1.3 Likelihoods for Regression Models

We will start with linear regression and then talk about more general models.

↓
nonlinear
GLM

1.3.1 Linear Model

Consider the familiar linear model

$$Y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \epsilon_i, \qquad i = 1, \ldots, n,$$

where $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are known nonrandom vectors.

$E[\varepsilon_i] = 0$ and $Var[\varepsilon_i] = \sigma^2$

often estimate β by $\hat{\beta}_{OLS}$, which does not require a distribution for ε_i.

For likelihood-based estimation, we need a distribution for ε_i! Start w/ $\varepsilon_i \sim N(0, \sigma^2)$.

$$\Rightarrow L(\boldsymbol{\beta}, \sigma | \{Y_i, \boldsymbol{x}_i\}_{i=1}^n) = \prod_{i=1}^n \left(\frac{1}{\sqrt{2\pi}\,\sigma}\right) \exp\left(-\frac{(Y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2}{2\sigma^2}\right)$$

$$= \left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^n \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (Y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2\right)$$

take log, derivatives, set $= 0$, solve $\longrightarrow$

$\hat{\boldsymbol{\beta}}_{MLE} = (X^\top X)^{-1} X^\top \boldsymbol{Y}$ same as $\hat{\beta}_{OLS}$!

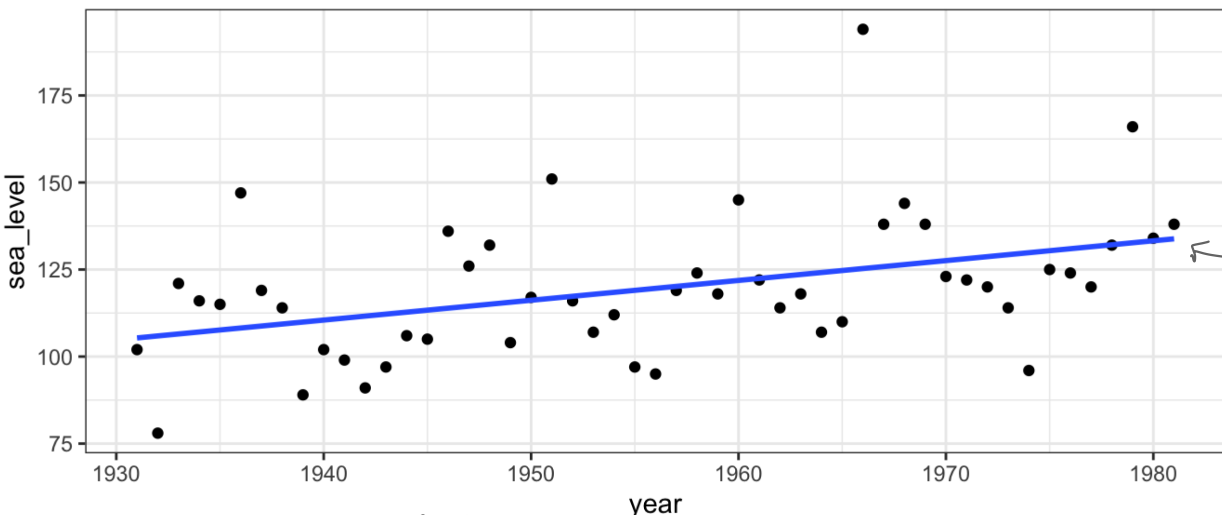
$\hat{\sigma}^2_{MLE} = \frac{1}{n} \sum_{i=1}^n (Y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2$ (only asymptotically unbiased).

What do you do when ϵ_i are not Gaussian?

- transform data so ε_i look Gaussian.
- Use a different distribution for ε_i!

Example (Venice sea levels): The annual maximum sea levels in Venice for 1931−1981 are :

We know maxima are not Gaussian!



Approach 1: OLS $E[\varepsilon_i]=0$, $Var[\varepsilon_i]=\sigma^2$ No distributional assumption.

Approach 2: Assume $\varepsilon_i \sim$ Gumbel (extreme value dsn), use ML

$$f(y;\sigma) = \frac{1}{\sigma}\exp\left(-\frac{y}{\sigma}\right)\exp\left(-\exp\left(-\frac{y}{\sigma}\right)\right).$$

$$\Rightarrow L(\beta,\sigma \mid \{y_i, \underline{x}_i\}_{i=1}^n) = \prod_{i=1}^n f(y_i - \underline{x}_i^T\beta) = \prod_{i=1}^n \frac{1}{\sigma}\exp\left(-\frac{y_i - \underline{x}_i^T\beta}{\sigma}\right)\exp\left(-\exp\left(-\frac{y_i-\underline{x}_i^T\beta}{\sigma}\right)\right)$$

YOUR TURN: Fit both approaches to the venice data.

OLS

$\hat\beta_0 = 104.8$ $\hat\beta_1 = .56$ (SE .177)

MLE, GUMBEL

$\hat\beta_0 = 96.8$, $\hat\beta_1 = 0.56$ (SE .136)

OLS vs MLE? If EV model is correct, more efficient (note: standard errors).

β_0 difference: $\overset{\text{Gumbel}}{E}[\varepsilon_i] = 0.577\sigma = 0.577\hat\sigma_{MLE} = 0.577(14.5) \neq 0$

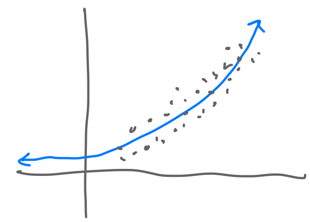
$96.8 + .577(14.5) = 105.1$

1.3.2 Additive Errors Nonlinear Model

Previous example had (1) linear trend, (2) Non-Gaussian errors.

Example: exponential growth model

$g(x, \underline{\beta}) = \beta_0 \exp(\beta_1 x)$



Non-linear additive model:

$$Y_i = g(\underline{x}_i, \underline{\beta}) + \varepsilon_i$$

Often interested in $\varepsilon_i \sim N(0, \sigma^2)$ but $g(\underline{x}_i, \underline{\beta}) \neq \underline{x}_i^T \underline{\beta} \Rightarrow$ ML required.

(1) non-linear trend, (2) Gaussian errors.

1.3.3 Generalized Linear Models

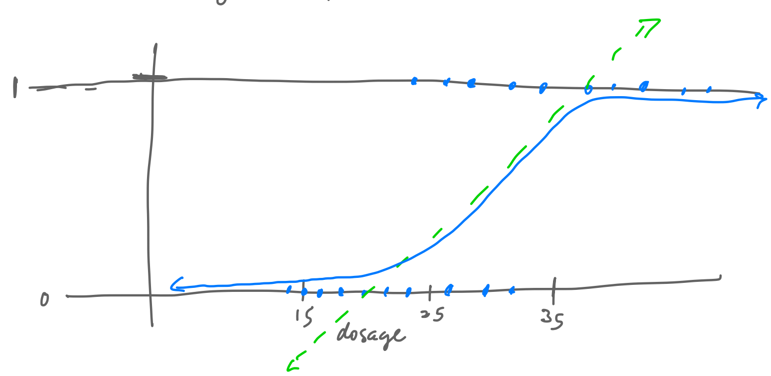
Regression: build a relationship between a parameter (mean) of response & covariates.

LM's: stochastic element is additive w/ mean.

GLM's: stochastic element is different.

Imagine an experiment where individual mosquitos are given some dosage of pesticide. The response is whether the mosquito $\underbrace{\text{lives}}_{=0}$ or $\underbrace{\text{dies}}_{=1}$. The data might look something like:

x (dosage)	y (0 = lives, 1 = dies)
15	0
17	0
18	1
20	0
21	1
⋮	⋮



Goal: Model the relationship between the predictor and response.

dosage dead/alive.

Sounds like regression!

Big difference: Y_i's are not continuous. They only take values of 0 or 1.

Question: What would a curve of best fit look like? Would we want a function that only takes values in $\{0, 1\}$.

It seems sensible to have a curve which takes values near 0 for low doses & near 1 for high doses and intermediate values for middle doses.

what does this curve represent? Probability.

Refined Goal: Model relationship betwen predictor (dosage) + probability of success in response (mosquito dies).

Let's build a sensible model. Note: We don't observe the probability.

Step 1: Find a function that behaves the way we want.

like the blue curve.

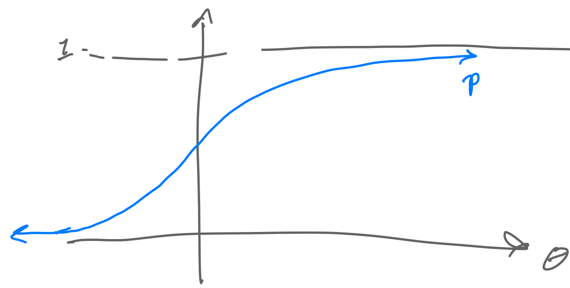
Consider the logistic function,

$$p = \frac{\exp(\theta)}{1+\exp(\theta)}.$$

As $\theta \to \infty$, $p \to 1$

$\theta \to -\infty$, $p \to 0$

$\theta = 0$, $p = \frac{1}{2}$

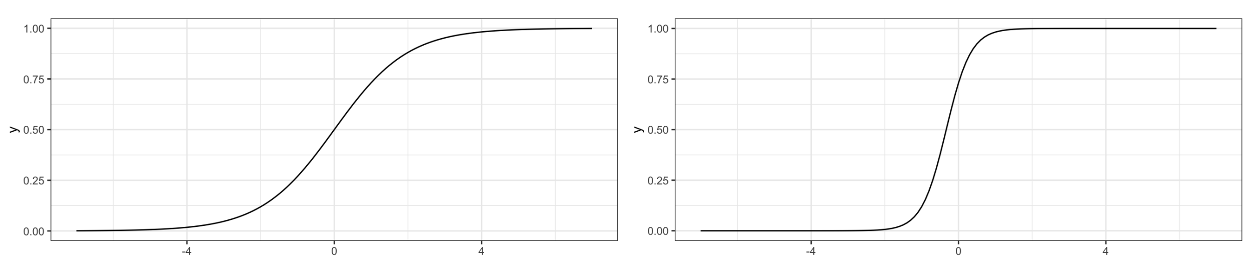


By changing θ, we can change location, slope, direction of this function.

Let $\theta = \beta_0 + \beta_1 x \Rightarrow p = \dfrac{\exp(\beta_0 + \beta_1 x)}{1+\exp(\beta_0+\beta_1 x)}$.

```r
# understanding the logistic function
# first, theta just equals x
x <- seq(-7, 7, .1)
theta <- x
y <- exp(theta)/(1 + exp(theta))
ggplot() + geom_line(aes(x, y))

# now, let theta be a linear function of x
theta <- 1 + 3*x
y <- exp(theta)/(1 + exp(theta))
ggplot() + geom_line(aes(x, y))
```



Now we can connect probabilities to covariate x!

We'd be done if we observed probabilities, but our response only takes values of 0 and 1.

Step 2: Build a stochastic mechanism to relate to a binary response.

Recall the Bernoulli distribution

$$Y = \begin{cases} 0 & \text{w.p. } 1-p \\ 1 & \text{w.p. } p. \end{cases}$$

Coin flip example w/ $p = 0.75$. Flip coin, you will observe 0 (tails) or 1 (heads).
biased

Aside: We could instead think about binomial dsn, which counts # of successes for n trials.

$X = \sum_{i=1}^{n} Y_i$, $Y_i \overset{iid}{\sim} \text{Bern}(p)$. X takes values in $\{0, 1, \dots, n\}$.

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

Step 3: Put Step 1 and Step 2 together.

$$Y_i \overset{ind}{\sim} \text{Bernoulli}(p_i) \quad + \quad p_i = \frac{\exp(\theta_i)}{1 + \exp(\theta_i)}, \quad \theta_i = \beta_0 + \beta_1 x_i$$

outcome of i^{th} observation (observed).

prob of i^{th} observation having success (unobserved).

Goal: estimate β_0 and β_1. Find the "best" estimates.

Fitting our model: Does OLS make sense? No.

What else can we do? Maximum likelihood!

↳ Find the parameters (βs) which make the density agree best w/ data we observed!

pmf of Bernoulli: $f(y_i; p_i) = p_i^{y_i} (1-p_i)^{1-y_i}$

> take y_i's to estimate p_i's.

Consider the likelihood contribution.

$$L_i(p_i|Y_i) = p_i^{Y_i}(1-p_i)^{1-Y_i} \qquad (Y_i\text{'s are } 0 \text{ or } 1).$$

So the log-likelihood contribution is

$$\ell_i(p_i) = Y_i \log p_i + (1-Y_i)\log(1-p_i) = \underbrace{\log(1-p_i) + Y_i \log\frac{p_i}{1-p_i}}_{(*)}$$

Recall, we said $\underbrace{p_i = \frac{\exp(\theta_i)}{1+\exp(\theta_i)}}$ was sensible.

Manipulating: $p_i + p_i\exp(\theta_i) = \exp(\theta_i)$

$$p_i = (1-p_i)\exp(\theta_i).$$

OR

$$\frac{p_i}{1-p_i} = \exp(\theta_i)$$

$$\log\left(\frac{p_i}{1-p_i}\right) = \theta_i \quad (1)$$

$$\underline{p_i}\exp(-\theta_i) = 1-p_i$$

$$\frac{\exp(\theta_i)}{1+\exp(\theta_i)}\exp(-\theta_i) = 1-p_i$$

$$\frac{1}{1+\exp(\theta_i)} = 1-p_i$$

$$-\log(1+\exp(\theta_i)) = \log(1-p_i) \quad (2).$$

Pluggin in (1) and (2) into (*).

Which gives us,

$$\ell_i(\theta_i) = -\log(1+\exp(\theta_i)) + Y_i\theta_i \qquad (\text{now in terms of } \theta_i \text{ not } p_i)$$

notice now the term w/ the data is "nice" for MLE things.
Why? Because log-likelihood + "sensible" function $p_i = \frac{\exp(\theta_i)}{1+\exp(\theta_i)}$
work well together.

<u>NOT</u> a coincidence.

So the log-likelihood is

$$\ell(\theta_1,\ldots,\theta_n) = \sum_{i=1}^{n} \ell_i(\theta_i)$$

$$= \sum_{i=1}^{n}\left\{-\log(1+\exp(\theta_i)) + Y_i\theta_i\right\}$$

$$\Rightarrow \ell(\beta_0,\beta_1) = \sum_{i=1}^{n}\left\{-\log(1+\exp(\beta_0+\beta_1 x_i)) + Y_i(\beta_0+\beta_1 x_i)\right\}$$

To optimize? Must be done numerically.

```r
## data on credit default
data("Default", package = "ISLR")
head(Default)
```

```
##   default student   balance    income
## 1      No      No  729.5265 44361.625
## 2      No     Yes  817.1804 12106.135
## 3      No      No 1073.5492 31767.139
## 4      No      No  529.2506 35704.494
## 5      No      No  785.6559 38463.496
## 6      No     Yes  919.5885  7491.559
```

optimizing likelihood numerically.

```r
## fit model with ML
m0 <- glm(default ~ balance, data = Default, family = binomial)
tidy(m0) |> kable()
```

data y's are 0, 1

broom package.

$\hat{\beta}_0, \hat{\beta}_1$ $se(\hat{\beta}_0), se(\hat{\beta}_1)$ $\frac{\hat{\beta}_i}{se(\hat{\beta}_i)}$.

term	estimate	std.error	statistic	p.value
(Intercept)	-10.6513306	0.3611574	-29.49221	0
balance	0.0054989	0.0002204	24.95309	0

$H_0: \beta_i = 0$ $i=1,2.$
$H_a: \beta_i \neq 0$

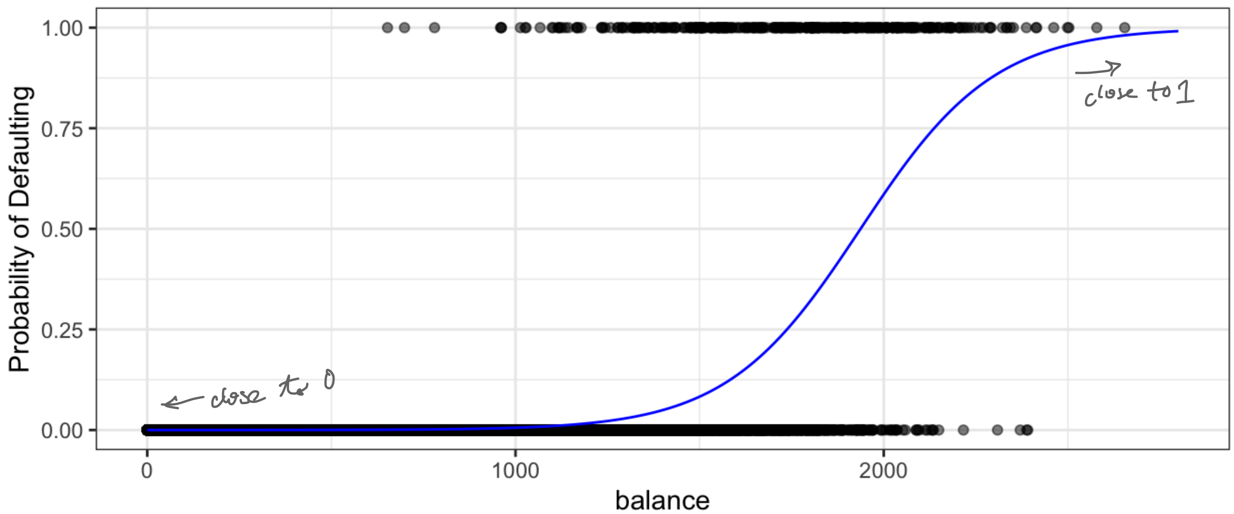
```r
glance(m0) |> kable()
```

null.deviance	df.null	logLik	AIC	BIC	deviance	df.residual	nobs
2920.65	9999	-798.2258	1600.452	1614.872	1596.452	9998	10000

$\ell(\hat{\beta}_0, \hat{\beta}_1)$.

```r
## plot the curve
x_new <- seq(0, 2800, length.out = 200)
theta <- m0$coefficients[1] + m0$coefficients[2]*x_new
p_hat <- exp(theta)/(1 + exp(theta))

ggplot() +
  geom_point(aes(balance, as.numeric(default) - 1), alpha = 0.5, data
        = Default) +
  geom_line(aes(x_new, p_hat), colour = "blue") +
  ylab("Probability of Defaulting")
```



never outside of [0,1] → valid probabilities!

In general, a GLM is three pieces:

1. The random component

   probability dsn from exponential family.

   Ex: Logistic Regression

   $Y_i \sim \text{Binom}(p_i)$

   Explanation: decribe the generating mechanism of observed data

2. The systemic component

   link function.

   A function relating the parameter of interest (mean!) to θ

   $E[Y] = g^{-1}(\eta)$

   $p_i = \dfrac{\exp(\theta_i)}{1+\exp(\theta_i)} = g^{-1}(\theta_i)$.

   $\uparrow$

   Note $Y_i \sim \text{Bern}(p_i)$

   $E[Y_i] = p_i$

   transforming linear relationship to be on a scale that makes sense for the parameter of interest

   "linking" linear relationship to mean.

3. A linear predictor

   $\theta = X\beta$.

   $\theta_i = x_i \beta$

   describing how θ is a linear function of predictor variables.

Remarks:

(1) Standard formulation denotes function by g^{-1}: $p = g^{-1}(\theta) = \dfrac{\exp(\theta)}{1+\exp(\theta)}$.

$$\Rightarrow \theta = g(p) = \log\left(\frac{p}{1-p}\right).$$

(2) Parameter of interest is still the mean, just like linear regression.

(3). Theoretical reasons for exponential family ... relationship btw/ param of interest & Variance.

for Count data.

Example (Poisson regression):

(1) Poisson (λ).

(3) $\theta = X\beta$

(2) $\lambda = g^{-1}(\theta) = \exp(\theta)$ $(\lambda > 0)$.

$\Downarrow$

$\theta = \log(\lambda)$.

Some background on choice of g.

Consider a general family of distributions:

subfamily of exponential family dsns that includes Binomial, Poisson, etc.

$$\log f(y_i;\theta_i,\phi) = \frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i,\phi).$$

$$f(y_i;\theta_i,\phi) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i,\phi)\right\}$$

$$= \exp\left\{\underbrace{\frac{y_i\theta_i}{a_i(\phi)} + c(y_i;\phi)}_{*} - \underbrace{\frac{b(\theta_i)}{a_i(\phi)}}_{**}\right\}.$$

recall exponential family w/ parameter $\underline{\theta} = (\theta_1,\dots,\theta_s)^T$ is of the form:

$$f(y;\underline{\theta}) = h(y)\exp\left\{\underbrace{\sum_{j=1}^{s} g_j(\underline{\theta})T_j(y)}_{*} - \underbrace{B(\underline{\theta})}_{**}\right\}$$

assumes $T_1(y_i) = y_i$ subfamily of exponential family.

$g_1(\underline{\theta}) = \frac{\theta_i}{a_i(\phi_i)}$ similar to single param exp family except dispersion term $a_i(\phi)$.

Example (Normal model):

We can learn something about this distribution by considering it's mean and variance. Because we don't have an explicit form of the density, we rely on two facts:

1. $\mathrm{E}\left[\frac{\partial \log f(Y_i;\theta_i,\phi)}{\partial\theta_i}\right] = 0.$

2. $\mathrm{E}\left[\frac{\partial^2 \log f(Y_i;\theta_i,\phi)}{\partial\theta_i^2}\right] + \mathrm{E}\left[\left(\frac{\partial \log f(Y_i;\theta_i,\phi)}{\partial\theta_i}\right)^2\right] = 0.$

For $\log f(y_i;\theta_i,\phi) = \frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i,\phi)$,

Example (Bernoulli model):

$$f(y_i;p_i) = p_i^{y_i}(1-p_i)^{1-y_i}$$

Finally, back to modelling. Our **goal** is to build a relationship between the mean of Y_i and covariates $\boldsymbol{x}_i$.

Example (Bernoulli model, cont'd):

1.4 Marginal and Conditional Likelihoods

Consider a model which has $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$, where $\boldsymbol{\theta}_1$ are the parameters of interest and $\boldsymbol{\theta}_2$ are nuisance parameters.

One way to improve estimation for $\boldsymbol{\theta}_1$ is to find a one-to-one transformation of the data $\boldsymbol{Y}$ to $(\boldsymbol{V}, \boldsymbol{W})$ such that either

The key feature is that one component of each contains only the parameter of interest.

Example (Neyman-Scott problem): Let $Y_{ij}, i = 1, \ldots, n, j = 1, 2$ be intependent normal random variables with possible different means μ_i but the same variance σ^2.

Our goal is to estimate σ^2. Should we be able to?

Following the usual arguments,

$$
\begin{aligned}
\hat{\mu}_{i,\text{MLE}} &= \frac{Y_{i1} + Y_{i2}}{2} \\
\hat{\sigma}^2_{\text{MLE}} &= \frac{1}{2n} \sum_{i=1}^{n} \sum_{j=1}^{2} (Y_{ij} - \hat{\mu}_{i,\text{MLE}})^2
\end{aligned}
$$

$$\mathrm{E}[\hat{\sigma}^2_{\mathrm{MLE}}] =$$

A reworking of the data seems more promising. Let,

$$V_i = \frac{Y_{i1} - Y_{i2}}{\sqrt{2}} \qquad \text{and} \qquad W_i = \frac{Y_{i1} + Y_{i2}}{\sqrt{2}}$$

For conditional likelihoods, we can often exploit the existence of sufficient statistics for the nuisance parameters under the assumption that the parameter of interest is known.

Example (Exponential Families): The structure of exponential families is such that it is often possible to exploit their properties to eliminated nuisance parameters. Let Y have a density of the form

$$f(y;\boldsymbol{\eta}) = h(y)\exp\left\{\sum_{i=1}^{s}\eta_i T_i(y) - A(\boldsymbol{\eta})\right\},$$

then

Thus, exponential families often provide an automatic procedure for finding $\boldsymbol{W}$ and $\boldsymbol{U}$.

Example (Logistic Regression): For binary Y_i, the standard logistics regression model is

$$P(Y_i = 1) = p_i(\boldsymbol{x}_i, \boldsymbol{\beta}) = \frac{\exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})}$$

and the likelihood is

$$L(\boldsymbol{\beta} | \boldsymbol{Y}, \boldsymbol{X}) =$$