Project 2

Vinay Srinivasan

May 29, 2018

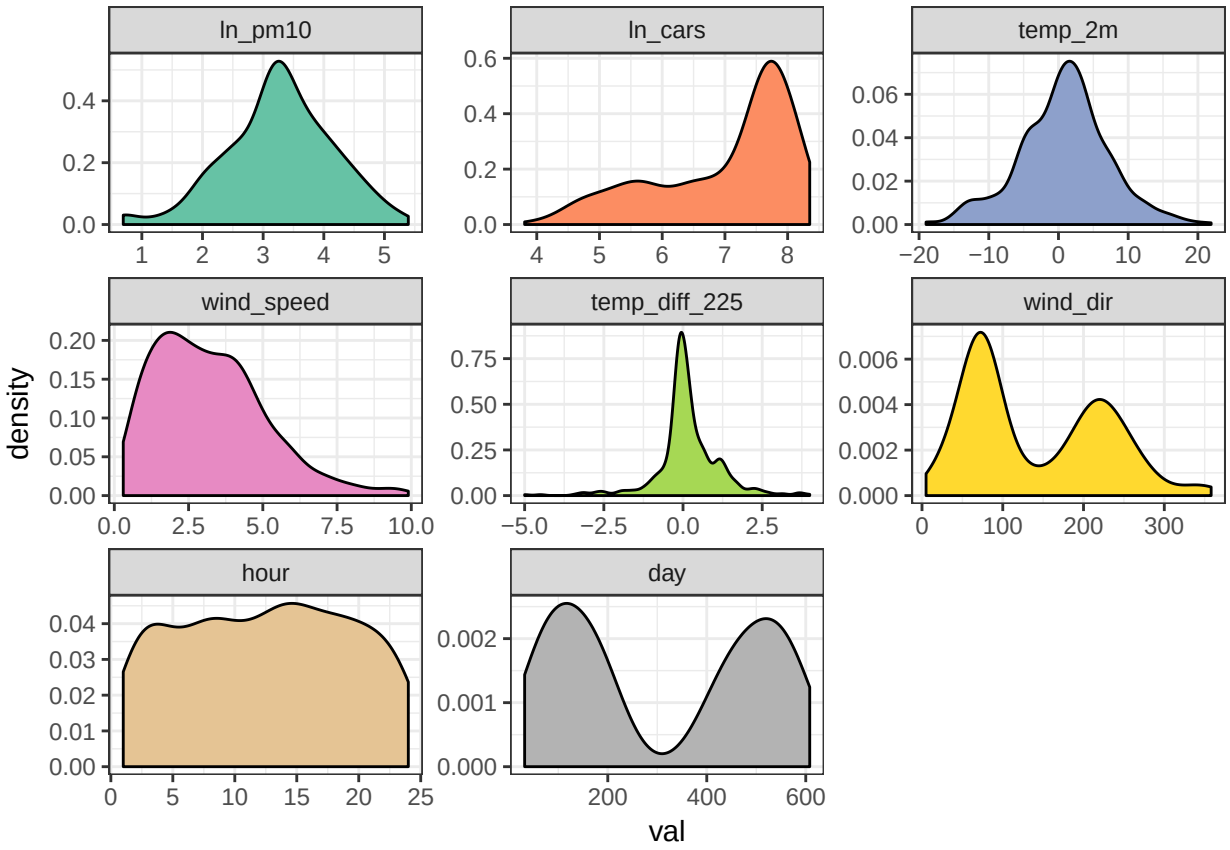
Exploring Data

```
# Loading data
datadf <- fread("../PM10.txt")
setnames(datadf, 1:8, c("ln_pm10", "ln_cars", "temp_2m", "wind_speed",
    "temp_diff_225", "wind_dir", "hour", "day"))

# Creating continuous time variable
datadf <- datadf[order(day, hour)]
datadf[, `:=`(t, 24 * day + hour)]
```

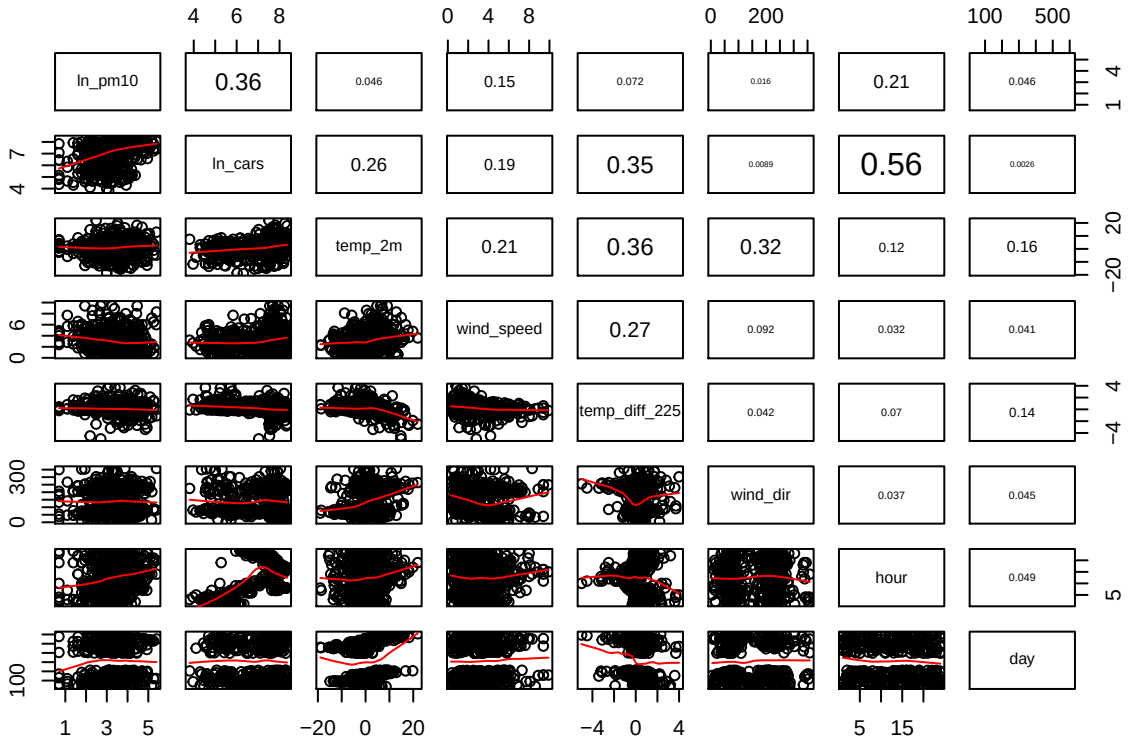
```
# Examing distributions of varibles in dataset
distdata <- melt(datadf[, !c("t")], variable.name = "var", value.name = "val",
    measure.vars = grep("^t$", names(datadf), invert = T, value = T))
```

```
## Warning in melt.data.table(datadf[, !c("t")], variable.name = "var",
## value.name = "val", : 'measure.vars' [ln_pm10, ln_cars, temp_2m,
## wind_speed, ...] are not all of the same type. By order of hierarchy, the
## molten data value column will be of type 'double'. All measure variables
## not of type 'double' will be coerced to. Check DETAILS in ?melt.data.table
## for more on coercion.
```

```
ggplot(data = distdata) + geom_density(aes(x = val, fill = var)) +
    scale_fill_brewer(palette = "Set2") + facet_wrap(~var, scales = "free") +
    guides(fill = F) + theme_bw()
```




```
# Examine bivariate relationships

panel.cor <- function(x, y, digits = 2, prefix = "", cex.cor,
    ...) {
    usr <- par("usr")
    on.exit(par(usr))
    par(usr = c(0, 1, 0, 1))
    r <- abs(cor(x, y))
    txt <- format(c(r, 0.123456789), digits = digits)[1]
    txt <- paste(prefix, txt, sep = "")
    if (missing(cex.cor))
        cex.cor <- 0.8/strwidth(txt)
    text(0.5, 0.5, txt, cex = 0.4 + cex.cor * r)
}

pairs(datadf[, !"t"], lower.panel = panel.smooth, upper.panel = panel.cor)
```




```
## Creating all combinations of formulas
form_base <- as.formula("ln_pm10~1 + f(day, model = 'iid') + f(hour, model = 'rw1')")

listcombo <- unlist(sapply(0:4, function(x) combn(4, x, simplify = FALSE)),
    recursive = FALSE)
predterms <- lapply(listcombo, function(x) paste(c(form_base,
    c("ln_cars", "temp_2m", "wind_speed", "temp_diff_225", "wind_dir")[x]),
    collapse = " + ")) %>% unlist

predterms <- c("ln_pm10~1", predterms)

# Model
resultsdf <- {
}

for (s in sample(10000, 5)) {

    message(paste0("Setting seed to ", s))

    ## Splitting data into train and test sets
    set.seed(s)
    datadf[, `:=`(samp, sample(.N, replace = F))]
    datadf <- datadf[order(samp)]
    designdf <- copy(datadf)[samp > 400, `:=`(ln_pm10, NA)]

    for (i in 1:length(predterms)) {
```

```
        message(paste0("FITTING ", predterms[i]))
        mod <- inla(formula = as.formula(predterms[i]), data = designdf,
            control.predictor = list(compute = T), control.compute = list(dic = T,
                waic = T), family = "gaussian")

        rmse <- (mod$summary.fitted.values[401:500, 1] - datadf[samp %in%
            401:500, ln_pm10])^2 %>% mean %>% sqrt

        results <- data.table(seed = s, model_id = i, model_form = predterms[i],
            waic = mod$waic$waic, dic = mod$dic$dic, oos_rmse = rmse)

        resultsdf <- rbind(resultsdf, results, use.names = T,
            fill = T)

    }

}
```

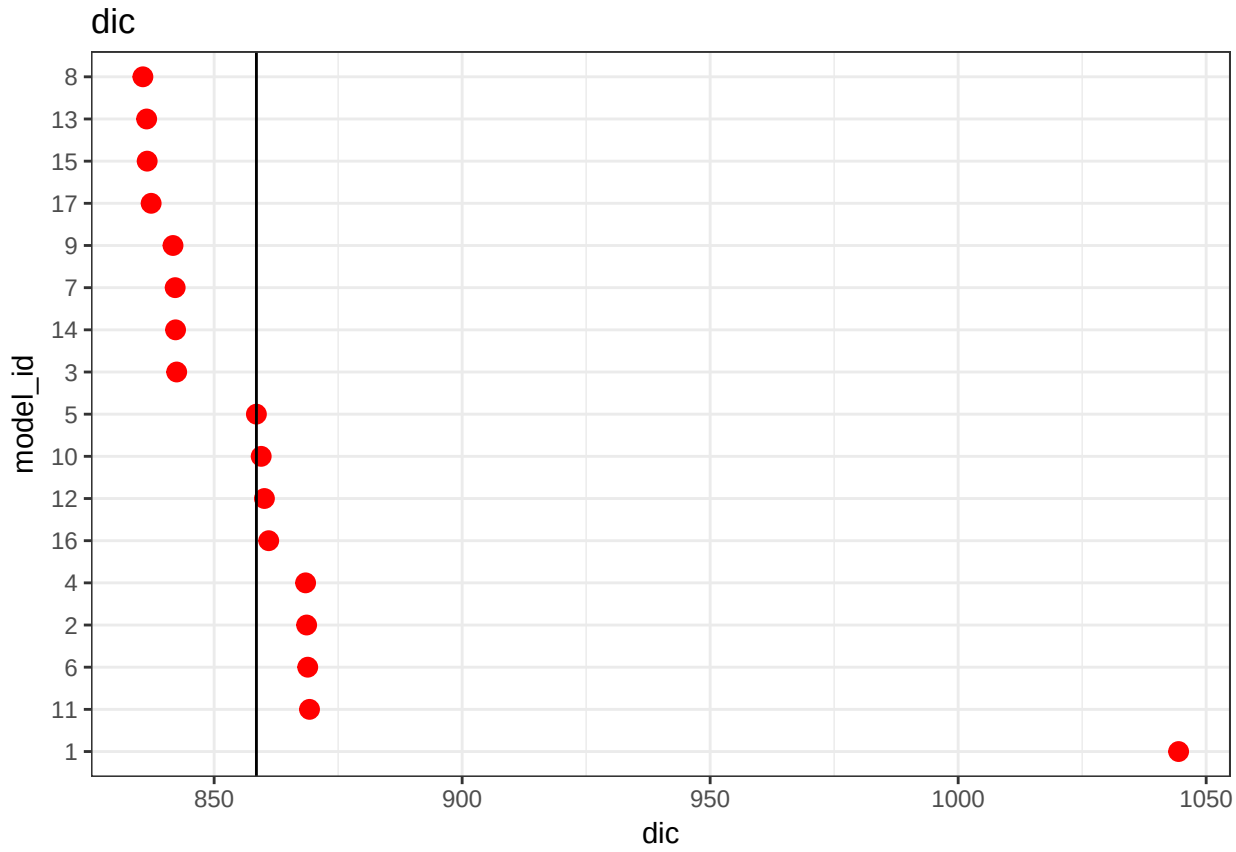
```
## Setting seed to 6032
## FITTING ln_pm10~1
## FITTING ln_pm10 ~ 1 + f(day, model = "iid") + f(hour, model = "rw1")
## FITTING ln_pm10 ~ 1 + f(day, model = "iid") + f(hour, model = "rw1") + ln_cars
## FITTING ln_pm10 ~ 1 + f(day, model = "iid") + f(hour, model = "rw1") + temp_2m
## FITTING ln_pm10 ~ 1 + f(day, model = "iid") + f(hour, model = "rw1") + wind_speed
## FITTING ln_pm10 ~ 1 + f(day, model = "iid") + f(hour, model = "rw1") + temp_diff_225
## FITTING ln_pm10 ~ 1 + f(day, model = "iid") + f(hour, model = "rw1") + ln_cars + temp_2m
## FITTING ln_pm10 ~ 1 + f(day, model = "iid") + f(hour, model = "rw1") + ln_cars + wind_speed
## FITTING ln_pm10 ~ 1 + f(day, model = "iid") + f(hour, model = "rw1") + ln_cars + temp_diff_225
## FITTING ln_pm10 ~ 1 + f(day, model = "iid") + f(hour, model = "rw1") + temp_2m + wind_speed
## FITTING ln_pm10 ~ 1 + f(day, model = "iid") + f(hour, model = "rw1") + temp_2m + temp_diff_225
## FITTING ln_pm10 ~ 1 + f(day, model = "iid") + f(hour, model = "rw1") + wind_speed + temp_diff_225
## FITTING ln_pm10 ~ 1 + f(day, model = "iid") + f(hour, model = "rw1") + ln_cars + temp_2m + wind_speed
## FITTING ln_pm10 ~ 1 + f(day, model = "iid") + f(hour, model = "rw1") + ln_cars + temp_2m + temp_diff
## FITTING ln_pm10 ~ 1 + f(day, model = "iid") + f(hour, model = "rw1") + ln_cars + wind_speed + temp_di
## FITTING ln_pm10 ~ 1 + f(day, model = "iid") + f(hour, model = "rw1") + temp_2m + wind_speed + temp_di
## FITTING ln_pm10 ~ 1 + f(day, model = "iid") + f(hour, model = "rw1") + ln_cars + temp_2m + wind_speed
## Setting seed to 7775
## FITTING ln_pm10~1
## FITTING ln_pm10 ~ 1 + f(day, model = "iid") + f(hour, model = "rw1")
## FITTING ln_pm10 ~ 1 + f(day, model = "iid") + f(hour, model = "rw1") + ln_cars
## FITTING ln_pm10 ~ 1 + f(day, model = "iid") + f(hour, model = "rw1") + temp_2m
## FITTING ln_pm10 ~ 1 + f(day, model = "iid") + f(hour, model = "rw1") + wind_speed
```

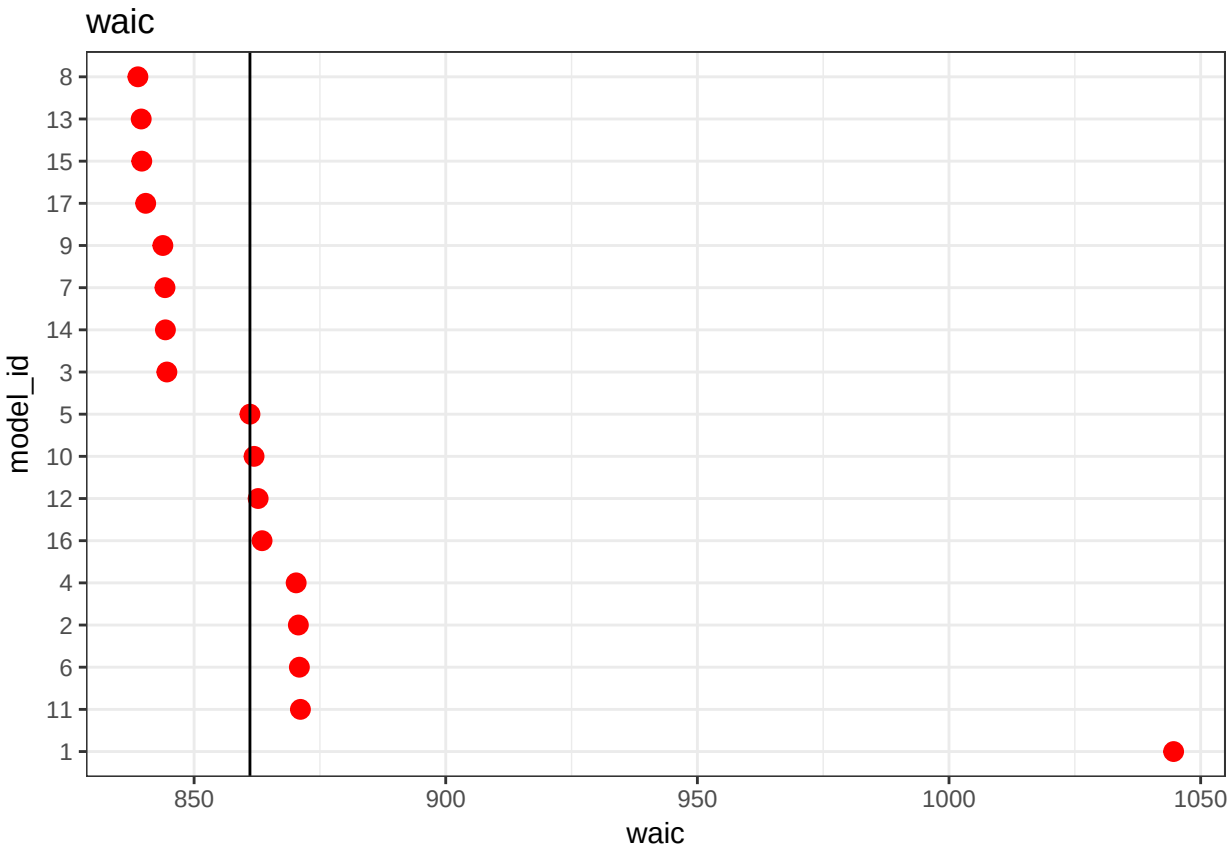
```
## FITTING ln_pm10 ~ 1 + f(day, model = "iid") + f(hour, model = "rw1") + temp_diff_225
## FITTING ln_pm10 ~ 1 + f(day, model = "iid") + f(hour, model = "rw1") + ln_cars + temp_2m
## FITTING ln_pm10 ~ 1 + f(day, model = "iid") + f(hour, model = "rw1") + ln_cars + wind_speed
## FITTING ln_pm10 ~ 1 + f(day, model = "iid") + f(hour, model = "rw1") + ln_cars + temp_diff_225
## FITTING ln_pm10 ~ 1 + f(day, model = "iid") + f(hour, model = "rw1") + temp_2m + wind_speed
## FITTING ln_pm10 ~ 1 + f(day, model = "iid") + f(hour, model = "rw1") + temp_2m + temp_diff_225
## FITTING ln_pm10 ~ 1 + f(day, model = "iid") + f(hour, model = "rw1") + wind_speed + temp_diff_225
## FITTING ln_pm10 ~ 1 + f(day, model = "iid") + f(hour, model = "rw1") + ln_cars + temp_2m + wind_speed
## FITTING ln_pm10 ~ 1 + f(day, model = "iid") + f(hour, model = "rw1") + ln_cars + temp_2m + temp_diff_
## FITTING ln_pm10 ~ 1 + f(day, model = "iid") + f(hour, model = "rw1") + ln_cars + wind_speed + temp_di
## FITTING ln_pm10 ~ 1 + f(day, model = "iid") + f(hour, model = "rw1") + temp_2m + wind_speed + temp_di
## FITTING ln_pm10 ~ 1 + f(day, model = "iid") + f(hour, model = "rw1") + ln_cars + temp_2m + wind_speed
## Setting seed to 436
## FITTING ln_pm10~1
## FITTING ln_pm10 ~ 1 + f(day, model = "iid") + f(hour, model = "rw1")
## FITTING ln_pm10 ~ 1 + f(day, model = "iid") + f(hour, model = "rw1") + ln_cars
## FITTING ln_pm10 ~ 1 + f(day, model = "iid") + f(hour, model = "rw1") + temp_2m
## FITTING ln_pm10 ~ 1 + f(day, model = "iid") + f(hour, model = "rw1") + wind_speed
## FITTING ln_pm10 ~ 1 + f(day, model = "iid") + f(hour, model = "rw1") + temp_diff_225
## FITTING ln_pm10 ~ 1 + f(day, model = "iid") + f(hour, model = "rw1") + ln_cars + temp_2m
## FITTING ln_pm10 ~ 1 + f(day, model = "iid") + f(hour, model = "rw1") + ln_cars + wind_speed
## FITTING ln_pm10 ~ 1 + f(day, model = "iid") + f(hour, model = "rw1") + ln_cars + temp_diff_225
## FITTING ln_pm10 ~ 1 + f(day, model = "iid") + f(hour, model = "rw1") + temp_2m + wind_speed
## FITTING ln_pm10 ~ 1 + f(day, model = "iid") + f(hour, model = "rw1") + temp_2m + temp_diff_225
## FITTING ln_pm10 ~ 1 + f(day, model = "iid") + f(hour, model = "rw1") + wind_speed + temp_diff_225
## FITTING ln_pm10 ~ 1 + f(day, model = "iid") + f(hour, model = "rw1") + ln_cars + temp_2m + wind_speed
## FITTING ln_pm10 ~ 1 + f(day, model = "iid") + f(hour, model = "rw1") + ln_cars + temp_2m + temp_diff_
## FITTING ln_pm10 ~ 1 + f(day, model = "iid") + f(hour, model = "rw1") + ln_cars + wind_speed + temp_di
## FITTING ln_pm10 ~ 1 + f(day, model = "iid") + f(hour, model = "rw1") + temp_2m + wind_speed + temp_di
## FITTING ln_pm10 ~ 1 + f(day, model = "iid") + f(hour, model = "rw1") + ln_cars + temp_2m + wind_speed
## Setting seed to 5950
## FITTING ln_pm10~1
## FITTING ln_pm10 ~ 1 + f(day, model = "iid") + f(hour, model = "rw1")
## FITTING ln_pm10 ~ 1 + f(day, model = "iid") + f(hour, model = "rw1") + ln_cars
## FITTING ln_pm10 ~ 1 + f(day, model = "iid") + f(hour, model = "rw1") + temp_2m
## FITTING ln_pm10 ~ 1 + f(day, model = "iid") + f(hour, model = "rw1") + wind_speed
```

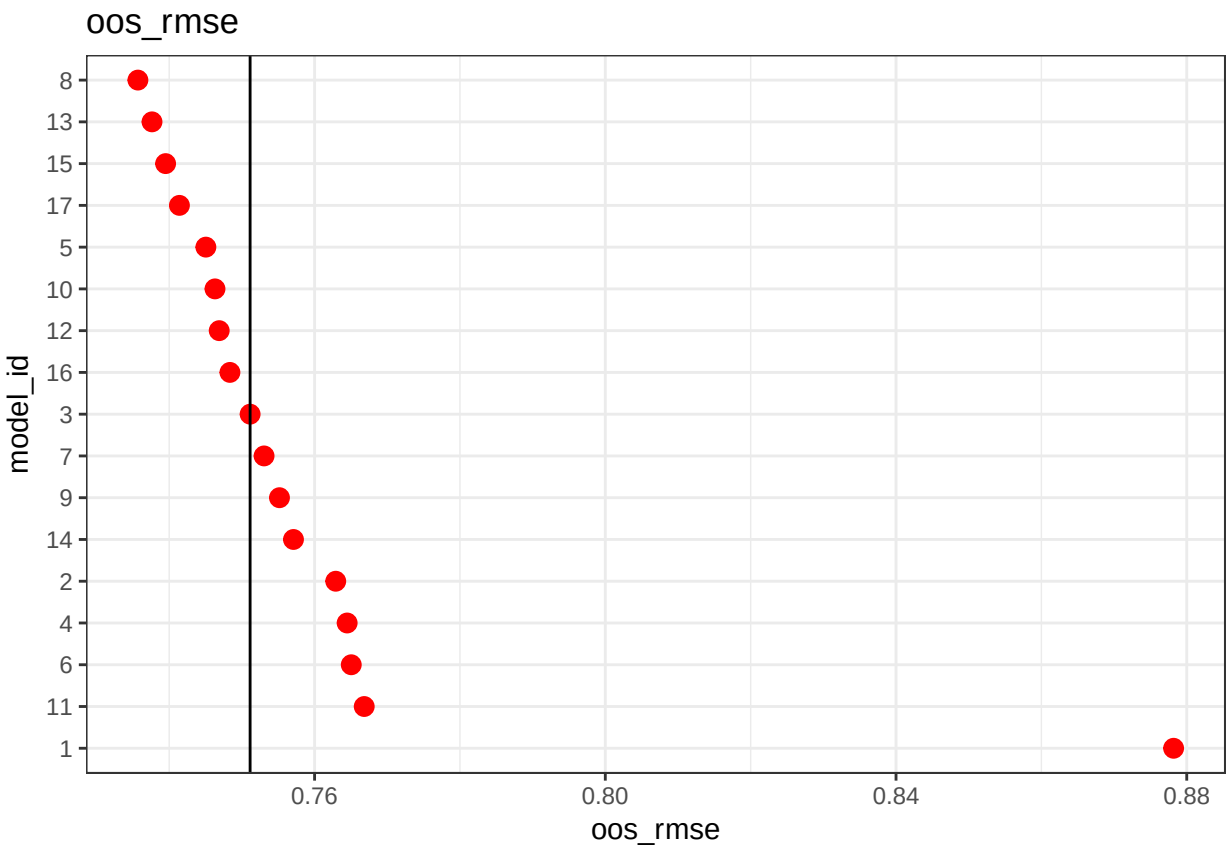
```
## FITTING ln_pm10 ~ 1 + f(day, model = "iid") + f(hour, model = "rw1") + temp_diff_225
## FITTING ln_pm10 ~ 1 + f(day, model = "iid") + f(hour, model = "rw1") + ln_cars + temp_2m
## FITTING ln_pm10 ~ 1 + f(day, model = "iid") + f(hour, model = "rw1") + ln_cars + wind_speed
## FITTING ln_pm10 ~ 1 + f(day, model = "iid") + f(hour, model = "rw1") + ln_cars + temp_diff_225
## FITTING ln_pm10 ~ 1 + f(day, model = "iid") + f(hour, model = "rw1") + temp_2m + wind_speed
## FITTING ln_pm10 ~ 1 + f(day, model = "iid") + f(hour, model = "rw1") + temp_2m + temp_diff_225
## FITTING ln_pm10 ~ 1 + f(day, model = "iid") + f(hour, model = "rw1") + wind_speed + temp_diff_225
## FITTING ln_pm10 ~ 1 + f(day, model = "iid") + f(hour, model = "rw1") + ln_cars + temp_2m + wind_speed
## FITTING ln_pm10 ~ 1 + f(day, model = "iid") + f(hour, model = "rw1") + ln_cars + temp_2m + temp_diff_
## FITTING ln_pm10 ~ 1 + f(day, model = "iid") + f(hour, model = "rw1") + ln_cars + wind_speed + temp_di
## FITTING ln_pm10 ~ 1 + f(day, model = "iid") + f(hour, model = "rw1") + temp_2m + wind_speed + temp_di
## FITTING ln_pm10 ~ 1 + f(day, model = "iid") + f(hour, model = "rw1") + ln_cars + temp_2m + wind_speed
## Setting seed to 5809
## FITTING ln_pm10~1
## FITTING ln_pm10 ~ 1 + f(day, model = "iid") + f(hour, model = "rw1")
## FITTING ln_pm10 ~ 1 + f(day, model = "iid") + f(hour, model = "rw1") + ln_cars
## FITTING ln_pm10 ~ 1 + f(day, model = "iid") + f(hour, model = "rw1") + temp_2m
## FITTING ln_pm10 ~ 1 + f(day, model = "iid") + f(hour, model = "rw1") + wind_speed
## FITTING ln_pm10 ~ 1 + f(day, model = "iid") + f(hour, model = "rw1") + temp_diff_225
## FITTING ln_pm10 ~ 1 + f(day, model = "iid") + f(hour, model = "rw1") + ln_cars + temp_2m
## FITTING ln_pm10 ~ 1 + f(day, model = "iid") + f(hour, model = "rw1") + ln_cars + wind_speed
## FITTING ln_pm10 ~ 1 + f(day, model = "iid") + f(hour, model = "rw1") + ln_cars + temp_diff_225
## FITTING ln_pm10 ~ 1 + f(day, model = "iid") + f(hour, model = "rw1") + temp_2m + wind_speed
## FITTING ln_pm10 ~ 1 + f(day, model = "iid") + f(hour, model = "rw1") + temp_2m + temp_diff_225
## FITTING ln_pm10 ~ 1 + f(day, model = "iid") + f(hour, model = "rw1") + wind_speed + temp_diff_225
## FITTING ln_pm10 ~ 1 + f(day, model = "iid") + f(hour, model = "rw1") + ln_cars + temp_2m + wind_speed
## FITTING ln_pm10 ~ 1 + f(day, model = "iid") + f(hour, model = "rw1") + ln_cars + temp_2m + temp_diff_
## FITTING ln_pm10 ~ 1 + f(day, model = "iid") + f(hour, model = "rw1") + ln_cars + wind_speed + temp_di
## FITTING ln_pm10 ~ 1 + f(day, model = "iid") + f(hour, model = "rw1") + temp_2m + wind_speed + temp_di
## FITTING ln_pm10 ~ 1 + f(day, model = "iid") + f(hour, model = "rw1") + ln_cars + temp_2m + wind_speed
```

```r
# Summarize
summdf <- resultsdf[, lapply(.SD, mean), by = .(model_id, model_form),
    .SDcols = c("dic", "waic", "oos_rmse")]

# Evaluate

for (plotvar in c("dic", "waic", "oos_rmse")) {

    p <- ggplot(data = summdf) + geom_point(aes(x = get(plotvar),
```

```
        y = factor(model_id, levels = summdf[order(-get(plotvar))]$model_id)),
        color = "red", size = 3) + geom_vline(xintercept = median(summdf[,
        get(plotvar)])) + theme_bw() + labs(title = plotvar,
        x = plotvar, y = "model_id")

    print(p)

}
```




waic
model_id
8
13
15
17
9
7
14
3
5
10
12
16
4
2
6
11
1
850
900
950
1000
1050
waic



The best model based on in-sample fit, out-of-sample performance, and parsimony seems to be ln_pm10 ~ 1 + f(day, model = "iid") + f(hour, model = "rw1") + ln_cars + wind_speed, ln_pm10 ~ 1 + f(day, model = "iid") + f(hour, model = "rw1") + ln_cars + wind_speed, ln_pm10 ~ 1 + f(day, model = "iid") + f(hour, model = "rw1") + ln_cars + wind_speed, ln_pm10 ~ 1 + f(day, model = "iid") + f(hour, model = "rw1") + ln_cars + wind_speed, ln_pm10 ~ 1 + f(day, model = "iid") + f(hour, model = "rw1") + ln_cars + wind_speed.