Pricing and Severity Modeling

Ryan Gomberg

Introduction & Goals

The purpose of this project is to explore primary determinants of auto insurance risk through multiple severity and frequency models. Modeling these features independently enables a more nuanced understanding of how policyholder features affect loss amounts. Then, by combining the two, we obtain a systematic premium model that uncovers the effect of discrete, rating variables on expected losses and informed pricing decisions.

To achieve this, a dataset of 5000 policyholders with 8 features will be simulated and decisively constructed to reflect patterns in policyholders and claim frequency/severity.

Required Packages

- MASS - fitting our GLM models
- actuar - loss simulation and severity modeling
- car - statistical diagnostics for GLMs
- tibble - generating the dataset

```r
knitr::opts_chunk$set(warning = FALSE, message = FALSE)
library(tibble)
library(dplyr)
library(ggplot2)
library(MASS)
library(actuar)
library(car)
set.seed(123)
```

Simulating Policy Data

We will generate a sample of 5000 policyholders, each containing the following features:

1. Age (between 18-90 years).
2. Vehicle Age (between 0-15 years).
3. Territory (T): Urban (U), Suburban (S), and Rural (R), with respective probabilities 0.5, 0.3, and 0.2.
4. Prior Claims: Follows a Poisson Distribution with mean 0.4, with the result rounded to the nearest integer. This was chosen to reflect most insurance claims: most policyholders have no prior claims, but a small population having at least 1.
5. Exposure (percentage of year in which the policy is active): Uniformly distributed on $(0.5, 1)$, making exposure periods between 6-12 months equally likely. This was chosen to simulate cases in which people purchase or cancel mid-year, and if policy periods start late.
6. Claim Count: Follows a Poisson distribution, the mean is a linear regression model:

   $$\lambda = \exp\left(-2 + 0.015 * age + 0.05 * vehicleage + 0.25 * priorclaims\right)$$

   ...rounded to the nearest integer. The exponent is used to keep $\lambda > 0$, a requirement for Poisson distributions, and because we will be using GLMs with log-links, exponentiating $\ln(\lambda)$ will transform our model back to its original scale. For instance, a 75-year old policyholder with a registered vehicle of 10 years and 2 prior claims would have a $e^{0.025} \approx 1.133 = 11.33\%$ increase in average claim count. We can expect a relatively small fraction of policyholders to not have 0 claims, since the argument inside the exponent must be at least 0, which is difficult to achieve unless the policyholder is older, has an older vehicle, and/or has a lot of prior claims. Without exponentiating, we would have a lot of negative and invalidated policyholders in our dataset.
7. Severity and Average Severity: We apply a lognormal distribution to generate the severity of each claim, where

   $$\mu = 5 + 0.02 * age + 0.05 * vehicleage + 0.15I(T = S) + 0.3I(T = R) + 0.4 * priorclaims$$

   $$\sigma = 0.45$$

   using urban as the base group for territory. The true mean of the severity is

$$E[S] = \exp\left(\mu + \frac{1}{2}\sigma^2\right) = \exp(\mu + 0.10125)$$

If we have a 75-year old policyholder, from a suburban region, with a registered vehicle of 10 years and 1 prior claim, then we expect their severity to be $e^{7.55} \approx \$1900.74$.

Average severity is taken as the ratio of severity per claim, or the expected dollar amount per claim. Since the number of claims is most likely to be 0 or 1, it is therefore likely for the average severity to equal the severity, but not impossible.

8. Claim Amount: The total dollar amount in claims for each policyholder. We simplify this calculation by making this quantity the product of average severity and claim count.

The dataset was constructed using a different .qmd file and exported as a .csv file to avoid complications when rendering.

```r
df <- read.csv("PricingSeverityModel.csv")

summary(df)
```

```
      age         vehicle_age      territory          prior_claims  
 Min.   :18.00   Min.   : 0.000   Length:5000        Min.   :0.000  
 1st Qu.:36.00   1st Qu.: 4.000   Class :character   1st Qu.:0.000  
 Median :54.00   Median : 7.000   Mode  :character   Median :0.000  
 Mean   :53.81   Mean   : 7.458                      Mean   :0.395  
 3rd Qu.:72.00   3rd Qu.:11.000                      3rd Qu.:1.000  
 Max.   :90.00   Max.   :15.000                      Max.   :5.000  
                                                                    
    exposure       claim_count        severity        avg_severity   
 Min.   :0.5001   Min.   :0.0000   Min.   :  104.6   Min.   : 104.6  
 1st Qu.:0.6280   1st Qu.:0.0000   1st Qu.:  594.4   1st Qu.: 490.5  
 Median :0.7497   Median :0.0000   Median :  936.6   Median : 810.1  
 Mean   :0.7520   Mean   :0.4184   Mean   : 1225.3   Mean   :1050.1  
 3rd Qu.:0.8813   3rd Qu.:1.0000   3rd Qu.: 1531.0   3rd Qu.:1326.4  
 Max.   :1.0000   Max.   :4.0000   Max.   :13054.0   Max.   :7352.0  
                                   NA's   :3339      NA's   :3339    
  claim_amount    
 Min.   :  104.6  
 1st Qu.:  594.4  
 Median :  936.6  
 Mean   : 1225.3  
 3rd Qu.: 1531.0  
 Max.   :13054.0  
 NA's   :3339
```

Some observations:

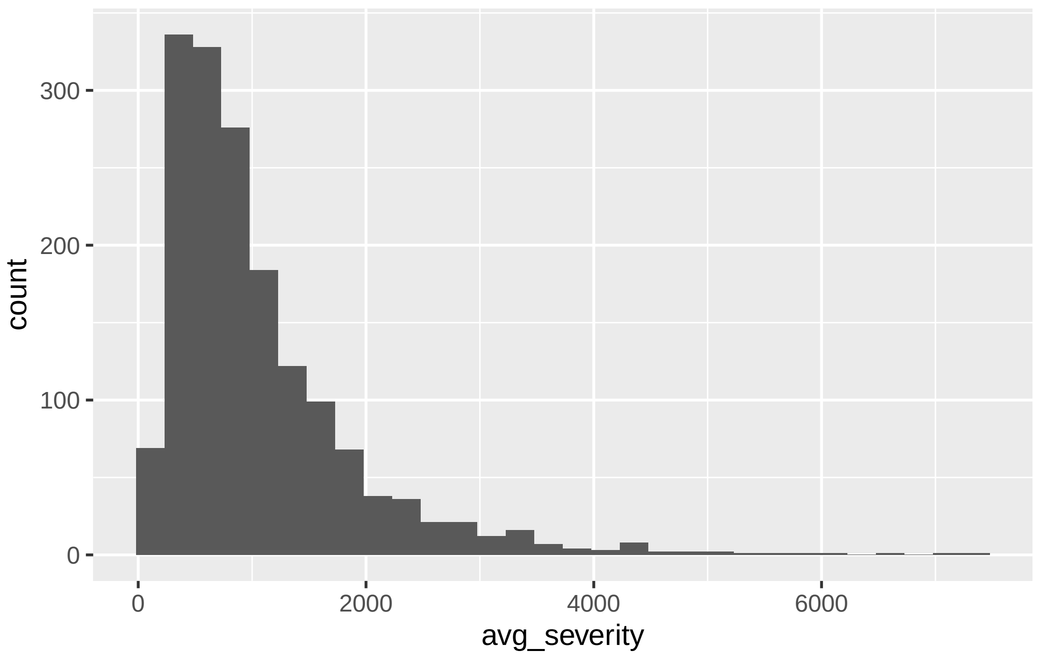
- The mean of the sample of prior_claims, exposure, and claim_count are unsurprisingly close to the mean of their distributions.
- Within the sample, the maximum amount of prior claims is 5. Since prior_claims was generated with a Poisson distribution of mean 0.4, we can find that

  $$P(priorclaims \leq 5) \approx 0.999996 = p.$$

  Therefore, the probability that all 5000 policyholders have at most 5 prior claims is $p^{5000} \approx 0.98$, meaning that there is only a 2% chance that the maximum amount of prior claims (for all 5000 policyholders) is at least 6. We could apply the same logic to the claim_count feature because its λ is small.
- The maximum amount of claim counts among all policyholders is 4, which is quite surprising. For this to happen, the argument inside the exponent must be at least 1.26, and the maximum possible value is 1.35. As indicated by the "NAs," 3,339 of the policyholders have no claims, and 1,661 have exactly at least 1. Therefore the maximum expected dollar amount per claim is equal to the claim amount and severity ($13,054.00).

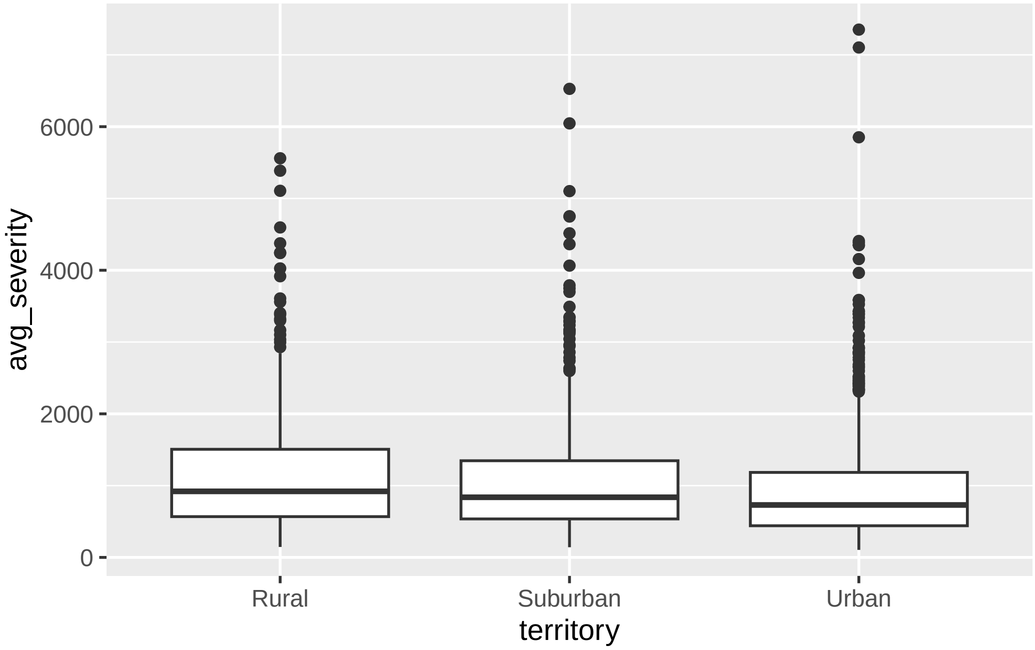
Exploratory Data Analysis

```r
# Severity density plot
ggplot(df, aes(x = avg_severity)) + geom_histogram(bins = 30)
```



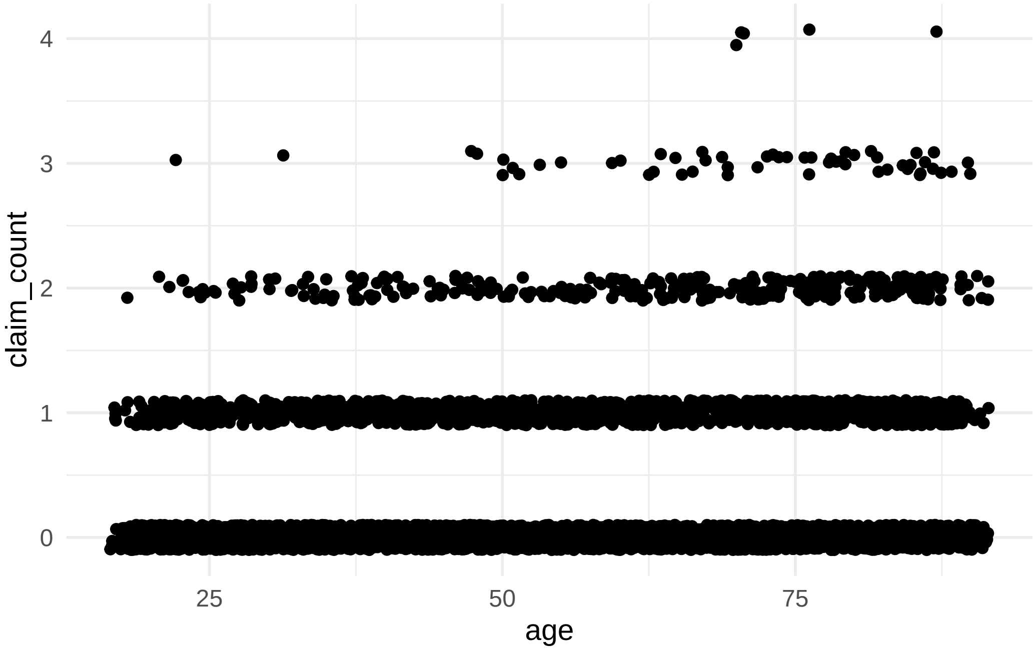
This loosely matches the lognormal distribution we are looking for. We can observe that the mode of ~325 occurs close to the first quartile (roughly 490), with an exponential-like decay after. Recall that the histogram is only accounting for policyholders with claims, leaving us with a sample of 1,661 policyholders.

```r
# Territory and severity box and whisker plot
ggplot(df, aes(x = territory, y = avg_severity)) + geom_boxplot()
```



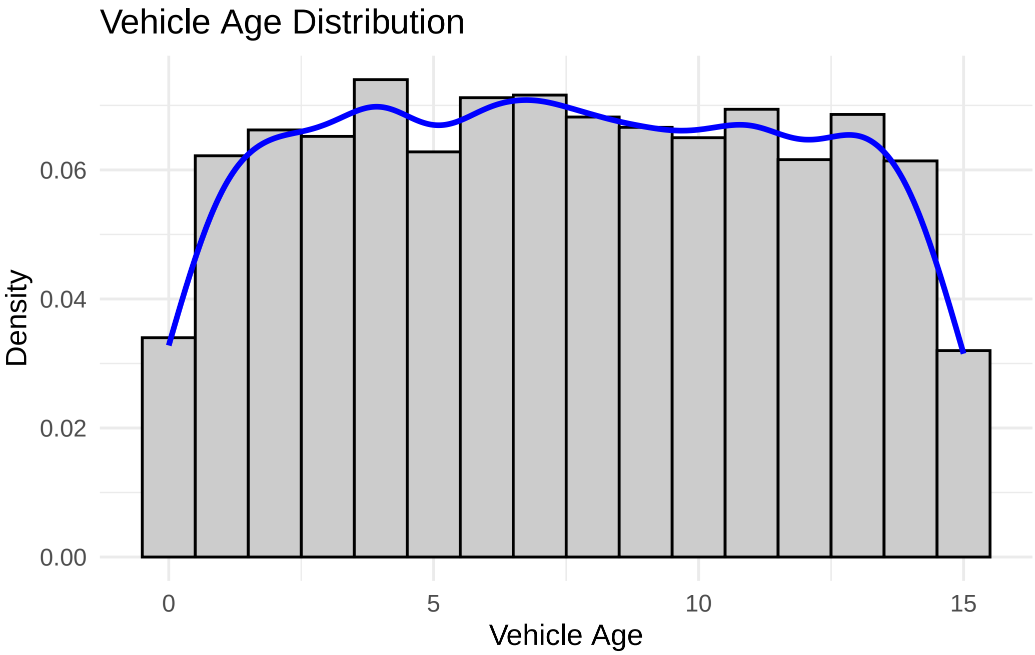
The median expected amount per claim is distributed in decreasing order for the Rural, Suburban, and Urban regions. Rural regions have outliers that significantly widen the overall range, likely due to the lower percentage of policyholders in this region. Therefore, the mean is pulled upward by extreme outliers in this territory. Lastly, it is worth noting that the median for all 3 areas are closer to the 25th percentile compared to the 75th percentile.

```
ggplot(df, aes(x = age, y = claim_count)) +
  geom_jitter(width = 1.5, height = 0.1) +
  theme_minimal()
```



- Age appears to be well distributed.
- The majority of policyholders have at most 2 claims, while the minority have at least 3.
- Given how we constructed claim_count, it is no surprise that older policyholders are more likely to have more claims. The impact is not large, however, given there are other covariates in the regression model for claim_count's λ.

```
# Vehicle age density plot
ggplot(df, aes(x = vehicle_age)) +
  geom_histogram(aes(y = ..density..), bins = 16, fill = "gray80", color = "black") +
  geom_density(color = "blue", size = 1) +
  labs(title = "Vehicle Age Distribution",
       x = "Vehicle Age",
       y = "Density") +
  theme_minimal()
```



As seen from the density curve, vehicle age is mostly well-distributed, with some minor fluctuations. The mode of vehicle age is at 4 years, and the most uncommon vehicle ages are 0 and 15 years. All vehicle ages have densities between 3% and 7.5%.

GLM and Regression Implementations

Frequency Models

There are four main objectives:

1. Fit a frequency model to a Poisson GLM and verify statistical significance among all features.
2. Check for overdispersion. If such overdispersion exists, try another model.
3. If no overdispersion exists, construct a QQ-plot to measure goodness of fit.
4. Depending on the QQ-plot, try another model.

For our Poisson model, we have that

$$\ln(\mu_i) = \ln(E_i) + \beta_0 + \beta_1(age_i) + \beta_2(vehicleage_i) + \beta_3(priorclaims_i) + \beta_4(T_i = S) + \beta_5(T_i = U)$$

where the offset term $\ln(E_i)$ is adjusting for differences in policy exposure. On top of developing a model, we apply a GLM ANOVA (likelihood ratio Chi-squared) to test whether all features are statistically significant as a group at the $\alpha = 0.05$ significance level.

- Null Hypothesis H_0: Removing all of the features leads to a better fit, or $\beta_i = 0$ for $i = 1, 2, 3, 4, 5$.
- Alternative Hypothesis H_a: Removing none of the features leads to a better fit, or $\beta_i \neq 0$ for at least one $i \in \{1, 2, 3, 4, 5\}$.

```r
# Poisson GLM Construction
freq_model <- glm(
  claim_count ~ age + vehicle_age + prior_claims + territory,
  offset = log(exposure),
  family = poisson,
  data = df
)
summary(freq_model)
```

```
Call:
glm(formula = claim_count ~ age + vehicle_age + prior_claims + 
    territory, family = poisson, data = df, offset = log(exposure))

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)    
(Intercept)       -1.827727   0.093299 -19.590  < 2e-16 ***
age                0.015423   0.001086  14.198  < 2e-16 ***
vehicle_age        0.044241   0.005144   8.600  < 2e-16 ***
prior_claims       0.236442   0.030868   7.660 1.86e-14 ***
territorySuburban -0.075415   0.061377  -1.229    0.219    
territoryUrban    -0.136705   0.055607  -2.458    0.014 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 4828.2  on 4999  degrees of freedom
Residual deviance: 4496.5  on 4994  degrees of freedom
AIC: 8079.2

Number of Fisher Scoring iterations: 6
```

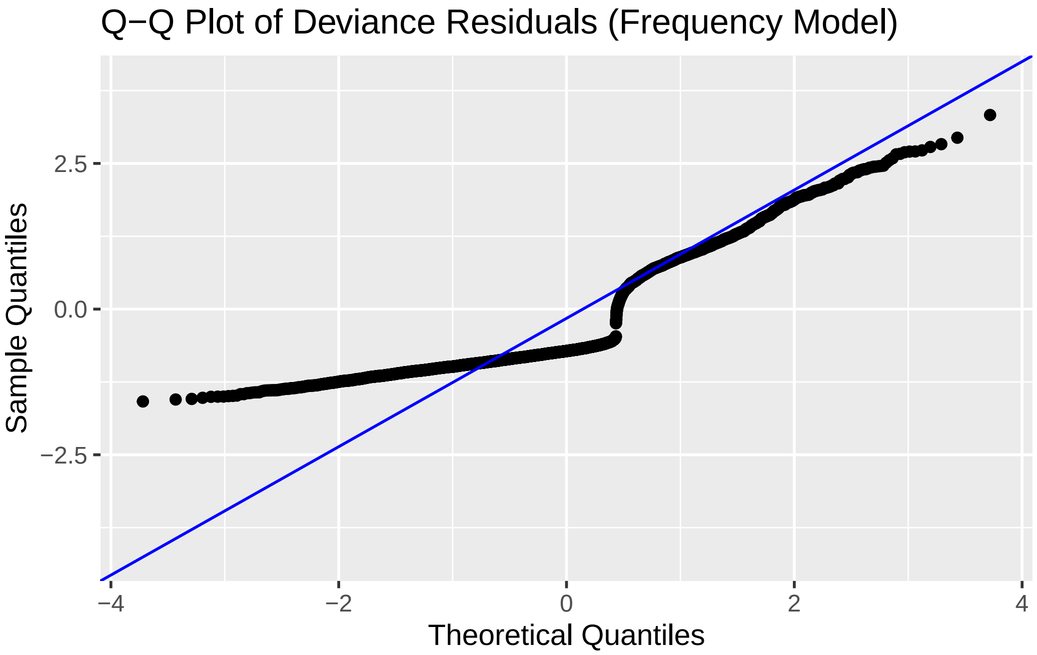
```r
# Testing for overdisperion
overdispersion = freq_model$deviance / freq_model$df.residual
overdispersion
```

```
[1] 0.9003834
```

```r
# Testing for statistical significance
anova(freq_model, test = "Chisq")
```

```
Analysis of Deviance Table

Model: poisson, link: log

Response: claim_count

Terms added sequentially (first to last)

             Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                          4999     4828.2
age           1  199.211      4998     4629.0 < 2.2e-16 ***
vehicle_age   1   71.741      4997     4557.3 < 2.2e-16 ***
prior_claims  1   54.629      4996     4502.6 1.456e-13 ***
territory     2    6.107      4994     4496.5   0.04719 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
library(ggplot2)

res_dev <- residuals(freq_model, type = "deviance")

ggqqplot <- ggplot(data.frame(res = res_dev), aes(sample = res)) +
  stat_qq() +
  stat_qq_line(color = "blue") +
  labs(title = "Q-Q Plot of Deviance Residuals (Frequency Model)",
       x = "Theoretical Quantiles",
       y = "Sample Quantiles")

ggqqplot
```



Although the p-value for the Suburban territory is greater than 0.05, given that the other features are in the model, the ANOVA test verifies that each coefficient is statistically significant, and therefore all features result in the best fit.

Some observations:

- The lowest expected claim frequency of 0.21222 is achieved when an 18-year old has no vehicle (or has bought one in the past year), has no prior claims, and lives in an Urban region. This makes sense, for an 18-year old probably has no registered vehicle.
- Living in an Urban or Suburban region decreases the claim frequency by 12.78% and 7.26%, respectively, compared to living in a Rural region. As a result, this dataset assumes rural drivers face higher frequency, which can be attributed to longer travel, wildlife, or harsher road/weather conditions.
- A 1 unit increase in prior claims multiplies the expected claim count by $e^{0.236442} \approx 1.2667$. That is to say, each past claim increases the frequency claim by 26.67%.
- A 10 unit increase in age and 3 unit increase in vehicle age will result in an expected change of $e^{0.015423(10)+0.044241(3)} \approx 1.3324$. That is to say that a policyholder that is 10 years older and whose vehicle is 3 years older will have an expected claim count that is 1.3324 times that of similar policyholders who have the original age and vehicle age, keeping all other factors constant.
- The deviance reduction of 331.7 indicate that the features in the model result in a better fit.

- An AIC of 8079.2 serves as a baseline if we need to consider other models. The poisson model has an overdispersion coefficient that is less than 1 (0.9), suggesting that the variance is handled well. However, this cannot be verified without looking at a QQ-plot.

On the upper-half, the QQ-plot for the Poisson model fits the pattern well and has low variance, which follows the 45-degree line we are looking for. However, the line appears to fall off as the claim count approaches 0 and then follows a line different from the desired 45-degree line. Not only does the QQ-plot weaken the nonexistent overdispersion claim, but more importantly, it suggests that our dataset has too many policyholders with zero claims compared to what the model predicts.

Next, we will try a negative binomial model and identify if the problem lies in the Poisson model or the dataset.

```r
freq_nb <- glm.nb(claim_count ~ age + vehicle_age + prior_claims + territory, data=df)
summary(freq_nb)
```

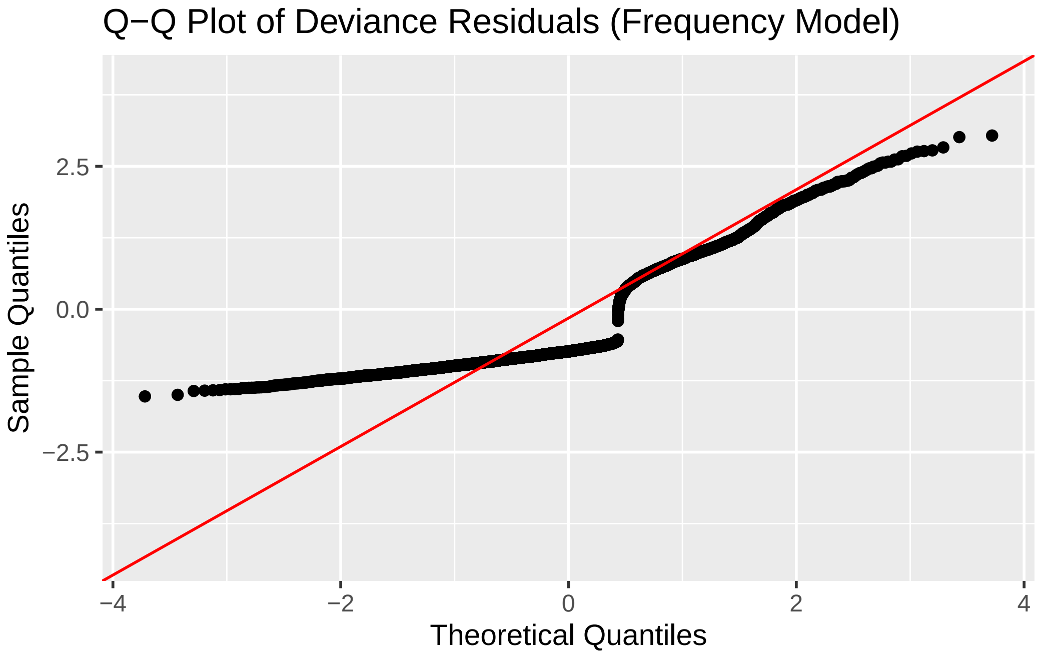
```

Call:
glm.nb(formula = claim_count ~ age + vehicle_age + prior_claims +
    territory, data = df, init.theta = 2064.336828, link = log)

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)       -2.114153   0.093345 -22.649  < 2e-16 ***
age                0.015475   0.001086  14.249  < 2e-16 ***
vehicle_age        0.044255   0.005143   8.605  < 2e-16 ***
prior_claims       0.232409   0.030816   7.542 4.63e-14 ***
territorySuburban -0.069896   0.061381  -1.139   0.2548
territoryUrban    -0.140188   0.055609  -2.521   0.0117 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(2064.337) family taken to be 1)

    Null deviance: 4910.9  on 4999  degrees of freedom
Residual deviance: 4580.6  on 4994  degrees of freedom
AIC: 8166.3

Number of Fisher Scoring iterations: 1

              Theta:  2064
          Std. Err.:  11364
```

```
Warning while fitting theta: iteration limit reached

 2 x log-likelihood:  -8152.295
```

```
overdispersion_nb = freq_nb$deviance / freq_nb$df.residual
overdispersion_nb
```

```
[1] 0.9172222
```

```
res_dev_nb <- residuals(freq_nb, type = "deviance")

ggqqplot_nb <- ggplot(data.frame(res = res_dev_nb), aes(sample = res)) +
  stat_qq() +
  stat_qq_line(color = "red") +
  labs(title = "Q-Q Plot of Deviance Residuals (Frequency Model)",
       x = "Theoretical Quantiles",
       y = "Sample Quantiles")

ggqqplot_nb
```



Both the Poisson and Negative Binomial fits the positive claim counts well, which is why all model features are statistically significant. However, the deviation in the QQ-plot is almost entirely concentrated on the lower tail, corresponding to the 66.78% of policyholders with no claims. All signs point to zero-inflation: the model greatly over- or under-predicts the probability of having no claims, but accurately models the claim frequency among policyholders with existing claims.

Severity Modeling

We test four different models: Gamma, Lognormal, Weibull, Exponential, and determine which is the best to model our average severity through AIC and QQ-plots.

```r
# Gamma GLM
sev_data <- df[df$claim_count > 0, ]
glm_gamma <- glm(avg_severity ~ age + territory + vehicle_age,
               family = Gamma(link = "log"),
               data = sev_data)

# Lognormal model
glm_lognorm <- lm(
  log(avg_severity) ~ age + territory + vehicle_age,
  data = sev_data
)


# Weibull model
library(survival)

glm_weibull <- survreg(
  Surv(avg_severity) ~ age + territory + vehicle_age,
  dist = "weibull",
  data = sev_data
)


# Exponential model
glm_exp <- survreg(
  Surv(avg_severity) ~ age + territory + vehicle_age,
  dist = "exponential",
  data = sev_data
)

# Displaying model and diagnostics
print(summary(glm_gamma))
```

```
Call:
glm(formula = avg_severity ~ age + territory + vehicle_age, family = Gamma(link = "log"),
    data = sev_data)

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        5.6437072  0.0671339  84.066  < 2e-16 ***
age                0.0178826  0.0008111  22.048  < 2e-16 ***
territorySuburban -0.0893531  0.0460661  -1.940   0.0526 .
territoryUrban    -0.2394535  0.0413002  -5.798 8.03e-09 ***
vehicle_age        0.0404697  0.0038390  10.542  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Gamma family taken to be 0.4390047)

    Null deviance: 867.99  on 1660  degrees of freedom
Residual deviance: 609.97  on 1656  degrees of freedom
AIC: 25405

Number of Fisher Scoring iterations: 5
```

```r
print(summary(glm_lognorm))
```

```
Call:
lm(formula = log(avg_severity) ~ age + territory + vehicle_age,
    data = sev_data)

Residuals:
     Min       1Q   Median       3Q      Max
-2.06995 -0.41439  0.01265  0.39940  2.25287

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        5.534764   0.061662  89.760  < 2e-16 ***
age                0.017105   0.000745  22.961  < 2e-16 ***
territorySuburban -0.125851   0.042311  -2.974  0.00298 **
territoryUrban    -0.270436   0.037934  -7.129  1.5e-12 ***
vehicle_age        0.040043   0.003526  11.356  < 2e-16 ***
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6086 on 1656 degrees of freedom
Multiple R-squared:  0.2913,    Adjusted R-squared:  0.2895
F-statistic: 170.1 on 4 and 1656 DF,  p-value: < 2.2e-16
```

```
AIC(glm_lognorm)
```

```
[1] 3070.845
```

```
print(summary(glm_weibull))
```

```


Call:
survreg(formula = Surv(avg_severity) ~ age + territory + vehicle_age,
    data = sev_data, dist = "weibull")
                      Value Std. Error      z       p
(Intercept)        5.735934   0.061019  94.00 < 2e-16
age                0.018326   0.000734  24.96 < 2e-16
territorySuburban -0.066304   0.041966  -1.58    0.11
territoryUrban    -0.220440   0.037597  -5.86 4.5e-09
vehicle_age        0.038682   0.003421  11.31 < 2e-16
Log(scale)        -0.506329   0.017454 -29.01 < 2e-16

Scale= 0.603

Weibull distribution
Loglik(model)= -12776.8   Loglik(intercept only)= -13069.8
    Chisq= 585.89 on 4 degrees of freedom, p= 1.8e-125
Number of Newton-Raphson Iterations: 5
n= 1661
```

```
AIC(glm_weibull)
```

```
[1] 25565.64
```

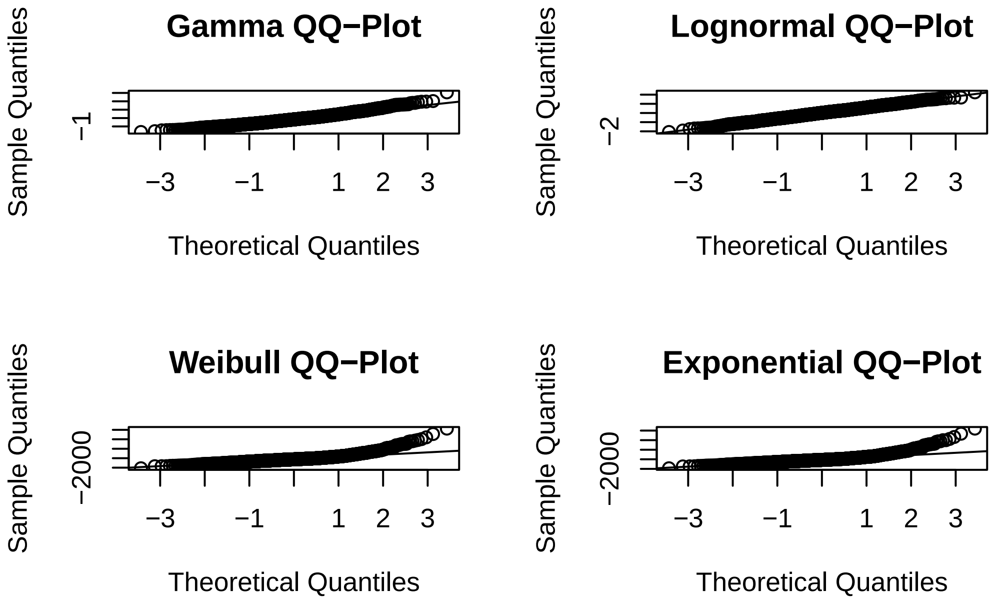
```
print(summary(glm_exp))
```

```
Call:
survreg(formula = Surv(avg_severity) ~ age + territory + vehicle_age,
    data = sev_data, dist = "exponential")
                     Value Std. Error     z       p
(Intercept)        5.64371    0.10095 55.91 < 2e-16
age                0.01788    0.00122 14.70 < 2e-16
territorySuburban -0.08935    0.06956 -1.28 0.19893
territoryUrban    -0.23945    0.06236 -3.84 0.00012
vehicle_age        0.04047    0.00575  7.03   2e-12

Scale fixed at 1

Exponential distribution
Loglik(model)= -13087   Loglik(intercept only)= -13216
    Chisq= 258.01 on 4 degrees of freedom, p= 1.2e-54
Number of Newton-Raphson Iterations: 4
n= 1661
```

```r
AIC(glm_exp)
```

```
[1] 26184.03
```

```r
par(mfrow = c(2, 2))

# Residuals and QQ-plot generation
qqnorm(residuals(glm_gamma),
      main = "Gamma QQ-Plot")
qqline(residuals(glm_gamma))

qqnorm(residuals(glm_lognorm),
       main = "Lognormal QQ-Plot")
qqline(residuals(glm_lognorm))

qqnorm(residuals(glm_weibull),
       main = "Weibull QQ-Plot")
qqline(residuals(glm_weibull))

qqnorm(residuals(glm_exp),
       main = "Exponential QQ-Plot")
qqline(residuals(glm_exp))
```



Note: Even though a lognormal model was used to generate the severity data, we can not only show that a lognormal regression is the correct model, but we can identify how other distributions fare in accuracy and reliability.

Given the right-skewed distribution of average severity values and multiplicative effect of covariates, a lognormal model is intuitively appropriate. Model diagnostics (AIC, QQ-plots) only reinforce this, for an AIC of 3070.85 is astoundingly low compared to the other models and the QQ-plot shows the best convergence to the 45-degree line.

Although lognormal regression offers the best fit given the lognormal nature of average severity in our dataset, the Gamma GLM comes up in second, albeit from pretty far, but it could still be an acceptable model. Its AIC does compare closer to Weibull and Exponential distributions; however, the QQ-plot does track the 45-degree line well, indicating that the overall shape is captured. The curve does lie slightly above the line with some tails on the upper and lower half, suggesting that the model overestimates the average severity. While the lognormal model is undoubtedly the one to pick, the Gamma GLM does pose as a reasonable fit as well, likely due to their similar shape.

The Weibull and Exponential distributions have higher AIC values compared to the other two, and their QQ-plot diverges on the upper-tail, making them very poor fits for our severity model.

Because our Gamma and Lognormal models are both good fits, we'll compare average severity amounts depending on different changes in our features.

Our Gamma (log-link) model states

$$\ln(\mu_i) = 5.6437 + 0.0179(age_i) + 0.0405(vehicleage_i) - 0.0894I(T = S) - 0.2395I(T = U)$$

Our Lognormal model states

$$\ln(Y_i) = 5.5348 + 0.0171(age_i) + 0.0492(vehicleage_i) - 0.1259I(T = S) - 0.2704I(T = U)$$

(1) Controlling for: Age 30 years, Vehicle Age of 5 years. Comparing over all 3 territories:

- **Lognormal**: Rural = $e^{6.2938} \approx$ \$541.21, Suburban = $e^{6.1679} \approx$ \$477.18, Urban = $e^{6.0234} \approx$ \$412.98.
- **Gamma GLM**: Rural = $e^{6.3832} \approx$ \$591.82, Suburban = $e^{6.2938} \approx$ \$541.21, Urban = $e^{6.1437} \approx$ \$465.77.

(2) Controlling for: Age 60 years, Vehicle Age of 10 years. Comparing over all 3 territories:

- **Lognormal**: Rural = $e^{7.0528} \approx$ \$1156.09, Suburban = $e^{6.9269} \approx$ \$1019.33, Urban = $e^{6.7824} \approx$ \$882.18.
- **Gamma GLM**: Rural = $e^{7.1227} \approx$ \$1239.79, Suburban = $e^{7.0333} \approx$ \$1133.77, Urban = $e^{6.8832} \approx$ \$975.74.

General observations:

1. Both higher age and vehicle age lead to higher severity, as proven by their coefficients.
2. Compared to Rural areas, living in a Suburban area reduces severity by 8.5-12% and living in an Urban further reduces the severity by 13-14% (or 21.5-26% compared to Rural). The effects on territory are moderate, but significant enough.
3. Gamma GLMs will more likely predict higher severities because it predicts the mean, average severity. Lognormal regressions predict the median, and the coefficient estimates are more conserved. The intercept and coefficients are all larger in magnitude for the Gamma model, meaning it will grow faster compared to the lognormal regression model.
4. In Sample 1, the Gamma GLM has severity estimates that are 1.09-1.13 times that of our lognormal regression model. In Sample 2, this estimate slightly to a factor of 1.07-1.11.
5. Despite this, the difference between lognormal regression and Gamma GLM gets larger as policyholder age and vehicle age increases. More precisely, when increasing policyholder age by 30 years and vehicle age by 5 years, this difference increases by 165.38% for Rural areas, 178.72% for Suburban areas, and 177.23% for Urban areas.

6. When increasing policyholder age by 30 years and vehicle age by 5 years, the average severity increases by 213.60% for lognormal models and 209.50% for our Gamma GLM, reinforcing that the lognormal regression model is less extreme in its approximation.

As for which model to use when making predicted average severity amounts, we should steer closer to the lognormal regression model. While we could take the given appproximations for the lognormal regression and be done, we can also exact a tighter approximation through an ensemble approach, where we use the lognormal model as the primary predictor and add a small contribution from the Gamma predictor to adjust for different biases in both models (lognormal tends to adjust for median and Gamma GLM tends to adjust for mean). If we let μ_γ measure the average severity of the Gamma GLM and μ_L measure the average severity of the lognormal regression model, we could calibrate a convex system

$$\mu = w\mu_\gamma + (1-w)\mu_L$$

and find a $w \in [0,1]$ that minimizes the residual mean squared error on policyholders with claims.

Effect of Both Models

Now that we've analyzed both frequency and severity models separately, we should think about both archetypes on a broader perspective. That is, are there patterns that result from piecing together both models?

- Regarding performance and fit, the negative binomial model technically beats out our Poisson model; our diagnostics suggest that the difference in their overall performance is very small. As for modeling severity, both lognormal and Gamma provide reasonably good fits, while Weibull and exponential are objectively weaker choices.
- Across both model types, the direction of associations is consistent:
  - Claim frequency and severity both increase as the policyholder's age increases, holding all other factors constant. Older policyholders file slightly more claims which generally cost more, producing a multiplicative effect on the total incurred loss, thereby making age a strong predictive feature.
  - Claim frequency and severity both increase as the policyholder's vehicle age increases, holding all other factors constant. Older vehicles are prone to more mechanical and technical failures, which can lead to more claims. Moreover, it may be harder to repair and/or replace parts on an older vehicle, making the total loss larger. The net impact on the total incurred loss is also multiplicative, which makes vehicle age a powerful risk factor.

  - Compared to rural regions, claim frequency decreases for urban and suburban regions (Urban < Suburban < Rural) and average severity decreases for urban and suburban regions (Urban < Suburban < Rural). This reflects a pattern where urban claims, for instance, are less frequent and typically lower-cost, but rural claims are more frequent and more severe (weather/road conditions, high speeds).

- The combined effect on the total premium varies on condition. The total premium is the product of frequency and severity.

  - Age: Increases total premium as policyholders get older (frequency and severity increase).

  - Vehicle age: Increases total premium for older vehicles (frequency and severity increase).

  - Territory: Net effect depends on the magnitudes of both severity and frequency. Total premiums are more likely to increase in Rural areas due to its high frequency and cost. Suburban and Urban areas are likely to have a smaller change in comparison because it has less claim frequency and lower severity.

  - Since claim frequency and severity move in the same direction, these effects are compounded, and we have

    $$premium_U < premium_S < premium_R$$

    . This is, in fact, **what is shown in the Exploratory Data Analysis.**

    * Urban: Fewer accidents and lower speeds = lowest expected total premium.

    * Suburban: Moderate accident frequency and severity = moderate total premium.

    * Rural: Highest accident frequency (higher speeds, wildlife, harsher weather/road conditions) and has the highest severity (given these conditions) = highest total premium.

- The features are therefore consistent among frequency and severity.

Conclusion

Final Statement

The frequency and severity models together describe a consistent and intuitive understanding of how risk varies among policyholders. The frequency model implies that, while holding other factors constant, rural policyholders exhibit the highest claim frequency, followed behind by suburban, then urban drivers having the lowest. The severity model follows the same pattern:

rural policyholders have the most severe losses, suburban are moderately lower, and urban areas experience the smallest claim severity. These results reflect situations in which rural areas are prone to harsher conditions and higher driving speeds compared to urban and suburban areas.

Because both models move along the same direction across all 3 territories, their combined effect strengthens the total premium. Therefore, urban policyholders have the lowest expected loss, suburban policyholders fall in the middle, and rural policyholders have the largest total premium. This compounding effect on expected, aggregate loss is largely influenced by the policyholder's age and vehicle age. To conclude, the combined modeling offers a mostly representative structure for pricing, with the given features reinforcing the risk patterns observed with this set of policyholders.

Improvements

While this project was primarily meant as a way of exposing myself to actuarial work and not to perfectly mimic it, the limitations and areas of improvement are clear. Here, I provide some of said limitations:

- Lack of real datasets: For obvious reasons, companies must keep their data on their policyholders confidential and only be used within company ground. Therefore, there aren't really many datasets that would include all of this information without it being anonymous and/or generated. If time allowed, I would have liked to look for a dataset that used information from real policyholders.
- Generated data and biases: To piggyback off the first reason, I had to generate a random sample as the dataset and find ways to "randomize" the range of each feature.
    - Purely randomizing age and vehicle age is fine, but I would have liked to construct a formula that shows a trend of slightly more policyholders having newer vehicles.
    - I spent too much time simulating claim count to make it "reasonable" in a real-world context. In the end, I settled with using exponential growth for the mean of a Poisson distribution to show that a high amount of claims are extremely unlikely. There are definitely better ways to simulate this, but for this project, I'd say it captures the probability of having x claims to the extent where it exhibits the desired pattern and does not appear random.
    - Making the severity lognormally distributed was a clear instance of bias and evidently came through when modeling it. This would not be the case in a real-world dataset, but a lognormal regression model would be very reliable by construction. However, I did not want to make these values random, nor did I want to use a lognormal distribution and call it a day. So, I use a lognormal distribution for the severity and measured how other potential models would compare.

- Once again, if time allowed, I would have liked to try the convex system for a refined average severity model, but I'm unfamiliar with optimization practices in R.