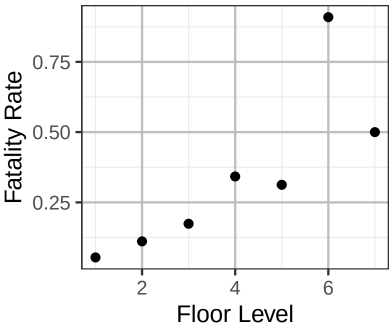
STAT 226 Lecture 9

Sections 3.1-3.2 Generalized Linear Models (GLM)

Yibi Huang

Example — Fatality in Falling Accidents[1]

floor level x	fatal falls y_x	total falls n_x	observed fatality rate $\widehat{\pi}_x = y_x/n_x$
1	2	37	$2/37 \approx 0.05$
2	6	54	$6/54 \approx 0.11$
3	8	46	$8/46 \approx 0.17$
4	13	38	$13/38 \approx 0.34$
5	10	32	$10/32 \approx 0.31$
6	10	11	$10/11 \approx 0.91$
7	1	2	$1/2 \approx 0.50$



If the falls were indep. of each other, and if the chance of fatality depended only on the floor level from which the victims fell, then

$$y_x \sim \text{Binomial}(n_x, \pi(x)).$$

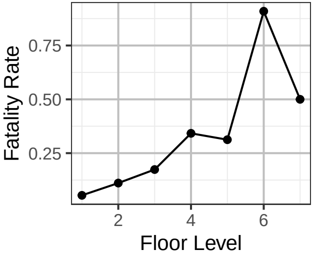
The MLE of $\pi(x)$ is $\widehat{\pi}_x = y_x/n_x$.

[1] Courtesy of Prof. Stephen M. Stigler

Why Modeling?

Without modeling, we can estimate $\pi(x)$ at $x = 1, 2, \ldots, 7$ using the sample fatality rate y_x/n_x, but there are a few problems.

- cannot estimate $\pi(x)$ at x's with no observation, e.g., $x = 8$ or 1.5.



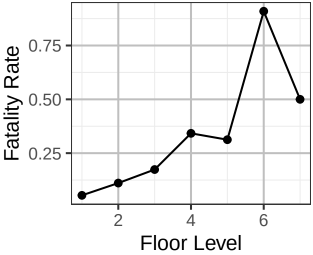
Why Modeling?

Without modeling, we can estimate $\pi(x)$ at $x = 1, 2, \ldots, 7$ using the sample fatality rate y_x/n_x, but there are a few problems.

- cannot estimate $\pi(x)$ at x's with no observation, e.g., $x = 8$ or 1.5.
- Fatality rate $\pi(x)$ should increase with floor level x. However, . . .

$$\widehat{\pi}(4) \approx 0.34 > \widehat{\pi}(5) \approx 0.31,$$

$$\widehat{\pi}(6) \approx 0.91 > \widehat{\pi}(7) \approx 0.50,$$



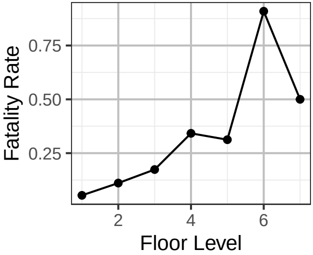
Why Modeling?

Without modeling, we can estimate $\pi(x)$ at $x = 1, 2, \ldots, 7$ using the sample fatality rate y_x/n_x, but there are a few problems.

- cannot estimate $\pi(x)$ at x's with no observation, e.g., $x = 8$ or 1.5.
- Fatality rate $\pi(x)$ should increase with floor level x. However, ...

$$\widehat{\pi}(4) \approx 0.34 > \widehat{\pi}(5) \approx 0.31,$$

$$\widehat{\pi}(6) \approx 0.91 > \widehat{\pi}(7) \approx 0.50,$$



- By modeling, we can incorporate prior knowledge about $\pi(x)$ to improve the accuracy of estimation.
  E.g., we can model $\pi(x)$ as an increasing function of x

$$\pi(x) = \alpha + \beta x, \quad \text{or} \quad \pi(x) = \frac{e^{\alpha+\beta x}}{1 + e^{\alpha+\beta x}} \quad \text{with } \beta > 0.$$

First Model — Linear Least-Square Regression

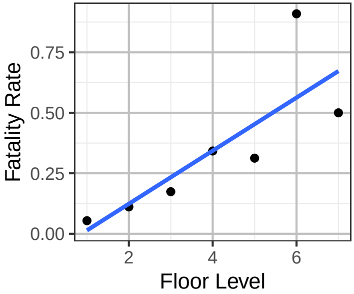
Suppose we model $\pi(x)$ as

$$\pi(x) = \alpha + \beta x,$$

how to estimate α and β? Let's try least-square regression with

- response = the observed fatality rates $\widehat{\pi}_x = y_x/n_x$, and
- predictor = the floor level x

floor level x	fatal falls y_x	total falls n_x	fatality rate $\widehat{\pi}_x$
1	2	37	0.05
2	6	54	0.11
3	8	46	0.17
4	13	38	0.34
5	10	32	0.31
6	10	11	0.91
7	1	2	0.50



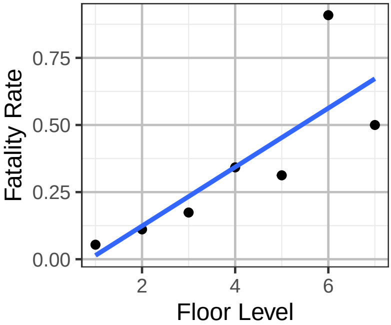
First Model — Linear Least Square Regression

Fitting a linear regression model, we get

$$\widehat{\pi(x)} = -0.0957 + 0.1097x,$$

which means, if the fall occurs one floor higher, the chance for it to be fatal increases by about 11%.

Any problem with this model?

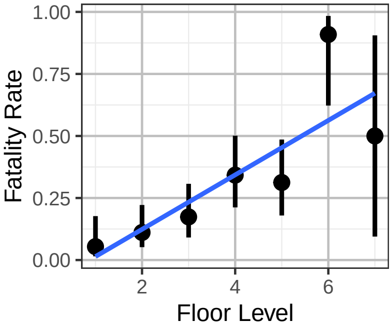


Problems of the Linear Least Square Regression

1. **Non-normality** of the response $\widehat{\pi}_x$
   - not a big issue since least square regression is robust to non-normality
2. **Non-constant variance** of the response: $\mathrm{SE}(\widehat{\pi}_x) = \sqrt{\frac{\widehat{\pi}_x(1-\widehat{\pi}_x)}{n_x}}$

Regression models assume constant variability.

Points w/ smaller SEs should be more influential to the fitted line as they are more accurate.



(Error bars are 95% score CIs for $\pi(x)$).

Problems of the Linear Least Square Regression

3. For probabilities,
   the diff. of $\pi_1 = 0.01$ and $\pi_2 = 0.0001$ is important, but
   the diff. of $\pi_1 = 0.51$ and $\pi_2 = 0.5001$ is often negligible.
   - Least square method regards the two differences equal,
   - Likelihood methods can reflect the distinction of the two differences.

4. $\pi(x) = \alpha + \beta x$ may not stay between 0 and 1

Second Attempt — Likelihood Methods

As $y_x \sim$ Binomial$(n_x, \pi(x))$, the likelihood of $\pi(x)$ is

$$\ell = \prod_{x=1}^{7} \binom{n_x}{y_x} [\pi(x)]^{y_x} [1 - \pi(x)]^{n_x - y_x} = C \prod_{x=1}^{7} [\pi(x)]^{y_x} [1 - \pi(x)]^{n_x - y_x}$$

where $C = \prod_{x=1}^{7} \binom{n_x}{y_x}$ is a constant involving no parameters, having no effect on parameter inference, and hence is often ignored.

For the linear probability model

$$\pi(x) = \alpha + \beta x,$$

the likelihood of α, β is

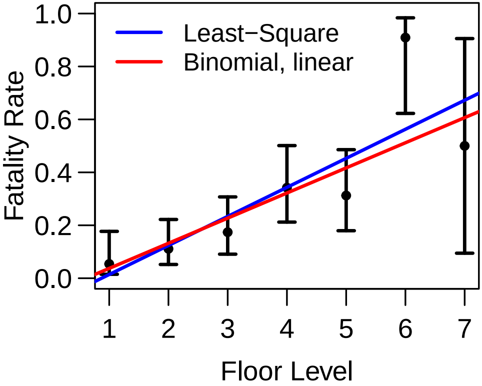
$$\ell(\alpha, \beta) = C \prod_{x=1}^{7} [\alpha + \beta x]^{y_x} [1 - \alpha - \beta x]^{n_x - y_x}.$$

- No close form formula for the MLEs $\widehat{\alpha}$ and $\widehat{\beta}$.
  R gives their values as $\widehat{\alpha} = -0.0577$, $\widehat{\beta} = 0.0949$.

Compare the two fitted lines founded using regression and binomial likelihoods.

$$\text{Linear Least Square : } \widehat{\pi(x)} = -0.0957 + 0.1097x$$

$$\text{Binomial likelihoods : } \widehat{\pi(x)} = -0.0577 + 0.0949x$$



(Error bars are 95% score CIs for $\pi(x)$)

Why Likelihood Methods Better Than Least-Square Estimates?

$$\text{likelihood} : C \prod_x [\pi(x)]^{y_x} [1 - \pi(x)]^{n_x - y_x}$$

$$\text{log-likelihood} : \log C + \sum_x \{y_x \log \pi(x) + (n_x - y_x) \log[1 - \pi(x)]\}$$

Contribution of an observation (x, n_x, y_x) to the log-likelihood is

$$y_x \log \pi(x) + (n_x - y_x) \log[1 - \pi(x)].$$

- Observations with larger n_x are more influential as they have greater contributions to log-likelihood

Why Likelihood Methods Better Than Least-Square Estimates?

likelihood : $C\prod_x [\pi(x)]^{y_x}[1-\pi(x)]^{n_x-y_x}$

log-likelihood : $\log C + \sum_x \{y_x \log \pi(x) + (n_x - y_x)\log[1-\pi(x)]\}$

Contribution of an observation (x, n_x, y_x) to the log-likelihood is

$$y_x \log \pi(x) + (n_x - y_x)\log[1-\pi(x)].$$

- Observations with larger n_x are more influential as they have greater contributions to log-likelihood
- Each single $y_x \log \pi(x) + (n_x - y_x)\log[1-\pi(x)]$ reach its max. at $\pi(x) = y_x/n_x$. Likelihood methods will make the fitted $\widehat{\pi}(x)$ as close to y_x/n_x as possible.

Why Likelihood Methods Better Than Least-Square Estimates?

likelihood : $C \prod_x [\pi(x)]^{y_x} [1 - \pi(x)]^{n_x - y_x}$

log-likelihood : $\log C + \sum_x \{y_x \log \pi(x) + (n_x - y_x) \log[1 - \pi(x)]\}$

Contribution of an observation (x, n_x, y_x) to the log-likelihood is

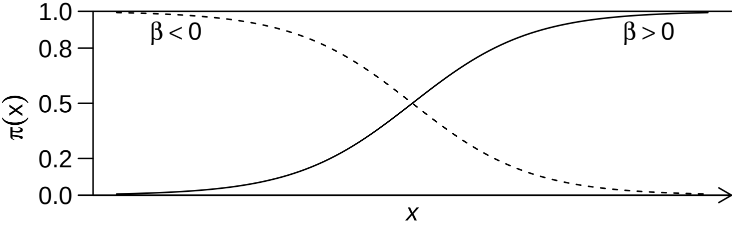
$$y_x \log \pi(x) + (n_x - y_x) \log[1 - \pi(x)].$$

- Observations with larger n_x are more influential as they have greater contributions to log-likelihood
- Each single $y_x \log \pi(x) + (n_x - y_x) \log[1 - \pi(x)]$ reach its max. at $\pi(x) = y_x/n_x$. Likelihood methods will make the fitted $\widehat{\pi}(x)$ as close to y_x/n_x as possible.
- log-likelihood changes a little when $\pi(x)$ changes from 0.51 to 0.501, log-likelihood changes a lot when $\pi(x)$ changes from 0.01 to 0.001.

S-shaped Relationships

In practice, $\pi(x)$ often increases or decreases slower as $\pi(x)$ gets closer to 0 or 1.

The S-shaped curves below are often (close to) realistic.



The most commonly used S-shaped function for modeling $\pi(x)$ is

$$\pi(x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)} = \frac{e^{\alpha+\beta x}}{1 + e^{\alpha+\beta x}}$$

Logistic Regression Models

The logistic regression model models the success probability $\pi(x)$ for the binomial response as

$$\pi(x) = \frac{e^{\alpha+\beta x}}{1 + e^{\alpha+\beta x}},$$

or equivalently,

$$\log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \alpha + \beta x.$$

- It ensures $\pi(x)$ staying between 0 and 1 regardless of the values of α, β, and x
- $g(\pi) = \log(\frac{\pi}{1-\pi})$ is called the *logit* function
- Interpretation: $\log(\text{odds}) = \alpha + \beta x$
  the odds increases by a factor of e^{β} whenever x increases by 1
- More details in Chapter 4 & 5

For the fatal fall example, the likelihood of α and β is

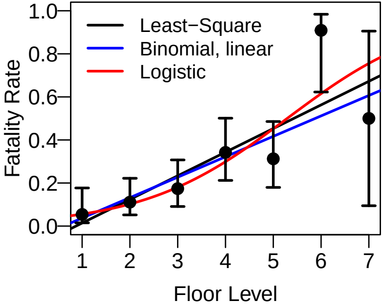
$$\ell(\alpha,\beta) = C \prod_{x=1}^{7} \left(\frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}\right)^{y_x} \left(\frac{1}{1+e^{\alpha+\beta x}}\right)^{n_x - y_x} .$$

MLEs for α and β:

$$\widehat{\alpha} \approx -3.492, \quad \widehat{\beta} \approx 0.660$$

The fitted model is

$$\widehat{\pi}(x) \approx \frac{e^{-3.492+0.660x}}{1+e^{-3.492+0.660x}}.$$



Estimated fatality rate for falls from the first floor is

$$\widehat{\pi}(1) \approx \frac{e^{-3.492+0.660\times 1}}{1+e^{-3.492+0.660\times 1}} \approx 0.0556 \approx 5.6\%.$$

Odds of death become $e^{0.660} \approx 1.93$ times as large if the falling accidents occurred one floor higher

Three Components of Generalized Linear Models

- **Random component** Y

  — the response variable with indep. obs. $Y_1, Y_2, \ldots, Y_n$ from a common prob. dist. (e.g., normal, binomial, Poisson)
- **Linear Predictor** — the explanatory variables of a linear structure

$$\alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

  Some x_j can be based on others x_k's, e.g., $x_3 = x_1 x_2$, $x_4 = x_1^2$
- **Link function** $g(\mu)$

  — connecting $\mu = \mathrm{E}[Y]$ and $\alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$ by a function

$$g(\mu) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

The same maximum likelihood (ML) fitting procedure is used to estimate the coefficients $\alpha, \beta_1, \ldots, \beta_k$ for all GLMs.

Linear Regression Models Are GLMs

Recall the ordinary linear regression models assume

$$Y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

where the noise ε has a normal distribution $N(0, \sigma^2)$

- The random component Y has a normal distribution
- $\alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$ is the linear predictor
- The link function is the identity link $g(\mu) = \mu$

$$g(\mu) = \mu = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

- The ML fitting procedure for estimating $\alpha, \beta_1, \ldots, \beta_k$ reduces to the **least square method** when the response variable has a normal distribution.

Commonly Used Link Functions

Link functions are usually continuous and strictly monotone.

- Identity link: $g(\mu) = \mu$
  - used when $Y \sim$ Normal, linear regression
- Log link: $g(\mu) = \log(\mu)$

$$\log(\mu) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

  - used when $Y \sim$ Poisson. See Section 3.3 and Ch 7
- Logit link $g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$
  - used for binary response models. See Chapter 4 and 5
- Other commonly used link functions for binary response models: probit, log-log, complementary log-log
  - not covered in STAT 226

How to Fit GLM in R

Loading data:

```
ff = read.table(
  "https://www.stat.uchicago.edu/~yibi/s226/falls.txt",
  h=T)
ff
```

```
  floor fatal live
1     1     2   35
2     2     6   48
3     3     8   38
4     4    13   25
5     5    10   22
6     6    10    1
7     7     1    1
```

GLM in R

```
ff.lin = glm(cbind(fatal,live) ~ floor,
               family=binomial(link="identity"),data=ff)
ff.lin$coef
(Intercept)       floor
   -0.05771     0.09491
ff.logit = glm(cbind(fatal,live) ~ floor,
                 family=binomial(link="logit"), data=ff)
ff.logit$coef
(Intercept)       floor
     -3.492       0.660
```

Fitted binomial model w/ identity link: $\widehat{\pi}(x) = -0.05771 + 0.09491x$.

Fitted logistic regression model: $\widehat{\pi}(x) = \dfrac{e^{-3.492+0.660x}}{1 + e^{-3.492+0.660x}}$.

Another way to fit a glm model

```
ff$total = ff$fatal+ff$live
ff$percent = ff$fatal/ff$total
ff.logit2 = glm(percent ~ floor, family=binomial(link="logit"),
                  weight = total, data=ff)
ff.logit2$coef                    # same fitted coefficients!
(Intercept)       floor
     -3.492       0.660
ff.logit$coef
(Intercept)       floor
     -3.492       0.660
```

Ungrouped Data and Grouped Data

Sometimes the data are ungrouped . . .

Ungrouped Data:

file: `fallsUG.txt`

no.	floor	outcome
1	2	live
2	5	live
3	5	live
4	2	live
5	1	live
6	4	live
7	5	fatal
8	1	live
9	4	live
10	3	live
11	4	live
12	4	fatal

Grouped Data:

file: `falls.txt`

floor	fatal	live
1	2	35
2	6	48
3	8	38
4	13	25
5	10	22
6	10	1
7	1	1

Fitting GLM for Ungrouped Data

```
ffug = read.table(
  "https://www.stat.uchicago.edu/~yibi/s226/fallsUG.txt",
  header=TRUE)
ffug.logit = glm((outcome == "fatal") ~ floor,
                  family=binomial(link="logit"), data=ffug)

ffug.logit$coef                 # same fitted coefficients!
(Intercept)       floor
     -3.492       0.660
ff.logit$coef
(Intercept)       floor
     -3.492       0.660
```

Fitted Values $\widehat{\pi}(x)$

Estimated $\widehat{\pi}(x)$ for the Binomial model with identity link

```
ff.lin$fit
      1       2       3       4       5       6       7
0.03719 0.13210 0.22701 0.32191 0.41682 0.51172 0.60663
```

Estimated $\widehat{\pi}(x)$ for the Binomial model with logit link (logistic regression)

```
ff.logit$fit
      1       2       3       4       5       6       7
0.05562 0.10230 0.18065 0.29903 0.45218 0.61495 0.75550
```

```r
library(binom)
ci = binom.confint(ff$fatal,ff$total,conf.level=0.95,method="wilson")
ff$lower = ci$lower;  ff$upper = ci$upper
ff$lsfit = fflm1$fit; ff$linfit = ff.lin$fit;
ff$logitfit = ff.logit$fit; ff
```

95% score CI

	floor	fatal	live	total	percent	lower	upper	lsfit	linfit	logitfit
1	1	2	35	37	0.054	0.015	0.18	0.014	0.037	0.056
2	2	6	48	54	0.111	0.052	0.22	0.124	0.132	0.102
3	3	8	38	46	0.174	0.091	0.31	0.234	0.227	0.181
4	4	13	25	38	0.342	0.212	0.50	0.343	0.322	0.299
5	5	10	22	32	0.312	0.180	0.49	0.453	0.417	0.452
6	6	10	1	11	0.909	0.623	0.98	0.563	0.512	0.615
7	7	1	1	2	0.500	0.095	0.91	0.672	0.607	0.756