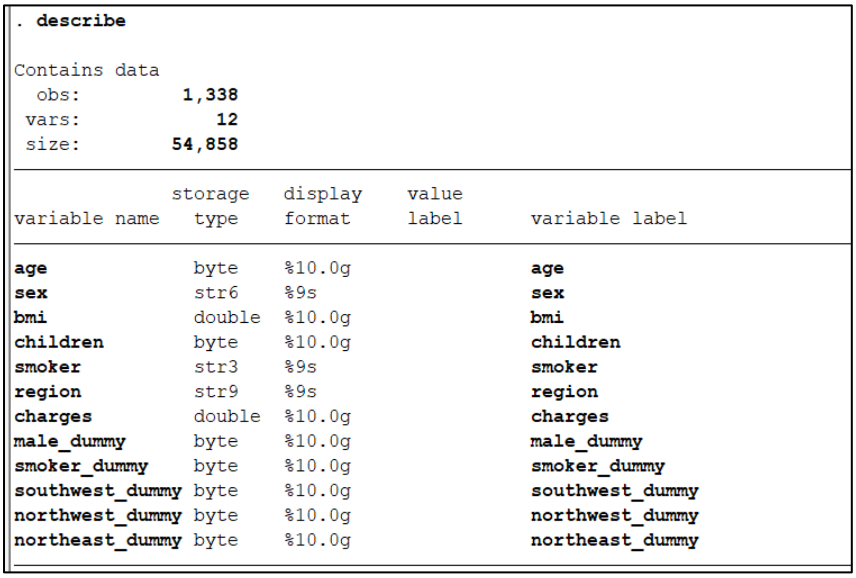
1. Data Exploration and Cleaning:

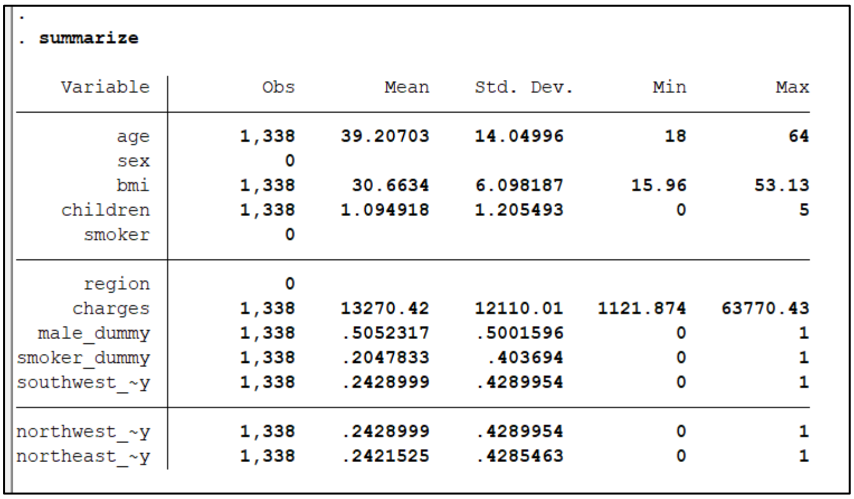
We first conducted a data exploration to identify the data types of each variable, after which we generated dummy variables for the categorical data as needed.



```
. describe

Contains data
  obs:         1,338
 vars:            12
 size:        54,858

              storage   display    value
variable name   type    format     label      variable label

age             byte    %10.0g                age
sex             str6    %9s                   sex
bmi             double  %10.0g                bmi
children        byte    %10.0g                children
smoker          str3    %9s                   smoker
region          str9    %9s                   region
charges         double  %10.0g                charges
male_dummy      byte    %10.0g                male_dummy
smoker_dummy    byte    %10.0g                smoker_dummy
southwest_dummy byte    %10.0g                southwest_dummy
northwest_dummy byte    %10.0g                northwest_dummy
northeast_dummy byte    %10.0g                northeast_dummy
```

2. Descriptive Statistics:

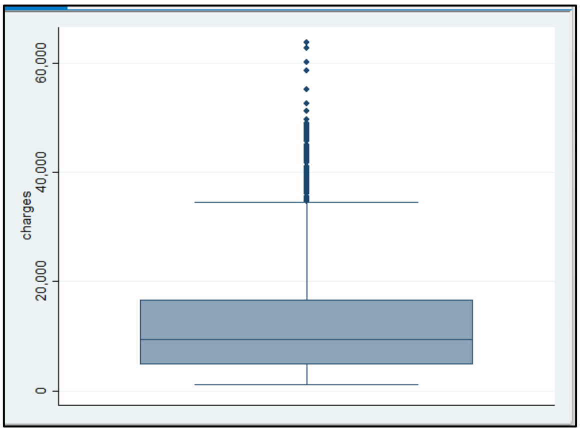
Using the summarize command, we determined the central tendencies of each variable.

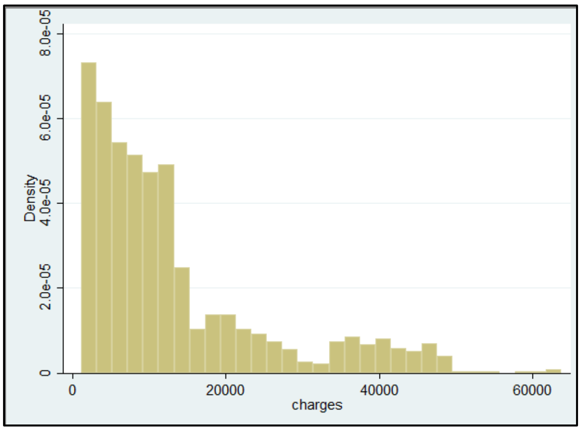


```
.
. summarize

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
         age |      1,338    39.20703    14.04996         18         64
         sex |          0
         bmi |      1,338     30.6634    6.098187      15.96      53.13
    children |      1,338    1.094918    1.205493          0          5
      smoker |          0
-------------+---------------------------------------------------------
      region |          0
     charges |      1,338    13270.42    12110.01   1121.874   63770.43
  male_dummy |      1,338    .5052317    .5001596          0          1
smoker_dummy |      1,338    .2047833     .403694          0          1
southwest_~y |      1,338    .2428999    .4289954          0          1
-------------+---------------------------------------------------------
northwest_~y |      1,338    .2428999    .4289954          0          1
northeast_~y |      1,338    .2421525    .4285463          0          1
```

Box plots and histograms were utilized to identify outliers and assess the distribution of the data.

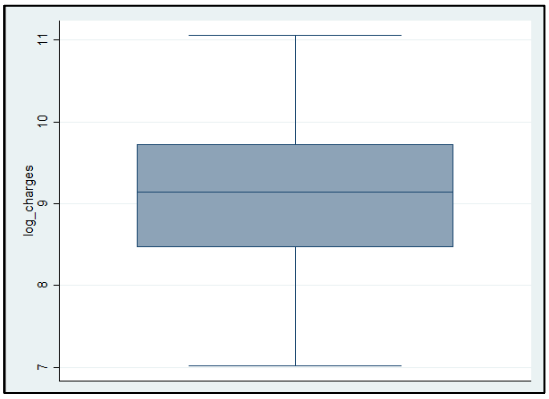
Graph box charges
hist charges

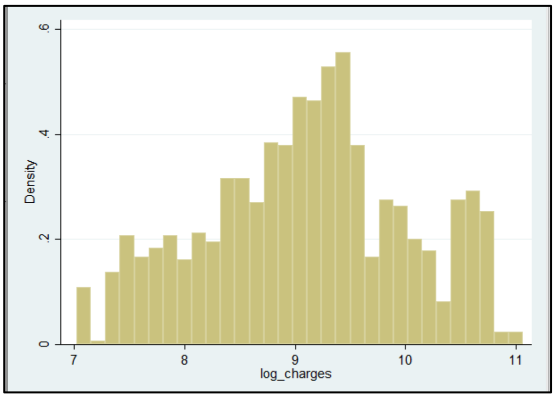




Given that the data is highly skewed with numerous outliers and is significantly larger in scale compared to the independent variable, it is advisable to apply a log transformation to normalize the distribution.

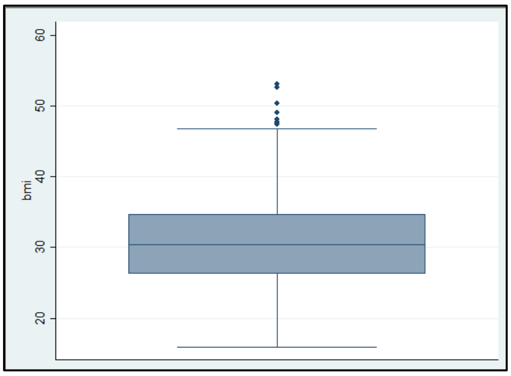
gen log_charges = log(charges)

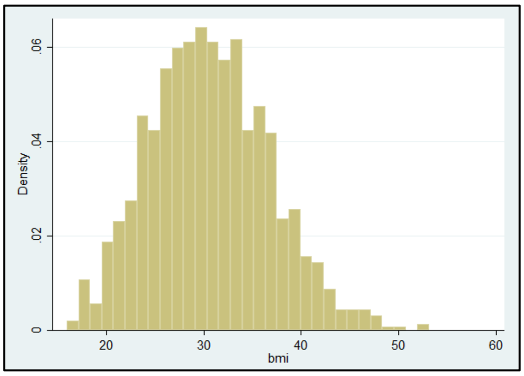




Outliers and the distribution of BMI have now been examined.

Hist bmi
graph box bmi



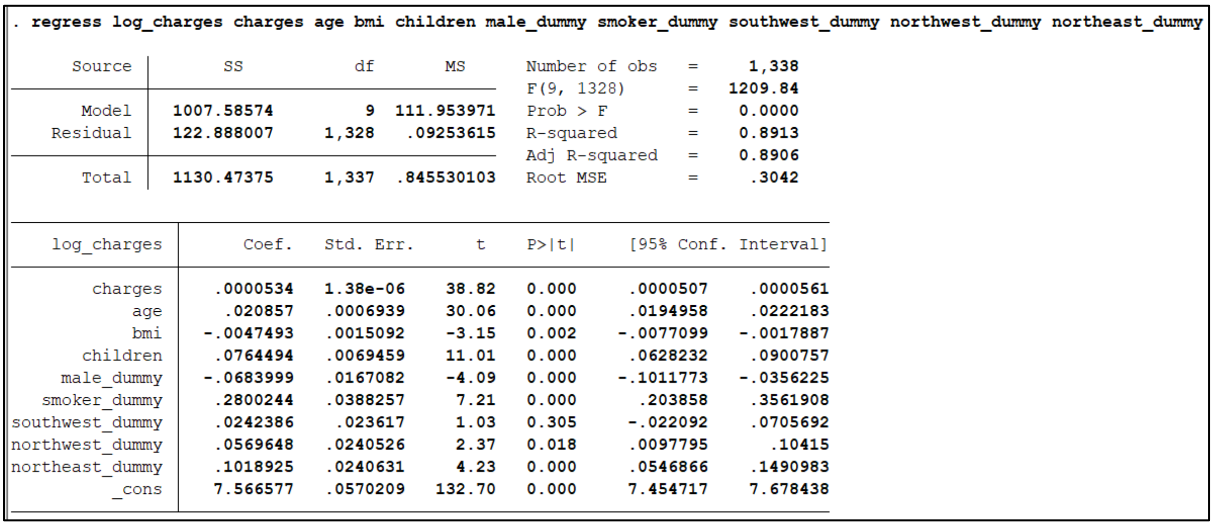


Outliers were detected in BMI; however, since they represent real-world data, they will not be removed from the dataset. Additionally, as the BMI data follows a normal distribution, a log transformation cannot be applied.

3. Initial Regression Model:

Run a linear regression

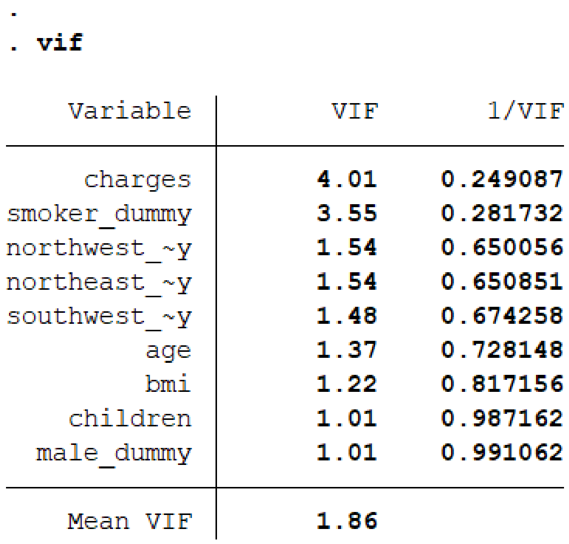
regress log_charges charges age bmi children male_dummy smoker_dummy southwest_dummy northwest_dummy northeast_dummy



```
. regress log_charges charges age bmi children male_dummy smoker_dummy southwest_dummy northwest_dummy northeast_dummy

      Source |       SS           df       MS      Number of obs   =     1,338
-------------+----------------------------------   F(9, 1328)      =   1209.84
       Model |  1007.58574         9  111.953971   Prob > F        =    0.0000
    Residual |  122.888007     1,328  .09253615    R-squared       =    0.8913
-------------+----------------------------------   Adj R-squared   =    0.8906
       Total |  1130.47375     1,337  .845530103   Root MSE        =     .3042

------------------------------------------------------------------------------
  log_charges |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
--------------+----------------------------------------------------------------
      charges |   .0000534   1.38e-06    38.82   0.000     .0000507    .0000561
          age |    .020857   .0006939    30.06   0.000     .0194958    .0222183
          bmi |  -.0047493   .0015092    -3.15   0.002    -.0077099   -.0017887
     children |   .0764494   .0069459    11.01   0.000     .0628232    .0900757
   male_dummy |  -.0683999   .0167082    -4.09   0.000    -.1011773   -.0356225
 smoker_dummy |   .2800244   .0388257     7.21   0.000      .203858    .3561908
southwest_dummy |   .0242386    .023617     1.03   0.305     -.022092    .0705692
northwest_dummy |   .0569648   .0240526     2.37   0.018     .0097795      .10415
northeast_dummy |   .1018925   .0240631     4.23   0.000     .0546866    .1490983
        _cons |   7.566577   .0570209   132.70   0.000     7.454717    7.678438
------------------------------------------------------------------------------
```

4. Check for Assumptions:

Multicollinearity: Use the **Variance Inflation Factor (VIF)** to check for multicollinearity among predictors.

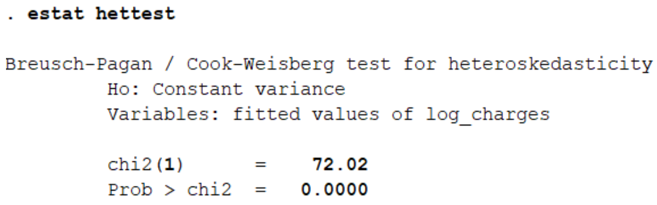


```
.
. vif

    Variable |       VIF       1/VIF  
-------------+----------------------
     charges |      4.01    0.249087
smoker_dummy |      3.55    0.281732
northwest_~y |      1.54    0.650056
northeast_~y |      1.54    0.650851
southwest_~y |      1.48    0.674258
         age |      1.37    0.728148
         bmi |      1.22    0.817156
    children |      1.01    0.987162
  male_dummy |      1.01    0.991062
-------------+----------------------
    Mean VIF |      1.86
```

Mean VIF: The mean VIF is 1.86, which indicates that overall, multicollinearity is not a serious issue in your model.

Heteroscedasticity: Perform the **Breusch-Pagan/Cook-Weisberg test** for heteroscedasticity.

estat hettest



```
. estat hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity 
         Ho: Constant variance
         Variables: fitted values of log_charges

         chi2(1)      =    72.02
         Prob > chi2  =   0.0000
```

Test Interpretation:

- **Null Hypothesis (Ho):** Constant variance (homoscedasticity) of the residuals.
    - The test assumes that the variance of the residuals is constant across all levels of the predicted values (log_charges in this case).
- **Test Statistic (chi2(1)):** 72.02
    - This is the chi-squared test statistic calculated by the Breusch-Pagan test.
- **p-value (Prob > chi2):** 0.0000
    - This p-value is very small ($p < 0.05$), leading you to reject the null hypothesis.

Conclusion:

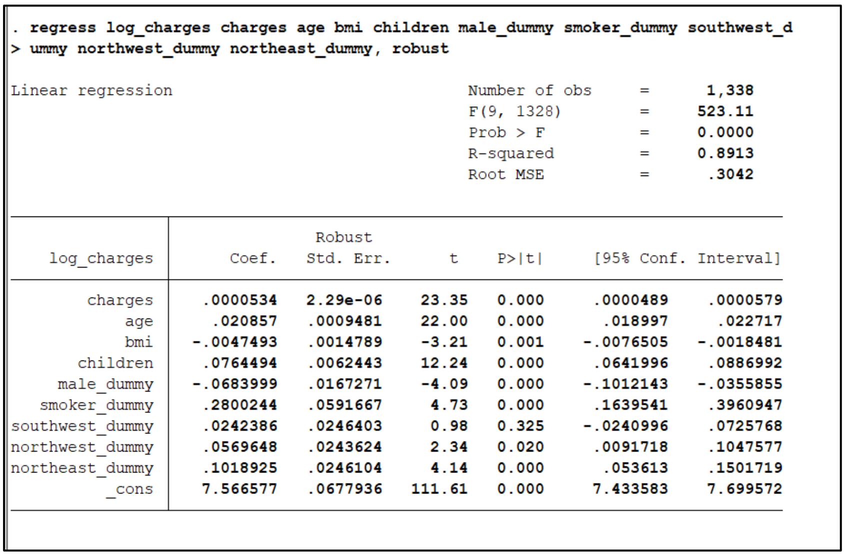
- **Reject the null hypothesis of constant variance.**

- The p-value of 0.0000 indicates that heteroskedasticity is present in the model. This means the variance of the residuals is not constant, which violates one of the key assumptions of ordinary least squares (OLS) regression.

Addressing Heteroskedasticity:

Heteroskedasticity can lead to biased standard errors, making statistical inferences (like p-values and confidence intervals) unreliable. We can address this issue using **Robust Standard Errors.**

The most straightforward solution is to use robust standard errors, which are adjusted to account for heteroskedasticity. This approach ensures your coefficient estimates remain unbiased, and your standard errors are valid even in the presence of heteroskedasticity.

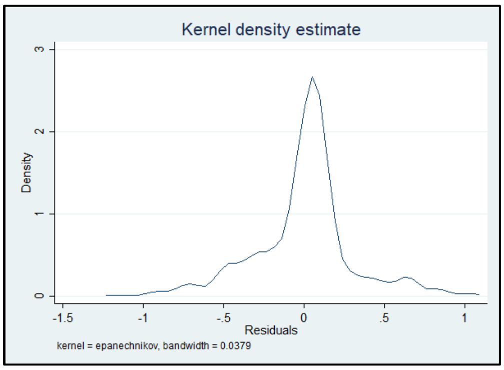


```
. regress log_charges charges age bmi children male_dummy smoker_dummy southwest_d
> ummy northwest_dummy northeast_dummy, robust

Linear regression                               Number of obs     =      1,338
                                                F(9, 1328)        =     523.11
                                                Prob > F          =     0.0000
                                                R-squared         =     0.8913
                                                Root MSE          =      .3042

------------------------------------------------------------------------------
                |               Robust
    log_charges |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
----------------+-------------------------------------------------------------
        charges |   .0000534   2.29e-06    23.35   0.000     .0000489    .0000579
            age |    .020857   .0009481    22.00   0.000      .018997     .022717
            bmi |  -.0047493   .0014789    -3.21   0.001    -.0076505   -.0018481
       children |   .0764494   .0062443    12.24   0.000     .0641996    .0886992
     male_dummy |  -.0683999   .0167271    -4.09   0.000    -.1012143   -.0355855
   smoker_dummy |   .2800244   .0591667     4.73   0.000     .1639541    .3960947
southwest_dummy |   .0242386   .0246403     0.98   0.325    -.0240996    .0725768
northwest_dummy |   .0569648   .0243624     2.34   0.020     .0091718    .1047577
northeast_dummy |   .1018925   .0246104     4.14   0.000      .053613    .1501719
          _cons |   7.566577   .0677936   111.61   0.000     7.433583    7.699572
------------------------------------------------------------------------------
```

5. Model Diagnostics:

Normality of residuals: Check whether the residuals are normally distributed.

predict residuals, resid
kdensity residuals

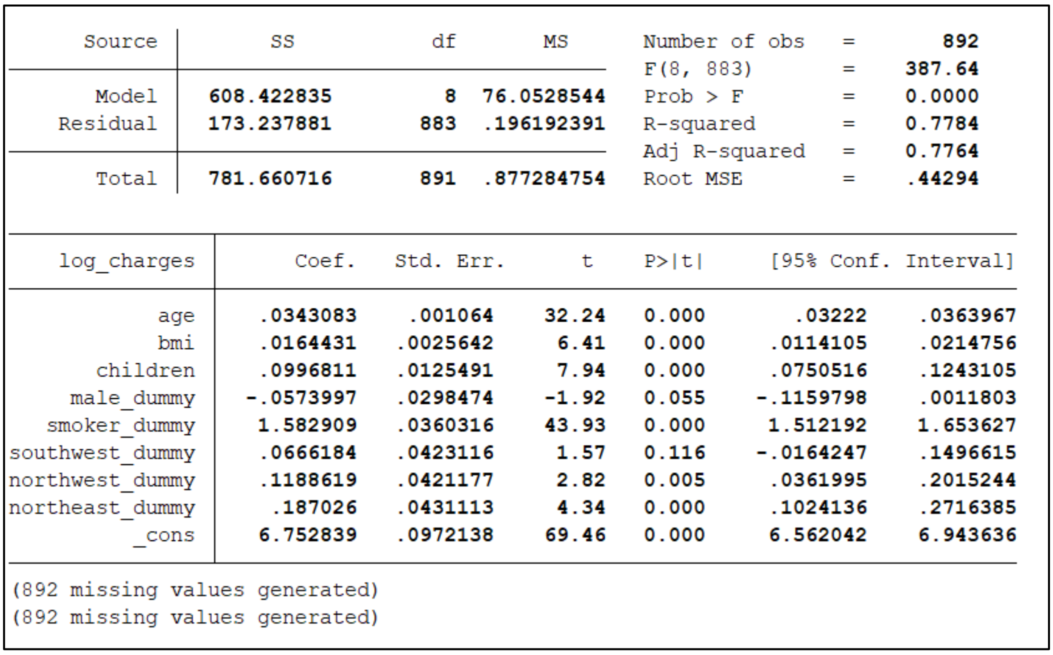


It can be clearly observed that the residuals are normally distributed

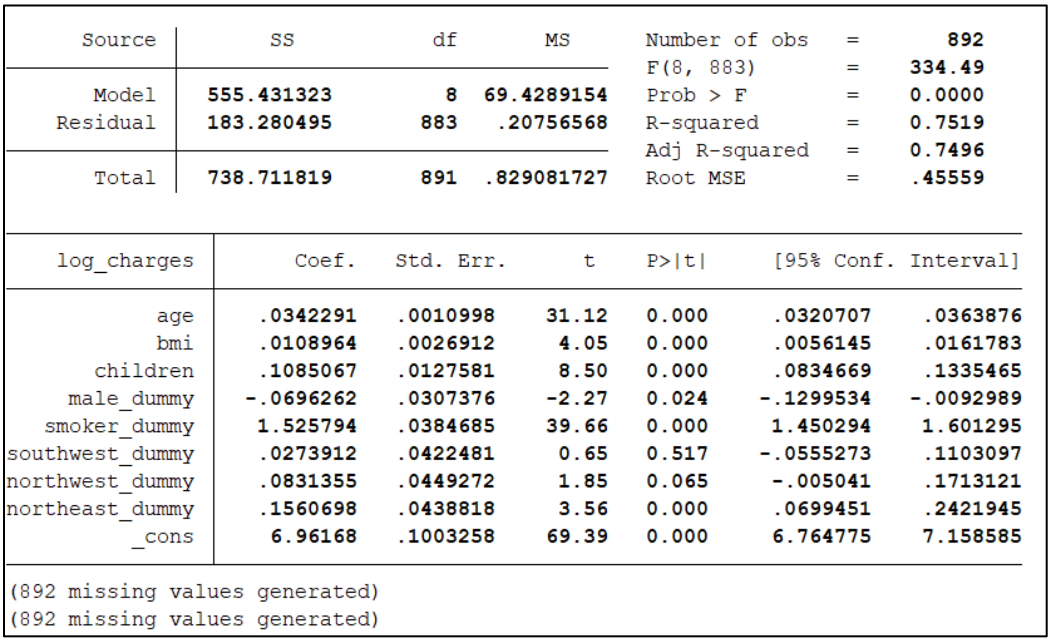
6. Cross-Validation:

- **K-fold Cross-Validation:** Perform cross-validation to assess the model's predictive performance.
    - **Divide the data into k folds**, train the model on k-1 folds, and validate on the remaining fold (k=3, in this case)

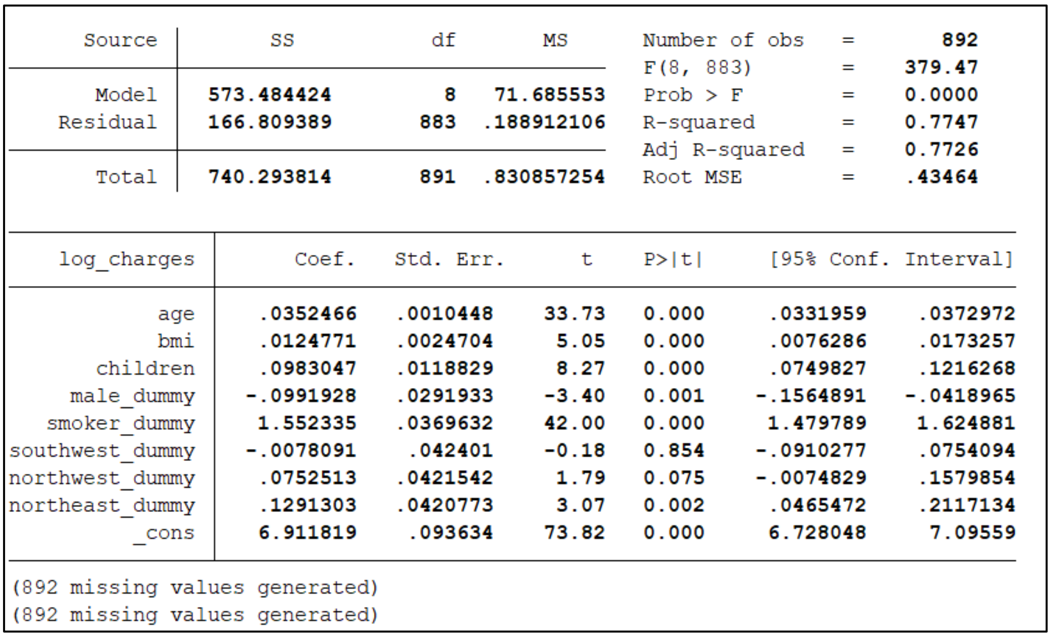
```
set seed 12345 // Set a seed for reproducibility

gen fold = mod(_n, 3) + 1  // Randomly assign fold numbers (1, 2, or 3)

local rmse_total = 0  // Initialize a variable to accumulate RMSE across folds

forval i = 1/3 {

   // Train the model on the data excluding fold `i'

   regress log_charges age bmi children male_dummy smoker_dummy southwest_dummy northwest_dummy northeast_dummy if fold != `i'


   // Predict values on the held-out fold `i'

   predict yhat if fold == `i', xb


   // Calculate squared errors for fold `i'

   gen se_`i' = (log_charges - yhat)^2 if fold == `i'


   // Summarize squared errors to calculate RMSE for fold `i'

   summ se_`i' if fold == `i', meanonly

   local mse_`i' = r(mean)  // Store the mean squared error for fold `i'

   local rmse_`i' = sqrt(`mse_`i'')  // Calculate RMSE for fold `i'


   // Add to total RMSE

   local rmse_total = `rmse_total' + `rmse_`i''


   // Drop temporary prediction and error variables for fold `i'

   drop yhat se_`i'

}

di "Average RMSE across 3 folds: " `rmse_total' / 3
```



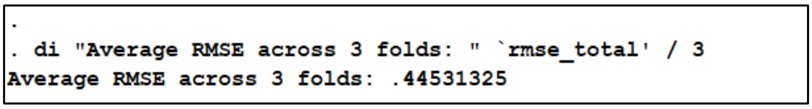
```
      Source |       SS           df       MS      Number of obs   =       892
-------------+----------------------------------   F(8, 883)       =    387.64
       Model |  608.422835         8  76.0528544   Prob > F        =    0.0000
    Residual |  173.237881       883  .196192391   R-squared       =    0.7784
-------------+----------------------------------   Adj R-squared   =    0.7764
       Total |  781.660716       891  .877284754   Root MSE        =    .44294

----------------------------------------------------------------------------------
     log_charges |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----------------+----------------------------------------------------------------
             age |   .0343083    .001064    32.24   0.000       .03222    .0363967
             bmi |   .0164431   .0025642     6.41   0.000     .0114105    .0214756
        children |   .0996811   .0125491     7.94   0.000     .0750516    .1243105
      male_dummy |  -.0573997   .0298474    -1.92   0.055    -.1159798    .0011803
    smoker_dummy |   1.582909   .0360316    43.93   0.000     1.512192    1.653627
southwest_dummy |   .0666184   .0423116     1.57   0.116    -.0164247    .1496615
northwest_dummy |   .1188619   .0421177     2.82   0.005     .0361995    .2015244
northeast_dummy |    .187026   .0431113     4.34   0.000     .1024136    .2716385
           _cons |   6.752839   .0972138    69.46   0.000     6.562042    6.943636
----------------------------------------------------------------------------------

(892 missing values generated)
(892 missing values generated)
```



```
      Source |       SS           df       MS      Number of obs   =       892
-------------+----------------------------------   F(8, 883)       =    334.49
       Model |  555.431323         8  69.4289154   Prob > F        =    0.0000
    Residual |  183.280495       883   .20756568   R-squared       =    0.7519
-------------+----------------------------------   Adj R-squared   =    0.7496
       Total |  738.711819       891  .829081727   Root MSE        =    .45559

----------------------------------------------------------------------------------
     log_charges |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----------------+----------------------------------------------------------------
             age |   .0342291   .0010998    31.12   0.000     .0320707    .0363876
             bmi |   .0108964   .0026912     4.05   0.000     .0056145    .0161783
        children |   .1085067   .0127581     8.50   0.000     .0834669    .1335465
      male_dummy |  -.0696262   .0307376    -2.27   0.024    -.1299534   -.0092989
    smoker_dummy |   1.525794   .0384685    39.66   0.000     1.450294    1.601295
southwest_dummy |   .0273912   .0422481     0.65   0.517    -.0555273    .1103097
northwest_dummy |   .0831355   .0449272     1.85   0.065     -.005041    .1713121
northeast_dummy |   .1560698   .0438818     3.56   0.000     .0699451    .2421945
           _cons |    6.96168   .1003258    69.39   0.000     6.764775    7.158585
----------------------------------------------------------------------------------

(892 missing values generated)
(892 missing values generated)
```



```
      Source |       SS           df       MS      Number of obs   =       892
-------------+----------------------------------   F(8, 883)       =    379.47
       Model |  573.484424         8   71.685553   Prob > F        =    0.0000
    Residual |  166.809389       883  .188912106   R-squared       =    0.7747
-------------+----------------------------------   Adj R-squared   =    0.7726
       Total |  740.293814       891  .830857254   Root MSE        =    .43464

----------------------------------------------------------------------------------
    log_charges |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
----------------+-----------------------------------------------------------------
            age |   .0352466   .0010448    33.73   0.000     .0331959    .0372972
            bmi |   .0124771   .0024704     5.05   0.000     .0076286    .0173257
       children |   .0983047   .0118829     8.27   0.000     .0749827    .1216268
     male_dummy |  -.0991928   .0291933    -3.40   0.001    -.1564891   -.0418965
   smoker_dummy |   1.552335   .0369632    42.00   0.000     1.479789    1.624881
southwest_dummy |  -.0078091    .042401    -0.18   0.854    -.0910277    .0754094
northwest_dummy |   .0752513   .0421542     1.79   0.075    -.0074829    .1579854
northeast_dummy |   .1291303   .0420773     3.07   0.002     .0465472    .2117134
          _cons |   6.911819    .093634    73.82   0.000     6.728048     7.09559
----------------------------------------------------------------------------------
(892 missing values generated)
(892 missing values generated)
```



```
.
. di "Average RMSE across 3 folds: " `rmse_total' / 3
Average RMSE across 3 folds: .44531325
```

Interpretation of Results:

I. **Average RMSE across 3 folds:**

   a. The **average RMSE** of **0.4453** across the three folds indicates the average prediction error in terms of the log-transformed medical charges

II. **Cross-validation in Stata:**

   a. This process ensures that the model is not overfitting the data. By splitting the data into three different subsets (folds) and training the model on each combination of two folds while testing on the third, you gain insights into the model's predictive performance.

III. **Coefficients:**

   a. The coefficients in the regression tables for each fold are fairly consistent, indicating stability across the folds. Most variables, like age, BMI, children, smoker status, and certain regions (e.g., northeast), have statistically significant effects on the log-transformed medical charges.

IV. **RMSE Consistency:**

   a. The RMSE for each fold (indicated by the different regression runs) is consistent, showing that the model is reasonably stable across different subsets of the data.

The southwest region is coming out to be insignificant across the different model. However, it doesn't necessarily mean that we should drop it automatically. Here are some considerations:

Insignificance vs. Relevance:

Insignificant variables don't contribute much to explaining the variation in the dependent variable. However, even insignificant variables can be theoretically important. Dropping a region could alter the interpretation of the model since regional effects are conceptually meaningful for this analysis.

7. Comparing different Models from Cross-Fold Validation:

i. **Evaluate the RMSE for Each Model:**

    a. **RMSE (Root Mean Squared Error)** gives us an idea of the average prediction error. The model with the lowest RMSE has better predictive accuracy.

    b. Comparing the RMSE across the models:

        i. **Fold 1 RMSE:** ≈0.45559\approx 0.45559≈0.45559

        ii. **Fold 2 RMSE:** ≈0.44294\approx 0.44294≈0.44294

        iii. **Fold 3 RMSE:** ≈0.43464\approx 0.43464≈0.43464

        iv. **Average RMSE:** 0.44530.44530.4453

ii. **Examine R-squared Values:**

    a. The **R-squared** measures the proportion of variance explained by the model. Higher R-squared values indicate better model fit. Compare both R-squared and Adjusted R-squared:

        i. **Fold 1:** R2 =0.7519, Adj.R2=0.7496

        ii. **Fold 2:** R2=0.7784, Adj.R2=0.7764

        iii. **Fold 3:** R2=0.7747, Adj.R2=0.7726

iii. Models from Fold 2 and Fold 3 have slightly better fit based on R-squared and Adjusted R-squared values.

iv. **Examine the Consistency of Coefficients:**

    a. A stable model should have similar coefficients across folds. By reviewing the coefficients for key variables (e.g., age, BMI, smoker status, regions) we can see that the coefficients are consistence across the different models.

    b. For example, the coefficient for age across folds is around 0.034, 0.034, 0.034, while the coefficient for smoker_dummy is consistently around 1.55−1.58, 1.55 - 1.58, 1.55−1.58. This consistency is a good sign that the model is stable across different subsets of the data.

v. **Examine Statistical Significance Across Folds:**

a. Variables like age, bmi, children, and smoker_dummy are consistently statistically significant (p-values close to 0). However, region dummy southwest_dummy, northwest_dummy have fluctuating significance levels across folds.

vi. **Compare the Coefficient of Variation of RMSE:**

a. We can compute the **Coefficient of Variation (CV)** of RMSE across the folds to measure variability in model performance. Lower variability suggests the model generalizes well across the different folds.

The formula for the coefficient of variation is:

CV=(RMSE / Mean of log_charges)×100

Model 1:

- RMSE = 0.45559
- Mean of log_charges = 9.098659

CV1=5.01%

Model 2:

- RMSE = 0.44294
- Mean of log_charges = 9.098659

CV2=4.87%

Model 3:

- RMSE = 0.43464
- Mean of log_charges = 9.098659

CV3=4.78%

Thus, the coefficient of variation (CV) of RMSE for each model is:

- **Model 1**: 5.01%
- **Model 2**: 4.87%
- **Model 3**: 4.78%

1. **Model Selection:**

   - Based on the evaluation metrics (RMSE, R-squared, consistency of coefficients, significance levels, and CV of RMSE), select the model that provides the best balance of predictive performance and generalizability.

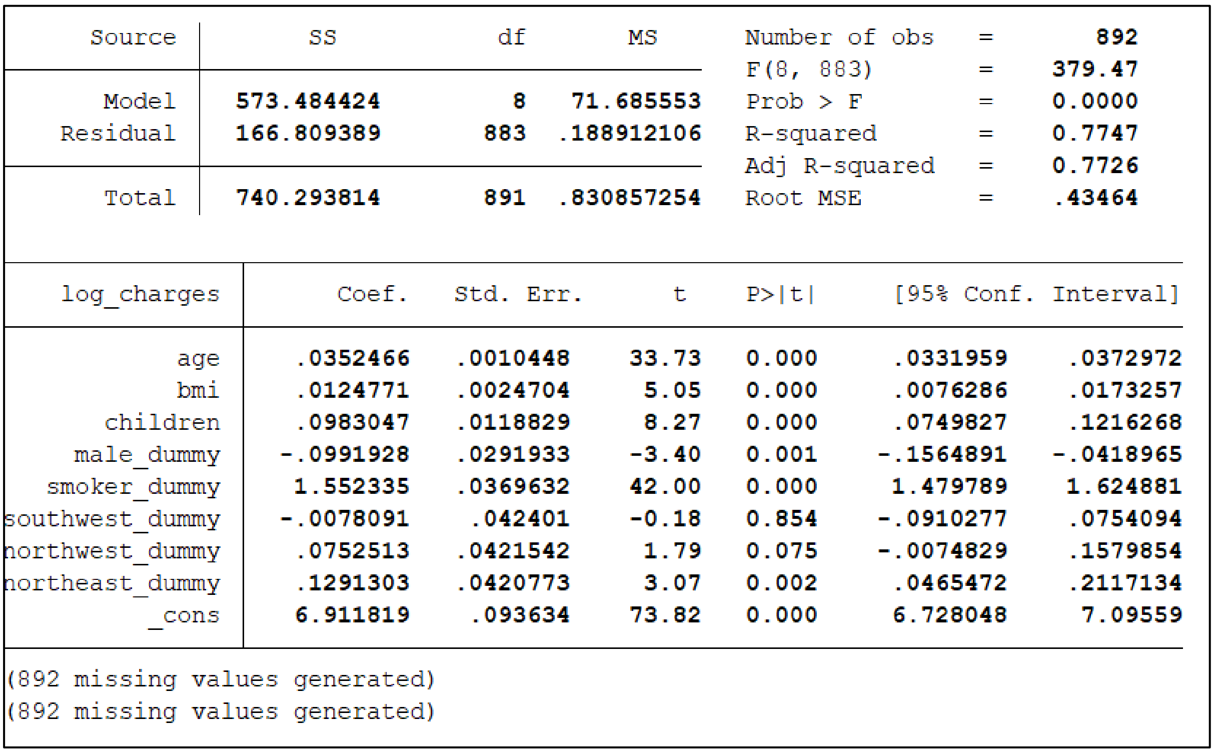
Summary of Comparison:

- **Fold 1:** RMSE=0.45559, R2=0.7519, CV1 = 5.01%
- **Fold 2:** RMSE=0.44294, R2=0.7784, CV2 = 4.87%
- **Fold 3:** RMSE=0.43464, R2=0.7747, CV3 = 4.78%

Best Model:

- **Fold 3 Model** seems to be the best based on the lowest RMSE (0.43464, 0.43464, 0.43464) and high R-squared (0.7747, 0.7747, 0.7747) and lowest CV of RMSE (5.01%, 4.87%, 4.78%) The coefficients are consistent across all three models, and Fold 3 has a good balance of fit and prediction accuracy.

8. Interpreting the comparatively best model:



```
      Source |       SS           df       MS      Number of obs   =       892
-------------+----------------------------------   F(8, 883)       =    379.47
       Model |  573.484424         8  71.685553   Prob > F        =    0.0000
    Residual |  166.809389       883  .188912106   R-squared       =    0.7747
-------------+----------------------------------   Adj R-squared   =    0.7726
       Total |  740.293814       891  .830857254   Root MSE        =    .43464

------------------------------------------------------------------------------
 log_charges |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0352466   .0010448    33.73   0.000     .0331959    .0372972
         bmi |   .0124771   .0024704     5.05   0.000     .0076286    .0173257
    children |   .0983047   .0118829     8.27   0.000     .0749827    .1216268
 male_dummy |  -.0991928   .0291933    -3.40   0.001    -.1564891   -.0418965
smoker_dummy |   1.552335   .0369632    42.00   0.000     1.479789    1.624881
southwest_dummy |  -.0078091    .042401    -0.18   0.854    -.0910277    .0754094
northwest_dummy |   .0752513   .0421542     1.79   0.075    -.0074829    .1579854
northeast_dummy |   .1291303   .0420773     3.07   0.002     .0465472    .2117134
       _cons |   6.911819    .093634    73.82   0.000     6.728048     7.09559
------------------------------------------------------------------------------
(892 missing values generated)
(892 missing values generated)
```

Overall Model Statistics:

- **Number of Observations (obs)**: 892
    - The model is based on 892 observations.
- **F-statistic**: 379.47
    - This indicates that the overall model is highly significant, meaning that the independent variables jointly explain the variation in the dependent variable (log_charges). A high F-statistic with a p-value of 0.0000 confirms that the model as a whole is statistically significant.
- **Prob > F**: 0.0000

  - The p-value associated with the F-statistic is 0.0000, meaning the overall model is significant at any conventional level (e.g., 1%, 5%, 10%).

- **R-squared**: 0.7747
  - About 77.47% of the variance in the dependent variable (log_charges) is explained by the independent variables. This indicates a good fit of the model.

- **Adjusted R-squared**: 0.7726
  - This adjusts the R-squared for the number of predictors and sample size. It is slightly lower than the R-squared, which is expected when multiple variables are included. It shows that approximately 77.26% of the variance is explained after adjusting for the number of predictors.

- **Root MSE**: 0.43464
  - The Root Mean Square Error (RMSE) is 0.43464, indicating the standard deviation of the residuals. It shows the average distance between the observed values and the predicted values from the model.

Coefficients and Individual Variable Interpretation:

- **Dependent Variable**: log_charges
  - This is the natural logarithm of medical charges. The coefficients represent the expected percentage change in medical charges for a one-unit change in the independent variable.

1. **Age**:
   - Coefficient: 0.0352
   - Interpretation: For every one-unit increase in age, log_charges is expected to increase by 0.0352, holding all other variables constant. Since this is log-transformed, this roughly corresponds to a 3.52% increase in charges for each additional year of age. This effect is statistically significant with a p-value of 0.000.
2. **BMI**:
   - Coefficient: 0.0125
   - Interpretation: For every one-unit increase in BMI, log_charges are expected to increase by 0.0125, or about 1.25%. This effect is also statistically significant with a p-value of 0.000.
3. **Children**:
   - Coefficient: 0.0983
   - Interpretation: Having one additional child is associated with an increase of approximately 9.83% in log_charges. This effect is statistically significant with a p-value of 0.000.
4. **Male Dummy**:
   - Coefficient: -0.0992

    - Interpretation: Being male is associated with a decrease of approximately 9.92% in log_charges compared to being female, holding all other variables constant. This effect is statistically significant with a p-value of 0.001.

5. **Smoker Dummy**:
    - Coefficient: 1.5523
    - Interpretation: Being a smoker is associated with an increase of approximately 155.23% in log_charges compared to non-smokers, holding other variables constant. This large effect is highly significant with a p-value of 0.000.
6. **Southwest Dummy**:
    - Coefficient: -0.0078
    - Interpretation: Living in the Southwest region is associated with a decrease of approximately 0.78% in log_charges compared to the reference region (likely Southeast), but this effect is not statistically significant (p-value = 0.854).
7. **Northwest Dummy**:
    - Coefficient: 0.0753
    - Interpretation: Living in the Northwest region is associated with an increase of approximately 7.53% in log_charges compared to the reference region, but this effect is marginally significant (p-value = 0.075).
8. **Northeast Dummy**:
    - Coefficient: 0.1291
    - Interpretation: Living in the Northeast region is associated with an increase of approximately 12.91% in log_charges compared to the reference region, and this effect is statistically significant (p-value = 0.002).
9. **Constant (_cons)**:
    - Coefficient: 6.9118
    - Interpretation: This is the expected value of log_charges when all independent variables are zero. This value is statistically significant with a p-value of 0.000, although it typically serves more as a point of reference in the context of the model.

Summary:

- The model explains a significant portion of the variance in log_charges (R-squared of 77.47%).
- Key predictors like age, BMI, number of children, gender, and smoking status are significant, with smoking having the largest positive effect on charges.
- The effect of living in different regions varies, with the Northeast region having a statistically significant positive impact on log_charges while the Southwest region does not show a significant impact.