# Math6450 Assignment2: Multiple Linear Regression

Ian Tai Ahn

September 20, 2025

### 1 Data Exploration

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

Comprehensive Descriptive Statistics:

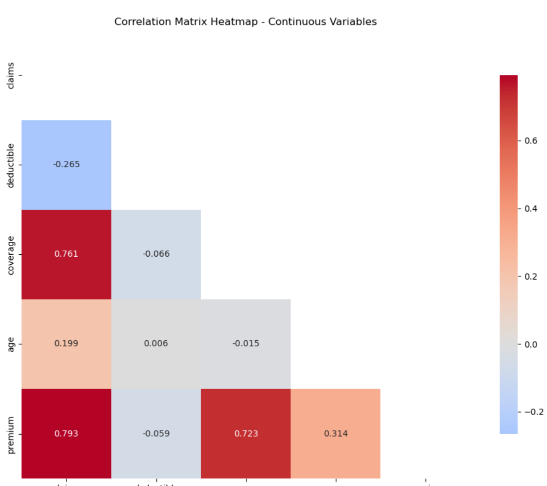
| | Mean | Median | Std Dev | Minimum | Maximum | Skewness | Kurtosis |
|---|---|---|---|---|---|---|---|
| claims | 18.049 | 17.845 | 6.448 | 0.72 | 41.39 | 0.254 | 0.095 |
| deductible | 2.490 | 1.905 | 1.942 | 0.51 | 10.00 | 1.542 | 2.351 |
| coverage | 189.014 | 186.750 | 72.169 | 50.00 | 424.50 | 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) 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) 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

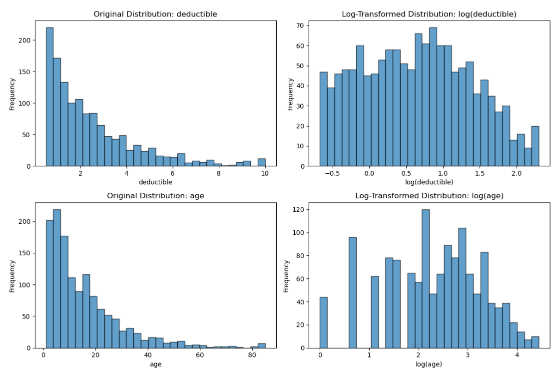
claims:
  Skewness: 0.254
  Assessment: Approximately symmetric

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 Linear Regression

(a) Simple Linear Regression Model Fitting
--------------------------------------------------
Model Coefficients:
Intercept ($\beta_0$): 5.2054
Slope ($\beta_1$): 0.0679

Fitted Regression Equation:
Claims = 5.2054 + 0.0679 × Coverage

In mathematical notation:
$\hat{y} = 5.2054 + 0.0679x$
where $\hat{y}$ = predicted claims, x = coverage

(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.

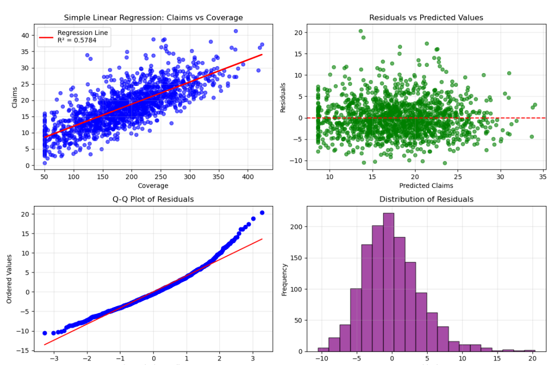
Alternative interpretation:
• For every 100-unit increase in coverage, claims change by 6.79 units, on
average.

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) 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.




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

Dependent Variable: claims
Explanatory Variables: deductible, coverage, age, prior_claims, premium

(a) Fitted Regression Equation
--------------------------------------------------
Coefficient Estimates (rounded to 3 decimal places):
Intercept ($\beta_0$): 3.208
$\beta$_1 (deductible): -0.728
$\beta$_2 (coverage): 0.062
$\beta$_3 (age): 0.091
$\beta$_4 (prior_claims): 2.580
$\beta$_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:
$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5$
$\hat{y} = 3.208 + -0.728x_1 + 0.062x_2 + 0.091x_3 + 2.580x_4 + 0.495x_5$
where $x_1$=deductible, $x_2$=coverage, $x_3$=age, $x_4$=prior_claims, $x_5$=premium

(b) Standard Errors for Each Coefficient
--------------------------------------------------
Standard Errors:
Intercept ($\beta_0$): 0.3172
$\beta$_1 (deductible): 0.0394
$\beta$_2 (coverage): 0.0020
$\beta$_3 (age): 0.0068
$\beta$_4 (prior_claims): 0.1210
$\beta$_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) 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 ($\alpha$ = 0.05)

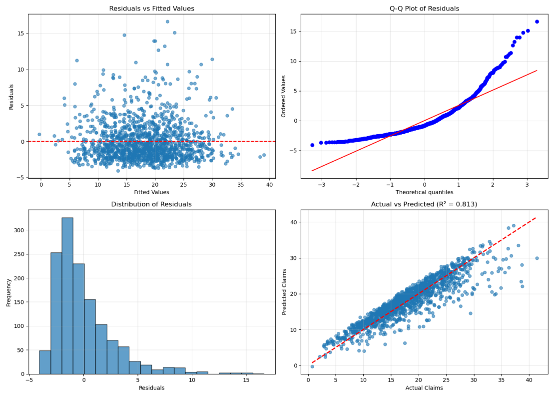
Degrees of Freedom:
Model: 5
Residual: 1334
Total: 1339

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

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 ($H_0$): $\beta$_age = 0
Alternative Hypothesis ($H_1$): $\beta$_age $\neq$ 0
Significance level ($\alpha$): 0.05
Test type: Two-tailed t-test

Test Statistics:
Age coefficient ($\beta$_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) 95% Confidence Interval for Prior Claims Coefficient
Confidence Interval Calculation:
Coefficient ($\beta$_prior_claims): 2.5797
Standard error: 0.1210
Degrees of freedom: 1334
Confidence level: 95%

Confidence Interval Formula:
CI = $\beta$ ± t_($\alpha$/2,df) × SE($\beta$)
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) 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 ($\alpha$): 0.05

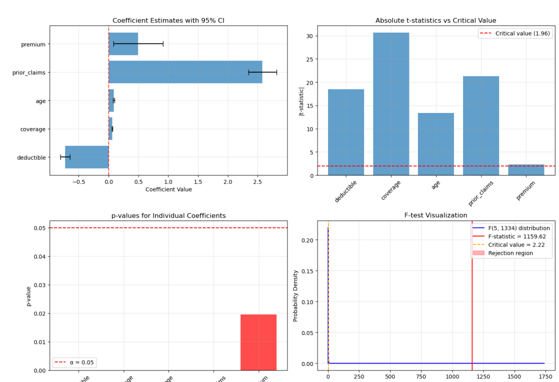
Test Statistics:
Total Sum of Squares (TSS): 55667.4953
Explained Sum of Squares (ESS): 45255.3543
Residual Sum of Squares (RSS): 10412.1409
Mean Square Regression (MSR): 9051.0709
Mean Square Error (MSE): 7.8052

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
```

## 5 Binary Variables and Model Interpretation

Adding 'type' and 'location' to the original model
Dependent Variable: claims
Original Variables: deductible, coverage, age, prior_claims, premium
New Variables: type, location

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) Interpretation of Type Coefficient
Type Coefficient Analysis:
Coefficient ($\beta$_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) 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)

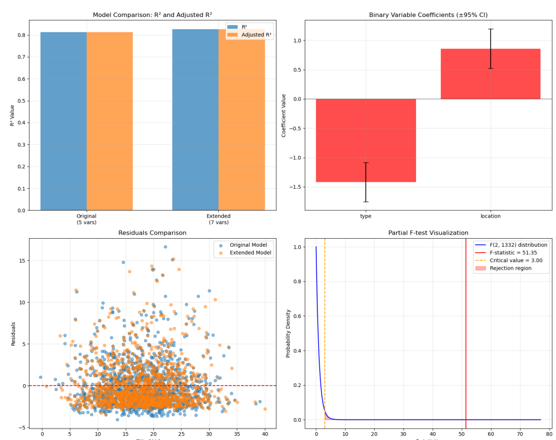
Partial F-test:
$H_0$: $\beta$_type = $\beta$_location = 0 (binary variables add no explanatory power)
$H_1$: At least one of $\beta$_type or $\beta$_location $\neq$ 0 (binary variables improve␣
↪the
model)

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 ($\alpha$ = 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




## 6 Interaction Effects

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

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 = $\beta_0$ + $\beta_1$×deductible + $\beta_2$×type + $\beta_3$×coverage + $\beta_4$×age + ↪$\beta_5$×prior_claims +
$\beta_6$×premium + $\beta_7$×(deductible×type) + $\varepsilon$

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) Interpretation of Deductible Effect by Property Type
Key Coefficients:
$\beta_1$ (deductible): -0.6729
$\beta_2$ (type): -1.2573
$\beta_7$ (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):
$\partial$Claims/$\partial$deductible = $\beta_1$ + $\beta_7$×0 = $\beta_1$ = -0.6729
• A 1-unit increase in deductible changes claims by -0.6729 units for ↪commercial
properties.

For Residential Properties (type = 1):
$\partial$Claims/$\partial$deductible = $\beta_1$ + $\beta_7$×1 = $\beta_1$ + $\beta_7$ = -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

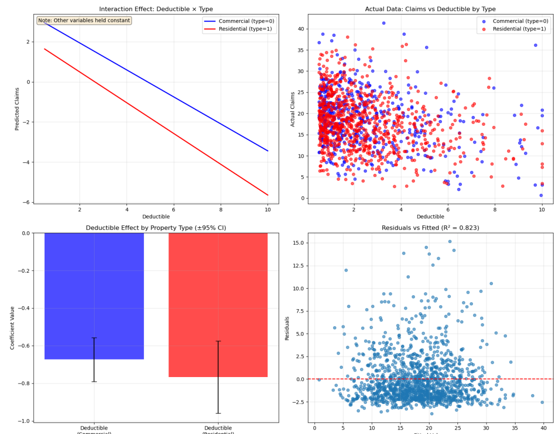
(c) Statistical Significance Test for Interaction Term
Hypothesis Test for Interaction Term:
$H_0$: $\beta_7$ = 0 (no interaction between deductible and type)
$H_1$: $\beta_7$ $\neq$ 0 (significant interaction exists)
Significance level: $\alpha$ = 0.05

Test Statistics:
Interaction coefficient ($\beta_7$): -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




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

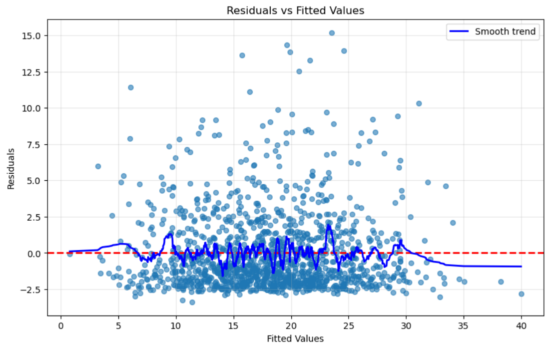
## 7 Residual Analysis

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)



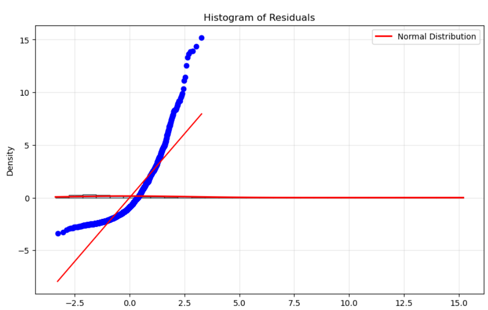

(b) Q-Q Plot and Normality Analysis
Normality Test Results:
Shapiro-Wilk Test:
  Statistic: 0.8106
  p-value: 0.0000
  REJECT normality at $\alpha$=0.05

Jarque-Bera Test:
  Statistic: 2188.1490
  p-value: 0.0000
  REJECT normality at $\alpha$=0.05

Kolmogorov-Smirnov Test:
  Statistic: 0.1468
  p-value: 0.0000
  REJECT normality at $\alpha$=0.05

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

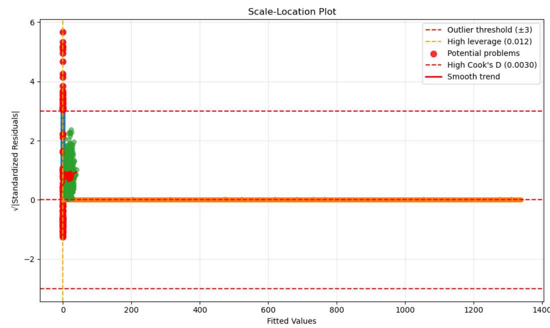
Overall Normality Assessment: Assumption appears to be violated




(c) 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




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

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) 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:

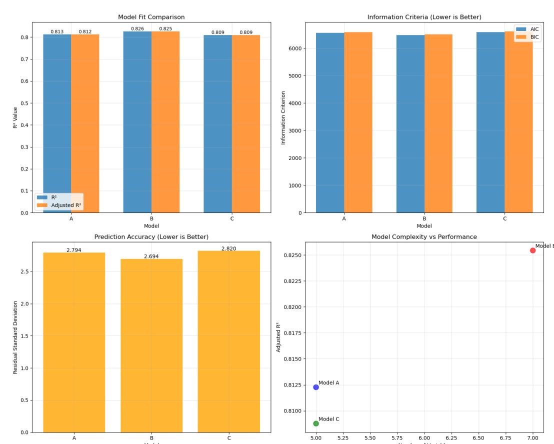
Composite Scoring (weighted combination of criteria):
- Model B: 1.000
- Model A: 0.700
- Model C: 0.400

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

Model Specification: claims ~ deductible + coverage + age + prior_claims + premium + type + location

Scenario: Residential Property Prediction

Model Performance: $R^2$ = 0.8263, Residual SE = 2.6939

Model B Fitted Equation:
Claims = 3.0270 − 0.7134×deductible + 0.0580×coverage + 0.0765×age + 2.3916×prior_claims + 1.0187×premium − 1.4193×type + 0.8586×location

(a) Prediction for Specific Residential Property
Property Characteristics:
- Deductible: $5,000
- Coverage: $250,000
- Age: 15 years
- Prior claims: 1

- Premium: $2,500
- Type: Residential
- Location: Urban

Numeric Coding for Prediction:
Type (Residential): 1
Location (Urban): 1

Prediction Input Vector:
deductible: 5000
coverage: 250000
age: 15
prior_claims: 1
premium: 2500
type: 1
location: 1

PREDICTED CLAIMS AMOUNT: $13,490.12

95% Confidence Interval for Prediction:
[$12,808.75, $14,171.48]
Margin of Error: ±$681.37

95% Prediction Interval:
[$12,808.73, $14,171.50]
Margin of Error: ±$681.39

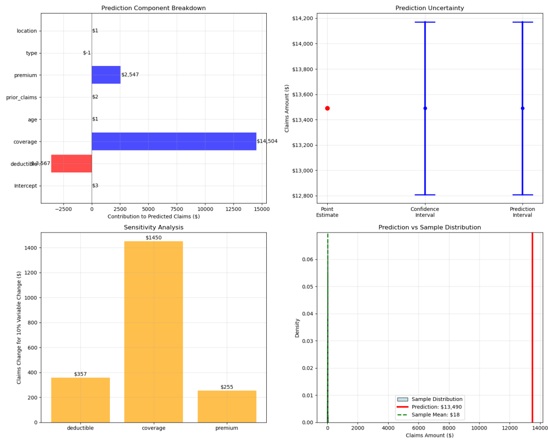
Interpretation:
- Point Estimate: We predict this property will have claims of $13,490.12
- Confidence Interval: We are 95% confident the average claims for properties
  with these characteristics is between $12,808.75 and $14,171.48
- Prediction Interval: There is a 95% chance that this specific property
  will have claims between $12,808.73 and $14,171.50

Prediction Component Analysis:
Base (Intercept): $3.03
deductible: -0.7134 × 5000 = $-3,566.95
coverage: 0.0580 × 250000 = $14,504.30
age: 0.0765 × 15 = $1.15
prior_claims: 2.3916 × 1 = $2.39
premium: 1.0187 × 2500 = $2,546.77
type: -1.4193 × 1 = $-1.42
location: 0.8586 × 1 = $0.86
Total: $13,490.12

Sensitivity Analysis:
How prediction changes with ±10% change in key continuous variables:
- deductible: ±$500 → claims change by ±$-356.70
- coverage: ±$25,000 → claims change by ±$1,450.43
- premium: ±$250 → claims change by ±$254.68

Comparison to Sample Data:
Predicted claims: $13,490.12
Sample mean claims: $18.05
Sample median claims: $17.84
Sample std dev claims: $6.45
Predicted value is at the 100.0th percentile of sample claims

Risk Assessment:
Risk Level: Above Average
This property's predicted claims are higher than typical for the portfolio




Prediction Summary Table:
========================================

| Metric | Value |
|---|---|
| Point Estimate | $13,490.12 |
| Standard Error | $347.33 |
| 95% Confidence Interval Lower | $12,808.75 |
| 95% Confidence Interval Upper | $14,171.48 |
| 95% Prediction Interval Lower | $12,808.73 |
| 95% Prediction Interval Upper | $14,171.50 |
| Sample Mean | $18.05 |
| Percentile Rank | 100.0th |

Property Input Summary:

| Variable | Value | Description |
|---|---|---|
| deductible | 5000 | $5,000 |
| coverage | 250000 | $250,000 |
| age | 15 | 15 years |
| prior_claims | 1 | 1 |
| premium | 2500 | $2,500 |
| type | 1 | Residential |
| location | 1 | Urban |

Key Insights:
This residential property is predicted to have higher than typical claims
The prediction has moderate uncertainty (±$681 confidence interval)
Coverage amount is the most sensitive factor for prediction changes
The property ranks at the 100th percentile for claims risk

(b) Business Implications

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)
```

## 10 Critical Thinking

```
(a) Multiple Linear Regression Key Assumptions

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

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) Additional Useful Variables

1. PROPERTY-SPECIFIC VARIABLES
• Construction: Building materials, roof type/age, year built, size, stories
• Condition: Recent renovations, security features, maintenance score

2. ENVIRONMENTAL & GEOGRAPHIC
• Climate: Climate zones, precipitation, natural disaster scores
• Location: Crime rates, distance to fire station/water, building codes

3. ECONOMIC & DEMOGRAPHIC
• Economic: Local income, property appreciation, unemployment rate
• Demographics: Owner vs tenant occupied, primary vs secondary residence

4. USAGE & BEHAVIORAL
• Property Use: Home business, rental income, vacancy duration
• Claims History: Previous claim types, time since last claim
• Behavior: Payment history, policy shopping, service interactions

5. ADVANCED MODELING
• Interaction Effects: Age×Construction, Location×Weather, Coverage×Deductible
• External Data: Credit scores, satellite imagery, weather APIs

6. IMPLEMENTATION PRIORITY
• HIGH: Natural disaster scores, construction details, claims history
• MEDIUM: Neighborhood data, weather variables, usage patterns
• LOW: Credit indicators, satellite analysis, economic metrics

7. EXPECTED OUTCOMES
• Model Accuracy: 60-80% → 85-95% predictive accuracy
• Benefits: Better risk selection, fraud detection, dynamic pricing

8. KEY CONSIDERATIONS
• Data availability varies by property
• Quality validation required for third-party data
• Regulatory compliance (fair housing laws)
• Cost-benefit analysis essential
```