Loglinear Regression for Rates

Grinnell College

November 11, 2026

Counts Need Context

Two road segments recorded crashes over the past year.

	Road A	Road B
Crashes	20	10
Million vehicle-miles	10	2
Crashes per million vehicle-miles	2	5

Which road has more crashes? Which has the higher crash rate?

A count combines the underlying rate with the amount of exposure.

From Counts to Rates

Recall that a Poisson distribution describes a distribution of counts over an interval at a constant rate, with

$$Y \sim \text{Poisson}(\mu)$$

The expected count, μ, can be expressed as a rate over a period of exposure

$$\mu = \lambda t$$

Equivalently, our rate can be expressed as the count over the period of exposure

$$\lambda = \frac{\mu}{t}$$

What Is a Vehicle-Mile?

One vehicle traveling one mile contributes one vehicle-mile.

- 1,000 vehicles each traveling 5 miles contribute 5,000 vehicle-miles.
- Exposure combines traffic volume, distance, and observation time.

Let Y_i be the crash count and $t_i > 0$ the exposure for road i.

$$\text{Observed rate} = \frac{Y_i}{t_i}, \qquad \text{Expected rate} = \frac{\mu_i}{t_i}, \quad \mu_i = E(Y_i).$$

Here, t_i is measured in **millions of vehicle-miles**.

From Counts to Rates

Previously, a Poisson regression modeled an expected count:

$$Y_i \sim \text{Poisson}(\mu_i), \qquad \log \mu_i = \beta_0 + \beta_1 x_i.$$

For unequal exposure, model the expected *rate*:

$$\log(\lambda_i) = \log\left(\frac{\mu_i}{t_i}\right) = \beta_0 + \beta_1 x_i.$$

Equivalently,

$$\boxed{\log \mu_i = \log t_i + \beta_0 + \beta_1 x_i.}$$

The Offset

$$\log \mu_i = \underbrace{\log t_i}_{\text{offset}} + \beta_0 + \beta_1 x_i$$

An offset is a known contribution to the linear predictor.

- Its coefficient is fixed at 1; it is not estimated – why?
- We still model the integer crash count Y_i.
- Exposure enters through $\log t_i$, not as the response Y_i / t_i.

R uses `offset(log(exposure))` on the right-hand side or the `offset` argument in `glm()`

What Does the Offset Assume?

Exponentiating gives

$$\mu_i = t_i \exp(\beta_0 + \beta_1 x_i).$$

Holding x_i fixed, doubling exposure doubles the expected count:

$$\frac{\mu(2t, x)}{\mu(t, x)} = 2.$$

The expected rate stays the same.

This is an assumption about how exposure operates. Adding $\log t_i$ as an ordinary predictor would estimate its coefficient and allow a different relationship.

Offset Coef

In particular, if we did allow a coefficient γ for the offset term, we would have

$$\mu_i = t_i^{\gamma} \exp(\beta_0 + \beta_1 x_i).$$

Then, holding x_i fixed and doubling the exposure would give us

$$\frac{\mu(2t, x)}{\mu(t, x)} = 2^{\gamma - 1}$$

Now, doubling the exposure increases the rate by an exponential factor

Non-constant rates

Suppose we did permit fitting a coefficient γ so that

$$\log(\mu_i) = \beta_0 + \gamma \log(t_i),$$

with our rate expressed as a function of exposure

$$\frac{\mu_i}{t_i} = t^{\gamma - 1} \exp(\beta_0)$$

This may be reasonable, say, if we believe that the per-capita murder rate in city is influenced by the size of the city

- $\gamma = 1$: Per-capita murder rates are constant across population sizes
- $\gamma > 1$: Larger cities have higher per-capita murder rate
- $\gamma < 1$: Larger cities have lower per-capita murder rates

An Crash Study

We observe 80 road segments for one year each.

- Response: number of crashes.
- Exposure: millions of vehicle-miles traveled.
- Continuous predictor: mean operating speed in mph.
- Categorical predictor: urban or nonurban road.

The Data

```
> head(crash[, c("crashes", "million_miles", "speed",
    "urban")])
  crashes million_miles speed urban
1       6           5.2  34.0    No
2      13           9.1  36.6    No
3       4           2.8  62.8    No
4      16          11.4  42.2    No
5      16           7.4  59.2    No
6      20           8.6  59.0    No
```

Each row is a road, not an individual vehicle or crash.

A road can have zero crashes and still have positive exposure.

A Continuous Predictor

Let x_i be mean operating speed and $u_i = 1$ for an urban road.

$$\log\left(\frac{\mu_i}{t_i}\right) = \beta_0 + \beta_1 x_i + \beta_2 u_i.$$

For two roads of the same urban status, differing by c mph:

$$\frac{\mu(x+c,u)/t}{\mu(x,u)/t} = \frac{e^{\beta_0+\beta_1(x+c)+\beta_2 u}}{e^{\beta_0+\beta_1 x+\beta_2 u}} = e^{c\beta_1}.$$

e^{β_1} is a **rate ratio** for a one-mph difference.

GLM with Offset

```
> summary(fit)

Call:
glm(formula = crashes ~ speed + urban, family = poisson, data = crash,
    offset = log(million_miles))

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.53770    0.12091   -4.45  8.7e-06 ***
speed        0.02344    0.00233   10.05  < 2e-16 ***
urbanYes     0.30290    0.05060    5.99  2.1e-09 ***
---

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 251.898  on 79  degrees of freedom
Residual deviance:  71.571  on 77  degrees of freedom
AIC: 454.3
```

Here, $\exp(0.02344) = 1.024$ indicates a rate of crash increase of about 2.4% per additional average mph

GLM with Offset (Scaled)

We can alternatively scale our predictors to express rate changes in different units

```
> summary(fit)

Call:
glm(formula = crashes ~ I(speed/10) + urban, family = poisson,
    data = crash, offset = log(million_miles))

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.5377     0.1209   -4.45  8.7e-06 ***
I(speed/10)   0.2344     0.0233   10.05  < 2e-16 ***
urbanYes      0.3029     0.0506    5.99  2.1e-09 ***
---

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 251.898  on 79  degrees of freedom
Residual deviance:  71.571  on 77  degrees of freedom
AIC: 454.3
```

Our fitted values (and deviance) do not change, but now the rate ratio is expressed in increases of 10mph

GLM with Offset (Scaled)

Regardless of scaling, the intercept term describes the log rate of crashes when `Urban = No` and speed is 0 mpg

If we both center and scale our covariate for speed, though, the intercept tells us the rate of crashes at the average speed in rural settings

Our coefficient for scaled speed then gives us the change in rate for a 1 SD change in speed, which may be more or less useful, depending on context

The Same Model in New Units

Since $x_i = \bar{x} + s_x z_i$,

$$\begin{aligned}\beta_0 + \beta_1 x_i + \beta_2 u_i &= \beta_0 + \beta_1(\bar{x} + s_x z_i) + \beta_2 u_i \\ &= \underbrace{(\beta_0 + \beta_1 \bar{x})}_{\alpha_0} + \underbrace{s_x \beta_1}_{\alpha_1} z_i + \beta_2 u_i.\end{aligned}$$

- Centering changes the interpretation of the intercept.
- Scaling changes the numerical slope.
- Fitted counts, rates, likelihood, and AIC are unchanged.

GLM with Offset (Scaled)

This changes the coefficients for the intercept and speed, but leave the rate ratio for Urban status unchanged

```
> summary(fit)

Call:
glm(formula = crashes ~ I(scale(speed)) + urban, family = poisson,
    data = crash, offset = log(million_miles))

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)       0.6119     0.0383   15.97  < 2e-16 ***
I(scale(speed))   0.2599     0.0259   10.05  < 2e-16 ***
urbanYes          0.3029     0.0506    5.99  2.1e-09 ***
---

> exp(coef(fit))
    (Intercept) I(scale(speed))        urbanYes
          1.844           1.297           1.354
```

Read the Coefficients

$$\log(\hat{\mu}_i/t_i) = 0.612 + 0.260 z_i + 0.303 u_i$$

- At the mean speed, a nonurban road has fitted rate $e^{0.612} = 1.844$ crashes per million vehicle-miles.
- One SD higher speed multiplies the fitted rate by $e^{0.260} = 1.297$, holding urban status fixed.
- Urban versus nonurban multiplies the fitted rate by $e^{0.303} = 1.354$, holding speed fixed.

Rate Ratios and an Interval

```
> exp(confint(fit))

                  2.5 % 97.5 %
(Intercept)       1.709  1.986
I(scale(speed)) 1.233  1.364
urbanYes          1.226  1.495
```

`confint` uses the profile likelihood on the log-rate scale. Exponentiating gives an interval for the rate ratio.

Exponentiating the intercept gives a *rate*; exponentiating a slope gives a *rate ratio*.

Does Speed Help Explain the Rate?

Just as before, we can use differences in deviance to test $H_0 : \beta_1 = 0$, retaining urban status and the offset.

```
> no_speed <- update(fit, . ~ . - speed_z)
> anova(no_speed, fit, test = "LRT")
Analysis of Deviance Table

Model 1: crashes ~ urban + offset(log(million_miles))
Model 2: crashes ~ speed_z + urban + offset(log(million_miles))
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        78      173.3
2        77       71.6  1      102   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This compares nested Poisson models fit to the same counts and exposures.

Scaling the Offset

We might also be interested in expressing our rate in different units

Let t_i be million vehicle-miles, and $t_i^* = 10^6 t_i$ be vehicle-miles.

$$\log \mu_i = \log t_i + \alpha_0 + \beta_1 z_i + \beta_2 u_i$$

becomes

$$\log \mu_i = \log t_i^* + \underbrace{(\alpha_0 - \log 10^6)}_{\text{new intercept}} + \beta_1 z_i + \beta_2 u_i.$$

- The rate is now expressed per vehicle-mile.
- Slopes and rate ratios do not change.
- Fitted counts do not change.

Scaling the Offset

```
glm(formula = crashes ~ I(scale(speed)) + urban, family = poisson,
    data = crash, offset = log(million_miles))

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)       0.6119     0.0383   15.97  < 2e-16 ***
I(scale(speed))   0.2599     0.0259   10.05  < 2e-16 ***
urbanYes          0.3029     0.0506    5.99  2.1e-09 ***
---
Residual deviance:  71.571  on 77  degrees of freedom

glm(formula = crashes ~ I(scale(speed)) + urban, family = poisson,
    data = crash, offset = log(10^6 * million_miles))

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     -13.2036     0.0383 -344.55  < 2e-16 ***
I(scale(speed))   0.2599     0.0259   10.05  < 2e-16 ***
urbanYes          0.3029     0.0506    5.99  2.1e-09 ***
---

Residual deviance:  71.571  on 77  degrees of freedom
```

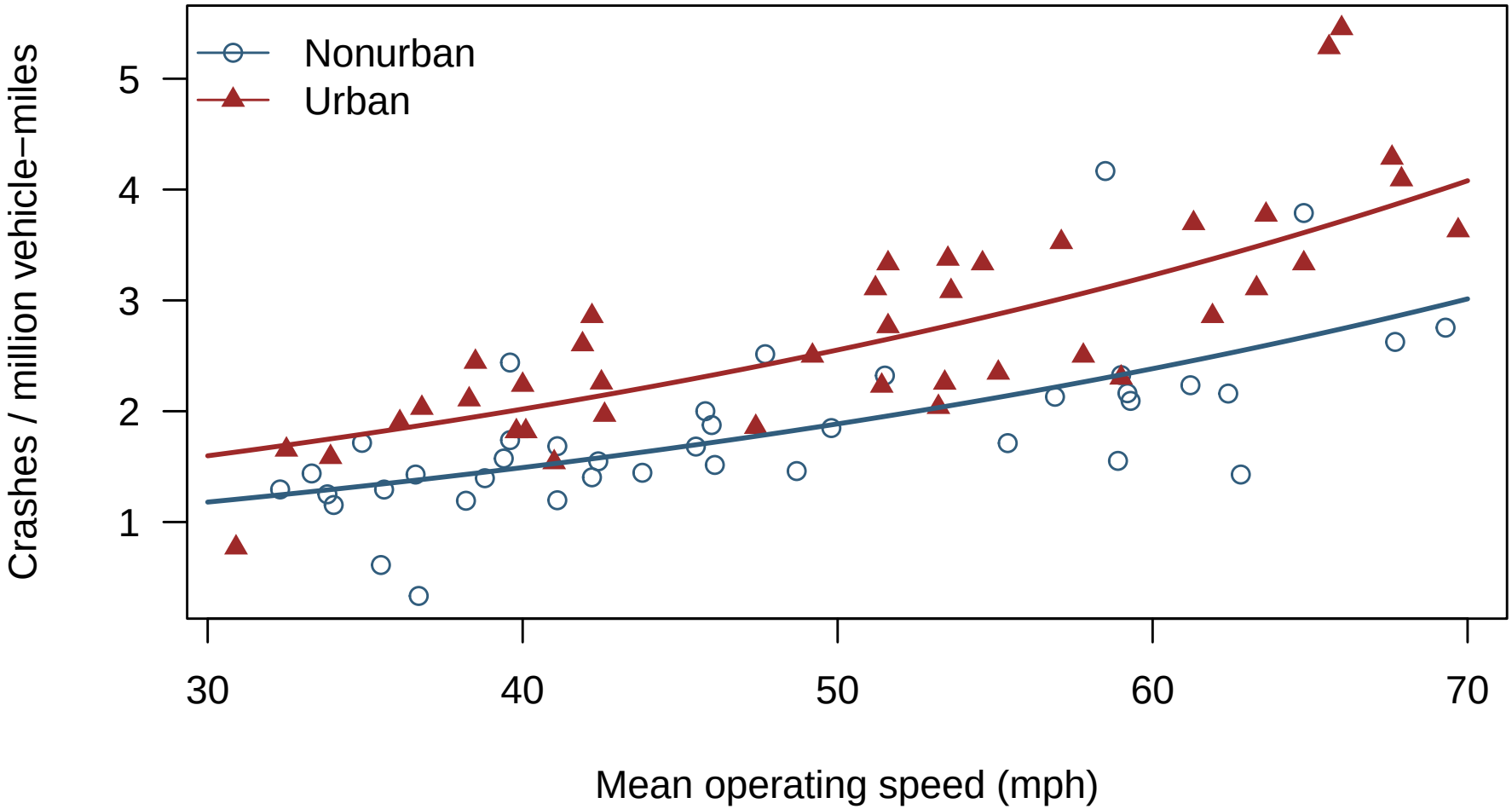
Predict a Rate or a Count

Just as with other models, we use `predict()` to get predicted counts

```
> new <- data.frame(speed_z = c(0, 1),
    urban = factor(c("No", "No"), levels = levels(d$urban)),
    million_miles = c(1, 1))
> predict(fit, newdata = new, type = "response")
    1     2
1.844 2.391
> new$million_miles <- c(10, 10)
> predict(fit, newdata = new, type = "response")
    1     2
18.44 23.91
```

With exposure 1, predictions are counts per million vehicle-miles. With exposure 10, they are expected counts over 10 million vehicle-miles.

Fitted Rates



What Else Did We Assume?

For the Poisson model,

$$\text{Var}(Y_i \mid z_i, u_i, t_i) = \mu_i.$$

The model specifies both a mean and a variance.

- The log rate is linear in the predictors.
- Counts are conditionally independent.
- Conditional variance equals conditional mean.

Even with the same speed, location, and exposure, might some roads have higher accident rates because of truck traffic, road design, or weather?

Unmodeled differences in rates can produce more variability than the Poisson model allows.

Before Friday

Two different questions:

1. How does the **expected rate** change with predictors?
2. How much do counts **vary around that expectation**?

Offsets and slopes address the first question.

Friday: how do we accommodate greater observed variability than the Poisson model allows?