Homework 7 - Predictive Modeling in Finance and Insurance

Dennis Goldenberg

2024-03-25

1. Generalized Linear Model

a. Showing pdf belongs to exponential family

I can modify the p.d.f and show:

$$\begin{aligned} f(y;\theta) = \theta y^{-\theta-1} = \theta e^{\ln(y^{-\theta-1})} &= \theta e^{-\ln(y)*\theta-\ln(y)} \\ &= e^{\ln(\theta)} * e^{-\ln(y)*\theta-\ln(y)} \\ &= e^{-\ln(y)\theta+\ln(\theta)-\ln(y)} \end{aligned}$$

Thus, we have that $a(y) = -\ln(y)$, $b(\theta) = \theta$, $c(\theta) = \ln(\theta)$, $d(y) = -\ln(y)$.

b. Natural Exponential family or Exponential Dispersion Family

Note that $a(y) \neq y$; thus the Pareto distribution is NOT part of the natural exponential family. However, if you restructure the p.d.f:

$$f(y;\theta) = e^{-\ln(y)\theta+\ln(\theta)-\ln(y)} = e^{-(\theta+1)\ln(y)+(1)\ln(\theta)}$$

You can note that the Pareto distribution is also part of the Exponential dispersion family, with $\lambda = 1$.

c. Score statistic

To find the score statistic, I need log-likelihood function:

$$\ell(\theta;y) = \ln(f(y;\theta)) = -\ln(y)*\theta + \ln(\theta) - \ln(y)$$

The score function is just the derivative of the log-likelihood:

$$S(\theta;y) = \frac{d}{d\theta}\ell(\theta;y) = -\ln(y) + \frac{1}{\theta}$$

As this Pareto is in the EDF, $\mathbb{E}[S(\theta;y)] = 0$. The Fischer information is simply the expectation of the second derivative:

$$i(\theta) = -\mathbb{E}\left[\frac{d^2}{d\theta^2}\ell(\theta;y)\right] = -\mathbb{E}\left[-\frac{1}{\theta^2}\right] = \frac{1}{\theta^2}$$

d. Pareto distribution and Exponential Family

i. Does it belong to the exponential family?

Y_i **does belong** to the exponential family. It has the Pareto distribution with $\alpha = 1$, which we showed in (a) belongs to the Exponential family.

ii. Is it a GLM?

This model is **NOT** a generalized linear model. The model is NOT linear in its parameters, so the systematic component does not match the $x^T\beta$ linear format.

2. Logistic Regression

Given the model, I have $\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + 1(\beta_{>1}) + 1(\beta_{[0\%,10\%]})$. Therefore:

$$\begin{aligned}
&\log\left(\frac{\pi}{1-\pi}\right) = 1.53 + 0.735 - 0.031 = 2.234 \\
&\rightarrow \frac{\pi}{1-\pi} = e^{2.234} \\
&\rightarrow \pi = (1-\pi)e^{2.234} \\
&\rightarrow \hat{\pi} = \frac{e^{2.234}}{1+e^{2.234}} = \mathbf{0.9033}
\end{aligned}$$

3. Logistic Regression Part 2

a. Coefficients from Models

I generate the data by creating a data frame, and then run the model:

```
anther <- c(0,0,0,0,0,0,1,1,1,1,1,1)
treatment <- c(0,0,0,1,1,1,0,0,0,1,1,1)
freq <- c(102-55,99-52,108-57,76-55,81-50,90-50,
          55,52,57,55,50,50)
force <- c(40,150,350,40,150,350,40,150,350,40,150,350)
dataEmb <- data.frame(cbind(anther, treatment, force, freq))
dataEmb$treatment <- factor(dataEmb$treatment)

# for Model 1
model1 <- glm('anther ~ treatment + force + treatment*force',
              data = dataEmb, family = "binomial"(link = 'logit'),
              weights = freq)
coef1 <- summary(model1)$coefficients[,1]

# for Model 2
model2 <- glm('anther ~ treatment + force',
              data = dataEmb, family = "binomial"(link = 'logit'),
              weights = freq)
coef2 <- summary(model2)$coefficients[,1]

# for Model 3
model3 <- glm('anther ~ force',
              data = dataEmb, family = "binomial"(link = 'logit'),
              weights = freq)
coef3 <- summary(model3)$coefficients[,1]
print("Model 1 Coefficients:")
```

```
## [1] "Model 1 Coefficients:"
```

```
print(coef1)
```

```
##      (Intercept)       treatment1            force treatment1:force
##     0.1456719125     0.7963143307    -0.0001227259    -0.0020493450
```

```
print("Model 2 Coefficients:")
```

```
## [1] "Model 2 Coefficients:"
```

```
print(coef2)
```

```
##   (Intercept)    treatment1         force
##  0.3066430701  0.4055543471 -0.0009970257
```

```
print("Model 3 Coefficients:")
```

```
## [1] "Model 3 Coefficients:"
```

```
print(coef3)
```

```
##   (Intercept)         force
##  0.4759286430 -0.0009553572
```

b. Probability Estimates from models

To calculate the probability estimates, I note that:

$$\hat{\pi} = \frac{e^{x^T\hat{\beta}}}{1 + e^{x^T\hat{\beta}}}$$

From this, I calculate the probability estimates from the models

```
# for model 1
x_s = dataEmb[1:6,2:3]
x_s$inter <- x_s[,2]*c(0,0,0,1,1,1)
desMat <- data.matrix(cbind(1,x_s))
exTB1 <-exp(desMat %*% matrix(coef1))
pi_1 <- exTB1/(1 + exTB1)
pi_1 <- round(c(pi_1),4)

#for model 2
x_s2 = dataEmb[1:6,2:3]
desMat2 <- data.matrix(cbind(1,x_s2))
exTB2 <-exp(desMat2 %*% matrix(coef2))
pi_2 <- exTB2/(1 + exTB2)
pi_2 <- round(c(pi_2),4)

#for model 3
x_s3 = dataEmb[1:3,3]
desMat3 <- data.matrix(cbind(1,x_s3))
exTB3 <-exp(desMat3 %*% matrix(coef3))
pi_3 <- exTB3/(1 + exTB3)
pi_3 <- round(c(pi_3, pi_3),4)
```

I then create a matrix to show the $\hat{\pi}|\text{force}, \text{treatment}$ under the 6 different possible conditions, where control $= C$, treatment $= T$:

```
pred_pis <- t(cbind(pi_1, pi_2,pi_3))
colnames(pred_pis) <- c("40, C","150, C","350, C",
                        "40, T", "150,T", "350, T")
rownames(pred_pis) <- c("Model 1", "Model 2", "Model 3")
pred_pis
```

```
##          40, C 150, C 350, C  40, T  150,T 350, T
## Model 1 0.7185 0.7158 0.7108 0.8391 0.8042 0.7267
## Model 2 0.6620 0.6371 0.5898 0.7461 0.7248 0.6833
## Model 3 0.6077 0.5824 0.5353 0.6077 0.5824 0.5353
```

c. Expeced Values from 3 Models

```
data.matrix(pred_pis) * c(102,99,108,76,81,90)
```

```
##           40, C  150, C  350, C   40, T   150,T  350, T
## Model 1 73.2870 54.4008 72.5016 63.7716 82.0284 55.2292
## Model 2 65.5380 51.6051 58.3902 60.4341 71.7552 55.3473
## Model 3 65.6316 52.4160 57.8124 54.6930 62.8992 48.1770
```

d. Pearson residuals

e. Goodness of fit statistics

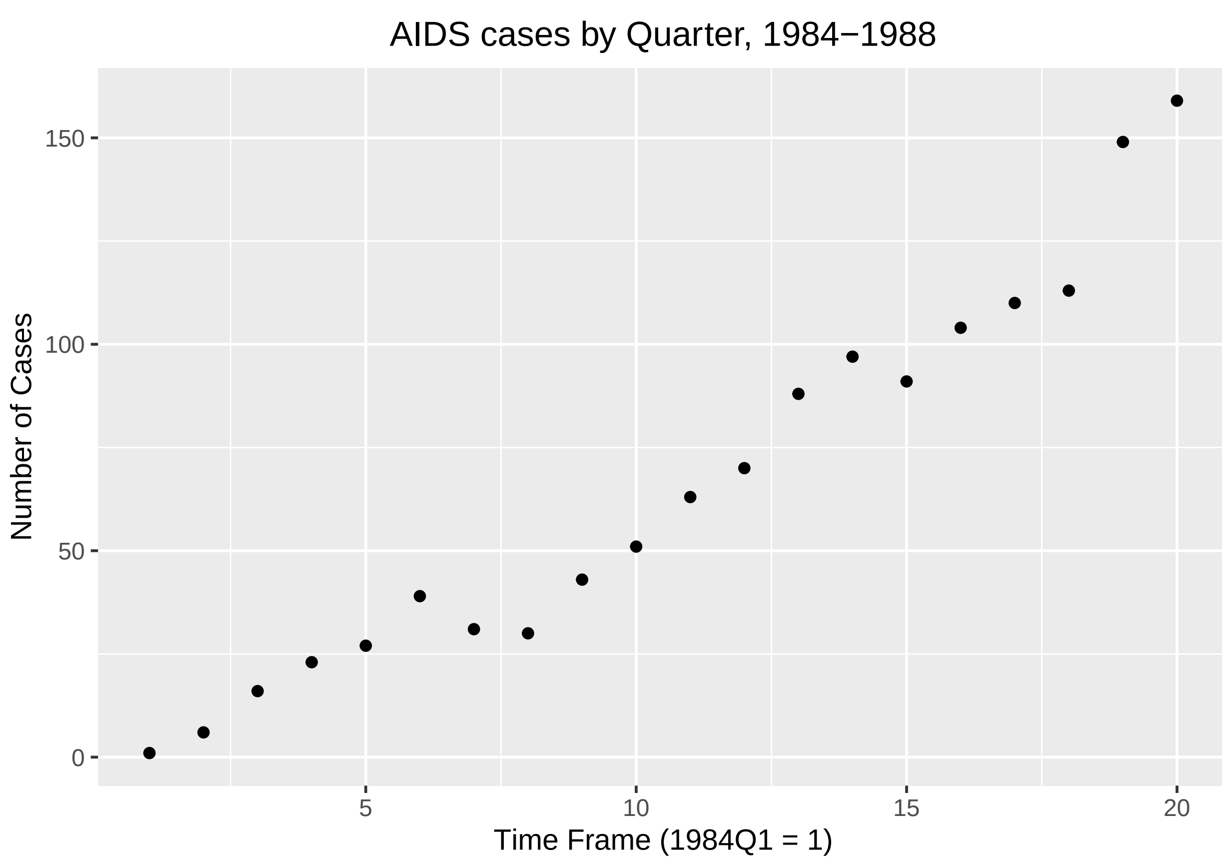
4. Maximum Likelihood Estimation approximation

```
library(ggplot2)
```

```
Year <- c(1984, 1985, 1986, 1987, 1988)
Q1 <- c(1,27,43,88,110)
Q2 <- c(6, 39, 51, 97, 113)
Q3 <- c(16, 31, 63, 91, 149)
Q4 <- c(23, 30, 70, 104, 159)
dataAIDS <- data.frame(cbind(Year, Q1,Q2,Q3,Q4))
```

a. Plot of number of cases against time period

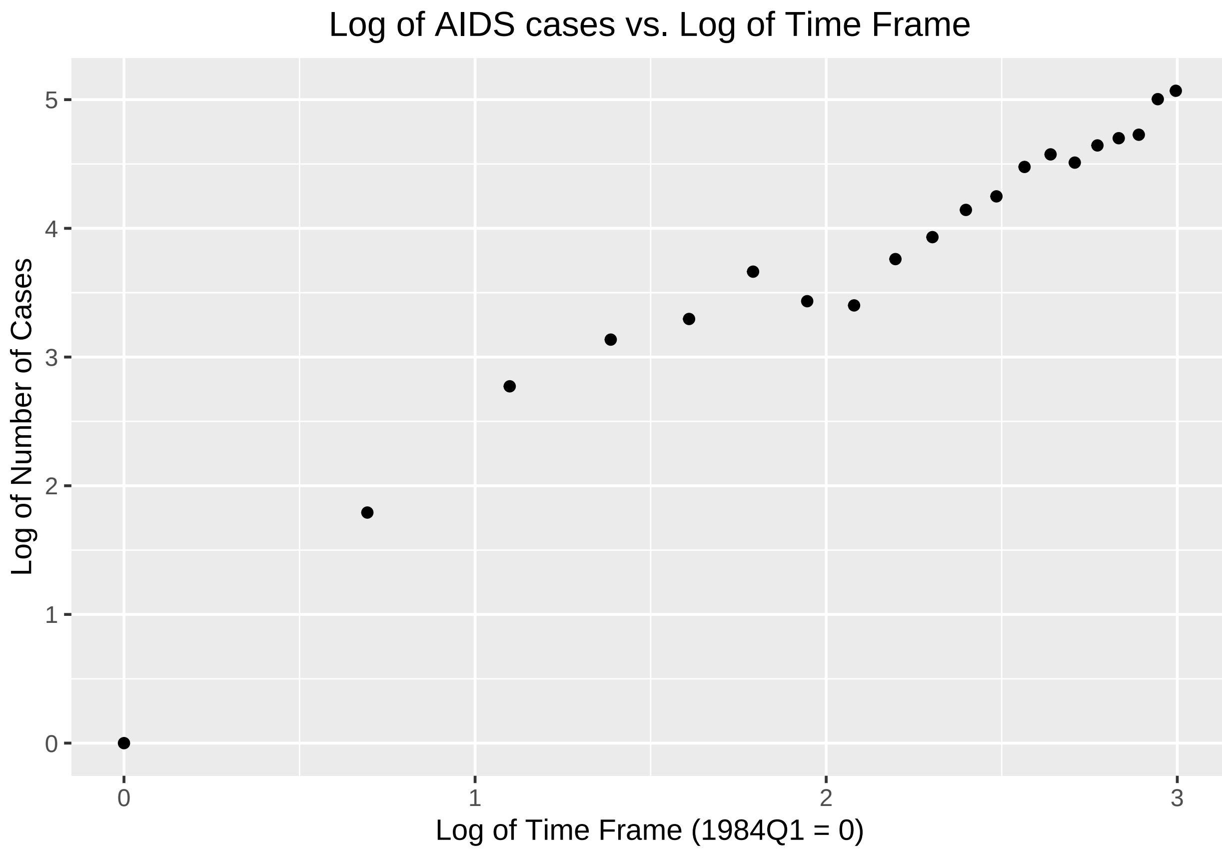
```
qMat <- t(data.matrix(dataAIDS[,2:5]))
timeFrame <- data.frame(cbind(1:(dim(qMat)[1] * dim(qMat)[2]),as.vector(qMat)))
ggplot(data = timeFrame) + geom_point(aes(X1,X2)) +
  xlab("Time Frame (1984Q1 = 1)") + ylab ("Number of Cases") +
  ggtitle("AIDS cases by Quarter, 1984-1988") +
  theme(plot.title = element_text(hjust = 0.5))
```



b. Plot against log i

```
timeFrameLog <- data.frame(cbind(log(timeFrame$X1),log(timeFrame$X2)))
ggplot(data = timeFrameLog) + geom_point(aes(X1,X2)) +
  xlab("Log of Time Frame (1984Q1 = 0)") + ylab ("Log of Number of Cases") +
```

```r
ggtitle("Log of AIDS cases vs. Log of Time Frame") +
theme(plot.title = element_text(hjust = 0.5))
```



There seems to be a linear relationship between $\log(y_i)$ and $\log(i)$, suggesting that this model may be appropriate.

c. Fitting GLM

Note that, here, we are using the canonical link function for the poisson distribution. Therefore:

$$g'(\mu)V(\mu) = 1 \rightarrow \mathbf{W} = \text{diag}\{1\} \text{ and } \theta = \lambda = g^{-1}(x^T\beta) = e^{x^T\beta}$$

I find the log-likelihood and Score function of β:

$$L(\lambda; y) = \prod_{i=1}^{n} \frac{e^{-\lambda_i} * \lambda_i^{y_i}}{y_i!} = e^{-\sum_{i=1}^{n}\lambda_i} * \frac{\prod_{i=1}^{n} \lambda_i^{y_i}}{\prod_{i=1}^{y_i}}$$

$$\rightarrow \ell(\lambda; y) = -\sum_{i=1}^{n} \lambda_i + \sum_{i=1}^{n} y_i \ln(\lambda_i) - \sum_{i=1}^{n} \ln(y_i!)$$

$$\rightarrow \ell(\beta; y) = -\sum_{i=1}^{n} e^{x_i^T\beta} + \sum_{i=1}^{n} y_i(x_i^T\beta) - \sum_{i=1}^{n} \ln(y_i!)$$

$$\rightarrow S(\beta; y) = \begin{bmatrix} \frac{d\ell}{d\beta_0} \\ \frac{d\ell}{d\beta_1} \end{bmatrix} = \begin{bmatrix} -\sum_{i=1}^{n} e^{x_i^T\beta} + \sum_{i=1}^{n} y_i \\ -\sum_{i=1}^{n} x_i e^{x_i^T\beta} + \sum_{i=1}^{n} x_i y_i \end{bmatrix}$$

Further, from the slides, again from the canonical link:

$$i(\beta, y) = x^T \text{diag}\{V(\mu_i)\} x = x^T \text{diag}\left\{e^{x_i^T\beta}\right\} x$$

Initial Step

Before beginning, it is important to note:

```
sum(timeFrame$X2)
```

```
## [1] 1311
```

```
sum(timeFrameLog$X1 * timeFrame$X2)
```

```
## [1] 3396.379
```

So, $\sum_{i=1}^{20} \log(i) = 1311$ and $\sum_{i=1}^{20} y_i \log(i) = 3396.379$. Therefore:

$$S(\beta; y) = \begin{bmatrix} -\sum_{i=1}^{n} e^{x_i^T \beta} + 1311 \\ -\sum_{i=1}^{n} x_i e^{x_i^T \beta} + 3396.379 \end{bmatrix}$$

Now for the initial step. I do some preliminary calculations:

```
e_xi <- sum(exp(timeFrameLog$X1))
x_ie_exi <- sum(exp(timeFrameLog$X1) %*% timeFrameLog$X1)
e_xi
```

```
## [1] 210
```

```
x_ie_exi
```

```
## [1] 529.6022
```

```
i_0 <- t(data.matrix(timeFrameLog)) %*% diag(exp(timeFrameLog$X1)) %*%
  data.matrix(timeFrameLog)
i_0
```

```
##          X1       X2
## X1 1386.477 2362.725
## X2 2362.725 4037.512
```

- I start with $\beta^0 = \begin{bmatrix} 0 & 1 \end{bmatrix}^T$.
- $S_0(\beta^0; y) = \begin{bmatrix} -\sum_{i=1}^{20} e^{x_i} + 1311 \\ -\sum_{i=1}^{n} x_i e^{x_i} + 3396.379 \end{bmatrix} = \begin{bmatrix} -210 + 1311 \\ -529 + 3396.379 \end{bmatrix} = \begin{bmatrix} 1101 \\ 2866.7768 \end{bmatrix}$.
- $i_0(\beta^0) = x^T \text{diag}\{e^{x_i}\} x = \begin{bmatrix} 1386.477 & 2362.725 \\ 2362.725 & 4037.512 \end{bmatrix}$
- $W = \text{diag}\{1\}$ as previously mentioned, and

I iterate once for step 2:

```
invI_0 <- solve(i_0)
beta_1 <- c(0,1) + invI_0%*%c(-1 * e_xi + sum(timeFrame$X2),
                           -1 * x_ie_exi + sum(timeFrameLog$X1 * timeFrame$X2))
rownames(beta_1) <- c("b_0","b_1")
beta_1
```

```
##           [,1]
## b_0 -150.68715
## b_1   89.89113
```

```
S_11 <- -1 *sum(exp(data.matrix(timeFrameLog)%*% beta_1)) + sum(timeFrame$X2)
S_11
```

```
## [1] -2.20635e+36
```

- $\beta^1 = \begin{bmatrix} -150.687 & 89.891 \end{bmatrix}$.
-

5. Deviance

a. Calculating Deviance

First, I calculate $\widehat{\ln(\theta)} = x^T \beta$:

```
design = data.matrix(cbind(c(1,1,1,1), c(1,0,1,1),
               c(2.5,1.5,2.9,3.0),
               c(6.25, 2.25, 8.41, 9),
               c(1,1,1,0)))
betaHat <- c(2.99, -0.27, -0.67, 0.16, 0.91)
hatLnTheta <- design %*% betaHat
hatLnTheta
```

```
##        [,1]
## [1,] 2.9550
## [2,] 3.2550
## [3,] 3.0326
## [4,] 2.1500
```

Therefore, $\mathbb{E}[Y] = \theta = e^{\ln(\theta)}$

```
predMean <- exp(hatLnTheta)
predMean
```

```
##           [,1]
## [1,] 19.201723
## [2,] 25.919615
## [3,] 20.751115
## [4,]  8.584858
```

Note that unit deviance is:

$$d(y;\theta) = 2 \max_{\mu} \ell(\mu; y) - 2\ell(\mu; y)$$

I derive $\hat{\mu}_{\text{M.L.E}}$ given an observed value y:

$$\begin{aligned}
L(\theta; y) = \frac{e^{-\frac{y}{\theta}}}{\theta} &\rightarrow \ell(\theta; y) = -\frac{y}{\theta} - \ln(\theta) \\
&\rightarrow \ell'(\theta; y) = \frac{y}{\theta^2} - \frac{1}{\theta} = 0 \\
&\rightarrow \theta^2 \left(\frac{y}{\theta^2} - \frac{1}{\theta} \right) = 0 * \theta^2 \\
&\rightarrow y - \theta = 0 \\
&\rightarrow \hat{\theta}_{\text{M.L.E}} = y
\end{aligned}$$

Therefore, in this case:

$$d(y;\hat{\theta}) = 2\left(\ell(y|y) - \ell(\hat{\theta}|y)\right) = 2\left(-1 - \ln(y) - \left(-\frac{y}{\hat{\theta}} - \ln(\hat{\theta})\right)\right)$$

Using this formula, I calculate the unit deviance for each observation, and then sum them up:

```
ys <- c(15,85,10,40)
unitD <- 2 * (-1 - log(ys) + (ys/predMean) + log(predMean))
Deviance <- sum(unitD)
sprintf("Deviance: %.4f", Deviance)
```

```
## [1] "Deviance: 6.9045"
```

b. Calculating Pearson residual

The formula for the pearson residual is $r_i = \frac{y_i - \hat{y_i}}{\sqrt{Var[\hat{y_i}]}} = \frac{y_i - \hat{y_i}}{\hat{y_i}}$. So, I calculate to get

```
pearson_2 <- (ys[2] - predMean[2])/predMean[2]
sprintf("Pearson Residual: %.5f",pearson_2)
```

```
## [1] "Pearson Residual: 2.27937"
```

c. Calculating deviance residual

The formula for the second deviance residual is $\text{sign}(y_i - \hat{y_i})\sqrt{d_i}$. So, I calculate:

```
dev2 <- sign(ys[2] - predMean[2])*sqrt(unitD[2])
sprintf("Deviance Residual: %.5f", dev2)
```

```
## [1] "Deviance Residual: 1.47765"
```

6. Nominal Regression

a. Odds ratio

All else equal, the odds ratio of females is just $e^{\beta_{\text{Female, Van}}} = e^{-.18} = \mathbf{0.8353}$.

b. Probability Calculation

In this case, $\text{Female}_i = 0$ and $I_{\text{age} < 25} = 1$ and $I_{\text{age} > 45} = 0$. Therefore, from slide 12 in the Nominal and Ordinal regression slides, I calculate:

$$\hat{\pi}_{\text{SUV}} = \frac{e^{x_{\text{SUV}}^T \hat{\beta}_{\text{SUV}}}}{1 + \sum_{j=2}^{J} e^{x_{\text{J}}^T \hat{\beta}_{\text{SUV}}}} = \frac{e^{x_{\text{SUV}}^T \hat{\beta}_{\text{SUV}}}}{1 + e^{x_{\text{SUV}}^T \hat{\beta}_{\text{SUV}}} + e^{x_{\text{Van}}^T \hat{\beta}_{\text{Van}}}} = \frac{e^{0.18}}{1 + e^{0.18} + e^{-.11}} = \mathbf{0.387}$$

7. Ordinal Regression

As medium risk is the second category, I use the formula in slide 32 from the Nominal and ordinal regression slides:

$$\begin{aligned}
\hat{\pi}_2 &= \frac{e^{\hat{\eta}_2}}{1 + e^{\hat{\eta}_2}} - \frac{e^{\hat{\eta}_1}}{1 + e^{\hat{\eta}_1}} \\
&= \frac{e^{2.05-(-0.12)}}{1 + e^{2.05-(-0.12)}} - \frac{e^{1.30-0.23}}{1 + e^{1.30-0.23}} \\
&= \frac{e^{2.17}}{1 + e^{2.17}} - \frac{e^{1.07}}{1 + e^{1.07}} \\
&= \mathbf{0.153}
\end{aligned}$$