Math6450_Assignment2

September 17, 2025

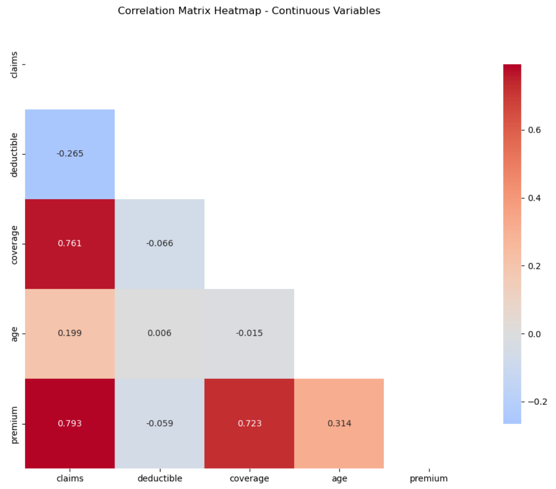
1 Data Exploration

(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
```

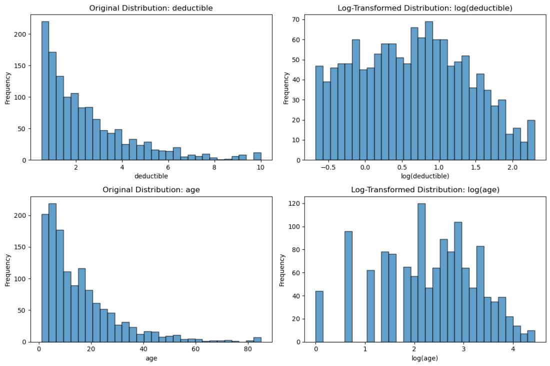
(b) Create a correlation matrix for all continuous variables. Which variable has the strongest linear relationship with claims?

```
(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.

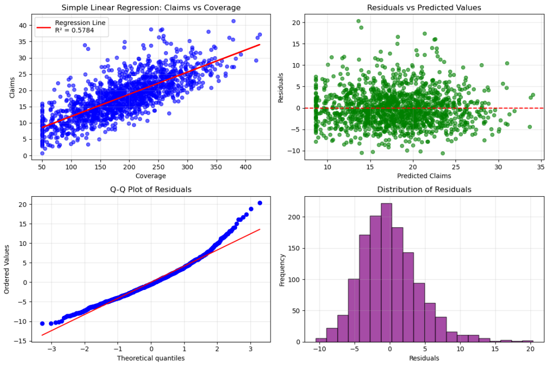
```
Simple Linear Regression Analysis: Claims vs Coverage
============================================================
Dataset Information:
Total observations: 1,340
Observations used in regression: 1,340
Missing values removed: 0

(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 (β_0) 5.2054            Expected claims when coverage = 0
     Slope (β_1) 0.0679 Change in claims per unit increase in coverage
             R^2 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 (β_0): 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 = β_0 + β_1x_1 + β_2x_2 + β x + β x + β x
ŷ = 3.208 + -0.728x_1 + 0.062x_2 + 0.091x + 2.580x + 0.495x
where x_1=deductible, x_2=coverage, x =age, x =prior_claims, x =premium
```

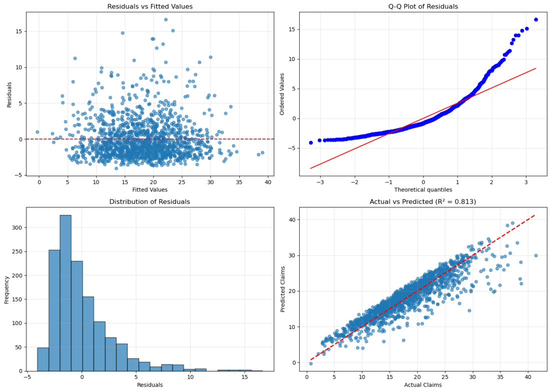
(b) Report the standard errors for each coefficient.

```
(b) Standard Errors for Each Coefficient
----------------------------------------
Standard Errors:
Intercept (β_0): 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
----------------------------------------
R^2 (Coefficient of Determination): 0.8130
Adjusted R^2: 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
                   R^2     0.8130
         Adjusted R^2     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 (R^2 = 0.
 ↪8130)
 Adjusted R^2 = 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
R^2: 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 (H_0): β_age = 0
Alternative Hypothesis (H_1): 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 H_0 if |t-statistic| > 1.9617 OR if p-value < 0.
 ↪05

Conclusion:
 REJECT H_0: 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.
```

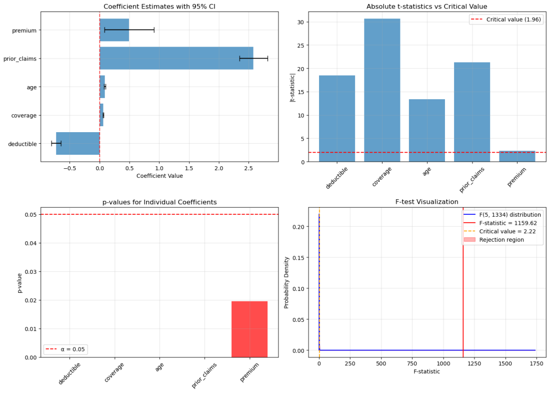
(b) Construct a 95% confidence interval for the coefficient of prior claims. Interpret this interval in practical terms.

```
(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$_0$): $\beta_1$ = $\beta_2$ = $\beta$ = $\beta$ = $\beta$ = 0
  (All explanatory variables have no effect on claims)
Alternative Hypothesis (H$_1$): At least one $\beta$ $\neq$ 0
  (At least one explanatory variable has a␣
 ↪significant effect)
Significance level ($\alpha$): 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 ($\alpha$ = 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.

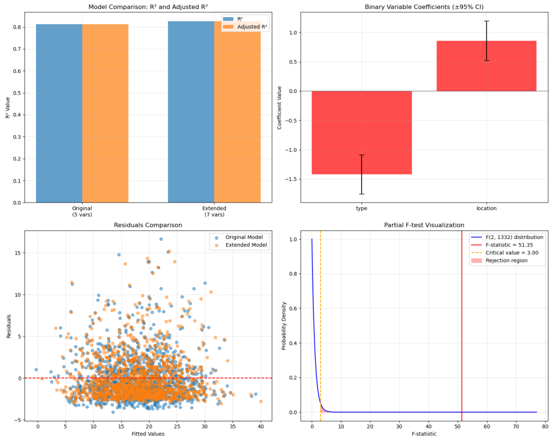
```
Extended Multiple Linear Regression Analysis with␣
 ↪Binary Variables
======================================================================
Adding 'type' and 'location' to the original model
Dependent Variable: claims
Original Variables: deductible, coverage, age,␣
 ↪prior_claims, premium
New Variables: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

Data Summary:
Original model observations: 1,340
Extended model observations: 1,340
Extended Model Summary:
Observations: 1340
Variables: 7
R^2: 0.8263
Adjusted R^2: 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: [0, 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               R^2          Adj R^2       Variables  RSS
----------------------------------------------------------------------
Original            0.8130       0.8123        5          10412.1409
Extended            0.8263       0.8254        7          9666.7444

R^2 Improvement: 0.0134 (1.34 percentage points)

Partial F-test:
H_0: β_type = β_location = 0 (binary variables add no explanatory power)
H_1: 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 H_0: 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:
• R^2 improved by 0.0134 (1.34 percentage points) - this is modest
• Extended model explains 82.6% vs 81.3% of variance
• Adjusted R^2 increased from 0.8123 to 0.8254
• The improvement in adjusted R^2 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: [0, 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
```

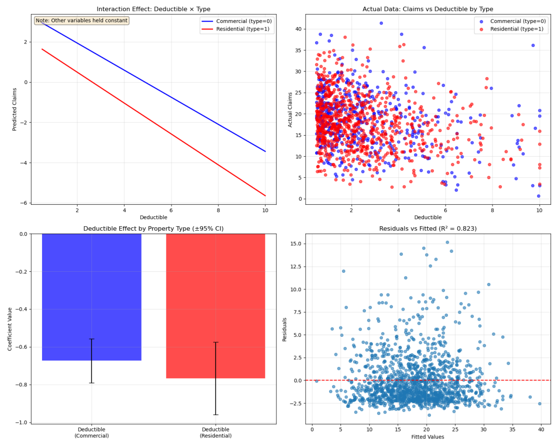
```
(a) Regression Function with Interaction Term
==================================================
General Form:
Claims = β_0 + β_1×deductible + β_2×type + β ×coverage␣
 ↪+ β ×age + β ×prior_claims +
β ×premium + β ×(deductible×type) +

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  (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 ×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 = $\beta_1$ + $\beta$ ×1 = $\beta_1$ + $\beta$  = -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:
$H_0$: $\beta$  = 0 (no interaction between deductible and␣
  ↪type)
$H_1$: $\beta$  ≠ 0 (significant interaction exists)
Significance level: $\alpha$ = 0.05

Test Statistics:
Interaction coefficient ($\beta$ ): -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 ≤ 1.9617
   p-value = 0.2245 ≥ 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
```

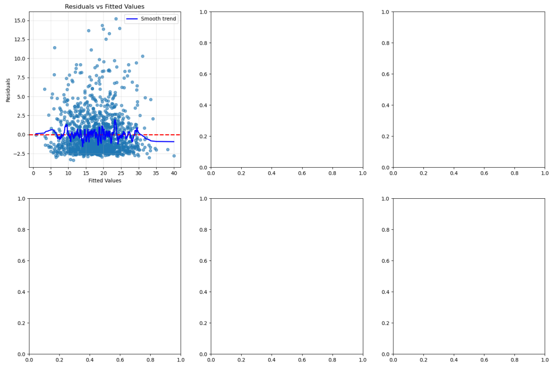
7 Residual Analysis

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

<Figure size 640x480 with 0 Axes>


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

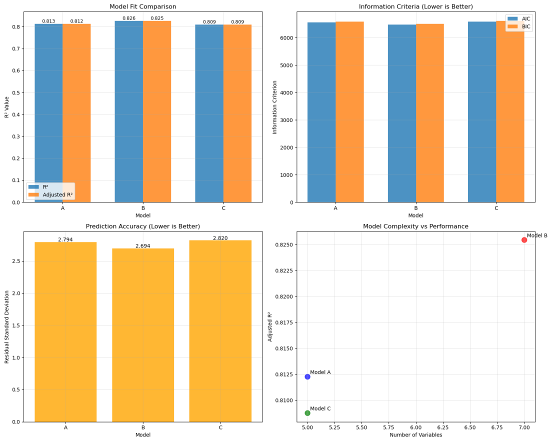
(a) Create a table comparing the R2, adjusted R2, and residual standard deviation for all three models.

```
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
R^2: 0.8130
Adjusted R^2: 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
R^2: 0.8263
Adjusted R^2: 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
R^2: 0.8095
Adjusted R^2: 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     R^2  Adj_R^2  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 R^2: Model B (0.8263)
• Highest Adjusted R^2: 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, R^2/var = 0.1626
Model B: 7 variables, R^2/var = 0.1180
Model C: 5 variables, R^2/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:
   • R^2 ranking: Model B > others
   • Adjusted R^2 ranking: Model B > others
   • R^2 improvement from A to B: 0.0134
   • Adjusted R^2 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): R^2 improves by 0.0134
   • Adjusted R^2 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 (R^2 = 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

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

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.

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