Analyzing Data With Regression Models

Professor Diane Lambert

June 2010

Supported by MOE-Microsoft Key Laboratory of Statistics and Information Technology and the Beijing International Center for Mathematical Research, Peking University.

With many thanks to Professor Bin Yu of University of California Berkeley, and Professor Yan Yao and Professor Ming Jiang of Peking University.

Structure of This Course

The Professor (me)

Past:

- Associate Professor of statistics at CMU
  - worked with environmental safety data
- Research Scientist, Bell Labs
  - network data, semiconductor data

Now:

- Research Scientist at Google in New York
  - network data, data on advertisements, advertisers, image search, and more

The Teaching Assistants

Xiaoxing Cheng (Daniel)

Tianxi Li

Jianghan Qu (Jenny)

Bihyuan Zhang (Elise)

Zoey Zhao

The teaching assistants will help you with the labs.

The Students?

Please send an email to

pkusms2010@gmail.com

With

- your name
- your university
- statistics courses you've taken
- what days you have exams this week
- have you used R before? a little? a lot?

What This Course Is About

- Data analysis with regression models
  - Linear regression
  - Logistic regression
  - Multilevel regression
    - what all these models are
    - how to work with them in R
    - how to interpret them

Organization of the Course

Lectures

- I will present the models, with examples
- You can ask questions during class
- Slides and R code used in the lectures will be available after class

Organization of the Course (continued)

Labs

A chance to analyze data in R using the models from lecture.

Assignments will be done in teams of 2 or 3 students.

The teaching assistants and I will help you.

The TAs will explain how to hand in assignments.

Only English will be used in the course.

This is a good opportunity to practice for visiting the US or studying in the US.

Schedule

Lectures 10:00 - 12:00 (with a break)

Labs 14:00 - 16:00

Tentative Plan

		10-12	2-4
June 15, 16	linear regression	√	√
June 17	logistic regression	√	
June 18	logistic regression		√
June 21	Google statistics		√
June 22	logistic regression	√	√
June 23, 24	multilevel models	√	√

Part 1

A regression model answers a question about data.

Basic Parts of A Regression Problem

1. An outcome Y that varies
2. Variables $X_1, X_2, \ldots, X_k$ that affect Y

   These are conditions that may affect Y

   Often called predictors or model terms

3. A question about Y

   what is the average Y under different conditions

   how do different conditions affect Y

   how to predict Y under different conditions

More simply

A regression problem tries to answer a question about an outcome Y with data on Y and $\mathbf{X} = (X_1, \ldots, X_K)$.

An Example Question

What affects the sleep a mammal needs?

Y: sleep (mean hours/day for a species)

X_1: body (mean body weight in kg)

X_2: brain (mean brain weight in g)

X_3: life (max lifetime in years)

X_4: predation (1 = low, 5 = high)

X_5: gestation (days until birth)

fact: Real data analysts give variables real names.

X_1, …, X_k are not real names.

Example Data

	sleep	body	brain	predation	danger
African_elephant	3.3	6654.0	5712.0	3	3
African_giant_pouched_rat	8.3	1.0	6.6	3	3
Arctic_Fox	12.5	3.4	44.5	1	1
Asian_elephant	3.9	2547.0	4603.0	3	4
Baboon	9.8	10.6	179.5	4	4
Big_brown_bat	19.7	0.0	0.3	1	1
Brazilian_tapir	6.2	160.0	169.0	4	4
Cat	14.5	3.3	25.6	1	1
Chimpanzee	9.7	52.2	440.0	1	1
Chinchilla	12.5	0.4	6.4	5	4
Cow	3.9	465.0	423.0	5	5
Donkey	3.1	187.1	419.0	5	5
Eastern_American_mole	8.4	0.1	1.2	1	1
Echidna	8.6	3.0	25.0	2	2

There are 51 species in the data (most not shown here).

A Statistical Regression Model

A statistical regression model consists of

1. a model for how the mean outcome Y changes with $X_1, \ldots, X_K$

$$E(Y \mid X_1, \ldots, X_K) = \mu(X_1, \ldots, X_K)$$

2. a model for how the variability in Y changes with $X_1, \ldots, X_K$

$$\text{var}(Y \mid X_1, \ldots, X_K) = \sigma^2(X_1, \ldots, X_K)$$

A regression model may or may not assume

3. a distribution for $Y \mid X_1, \ldots, X_K$.

Simplest Regression Model For Sleep

Mean

$$E(\text{sleep} \mid \mathbf{X}) = \mu$$

Variance

$$\text{var}(\text{sleep} \mid \mathbf{X}) = \sigma^2$$

Distribution

$$\text{sleep} \sim \text{Normal}(\mu, \sigma^2)$$

X isn't used to predict sleep in this model

Is this a good enough model for sleep across species of mammals?

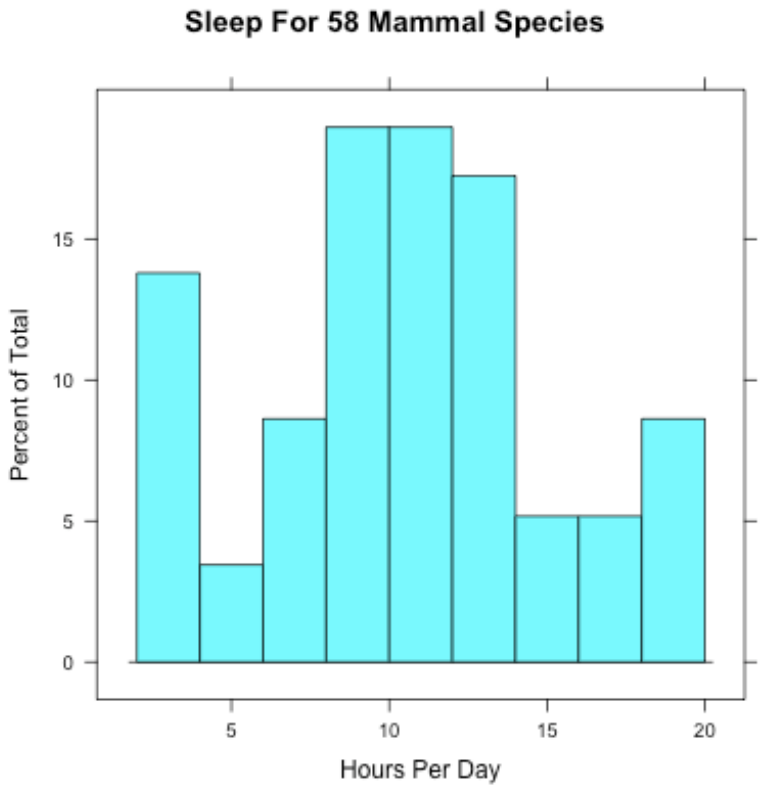
The Simplest Model For Sleep

$E(\text{sleep} \mid \mathbf{X}) = \mu$

$\text{var}(\text{sleep} \mid \mathbf{X}) = \sigma^2$

$\text{sleep} \sim \text{Normal}(\mu, \sigma^2)$

If the model is reasonable, the sleep data should look like a random sample from a normal distribution

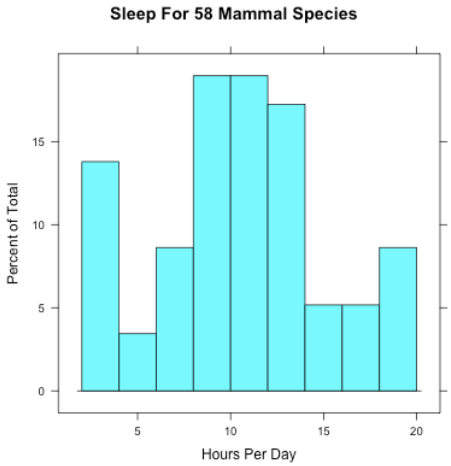


The histogram doesn't look like a normal curve, so the simplest model isn't very good.

To Get Better Models

We could use a different distribution

- we need more mass in the tails
- it is hard to find a distribution like that, and it won't give us much insight into why there is mass in the tails.



Instead, try to model the mean

- we can use the other information (like body weight) about the species

Making The Mean Depend on **X**

Linear regression has two assumptions.

1. Mean outcome Y is linear in the terms in **X**

$$\mu(\mathbf{X}) = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_K$$

2. Variance of the outcome is constant in **X**.

$$\text{var}(Y \mid \mathbf{X}) = \sigma^2$$

Usually, the distribution of Y|**X** is unimportant, as long as Y is numeric and has several possible values.

Y| **X** does not have to be normally distributed.

Back to The Example

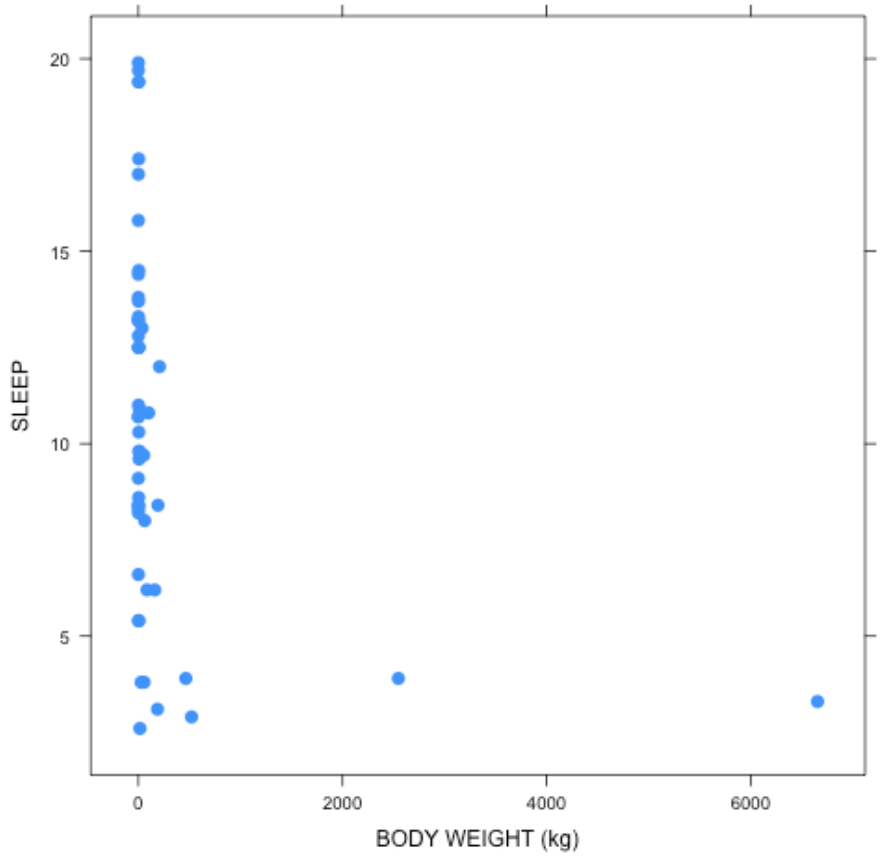
Is the mean sleep of a mammal species linear in body weight?

There is no theory here.

Look at the data. →

Answer:

A line doesn't describe how E(sleep | body) depends on body weight.

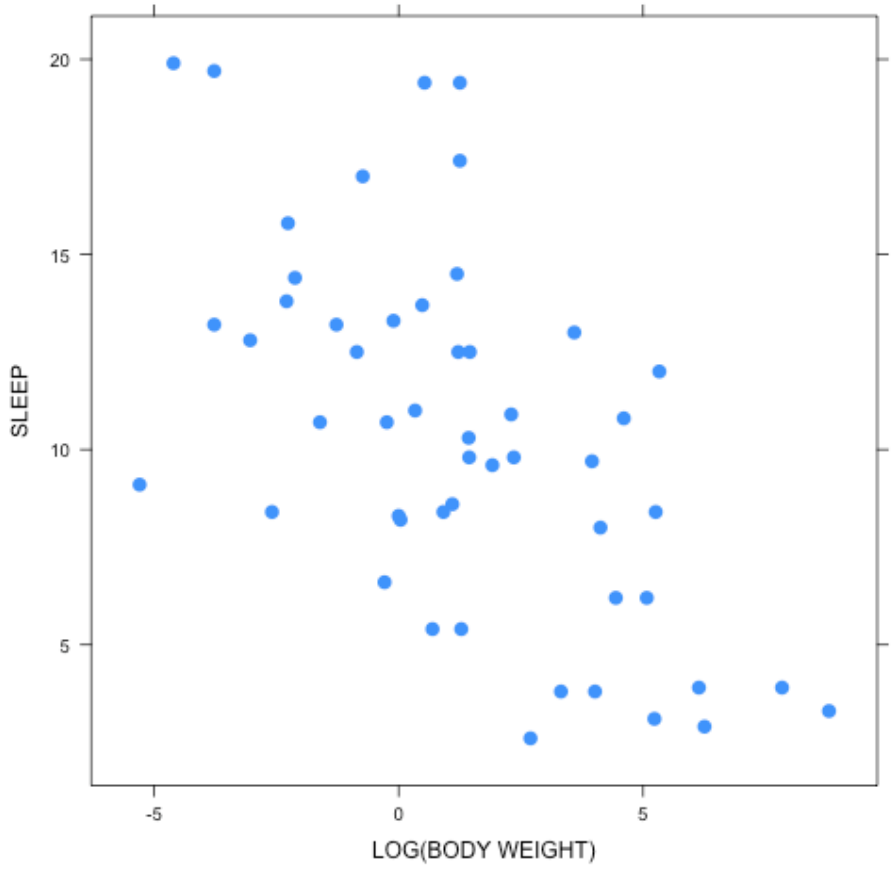


Linear regression of sleep on body weight is not a good model for mammals *even though* b_1 *is statistically significant here*. Looking at the regression coefficients won't tell you if the model fits.

Changing Predictors

In linear regression, the mean is linear in the predictors.

- An X can be transformed, or several can be combined like X_1X_2 and it's still linear regression.
- Body weight covers several orders of magnitude. In such cases, try log(X).



Linear regression of sleep on log(body weight) is sensible.

Part 2

Linear Regression Coefficients

Least Squares Coefficient Estimates

With estimated coefficients $b_0, \ldots, b_K$, the residuals or prediction errors are

$$e_i = Y_i - b_0 - b_1 X_1 - \ldots - b_k X_K$$

The least squares coefficients $b_0, \ldots, b_K$ minimize the sum of squared errors

$$\sum_i \left(Y_i - b_0 - b_1 X_1 - \ldots - b_k X_K\right)^2$$

FACT: The b_k's are also maximum likelihood estimates when $Y|(X_0, \ldots, X_K)$ has a normal distribution. In practice, estimates are often chosen to minimize a loss function like mean squared error without assigning a distribution to Y.

Linear Regression Without Predictors

The estimated intercept

b_0= sample mean

The residuals

$e_i = Y_i - b_0$

The estimated var(Y | X) [X is null]

$$s^2 = (n-1)^{-1} \sum_{i=1}^{n} (y_i - b_0)^2$$

For the sleep data,

$b_0 = 10.3, \quad s^2 = 21.9$

Computing Regression Coefficients

Any reasonable statistical software will compute the least squares coefficients (b's) and residuals (e's).

We'll use R -- it's commonly used in research

Many corporations use SAS -- I don't know SAS.

Fact (don't worry if you haven't seen it before):

$$\mathbf{b} = (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t \mathbf{Y}$$

where

b is the vector of K estimated coefficients

Y is the vector of n outcomes (one for each observation)

X is an n x K matrix, each row is a different observation and each column is a different predictor

Estimates For The Sleep Data

The estimated line

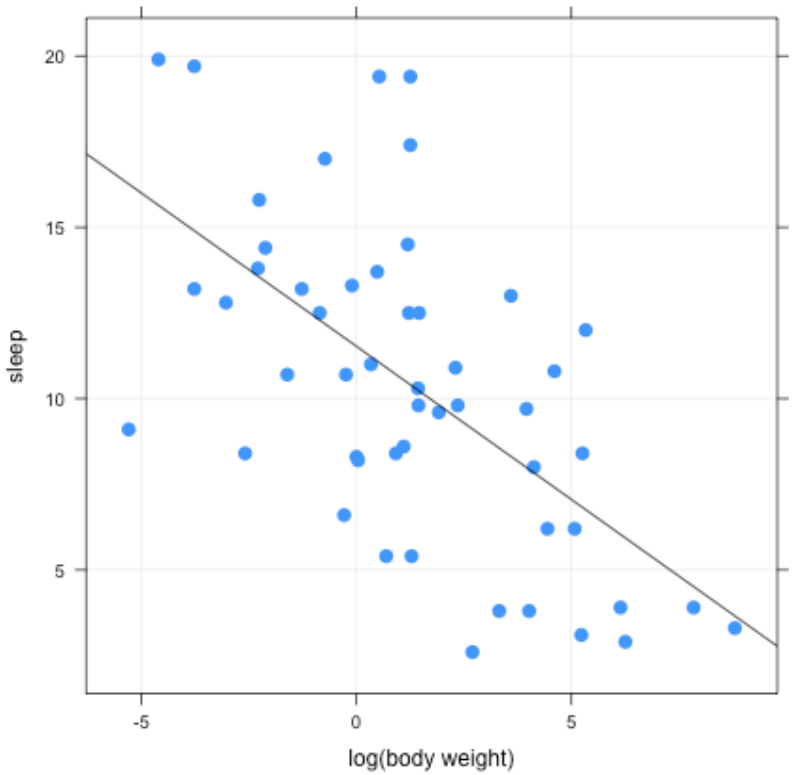
$$\mu(X_1) = b_0 + b_1X_1$$

has coefficients

$$(b_0, b_1) = (11.5, -0.9)$$

a species with mean weight 1 kg is estimated to sleep 11.5 hours on average

a species with a mean weight of 100 kg is estimated to sleep 7.4 hours on average.



Residuals for the Sleep Data

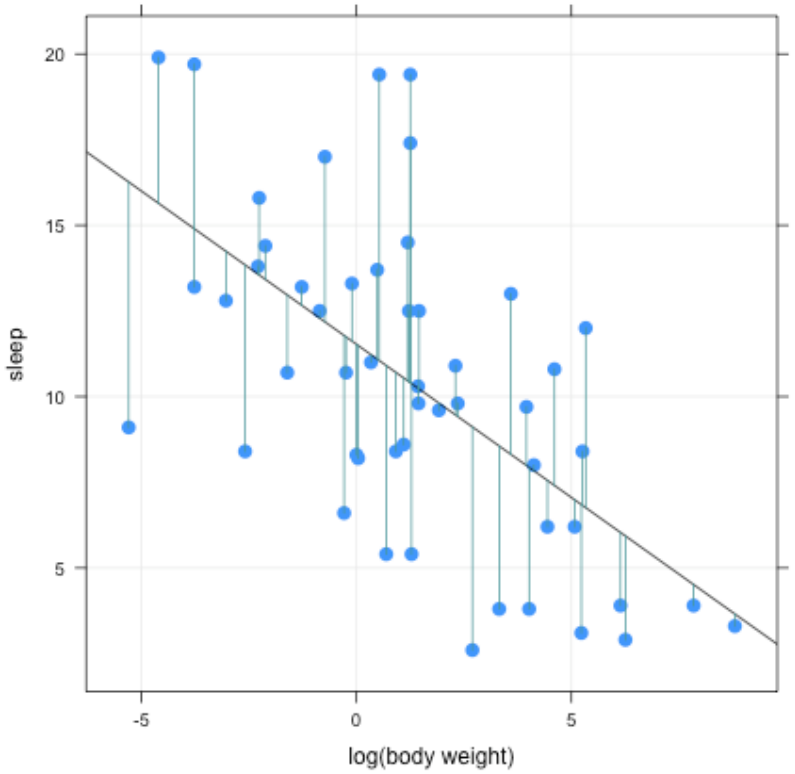
Gray segments show the residuals

$$e_i = Y_i - b_0 - b_1 \log(body_i)$$

If a segment is below the line, the outcome Y for that species is smaller than its estimated mean.

The closer the residuals are to zero, the better the regression fits the data.

Most measures of how well a model fits the data are based on residuals.



What About var(Y | **X**)?

In linear regression, the variance of sleep should not depend on **X**

$$\text{var}(Y \mid X_1, \ldots, X_k) = s^2$$

Example:

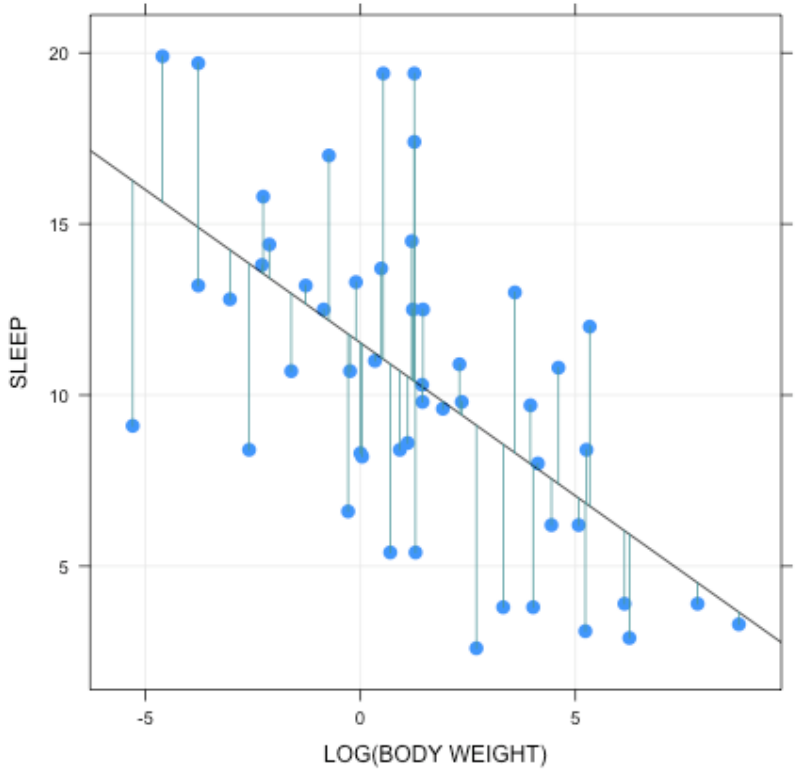
The variance of sleep for species with average weight 100 kg should be the same as the variance of sleep for species with average weight 1 kg.

Is this assumption plausible for the sleep data?

Variance of Y | X

If the residuals have a pattern in X, then var(Y| X) depends on X, so the model is bad.

Here, the residuals don't vary consistently with log(body weight).



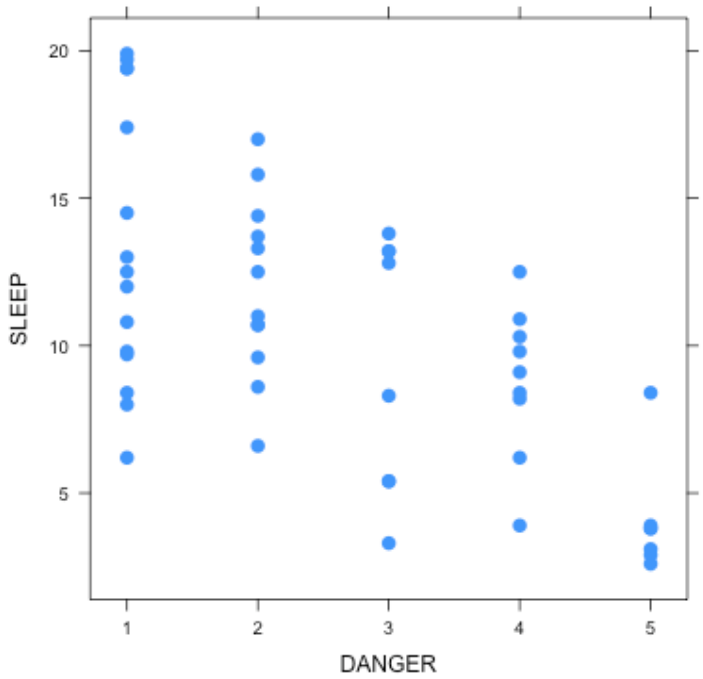
The assumption that variance of sleep is the same for all body weights passes the eyeball test.

Fact: var(Y|X) is estimated by $s^2 = (n-K)^{-1}\sum_{i=1}^{n} e_i^2$

Multiple Predictors

Including more predictors can improve a model

A quick look at the predictor danger →



Fit a regression model with log(body) and danger as predictors in R.

	intercept	log(body)	danger
old model	11.5	-.89	
new model	15.7	-.67	-1.68

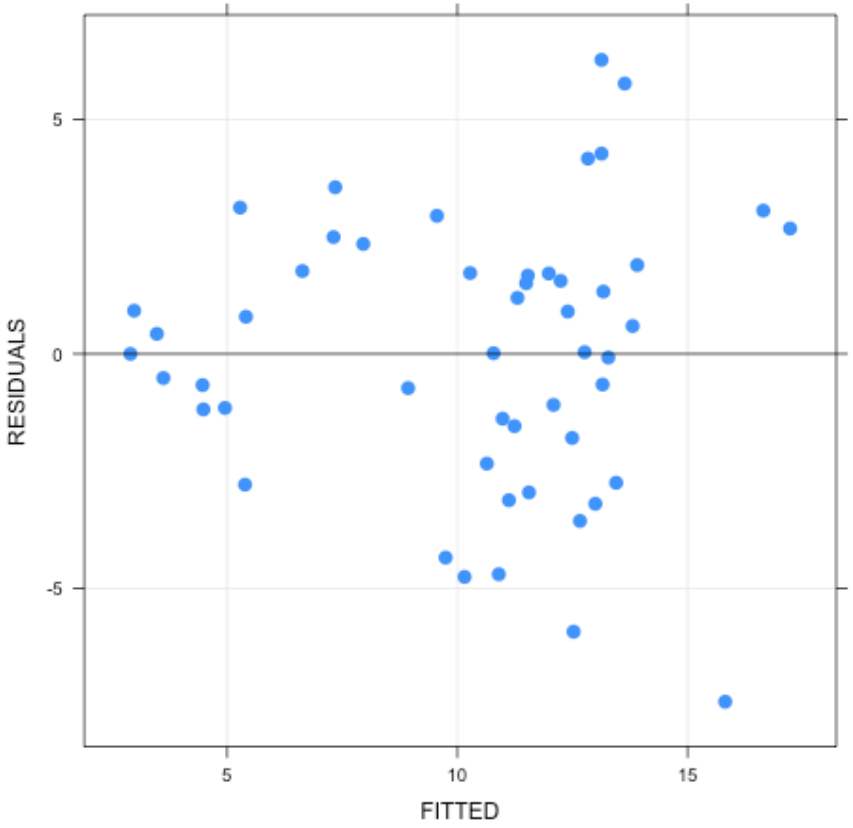
Including a new predictor can change our understanding of an old predictor.

Evaluating the Fit of Linear Regression Models

First Look At The Residuals

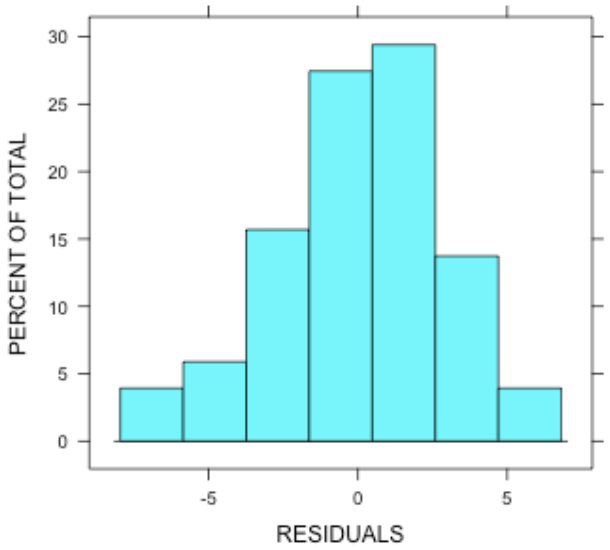
Residuals vs Estimated Means (the fitted values)

- a pattern suggests that the model doesn't fit
- sleep example with log(body) and danger as predictors
  - symmetric about zero over the range of the fitted, but a few points suggest taking a closer look at model fit.

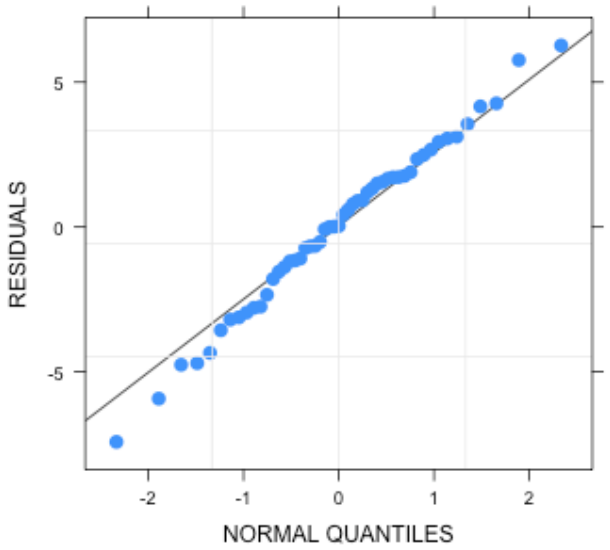


Residuals Should Be Close to Normal

Histogram of residuals looks reasonably similar to that of a random sample from a normal distribution with mean zero.



The ordered residuals look reasonably like the expected order statistics from a normal distribution with mean zero.



Classical Regression Statistics

The linear regression coefficients depend on the data (X_1, …, X_k, Y) so have standard errors.

R and SAS compute standard errors.

$b_k/se(b_k)$ is approximately t_{n-k} when $\beta_0=0$

Warning: The t-statistics assume that the other terms are in the model.

	b_k	$se(b_k)$	t-value	Pr(>|t|)
(Intercept)	15.6902	0.8771	17.889	< 2e-16
log(body)	-0.6993	0.1368	-5.112	5.50e-06
danger	-1.6830	0.3041	-5.534	1.28e-06

More Classical Regression Statistics

We can get confidence intervals for the mean Y and the prediction for a different species under conditions $X_1,\ldots,X_K$.

body	danger	estimate	meanLow	meanHigh	predLow	predHigh
0.1	1	15.6	14.1	17.1	9.5	21.8
1.0	1	14.0	12.7	15.3	7.9	20.1
10.0	1	12.4	11.0	13.8	6.3	18.5
100.0	1	10.8	9.1	12.5	4.6	17.0
0.1	5	8.9	6.8	11.0	2.6	15.2
1.0	5	7.3	5.5	9.1	1.0	13.5
10.0	5	5.7	4.0	7.3	-0.5	11.8
100.0	5	4.1	2.3	5.8	-2.2	10.3

The predicted value is the mean, but predicting is riskier than estimating a mean.