Week 2 assignment: likelihood

Your name here

Jan 17, 2015

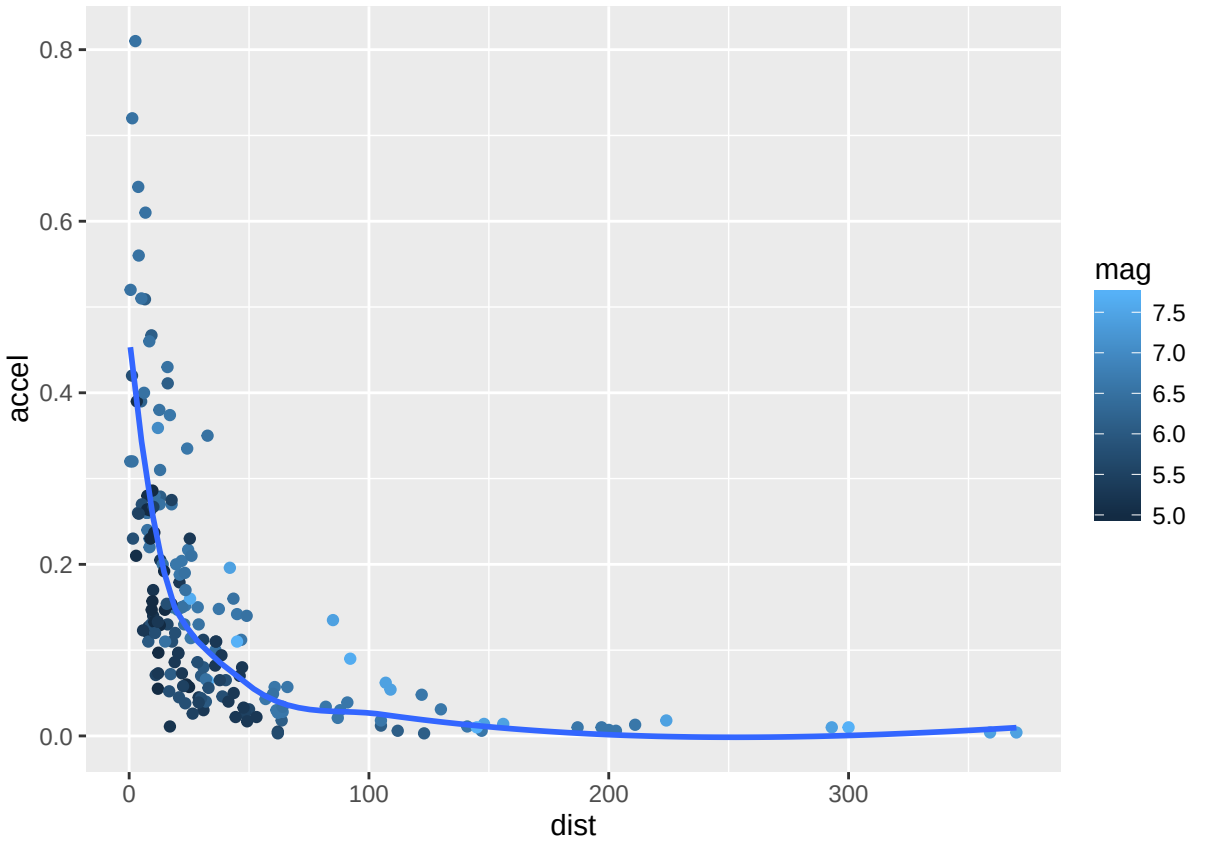
Problem 1

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.

```
# plot acceleration as a function of distance
library(ggplot2)
ggplot(data = attenu, aes(dist,accel)) +
  geom_point(aes(col = mag)) +
  geom_smooth(se = FALSE)
```

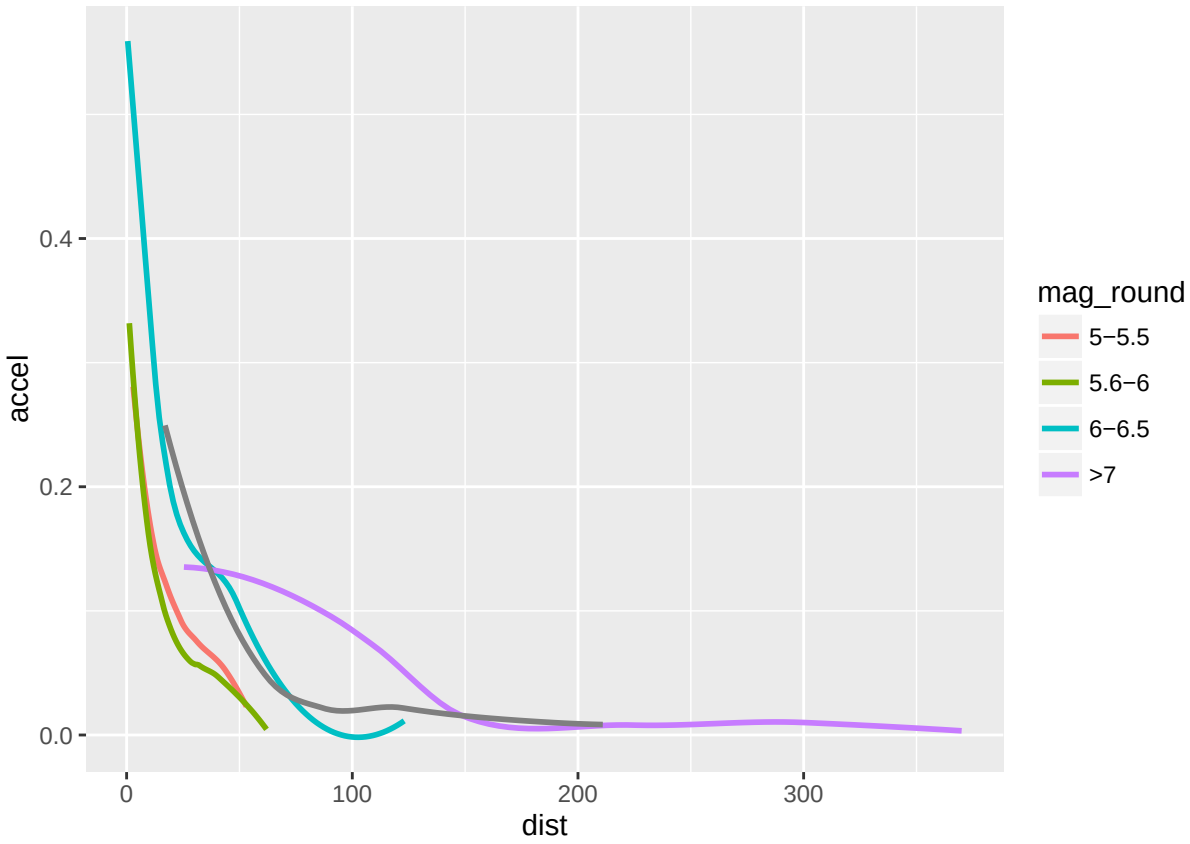



```
  # seems like an exponential decay overall, but are there differences in the
  # acceleration at the epicenter and the rate of decay for earthquakes of
  # different magnitudue?

# let's bin the magnitudes to make visualization a little easier
attenu[attenu$mag >= 5 & attenu$mag <= 5.5,"mag_round"] <- "5-5.5 "
attenu[attenu$mag > 5.5 & attenu$mag <=6,"mag_round"] <- "5.6-6 "
attenu[attenu$mag > 6 & attenu$mag <=6.5,"mag_round"] <- "6-6.5 "
attenu[attenu$mag > 6.6 & attenu$mag <=7,"mag_round"] <- "6.6-7 "
attenu[attenu$mag > 7, "mag_round"] <- ">7"
table(attenu$mag_round)
```

```
##
## 5-5.5  5.6-6  6-6.5  6.6-7      >7
##     58     26     48      1     16
```

```
# re-plot acceleration as a function of distance
ggplot(data = attenu, aes(dist,accel)) +
  geom_smooth(se = FALSE, aes(color = mag_round))
```




```
  # doesn't look like a huge difference in the rate of decay until magnitudes
  # greater than 7. However, the epicenter acceleration seems maximized for
  # intermediate values of magnitude.
```

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

$y_i \sim N((\beta m_i + \gamma m_i^2) * e^{-\frac{d_i}{m_i \lambda}}, \sigma)$

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

Here, y_i is the predicted acceleration, m_i is the magnitude and d_i is the distance from the epicenter. The parameters β and γ together determine the amount of acceleration when distance equals zero. Because γ is multiplied by the square of the magnitude, this means that the acceleration at the epicenter will be quadratic (rather than linear) function of the magnitude of the earthquake. Hopefully, this accounts for the in larger accelerations at the epicenter for moderately magnitude earthquakes. λ is the exponential decay constant, which itself is proportional to the magnitude of the earthquake. This should reduce the the rate of decay of earthquakes of higher magnitude (slightly).

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){
  beta <- theta['beta']
  gamma <- theta['gamma']
  lambda <- theta['lambda']
```

```
  sigma <- exp(theta['lsigma'])
  mu <- (beta * data[,'mag'] + gamma * (data[,'mag'])^2) *  exp(- data[, 'dist']
   / (lambda * data[,'mag']))
  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.

```
# intial values
inits <- c(beta = 0, gamma = .1, lambda = .1, lsigma = .1)

# optimize
out <- optim(inits, nll, data = attenu)
out
```

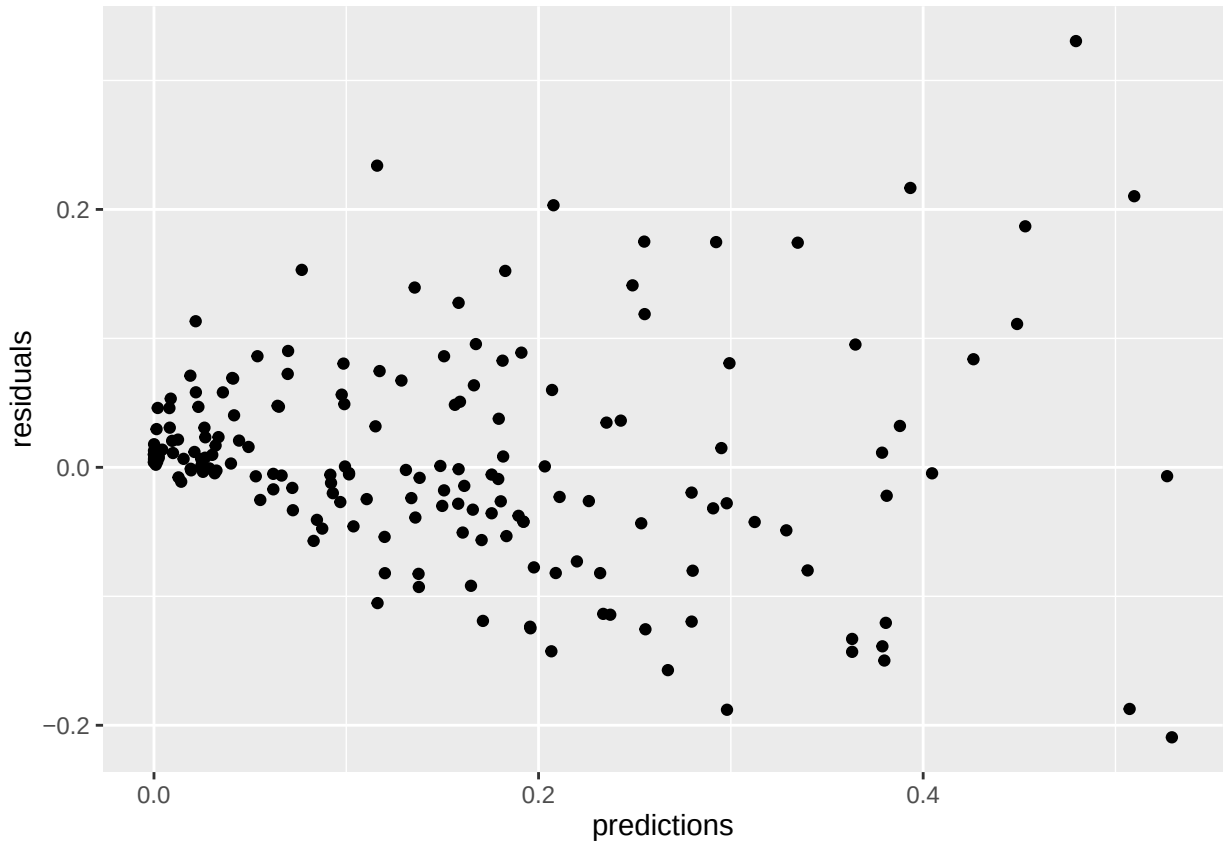
```
## $par
##        beta       gamma      lambda      lsigma
## -0.02915898  0.01731331  3.26421555 -2.50792412
##
## $value
## [1] -198.1889
##
## $counts
## function gradient
##      479       NA
##
## $convergence
## [1] 0
##
## $message
## NULL
```

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

Sure did. Convergence = 0

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?

Definately some heteroscedasticity present. There is much more spread around the larger predicted accelerations

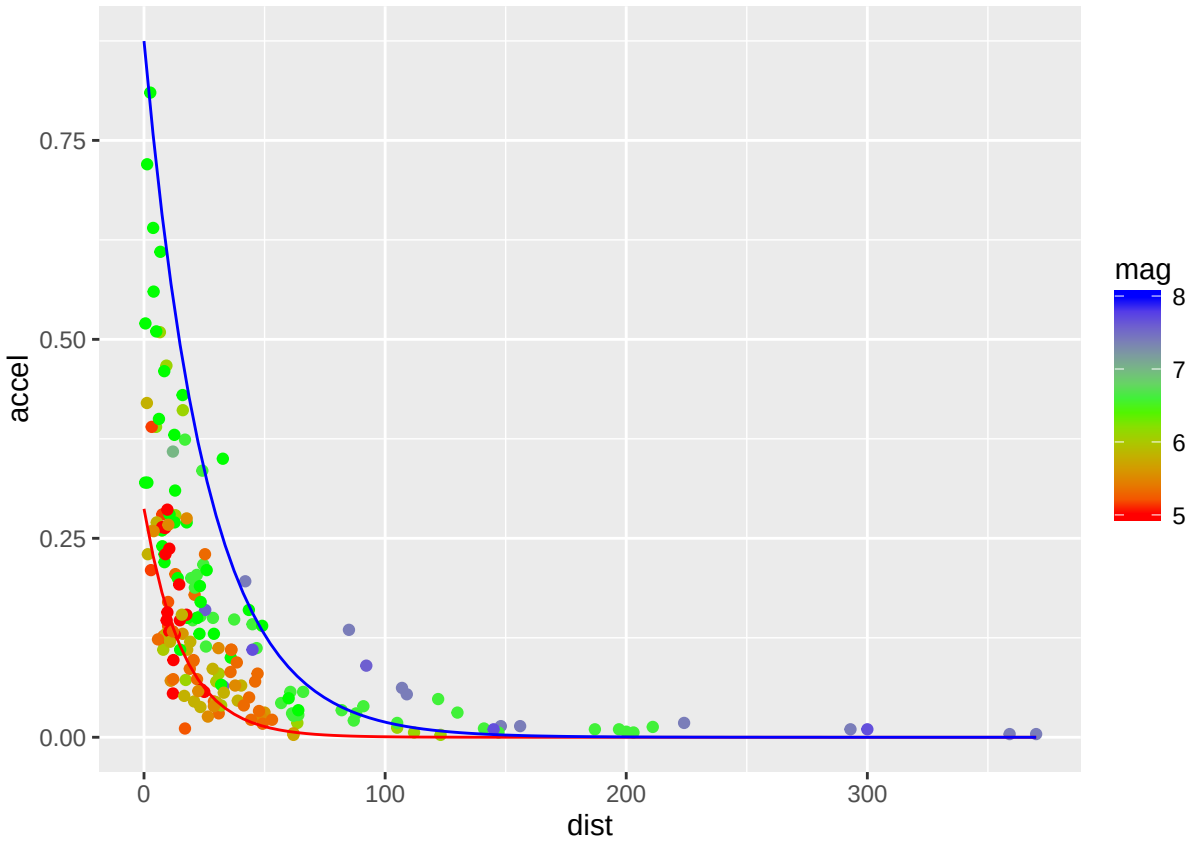
Plot the distance by acceleration data along with your predicted curves (starting at a distance of 0) for earthquakes of magnitude 5 and 8.

```
# 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)
```



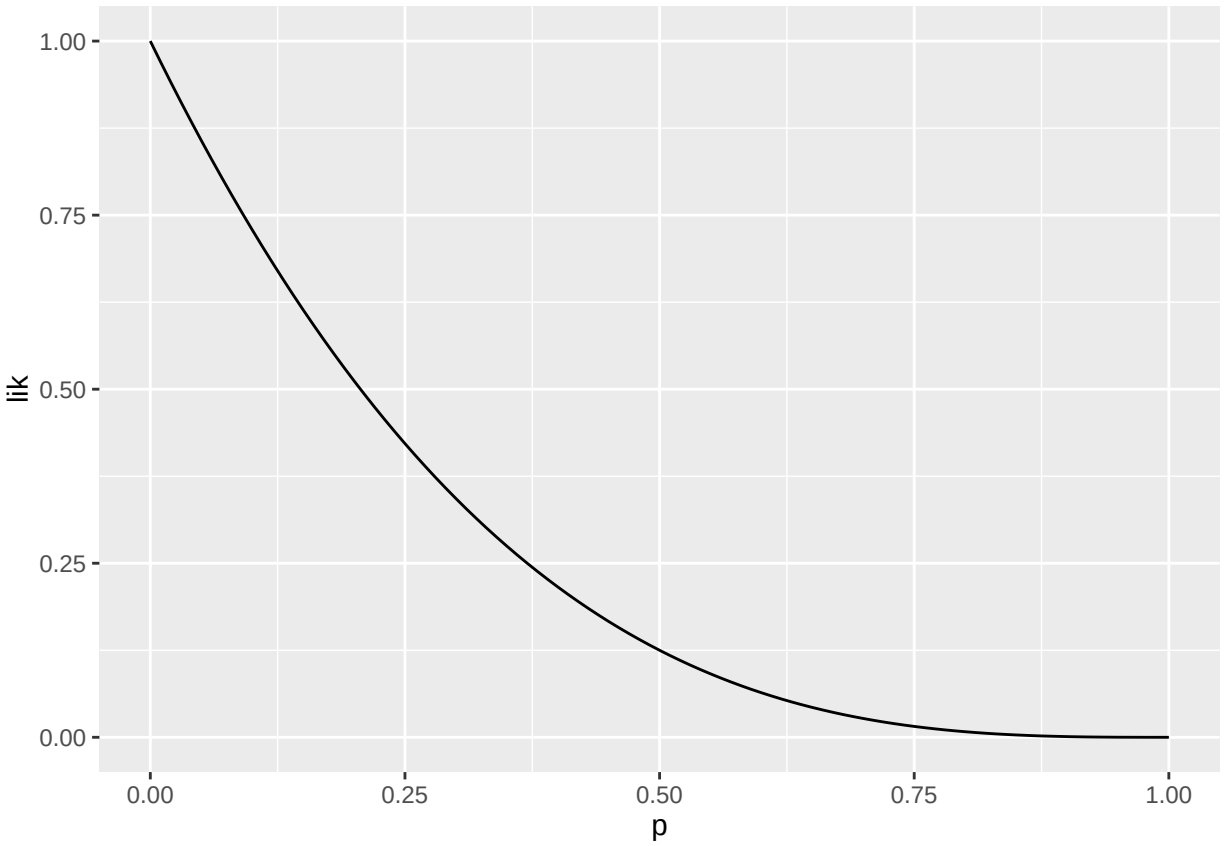
How do your predictions compare to the data? Which characteristics of the data are captured well, and which are captured poorly by your model?

Not too shabby. We did a good job of separating out the fits for low and high magnitude earthquakes, though we have little data for large magnitude earthquakes near the epicenter. We may be slightly underestimating the acceleration for large magnitude earthquakes as well.

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)
qplot(p, lik, geom = "line")
```



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

Well, technically speaking, we only saw Pat attempt three free throws. Clearly, the maximum likelihood estimate for the probablility that Pat has ever, or will ever, make a free throw is zero. We are completely objective scientists after all. Data don't lie