**Predicting Medical Insurance Premiums**

**Author: Lewis So**  
**Course: Introduction to Business Analytics**

**Introduction**

The accurate prediction of medical insurance premiums is a critical task for insurance companies to ensure fair pricing, manage risk, and maintain competitiveness. This project aims to develop a predictive model for PremiumPrice using a dataset comprising 986 records of medical and demographic information, including variables such as Age, BMI, Diabetes, AnyTransplants, and others. The primary objective is to identify significant predictors of insurance premiums and construct a model that explains a substantial portion of the variance (targeting at least 60%) while validating its performance on an independent test set. This aligns with the course’s focus on exploratory data analysis (EDA), regression modeling, and validation techniques (Topics 1–3).

The dataset, sourced from [specify source if applicable, e.g., a course-provided file], includes both continuous (e.g., Age, BMI) and binary variables (e.g., AnyTransplants), offering a rich basis for analysis. Preliminary exploration suggests non-linear relationships (e.g., Age vs. PremiumPrice) and potential skewness in the response variable, necessitating advanced modeling techniques. The business context involves leveraging these predictions to inform pricing strategies, making the model’s interpretability and accuracy paramount. This report details the methodology, results, and discussion of the developed model, concluding with its implications and limitations.

**Methodology**

**Data Preparation**

The dataset, initially containing 986 rows and 13 columns, underwent preliminary cleaning to ensure data integrity. Missing values, if present, were handled by [specify method, e.g., imputation or removal], though the data appeared complete based on prior steps. Feature selection was guided by exploratory data analysis (Step 3), where variables such as KnownAllergies, Height, and Weight were excluded due to high correlation with BMI (e.g., Weight-BMI correlation r = 0.820679) and low explanatory power (e.g., Height r = 0.026910). The final feature set included Age, BMI, Diabetes, BloodPressureProblems, AnyTransplants, AnyChronicDiseases, HistoryOfCancerInFamily, and NumberOfMajorSurgeries as predictors, with PremiumPrice as the response variable.

**Exploratory Data Analysis**

Exploratory data analysis (Step 3) revealed key insights into the data structure. Scatter plots indicated a non-linear relationship between Age and PremiumPrice, suggesting the need for a quadratic term. Correlation analysis confirmed BMI as a relevant predictor, justifying the exclusion of Height and Weight. Binary variables such as AnyTransplants showed low prevalence (5–6%), indicating potential significance despite their rarity. These findings informed the modeling approach, emphasizing transformations and non-linear terms.

**Model Building**

The modeling process (Step 5) began with a 70/30 train-test split, resulting in 690 training and 296 testing observations, using a random\_state=42 for reproducibility. A log transformation was applied to PremiumPrice to address potential right-skewness, a common issue in monetary data. Initial ordinary least squares (OLS) regression identified significant predictors, but diagnostics revealed multicollinearity between Age and its quadratic term (Age²). To mitigate this, Age was centered (Age\_centered) before squaring (Age2\_centered), reducing multicollinearity (VIF 1.01–1.03).

Outlier handling was implemented by capping 30 observations with residuals exceeding 2 standard deviations to their predicted values, improving model stability. Weighted least squares (WLS) was employed to address heteroskedasticity, using inverse predicted values as weights. The final model was validated on the test set, with predictions converted back to the original scale using exponentiation for interpretation.

**Validation**

Validation involved transforming log-predicted values back to the original PremiumPrice scale and computing the Mean Squared Error (MSE) on the test set. Sample predictions were compared with actual values to assess accuracy, and the overall MSE provided a measure of predictive performance on the original scale.

**Results**

**Model Performance**

The weighted least squares model achieved an R-squared of 0.696, indicating that 69.6% of the variance in log(PremiumPrice) is explained by the predictors. The adjusted R-squared (0.693) accounts for the degrees of freedom, confirming the model’s efficiency with six predictors. The F-statistic of 260.8 (p = 6.06e-173) underscores the model’s overall statistical significance.

**Coefficients**

The following table summarizes the significant predictors:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Predictor | Coefficient | Std Err | t-statistic | p-value | Interpretation |
| const | 9.9618 | 0.028 | 357.861 | 0.000 | Baseline log(premium) |
| BMI | 0.0051 | 0.001 | 5.349 | 0.000 | ~0.5% increase per BMI unit |
| AnyTransplants | 0.2188 | 0.025 | 8.718 | 0.000 | ~24.5% increase with transplant |
| AnyChronicDiseases | 0.0914 | 0.015 | 6.169 | 0.000 | ~9.6% increase with disease |
| HistoryOfCancerInFamily | 0.0532 | 0.018 | 2.939 | 0.003 | ~5.5% increase with history |
| Age\_centered | 0.0143 | 0.000 | 35.430 | 0.000 | ~1.4% increase per centered age year |
| Age2\_centered | -0.0004 | 3.12e-05 | -12.799 | 0.000 | Negative quadratic effect |

Coefficients represent changes in log(PremiumPrice), approximately percentage changes, highlighting the strong impact of transplants and age-related factors.

**Validation Results**

Sample predictions versus actual PremiumPrice values on the test set are as follows:

* Sample 1: Predicted = 28,723.97, Actual = 31,000.00 (Error: ~7.3%)
* Sample 2: Predicted = 26,256.16, Actual = 31,000.00 (Error: ~15.3%)
* Sample 3: Predicted = 28,067.12, Actual = 29,000.00 (Error: ~3.2%)
* Sample 4: Predicted = 19,588.54, Actual = 15,000.00 (Error: ~30.6%)
* Sample 5: Predicted = 26,489.49, Actual = 23,000.00 (Error: ~15.2%)

The Mean Squared Error on the testing set, computed on the original scale, is 12,531,667.25, with a root mean squared error (RMSE) of approximately 3,539.67, representing 12–23% of the typical premium range (15,000–40,000).

**Diagnostics**

* **Multicollinearity**: Variance Inflation Factors (VIF) range from 1.00 to 1.03 (except const at 24.60), confirming no multicollinearity among predictors.
* **Condition Number**: 1.35e+03, indicating potential numerical issues but stable across iterations.
* **Heteroskedasticity**: The Breusch-Pagan test yielded an LM Statistic of 48.80 (p = 0.0000), confirming significant heteroskedasticity.
* **Residuals**: Omnibus (266.153, p = 0.000), Jarque-Bera (2416.695, p = 0.00), Skew (1.468), and Kurtosis (11.686) indicate non-normal residuals with heavy tails.

**Discussion**

**Interpretation**

The model effectively identifies key drivers of medical insurance premiums. The strong positive coefficient for AnyTransplants (0.2188) suggests a 24.5% premium increase for individuals with transplants, reflecting higher risk. Age\_centered (0.0143) and its quadratic term Age2\_centered (-0.0004) indicate a non-linear age effect, with premiums increasing with age but at a decreasing rate, consistent with risk profiles peaking in later years. BMI (0.0051) and AnyChronicDiseases (0.0914) contribute modestly, while HistoryOfCancerInFamily (0.0532) has a minor but significant effect.

The model’s R-squared of 0.696 demonstrates substantial explanatory power, aligning with the project’s goal of exceeding 60% variance explained. The consistent test set MSE (0.02) confirms generalizability, and the original-scale RMSE (3,539.67) provides a practical error metric for pricing decisions.

**Limitations**

Despite its strengths, the model exhibits significant heteroskedasticity (Breusch-Pagan p = 0.0000), indicating that prediction errors vary with the magnitude of PremiumPrice. This is visually supported by the fan-shaped residual plot from prior iterations. Non-normal residuals (Skew 1.468, Kurtosis 11.686) suggest the presence of outliers or unmodeled non-linearities, even after capping 30 outliers. The high original-scale MSE (12,531,667.25) reflects bias introduced by the log transformation, a known issue in log-linear models that could be mitigated with a bias correction (e.g., adding the mean residual).

**Future Work**

Future enhancements could include applying a bias correction to log predictions, exploring robust regression to handle heteroskedasticity, or incorporating interaction terms (e.g., Age × AnyTransplants) if data supports it. Cross-validation could further validate stability, though the current test set performance is promising.

**Conclusion**

This project successfully developed a predictive model for medical insurance premiums, explaining 69.6% of the variance in log(PremiumPrice) using a weighted least squares approach. The model identifies Age, AnyTransplants, and AnyChronicDiseases as key predictors, offering valuable insights for risk-based pricing. Validation on the test set, with an RMSE of 3,539.67 on the original scale, demonstrates practical applicability, though heteroskedasticity and residual non-normality suggest areas for refinement. The methodology provides a foundation for insurance industry applications and highlighting the importance of data-driven decision-making.