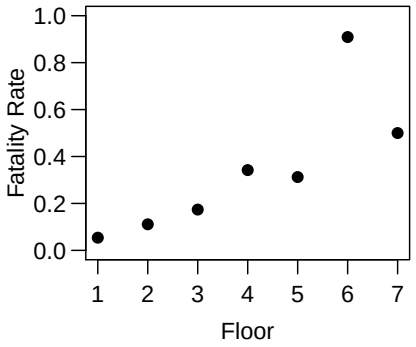
Chapter 3 Generalized Linear Models (GLM)

Example — Fatality in Falling Accidents[1]



floor level x	fatal falls y_x	total falls n_x	observed fatality rate $p_x = y_x/n_x$
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

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 $p_x = y_x/n_x$.

[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 those x with no observation, e.g., $x = 8$ or 1.5.
- We expect the fatality rate $\pi(x)$ to increase with the floor level x. However, the sample fatality rates $p_x = y_x/n_x$ are not monotone increasing in x:

$$p_4 = 0.34 > p_5 = 0.31,$$
$$p_6 = 0.91 > p_7 = 0.50,$$

which is not reasonable.

  - 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

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

First Attempt – Linear Regression

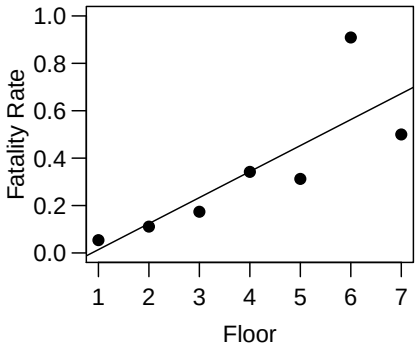
Suppose we model $\pi(x)$ as

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

how to estimate α and β? Let's try linear regression with

- response = the observed fatality rates $p_x = y_x/n_x$, and
- predictor = the floor level x

floor level x	fatal falls y_x	total falls n_x	fatality rate p_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

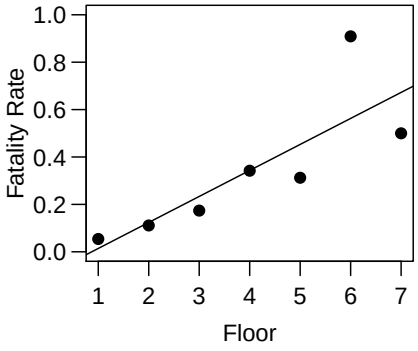


First Model – Linear Regression

Fitting a linear regression model, we get

$$\widehat{\pi(x)} = -0.09566 + 0.10973x,$$

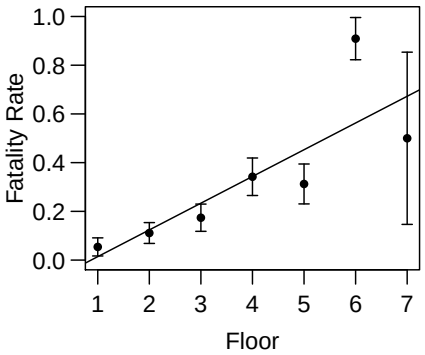
which means, if the fall occurs one floor higher, the chance for it to be fatal increases by about 11%.



Any problem in this model?

Problems of the Linear Regression Model

1. **Non-normality** of the response p_x
   - not a serious issue because regression models don't require the response to be normal
2. **Non-constant variance** of the response: $\text{SE}(p_x)=\sqrt{\frac{p_x(1-p_x)}{n_x}}$



Regression models assume constant variabilities of all points.

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

(The error bars go 1 SE above and below p_x).

Problems of the Linear Regression Model

3. For probabilities,
   the diff. of $\pi = 0.01$ and $\pi = 0.0001$ is important, but
   the diff. of $\pi = 0.51$ and $\pi = 0.5001$ 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 $y_x \sim \text{binomial}(n_x, \pi(x))$, 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

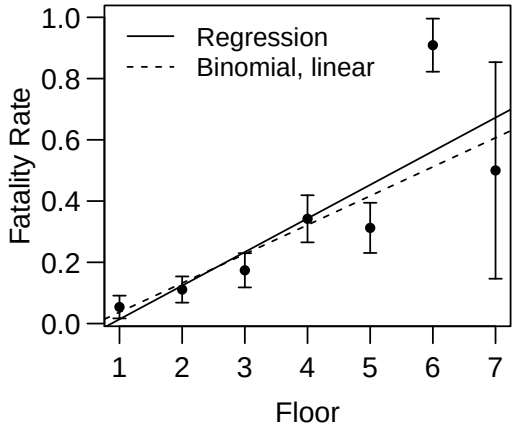
$$\ell(\alpha, \beta) = C \prod_{x=1}^{7} [\alpha + \beta x]^{y_x}[1-\alpha-\beta x]^{n_x-y_x}.$$

- No close form formula for the MLEs of α and β. Numerical tools give their values as

$$\widehat{\alpha} = -0.0577, \quad \widehat{\beta} = 0.0949.$$

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

$$\text{Regression} : \widehat{\pi(x)} = -0.0957 + 0.1097x$$
$$\text{Binomial likelihoods} : \widehat{\pi(x)} = -0.0577 + 0.0949x$$



Why Likelihood Methods Improve Over Regression?

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

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

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

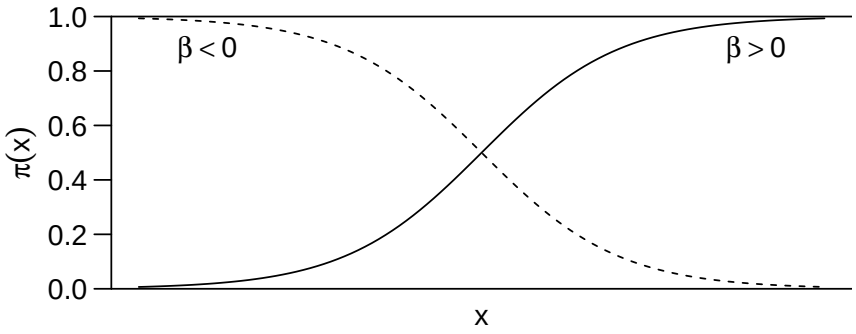
$$y_x \log \pi(x) + (n_x - y_x)\log[1-\pi(x)].$$

- Observations with larger n_x are more influential as they have greater contributions to log-likelihood
- 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 $\begin{matrix}\text{a little}\\ \text{a lot}\end{matrix}$ when $\pi(x)$ changes $\begin{matrix}\text{from .51 to .501,}\\ \text{from .01 to .001.}\end{matrix}$

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
- We'll discuss in detail in Chapter 4 & 5

For the example about accidental falls, the likelihood function of α and β is

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

The MLE of α and β is:

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

The fitted model is

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

- The model estimates that 5.6% of the falls from the first floor is fatal because
$$\widehat{\pi}(1) = \frac{e^{-3.492+0.660\times 1}}{1+e^{-3.492+0.660\times 1}} \approx 0.0556 \approx 5.6\%.$$
- If a victim had fell from somewhere one floor higher, the odds of death would have increased by a factor of $e^{0.660} \approx 1.93$.

Probit Regression Model

Another model that has the S-shaped curves is the *probit model*, which assumes

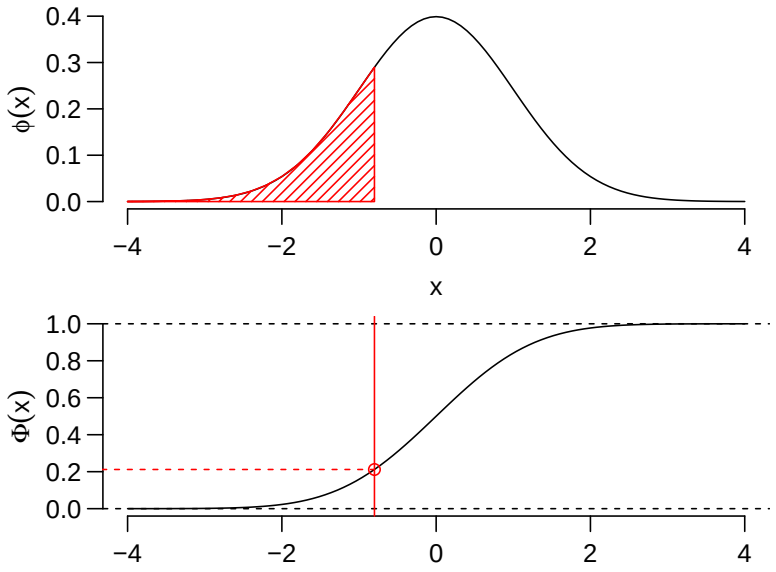
$$\pi(x) = \Phi(\alpha + \beta x)$$

where Φ is the cumulative distribution function of $N(0, 1)$,

$$\Phi(z) = \mathrm{P}(N(0,1) \leq z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx.$$

- What are the values of $\Phi(0)$, $\Phi(-1.96)$, $\Phi(1.96)$?

N(0,1) density function $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ (top panel) and cumulative distribution function $\Phi(x) = \int_{-\infty}^{x} \phi(z)dz$ (bottom panel).



Probit Regression Model

The fitted probit regression model for the accidental fall data (based on likelihood methods) is

$$\widehat{\pi}(x) = \Phi(\widehat{\alpha} + \widehat{\beta} x) = \Phi(-2.0241 + 0.3794x).$$

- The estimated fatality rate of falling from the first floor is

$$\widehat{\pi}(1) = \Phi(-2.0241 + 0.3794 \times 1) = \Phi(-1.6447) \approx 0.0500.$$

Complementary Log-Log Models

Both logit and probit models assume that $\pi(x)$ approaches 0 at the same rate as it approaches 1.

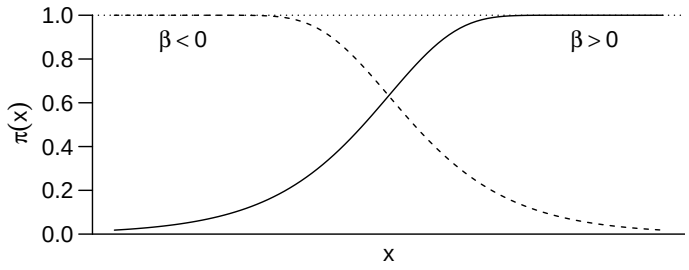
The complementary log-log models assume

$$\pi(x) = 1 - \exp(-\exp(\alpha + \beta x)),$$

or equivalently

$$\log(-\log(1 - \pi(x))) = \alpha + \beta x.$$

In this model, $\pi(x)$ approaches 0 fairly slowly but approaches 1 quite sharply.

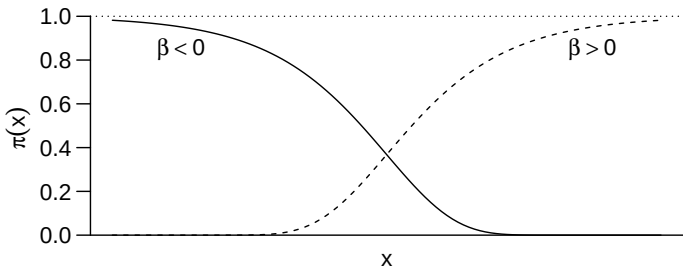


Log-log Models

On the contrary, the log-log models assume

$$\pi(x) = \exp(-\exp(\alpha + \beta x)), \quad \text{or} \quad \log(-\log(\pi(x))) = \alpha + \beta x$$

of which $\pi(x)$ approaches 0 quickly but approaches 1 slowly



- Neither the complementary log-log models nor the log-log models are included in the textbook.

  We include them here just for your reference.

Link Functions

All the models above assume a linear relationship between the explanatory variable x and some function $g(\pi)$ of the binomial proportion π.

- logit: $\pi = \dfrac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}} \Leftrightarrow g(\pi) = \log\left(\dfrac{\pi}{1-\pi}\right) = \alpha + \beta x$
- probit: $\pi = \Phi(\alpha + \beta x) \Leftrightarrow g(\pi) = \Phi^{-1}(\pi) = \alpha + \beta x$
- complementary log-log:

$$\pi = 1 - e^{-e^{\alpha+\beta x}} \Leftrightarrow g(\pi) = \log(-\log(1-\pi)) = \alpha + \beta x$$

- log-log:

$$\pi = e^{-e^{\alpha+\beta x}} \Leftrightarrow g(\pi) = \log(-\log(\pi)) = \alpha + \beta x$$

All models above belong to a large class of models .... the **generalized linear models**.

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)
- **System component** — 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 $\mu = \mathbb{E}[Y]$ 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 $N(0, \sigma^2)$

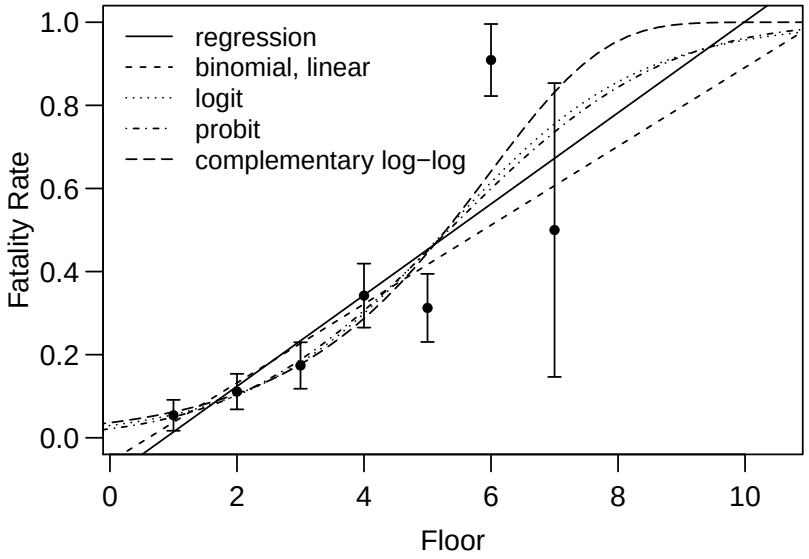
- The random component Y has a normal distribution
- $\alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$ is the systematic component
- The link function is the identity link $g(\mu) = \mu$

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

- 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.

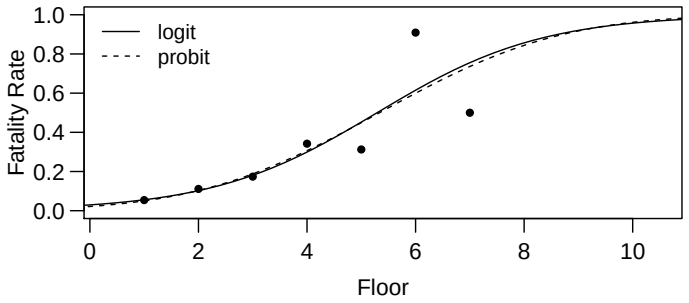
Back to the Example of Fatal Falls

Which model fits the data the best?



How to Choose a Link Function?

- Logit models and probit models usually give similar fitted curves



- Locally, both logit and probit link are close to linear
- Often, the conclusions made are not sensitive to the choice of the link function, though there are occasional exceptions.
- Logit models have nice interpretation (odds ratio) and hence are most commonly used for binomial response data

How to Fit GLM in R

Loading data:

```
> ff = read.table("falls.dat",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
```

Fitting a binomial model with identity link $\pi(x) = \alpha + \beta x$

```
> ff.lin = glm(cbind(fatal,live) ~ floor,
               family=binomial(link="identity"),data=ff)
> ff.lin$coef
(Intercept)       floor
-0.05771138  0.09490576
```

The fitted model is $\widehat{\pi}(x) = -0.05771138 + 0.09490576x$.

Fitting a logit, a probit, or a complementary log-log model:

```
> ff.logit = glm(cbind(fatal,live) ~ floor,
                 family=binomial(link="logit"), data=ff)
> ff.probit = glm(cbind(fatal,live) ~ floor,
                 family=binomial(link="probit"), data=ff)
> ff.cloglog = glm(cbind(fatal,live) ~ floor,
                  family=binomial(link="cloglog"), data=ff)
> ff.logit$coef
(Intercept)       floor
 -3.4920438   0.6600324
> ff.probit$coef
(Intercept)       floor
 -2.0241333   0.3793616
> ff.cloglog$coef
(Intercept)       floor
 -3.2997589   0.5540277
```

The fitted logit model is $\widehat{\pi}(x) = \dfrac{e^{-3.492+0.660x}}{1+e^{-3.492+0.660x}}$,

the fitted probit model is $\widehat{\pi}(x) = \Phi(-2.024 + 0.3794x)$, and

the fitted complementary log-log model is $\widehat{\pi}(x) = 1 - e^{-e^{-3.300+0.554x}}$.

Another syntax to fit a glm model

```
> total = ff$fatal+ff$live
> percent = ff$fatal/total
> ff.logit2 = glm(percent ~ floor, family=binomial(link="logit"),
                  weight = total, data=ff)

> ff.logit2$coef                    # same fitted coefficients!
(Intercept)       floor
 -3.4920438   0.6600324

> ff.logit$coef
(Intercept)       floor
 -3.4920438   0.6600324
```


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

Fitted values for $\widehat{\pi}(x)$ at data points, e.g., for the model with identity link

```
> ff.lin$fit
        1         2         3         4         5         6
0.03719438 0.13210014 0.22700590 0.32191166 0.41681743 0.51172319 0.60662
```

```
> round(ff.lin$fit, 3)
    1     2     3     4     5     6     7
0.037 0.132 0.227 0.322 0.417 0.512 0.607
```

for the models with logit and probit links

```
> round(ff.logit$fit,3)
    1     2     3     4     5     6     7
0.056 0.102 0.181 0.299 0.452 0.615 0.756
```

```
> round(ff.probit$fit,3)
    1     2     3     4     5     6     7
0.050 0.103 0.188 0.306 0.449 0.599 0.736
```


Ungrouped Data and Grouped Data

Sometimes the data are ungrouped ...

Ungrouped Data:
file: `fallsUG.dat`

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
⋮		
219	1	live
220	4	live

Grouped Data:
file: `falls.dat`

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("fallsUG.dat",h=T)
> ffug.logit = glm((outcome == "fatal") ~ floor,
                   family=binomial(link="logit"), data=ffug)

> ffug.logit$coef              # same fitted coefficients!
(Intercept)       floor
 -3.4920437   0.6600324

> ff.logit$coef
(Intercept)       floor
 -3.4920438   0.6600324

> round(ffug.logit$fit,3)      # estimated fatality rates
    1     2     3     4     5     6     7     8     9
0.102 0.452 0.452 0.102 0.056 0.299 0.452 0.056 0.299
(... omitted ...)
   214   215   216   217   218   219   220
 0.102 0.102 0.615 0.299 0.181 0.056 0.299
```

3.4 Statistical Inference for GLMs

The **Wald statistic** for testing H_0: $\beta = c$ is

$$z = \frac{\widehat{\beta} - c}{\text{SE}(\widehat{\beta})}$$

We omit the formula for $\text{SE}(\widehat{\beta})$, of which the value can be found in R as follows.

```
> ff.lin = glm(cbind(fatal,live) ~ floor,
               family=binomial(link="identity"),data=ff)
> summary(ff.lin)
...
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.05771    0.03927  -1.469    0.142
floor        0.09491    0.01567   6.057 1.38e-09 ***
```

The column `Std.Error` gives the desired SE.

Remark: The SE of $\widehat{\beta}$ depends on the unknown true value of β. The SE in the Wald statistic is evaluated at $\beta = \widehat{\beta}$, not at the value $\beta = c$ under H_0.

Wald statistic is approx. $N(0, 1)$ under H_0.

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.05771    0.03927  -1.469    0.142
floor        0.09491    0.01567   6.057 1.38e-09 ***
```

- R summary output gives the Wald statistics `z value` for testing H_0: $\beta = 0$ and the corresponding P-values.

$$\texttt{z value} = \frac{\texttt{Estimate}}{\texttt{Std. Error}} = \frac{\widehat{\beta}}{\text{SE}} = \frac{0.09491}{0.01567} = 6.057.$$

- To test H_0: $\beta = 0.05$,

$$\text{Wald statistic } z = \frac{\widehat{\beta} - 0.05}{\text{SE}(\widehat{\beta})} = \frac{0.09491 - 0.05}{0.01567} \approx 2.866$$

The P-value is about 0.004.

Wald CIs

The **Wald** $(1-\alpha)100\%$ **CIs** for β are

$$\widehat{\beta} \pm z_{\alpha/2}\text{SE}(\widehat{\beta}).$$

e.g., 95% CI for β:

$$0.09491 \pm 1.96 \times 0.01567 \approx (0.064, 0.126).$$

```
> summary(ff.lin)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.05771    0.03927  -1.469    0.142
floor        0.09491    0.01567   6.057 1.38e-09 ***
```

R command command `confint.default()` gives the Wald CIs.

```
> confint.default(ff.lin, level=0.95)
                  2.5 %     97.5 %
(Intercept) -0.13468514 0.01926237
floor        0.06419697 0.12561455
```


Likelihood Ratio Tests

To test H_0: $\beta = 0$ vs H_a: $\beta \neq 0$

$$\ell_0 = \text{ max. likelihood when } \beta = 0,$$
$$\ell_1 = \text{ max. likelihood over all possible } \beta$$

The likelihood ratio test statistic is

$$\begin{aligned} LRT &= -2\log(\ell_0/\ell_1) \\ &= -2\left[\log(\ell_0) - \log(\ell_1)\right] \\ &= -2(L_0 - L_1) \sim \chi^2_1 \quad \text{when sample size is large} \end{aligned}$$

where $L_i = \log(\ell_i)$.

Example (Fatal Falls) . For identity link $\pi(x) = \alpha + \beta x$,

- under H_0: $\beta = 0$, $\pi(x) = \alpha$, $L_0 = -117.9112$
- under H_a: $\beta \neq 0$, $\pi(x) = \alpha + \beta x$, $L_1 = -102.4135$

$$LRT = -2(L_0 - L_1) = 30.995, \quad df = 1,$$
$$P\text{-value} = 2.6 \times 10^{-8}$$

Likelihood Ratio Tests in R

The `drop1()` command in R can do LR tests for coefficients.

```
> drop1(ff.lin, test="Chisq")
Single term deletions

Model:
cbind(fatal, live) ~ floor
       Df Deviance    AIC    LRT  Pr(>Chi)
<none>      11.037 35.959
floor   1   42.032 64.955 30.995 2.586e-08 ***
```


Remarks

- For very large n, Wald and LR tests are approx. equivalent, but for small to moderate n, the LR test is more reliable and powerful.
- R does not report the maximized log-likelihood of a model but report its "*deviance*", which we will introduce in Section 3.4.3 and 5.2.2. Now just keep in mind that

$$\text{Deviance} = -2(\text{max. log-likelihood}) + \textit{constant}$$

  where the constant just depends on data but not the model. Thus

$$\text{LR statistic} = \text{difference in deviances}$$

LR Confidence Intervals

LR method also extends to CIs:
$(1-\alpha)100\%$ CI is set of β^* for which P-value $> \alpha$ in LR test of H_0: $\beta = \beta^*$. Computed by `confint()` function in R.

```
> confint(ff.lin, level=0.95)          # LR confidence intervals
Waiting for profiling to be done...
                 2.5 %     97.5 %
(Intercept) -0.10588159 0.02906596
floor        0.06389293 0.12235797
There were 11 warnings (use warnings() to see them)
```

compared with Wald confidence intervals

```
> confint.default(ff.lin, level=0.95)  # Wald confidence intervals
                 2.5 %     97.5 %
(Intercept) -0.13468514 0.01926237
floor        0.06419697 0.12561455
```


We defer the following sections now and will come back to them later.

- Section 3.3 introduces GLMs for count data (rather than binary or binomial data), including *Poisson regression* and *negative binomial regression*. We will come back to this section when we reach Chapter 7.
- Section 3.4.3-3.4.5 introduce deviance, model comparisons, and residuals generally for all GLMs. We will discuss deviance and residuals for binary response models in Chapter 5 and how to use them to do model selection.