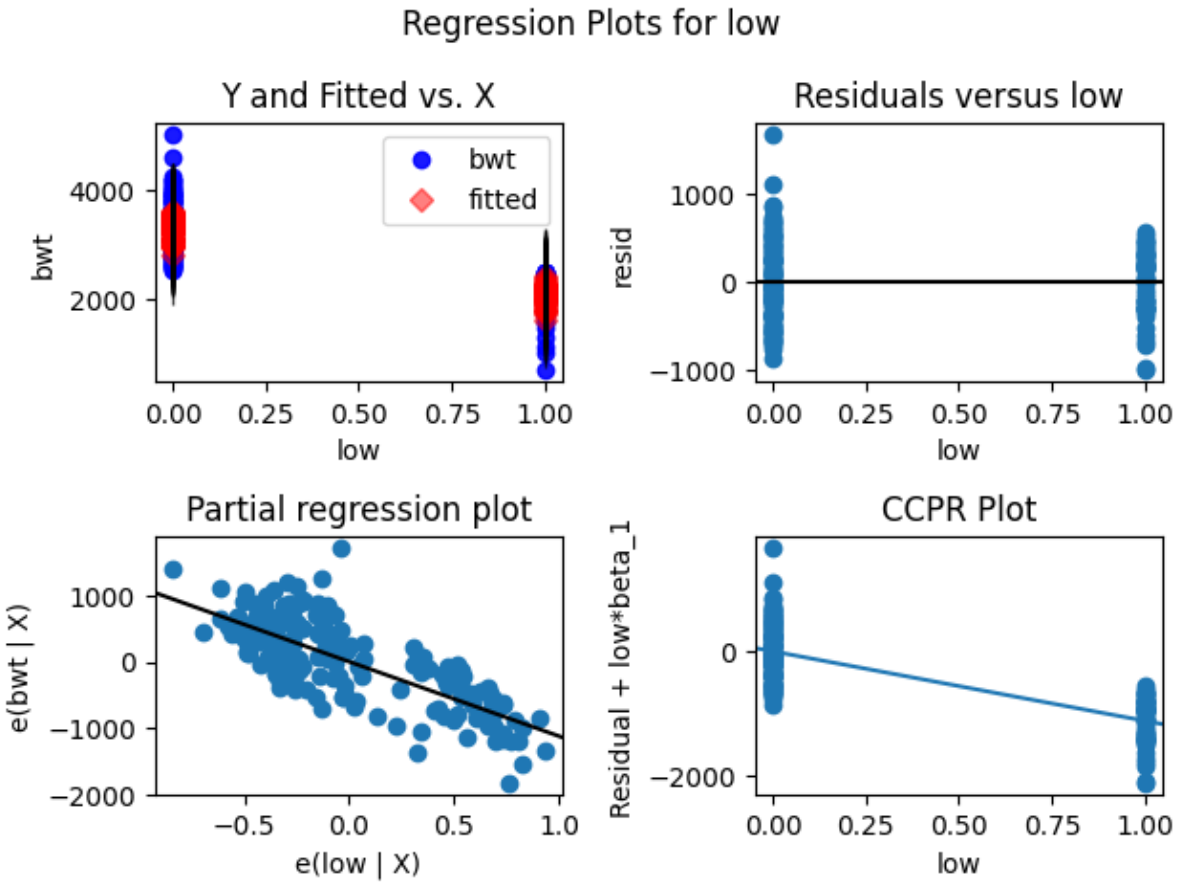
Significant predictors ($p < 0.05$):

- Low birth weight indicator (strong negative effect: -1118.64 grams)
- Smoking status (negative effect: -172.72 grams)
- Uterine irritability (negative effect: -339.38 grams)
- Race categories (all significant, with varying positive effects)

The model shows that the low birth weight indicator has the strongest effect, followed by race categories. Smoking and uterine irritability also have substantial negative impacts on birth weight.

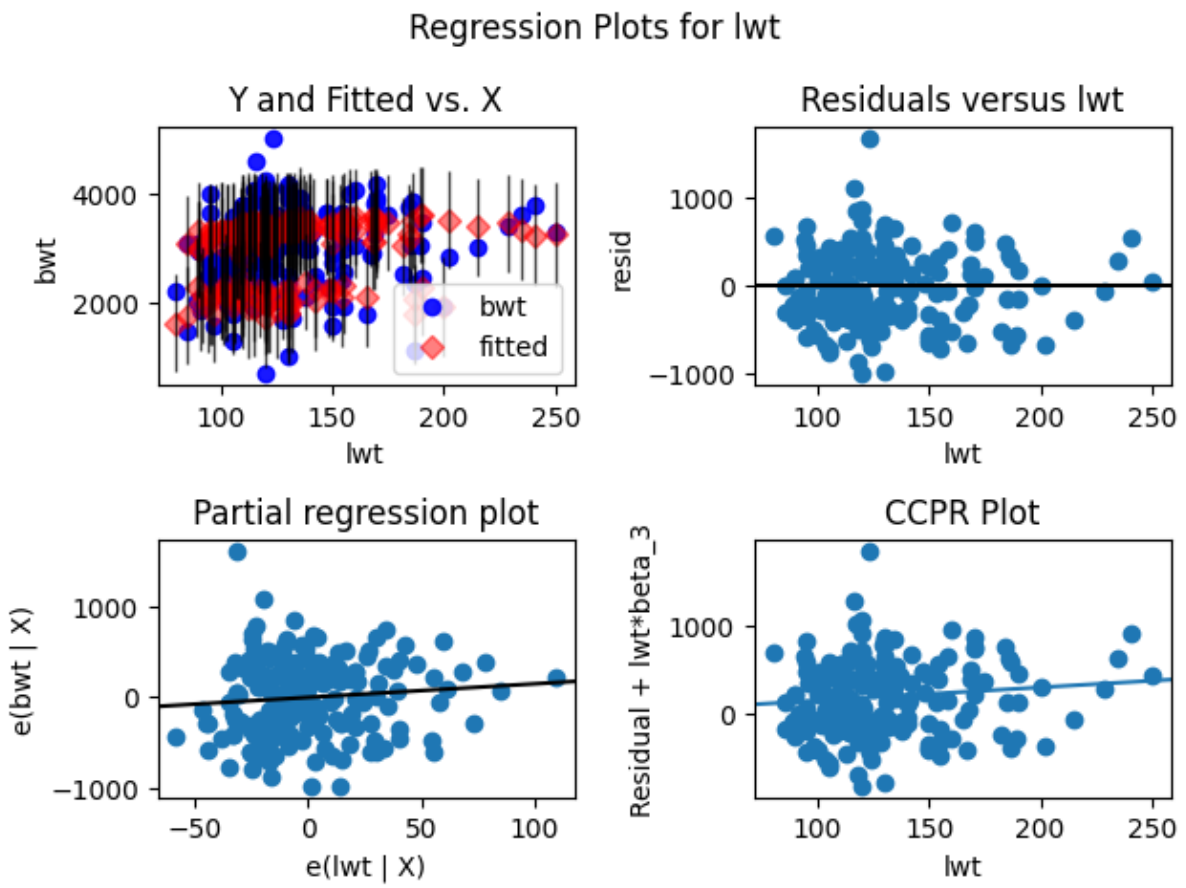
2. Classical Assumptions Tests

2.1 Linearity

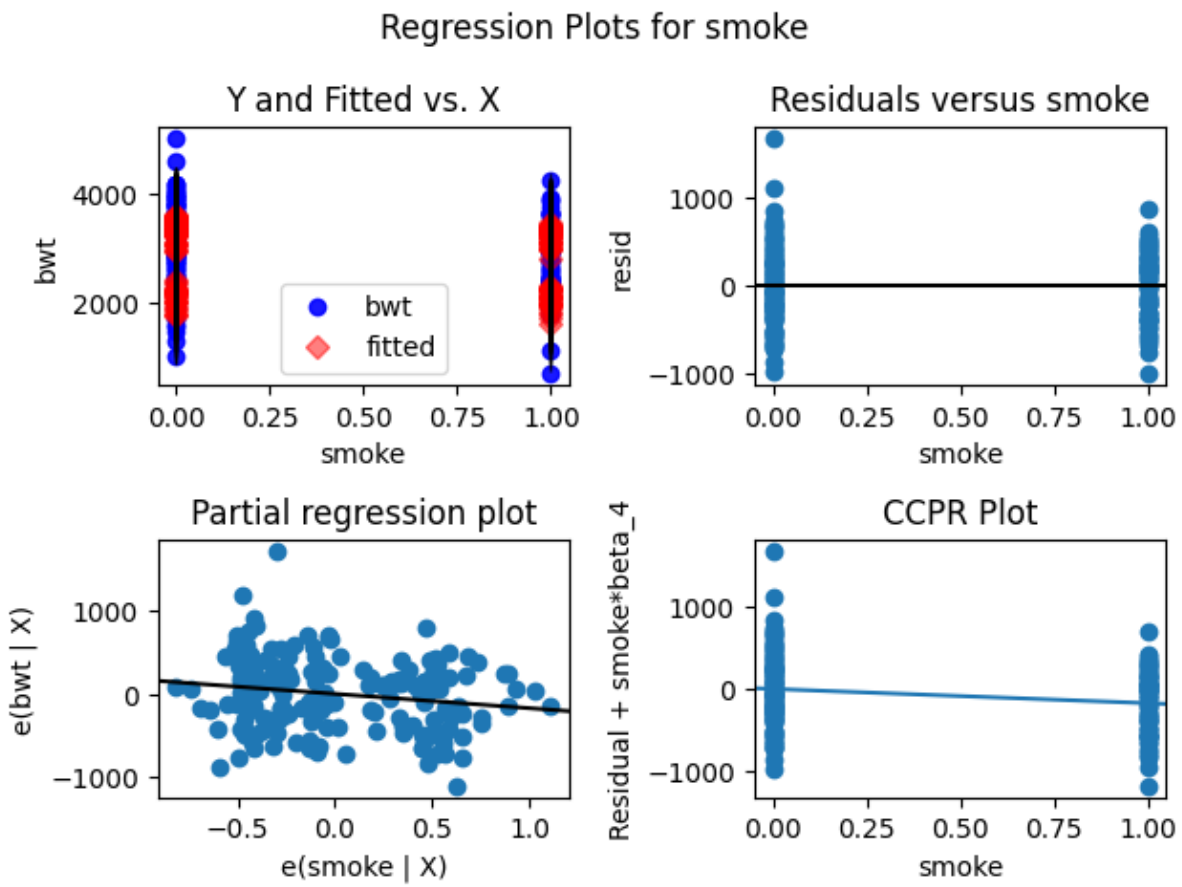


- The plot shows two distinct clusters, with the low birth weight group (1) significantly lower than the normal group (0).
- There's a clear separation between the groups, confirming the strong negative effect of the low birth weight indicator.
- The relationship is clearly non-linear, justifying its importance in the model.

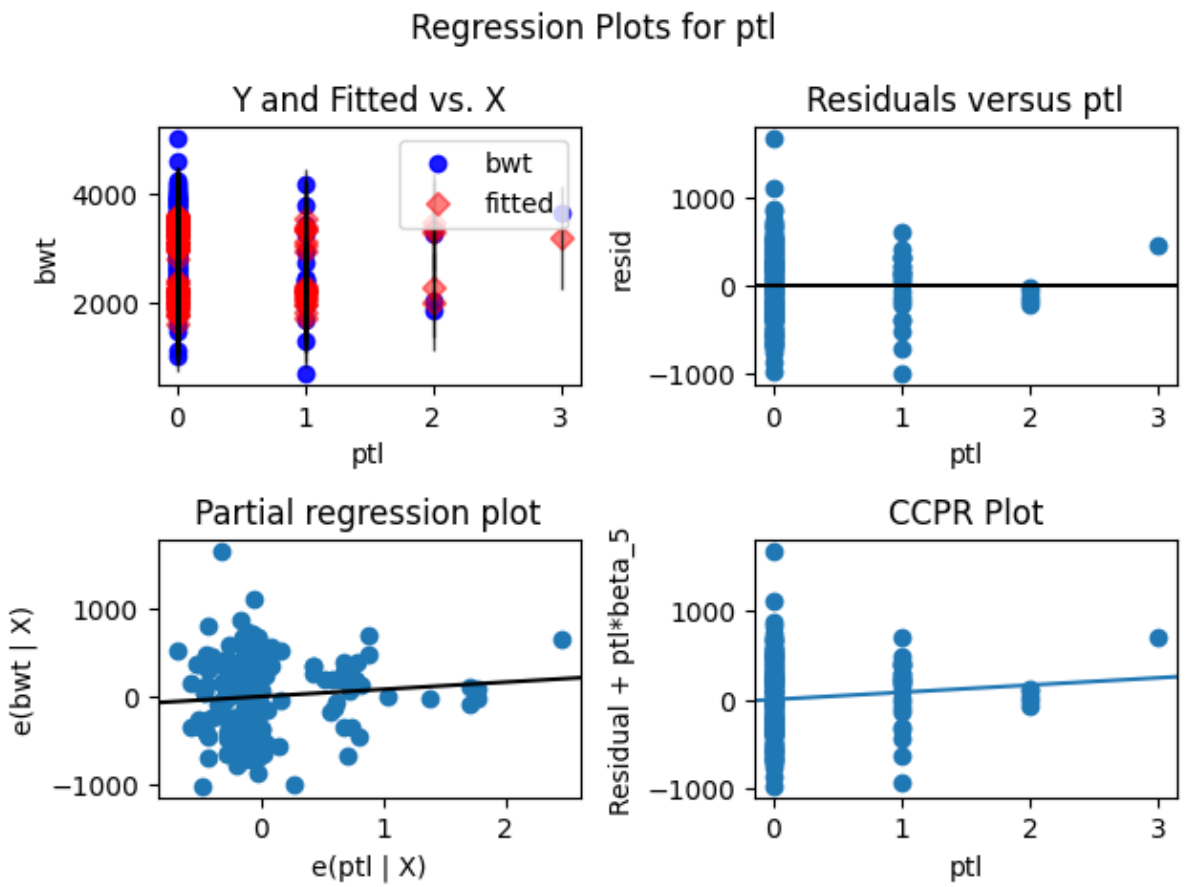
- There's a slight positive trend, but with considerable scatter.
- The relationship appears somewhat non-linear, with birth weight increasing more rapidly in early adulthood and then leveling off.
- This supports the finding from the polynomial model that age has a complex relationship with birth weight.



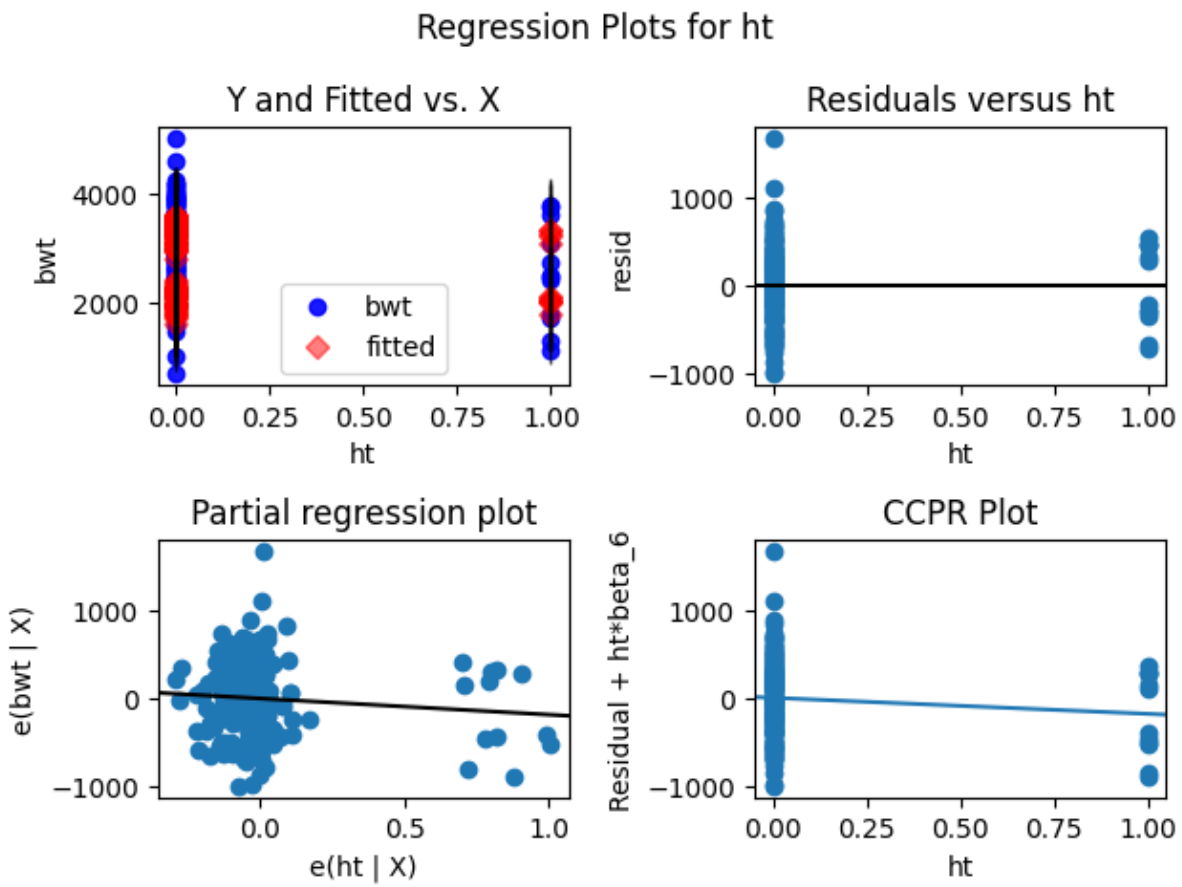
- There's a positive trend, indicating that higher maternal weight is associated with higher birth weight.
- The relationship appears roughly linear, but with substantial variability.
- Some outliers are visible, particularly for very high maternal weights.



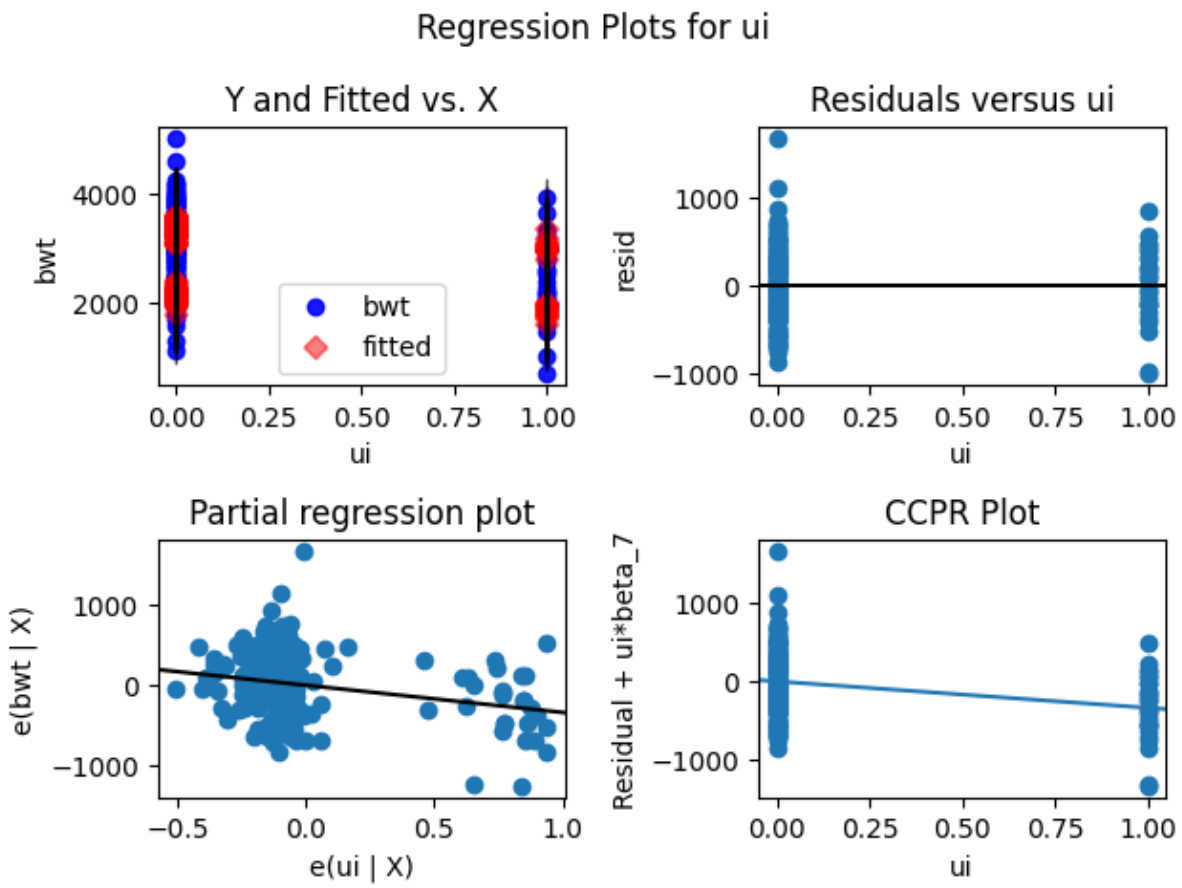
- Two distinct groups are visible (0 for non-smokers, 1 for smokers).
- The smoking group shows generally lower birth weights, confirming the negative effect of smoking.
- There's considerable overlap between the groups, suggesting smoking isn't the only important factor.



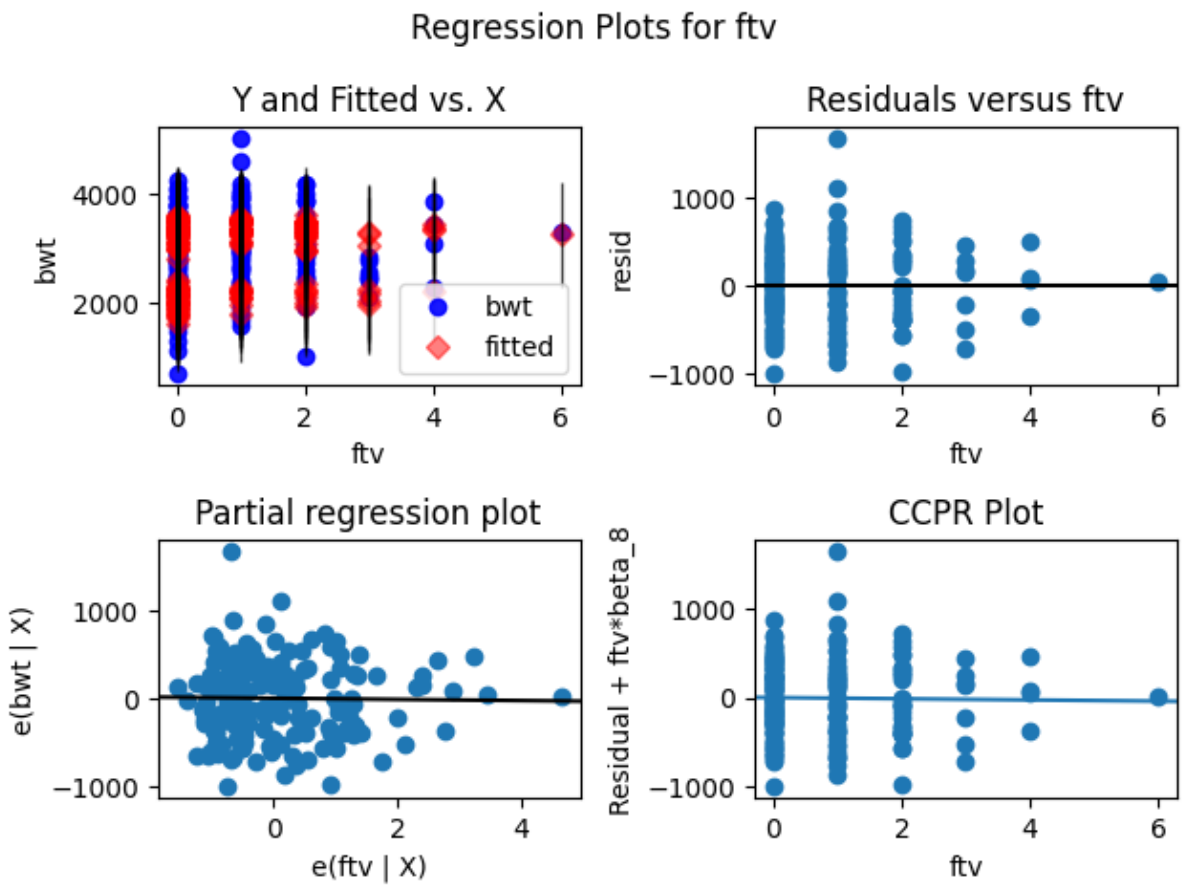
- There's a slight negative trend, with higher values of ptl associated with lower birth weights.
- The relationship is weak and there's significant overlap between groups.
- This explains why ptl wasn't a strong predictor in the stepwise model.



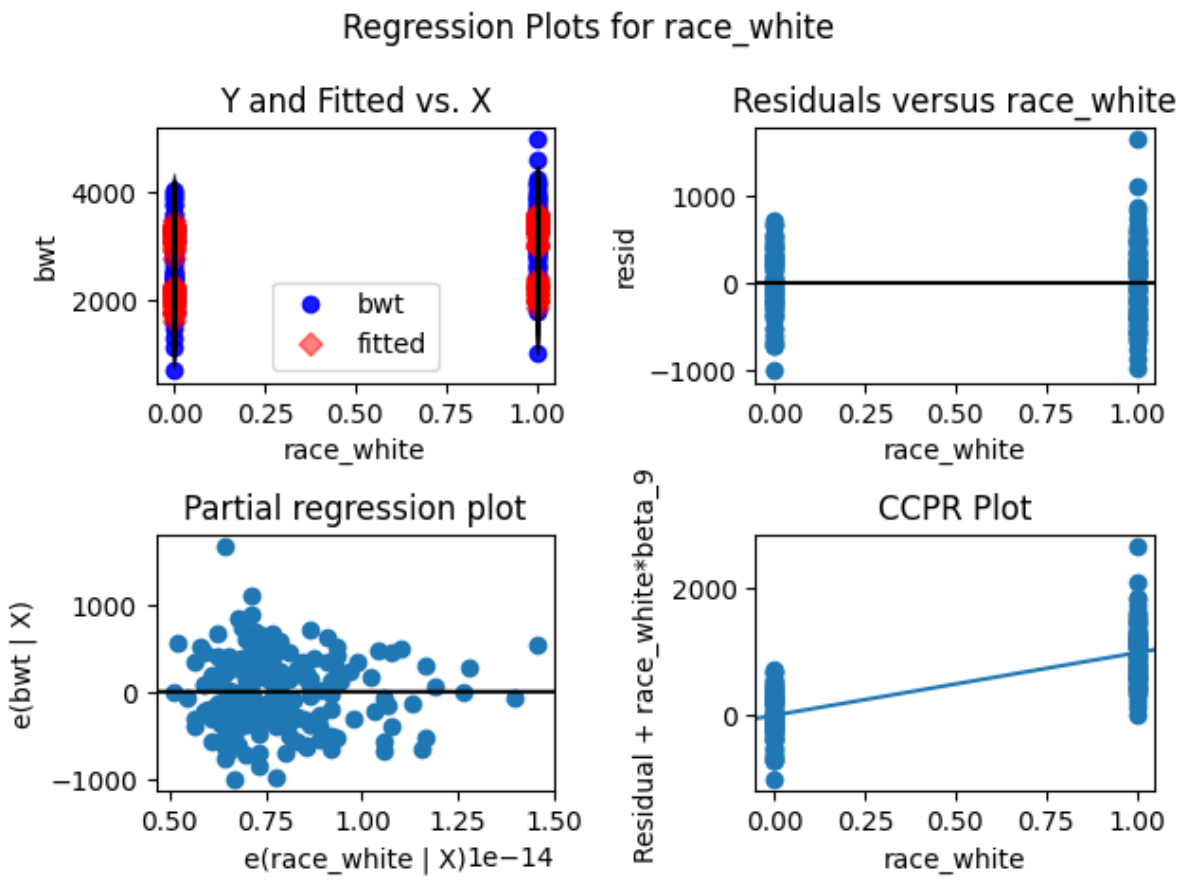
- Two groups are visible (0 for no hypertension, 1 for hypertension).
- The hypertension group shows slightly lower birth weights on average.
- There's substantial overlap between the groups, indicating hypertension alone isn't a definitive predictor.



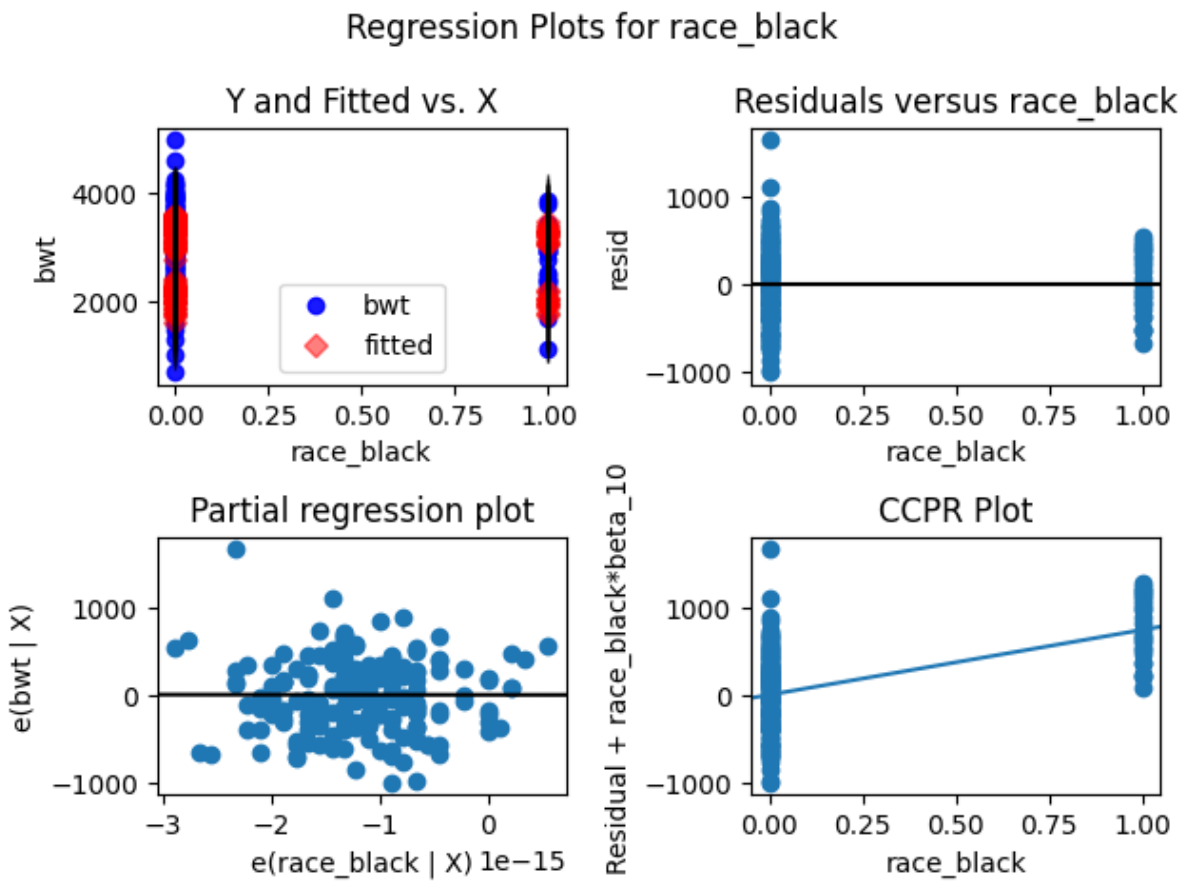
- Two distinct groups are visible (0 for no UI, 1 for UI).
- The UI group shows noticeably lower birth weights.
- This clear separation supports this variable's importance in the final model.



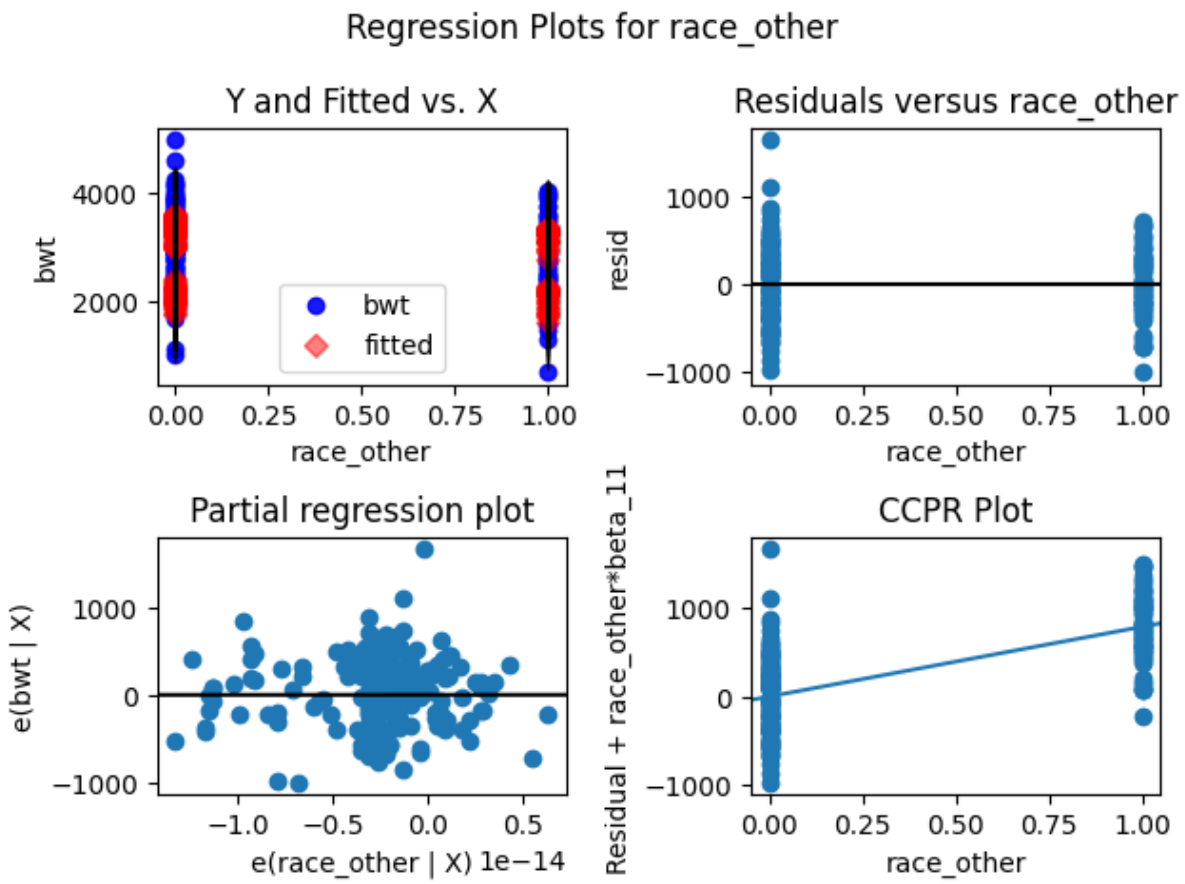
- There's no clear trend visible.
- Birth weights are widely scattered across all values of ftv.
- This lack of clear relationship explains why ftv wasn't a strong predictor.



- Two distinct groups (0 for non-white, 1 for white).
- The white group shows higher average birth weights.
- There's some overlap between the groups, indicating race isn't the only important factor.



- Two groups (0 for non-black, 1 for black).
- The black group shows slightly lower average birth weights.
- Considerable overlap exists between the groups.



- Two groups (0 for not other race, 1 for other race).
- The 'other' race group shows average birth weights between white and black groups.
- Significant overlap exists between the groups.

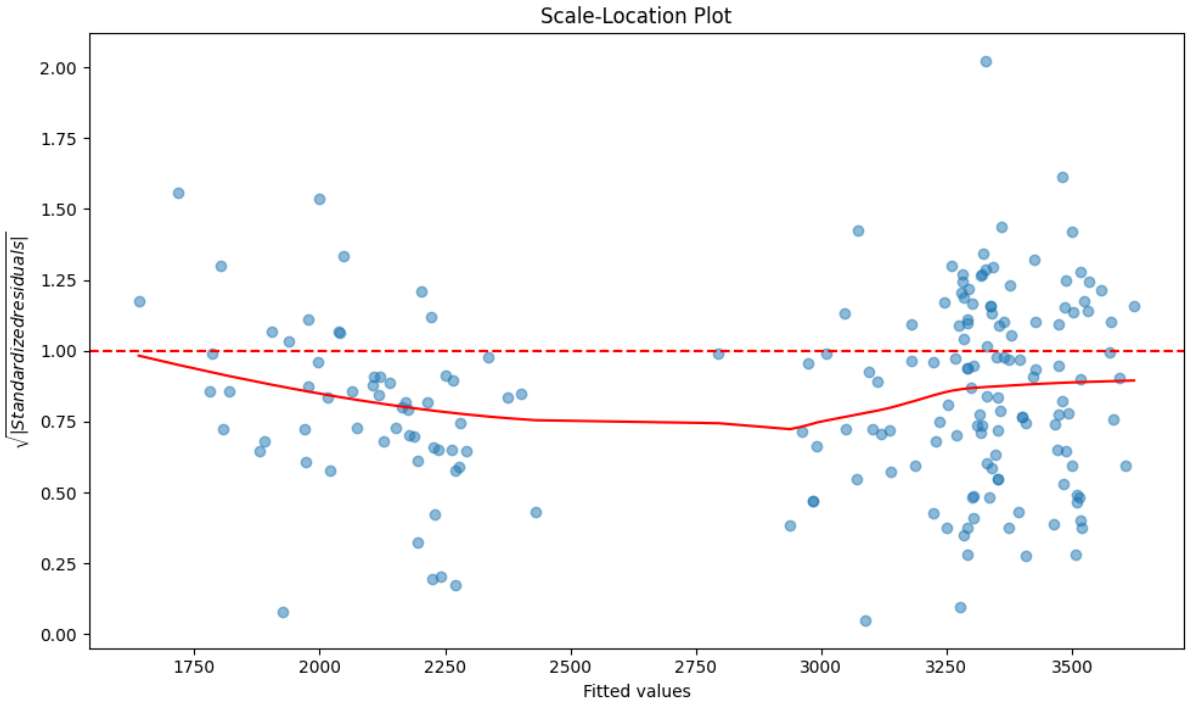
2.2 Multicollinearity

Variance Inflation Factors (VIF) are calculated:

- Most variables show low VIF values (< 5), indicating no severe multicollinearity.
- Race variables show high VIF values:
    - race_white: 22.97
    - race_black: 6.99
    - race_other: 13.68

The high VIF for race variables suggests strong multicollinearity, which is expected due to their mutually exclusive nature. This doesn't invalidate the model but may affect the interpretation of individual race effects.

2.3 Homoscedasticity



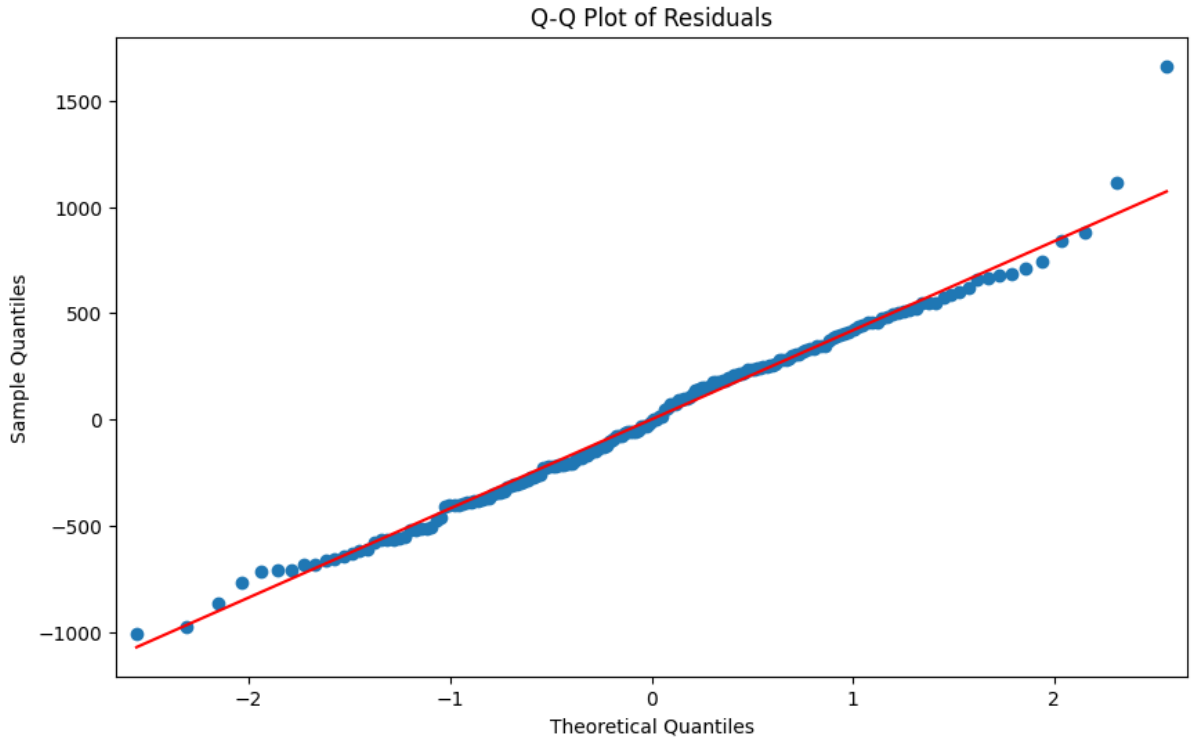
- The trnd line is mostly flat, but it slightly dips below 1 for fitted values between 1750 and 2500 and rises again toward the higher fitted values (beyond 3250).
- This slight trend might indicate a small deviation from homoscedasticity, but it's not extreme.
- There seems to be more spread in the residuals at the lower and higher ends of the fitted values, especially at around 1750 and beyond 3250. The middle range (around 2250 to 3000) shows less spread.

The Breusch-Pagan test results:

- LM Statistic : 24.65
- LM-Test p-value : 0.0103
- F-Statistic : 2.67
- F-Test p-value : 0.0046

The low p-values (< 0.05) suggest the presence of heteroscedasticity, indicating that the variance of residuals is not constant across all levels of the predictors.

2.4Normality of Residuals



- The majority of the residuals, especially in the middle range of quantiles (around 0), lie close to the red line, indicating that for most of the observations, the residuals are reasonably normally distributed.
- At both extremes (the lower quantiles on the left and the higher quantiles on the right), there is some deviation from the red line.
- On the right (positive residuals), the residuals show a strong upward deviation, suggesting the presence of outliers or that the residuals are skewed.
- On the left (negative residuals), there is a smaller but noticeable deviation below the red line, indicating potential outliers or non-normality in the lower tail as well.

Normality test p-value: 0.15944561131318238

The normality test yields a p-value of 0.159, suggesting that we fail to reject the null hypothesis of normality. The residuals appear to be approximately normally distributed.

2.5 Autocorrelation

Durbin-Watson statistic: 0.49875306281225185

The Durbin-Watson statistic is 0.499, suggesting positive autocorrelation in the residuals.

3. Model Selection

3.1 Stepwise Selection

A stepwise selection process is performed, resulting in a simplified model with only two predictors:

1. Low birth weight indicator
2. Uterine irritability

These two variables are the most important predictors of birth weight in this dataset.

3.2 Final Model Summary

Dep. Variable:	bwt	**R-squared:**	0.640
Model:	OLS	**Adj. R-squared:**	0.636
Method:	Least Squares	**F-statistic:**	165.0
Date:	Sat, 28 Sep 2024	**Prob (F-statistic):**	6.16e-42
Time:	22:32:17	**Log-Likelihood:**	-1417.1
No. Observations:	189	**AIC:**	2840.
Df Residuals:	186	**BIC:**	2850.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	3362.9472	39.831	84.431	0.000	3284.369	3441.525
low	-1190.5889	70.085	-16.988	0.000	-1328.854	-1052.324
ui	-317.2242	91.418	-3.470	0.001	-497.574	-136.874

Omnibus:	2.265	**Durbin-Watson:**	0.353
Prob(Omnibus):	0.322	**Jarque-Bera (JB):**	1.861
Skew:	0.185	**Prob(JB):**	0.394
Kurtosis:	3.316	**Cond. No.**	3.12

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

- R-squared: 0.640
- Adjusted R-squared: 0.636
- F-statistic: 165.0 (p-value: 6.16e-42)

Both predictors are highly significant ($p < 0.001$):

- Low birth weight indicator: coefficient -1190.59
- Uterine irritability: coefficient -317.22

The final model explains 64% of the variance in birth weight, which is only slightly lower than the full model (66.8%), despite using only two predictors.

4. Non-linear Relationships

A polynomial model is fitted to explore potential non-linear relationships, particularly for age and mother's weight (lwt)

Dep. Variable:	bwt	**R-squared:**	0.692
Model:	OLS	**Adj. R-squared:**	0.670
Method:	Least Squares	**F-statistic:**	30.31
Date:	Sat, 28 Sep 2024	**Prob (F-statistic):**	4.34e-38
Time:	22:43:31	**Log-Likelihood:**	-1402.1
No. Observations:	189	**AIC:**	2832.
Df Residuals:	175	**BIC:**	2878.
Df Model:	13		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	2843.6698	621.507	4.575	0.000	1617.057	4070.283
age	-89.8146	43.151	-2.081	0.039	-174.978	-4.651
lwt	9.5841	7.867	1.218	0.225	-5.942	25.110
age^2	2.7373	0.787	3.477	0.001	1.184	4.291
age lwt	-0.4176	0.196	-2.127	0.035	-0.805	-0.030
lwt^2	0.0070	0.023	0.308	0.758	-0.038	0.052

low	-1103.1933	72.110	-15.299	0.000	-1245.511	-960.875
smoke	-164.3423	71.053	-2.313	0.022	-304.573	-24.112
ptl	93.1607	66.455	1.402	0.163	-37.996	224.318
ht	-189.4524	135.000	-1.403	0.162	-455.890	76.985
ui	-352.3995	91.194	-3.864	0.000	-532.381	-172.418
ftv	-7.0621	30.137	-0.234	0.815	-66.542	52.418
race_white	1094.6066	217.352	5.036	0.000	665.637	1523.576
race_black	823.8514	219.262	3.757	0.000	391.114	1256.589
race_other	925.2118	205.523	4.502	0.000	519.589	1330.834

Omnibus:	2.012	**Durbin-Watson:**	0.467
Prob(Omnibus):	0.366	**Jarque-Bera (JB):**	1.601
Skew:	0.035	**Prob(JB):**	0.449
Kurtosis:	2.555	**Cond. No.**	1.04e+20

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 7.16e-30. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.

- Age shows a quadratic relationship with birth weight (age^2 term is significant, p = 0.001)
- Interaction between age and mother's weight is significant (p = 0.035)
- The polynomial model slightly improves R-squared to 0.692

The relationship between mother's age and birth weight is more complex than a simple linear association, and that the effect of mother's weight may depend on her age.

5. Model Comparison

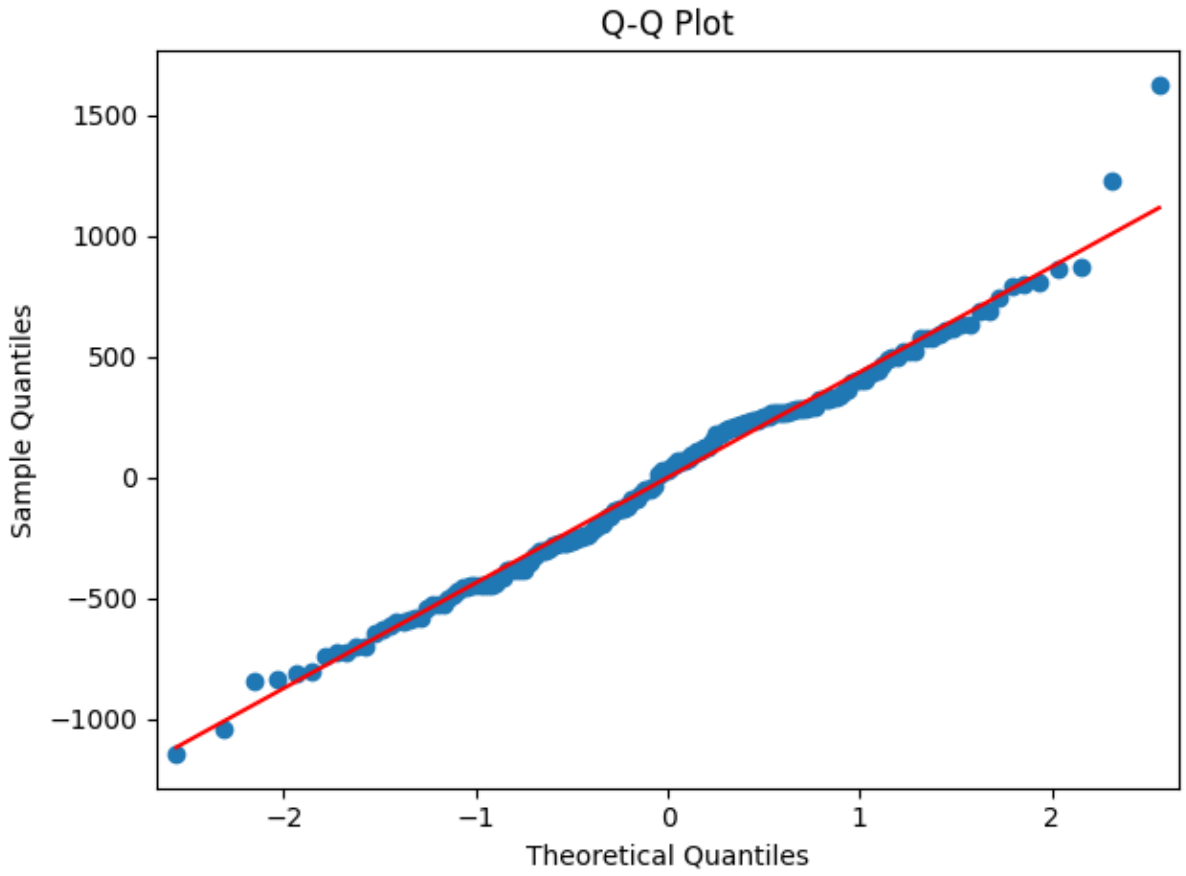
An ANOVA comparison of the initial full model, stepwise-selected model, and polynomial model is performed.

df_resid	ssr	df_diff	ss_diff	F	Pr(>F)
178	3.32E+07	0	NaN	NaN	NaN
186	3.60E+07	-8	-2.80E+06	1.994833	NaN
175	3.07E+07	11	5.29E+06	2.737041	0.00273

1. The stepwise-selected model, despite its simplicity, is not significantly worse than the full model.
2. The polynomial model shows a significant improvement over the other models ($F = 2.737$, $p = 0.00273$), indicating that including non-linear terms provides a better fit to the data.

6. Final Model Diagnostics

Q-Q Plot



- The majority of the residuals are normally distributed.
- The deviations at the ends (especially the upper tail) suggest that there may be some outliers.

7. Conclusion

The low birth weight indicator and uterine irritability are the most significant predictors of birth weight in this dataset. The stepwise-selected model, using only these two predictors, explains 64% of the variance in birth weight, nearly as much as the full model with all variables.

However, the significance of polynomial terms, particularly for mother's age, suggests that the relationships between predictors and birth weight are more complex than simple linear associations.