Inclass_week10

Will Doyle

3/26/2019

```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------- tidyverse 1.2.1 --
```

```
## v ggplot2 3.1.0     v purrr   0.2.5
## v tibble  2.0.1     v dplyr   0.7.8
## v tidyr   0.8.2     v stringr 1.3.1
## v readr   1.3.1     v forcats 0.3.0
```

```
## -- Conflicts ------------------------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(modelr)
library(caret)
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'caret'
```

```
## The following object is masked from 'package:purrr':
##
##     lift
```

```
load("els.Rdata")
```

```
## Define the model
mod1_formula<-formula(bynels2m~byses1+ ##DEFINE YOUR OWN MODEL HERE
                        bynels2r)
## Run the model against all of the data
basic.mod<-lm(mod1_formula,
              data=els); summary(basic.mod)
```

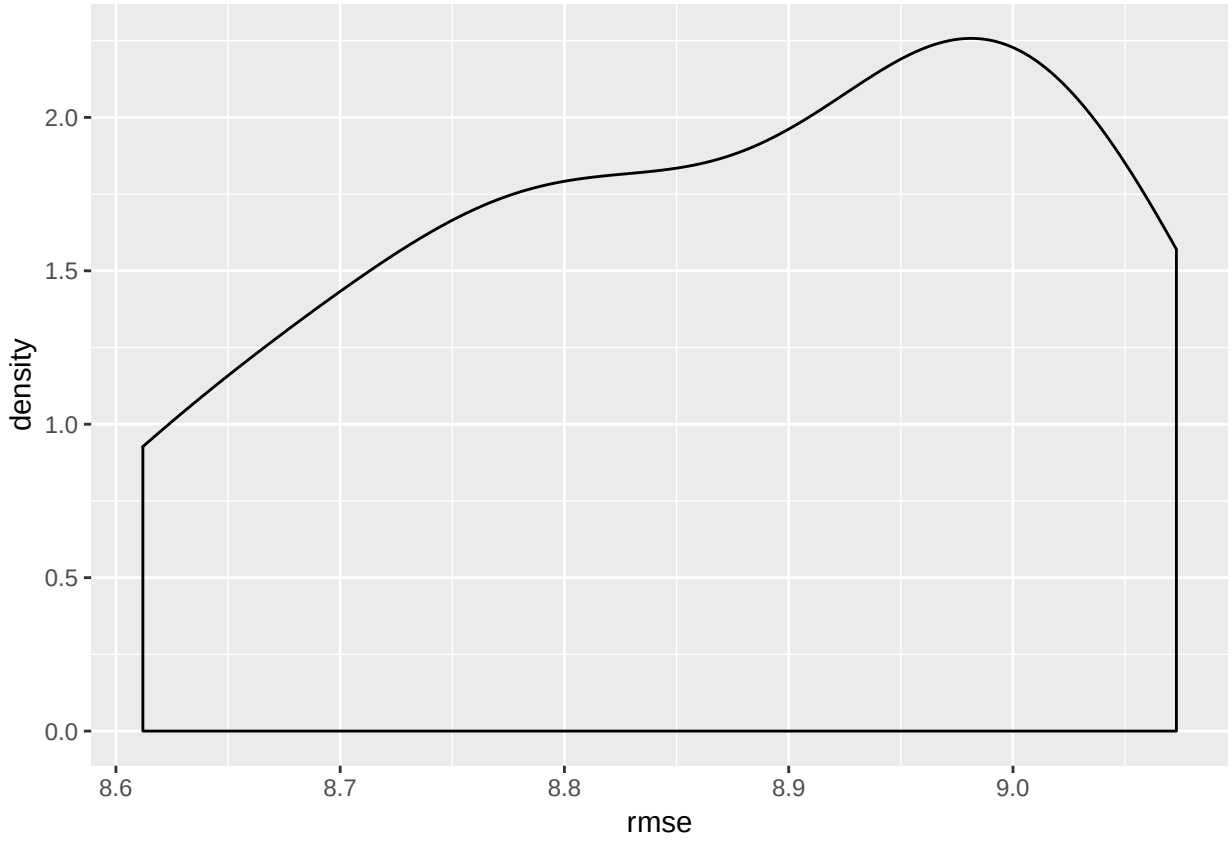
```
##
## Call:
## lm(formula = mod1_formula, data = els)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.212  -5.992  -0.165   5.853  46.062
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 15.844996   0.256915   61.67   <2e-16 ***
## byses1       2.493117   0.106930   23.32   <2e-16 ***
## bynels2r     0.991916   0.008382  118.34   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.871 on 15322 degrees of freedom
```

```
##   (964 observations deleted due to missingness)
## Multiple R-squared:  0.5759, Adjusted R-squared:  0.5758
## F-statistic: 1.04e+04 on 2 and 15322 DF,  p-value: < 2.2e-16
```

```
els_cf<-els%>% ##SUBSTITUTE YOUR OWN DATA HERE
  crossv_kfold(10)
```

```
rmse_mod1<-els_cf %>%
  mutate(train = map(train, as_tibble)) %>% ## Convert to tibbles
  mutate(model = map(train, ~ lm(mod1_formula,
                                 data = .))) %>%
  mutate(rmse = map2_dbl(model, test, rmse)) %>% ## apply model, get rmse
  select(.id, rmse) ## pull just id and rmse
```

```
gg<-ggplot(data=rmse_mod1,aes(x=rmse))
gg<-gg+geom_density()
gg
```




```
els_cv<-els%>% ##SUBSTITUTE YOUR OWN DATA HERE
  crossv_mc(n=1000,test=.2)
els_cv
```

```
## # A tibble: 1,000 x 3
##    train          test           .id
##    <list>         <list>         <chr>
##  1 <S3: resample> <S3: resample> 0001
##  2 <S3: resample> <S3: resample> 0002
##  3 <S3: resample> <S3: resample> 0003
##  4 <S3: resample> <S3: resample> 0004
```

```
##  5 <S3: resample> <S3: resample> 0005
##  6 <S3: resample> <S3: resample> 0006
##  7 <S3: resample> <S3: resample> 0007
##  8 <S3: resample> <S3: resample> 0008
##  9 <S3: resample> <S3: resample> 0009
## 10 <S3: resample> <S3: resample> 0010
## # ... with 990 more rows
```

```
mod1_rmse_cv<-els_cv %>%
  mutate(train = map(train, as_tibble)) %>% ## Convert to tibbles
  mutate(model = map(train, ~ lm(mod1_formula, data = .)))%>%
  mutate(rmse = map2_dbl(model, test, rmse))%>%
  select(.id, rmse) ## pull just id and rmse
```

```
summary(mod1_rmse_cv$rmse)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   8.496   8.797   8.880   8.877   8.953   9.311
```

```
gg<-ggplot(mod1_rmse_cv,aes(x=rmse))
gg<-gg+geom_density()
gg
```