Likelihood, Deviance and AIC

Silwood Masters Statistics: Lecture Two

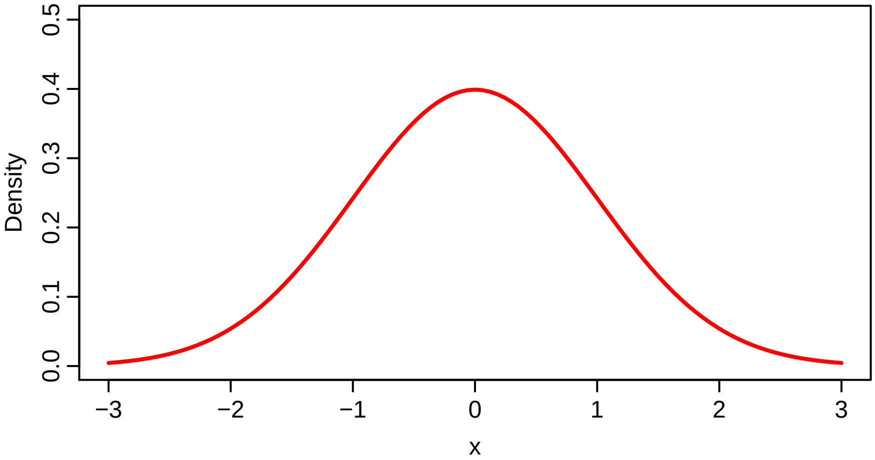
David Orme

Lecture outline

- Probability density curves
- Combining probabilities
- Likelihood and log-likelihood
- Akaike information criterion (AIC)
- Deviance

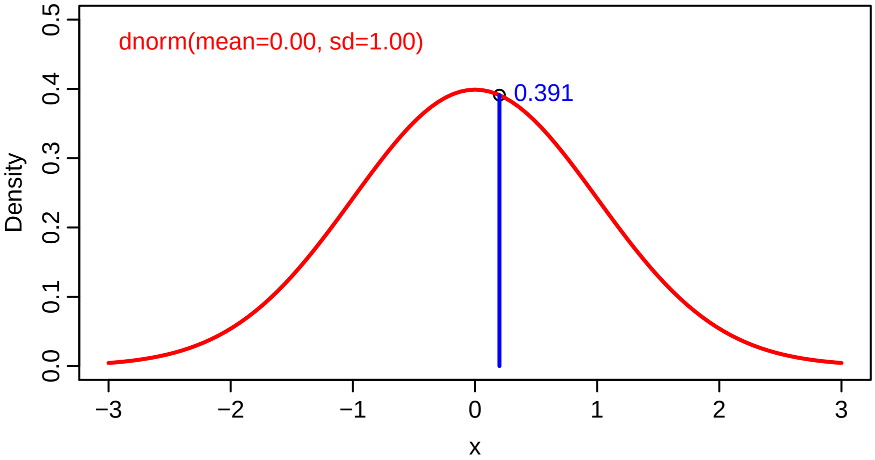
The normal probability density function

Probability of observing a given value of x.



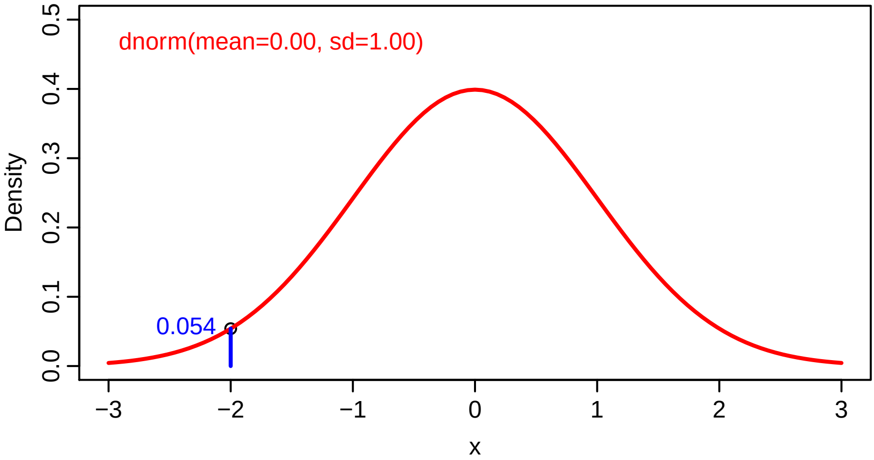
The normal probability density function

Some values of x have a high probability of being observed ...

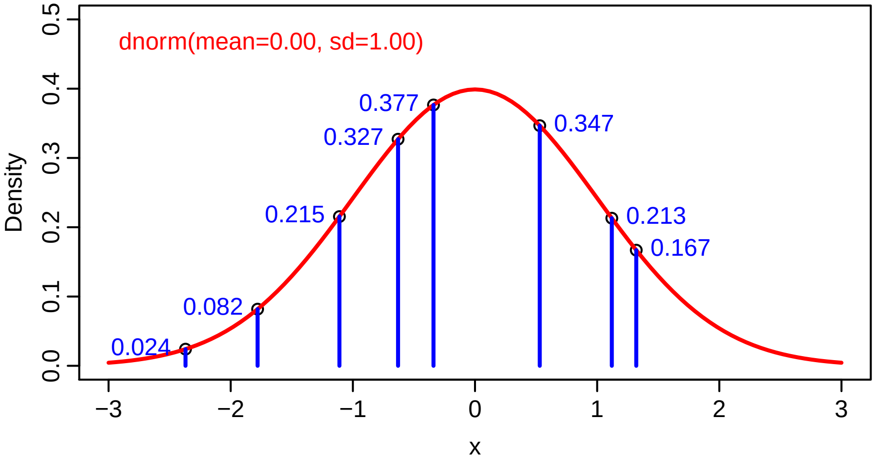


The normal probability density function

…while other values of x have a low probability of being observed.



What about multiple observations?



Combining probabilities: multiplication

The probability of multiple events is the product of the individual probabilities:

		⚀	⚁	⚂	⚃	⚄	⚅
		$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$
⚀	$\frac{1}{6}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
⚁	$\frac{1}{6}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
⚂	$\frac{1}{6}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
⚃	$\frac{1}{6}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
⚄	$\frac{1}{6}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
⚅	$\frac{1}{6}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$

Combining probabilities: logarithms

Multiplying probabilities ($0 \leq p \leq 1$) leads to very small numbers!

$$
\begin{aligned}
0.02 &= 0.0240556 \\
0.02 \times 0.08 &= 0.0019684 \\
0.02 \times 0.08 \times 0.22 &= 0.0004241 \\
0.02 \times 0.08 \times 0.22 \times 0.33 &= 0.0001387 \\
0.02 \times 0.08 \times 0.22 \times 0.33 \times 0.38 &= 0.0000522 \\
0.02 \times 0.08 \times 0.22 \times 0.33 \times 0.38 \times 0.35 &= 0.0000181 \\
0.02 \times 0.08 \times 0.22 \times 0.33 \times 0.38 \times 0.35 \times 0.21 &= 0.0000039 \\
0.02 \times 0.08 \times 0.22 \times 0.33 \times 0.38 \times 0.35 \times 0.21 \times 0.17 &= 0.0000006
\end{aligned}
$$

Combining probabilities: logarithms

- Multiplied probabilities are awkward to report
- Can lead to loss of precision on computers
- Add the logarithms of probabilities instead

$$\begin{aligned} e^2 \times e^5 \times e^3 = e^{(2+5+3)} &= e^{10} \\ (e \times e) \times (e \times e \times e \times e) \times (e \times e \times e) &= e^{10} \\ &= 22026.47 \end{aligned}$$

$$\begin{aligned} \ln_e(e^2) + \ln_e(e^5) + \ln_e(e^3) &= \ln_e(e^{10}) \\ 2 + 5 + 3 &= 10 \end{aligned}$$

$$\exp(10) = e^{10} = 22026.47$$

Combining probabilities: logarithms

Adding log probabilities is easier and more stable

$$-3.7 = -3.73$$
$$-3.7 + -2.5 = -6.23$$
$$-3.7 + -2.5 + -1.5 = -7.77$$
$$-3.7 + -2.5 + -1.5 + -1.1 = -8.88$$
$$-3.7 + -2.5 + -1.5 + -1.1 + -1.0 = -9.86$$
$$-3.7 + -2.5 + -1.5 + -1.1 + -1.0 + -1.1 = -10.92$$
$$-3.7 + -2.5 + -1.5 + -1.1 + -1.0 + -1.1 + -1.5 = -12.47$$
$$-3.7 + -2.5 + -1.5 + -1.1 + -1.0 + -1.1 + -1.5 + -1.8 = -14.26$$

$$\exp(-14.26) = e^{-14.26} = 0.0000006$$

Probability of observing multiple x

A vector of observed values (x):

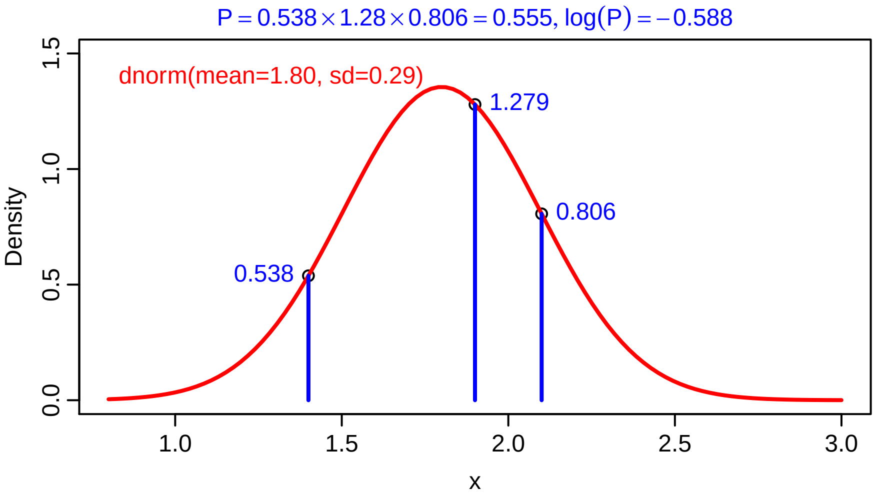
$$\text{Mean } (\bar{x}) = \frac{\sum x}{n} \qquad \text{SD } (\sigma) = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}}$$

```
# define our observed values
library(MASS)
x <- c(1.4, 1.9, 2.1)
fitdistr(x, 'normal')
```

```
##      mean       sd
##    1.800    0.294
##   (0.170)  (0.120)
```

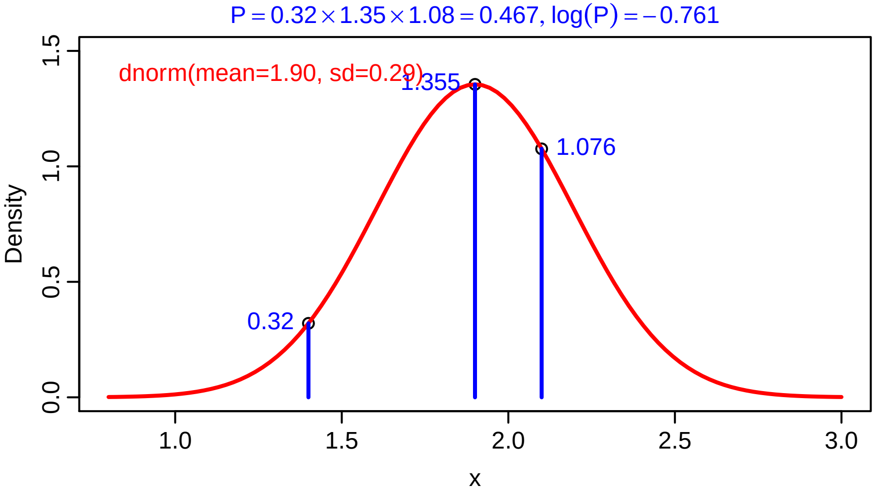
Probability of observing x

Probability at the observed mean and standard deviation



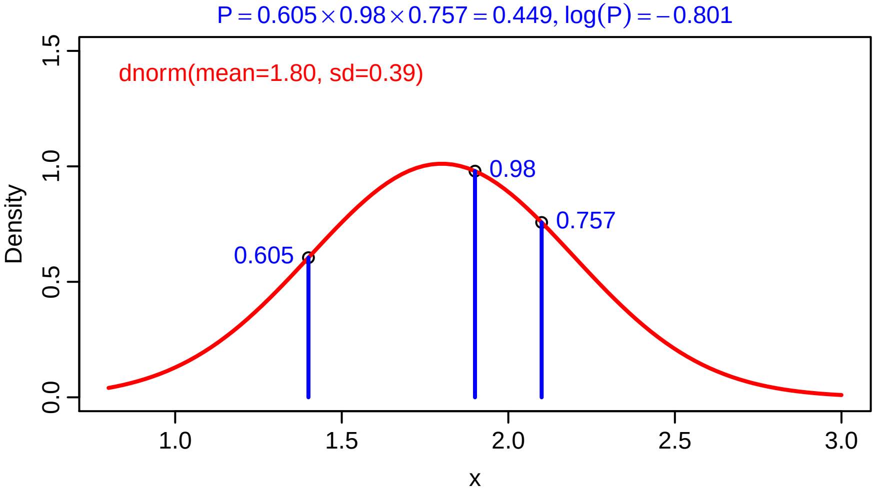
Probability of observing x

Probability at a slightly higher mean



Probability of observing x

Probability at slightly higher standard deviation

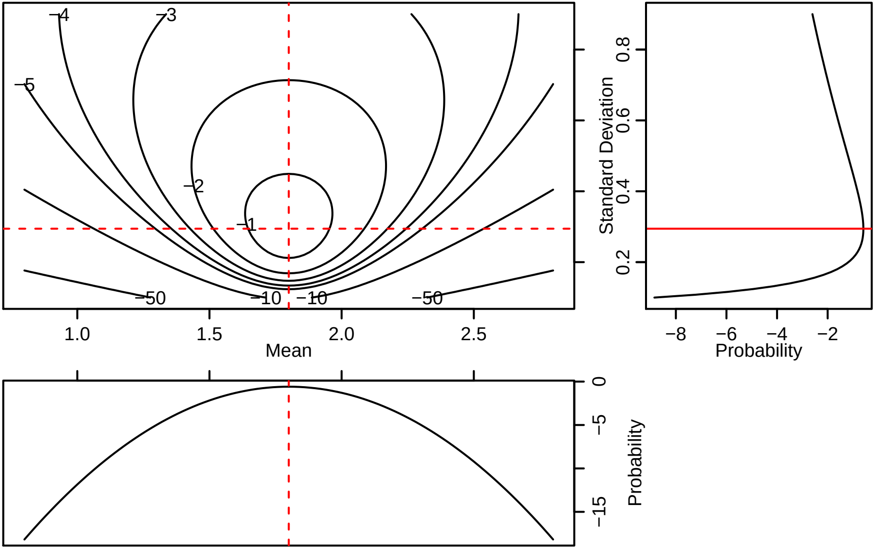


Finding the most probable distribution

What distribution gives the highest probability of the data:

- Given a distribution (e.g. normal)
- Under what combination of parameters (e.g. mean and standard deviation) is the data most probable?
- Smaller values are less probable.
- Find parameters that give the largest probability.

Probability surface



Probability and Likelihood

Given some observed data (O) and a probability function ($f()$) with some parameters (θ).

$P(O|\theta) =$ The **probability** of the observed data given the parameters.

$L(\theta|O) =$ The **likelihood** of the parameters given the observed data.

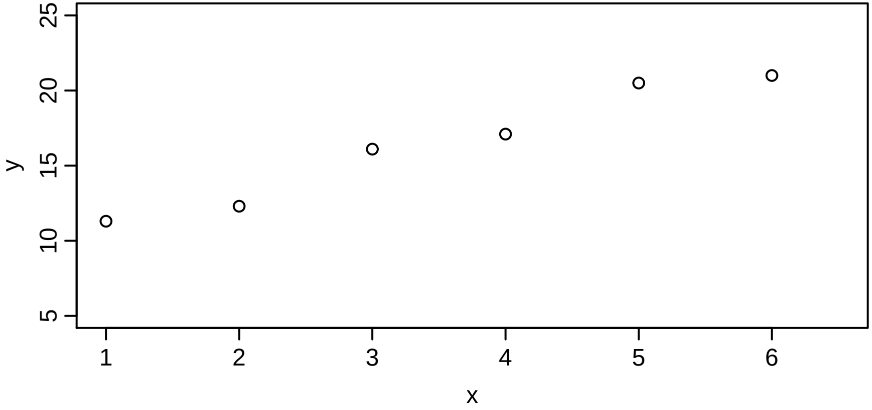
But,

$$L(\theta|O) = P(O|\theta)$$

Likelihood is just a change of perspective - *how likely is this model, given the data*.

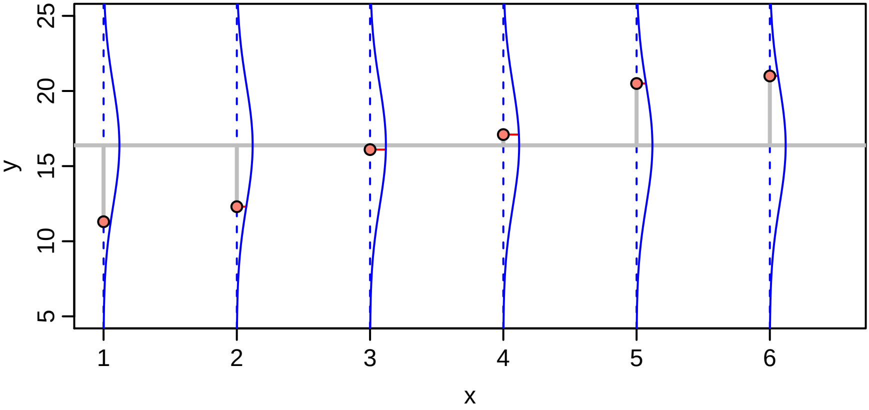
Likelihood in a linear model

Using a simple linear regression:



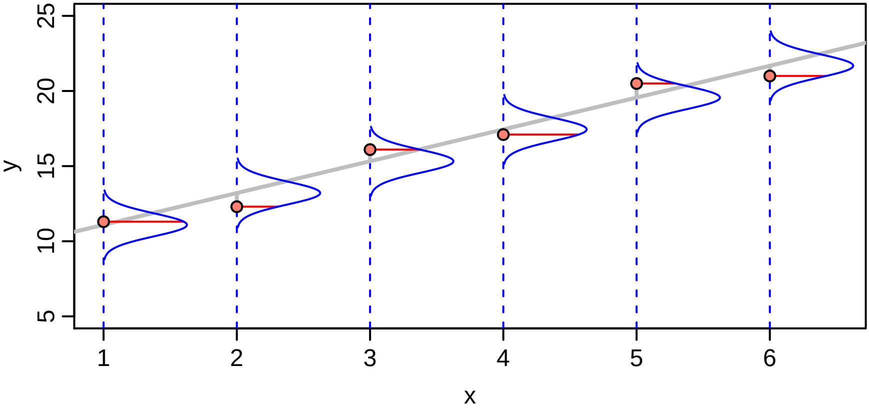
Likelihood in a linear model

- The **null** model ($y = a$)
- Residuals have high variance and low probabilities
- Log likelihood of the model is -16.335

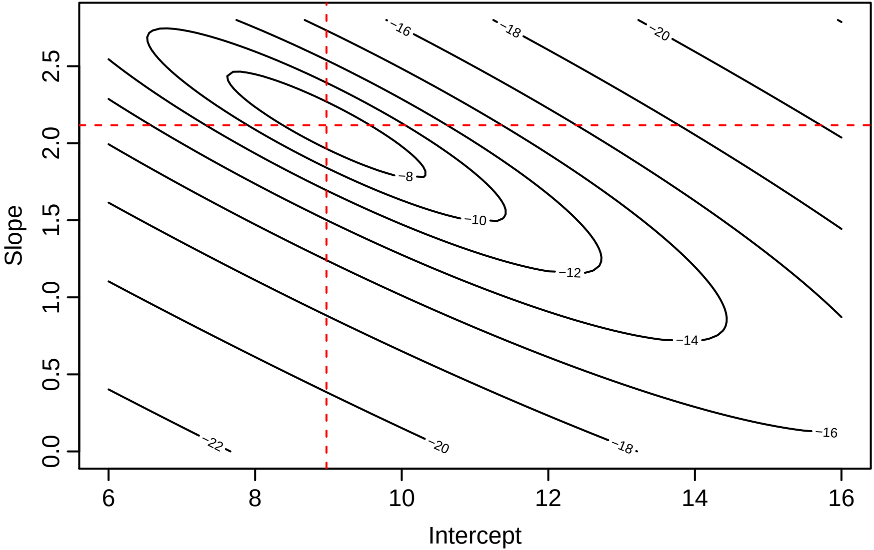


Likelihood in a linear model

- The **regression** model ($y = a + bx$)
- Residuals have low variance and higher probabilities
- Log likelihood of the model is -6.362

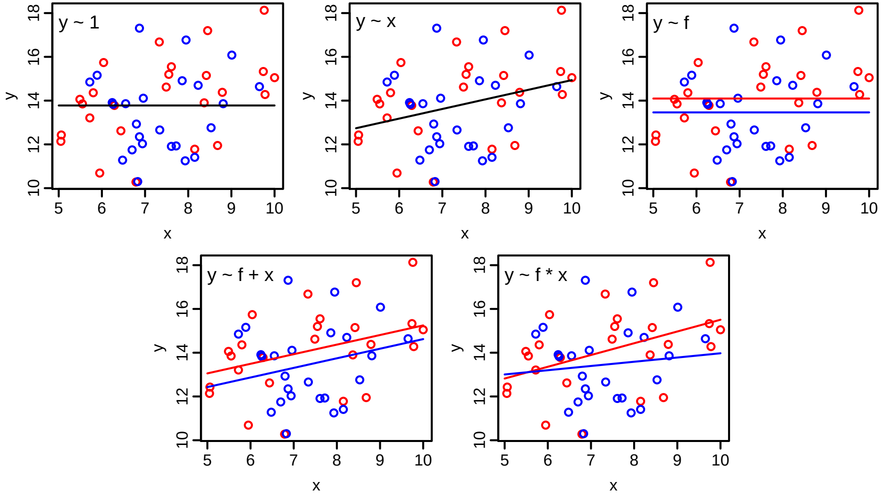


Likelihood profile for a model



Using likelihood to compare models

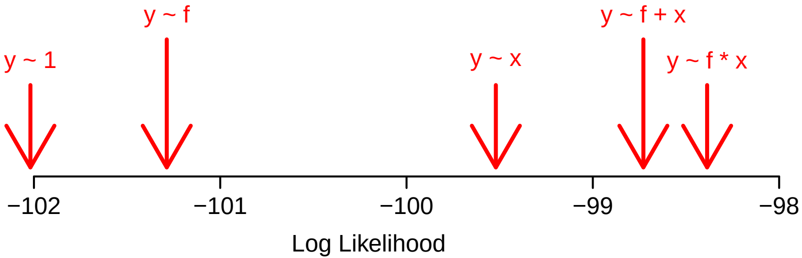
Five models of increasing complexity *fitted to the same data*:



Using likelihood to compare models

More complex models:

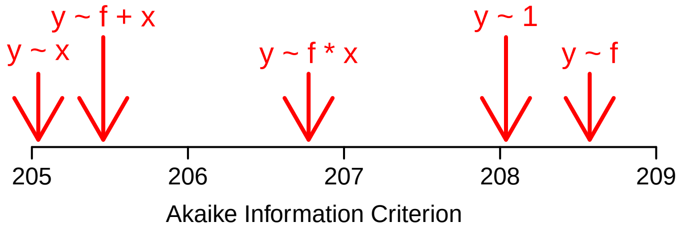
- Have more parameters
- Explain more variation in the data
- Have higher log likelihood



Using AIC to compare models

The Akaike Information Criterion (AIC) balances

- likelihood (L) – higher is better – and
- the number of parameters (k) – fewer is better.
- $AIC = 2k - 2\ln(L)$
- Note the signs of the parameters: low AIC is better



Most complex model summary

```
summary(modF)
##
## Call:
## lm(formula = y ~ f * x)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -3.50  -1.09   0.06   1.05   3.94
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   10.133      1.746    5.80  5.7e-07 ***
## f2             1.906      3.220    0.59    0.557
## x              0.537      0.232    2.32    0.025 *
## f2:x          -0.344      0.432   -0.80    0.430
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.8 on 46 degrees of freedom
## Multiple R-squared:  0.135, Adjusted R-squared:  0.0788
## F-statistic:  2.4 on 3 and 46 DF,  p-value: 0.0801
```

ANOVA comparison

```
anova(modN, modR, modA, modC, modF)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ 1
## Model 2: y ~ x
## Model 3: y ~ f
## Model 4: y ~ f + x
## Model 5: y ~ f * x
##   Res.Df RSS Df Sum of Sq    F Pr(>F)
## 1     49 173
## 2     48 157  1     16.48 5.06  0.029 *
## 3     48 168  0    -11.48
## 4     47 152  1     16.37 5.02  0.030 *
## 5     46 150  1      2.06 0.63  0.430
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```