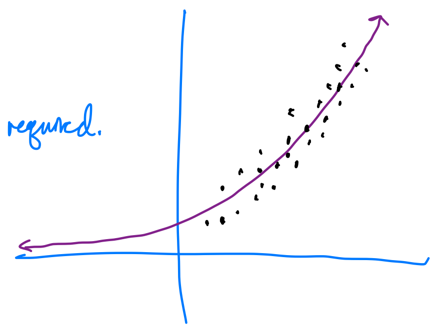
1.3.2 Additive Errors Nonlinear Model

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

Example: exponential growth

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



Nonlinear additive model:

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

usually $\varepsilon_i \sim N(0, \sigma^2)$ but $g(\underline{X}_i, \underline{\beta}) \neq \underline{x}_i^T \underline{\beta}$ $\Rightarrow$ max. likelihood required.

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

1.3.3 Generalized Linear Models

Regression: build a relationship w/ parameter (mean) + covariates.

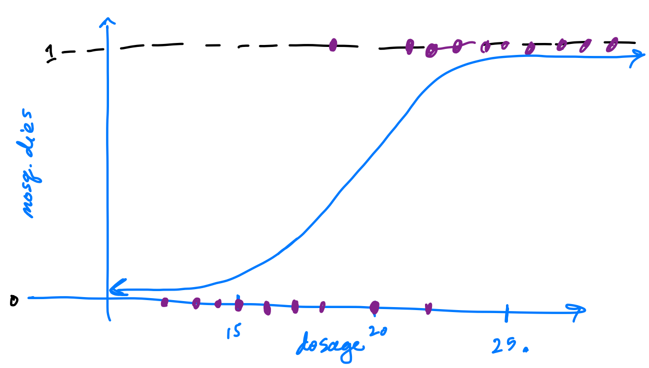
LM: stochastic element is additive w/ mean.

GLM: stochastic element be different.

Imagine an experiment where individual mosquitos are given some dosage of pesticide. The response is whether the mosquito lives or dies. The data might look something like:

=0 =1

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 (binary response).

Question: What would a curve of best fit look like? Would it only take values of 0 or 1?

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

What does this curve represent? Probability.

Refined Goal: Model relationship between predictor (dosage) & probability (of mosq. dies).

Let's build a sensible model. Note: We do not observe the probability!

If I flip a possibly biased coin & 3 heads do you know $P(\text{heads}) = .75$?

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

↳ like the blue curve.

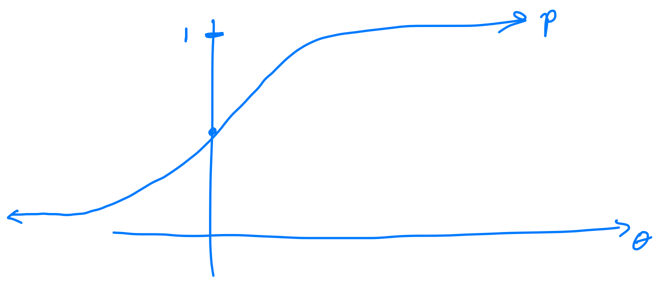
Consider the logistic function

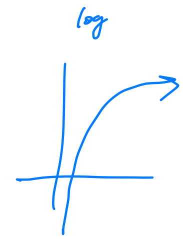
$$p = \frac{\exp(\theta)}{1+\exp(\theta)}$$

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

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

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



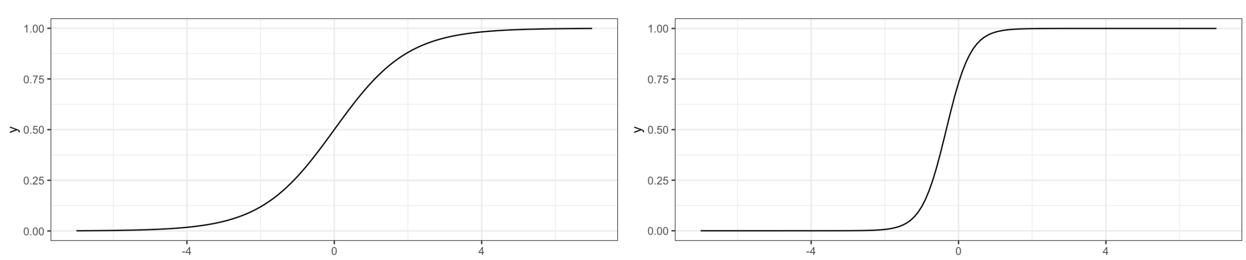


By changing θ, we can change location, slope, direction of this function.

Let $\theta = \beta_0 + \beta_1 x \implies p = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}$

```
# 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 have the ability to connect probabilities to covariate x!

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

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

Recall the Bernoulli distribution:

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

Back to biased coin flip example w/ $p = 0.75$. Flip once, you will get 0 (tails) or 1 (heads).

Aside: You may be more familiar w/ Binomial distribution, which counts # of successes for n trials.

X takes values in $\{0, 1, \ldots, n\}$

$X = \sum_{i=1}^{n} Y_i$, $Y_i \overset{iid}{\sim}$ Bernoulli (p).

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

Step 3: Put Step 1 and Step 2 together.

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

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

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

& $p_i = \dfrac{\exp(\theta_i)}{1+\exp(\theta_i)}$, $\theta_i = \beta_0 + \beta_1 x_i$

OR

$$p_i = \frac{\exp(\beta_0 + \beta_1 x_i)}{1+\exp(\beta_0 + \beta_1 x_i)}$$

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

Fitting our model: Does OLS make sense? NO.

What else can we do? Maximum likelihood!

↳ Find the parameters (β's) which make the density agree best w/ data!

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.

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

So the log-likelihood contribution is

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

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

Manipulating:

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

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

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

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

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

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

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

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

Plugging (1) & (2) into (*)

Which gives us,

$$\ell_i(\theta_i) = -\log(1+\exp(\theta_i)) + \underbrace{y_i\theta_i} \quad (\text{now in terms of } \theta_i \text{ not } p_i).$$

Notice now the term w/ data is "nice" for MLE things.

Not a coincidence! Because the "sensible function" $p_i = \frac{\exp(\theta_i)}{1+\exp(\theta_i)}$ works well together w/ log-likelihood.

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
```

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

↑ optimizing likelihood numerically.

$\hat\beta_{0,MLE}$ $\hat\beta_{1,MLE}$ $sd(\hat\beta_0), sd(\hat\beta_1)$

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

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

$\ell(\hat\beta_0, \hat\beta_1)$

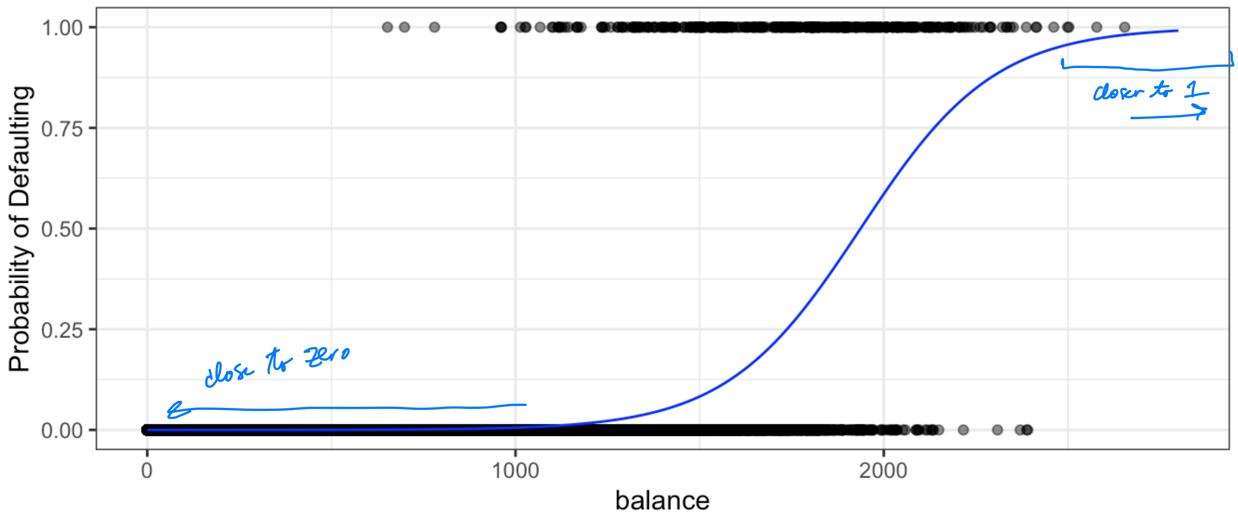
```r
## plot the curve
x_new <- seq(0, 2800, length.out = 200)
theta <- m0$coefficients[1] + m0$coefficients[2]*x_new
p_hat <- exp(theta)/(1 + exp(theta))
ggplot() +
  geom_point(aes(balance, as.numeric(default) - 1), alpha =
0.5, data = Default) +
```

$\hat\beta_0 + \hat\beta_1 x$

estimated probabilities

change "Yes", "No" to 1, 0

```
           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:

	Ex: logistic regression	Explanation:
1. The random component probability distribution from exponential family	$Y_i \sim \text{Bernoulli}(p_i)$	describes generating mechanism of observed data.
2. The systemic component A link function relating the parameter of interest (mean!) to θ $E[Y] = g^{-1}(\eta)$	$p_i = \frac{\exp(\theta_i)}{1+\exp(\theta_i)} = g^{-1}(\theta_i)$ Note if $Y \sim \text{Ben}(p)$, $E[Y] = p$.	transforms linear relationship to be on a scale that makes sense for parameter of interest "links" linear relationship to mean.
3. A linear predictor $\theta = X\beta$	$\theta = X\beta$	A linear relationship describing how θ is a linear function of predictors.

Remarks:

1. Standard formulation denotes link function by g^{-1}: $p = g^{-1}(\theta) = \frac{\exp(\theta)}{1+\exp(\theta)}$
   $\Rightarrow \theta = g(p) = \log\left(\frac{p}{1-p}\right)$
2. Parameter of interest is still the mean, just like linear regression.
3. Theoretical reasons for exponential family ... relationship btw/ param of interest & variance.

Example (Poisson regression):

1. $Y_i \sim \text{Poisson}(\lambda_i)$
2. $\lambda_i = \underbrace{\exp(\theta_i)}_{}$ ↙range is ≥0 $= g^{-1}(\theta_i)$
   $\theta_i = g(\lambda_i) = \log(\lambda_i)$ ← "log link"
3. $\theta_i = \underline{x}_i \beta$.