Week 2 assignment: likelihood

Example solutions

Jan 17, 2015

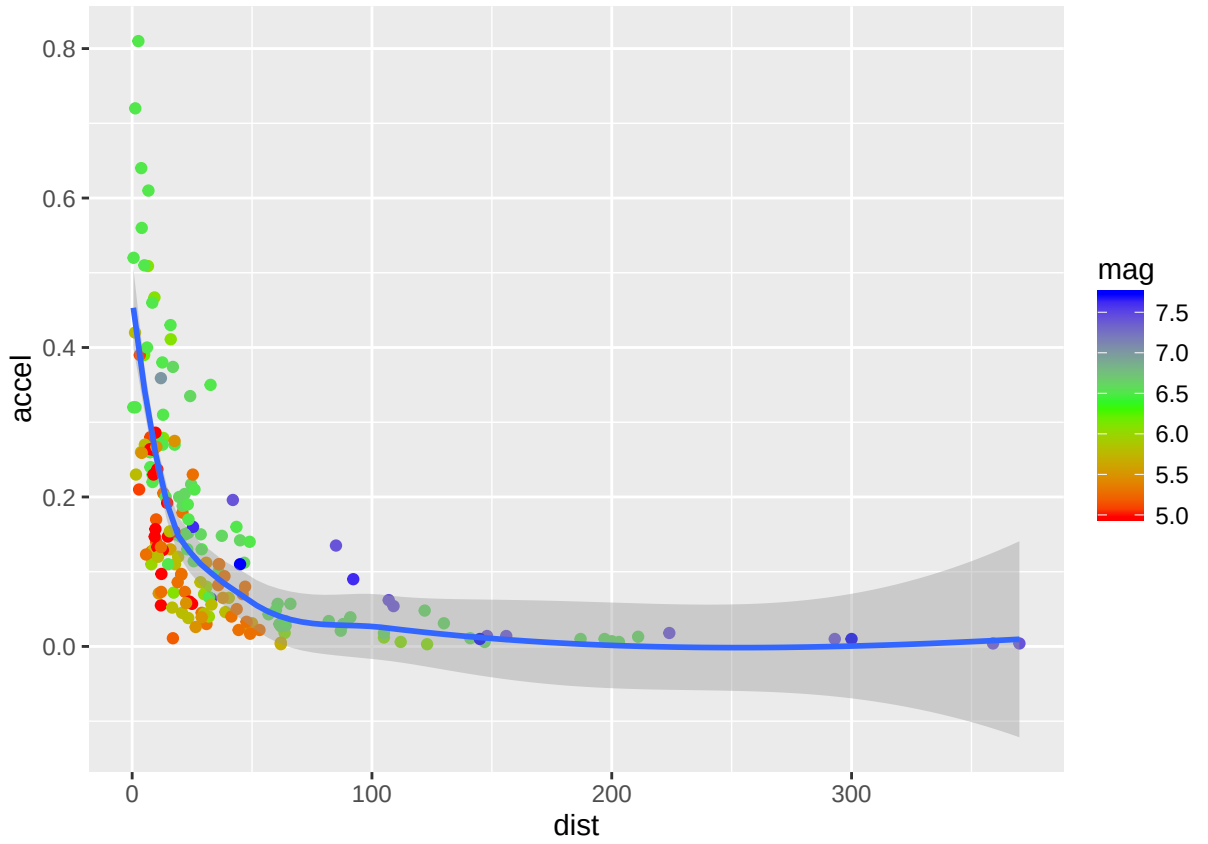
Problem 1

Earthquakes are most forceful at their epicenters, and this force attenuates with distance. R has an earthquake dataset `attenu` with measurements of peak horizontal ground acceleration by distance measured at multiple recording stations, with each earthquake coded in terms of magnitude. See the help file for more information on this dataset by typing `?attenu`.

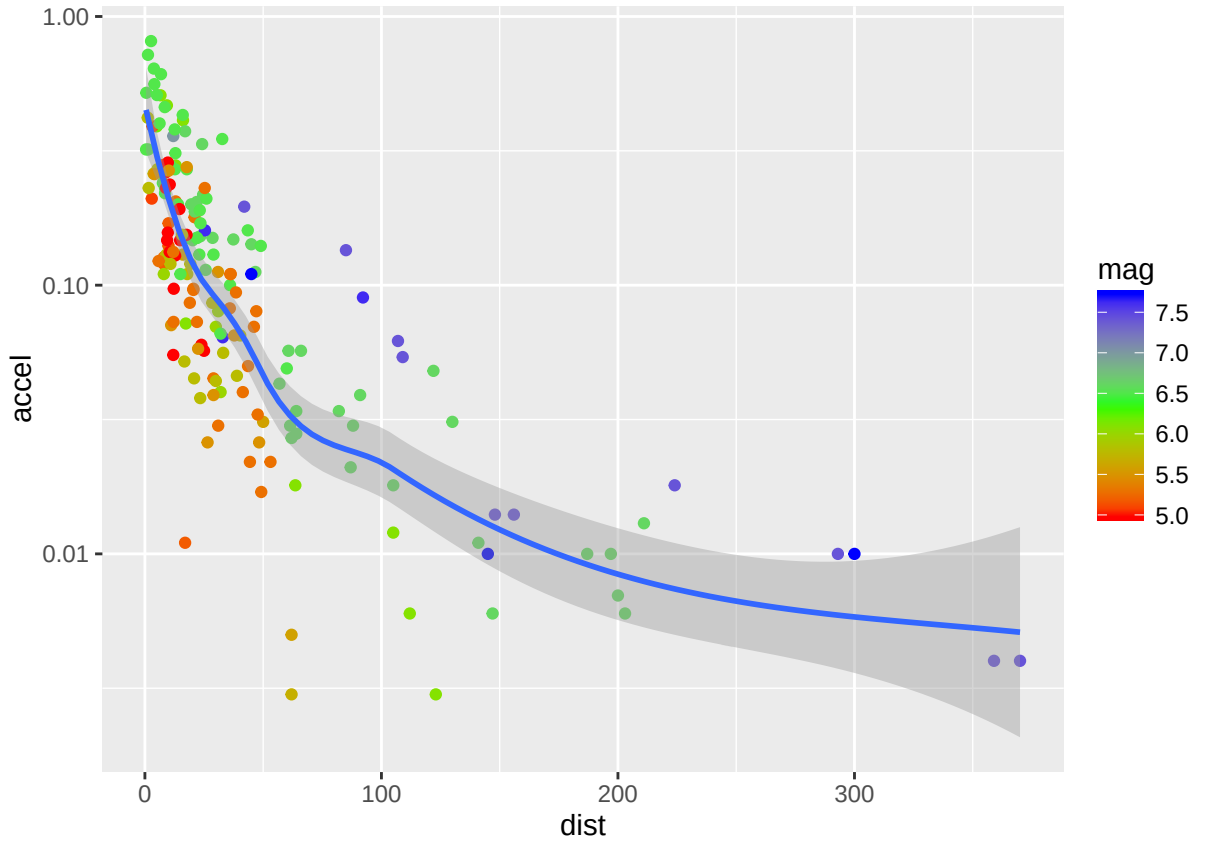
Your main task this week is to build one model that predicts both ground acceleration at the epicenter (distance = 0), and the acceleration by distance curve as a function of magnitude and distance from the epicenter. You will obtain maximum likelihood estimates for the parameters of the model using the `optim` function. The structure of your model is up to you. You can use a combination of intuition, imagination, first principles, research, and collaboration to construct your model. (Note: there are many possible models that one could construct!)

You will benefit from visualizing the data, specifically the relationships between the quantities of interest i.e., magnitude (`mag`), distance from epicenter (`dist`), and peak acceleration (`accel`). Include your visualization code and plots below.

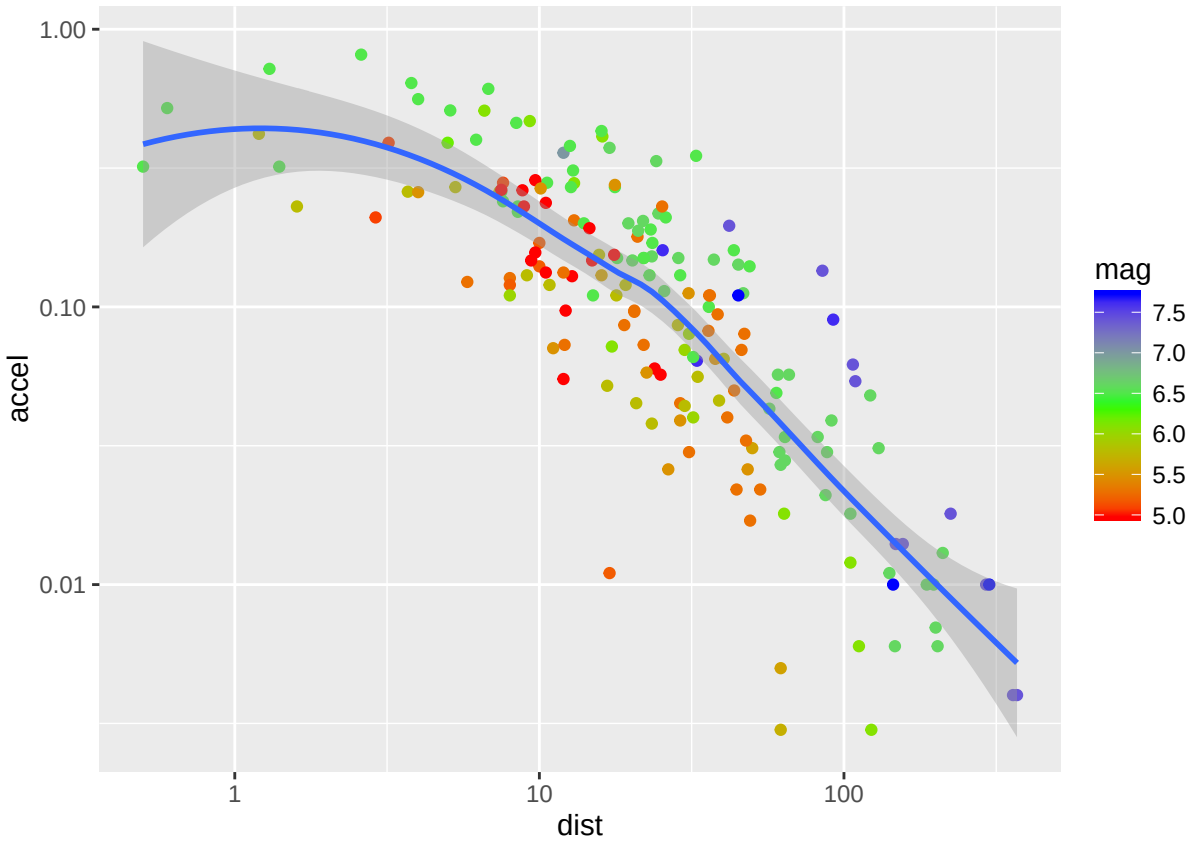
```
# your code here
library(ggplot2)
ggplot(attenu, aes(x=dist, y=accel, col=mag)) +
  geom_point() +
  scale_color_gradientn(colors = rainbow(3)) +
  geom_smooth()
```




```
ggplot(attenu, aes(x=dist, y=accel, col=mag)) +
  geom_point() +
  scale_color_gradientn(colors = rainbow(3)) +
  geom_smooth() +
  scale_y_log10()
```




```
ggplot(attenu, aes(x=dist, y=accel, col=mag)) +
  geom_point() +
  scale_color_gradientn(colors = rainbow(3)) +
  scale_y_log10() +
  scale_x_log10() +
  geom_smooth()
```



What is the equation for your model? Write it in LaTeX, bounded between dollar signs (e.g., $e = mc^2$), not in normal text.

$y_i \sim N(\gamma m_i e^{-\frac{d_i}{\phi}}, \sigma)$

Define all of the parameters, and explain why you formulated it in this way. What assumptions went into your model's construction?

There are a ton of potential answers here. For me, y_i is the peak ground acceleration, m_i is magnitude, d_i is distance from epicenter, γ is a proportionality parameter that determines peak acceleration at the epicenter, and ϕ is a distance decay parameter. I'm assuming that the acceleration decays exponentially with distance, and that the rate of decay is constant with respect to magnitude. The initial peak acceleration however is a function of magnitude (specifically some fraction of magnitude). I'm ignoring the information on different stations, assuming that there are no systematic differnece among stations other than distance to epicenter.

Write a function called `nll` that returns the negative log likelihood for your model. The arguments to this function should be `theta` (the parameters), and `data` (the data).

```
# function to return the expected value (the mean function)
predict_accel <- function(theta, data){
  gamma <- theta['gamma']
  phi <- theta['phi']
  sigma <- exp(theta['lsigma'])
  mu <- gamma * data[, 'mag'] * exp(- data[, 'dist'] / phi)
  mu
}

# function to return the negative log likelihood
nll <- function(theta, data){
  mu <- predict_accel(theta, data)
```

```
  -sum(dnorm(data[, 'accel'], mu, exp(theta['lsigma']), log = TRUE))
}
```

Use `optim` to obtain maximum likelihood estimates for your model parameters.

```
inits <- c(gamma = .1, phi = .1, lsigma = .1)
out <- optim(inits, nll, data = attenu)
out
```

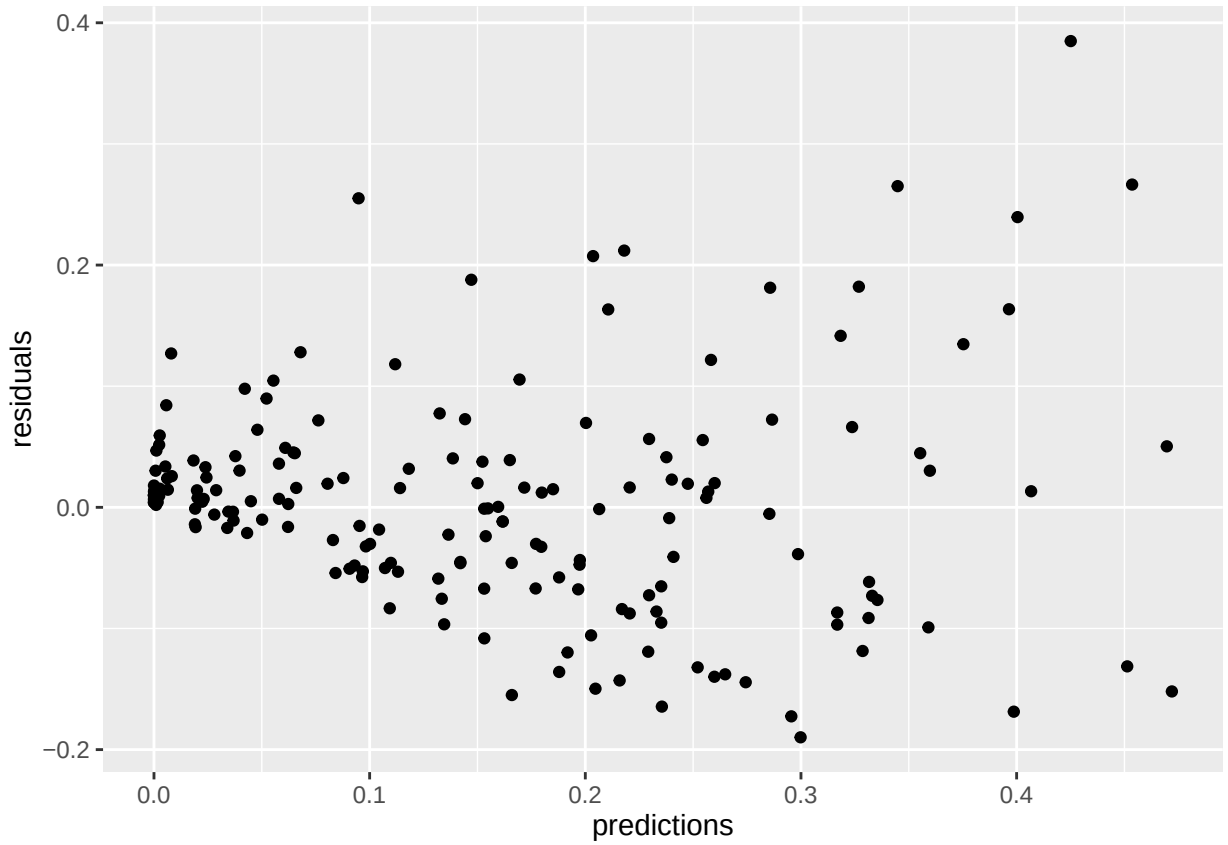
```
## $par
##       gamma         phi      lsigma
##  0.07444527 20.07108770 -2.41511642
##
## $value
## [1] -181.2913
##
## $counts
## function gradient
##      362       NA
##
## $convergence
## [1] 0
##
## $message
## NULL
```

Did `optim()` converge to a minimum? How do you know?

Yes, the convergence code is 0, indicating successful convergence to a minimum.

Create a scatterplot with fitted values vs. residuals.

```
attenu$predictions <- predict_accel(out$par, attenu)
attenu$residuals <- with(attenu, accel - predictions)
ggplot(attenu, aes(x = predictions, y = residuals)) +
  geom_point()
```



What stands out in the plot of fitted values vs. residuals? Are you worried about any violations of assumptions? Why or why not?

It looks like there is heteroscedasticity. The residual spread is much greater for higher predictions than smaller.

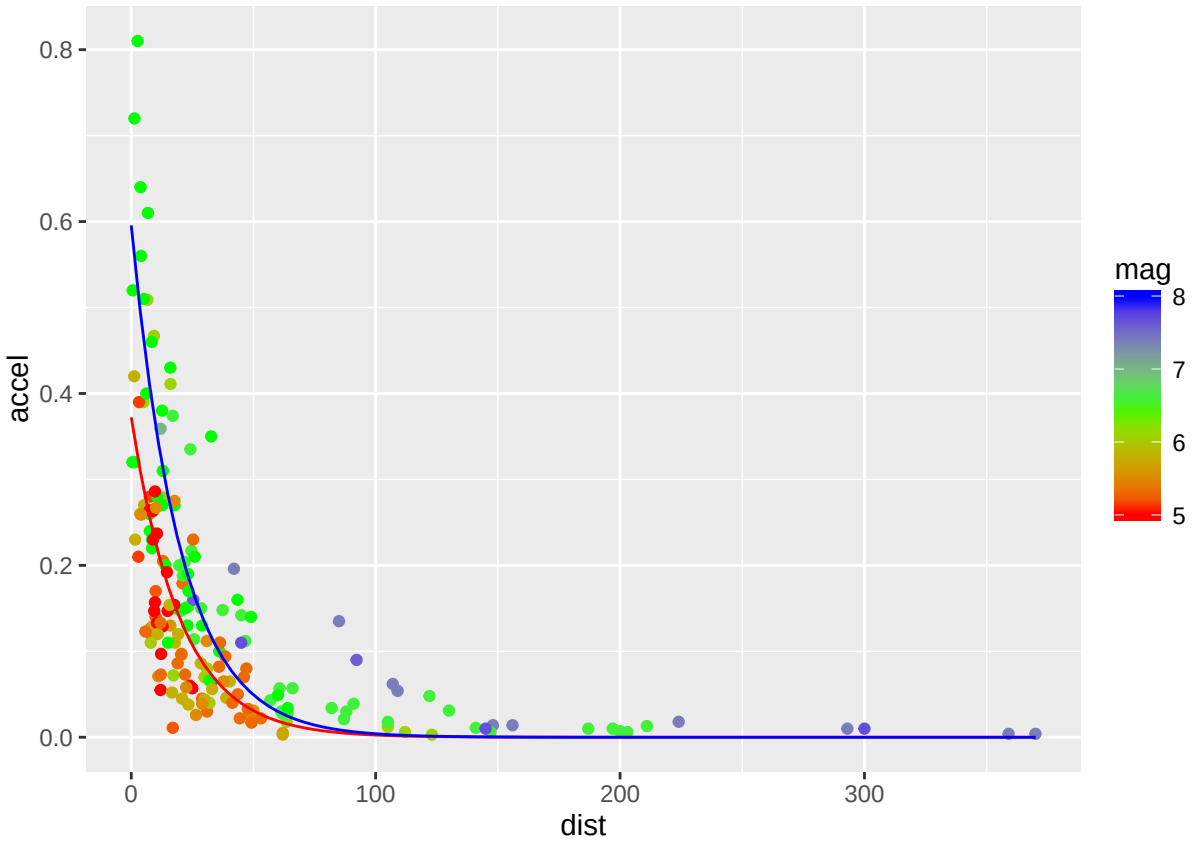
Plot the distance by acceleration data along with your predicted curves (starting at a distance of 0) for earthquakes of magnitude 5 and 8.

```
# create a vector of distances
lo <- 100
dists <- seq(0, max(attenu$dist), length.out=lo)

new_d5 <- data.frame(mag = 5, dist = dists)
new_d5$accel <- predict_accel(out$par, new_d5)

new_d8 <- data.frame(mag = 8, dist = dists)
new_d8$accel <- predict_accel(out$par, new_d8)

ggplot(attenu, aes(x=dist, y=accel, col=mag)) +
  geom_point() +
  scale_color_gradientn(colors = rainbow(3)) +
  geom_line(data = new_d5) +
  geom_line(data = new_d8)
```



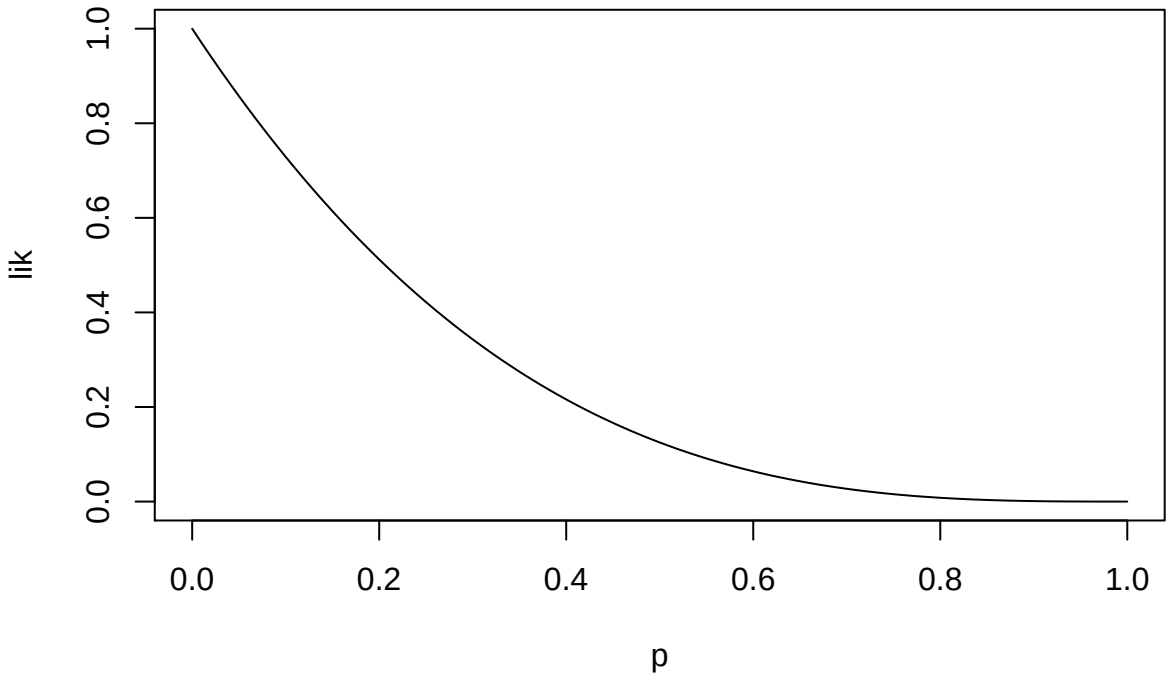
How do your predictions compare to the data? Which characteristics of the data are captured well, and which are captured poorly by your model?

The predictions match the decay of the data somewhat, though the predictions aren't as extreme as the data on either end. Also the model does not capture the positivity of acceleration - for low expected values the model does have positive probability for negative acceleration.

Problem 2

Pat loves to play basketball. You observe Pat practicing free throws at the recreation center one day. Pat misses 3 shots in a row. Generate a likelihood profile for p, the probability that Pat makes a free throw.

```
p <- seq(0, 1, .01)
lik <- dbinom(0, 3, p)
plot(p, lik, type = 'l')
```



What is your MLE for p, and does it make sense? Why or why not?

The MLE is 0. This seems unreasonable - it is very unlikely that Pat has a zero probability of making a free throw.