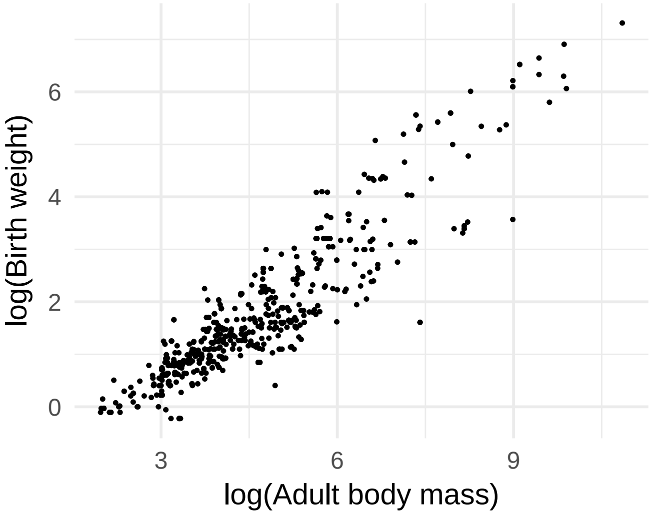
Intro to linear regression

Max Joseph

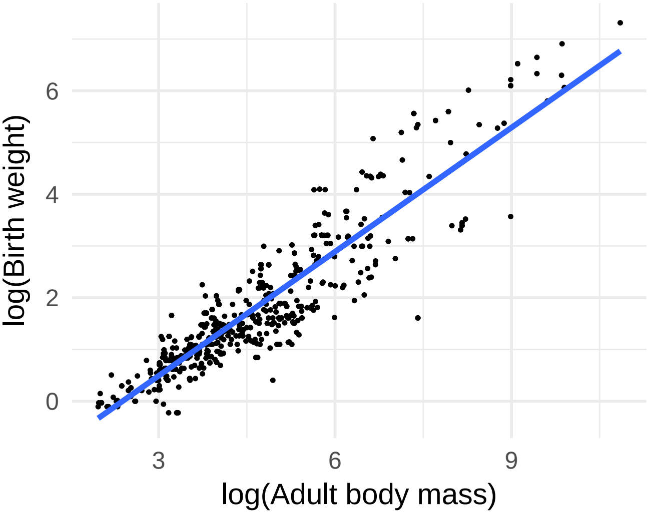
March 07, 2017

Situation

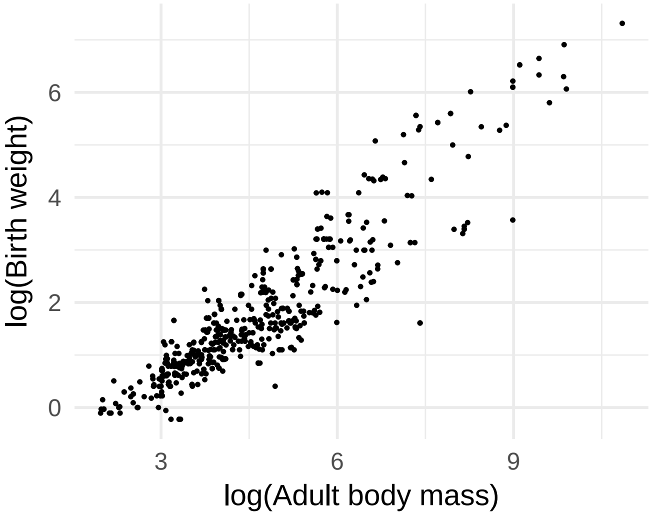
How to predict birth weight?



Spoiler alert



Which line to draw?



Simple linear regression

We have pairs (y_i, x_i) for $i = 1, 2, ..., n$

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

Assumption

1. **Linear relationship between x and y**

In context

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

Deterministic vs. stochastic parts

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

Deterministic

$$\hat{y}_i = \beta_0 + \beta_1 x_i$$

- for any x_i, the value of $\hat{y}_i$ is always the same.

Deterministic vs. stochastic parts

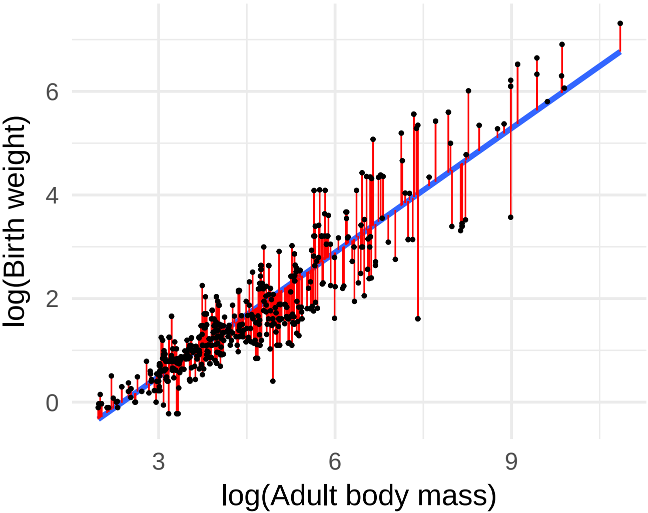
$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

Deterministic

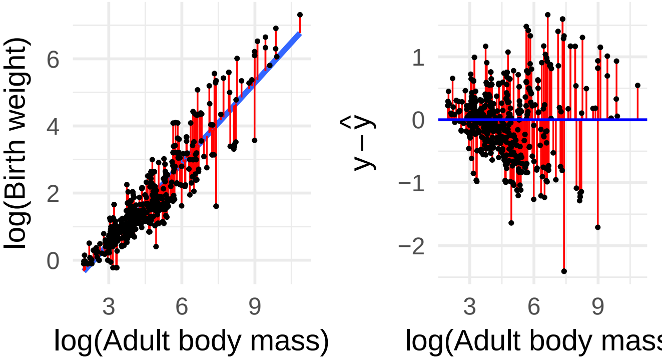
$$\hat{y}_i = \beta_0 + \beta_1 x_i$$

What is stochastic?

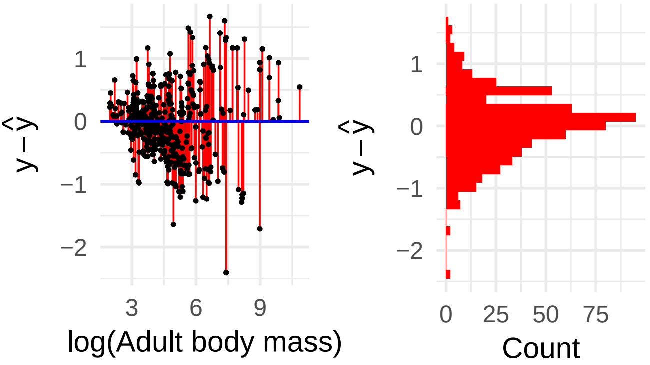
Stochastic error!



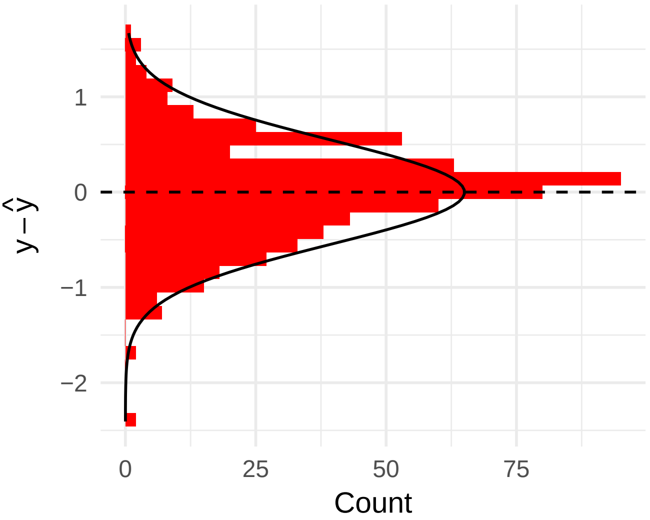
What distribution could work for the errors?



Error histogram



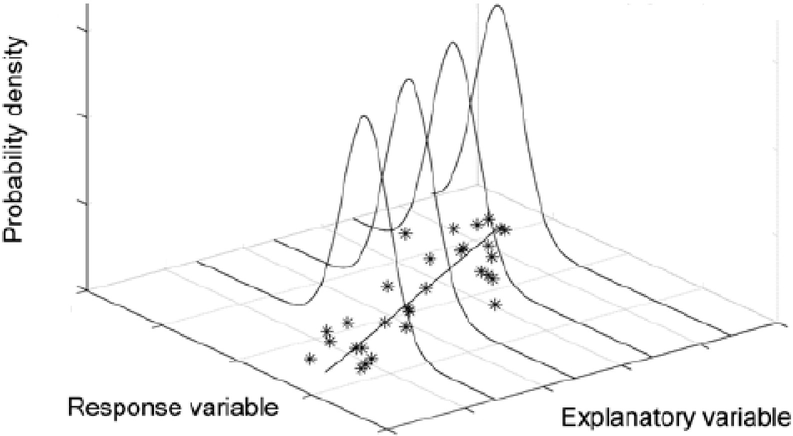
How about Normal error



Assumption

1. Linear relationship between x and y
2. **Normal error (variation around line)**

Another look



http://bolt.mph.ufl.edu/6050-6052/unit-4b/module-15/

Linear regression and normality

Deterministic component

$$\hat{y}_i = \beta_0 + \beta_1 x_i$$

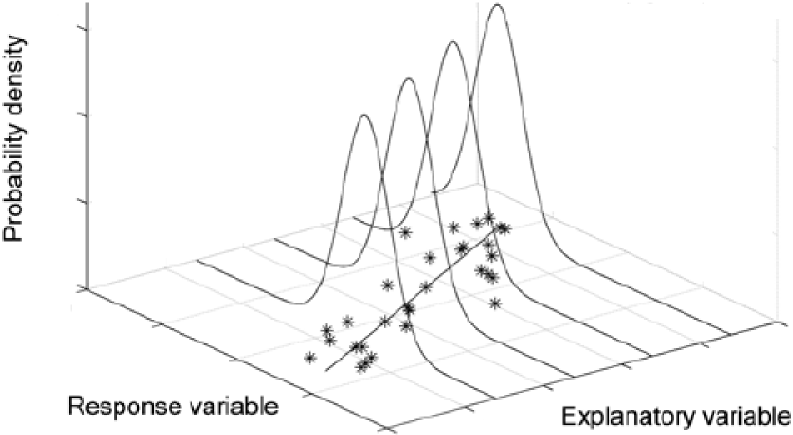
Stochastic component

$$y_i \sim \text{Normal}(\hat{y}_i, \sigma)$$

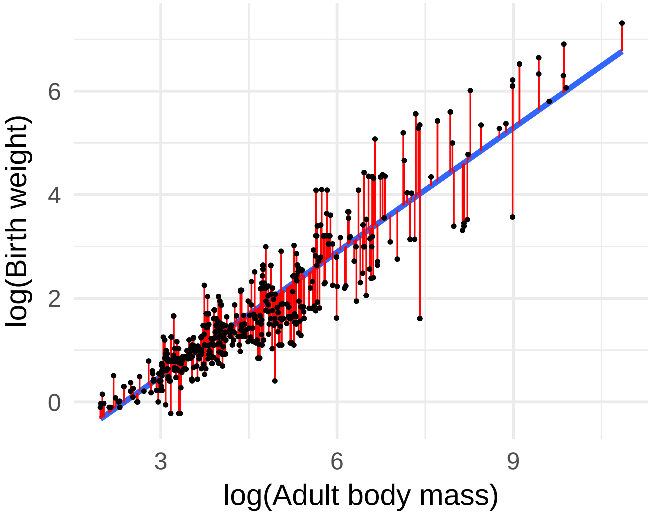
Assumption

1. Linear relationship between x and y
2. Normal error (variation around line)
3. **Homoskedasticity (errors have constant variance)**

What would this plot look like with heteroskedasticity?



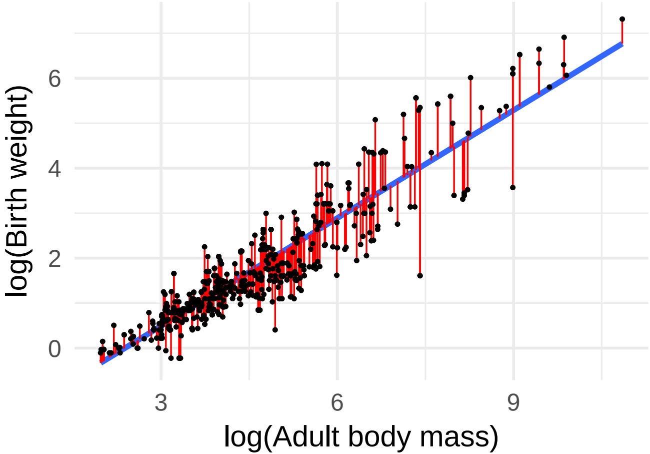
Aside: error in what?



Aside: error in what?

Error is **orthogonal** to x

- x is fixed and known without error
- **no assumptions** about the normality of x



Maximum likelihood estimation for linear regression

Goal

Maximize $L(y; \beta_0, \beta_1, \sigma)$

MLE for linear regression

Likelihood is the joint probability of observations $y_1, y_2, ..., y_n$:

$$L(y; \beta_0, \beta_1, \sigma) = p(y_1, y_2, ..., y_n; \beta_0, \beta_1, \sigma)$$

MLE for linear regression

Likelihood is the joint probability of observations $y_1, y_2, ..., y_n$:

$$L(y; \beta_0, \beta_1, \sigma) = p(y_1, y_2, ..., y_n; \beta_0, \beta_1, \sigma)$$

Observations are independent, so we multiply probabilities:

$$L(y; \beta_0, \beta_1, \sigma) = p(y_1; \beta_0, \beta_1, \sigma) \times p(y_2; \beta_0, \beta_1, \sigma) \times ... \times p(y_n; \beta_0, \beta_1, \sigma)$$

Assumption

1. Linear relationship between x and y
2. Normal error (variation around line)
3. Homoskedasticity (errors have constant variance)
4. **Observations are independent**

MLE for linear regression

Likelihood is the joint probability of observations $y_1, y_2, ..., y_n$:

$$L(y; \beta_0, \beta_1, \sigma) = p(y_1, y_2, ..., y_n; \beta_0, \beta_1, \sigma)$$

Observations are independent, so we multiply probabilities:

$$L(y; \beta_0, \beta_1, \sigma) = p(y_1; \beta_0, \beta_1, \sigma) \times p(y_2; \beta_0, \beta_1, \sigma) \times ... \times p(y_n; \beta_0, \beta_1, \sigma)$$

We are lazy, so we use product notation:

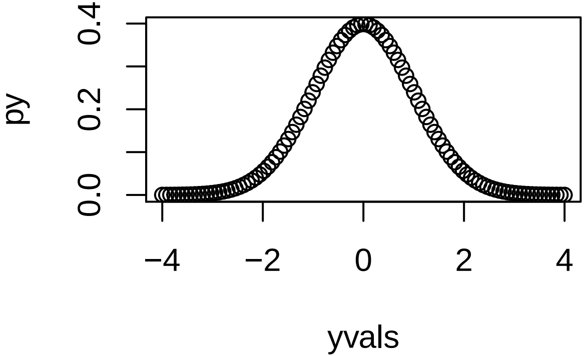
$$L(y; \beta_0, \beta_1, \sigma) = \prod_{i=1}^{n} p(y_i; \beta_0, \beta_1, \sigma)$$

But how to we get $p(y_i; \beta_0, \beta_1, \sigma)$?

$$L(y; \beta_0, \beta_1, \sigma) = \prod_{i=1}^{n} p(y_i; \beta_0, \beta_1, \sigma)$$

Getting normal probability densities in R

```
yvals <- seq(-4, 4, length.out = 100)
py <- dnorm(yvals, mean = 0, sd = 1)
plot(yvals, py)
```



Getting normal probability densities for linear regression

$$\hat{y}_i = \beta_0 + \beta_1 x_i$$

$$y_i \sim \text{Normal}(\hat{y}_i, \sigma)$$

```
y_hat <- beta0 + beta1 * x
dnorm(y, mean = y_hat, sd = sigma)
```

MLE for linear regression

$$L(y; \beta_0, \beta_1, \sigma) = \prod_{i=1}^{n} p(y_i; \beta_0, \beta_1, \sigma)$$

In R:

```
prod(dnorm(y, mean = yhat, sd = sigma))
```

Why can't we work with this directly?

$$L(y; \beta_0, \beta_1, \sigma) = \prod_{i=1}^{n} p(y_i; \beta_0, \beta_1, \sigma)$$

```
prod(dnorm(y, mean = yhat, sd = sigma))
```

Log-likelihood to avoid underflow

$$\log(L(y; \beta_0, \beta_1, \sigma)) = \log(\prod_{i=1}^{n} p(y_i; \beta_0, \beta_1, \sigma))$$

$$= \sum_{i=1}^{n} \log(p(y_i; \beta_0, \beta_1, \sigma))$$

because $\log(ab) = \log(a) + \log(b)$

```
sum(dnorm(y, mean = yhat, sd = sigma, log = TRUE))
```

One last thing...

We typically use the *negative log likelihood*:

$$-\log(L(y;\beta_0,\beta_1,\sigma)) = -\sum_{i=1}^{n}\log(p(y_i;\beta_0,\beta_1,\sigma))$$

Acquiring estimates in R

Specifying a negative log likelihood function

```r
nll <- function(pars, x, y) {
  b0 <- pars[1]
  b1 <- pars[2]
  sigma <- exp(pars[3])
  y_hat <- b0 + b1 * x
  -sum(dnorm(y, y_hat, sigma, log = TRUE))
}
```

Minimizing the negative log likelihood

```
fit <- optim(c(0, 0, 0), nll, x = x, y = y)
fit
```

```
## $par
## [1] -1.9062602  0.7994126 -0.6056837
##
## $value
## [1] 509.9306
##
## $counts
## function gradient
##      168       NA
##
## $convergence
## [1] 0
##
## $message
## NULL
```

Plotting the line

Assumptions

1. Linear relationship between x and y
2. Normal error (variation around line)
3. Homoskedasticity (errors have constant variance)
4. Observations are independent

Questions?

What we covered:

- linear model structure & assumptions
- defining a likelihood
- writing a negative log likelihood function in R
- minimizing the negative log likelihood with `optim()`