Lesson 7: Linear regression

Lesson 7.2

Data

As an example of linear regression, we'll look at the Leinhardt data from the `car` package in `R`.

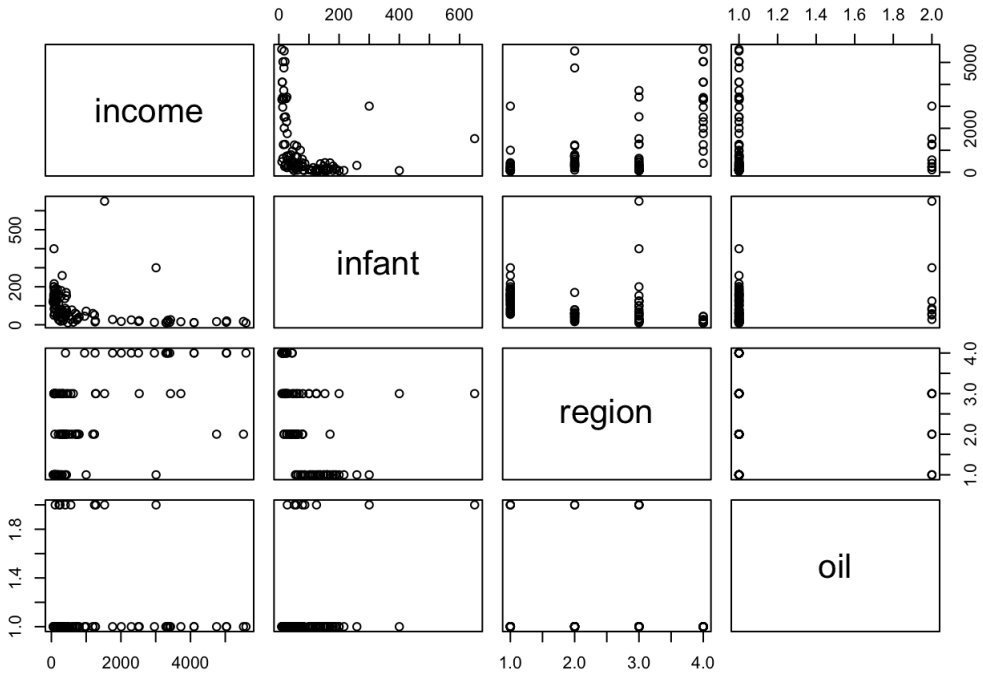
```
library("car")
data("Leinhardt")
?Leinhardt
head(Leinhardt)
```

```
##           income infant   region oil
## Australia   3426   26.7     Asia  no
## Austria     3350   23.7   Europe  no
## Belgium     3346   17.0   Europe  no
## Canada      4751   16.8 Americas  no
## Denmark     5029   13.5   Europe  no
## Finland     3312   10.1   Europe  no
```

```
str(Leinhardt)
```

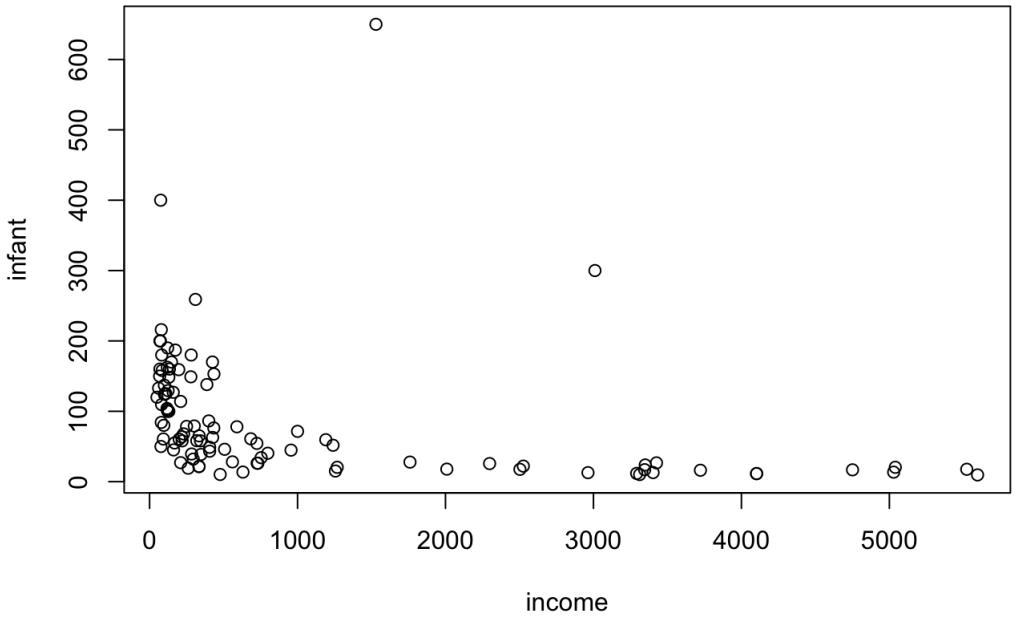
```
## 'data.frame':    105 obs. of  4 variables:
##  $ income: int  3426 3350 3346 4751 5029 3312 3403 5040 2009 2298 ...
##  $ infant: num  26.7 23.7 17 16.8 13.5 10.1 12.9 20.4 17.8 25.7 ...
##  $ region: Factor w/ 4 levels "Africa","Americas",..: 3 4 4 2 4 4 4 4 4 4 ...
##  $ oil   : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...
```

```
pairs(Leinhardt)
```

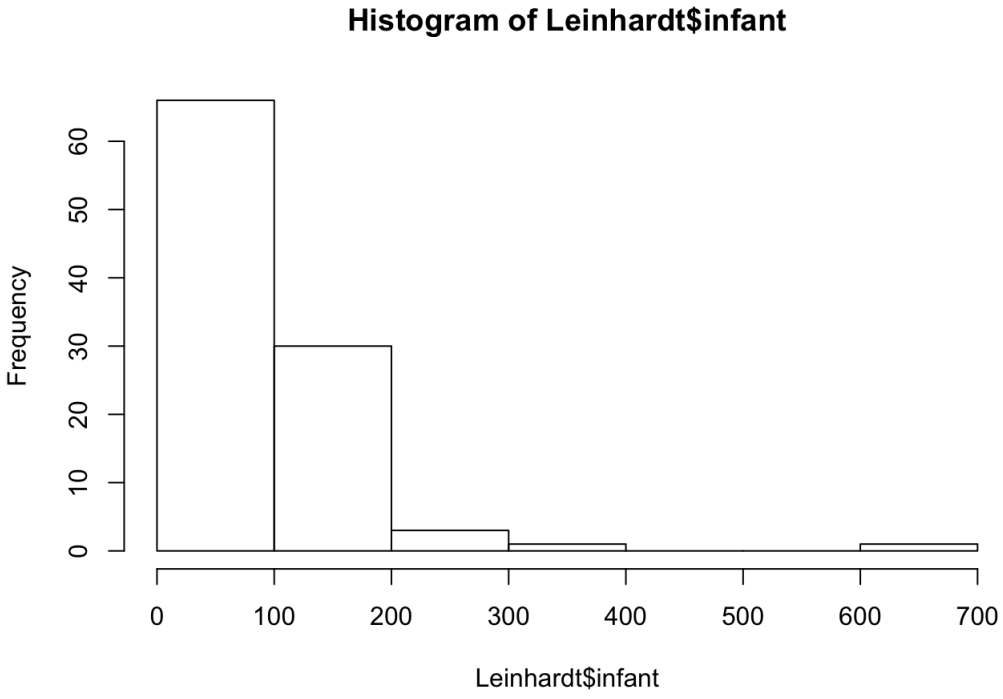


We'll start with a simple linear regression model that relates infant mortality to per capita income.

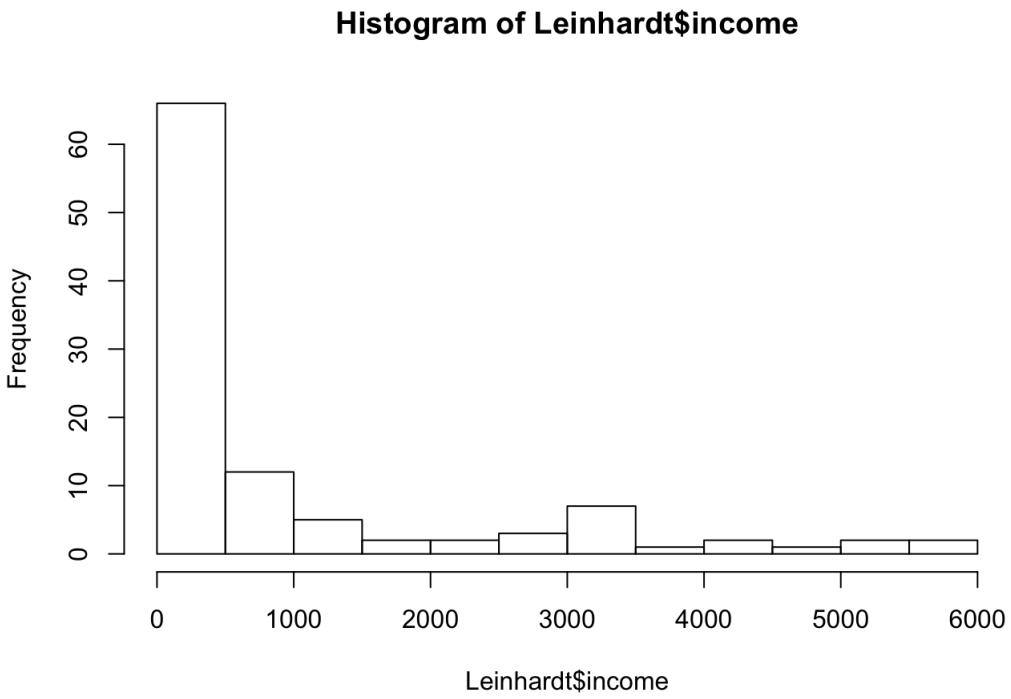
```
plot(infant ~ income, data=Leinhardt)
```



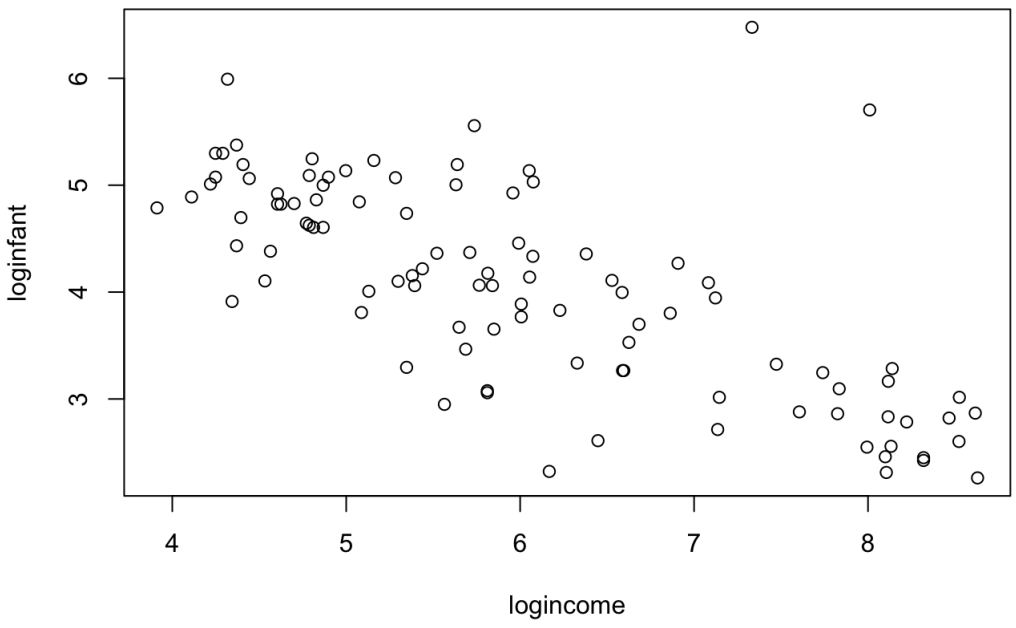

```
hist(Leinhardt$infant)
```




```
hist(Leinhardt$income)
```




```
Leinhardt$loginfant = log(Leinhardt$infant)
Leinhardt$logincome = log(Leinhardt$income)

plot(loginfant ~ logincome, data=Leinhardt)
```



Since infant mortality and per capita income are positive and right-skewed quantities, we consider modeling them on the logarithmic scale. A linear model appears much more appropriate on this scale.

Modeling

The reference Bayesian analysis (with a noninformative prior) is available directly in `R`.

```
lmod = lm(loginfant ~ logincome, data=Leinhardt)
summary(lmod)
```

```
##
## Call:
## lm(formula = loginfant ~ logincome, data = Leinhardt)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.66694 -0.42779 -0.02649  0.30441  3.08415
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.14582    0.31654  22.575   <2e-16 ***
## logincome   -0.51179    0.05122  -9.992   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6867 on 99 degrees of freedom
##   (4 observations deleted due to missingness)
## Multiple R-squared:  0.5021, Adjusted R-squared:  0.4971
## F-statistic: 99.84 on 1 and 99 DF,  p-value: < 2.2e-16
```

Leson 7.3

Model in `JAGS`

Now we'll fit this model in `JAGS`. A few countries have missing values, and for simplicity, we will omit those.

```
dat = na.omit(Leinhardt)

library("rjags")
```

```
## Loading required package: coda
```

```
## Linked to JAGS 4.2.0
```

```
## Loaded modules: basemod,bugs
```

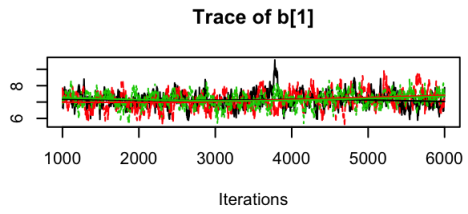
```
mod1_string = " model {
    for (i in 1:n) {
        y[i] ~ dnorm(mu[i], prec)
        mu[i] = b[1] + b[2]*log_income[i]
    }

    for (i in 1:2) {
        b[i] ~ dnorm(0.0, 1.0/1.0e6)
    }

    prec ~ dgamma(5/2.0, 5*10.0/2.0)
    sig2 = 1.0 / prec
    sig = sqrt(sig2)
} "

set.seed(72)
data1_jags = list(y=dat$loginfant, n=nrow(dat),
              log_income=dat$logincome)

params1 = c("b", "sig")

inits1 = function() {
    inits = list("b"=rnorm(2,0.0,100.0), "prec"=rgamma(1,1.0,1.0))
}

mod1 = jags.model(textConnection(mod1_string), data=data1_jags, inits=inits1, n.chains=3)
update(mod1, 1000) # burn-in

mod1_sim = coda.samples(model=mod1,
                        variable.names=params1,
                        n.iter=5000)

mod1_csim = do.call(rbind, mod1_sim) # combine multiple chains
```

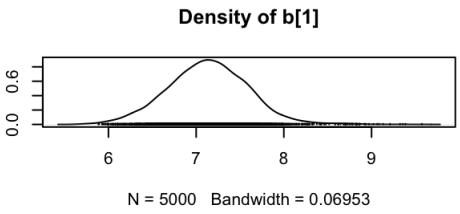
Lesson 7.4

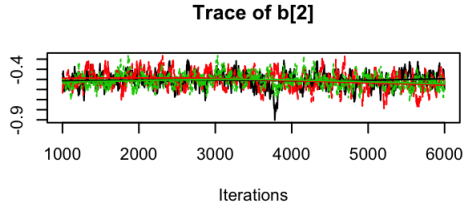
MCMC convergence

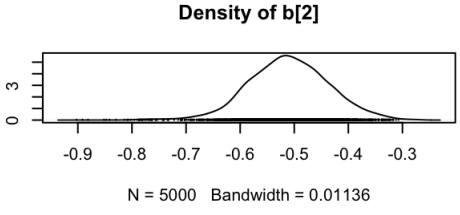
Before we check the inferences from the model, we should perform convergence diagnostics for our Markov chains.

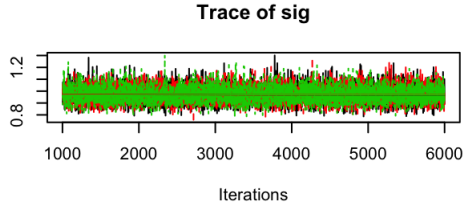
```
plot(mod1_sim)
```

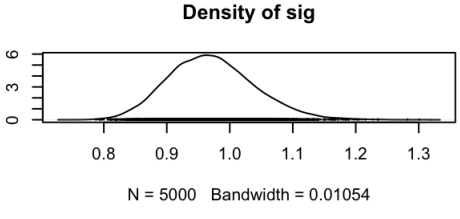












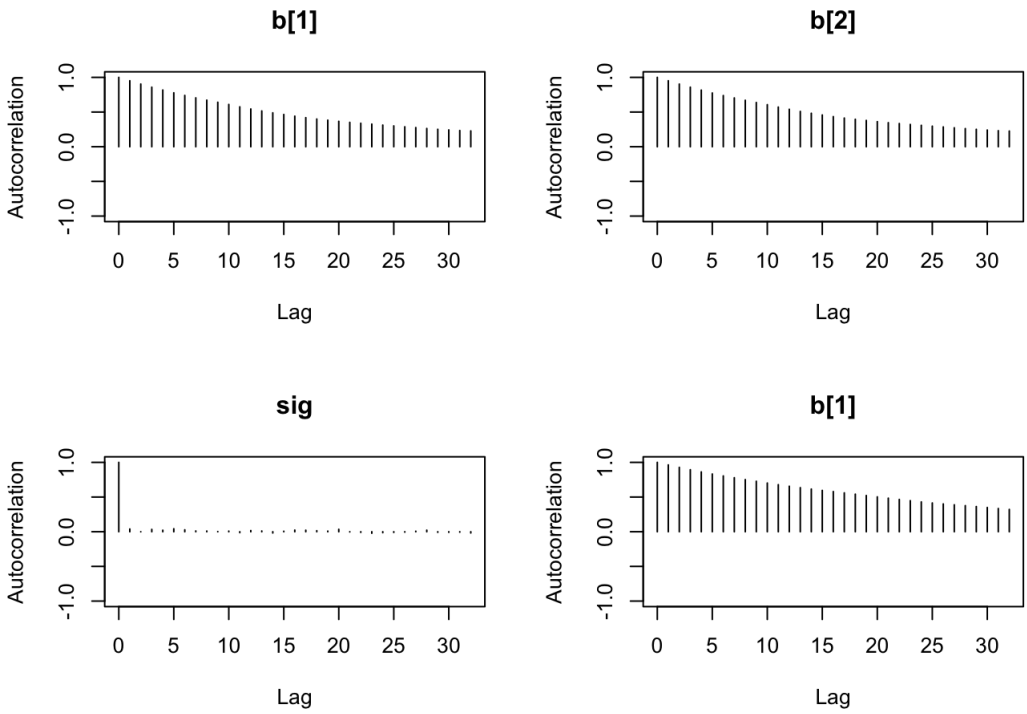
```
gelman.diag(mod1_sim)
```

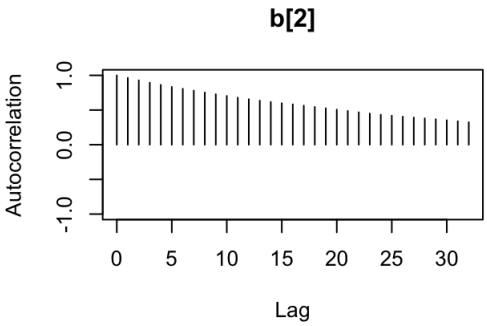
```
## Potential scale reduction factors:
##
##      Point est. Upper C.I.
## b[1]       1.01       1.03
## b[2]       1.01       1.03
## sig        1.00       1.00
##
## Multivariate psrf
##
## 1.01
```

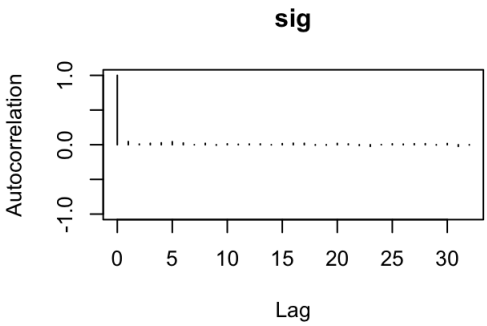
```
autocorr.diag(mod1_sim)
```

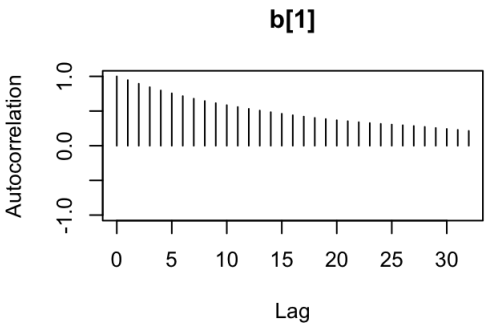
```
##              b[1]       b[2]          sig
## Lag 0   1.00000000 1.00000000  1.000000000
## Lag 1   0.95358813 0.95355035  0.029614363
## Lag 5   0.78817893 0.78777679  0.028026775
## Lag 10  0.63205187 0.63211729  0.009436514
## Lag 50  0.09129094 0.09266128 -0.008772217
```

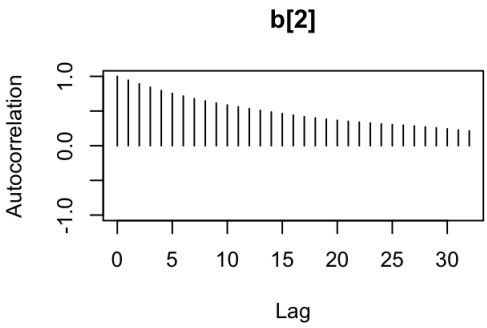
```
autocorr.plot(mod1_sim)
```

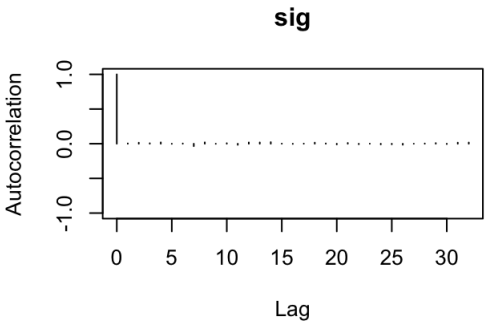













```
effectiveSize(mod1_sim)
```

```
##       b[1]       b[2]        sig
##   367.5878   370.5843 12313.6822
```

We can get a posterior summary of the parameters in our model.

```
summary(mod1_sim)
```

```
##
## Iterations = 1001:6000
## Thinning interval = 1
## Number of chains = 3
## Sample size per chain = 5000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##         Mean      SD  Naive SE Time-series SE
## b[1]  7.1345 0.46174 0.0037701      0.0249762
## b[2] -0.5102 0.07473 0.0006101      0.0040391
## sig   0.9712 0.06811 0.0005561      0.0006208
##
## 2. Quantiles for each variable:
##
##         2.5%     25%     50%     75%   97.5%
## b[1]  6.2271  6.8331  7.1348  7.4346  8.0456
## b[2] -0.6575 -0.5590 -0.5106 -0.4607 -0.3623
## sig   0.8490  0.9231  0.9674  1.0143  1.1153
```

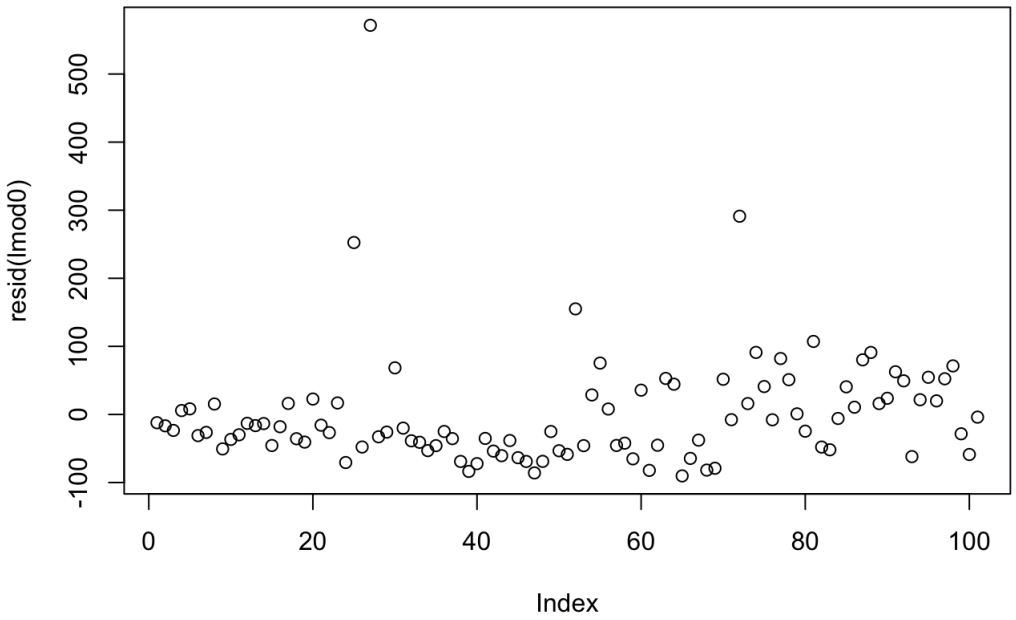
Don't forget that these results are for a regression model relating the logarithm of infant mortality to the logarithm of income.

Residual checks

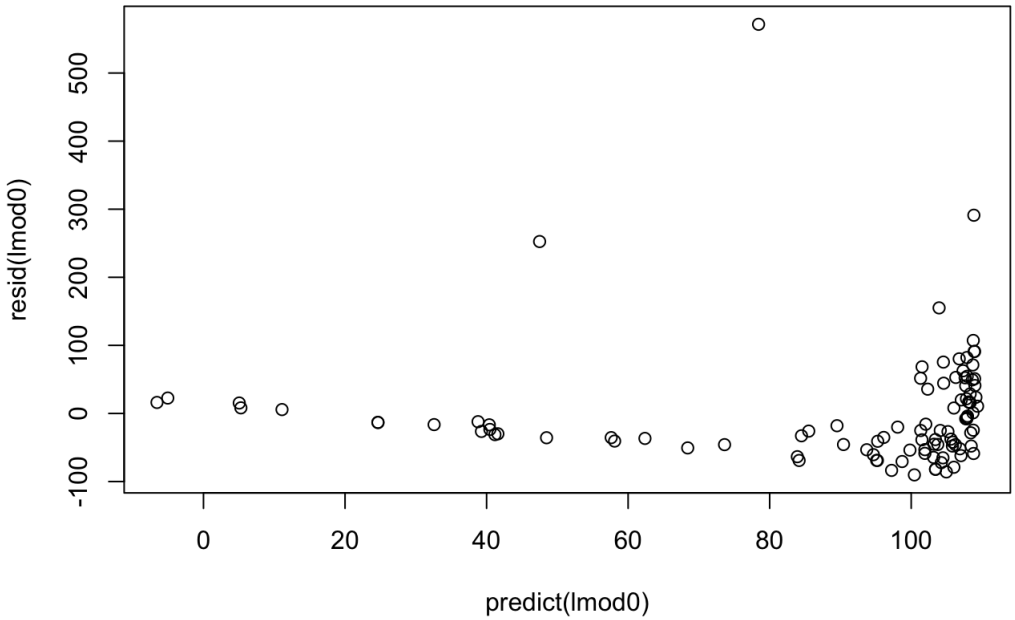
Checking residuals (the difference between the response and the model's prediction for that value) is important with linear models since residuals can reveal violations of the assumptions we made to specify the model. In particular, we are looking for any sign that the model is not linear, normally distributed, or that the observations are not independent (conditional on covariates).

First, let's look at what would have happened if we fit the reference linear model to the un-transformed variables.

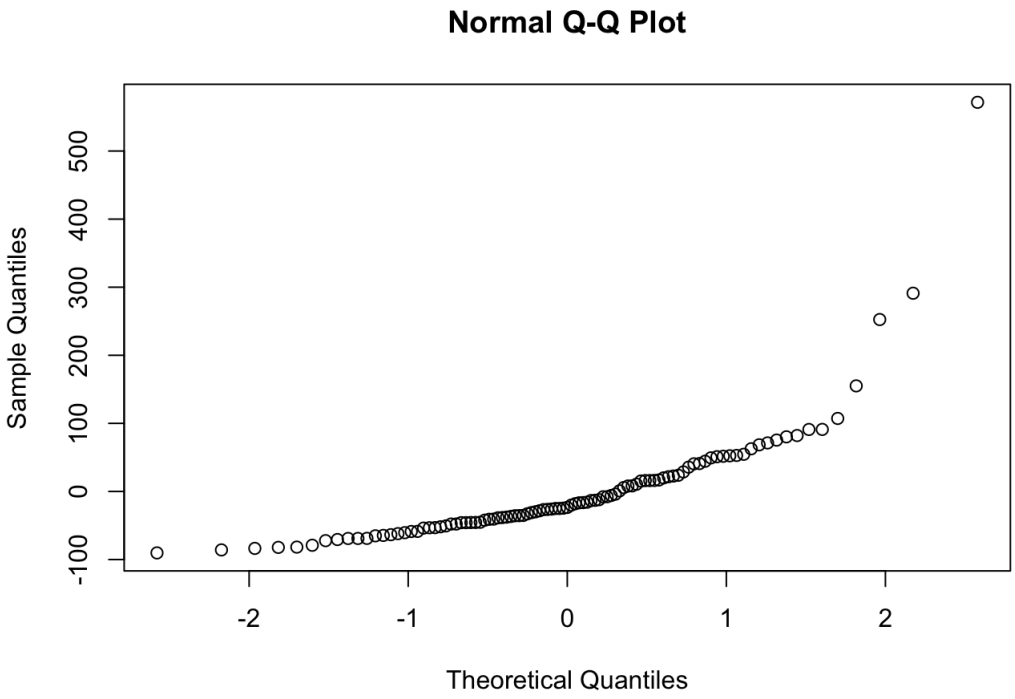
```
lmod0 = lm(infant ~ income, data=Leinhardt)
plot(resid(lmod0)) # to check independence (looks okay)
```




```
plot(predict(lmod0), resid(lmod0)) # to check for linearity, constant variance (looks bad)
```




```
qqnorm(resid(lmod0)) # to check Normality assumption (we want this to be a straight line)
```



Now let's return to our model fit to the log-transformed variables. In a Bayesian model, we have distributions for residuals, but we'll simplify and look only at the residuals evaluated at the posterior mean of the parameters.

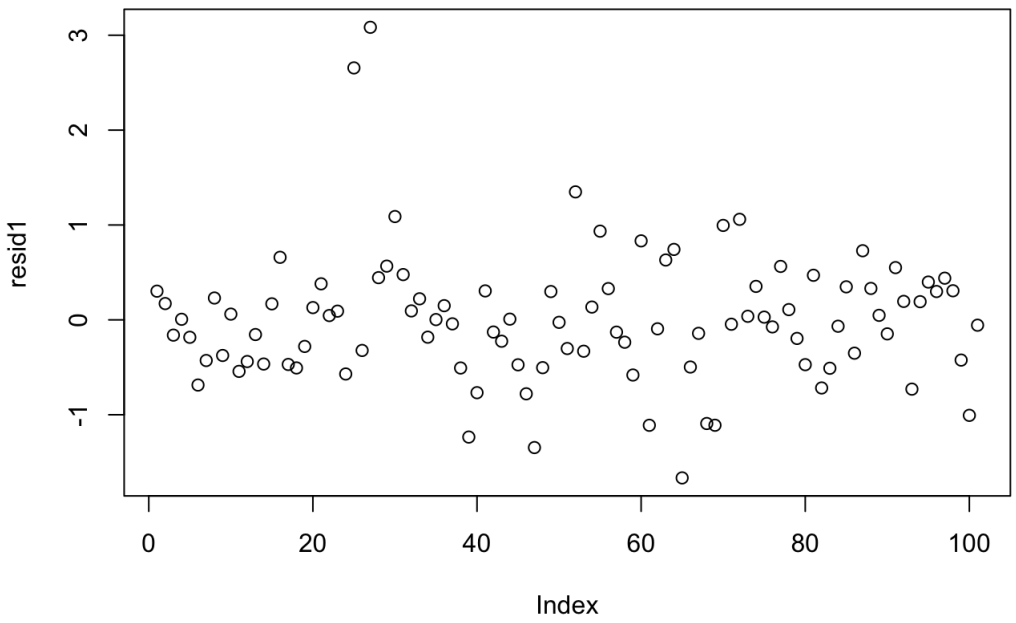
```
X = cbind(rep(1.0, data1_jags$n), data1_jags$log_income)
head(X)
```

```
##      [,1]     [,2]
## [1,]    1 8.139149
## [2,]    1 8.116716
## [3,]    1 8.115521
## [4,]    1 8.466110
## [5,]    1 8.522976
## [6,]    1 8.105308
```

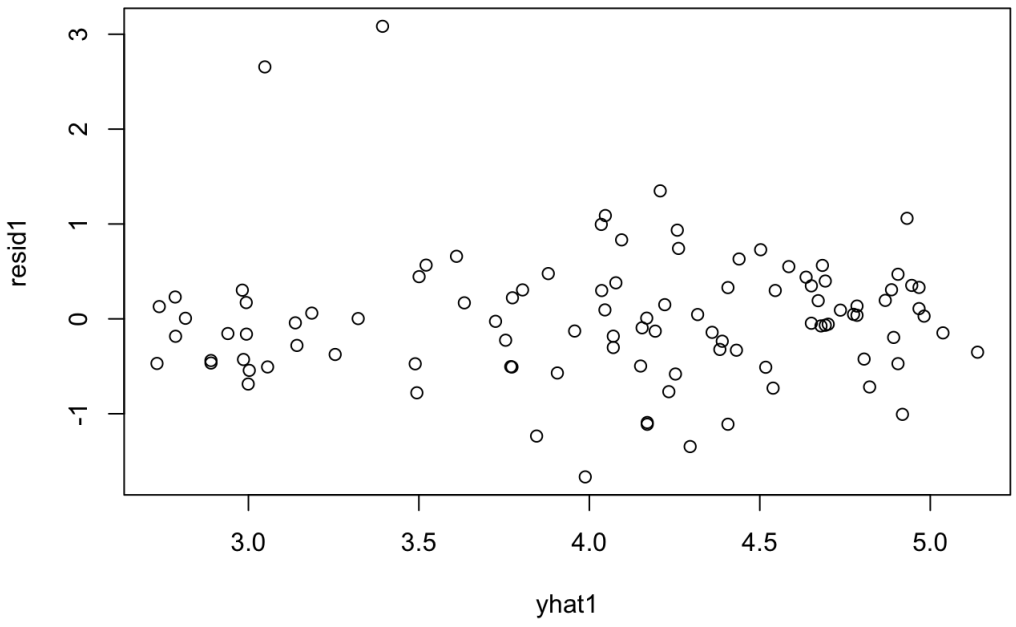
```
(pm_params1 = colMeans(mod1_csim)) # posterior mean
```

```
##       b[1]       b[2]        sig
##  7.1344761 -0.5101693  0.9711672
```

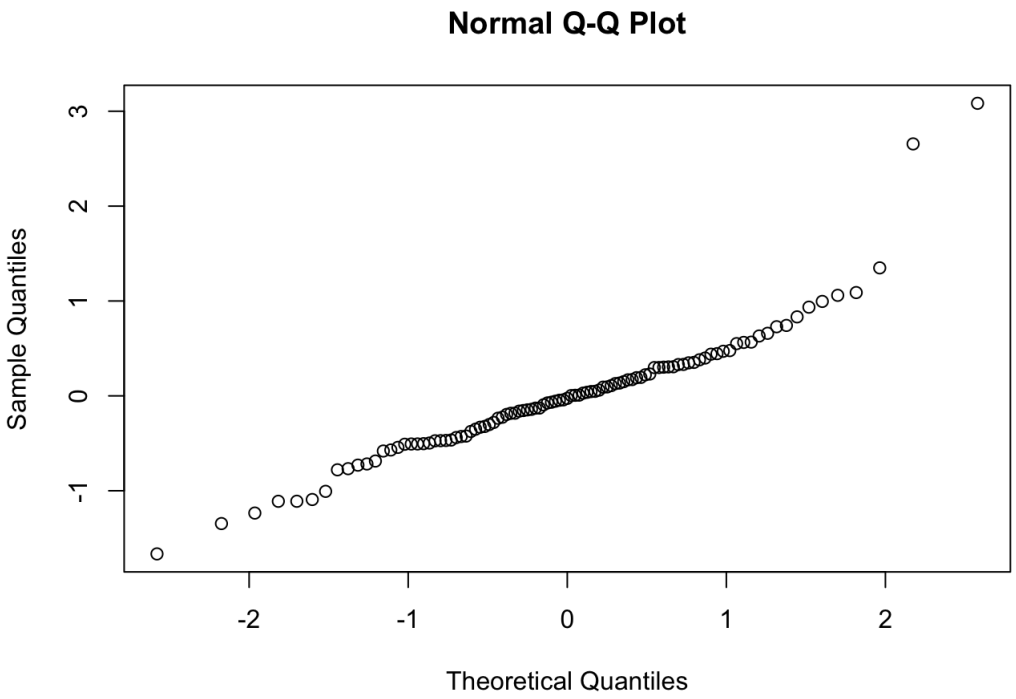
```
yhat1 = drop(X %*% pm_params1[1:2])
resid1 = data1_jags$y - yhat1
plot(resid1) # against data index
```



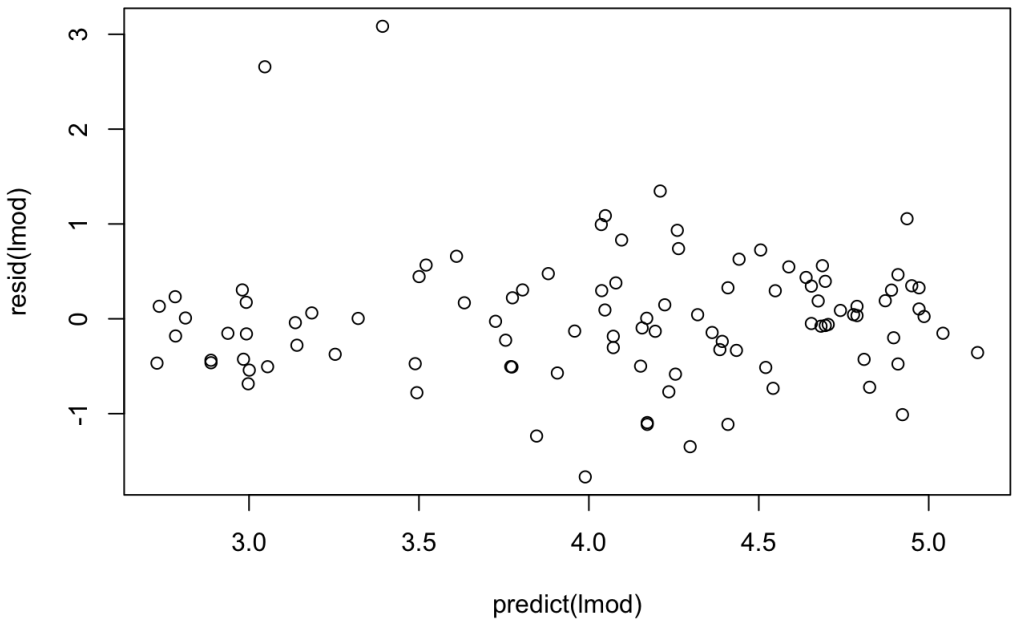

```
plot(yhat1, resid1) # against predicted values
```




```
qqnorm(resid1) # checking normality of residuals
```




```
plot(predict(lmod), resid(lmod)) # to compare with reference linear model
```




```
rownames(dat)[order(resid1, decreasing=TRUE)[1:5]] # which countries have the largest positive residuals?
```

```
## [1] "Saudi.Arabia" "Libya"        "Zambia"       "Brazil"
## [5] "Afganistan"
```

The residuals look pretty good here (no patterns, shapes) except for two strong outliers, Saudi Arabia and Libya. When outliers appear, it is a good idea to double check that they are not just errors in data entry. If the values are correct, you may reconsider whether these data points really are representative of the data you are trying to model. If you conclude that they are not (for example, they were recorded on different years), you may be able to justify dropping these data points from the data set.

If you conclude that the outliers are part of data and should not be removed, we have several modeling options to accommodate them. We will address these in the next segment.

Lesson 7.5

In the previous segment, we saw two outliers in the model relating the logarithm of infant mortality to the logarithm of income. Here we will discuss options for when we conclude that these outliers belong in the data set.

Additional covariates

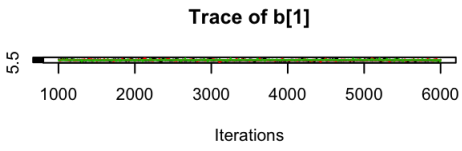
The first approach is to look for additional covariates that may be able to explain the outliers. For example, there could be a number of variables that provide information about infant mortality above and beyond what income provides.

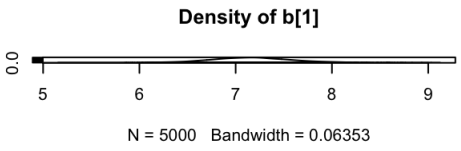
Looking back at our data, there are two variables we haven't used yet: `region` and `oil`. The `oil` variable indicates oil-exporting countries. Both Saudi Arabia and Libya are oil-exporting countries, so perhaps this might explain part of the anomaly.

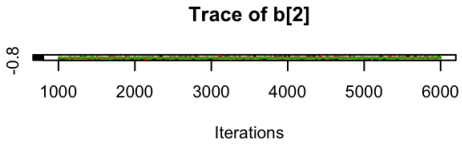
```
library("rjags")

mod2_string = " model {
    for (i in 1:length(y)) {
        y[i] ~ dnorm(mu[i], prec)
        mu[i] = b[1] + b[2]*log_income[i] + b[3]*is_oil[i]
    }

    for (i in 1:3) {
        b[i] ~ dnorm(0.0, 1.0/1.0e6)
    }

    prec ~ dgamma(5/2.0, 5*10.0/2.0)
    sig = sqrt( 1.0 / prec )
} "


set.seed(73)
data2_jags = list(y=dat$loginfant, log_income=dat$logincome,
                  is_oil=as.numeric(dat$oil=="yes"))
data2_jags$is_oil

params2 = c("b", "sig")

inits2 = function() {
    inits = list("b"=rnorm(3,0.0,100.0), "prec"=rgamma(1,1.0,1.0))
}

mod2 = jags.model(textConnection(mod2_string), data=data2_jags, inits=inits2, n.chains=3)
update(mod2, 1e3) # burn-in

mod2_sim = coda.samples(model=mod2,
                        variable.names=params2,
                        n.iter=5e3)

mod2_csim = as.mcmc(do.call(rbind, mod2_sim)) # combine multiple chains
```

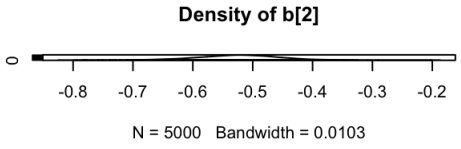
As usual, check the convergence diagnostics.

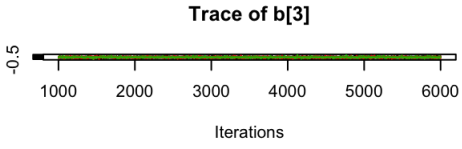
```
plot(mod2_sim)
```

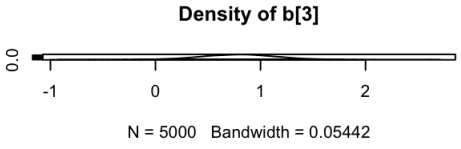


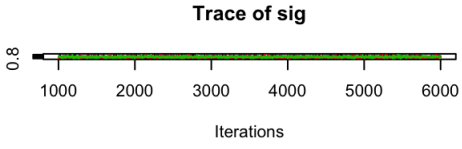


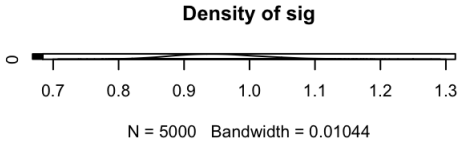




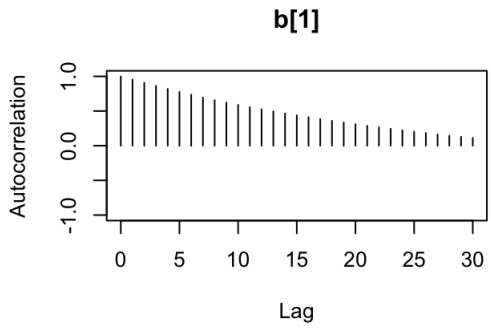


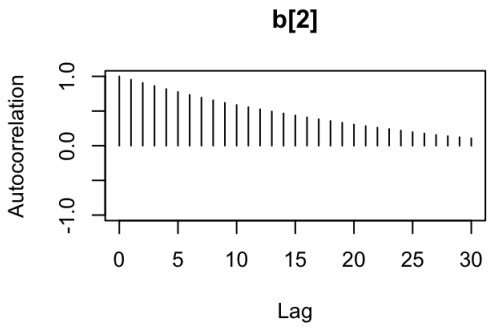


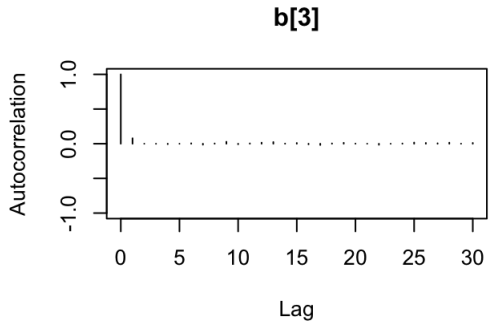


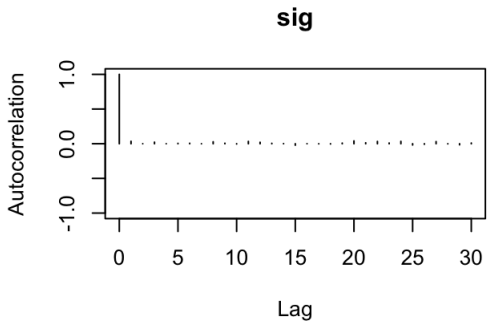


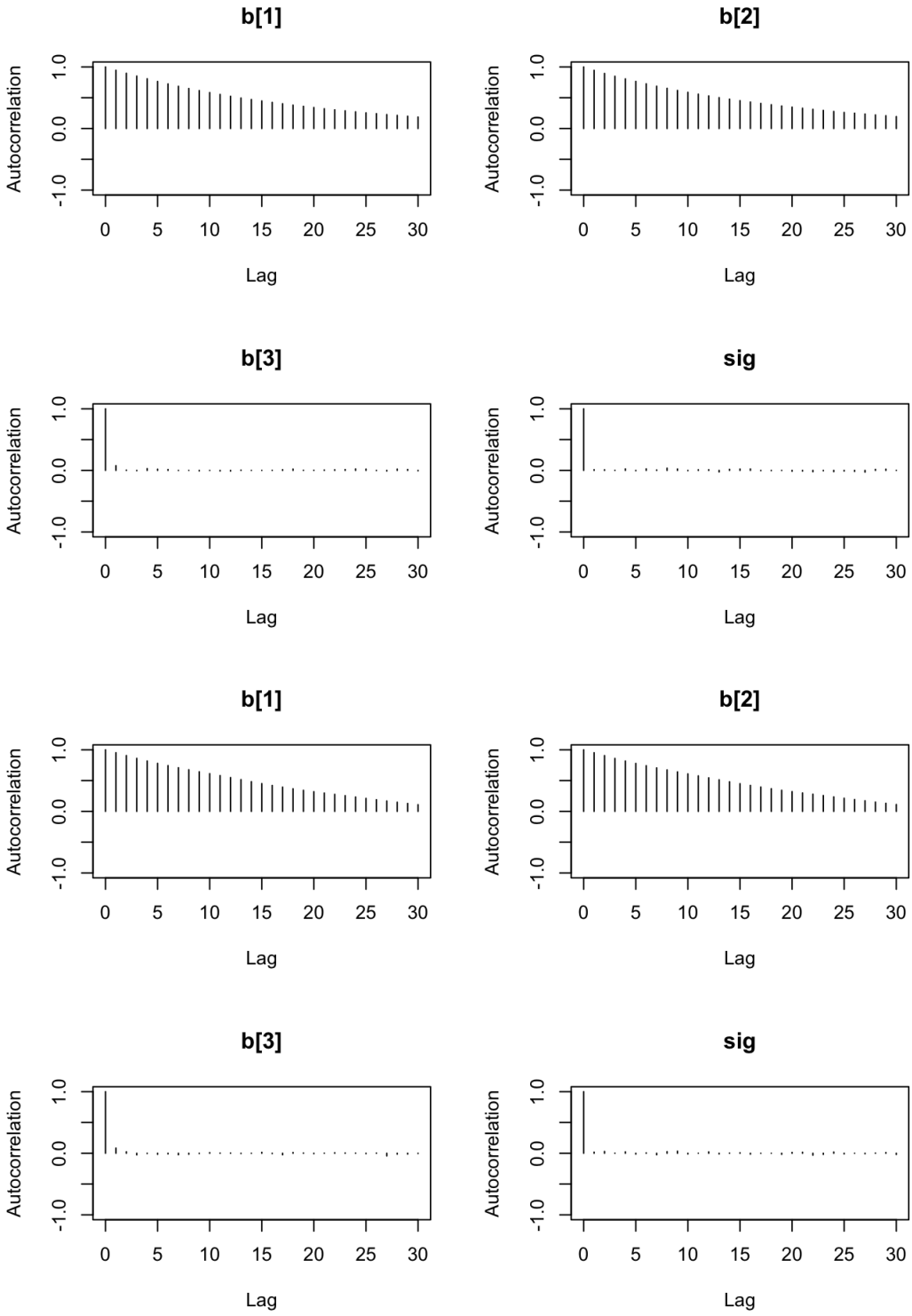

```
gelman.diag(mod2_sim)
autocorr.diag(mod2_sim)

autocorr.plot(mod2_sim)
```












```
effectiveSize(mod2_sim)
```

We can get a posterior summary of the parameters in our model.

```
summary(mod2_sim)
```

```
##
## Iterations = 1001:6000
## Thinning interval = 1
## Number of chains = 3
## Sample size per chain = 5000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##         Mean      SD  Naive SE Time-series SE
## b[1]  7.1691 0.42846 0.0034984      0.0217879
## b[2] -0.5260 0.06936 0.0005663      0.0035598
## b[3]  0.7934 0.35131 0.0028685      0.0030929
## sig   0.9531 0.06805 0.0005556      0.0005817
##
## 2. Quantiles for each variable:
##
##         2.5%     25%     50%     75%   97.5%
## b[1]  6.3480  6.8886  7.1582  7.4382  8.0577
## b[2] -0.6696 -0.5693 -0.5236 -0.4802 -0.3927
## b[3]  0.1018  0.5552  0.7957  1.0273  1.4936
## sig   0.8316  0.9057  0.9486  0.9961  1.0988
```

It looks like there is a positive relationship between oil-production and log-infant mortality. Because these data are merely observational, we cannot say that oil-production causes an increase in infant mortality (indeed that most certainly isn't the case), but we can say that they are positively correlated.

Now let's check the residuals.

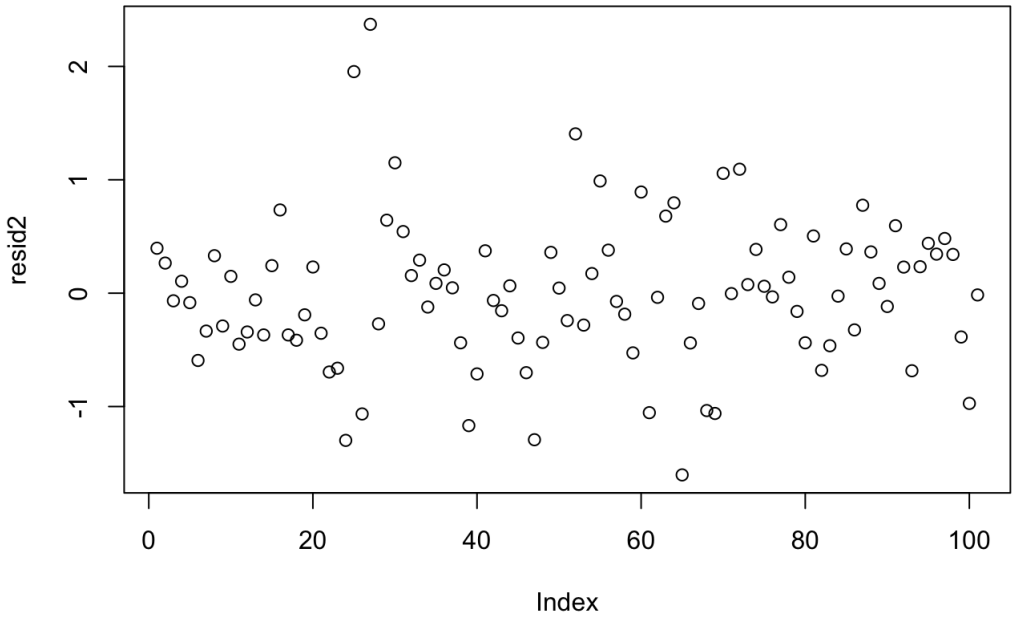
```
X2 = cbind(rep(1.0, data1_jags$n), data2_jags$log_income, data2_jags$is_oil)
head(X2)
```

```
##      [,1]     [,2] [,3]
## [1,]    1 8.139149    0
## [2,]    1 8.116716    0
## [3,]    1 8.115521    0
## [4,]    1 8.466110    0
## [5,]    1 8.522976    0
## [6,]    1 8.105308    0
```

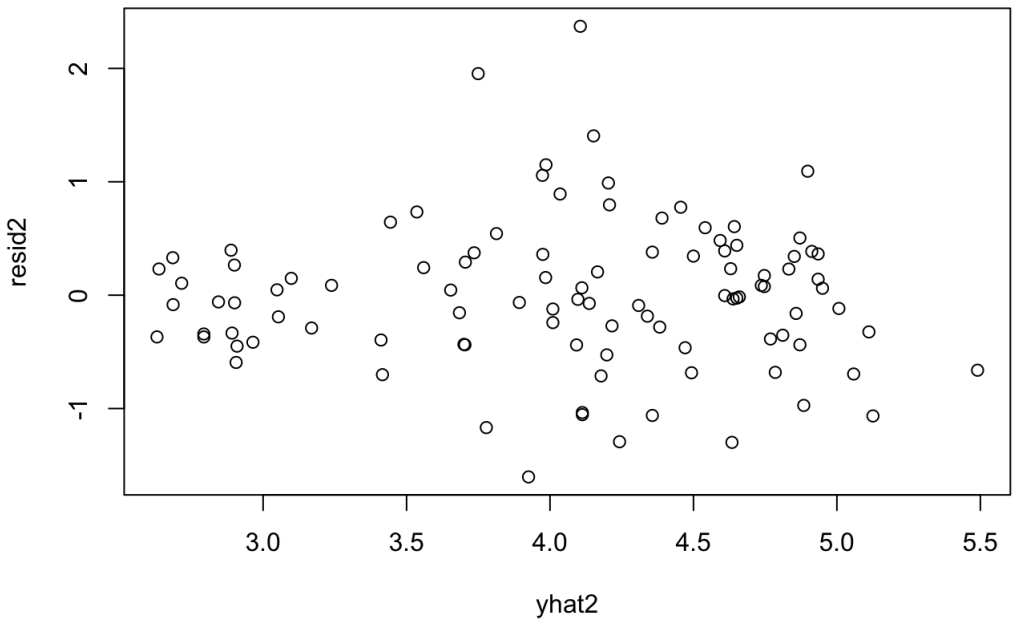
```
(pm_params2 = colMeans(mod2_csim)) # posterior mean
```

```
##       b[1]       b[2]       b[3]        sig
##  7.1691438 -0.5259763  0.7933962  0.9531215
```

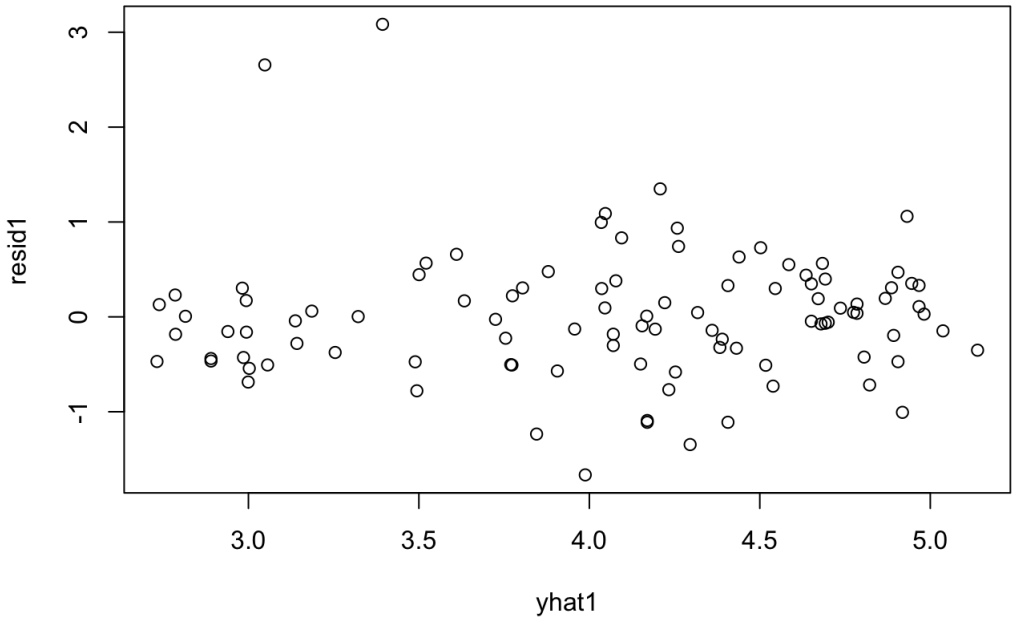
```
yhat2 = drop(X2 %*% pm_params2[1:3])
resid2 = data2_jags$y - yhat2
plot(resid2) # against data index
```




```
plot(yhat2, resid2) # against predicted values
```




```
plot(yhat1, resid1) # residuals from the first model
```




```
sd(resid2) # standard deviation of residuals
```

```
## [1] 0.6489021
```

These look much better, although the residuals for Saudi Arabia and Libya are still more than three standard deviations away from the mean of the residuals. We might consider adding the other covariate `region`, but instead let's look at another option when we are faced with strong outliers.

\(t\) likelihood

Let's consider changing the likelihood. The normal likelihood has thin tails (almost all of the probability is concentrated within the first few standard deviations from the mean). This does not accommodate outliers well. Consequently, models with the normal likelihood might be overly-influenced by outliers. Recall that the \(t\) distribution is similar to the normal distribution, but it has thicker tails which can accommodate outliers.

The \(t\) linear model might look something like this. Notice that the \(t\) distribution has three parameters, including a positive "degrees of freedom" parameter. The smaller the degrees of freedom, the heavier the tails of the distribution. We might fix the degrees of freedom to some number, or we can assign it a prior distribution.

```
mod3_string = " model {
    for (i in 1:length(y)) {
        y[i] ~ dt( mu[i], tau, df )
        mu[i] = b[1] + b[2]*log_income[i] + b[3]*is_oil
    }

    for (i in 1:3) {
        b[i] ~ dnorm(0.0, 1.0/1.0e6)
    }

    df = nu + 2.0 # we want degrees of freedom > 2 to guarantee existence of mean and variance
    nu ~ dexp(1.0)

    tau ~ dgamma(5/2.0, 5*10.0/2.0) # tau is close to, but not equal to the precision
    sig = sqrt( 1.0 / tau * df / (df - 2.0) ) # standard deviation of errors
} "
```

We will leave it up to you to fit this model.

Lesson 7.6

Compare models using Deviance Information Criterion

We have now proposed three different models. How do we compare their performance on our data? In the previous course, we discussed estimating parameters in models using the maximum likelihood method. Similarly, we can choose between competing models using the same idea.

We will use a quantity known as the deviance information criterion (DIC). It essentially calculates the posterior mean of the log-likelihood and adds a penalty for model complexity.

Let's calculate the DIC for our first two models:

the simple linear regression on log-income,

```
dic.samples(mod1, n.iter=1e3)
```

```
## Mean deviance:  231.6
## penalty 2.858
## Penalized deviance: 234.4
```

and the second model where we add oil production.

```
dic.samples(mod2, n.iter=1e3)
```

```
## Mean deviance:  225.4
## penalty 4.081
## Penalized deviance: 229.5
```

The first number is the Monte Carlo estimated posterior mean deviance, which equals \(-2\) times the log-likelihood (plus a constant that will be irrelevant for comparing models). Because of that \(-2\) factor, a smaller deviance means a higher likelihood.

Next, we are given a penalty for the complexity of our model. This penalty is necessary because we can always increase the likelihood of the model by making it more complex to fit the data exactly. We don't want to do this because over-fit models generalize poorly. This penalty is roughly equal to the effective number of parameters in your model. You can see this here. With the first model, we had a variance parameter and two betas, for a total of three parameters. In the second model, we added one more beta for the oil effect.

We add these two quantities to get the DIC (the last number). The better-fitting model has a lower DIC value. In this case, the gains we receive in deviance by adding the `is_oil` covariate outweigh the penalty for adding an extra parameter. The final DIC for the second model is lower than for the first, so we would prefer using the second model.

We encourage you to explore different model specifications and compare their fit to the data using DIC. Wikipedia provides a good introduction to DIC and we can find more details about the `JAGS` implementation through the `rjags` package documentation by entering `?dic.samples` in the `R` console.