Model diagnostics

In this exercise we'll fit a simple model and apply a variety of model diagnostics to it. Some may be very familiar ...

```
library(armlite) ## for sim(): you can use arm() instead
```

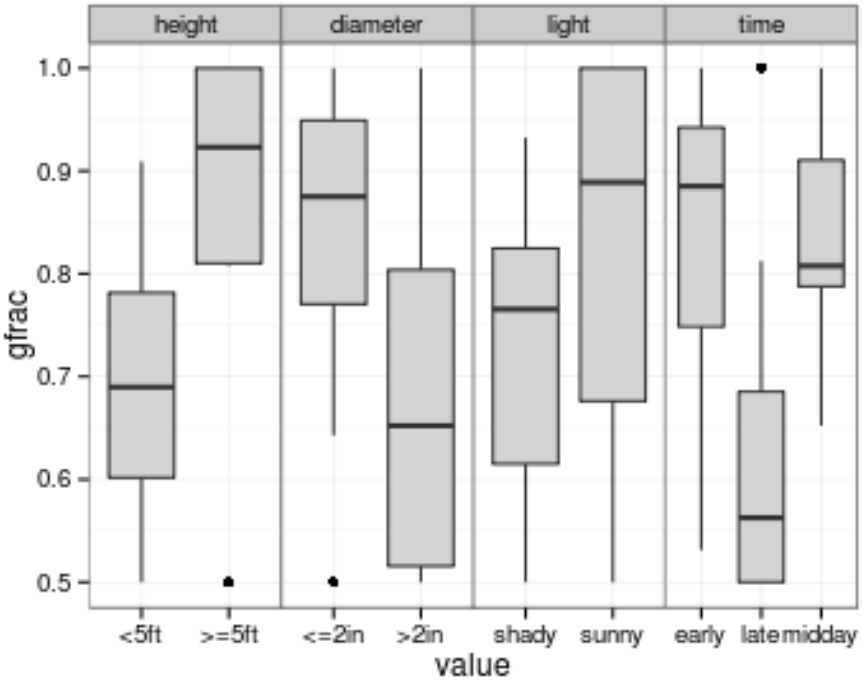
```
##
## armlite (Version 0.00-15, built: 2009-11-22)
## Working directory is /mnt/hgfs/bolker/Documents/meetings/nceas_summer/training/materials/
```

```
library(descr)  ## for LogRegR2
require("reshape2")
## graphics prettiness
library("ggplot2")
theme_set(theme_bw())
library("grid")
zmargin <- theme(panel.margin=unit(0,"lines"))
```

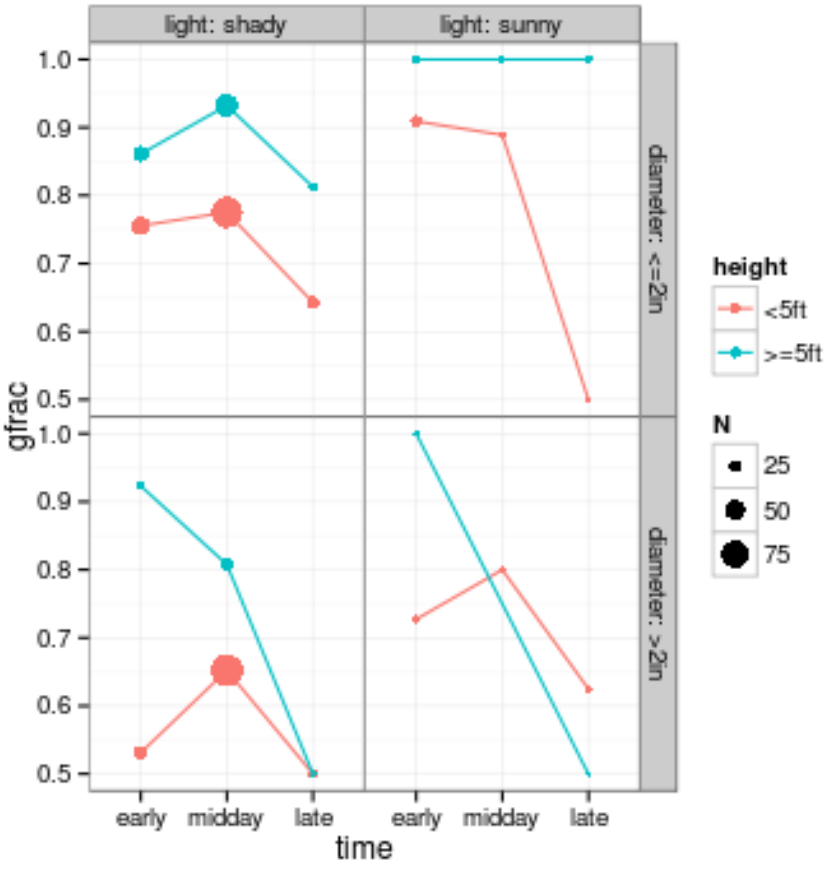
Data on lizard perching behaviour, from the `brglm` package (and before that from McCullagh and Nelder (McCullagh and Nelder 1989), ultimately from Schoener (1970)).

```
lizards <- read.csv("data/lizards.csv")
lizards$time <- factor(lizards$time,
                       levels=c("early","midday","late"))
```

A quick look at the data: response is fraction of *Anolis grahami* lizards found on perches in particular conditions. Plot univariate responses:

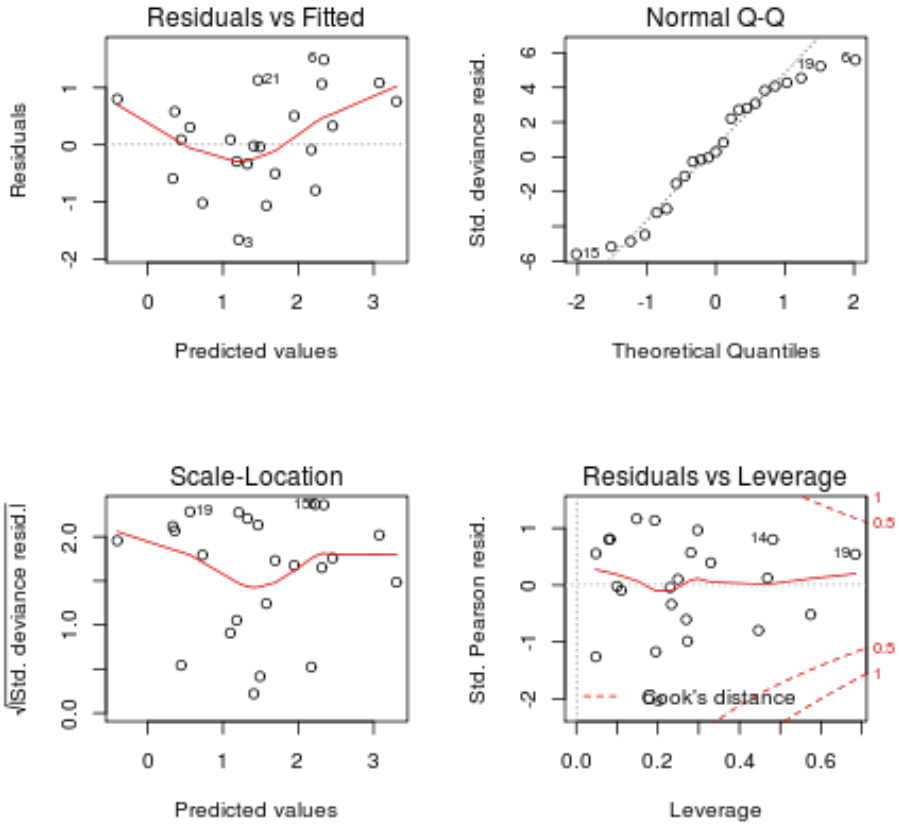



```
(g1 <- ggplot(lizards,
    aes(x=time,y=gfrac,colour=height))+
  geom_point(aes(size=N))+
  geom_line(aes(group=height))+
  facet_grid(diameter~light,labeller=label_both)+zmargin)
```




```
m1 <- glm(gfrac~time+height+light+diameter,
    weights=N,
    family="binomial",
    data=lizards)

op <- par(mfrow=c(2,2))
plot(m1)
```

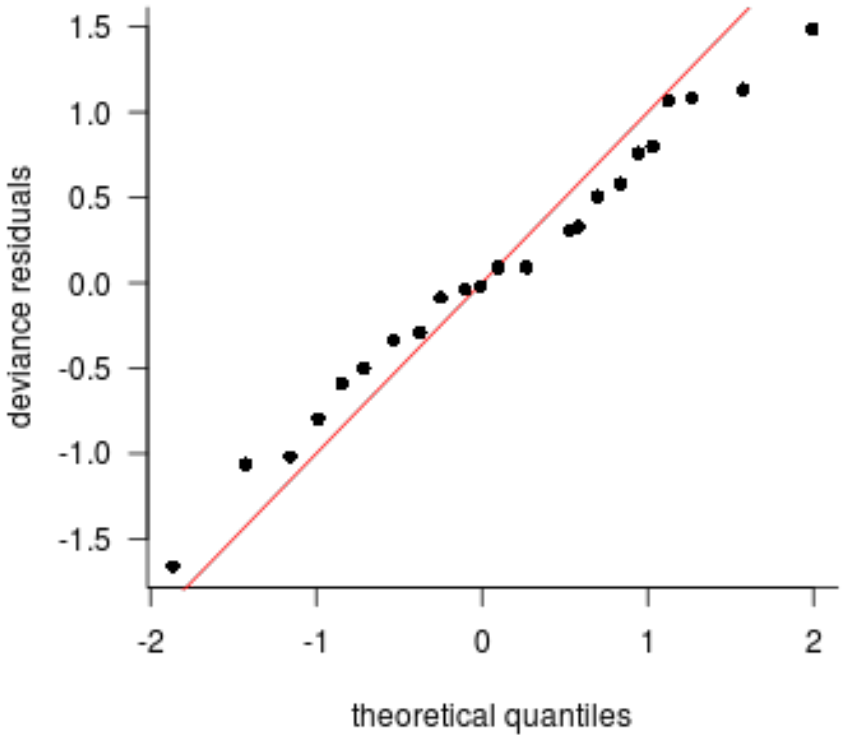



```
par(op) ## restore original parameters
```

test: [@augustin_quantile_2012]

An improved Q-Q plot, from Augustin et al. [@augustin_quantile_2012] by way of the `mgcv` package:

```
library(mgcv)
qq.gam(m1,pch=16)
```



Check for overdispersion:

```
resid.ssq <- sum(residuals(m1,type="pearson")^2)
resid.df <- nrow(lizards)-length(coef(m1))
resid.ssq/resid.df
```

```
## [1] 0.7406
```

Not overdispersed, apparently.

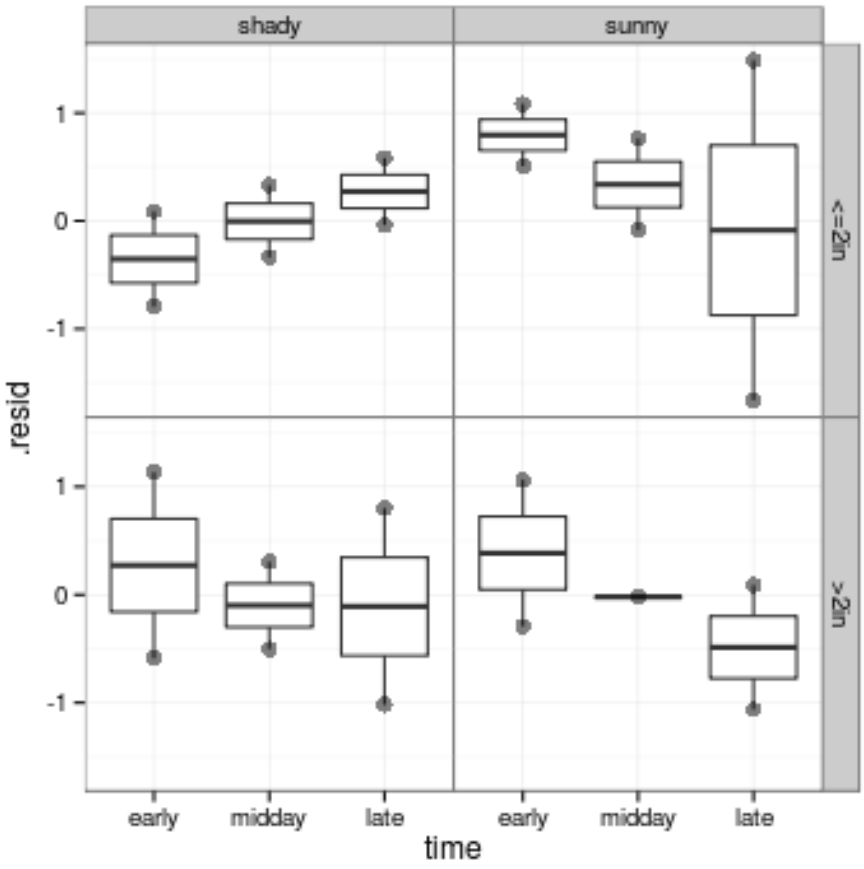
```
LogRegR2(m1)
```

```
## Chi2                 55.9
## Df                   5
## Sig.                 8.532e-11
```

```
## Cox & Snell Index    0.912
## Nagelkerke Index     0.9574
## McFadden's R2        0.7974
```

Use `fortify(model_fit)` to add the standard diagnostics (fitted values, residuals, standardized residuals, ...) to the data from a model

```
m1F <- fortify(m1)
ggplot(m1F,aes(x=time,y=.resid))+geom_boxplot()+
    geom_point(size=3,alpha=0.5)+
    facet_grid(diameter~light)+zmargin
```

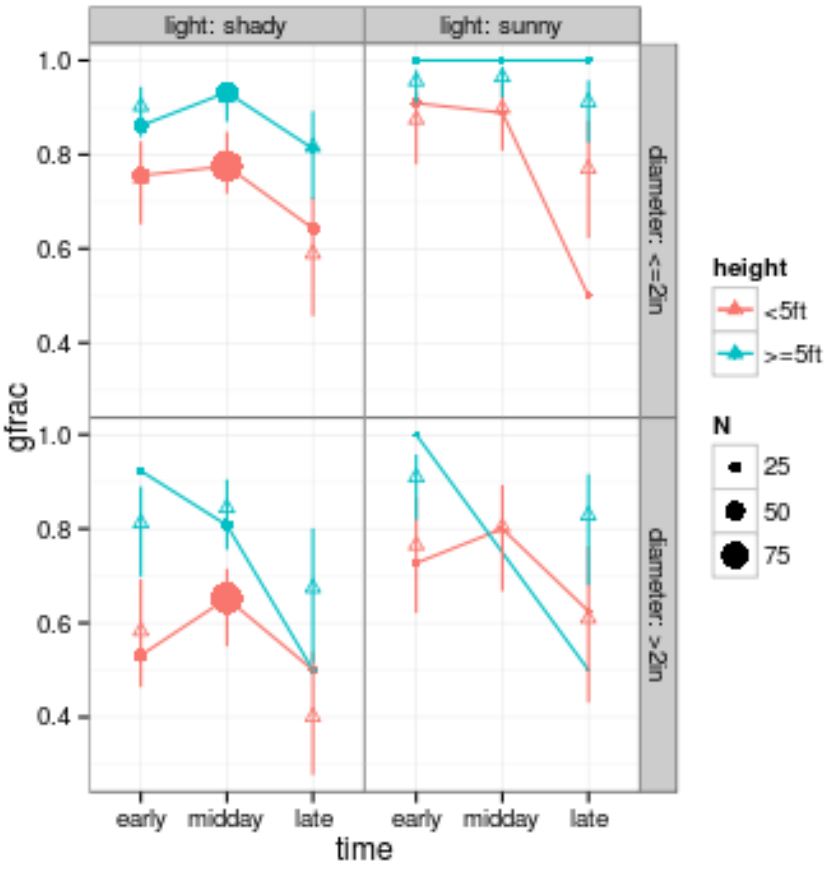


Uh-oh ...

Other tests of distribution are a bit harder.

Or we can plot predicted values.

```
lPred <- predict(m1,se.fit=TRUE)
lizardsX <- transform(lizards,pred=plogis(lPred$fit),
              lwr=plogis(lPred$fit-2*lPred$se.fit),
              upr=plogis(lPred$fit+2*lPred$se.fit))
g1 + geom_pointrange(data=lizardsX,shape=2,
             aes(y=pred,ymin=lwr,ymax=upr))
```



When you have continuous predictors or more complicated/unbalanced situations you will often want to construct your own data frame for predictions, e.g.

```
predframe <- with(lizards,
                  expand.grid(light=levels(light),
                              time=levels(time),
                              diameter="<=2in"))
predict(m1,newdata=predframe)
```

Warning signs of two problems:

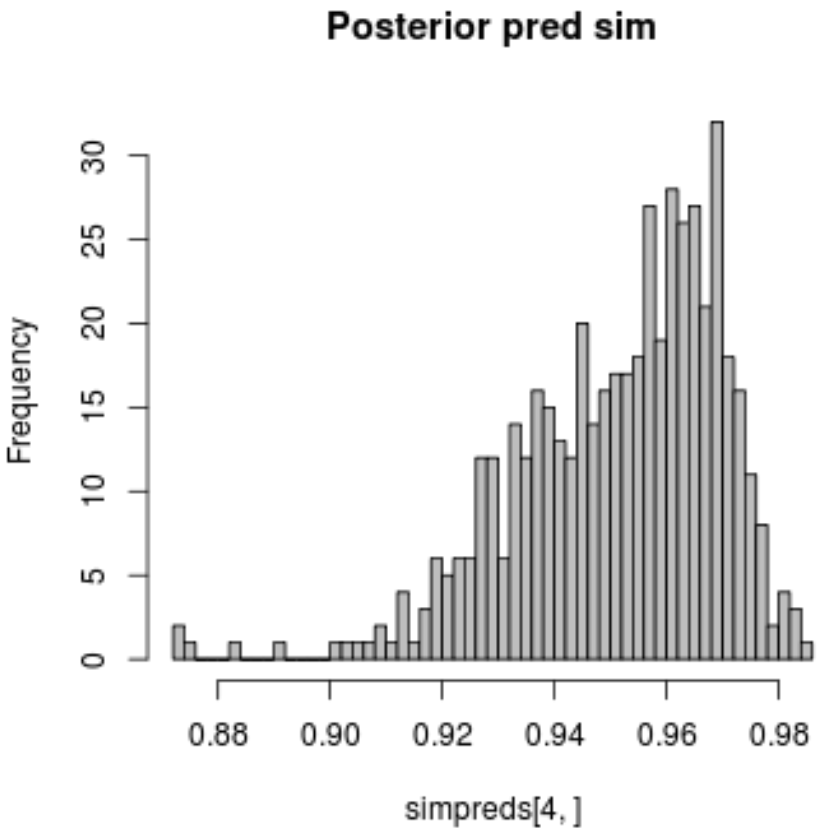
- *complete separation*: all-zero or all-one in some categories (bias-reduced regression via `logistf` or `brglm`, or regularization via Bayesian (`arm::bayesglm`) or other approaches)
- failure of the Wald approximation (*Hauck-Donner effect*, Hauck and Donner (1977))

Posterior predictive simulation

```
betasim <- sim(m1,n.sims=500)
X <- model.matrix(m1)
simpreds <- plogis(X %*% t(betasim$coef))
subset(lizards,gfrac==1.0)
```

```
##     X grahami opalinus height diameter light   time  N gfrac
## 4   4      13        0  >=5ft    <=2in sunny  early 13     1
## 5   5       8        0  >=5ft    <=2in sunny midday  8     1
## 6   6      12        0  >=5ft    <=2in sunny   late 12     1
## 10 10       6        0  >=5ft     >2in sunny  early  6     1
```

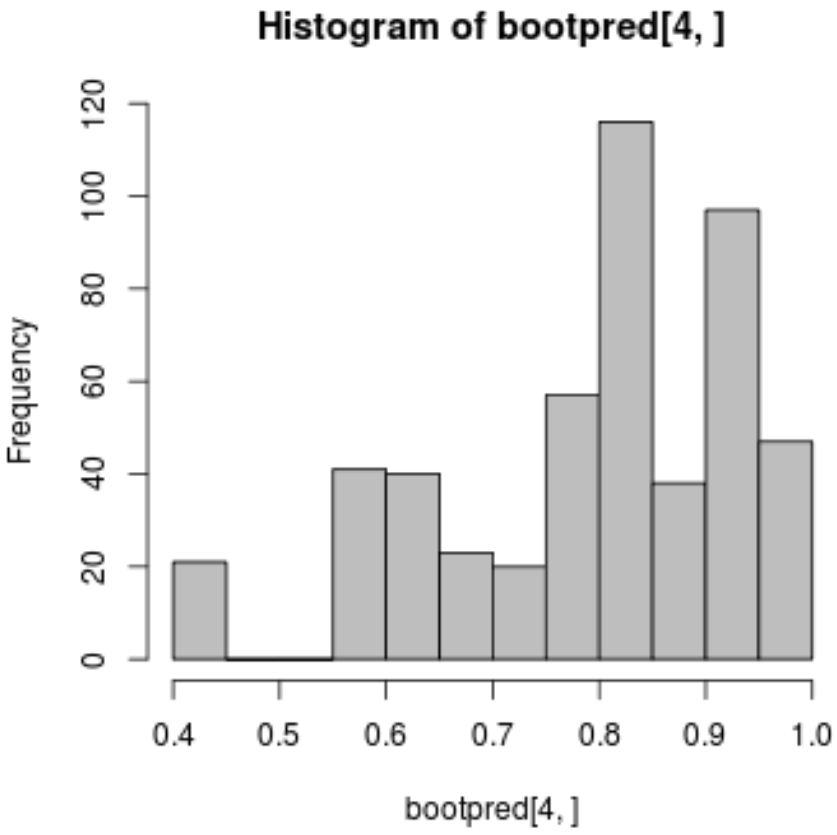
```
hist(simpreds[4,],breaks=50,col="gray",
     main="Posterior pred sim")
```



Bootstrapping

```
bootfun <- function() {

  bsamp <- sample(nrow(lizards),
                  size=nrow(lizards),
                replace=FALSE)
  bmodel <- update(m1,data=lizards[bsamp,])
  bpred <- predict(bmodel,type="response")
}
bootpred <- replicate(500,bootfun())
hist(bootpred[4,],breaks=20,col="gray")
```



Cross-validation

```
library(boot)
```

Need to define a *cost function* `cost(observed, fitted)`; default is avg squared error

```
cost <- function(r, pi = 0) mean(abs(r-pi)) ## use mean abs dev
cv1 <- cv.glm(lizards,m1)
str(cv1)
```

```
## List of 4
##  $ call : language cv.glm(data = lizards, glmfit = m1)
##  $ K    : num 23
```

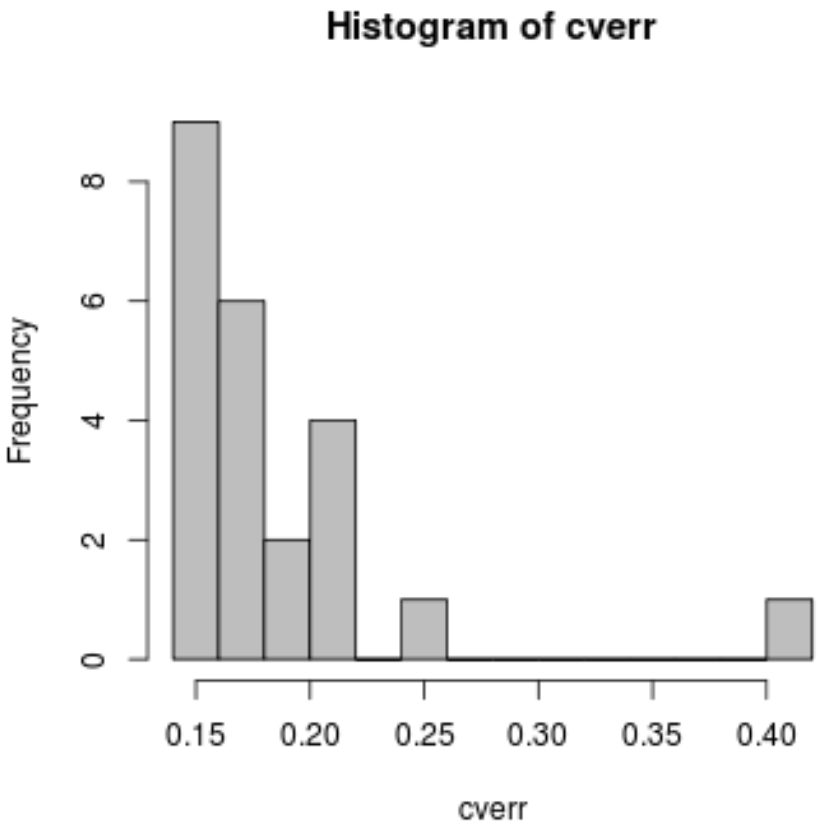
```
##  $ delta: num [1:2] 0.016 0.0159
##  $ seed : int [1:626] 403 258 1299843944 923835171 1909356612 1455123722 -1505139325 868
```

```
cv1$delta
```

```
## [1] 0.01602 0.01594
```

Or do it by hand:

```
cverr <- numeric(nrow(lizards))
for (i in 1:nrow(lizards)) {
  cvdata <- lizards[-i,]
  cvmodel <- update(m1,data=cvdata)
  predval <- predict(cvmodel,newdata=lizards[i,],
                     type="response")
  cverr[i] <- cost(lizards$gfrac,predval)
}
hist(cverr,breaks=10,col="gray")
```




```
mean(cverr)
```

```
## [1] 0.1844
```

This is *leave-one-out* cross-validation: K-fold is usually better (but maybe worth using `cv.glm` instead)

Exercises

- Change the model to incorporate two-way interactions (`m2 <- update(m1,.~.^2)`) and see if that seems to fix any problems we found in the model. Compare this with the statistical significance of the added terms (`summary(m2)` or `drop1(m2,test="Chisq")`)

Hauck, Walter W., and Allan Donner. 1977. "Wald's Test as Applied to Hypotheses in Logit Analysis." *Journal of the American Statistical Association* 72 (360) (dec): 851–853. doi:10.2307/2286473. http://www.jstor.org/stable/2286473.

McCullagh, P., and J. A. Nelder. 1989. *Generalized Linear Models*. London: Chapman and Hall.

Schoener, Thomas W. 1970. "Nonsynchronous Spatial Overlap of Lizards in Patchy Habitats." *Ecology* 51 (3) (may): 408–418. doi:10.2307/1935376. http://www.jstor.org/stable/1935376.