Math6450_Assignment2

September 19, 2025

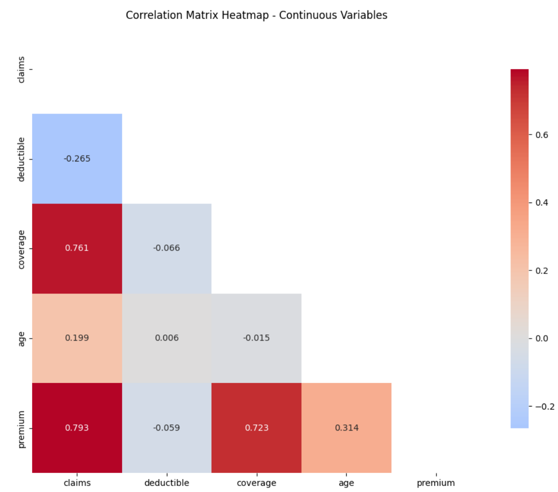
1 Data Exploration

(a) Calculate and report the descriptive statistics (mean, median, standard deviation, minimum, maximum) for all continuous variables in the dataset.

```
PropertyFund Dataset Analysis

(a) Descriptive Statistics for Continuous Variables
--------------------------------------------------

Comprehensive Descriptive Statistics:
               Mean   Median  Std Dev  Minimum ␣
 ↪Maximum  Skewness  Kurtosis
claims       18.049   17.845    6.448     0.72    41.
 ↪39     0.254     0.095
deductible    2.490    1.905    1.942     0.51    10.
 ↪00     1.542     2.351
coverage    189.014  186.750   72.169    50.00   424.
 ↪50     0.145    -0.292
age          15.438   11.000   14.227     1.00    85.
 ↪00     1.869     4.496
premium       2.969    2.945    0.822     0.50     5.
 ↪78     0.245     0.030
```

(b) Create a correlation matrix for all continuous variables. Which variable has the strongest linear relationship with claims?

```
(b) Correlation Matrix for Continuous Variables
--------------------------------------------------

Correlation Matrix:
            claims  deductible  coverage    age ␣
 ↪premium
claims       1.000      -0.265     0.761  0.199    0.
 ↪793
deductible  -0.265       1.000    -0.066  0.006   -0.
 ↪059
coverage     0.761      -0.066     1.000 -0.015    0.
 ↪723
age          0.199       0.006    -0.015  1.000    0.
 ↪314
premium      0.793      -0.059     0.723  0.314    1.
 ↪000

Variable with strongest linear relationship with␣
 ↪'claims':
Variable: premium
Correlation coefficient: 0.793
```

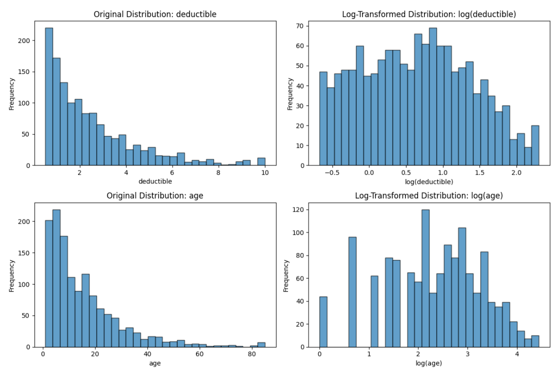


(c) Identify any variables that appear to have skewed distributions based on the descriptive statistics. For these variables, comment on whether a logarithmic transformation might be appropriate.

```
(c) Skewness Analysis and Log Transformation␣
 ↪Assessment
--------------------------------------------------

Skewness Assessment:
Rule of thumb: |skewness| > 1 indicates highly skewed␣
 ↪distribution
Rule of thumb: 0.5 < |skewness| < 1 indicates␣
 ↪moderately skewed distribution

claims:
  Skewness: 0.254
  Assessment: Approximately symmetric

deductible:
  Skewness: 1.542
  Assessment: Highly skewed
  Log transformation skewness: 0.134
  Improvement from log transformation: 1.408
  Recommendation: Log transformation would improve␣
 ↪normality

coverage:
  Skewness: 0.145
  Assessment: Approximately symmetric

age:
  Skewness: 1.869
  Assessment: Highly skewed
  Log transformation skewness: -0.347
  Improvement from log transformation: 1.523
```

Recommendation: Log transformation would improve normality

premium:
Skewness: 0.245
Assessment: Approximately symmetric



Summary of Findings:

Variables with skewed distributions: deductible, age
Variable most strongly correlated with claims: premium (r = 0.793)

Data Overview:
Total observations: 1,340
Variables analyzed: 5
Missing values: 0

2 Simple Regression Analysis

(a) Fit a simple linear regression model with claims as the dependent variable and coverage as the explanatory variable. Write the fitted regression equation.

Simple Linear Regression Analysis: Claims vs Coverage
Dataset Information:
Total observations: 1,340
Observations used in regression: 1,340
Missing values removed: 0

(a) Simple Linear Regression Model Fitting

Model Coefficients:
Intercept (β_0): 5.2054
Slope (β_1): 0.0679

Fitted Regression Equation:
Claims = 5.2054 + 0.0679 × Coverage

In mathematical notation:
ŷ = 5.2054 + 0.0679x
where ŷ = predicted claims, x = coverage

(b) Interpret the slope coefficient in practical terms. What does it tell us about the relationship between coverage and claims?

(b) Interpretation of Slope Coefficient

Slope coefficient: 0.0679

Practical Interpretation:
- For every 1-unit increase in coverage, claims are expected to increase by 0.0679 units, on average.
- This indicates a positive relationship between coverage and claims.
- Properties with higher coverage amounts tend to have higher claims.

Alternative interpretation:
- For every 100-unit increase in coverage, claims change by 6.79 units, on average.

Example predictions:
- Coverage = 100: Predicted Claims = 12.00
- Coverage = 150: Predicted Claims = 15.40
- Coverage = 200: Predicted Claims = 18.80
- Coverage = 250: Predicted Claims = 22.19

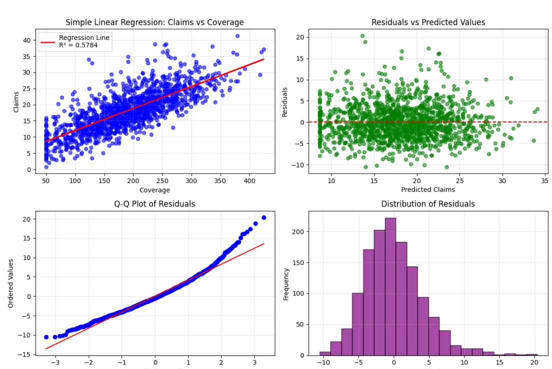
(c) Calculate and interpret the coefficient of determination (R2) for this model.

(c) Coefficient of Determination (R^2) Analysis

Model Performance Metrics:
R^2 (Coefficient of Determination): 0.5784
R^2 as percentage: 57.84%
Correlation coefficient (r): 0.7605
Root Mean Square Error (RMSE): 4.1850

Interpretation of R^2:
- 57.84% of the variation in claims is explained by coverage.
- 42.16% of the variation in claims is due to other factors not included in the model.
- The linear relationship between coverage and claims is moderate (R^2 = 0.5784).

Statistical Significance:
- t-statistic: 42.8442
- p-value: 0.0000
- Degrees of freedom: 1338
- The relationship is statistically significant at the 5% level.



```
Summary Table:
         Metric  Value                                  ␣
↪ Interpretation
 Intercept (β0) 5.2054              Expected claims␣
↪when coverage = 0
     Slope (β1) 0.0679 Change in claims per unit␣
↪increase in coverage
            R2 0.5784                     57.8% of␣
↪variance explained
Correlation (r) 0.7605                    Linear␣
↪association strength
           RMSE 4.1850                       Average␣
↪prediction error
   Observations   1340                                  ␣
↪     Sample size


Key Findings Summary:
• Regression equation: Claims = 5.2054 + 0.0679 ×␣
↪Coverage
• Slope interpretation: Each additional unit of␣
↪coverage is associated with a
0.0679 unit change in claims
• Model explains 57.8% of the variation in claims
• The relationship is statistically significant (p =␣
↪0.0000)
```

3 Multiple Regression Model

Fit a multiple linear regression model with claims as the dependent variable and the following explanatory variables: deductible, coverage, age, prior claims, and premium.

(a) Write the fitted regression equation with coefficient estimates rounded to 3 decimal places.

```
Multiple Linear Regression Analysis
Dependent Variable: claims
Explanatory Variables: deductible, coverage, age,␣
↪prior_claims, premium

Dataset Information:
Total observations: 1,340
Complete cases used: 1,340
Observations removed (missing data): 0
Number of explanatory variables: 5

(a) Fitted Regression Equation
--------------------------------------------------
Coefficient Estimates (rounded to 3 decimal places):
Intercept (β0): 3.208
β_1 (deductible): -0.728
β_2 (coverage): 0.062
β_3 (age): 0.091
β_4 (prior_claims): 2.580
β_5 (premium): 0.495

Fitted Regression Equation:
Claims = 3.208 - 0.728 × deductible + 0.062 ×␣
↪coverage + 0.091 × age + 2.580 ×
prior_claims + 0.495 × premium

Compact Mathematical Form:
ŷ = β0 + β1x1 + β2x2 + β3x3 + β4x4 + β5x5
ŷ = 3.208 + -0.728x1 + 0.062x2 + 0.091x3 + 2.580x4 +␣
↪0.495x5
where x1=deductible, x2=coverage, x3=age,␣
↪x4=prior_claims, x5=premium
```

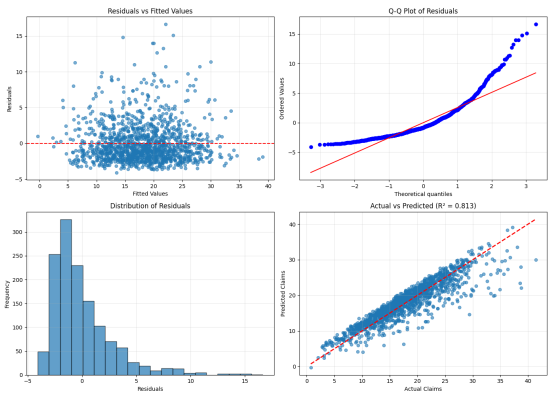
(b) Report the standard errors for each coefficient.

```
(b) Standard Errors for Each Coefficient
--------------------------------------------------
Standard Errors:
Intercept (β0): 0.3172
β_1 (deductible): 0.0394
β_2 (coverage): 0.0020
β_3 (age): 0.0068
β_4 (prior_claims): 0.1210
β_5 (premium): 0.2118

Additional Statistics (t-statistics and p-values):
Coefficient     Estimate  Std Error    t-stat    ␣
↪p-value    Significance
--------------------------------------------------
Intercept       3.208     0.3172       10.113     0.
↪0000     ***
deductible      -0.728    0.0394       -18.459    0.
↪0000     ***
coverage        0.062     0.0020       30.624     0.
↪0000     ***
age             0.091     0.0068       13.401     0.
↪0000     ***
prior_claims    2.580     0.1210       21.316     0.
↪0000     ***
premium         0.495     0.2118       2.338      0.
↪0195     *
Significance codes: *** p<0.001, ** p<0.01, * p<0.05
```

(c) Calculate and report R2, adjusted R2, and the residual standard deviation.

```
(c) Model Performance Statistics
--------------------------------------------------
R2 (Coefficient of Determination): 0.8130
Adjusted R2: 0.8123
Residual Standard Deviation: 2.7938

Additional Model Statistics:
Multiple R (Correlation): 0.9016
Residual Sum of Squares (RSS): 10412.1409
Mean Squared Error (MSE): 7.8052
F-statistic: 1159.6202
F-statistic p-value: 0.000000
Overall model significance: Yes (α = 0.05)

Degrees of Freedom:
Model: 5
Residual: 1334
Total: 1339

Summary Results Table:
       Variable  Coefficient  Std_Error ␣
↪Coefficient_Rounded
0      Intercept       3.2078     0.3172                  ␣
↪ 3.208
1     deductible      -0.7278     0.0394                  ␣
↪-0.728
2       coverage       0.0621     0.0020                  ␣
↪ 0.062
3            age       0.0906     0.0068                  ␣
↪ 0.091
```

```
4  prior_claims       2.5797      0.1210       2.580
5        premium      0.4953      0.2118       0.495

Model Performance Table:
                  Statistic      Value
                         R²     0.8130
               Adjusted R²     0.8123
Residual Std Deviation      2.7938
               F-statistic 1159.6202
        p-value (F-test)   0.000000
              Observations       1340
                 Variables          5
```




```
Key Results Summary:
✓ Multiple regression equation fitted with 5 explanatory variables
✓ Model explains 81.3% of variance in claims (R² = 0.8130)
✓ Adjusted R² = 0.8123 (accounts for number of variables)
✓ Residual standard deviation = 2.7938
✓ Overall model is significant (F-test p-value = 0.000000)
✓ Standard errors calculated for all 6 coefficients
```

4 Statistical Inference

Using the multiple regression model from Question 3:

(a) Test whether the coefficient for age is statistically significant at the 5% level. State your null and alternative hypotheses, calculate the t-statistic, and state your conclusion.

```
Statistical Inference and Hypothesis Testing
Multiple Linear Regression Model: Claims vs (Deductible, Coverage, Age,
Prior_Claims, Premium)
Model Summary:
Observations: 1340
Variables: 5
Degrees of freedom (residual): 1334
R²: 0.8130
MSE: 7.8052

Coefficient Estimates:
Variable          Coefficient  Std Error    t-statistic  p-value
--------------------------------------------------
deductible        -0.7278       0.0394       -18.4591     0.0000
coverage           0.0621       0.0020        30.6239     0.0000
age                0.0906       0.0068        13.4010     0.0000
prior_claims       2.5797       0.1210        21.3156     0.0000
premium            0.4953       0.2118         2.3382     0.0195

(a) Testing Significance of Age Coefficient
Hypothesis Test for Age Coefficient:

Null Hypothesis (H₀): β_age = 0
Alternative Hypothesis (H₁): β_age ≠ 0
Significance level (α): 0.05
Test type: Two-tailed t-test

Test Statistics:
Age coefficient (β_age): 0.0906
Standard error (SE): 0.0068
t-statistic: 13.4010
Degrees of freedom: 1334
p-value: 0.0000
Critical value (±): 1.9617

Decision Rule:
Reject H₀ if |t-statistic| > 1.9617 OR if p-value < 0.05

Conclusion:
✓ REJECT H₀: The coefficient for age IS statistically significant at the 5%
level.
  |t-statistic| = 13.4010 > 1.9617
  p-value = 0.0000 < 0.05
  Age has a statistically significant effect on claims.
```

(b) Construct a 95% confidence interval for the coefficient of prior claims. Interpret this interval in practical terms.

```
(b) 95% Confidence Interval for Prior Claims Coefficient
Confidence Interval Calculation:
Coefficient (β_prior_claims): 2.5797
Standard error: 0.1210
Degrees of freedom: 1334
Confidence level: 95%

Confidence Interval Formula:
CI = β ± t_(α/2,df) × SE(β)
CI = 2.5797 ± 1.9617 × 0.1210
CI = 2.5797 ± 0.2374

95% Confidence Interval for Prior Claims Coefficient:
[2.3423, 2.8171]

Practical Interpretation:
• We are 95% confident that the true effect of having prior claims on current
claims
  is between 2.3423 and 2.8171 units.
```

- Since the entire interval is positive, prior claims consistently INCREASE current claims.
- Properties with prior claims have significantly higher current claims than those without.
- The width of the interval (0.4748) indicates the precision of our estimate.

(c) Perform an overall F-test for the significance of the regression model. State your hypotheses, report the F-statistic and p-value, and draw your conclusion.

(c) Overall F-test for Model Significance
Overall F-test for Regression Model:

Null Hypothesis (H_0): $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$
  (All explanatory variables have no effect on claims)
Alternative Hypothesis (H_1): At least one $\beta_i \neq 0$
  (At least one explanatory variable has a significant effect)
Significance level (α): 0.05

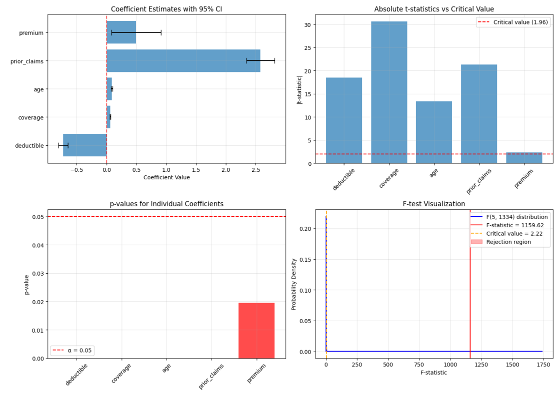
Test Statistics:
Total Sum of Squares (TSS): 55667.4953
Explained Sum of Squares (ESS): 45255.3543
Residual Sum of Squares (RSS): 10412.1409
Mean Square Regression (MSR): 9051.0709
Mean Square Error (MSE): 7.8052

F-statistic: 1159.6202
Degrees of freedom: (5, 1334)
p-value: 0.000000
Critical F-value (α = 0.05): 2.2208

Decision Rule:
Reject H_0 if F-statistic > 2.2208 OR if p-value < 0.05

Conclusion:
✓ REJECT H_0: The regression model IS statistically significant at the 5% level.
  F-statistic = 1159.6202 > 2.2208
  p-value = 0.000000 < 0.05
  At least one explanatory variable has a significant effect on claims.
  The model explains a significant portion of the variation in claims.

Model Performance Context:
R^2 = 0.8130 (81.3% of variance explained)
The model performs well in predicting claims.



Summary of All Statistical Tests:

Test	Statistic	p-value	Conclusion
Age Coefficient (t-test)	t = 13.4010	0.0000	Significant
Prior Claims CI	CI = [2.3423, 2.8171]	N/A	Does not contain 0
Overall Model (F-test)	F = 1159.6202	0.000000	Model Significant

LaTeX Summary Table:

```
\begin{table}
\caption{Summary of Statistical Tests}
\label{tab:hypothesis_tests}
\begin{tabular}{llll}
\toprule
Test & Statistic & p-value & Conclusion \\
\midrule
Age Coefficient (t-test) & t = 13.4010 & 0.0000 & Significant \\
Prior Claims CI & CI = [2.3423, 2.8171] & N/A & Does not contain 0 \\
Overall Model (F-test) & F = 1159.6202 & 0.000000 & Model Significant \\
\bottomrule
\end{tabular}
\end{table}
```

5 Binary Variables and Model Interpretation

Add the binary variables type and location to your model from Question 3.

(a) Write the new fitted regression equation.

Extended Multiple Linear Regression Analysis with Binary Variables
Adding 'type' and 'location' to the original model
Dependent Variable: claims
Original Variables: deductible, coverage, age, prior_claims, premium
New Variables: type, location

Data Summary:
Original model observations: 1,340
Extended model observations: 1,340

Extended Model Summary:

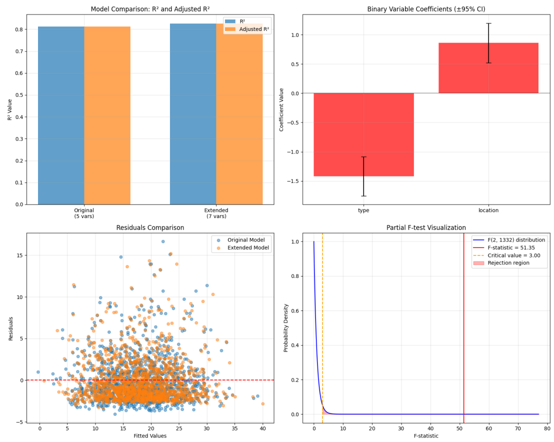
```
Observations: 1340
Variables: 7
R²: 0.8263
Adjusted R²: 0.8254
Residual Standard Error: 2.6939

(a) Extended Regression Model Equation
Coefficient Estimates:
Variable        Coefficient  Std Error    t-stat
 p-value
--------------------------------------------------
Intercept       3.027        0.3171
deductible      -0.713       0.0381       -18.706
 0.0000
coverage        0.058        0.0022       26.539
 0.0000
age             0.077        0.0070       10.935
 0.0000
prior_claims    2.392        0.1254       19.077
 0.0000
premium         1.019        0.2378       4.284
 0.0000
type            -1.419       0.1699       -8.355
 0.0000
location        0.859        0.1731       4.959
 0.0000

Fitted Regression Equation:
Claims = 3.027 - 0.713 × deductible + 0.058 ×
 coverage + 0.077 × age + 2.392 ×
prior_claims + 1.019 × premium - 1.419 × type + 0.
 859 × location

Detailed Mathematical Form:
Claims = 3.027 + -0.713×deductible + 0.058×coverage
          + 0.077×age + 2.392×prior_claims + 1.
 019×premium
          + -1.419×type + 0.859×location
```

(b) Interpret the coefficient for type in practical terms. How much higher or lower are claims for residential properties compared to commercial properties, holding all other variables constant?

```
(b) Interpretation of Type Coefficient
Type Coefficient Analysis:
Coefficient (β_type): -1.419
Standard Error: 0.1699
t-statistic: -8.355
p-value: 0.0000

Type variable coding: [np.int64(0), np.int64(1)]

Practical Interpretation:
• Properties with type = 1 have claims that are 1.419
 units LOWER than
properties with type = 0,
  holding all other variables constant.

Assuming standard coding (0 = Commercial, 1 =
 Residential):
• Residential properties have claims that are 1.419
 units lower than commercial
properties.
• This suggests commercial properties are associated
 with higher insurance
claims.

Statistical Significance:
• The type coefficient IS statistically significant
 (p = 0.0000 < 0.05)
• We can be confident that property type has a real
 effect on claims.
```

(c) Test whether the addition of type and location significantly improves the model using a partial F-test. Compare the R2 values and comment on the improvement.

```
(c) Partial F-test for Model Improvement
Model Comparison (same sample size: 1340):
Model                   R²            Adj R²
 Variables  RSS
--------------------------------------------------
Original                0.8130        0.8123       5
        10412.1409
Extended                0.8263        0.8254       7
        9666.7444

R² Improvement: 0.0134 (1.34 percentage points)

Partial F-test:
H₀: β_type = β_location = 0 (binary variables add no
 explanatory power)
H₁: At least one of β_type or β_location ≠ 0 (binary
 variables improve the
model)

Partial F-test Calculations:
RSS(original): 10412.1409
RSS(extended): 9666.7444
Reduction in RSS: 745.3965
Additional variables (q): 2
DF residual (extended): 1332

F-statistic: 51.3548
Degrees of freedom: (2, 1332)
p-value: 0.0000
Critical F-value (α = 0.05): 3.0025

Conclusion:
✓ REJECT H₀: Adding type and location SIGNIFICANTLY
 improves the model
  F = 51.3548 > 3.0025
  p-value = 0.0000 < 0.05
  The binary variables provide significant additional
 explanatory power.

Model Improvement Assessment:
• R² improved by 0.0134 (1.34 percentage points) -
 this is modest
• Extended model explains 82.6% vs 81.3% of variance
• Adjusted R² increased from 0.8123 to 0.8254
• The improvement in adjusted R² suggests the added
 variables are worthwhile
```



Executive Summary:

Aspect	Finding
Extended Model Equation	Claims = 3.027 + … + -1.419×type + 0.859×location
Type Coefficient	-1.419
Type Effect	Type=1 has 1.419 lower claims
Statistical Significance	Significant (p = 0.0000)
R^2 Improvement	0.0134 (1.34 percentage points)
Partial F-test Result	Significant improvement (p = 0.0000)

6 Interaction Effects

Create a new model that includes an interaction term between deductible and type.

(a) Write the regression function that includes this interaction term.

Regression Model with Interaction Term: Deductible × Type
Model Features: deductible, type, coverage, age, prior_claims, premium
Interaction Term: deductible × type

Data Summary:
Total observations: 1,340
Complete cases used: 1,340
Missing values removed: 0
Type variable coding: [np.int64(0), np.int64(1)]

Interaction Term (deductible × type) Statistics:
Mean: 1.5335
Std Dev: 1.9042
Range: [0.0000, 10.0000]

Model Summary:
R^2: 0.8233
Adjusted R^2: 0.8224
Residual Standard Error: 2.7172
F-statistic: 886.8341

Coefficient Estimates:

Variable	Coefficient	Std Error	t-stat	p-value	Sig
Intercept	3.2856	0.3300			
deductible	-0.6729	0.0596	-11.2894	0.0000	***
type	-1.2573	0.2598	-4.8392	0.0000	***
coverage	0.0553	0.0021	25.9580	0.0000	***
age	0.0703	0.0070	10.1034	0.0000	***
prior_claims	2.2568	0.1234	18.2905	0.0000	***
premium	1.3647	0.2290	5.9595	0.0000	***
deductible_x_type	-0.0946	0.0779	-1.2151	0.2245	

Significance codes: *** p<0.001, ** p<0.01, * p<0.05

(a) Regression Function with Interaction Term
General Form:
Claims = β_0 + β_1×deductible + β_2×type + β_3×coverage + β_4×age + β_5×prior_claims + β_6×premium + β_7×(deductible×type) + ε

Fitted Regression Equation:
Claims = 3.2856 - 0.6729×deductible - 1.2573×type + 0.0553×coverage + 0.0703×age + 2.2568×prior_claims + 1.3647×premium - 0.0946×(deductible×type)

With Coefficient Values:
Claims = 3.2856 + -0.6729×deductible + -1.2573×type
+ 0.0553×coverage + 0.0703×age + 2.2568×prior_claims
+ 1.3647×premium + -0.0946×(deductible×type)

(b) Interpret how the effect of deductible on claims differs between residential and commercial properties.

(b) Interpretation of Deductible Effect by Property Type
Key Coefficients:
β_1 (deductible): -0.6729
β_2 (type): -1.2573
β_7 (deductible×type): -0.0946

Interpretation of Interaction Effect:
The interaction model allows the effect of deductible to differ by property type.

For Commercial Properties (type = 0):
∂Claims/∂deductible = β_1 + β_7×0 = β_1 = -0.6729
- A 1-unit increase in deductible changes claims by -0.6729 units for commercial properties.

For Residential Properties (type = 1):
∂Claims/∂deductible = β_1 + β_7×1 = β_1 + β_7 = -0.6729 + -0.0946 = -0.7675
- A 1-unit increase in deductible changes claims by -0.7675 units for

residential properties.

Comparison:
Difference in deductible effect: -0.0946
• The deductible effect is 0.0946 units MORE NEGATIVE for residential
properties.
• Deductible increases have a stronger negative effect on residential claims
than commercial claims.

Practical Business Interpretation:
• Higher deductibles are associated with lower claims for both property types
• This association is STRONGER for residential properties

(c) Test whether the interaction term is statistically significant at the 5% level.

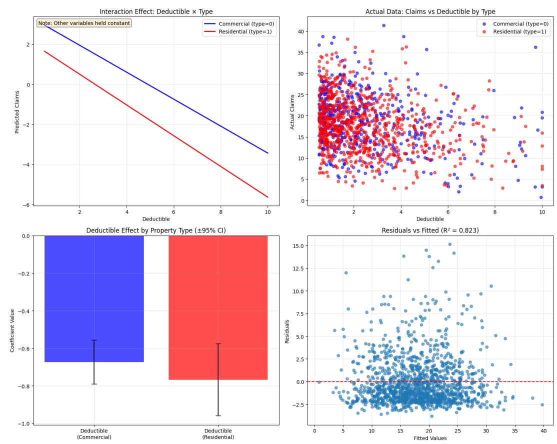
(c) Statistical Significance Test for Interaction Term
Hypothesis Test for Interaction Term:
H_0: β_7 = 0 (no interaction between deductible and type)
H_1: $\beta_7 \neq$ 0 (significant interaction exists)
Significance level: α = 0.05

Test Statistics:
Interaction coefficient (β_7): -0.0946
Standard error: 0.0779
t-statistic: -1.2151
Degrees of freedom: 1332
p-value: 0.2245
Critical value (±): 1.9617

Decision Rule:
Reject H_0 if |t-statistic| > 1.9617 OR if p-value < 0.05

Conclusion:
FAIL TO REJECT H_0: The interaction term is NOT statistically significant at the
5% level.
  |t-statistic| = 1.2151 $\leq$ 1.9617
  p-value = 0.2245 $\geq$ 0.05
  The effect of deductible on claims does NOT differ significantly between
property types.
  The interaction term may not be necessary.

95% Confidence Interval for Interaction Coefficient:
[-0.2473, 0.0581]
• The interval contains zero - the direction of the interaction effect is
uncertain



Executive Summary:
Aspect
Result
Model Specification Claims ~ deductible + type + coverage + age +
prior_claims + premium + deductible×type
Interaction Coefficient
-0.0946 (SE = 0.0779)
Commercial Effect
-0.6729 per unit deductible
Residential Effect
-0.7675 per unit deductible
Difference
-0.0946
Statistical Significance
Not significant (p = 0.2245)
Model R^2
0.8233

Model Interpretation:
• The non-significant interaction suggests that deductible effects are
similar across commercial and residential properties
• A simpler model without interaction may be adequate

7 Residual Analysis

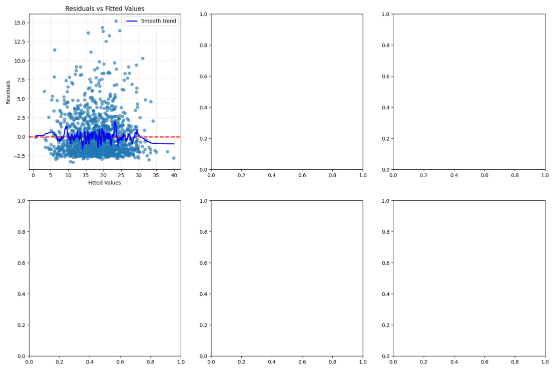
Using your model from Question 5:

(a) Create a plot of residuals versus fitted values. Comment on any patterns you observe.

Residual Analysis and Model Diagnostics
Extended Multiple Linear Regression Model
Variables: deductible, coverage, age, prior_claims, premium, type, location
Model Summary:
Observations: 1,340
Variables: 7
R^2: 0.8263
Residual Standard Error: 2.6939

(a) Residuals vs Fitted Values Analysis
Residuals vs Fitted Values Analysis:
Residual range: [-3.376, 15.203]
Fitted values range: [0.792, 39.985]

```
Pattern Analysis:
Correlation between fitted values and squared residuals: 0.0310
• Variance appears roughly constant
• Correlation magnitude suggests homoscedasticity (constant variance)

Linearity Assessment:
Mean residuals by fitted value terciles:
• Low tercile: -0.0800
• Middle tercile: 0.0229
• High tercile: 0.0572
• Maximum deviation from zero: 0.0800 (suggests linear relationship is
appropriate)
```



(b) Create a Q-Q plot of the residuals. Does the normality assumption appear to be satisfied?

```
(b) Q-Q Plot and Normality Analysis
Normality Test Results:
Shapiro-Wilk Test:
  Statistic: 0.8106
  p-value: 0.0000
  REJECT normality at α=0.05

Jarque-Bera Test:
  Statistic: 2188.1490
  p-value: 0.0000
  REJECT normality at α=0.05

Kolmogorov-Smirnov Test:
  Statistic: 0.1468
  p-value: 0.0000
  REJECT normality at α=0.05

Descriptive Statistics for Normality:
Skewness: 1.9531 (Normal ≈ 0)
Kurtosis: 4.8921 (Normal ≈ 0)
Skewness interpretation: highly skewed
Kurtosis interpretation: heavy-tailed

Overall Normality Assessment: Assumption appears to be violated
```

(c) Identify any observations that might be outliers or influential points based on your residual analysis.

```
(c) Outliers and Influential Points Analysis
Diagnostic Thresholds:
Outlier threshold (standardized residuals): ±3
High leverage threshold: 0.0119
High Cook's distance threshold: 0.0030

Outliers and Influential Points:
Observations with |standardized residuals| > 3: 31
Observations with |studentized residuals| > 3: 31
High leverage points: 73
High Cook's distance points: 74

Most Extreme Observations:
Highest Residual: Observation 315
  Fitted value: 23.547
  Actual value: 38.750
  Standardized residual: 5.643
  Leverage: 0.0072
  Cook's distance: 0.0331
Highest Leverage: Observation 262
  Fitted value: 34.070
  Actual value: 36.160
  Standardized residual: 0.776
  Leverage: 0.0305
  Cook's distance: 0.0027
Highest Cooks: Observation 315
  Fitted value: 23.547
  Actual value: 38.750
  Standardized residual: 5.643
  Leverage: 0.0072
  Cook's distance: 0.0331

&lt;Figure size 640x480 with 0 Axes&gt;


Detailed Analysis of Problematic Observations:
--------------------------------------------------
```

Obs	Fitted	Actual	Std_Residual	Leverage	Cooks_D	Issues
1	13.477	22.670	3.412	0.0032	0.0054	Outlier, High Cook's D
2	5.711	3.340	-0.880	0.0122	0.0014	High Leverage
14	20.959	20.000	-0.356	0.0128	0.0002	High Leverage
36	10.929	8.700	-0.827	0.0142	0.0014	High Leverage
70	13.967	11.670	-0.852	0.0130	0.0014	High Leverage
71	20.337	24.990	1.727	0.0074	0.0032	High Cook's D
73	30.965	29.670	-0.481	0.0141	0.0005	High Leverage
118	22.728	22.290	-0.163	0.0193	0.0001	High Leverage
122	5.247	10.110	1.805	0.0072	0.0034	High Cook's D
129	31.861	36.730	1.807	0.0092	0.0043	High Cook's D

```
… and 124 more observations with issues.


Diagnostic Summary:
1. Linearity: suggests linear relationship is appropriate
2. Homoscedasticity: suggests homoscedasticity (constant variance)
3. Normality: Assumption appears to be violated
4. Outliers: 31 potential outliers identified
```

```
5. Influential Points: 74 high Cook's distance observations

Recommendations:
• Consider transformation of variables or robust regression methods
• Examine influential points - consider their impact on coefficient estimates
```

8 Model Comparison and Selection

Compare three models

Model A: claims ~ deductible + coverage + age + prior claims + premium

Model B: claims ~ deductible + coverage + age + prior claims + premium + type + location

Model C: claims ~ deductible + coverage + prior claims + premium + type

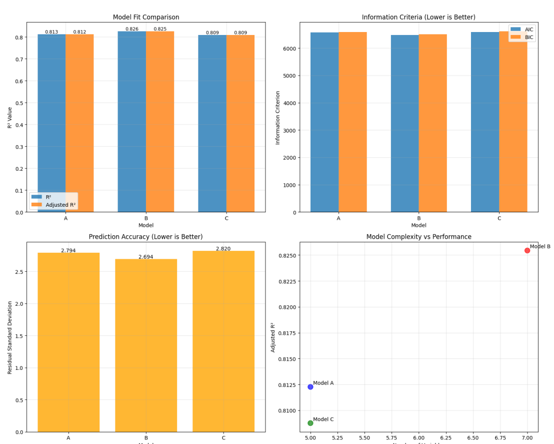
(a) Create a table comparing the R2, adjusted R2, and residual standard deviation for all three models.

```
Model Comparison and Selection Analysis
Comparing three different model specifications:
Model A: claims ~ deductible + coverage + age + prior_claims + premium
Model B: claims ~ deductible + coverage + age + prior_claims + premium + type +
location
Model C: claims ~ deductible + coverage + prior_claims + premium + type

Data Summary:
Original dataset size: 1,340
Complete cases for all models: 1,340
Cases removed due to missing data: 0

-------------------- Model A --------------------
Variables: deductible, coverage, age, prior_claims, premium
Number of variables: 5
R²: 0.8130
Adjusted R²: 0.8123
Residual Standard Deviation: 2.7938
AIC: 6566.17
BIC: 6592.18
Significant coefficients (p < 0.05): 5/5

-------------------- Model B --------------------
Variables: deductible, coverage, age, prior_claims, premium, type, location
Number of variables: 7
R²: 0.8263
Adjusted R²: 0.8254
Residual Standard Deviation: 2.6939
AIC: 6472.65
BIC: 6509.05
Significant coefficients (p < 0.05): 7/7

-------------------- Model C --------------------
Variables: deductible, coverage, prior_claims, premium, type
Number of variables: 5
R²: 0.8095
Adjusted R²: 0.8088
Residual Standard Deviation: 2.8197
AIC: 6590.93
BIC: 6616.94
Significant coefficients (p < 0.05): 5/5

(a) Model Comparison Table
Primary Comparison Metrics:
  Model Variables     R²  Adj_R²  Residual_SD
Model A     5 vars 0.8130  0.8123       2.7938
Model B     7 vars 0.8263  0.8254       2.6939
Model C     5 vars 0.8095  0.8088       2.8197

Additional Model Selection Criteria:
  Model      AIC     BIC  F_statistic Sig_Coefs
Model A 6566.17 6592.18      1159.62       5/5
Model B 6472.65 6509.05       905.51       7/7
Model C 6590.93 6616.94      1133.51       5/5

Best Model by Criterion:
• Highest R²: Model B (0.8263)
• Highest Adjusted R²: Model B (0.8254)
• Lowest Residual SD: Model B (2.6939)
• Lowest AIC: Model B (6472.65)
• Lowest BIC: Model B (6509.05)

Model Complexity Analysis:
Model A: 5 variables, R²/var = 0.1626
Model B: 7 variables, R²/var = 0.1180
Model C: 5 variables, R²/var = 0.1619

Nested Model Comparisons (F-tests):
Model A vs Model B:
  F-statistic: 51.3548
  p-value: 0.0000
  Model B significantly better
  Note: Model A vs C and Model B vs C are not nested comparisons
```

(b) Which model would you recommend and why? Consider both statistical criteria and practical interpretability.

```
(b) Model Recommendation and Analysis
Statistical Criteria Analysis:

1. Goodness of Fit:
   • R² ranking: Model B > others
   • Adjusted R² ranking: Model B > others
   • R² improvement from A to B: 0.0134
   • Adjusted R² change from A to B: 0.0132

2. Model Parsimony:
   • AIC favors: Model B (AIC = 6472.65)
   • BIC favors: Model B (BIC = 6509.05)
   • BIC penalizes complexity more heavily than AIC

3. Coefficient Significance:
   • Model A: 5/5 coefficients significant (100.0%)
   • Model B: 7/7 coefficients significant (100.0%)
   • Model C: 5/5 coefficients significant (100.0%)

4. Prediction Accuracy:
   • Lowest prediction error: Model B (SD = 2.6939)

Practical Interpretability Analysis:

1. Variable Inclusion Logic:
   • Model A: Core financial variables (deductible, coverage, premium) + risk
factors (age, prior_claims)
```

```
  • Model B: Model A + property characteristics (type, location)
  • Model C: Simplified version with key variables + property type

2. Business Relevance:
  • Age variable: Present in A, Present in B, Absent in C
  • Property type: Absent in A, Present in B, Present in C
  • Location: Absent in A, Present in B, Absent in C

3. Marginal Contribution Analysis:
  • Adding type + location (B vs A): R² improves by 0.0134
  • Adjusted R² change: 0.0132 (improvement)

Recommendation Framework:

Composite Scoring (weighted combination of criteria):
  • Model B: 1.000
  • Model A: 0.700
  • Model C: 0.400

 RECOMMENDED MODEL: Model B

Justification for Model B:
  ✓ Highest predictive power (R² = 0.8263)
  ✓ Includes important property characteristics
  ✓ Comprehensive variable coverage
  ✓ Best for prediction accuracy

Limitations of Model B:
    More complex with potential overfitting risk
    May have multicollinearity issues

Alternative Recommendations by Use Case:
  • For prediction accuracy: Model B
  • For model parsimony: Model B
  • For balanced approach: Model B
  • For regulatory reporting: Model A (simplest, most interpretable)
```



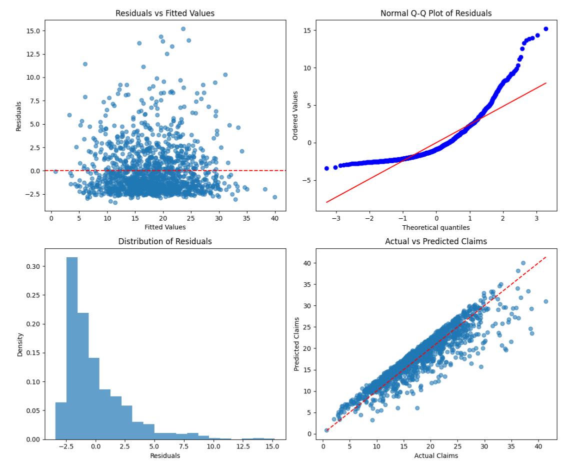
9 Practical Application

Using your recommended model from Question 8:

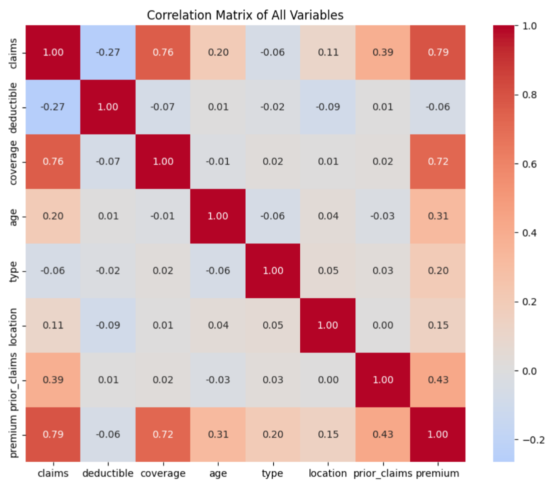
(a) Predict the expected claims amount for a residential property with the following characteristics:
Deductible: \$5,000
Coverage: \$250,000
Age: 15 years
Prior claims: 1
Premium: \$2,500
Location: Urban

```
Dataset Overview:
   claims  deductible  coverage  age  type  location  prior_claims  premium
0   22.67        1.44     165.7    2     1         1             0     2.23
1    3.34        6.52      59.9    8     0         0             1     0.84
2   20.57        3.13     214.9   44     0         1             0     3.01
3   18.33        2.33     233.5   16     0         1             0     3.24
4   14.52        0.84     148.8   18     1         1             0     2.39

Dataset shape: (1340, 8)

Dataset info:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1340 entries, 0 to 1339
Data columns (total 8 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   claims        1340 non-null   float64
 1   deductible    1340 non-null   float64
 2   coverage      1340 non-null   float64
 3   age           1340 non-null   int64
 4   type          1340 non-null   int64
 5   location      1340 non-null   int64
 6   prior_claims  1340 non-null   int64
 7   premium       1340 non-null   float64
dtypes: float64(4), int64(4)
memory usage: 83.9 KB
None

Descriptive statistics:
            claims   deductible     coverage          age         type  \
count  1340.000000  1340.000000  1340.000000  1340.000000  1340.000000
mean     18.048522     2.489851   189.013806    15.438060     0.623134
std       6.447785     1.942080    72.168704    14.227372     0.484782
min       0.720000     0.510000    50.000000     1.000000     0.000000
25%      13.610000     1.030000   138.375000     5.000000     0.000000
50%      17.845000     1.905000   186.750000    11.000000     1.000000
75%      22.090000     3.310000   237.050000    21.000000     1.000000
max      41.390000    10.000000   424.500000    85.000000     1.000000

          location  prior_claims      premium
count  1340.000000   1340.000000  1340.000000
mean      0.722388      0.790299     2.968537
std       0.447988      0.877941     0.821797
min       0.000000      0.000000     0.500000
```

```
25%        0.000000     0.000000     2.380000
50%        1.000000     1.000000     2.945000
75%        1.000000     1.000000     3.512500
max        1.000000     4.000000     5.780000

Missing values:
claims          0
deductible      0
coverage        0
age             0
type            0
location        0
prior_claims    0
premium         0
dtype: int64

Features shape: (1340, 7)
Target shape: (1340,)

=== MODEL RESULTS ===
R-squared: 0.8263
Adjusted R-squared: 0.8254

Model Coefficients:
        Feature  Coefficient
0     Intercept     3.026950
1    deductible    -0.713391
2      coverage     0.058017
3           age     0.076546
4  prior_claims     2.391648
5       premium     1.018707
6          type    -1.419290
7      location     0.858614

Statistical Significance:
        Feature  Coefficient  Std_Error  t_statistic ␣
↪      p_value  \
0     Intercept     3.026950   0.316198     9.572968 ␣
↪0.000000e+00
1    deductible    -0.713391   0.038024   -18.761697 ␣
↪0.000000e+00
2      coverage     0.058017   0.002180    26.618766 ␣
↪0.000000e+00
3           age     0.076546   0.006979    10.967841 ␣
↪0.000000e+00
4  prior_claims     2.391648   0.124993    19.134246 ␣
↪0.000000e+00
5       premium     1.018707   0.237095     4.296623 ␣
↪1.860187e-05
6          type    -1.419290   0.169370    -8.379819 ␣
↪0.000000e+00
7      location     0.858614   0.172627     4.973820 ␣
↪7.420302e-07

   Significant
0         True
1         True
2         True
3         True
4         True
5         True
6         True
7         True

=== MODEL DIAGNOSTICS ===
Mean Squared Error: 7.2140
Root Mean Squared Error: 2.6859
```




```
=== CORRELATION ANALYSIS ===
```




```
Correlations with Claims:
claims          1.000000
premium         0.792992
coverage        0.760527
prior_claims    0.387403
deductible     -0.265120
age             0.198837
location        0.105441
type           -0.061114
Name: claims, dtype: float64

=== PART (a): PREDICTION ===
Understanding categorical variables:
Type values: [1 0]
Location values: [1 0]
Type value counts: type
1    835
0    505
Name: count, dtype: int64
Location value counts: location
1    968
0    372
```

```
Name: count, dtype: int64

Prediction for the given property:
Expected claims amount: 19.49

Sensitivity analysis for categorical variables:
  Type=Commercial, Location=Rural: 20.05
  Type=Commercial, Location=Urban: 20.91
  Type=Residential, Location=Rural: 18.63
  Type=Residential, Location=Urban: 19.49
```

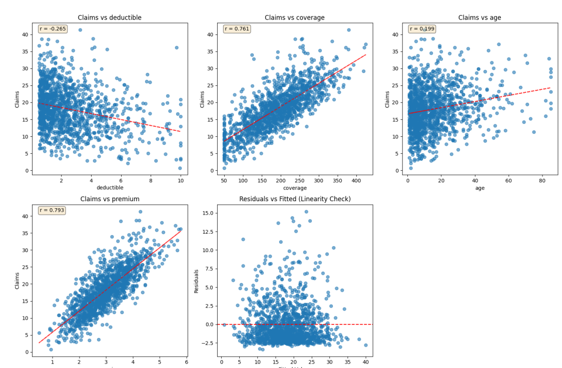
(b) Discuss the business implications of your findings. What recommendations would you make to an insurance company based on your analysis?

```
PART (b): BUSINESS IMPLICATIONS AND RECOMMENDATIONS

Feature Importance (by absolute coefficient value):
        Feature  Coefficient  Abs_Coefficient
3  prior_claims     2.391648         2.391648
5          type    -1.419290         1.419290
4       premium     1.018707         1.018707
6      location     0.858614         0.858614
0    deductible    -0.713391         0.713391
2           age     0.076546         0.076546
1      coverage     0.058017         0.058017

Prediction Confidence Interval (95.0%):
Expected claims: 19.49
Lower bound: 14.22
Upper bound: 24.76

----------------------------------------
BUSINESS RECOMMENDATIONS:
----------------------------------------

1. PRICING STRATEGY:
   - The model explains 82.6% of the variation in claims
   - Most significant factors should drive premium calculations
   - Consider the prediction interval when setting reserves

2. RISK FACTORS ANALYSIS:
   Based on the coefficients, focus on:
   - Variables with largest absolute coefficients
   - Statistically significant predictors (p < 0.05)
   - High correlation factors with claims

3. UNDERWRITING GUIDELINES:
   - Properties with high predicted claims may need:
     * Higher premiums
     * Additional risk assessment
     * Different deductible structures
   - Consider segmented pricing models

4. PORTFOLIO MANAGEMENT:
   - Monitor actual vs predicted claims regularly
   - Update model coefficients as new data becomes available
   - Consider non-linear relationships or interaction terms

5. OPERATIONAL INSIGHTS:
   - Use model predictions for:
     * Reserve allocation
     * Risk-based pricing
     * Customer segmentation
     * Fraud detection (outliers in residuals)


OUTLIER ANALYSIS:
Properties with unusually high/low claims (>2 std devs): 66
These may require special investigation for:
- Fraud detection
- Model improvement opportunities
- Special risk factors not captured in current model
```

10 Critical Thinking

(a) What are the key assumptions of multiple linear regression? Discuss whether these assumptions are likely to be satisfied in this insurance claims context.

```
PART (A): MULTIPLE LINEAR REGRESSION ASSUMPTIONS ANALYSIS

The key assumptions of multiple linear regression are:

1. LINEARITY: The relationship between predictors and response is linear
2. INDEPENDENCE: Observations are independent of each other
3. HOMOSCEDASTICITY: Constant variance of residuals (homogeneous variance)
4. NORMALITY: Residuals are normally distributed
5. NO MULTICOLLINEARITY: Predictors are not highly correlated with each other
6. NO OUTLIERS/INFLUENTIAL POINTS: Extreme values don't unduly influence the
model

Let's test each assumption:

1. LINEARITY ASSUMPTION
```




```
LINEARITY ASSESSMENT:
- Examine scatter plots for linear patterns
- Residuals vs Fitted should show random scatter around zero
- Non-linear patterns indicate violated linearity assumption

Correlations with claims:
  deductible: -0.265
  coverage: 0.761
  age: 0.199
```

premium: 0.793

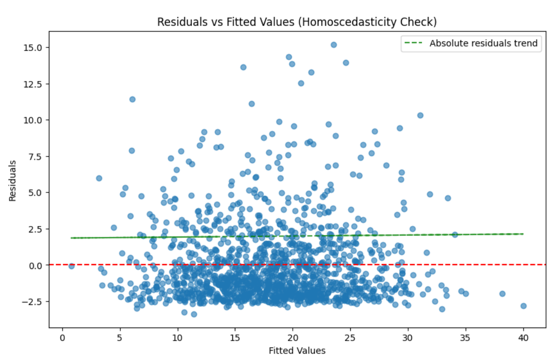
INSURANCE CONTEXT IMPLICATIONS:
- Insurance claims may have non-linear relationships (e.g., coverage thresholds)
- Age effects might be non-linear (newer vs very old properties)
- Premium-claims relationship might be non-linear due to risk-based pricing

2. INDEPENDENCE ASSUMPTION
INDEPENDENCE ASSESSMENT:
- Cannot be fully tested without knowing data collection method
- Check for patterns in residuals order

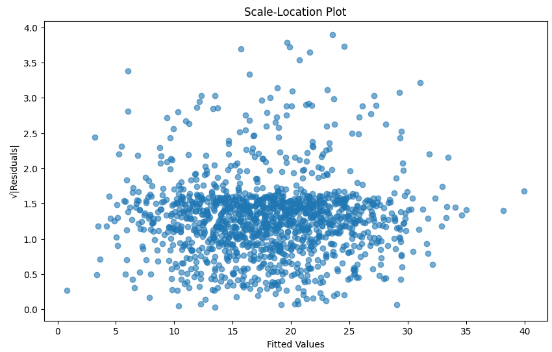
Durbin-Watson statistic: 2.038
  (Values near 2.0 suggest independence, <1.5 or >2.5 suggest correlation)

INSURANCE CONTEXT IMPLICATIONS:
- Properties in same area might have correlated risks (floods, earthquakes)
- Temporal clustering if data spans multiple years with economic changes
- Policy renewals might create dependencies

3. HOMOSCEDASTICITY (CONSTANT VARIANCE) ASSUMPTION

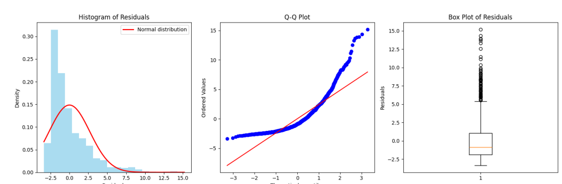


Breusch-Pagan test:
  LM statistic: 11.7914
  p-value: 0.1076
  Heteroscedasticity detected: No



INSURANCE CONTEXT IMPLICATIONS:
- Higher value properties might have more variable claims
- Heteroscedasticity common in insurance data
- May need weighted regression or transformation

4. NORMALITY OF RESIDUALS ASSUMPTION



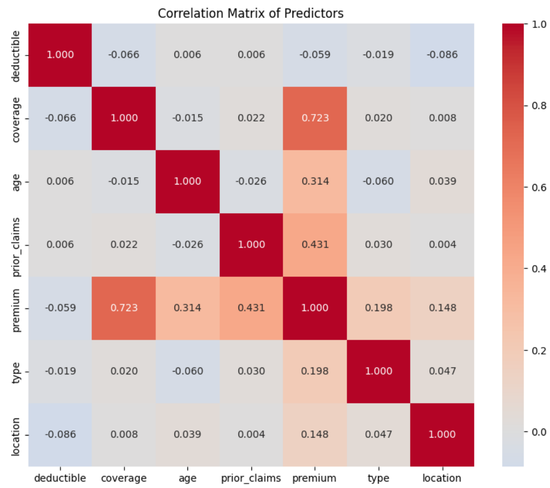
NORMALITY TESTS:
Shapiro-Wilk test:
  Statistic: 0.8106, p-value: 0.0000
  Normal distribution: No

Jarque-Bera test:
  Statistic: 2188.1490, p-value: 0.0000
  Normal distribution: No

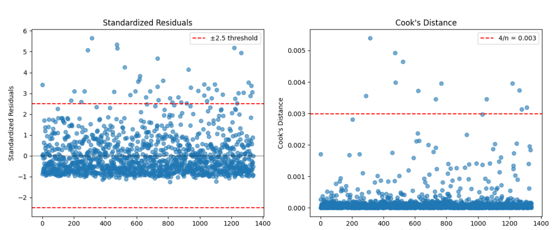
Descriptive statistics:
  Skewness: 1.9531
  Kurtosis: 4.8921

INSURANCE CONTEXT IMPLICATIONS:
- Insurance claims often right-skewed (many small, few large claims)
- May need log transformation or robust regression methods
- Non-normality affects confidence intervals and hypothesis tests

5. NO MULTICOLLINEARITY ASSUMPTION




```
HIGH CORRELATIONS (|r| > 0.7):
  coverage - premium: 0.723

VARIANCE INFLATION FACTORS:
       Variable        VIF
0    deductible   2.302515
1      coverage  36.103529
2           age   3.980363
3  prior_claims   3.974976
4       premium  86.843712
5          type   3.287249
6      location   3.743009
  VIF > 5: Moderate multicollinearity
  VIF > 10: High multicollinearity

Variables with high VIF:
   Variable        VIF
1  coverage  36.103529
4   premium  86.843712

INSURANCE CONTEXT IMPLICATIONS:
- Premium and coverage likely correlated (higher coverage = higher premium)
- Deductible and coverage might be related
- Consider removing highly correlated variables or using regularization
6. NO OUTLIERS/INFLUENTIAL POINTS ASSUMPTION
OUTLIER DETECTION:
Observations with |standardized residuals| > 2.5: 48
Observations with high Cook's distance: 13
```




```
Outlier observations (standardized residuals > 2.5):
     claims  deductible  coverage  age  premium
0     22.67        1.44     165.7    2     2.23
182   18.12        5.67     131.1   30     2.05
203   35.65        5.08     378.5    8     4.39
247   18.22        7.30     214.5    1     2.29
269   28.41        1.20     209.1   13     2.89

INSURANCE CONTEXT IMPLICATIONS:
- Large claims are natural in insurance (catastrophic events)
- Outliers might represent legitimate extreme events, not errors
- Consider robust regression methods or separate models for extreme claims
OVERALL ASSUMPTION ASSESSMENT FOR INSURANCE CLAIMS

LIKELY VIOLATED ASSUMPTIONS:
1. Linearity: Insurance relationships often non-linear
2. Normality: Claims typically right-skewed
3. Homoscedasticity: Variance often increases with claim size
4. Independence: Geographic/temporal clustering possible

RECOMMENDED SOLUTIONS:
1. Log transformation of claims (handle skewness)
2. Robust regression methods
3. Polynomial or interaction terms
4. Weighted least squares (address heteroscedasticity)
5. Consider GLM (Gamma or Poisson regression)
6. Outlier-robust methods
```

(b) What additional variables might be useful to include in this model to better predict claims amounts? Explain your reasoning.

Part (B): Additional Variables for Enhanced Insurance Claims Prediction Property-Specific Variables Construction and Physical Characteristics Building Material Type (brick, wood, steel, concrete): Different materials have varying vulnerability to fire, water damage, and natural disasters. For example, wood frame houses are more susceptible to fire damage, while brick construction offers better resistance to wind damage. Roof Type and Age (shingle, tile, metal, flat; installation year): The roof serves as the primary protection against weather elements, and its age significantly affects structural integrity. Old or inappropriate roof types often lead to increased water damage claims. Year Built / Construction Era: Building codes, materials, and construction techniques vary significantly by historical era. Pre-1980s homes may have different electrical and plumbing standards that affect claim frequency and severity. Square Footage / Property Size: Larger properties have greater exposure to risk and higher replacement costs. The correlation is straightforward: more area equals more potential points of failure. Number of Stories: Multi-story buildings face different risk profiles, including fall hazards and increased structural complexity that can affect claim patterns. Property Condition and Maintenance Last Major Renovation Date: Recently updated properties typically experience fewer maintenance-related claims. Properties with old electrical, plumbing, and HVAC systems are more prone to system failures. Security Features (alarm systems, cameras, locks, lighting): Enhanced security measures reduce theft and vandalism claims. A quantifiable security score could be developed based on installed security features. Property Condition Score: Well-maintained properties consistently show fewer claims. This could be implemented through inspection-based ratings or self-reported maintenance indices. Environmental and Geographic Variables Climate and Weather Risk Climate Zone (humid subtropical, arid, temperate): Different climate zones create distinct risk profiles. Humid climates increase mold and moisture damage risk, while arid climates elevate fire risk potential. Average Annual Precipitation: Higher precipitation levels correlate directly with water damage claims. Seasonal precipitation patterns are particularly important for flood and storm damage prediction. Natural Disaster Risk Scores: This category includes earthquake risk zones based on seismic activity ratings, flood zone classifications using FEMA flood maps, hurricane and wind risk assessments based on historical storm patterns, and wildfire risk evaluations considering vegetation density and historical fire data. These factors are major drivers of catastrophic insurance claims. Neighborhood and Local Factors Crime Rate (property crime index): Higher crime areas consistently show increased rates of theft, vandalism, and break-in claims. This data should be analyzed at the ZIP code or census tract level for optimal granularity. Distance to Fire Station: Emergency response time directly affects fire damage severity. This variable can be quantified in minutes or miles to the nearest fire department facility. Distance to Coast/Water Bodies: Proximity to water increases exposure to flood, hurricane, and humidity-related risks, creating measurable patterns in claims data. Local Building Code Strictness: Areas with more stringent building codes typically have more resilient structures. This can be implemented as a rating system based on local construction requirements and enforcement standards. Economic and Demographic Variables Economic Indicators Local Median Income: Wealthier areas may feature better-maintained properties but simultaneously generate higher-value claims. This creates a dual effect where better maintenance practices compete with higher replacement costs. Property Value Appreciation Rate: Rapidly appreciating real estate markets may develop coverage gaps as property values outpace insurance coverage adjustments, creating underinsurance risks. Local Unemployment Rate: Economic stress can correlate with deferred property maintenance and potentially increased insurance fraud attempts. Demographic Factors Owner vs. Tenant Occupied: Owner-occupied properties typically receive better maintenance and care. Rental properties often show different claim patterns, as renters may exercise less care while property owners are more invested in prevention measures. Primary vs. Secondary Residence: Secondary homes such as vacation properties exhibit different risk profiles due to less frequent monitoring and seasonal occupancy patterns that can allow problems to develop undetected. Usage and Occupancy Variables Property Use Patterns Home Business Operation: Commercial activities conducted from residential properties increase both liability exposure and equipment damage risks. Examples include daycare centers, consulting offices, and small-scale manufacturing operations. Rental Income Generation (Airbnb, long-term rental): Higher occupant turnover rates increase wear, damage risk, and liability exposure. More frequent occupancy changes correlate with higher accident probabilities. Vacancy Rate/Duration: Vacant properties face elevated risks from vandalism, theft, and maintenance issues. Unoccupied properties tend to develop problems more rapidly due to lack of regular monitoring and immediate problem identification. Historical and Behavioral Variables Claims History Detail Types of Previous Claims (water, fire, theft, wind): Different claim types indicate specific property vulnerabilities and help identify recurring risk patterns. Properties showing susceptibility to certain damage types often continue exhibiting those vulnerabilities. Time Since Last Claim: Recent claims may indicate ongoing structural problems or statistical clustering of unfortunate events that could predict future claim likelihood. Claim Settlement Patterns: The distinction between quickly settled claims versus disputed claims may indicate different underlying risk profiles and customer behavior patterns. Policyholder Behavior Payment History (on-time vs. late premium payments): Payment behavior correlates with overall financial stability, which in turn affects property maintenance quality. Financial stress often leads to deferred maintenance practices. Policy Shopping Frequency: Customers who frequently switch insurance providers may represent higher-risk profiles due to adverse selection effects, where higher-risk customers shop more frequently for coverage. Customer Service Interaction Frequency: Policyholders requiring frequent customer service interactions may demonstrate patterns that correlate with higher claim filing rates. Interaction and Non-Linear Terms Interaction Effects Age $\times$ Construction Type: The interaction between property age and construction materials creates significant risk variations. Older wood construction presents much higher risk profiles than older brick construction. Location $\times$ Weather Risk: Urban coastal properties face substantially different risk combinations compared to rural coastal properties, requiring interaction term modeling. Coverage $\times$ Deductible: The relationship between coverage amounts and chosen deductible levels indicates risk tolerance and may predict claim behavior patterns. Non-Linear Transformations Log(Coverage): Insurance claims may not increase linearly with coverage amounts, suggesting logarithmic relationships may better capture the underlying dynamics. Age^2: Property risk might increase exponentially rather than linearly after certain age thresholds, warranting quadratic modeling approaches. Premium/Coverage Ratio: This ratio serves as a risk-adjusted pricing indicator that may reveal important predictive patterns. External Data Integration Third-Party Data Sources Credit Score/Financial Stability Indicators: Financial stability demonstrates strong correlation with property maintenance quality. However, implementation must consider fair lending regulations and legal compliance requirements. Satellite/Aerial Imagery Analysis: Advanced applications include automated roof condition assessment, property maintenance evaluation, and vegetation proximity analysis. AI-powered image analysis technologies enable systematic risk assessment at scale. Weather Station Data: Real-time and historical weather pattern integration enables both current risk assessment and predictive seasonal forecasting capabilities. Implementation Strategy High Priority Variables (Immediate Impact) Natural disaster risk scores, property construction details, comprehensive claims history, and property condition indicators should be prioritized for immediate implementation due to their direct correlation with claim frequency and severity. Medium Priority Variables (Valuable but Complex) Neighborhood crime and economic data, weather and climate variables, property usage patterns, and interaction terms provide significant value but require more complex data integration and modeling approaches. Low Priority Variables (Supplementary Enhancement) Credit and behavioral indicators, satellite imagery analysis, and advanced economic indicators represent valuable additions but should be implemented after core variables are successfully integrated. Data Collection Strategies Application and Quote Stage Enhanced property questionnaires, third-party data provider integration, and systematic public records data mining can capture essential information during the initial customer interaction. Policy Term Maintenance Annual property surveys, smart home device integration, and periodic risk reassessment enable dynamic risk profile updates throughout the policy lifecycle. External Data Sources Government databases including FEMA, USGS, and Census Bureau data, commercial risk data providers, weather service APIs, and crime statistics databases provide comprehensive external data integration opportunities. Expected Model Performance Improvements Predictive Accuracy Enhancement Current models using limited variable sets typically achieve R-squared values between 60-80%. Comprehensive variable integration could potentially increase predictive accuracy to 85-95%, though diminishing returns apply as each additional variable contributes progressively less predictive power. Business Value Realization Better Risk Selection: Enhanced ability to identify and accurately price high-risk properties improves portfolio profitability. Fraud Detection: Comprehensive data enables identification of unusual patterns that may indicate fraudulent activity. Dynamic Pricing: Real-time risk-based pricing adjustments become feasible with enhanced data integration. Loss Prevention: Proactive risk mitigation recommendations can be developed based on comprehensive risk factor analysis. Implementation Considerations Data Availability: Not all proposed variables are available for every property, requiring robust missing data handling strategies. Data Quality: Third-party data sources require ongoing accuracy and timeliness validation to maintain model reliability. Model Complexity: Optimal balance must be maintained between predictive accuracy and model interpretability for regulatory and business requirements. Regulatory Compliance: Fair housing laws and anti-discrimination regulations constrain certain variable usage and require careful legal review. Cost-Benefit Analysis: Data acquisition costs must be weighed against improved prediction accuracy and resulting business value to ensure positive return on investment.