QTM 220 HW #6

Author

Veronica Vargas

QTM 220 HW #6

Exercise #1 - Cross Validation

```
# load packages
library(tidyverse)
```

```
Warning: package 'tidyverse' was built under R version 4.3.3
```

```
── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
✓ dplyr     1.1.3     ✓ readr     2.1.4
✓ forcats   1.0.0     ✓ stringr   1.5.0
✓ ggplot2   3.4.3     ✓ tibble    3.2.1
✓ lubridate 1.9.2     ✓ tidyr     1.3.0
✓ purrr     1.0.2
── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()
ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(mosaicData)
```

```
Warning: package 'mosaicData' was built under R version 4.3.3
```

```
library(leaps)
```

```
Warning: package 'leaps' was built under R version 4.3.3
```

```
library(caret)
```

```
Warning: package 'caret' was built under R version 4.3.3
```

```
Loading required package: lattice
```

```
Warning: package 'lattice' was built under R version 4.3.3
```

```

Attaching package: 'caret'
```

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

    lift
```

```
library(ISLR2)
```

```
Warning: package 'ISLR2' was built under R version 4.3.3
```

```
library(ggplot2)
```

```
data("HELPrct")
head(HELPrct)
```

```
  age anysubstatus anysub cesd d1 daysanysub dayslink drugrisk e2b female
1  37            1    yes   49  3        177      225        0  NA      0
2  37            1    yes   30 22          2       NA        0  NA      0
3  26            1    yes   39  0          3      365       20  NA      0
4  39            1    yes   15  2        189      343        0   1      1
5  32            1    yes   39 12          2       57        0   1      0
6  47            1    yes    6  1         31      365        0  NA      1
     sex g1b homeless i1 i2 id indtot linkstatus link       mcs      pcs pss_fr
1   male yes   housed 13 26  1     39          1  yes 25.111990 58.41369      0
2   male yes homeless 56 62  2     43         NA <NA> 26.670307 36.03694      1
3   male  no   housed  0  0  3     41          0   no  6.762923 74.80633     13
4 female  no   housed  5  5  4     28          0   no 43.967880 61.93168     11
5   male  no homeless 10 13  5     38          1  yes 21.675755 37.34558     10
6 female  no   housed  4  4  6     29          0   no 55.508991 46.47521      5
  racegrp satreat sexrisk substance treat avg_drinks max_drinks
1   black      no       4   cocaine   yes         13         26
2   white      no       7   alcohol   yes         56         62
3   black      no       2    heroin    no          0          0
4   white     yes       4    heroin    no          5          5
5   black      no       6   cocaine    no         10         13
6   black      no       5   cocaine   yes          4          4
  hospitalizations
1                3
2               22
3                0
4                2
5               12
6                1
```

```
names(HELPrct)
```

```
 [1] "age"              "anysubstatus"     "anysub"           "cesd"
 [5] "d1"               "daysanysub"       "dayslink"         "drugrisk"
 [9] "e2b"              "female"           "sex"              "g1b"
[13] "homeless"         "i1"               "i2"               "id"
[17] "indtot"           "linkstatus"       "link"             "mcs"
[21] "pcs"              "pss_fr"           "racegrp"          "satreat"
[25] "sexrisk"          "substance"        "treat"            "avg_drinks"
[29] "max_drinks"       "hospitalizations"
```

(a)

```
ggplot(HELPrct, aes(x = age, y = cesd)) +
  geom_point(aes(x = age, y = cesd,
                             color = factor(sex)),
             alpha = 0.3) +
  labs(
    title = "Age vs. CESD by Sex",
    x = "Age",
    y = "CESD",
    color = "Sex") +
  theme_minimal()
```

(b) Parallel Lines Model

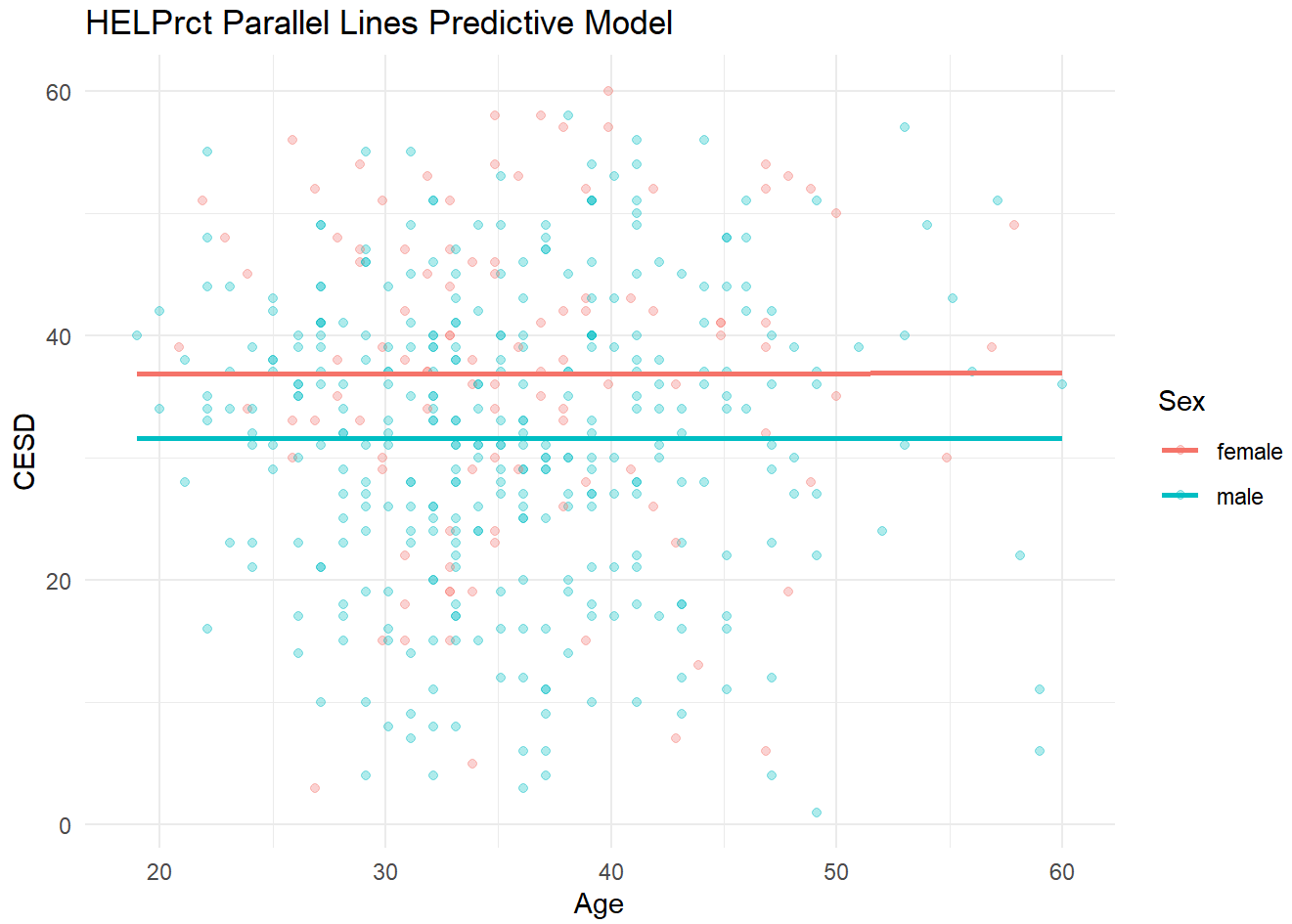
```
mod.coarsen1 <- lm(cesd ~ sex + age, data = HELPrct)
mod.coarsen1
```

```
Call:
lm(formula = cesd ~ sex + age, data = HELPrct)

Coefficients:
(Intercept)      sexmale          age
 36.8525597   -5.2888212    0.0009735
```

```
age_seq <- seq(min(HELPrct$age), max(HELPrct$age), by = 1)

pred_data <- expand.grid(
  age = age_seq,
  sex = c(0, 1))

pred_data$sex <- factor(pred_data$sex,
                        levels = c(0, 1),
                        labels = c("male", "female"))

pred_data$predicted_cesd <- predict(mod.coarsen1, newdata = pred_data)

ggplot() +
  geom_point(data = HELPrct, aes(x = age, y = cesd, color = sex),
             alpha = 0.3, position=position_dodge(width=0.5)) +
  geom_line(data = pred_data, aes(x = age, y = predicted_cesd,
```

```
                                color = sex), size = 1) +
  labs(
    title = "HELPrct Parallel Lines Predictive Model",
    x = "Age",
    y = "CESD",
    color = "Sex"
  ) +
  theme_minimal()
```

```
Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
ℹ Please use `linewidth` instead.
```

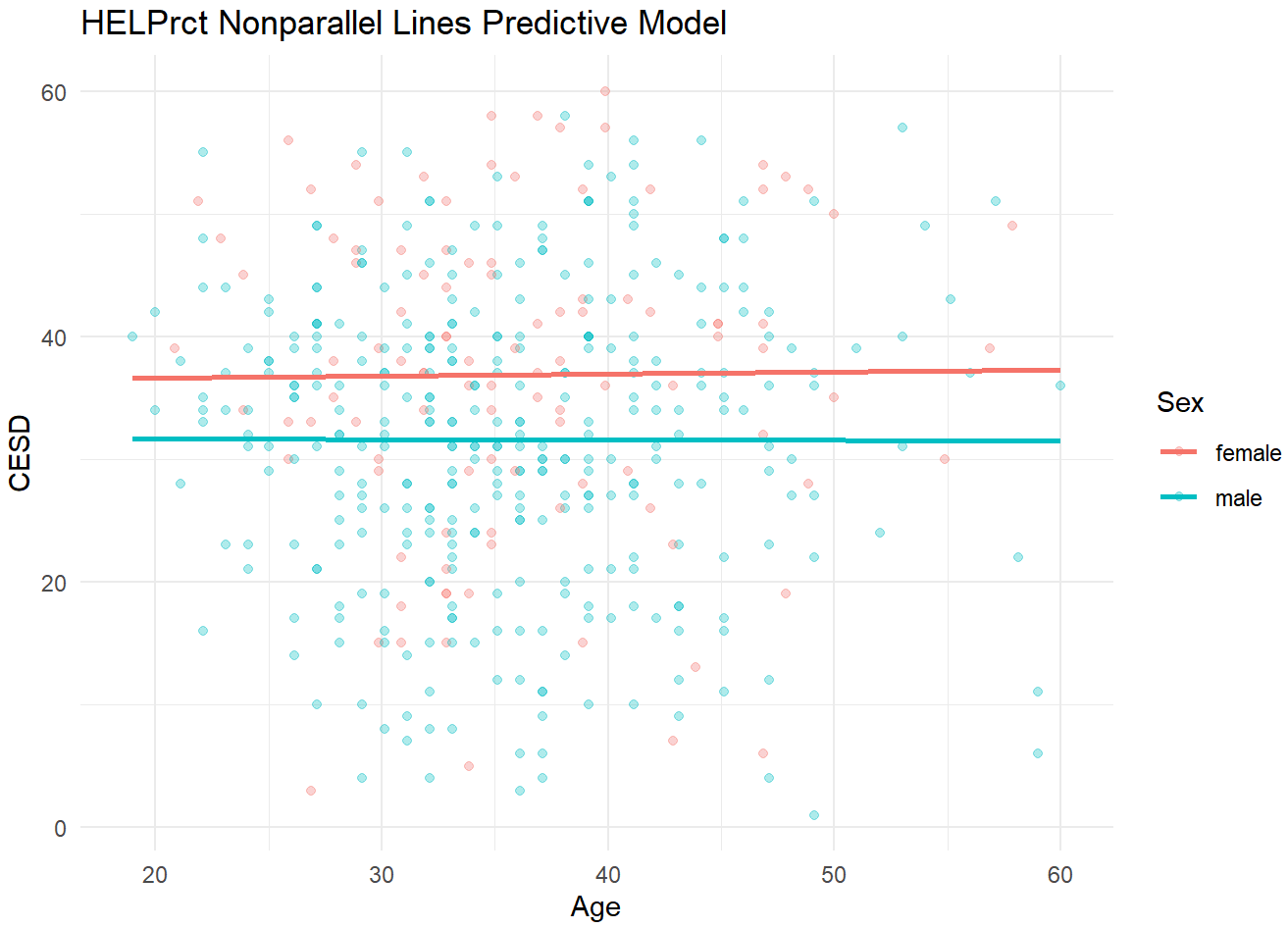


(d) Nonparallel Lines Model

```
mod.coarsen2 <- lm(cesd ~ age + sex + age*sex, data = HELPrct)
mod.coarsen2
```

```

Call:
lm(formula = cesd ~ age + sex + age * sex, data = HELPrct)

Coefficients:
(Intercept)          age       sexmale  age:sexmale
   36.28726      0.01657     -4.56076     -0.02018

age_seq <- seq(min(HELPrct$age), max(HELPrct$age), by = 1)

pred_data <- expand.grid(
  age = age_seq,
```

```
  sex = c(0, 1))

pred_data$sex <- factor(pred_data$sex,
                            levels = c(0, 1),
                            labels = c("male", "female"))

pred_data$predicted_cesd <- predict(mod.coarsen2, newdata = pred_data)

ggplot() +
  geom_point(data = HELPrct, aes(x = age, y = cesd, color = sex),
             alpha = 0.3, position=position_dodge(width=0.5)) +
  geom_line(data = pred_data, aes(x = age, y = predicted_cesd,
                                  color = sex), size = 1) +
  labs(
    title = "HELPrct Nonparallel Lines Predictive Model",
    x = "Age",
    y = "CESD",
    color = "Sex"
  ) +
  theme_minimal()
```



(f) Validation Set Approach

```
set.seed(42)
n <- nrow(HELPrct)
train_indices <- sample(1:n, size = 0.70 * n)
train_data <- HELPrct[train_indices, ]
test_data <- HELPrct[-train_indices, ]
```

```
# Model A: Parallel Lines Model
model_A_split <- lm(cesd ~ sex + age, data = train_data)
predictions_A_split <- predict(model_A_split, test_data)
rss_A_split <- sum((test_data$cesd - predictions_A_split)^2)
rss_A_split
```

```
[1] 21962.02
```

```
# Model B: Nonparallel Lines Model
model_B_split <- lm(cesd ~ age + sex + age*sex, data = train_data)
predictions_B_split <- predict(model_B_split, test_data)
rss_B_split <- sum((test_data$cesd - predictions_B_split)^2)
rss_B_split
```

```
[1] 21975.74
```

(g)

Repetition #1

```
set.seed(300)
train_indices <- sample(1:n, size = 0.75 * n)
train_data <- HELPrct[train_indices, ]
test_data <- HELPrct[-train_indices, ]
```

```
# Model A: Parallel Lines Model
model_A_split <- lm(cesd ~ sex + age, data = train_data)
predictions_A_split <- predict(model_A_split, test_data)
rss_A_split <- sum((test_data$cesd - predictions_A_split)^2)
rss_A_split
```

```
[1] 18660.69
```

```
# Model B: Nonparallel Lines Model
model_B_split <- lm(cesd ~ age + sex + age*sex, data = train_data)
predictions_B_split <- predict(model_B_split, test_data)
rss_B_split <- sum((test_data$cesd - predictions_B_split)^2)
rss_B_split
```

```
[1] 18704.95
```

Repetition #2

```
set.seed(45)
train_indices <- sample(1:n, size = 0.90 * n)
train_data <- HELPrct[train_indices, ]
test_data <- HELPrct[-train_indices, ]
```

```
# Model A: Parallel Lines Model
model_A_split <- lm(cesd ~ sex + age, data = train_data)
predictions_A_split <- predict(model_A_split, test_data)
rss_A_split <- sum((test_data$cesd - predictions_A_split)^2)
rss_A_split
```

```
[1] 6969.56
```

```
# Model B: Nonparallel Lines Model
model_B_split <- lm(cesd ~ age + sex + age*sex, data = train_data)
predictions_B_split <- predict(model_B_split, test_data)
```

```
rss_B_split <- sum((test_data$cesd - predictions_B_split)^2)
rss_B_split
```

```
[1] 7007.416
```

Repetition #3

```
set.seed(12345)
train_indices <- sample(1:n, size = 0.95 * n)
train_data <- HELPrct[train_indices, ]
test_data <- HELPrct[-train_indices, ]

# Model A: Parallel Lines Model
model_A_split <- lm(cesd ~ sex + age, data = train_data)
predictions_A_split <- predict(model_A_split, test_data)
rss_A_split <- sum((test_data$cesd - predictions_A_split)^2)
rss_A_split
```

```
[1] 3715.129
```

```
# Model B: Nonparallel Lines Model
model_B_split <- lm(cesd ~ age + sex + age*sex, data = train_data)
predictions_B_split <- predict(model_B_split, test_data)
rss_B_split <- sum((test_data$cesd - predictions_B_split)^2)
rss_B_split
```

```
[1] 3716.06
```

(i) LOOCV Approach

```
rss_summary <- function(data, lev = NULL, model = NULL) {
  residuals <- data$obs - data$pred
  rss <- sum(residuals^2)
  rmse <- sqrt(mean(residuals^2))
  return(c(RMSE = rmse, RSS = rss))
}

train_control_loocv <- trainControl(
  method = "LOOCV",
  summaryFunction = rss_summary,
  savePredictions = "all",
  classProbs = FALSE,
  allowParallel = FALSE
)

# Train Model A: Parallel Lines Model
set.seed(100)
model_A_caret_loocv <- train(
  cesd ~ sex + age,
  data = HELPrct,
  method = "lm",
  trControl = train_control_loocv,
  metric = "RMSE"
)

# Train Model B: Nonparallel Lines Model
set.seed(100)
model_B_caret_loocv <- train(
  cesd ~ age + sex + age*sex,
  data = HELPrct,
```

```
  method = "lm",
  trControl = train_control_loocv,
  metric = "RMSE"
)

model_A_caret_loocv$results
```

```
  intercept     RMSE      RSS
1      TRUE 12.38263 69458.28
```

```
model_B_caret_loocv$results
```

```
  intercept     RMSE      RSS
1      TRUE 12.41142 69781.69
```

(l) 10-Fold Validation Approach

```
train_control_kfold <- trainControl(
  method = "cv",
  number = 10,
  summaryFunction = rss_summary,
  savePredictions = "final",
  classProbs = FALSE,
  allowParallel = FALSE)

# Train Model A: Parallel Lines Model
set.seed(123)
model_A_caret <- train(
  cesd ~ sex + age,
  data = HELPrct,
  method = "lm",
  trControl = train_control_kfold,
  metric = "RMSE")

# Train Model B: Nonparallel Lines Model
set.seed(123)
model_B_caret <- train(
  cesd ~ age + sex + age*sex,
  data = HELPrct,
  method = "lm",
  trControl = train_control_kfold,
  metric = "RMSE")

model_A_caret$results
```

```
  intercept     RMSE      RSS    RMSESD   RSSSD
1      TRUE 12.33696 6915.772 0.7947347 834.725
```

```
model_B_caret$results
```

```
  intercept     RMSE      RSS    RMSESD   RSSSD
1      TRUE 12.36503 6947.187 0.7984796 837.477
```

(m) Validation Set Approach Bootstrapped

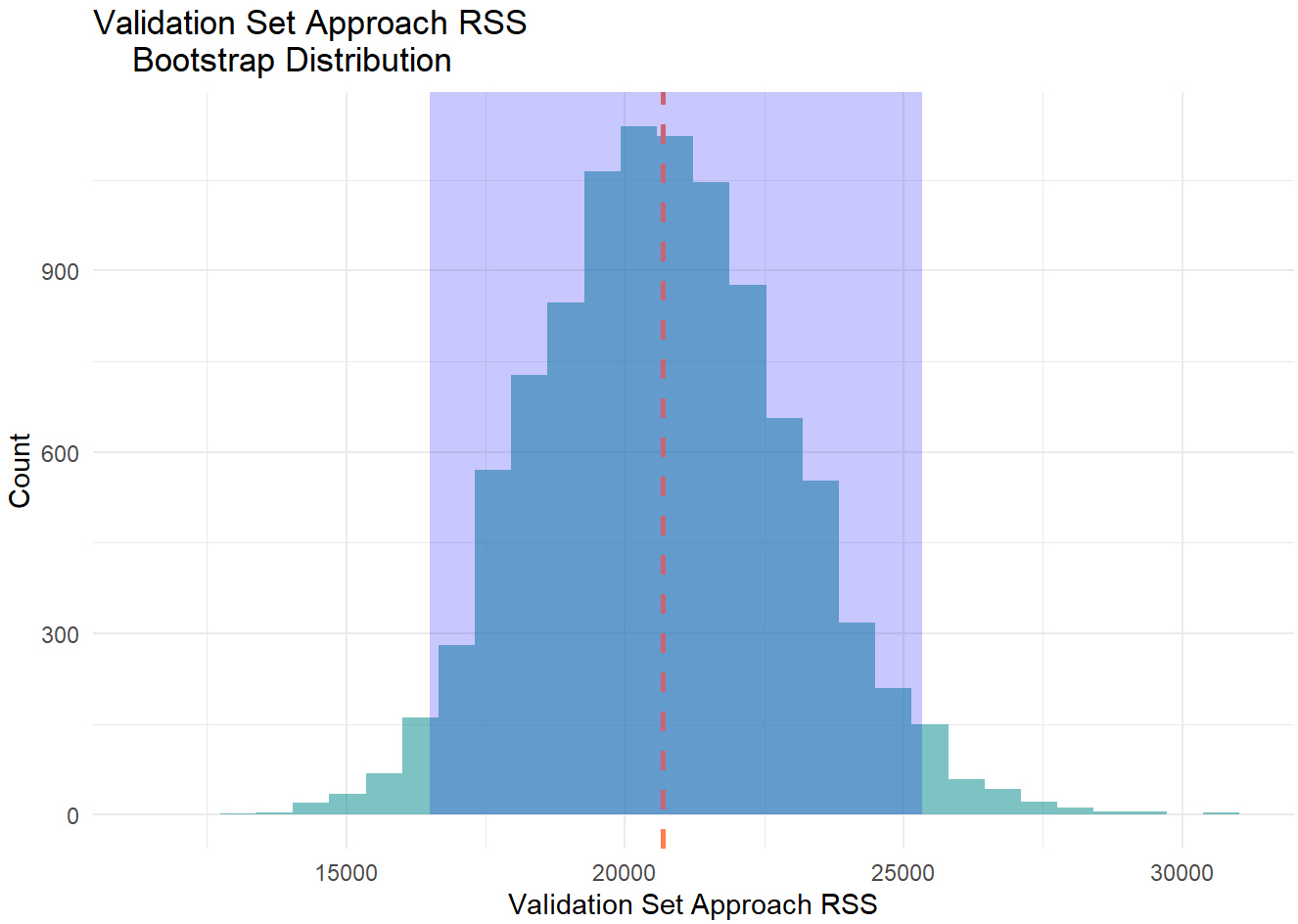
```
rss_A_split_boot <- function(data, B = 10000, train_size = 0.7, set_seed = 123) {

  set.seed(set_seed)
  rss_A_split_boot <- numeric(B)
  n <- nrow(data)
```

```
  for (b in 1:B) {
    boot_sample <- data[sample(1:nrow(data), size = n, replace = TRUE), ]

    train_indices <- sample(1:nrow(boot_sample), size = train_size * n)
    train_data <- boot_sample[train_indices, ]
    test_data <- boot_sample[-train_indices, ]

    model_A_split <- lm(cesd ~ sex + age, data = train_data)
    predictions_A_split <- predict(model_A_split, test_data)
    rss_A_split_boot[b] <- sum((test_data$cesd - predictions_A_split)^2)
  }
  return(rss_A_split_boot)
}

modelA_boot_rss <- rss_A_split_boot(HELPrct, B = 10000)
```

(o)

```
ggplot(data = data.frame(rss_A_split = modelA_boot_rss), aes(x = rss_A_split)) +
  geom_histogram(fill = "cyan4", alpha = 0.5, position = "identity") +
  geom_vline(xintercept = mean(modelA_boot_rss),
             linetype="dashed",
             color = "coral", linewidth=1) +
  annotate("rect", xmin = quantile(modelA_boot_rss, 0.025),
                   xmax = quantile(modelA_boot_rss, 0.975),
                   ymin = 0, ymax = Inf, fill = "blue",
                   alpha = 0.2) +
  labs(
    title = "Validation Set Approach RSS
    Bootstrap Distribution",
    x = "Validation Set Approach RSS",
    y = "Count") +
  theme_minimal()
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```


Validation Set Approach RSS
Bootstrap Distribution
Count
900
600
300
0
15000
20000
25000
30000
Validation Set Approach RSS