Veronica Vargas

Homework 5

Due date: 11/01/2024

October 24, 2024

1 Introduction

This exercise sheet contains a series of problems designed to test and enhance your understanding of the topics covered in the course. Please ensure that you attempt all problems and provide detailed solutions where necessary. If you have any questions or need clarification, feel free to reach out your TA.

2 Exercices

Exercise 1: Minimizing Sum of Squared Residuals [15 points] Solve for the *a* and *b* that minimize the sum of squared residuals.

$$\underset{a,b}{\operatorname{argmin}} \sum_{i=1}^{n} \{Y_i - (bX_i + a)\}^2$$

You may want to use the following facts:

- $\sum_{i=1}^{n} Y_i = n\bar{Y}$
- $\sum_{i=1}^{n} X_i = n\bar{X}$
- $(\sum_{i=1}^{n} X_i Y_i) - n\bar{X}\bar{Y} = \sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$
- $(\sum_{i=1}^{n} X_i^2) - n\bar{X}^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2$

QTM 220 HW #5

10/30/2024

Exercise #1

$$\underset{a\,b}{\operatorname{argmin}} \sum_{i=1}^{n} \{Y_i - (bX_i + a)\}^2$$

① $$\partial/\partial a = \sum_{i=1}^{n} 2(Y_i - bX_i - a)(-1) = 0 \rightarrow$$ ② $$\sum_{i=1}^{n} (Y_i - bX_i - a) = 0$$

③ $$= \sum_{i=1}^{n} (Y_i) - \sum_{i=1}^{n} bX_i - \sum_{i=1}^{n} a$$ ④ $$\rightarrow \sum_{i=1}^{n} a = \sum_{i=1}^{n} Y_i - b\sum_{i=1}^{n} X_i$$

⑤ $$na = \sum_{i=1}^{n} Y_i - b\sum_{i=1}^{n} X_i \rightarrow$$ ⑥ $$a = (1/n)\sum_{i=1}^{n} Y_i - b(1/n)\sum_{i=1}^{n} X_i$$

$$\boxed{a = \bar{Y} - b\bar{X}}$$

Part 2)

① $$\partial/\partial b = \sum_{i=1}^{n} 2(Y_i - bX_i - a)(-X_i) = 0 \rightarrow$$ ② $$\sum_{i=1}^{n} (Y_i - bX_i - a)(X_i) = 0$$

③ $$\sum_{i=1}^{n} (Y_i X_i) - \sum_{i=1}^{n} bX_i(X_i) - \sum_{i=1}^{n} aX_i = 0$$ (X_i → X_i^2)

④ $$\sum_{i=1}^{n} (Y_i X_i) - \sum_{i=1}^{n} bX_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} Y_i - b\frac{1}{n}\sum_{i=1}^{n} X_i\right)\sum_{i=1}^{n} X_i = 0$$

⑤ $$\sum_{i=1}^{n} (Y_i X_i) - b\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} Y_i \sum_{i=1}^{n} X_i\right) + b(1/n)\sum_{i=1}^{n} X_i \sum_{i=1}^{n} X_i = 0$$

QTM 220 Q1 - Continued

$$\sum_{i=1}^{n}(\gamma_i x_i) - \left(\tfrac{1}{n}\sum_{i=1}^{n}\gamma_i \sum_{i=1}^{n} x_i\right) = b\sum_{i=1}^{n} x_i^2 - b\left(\tfrac{1}{n}\right)\sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n}(\gamma_i x_i) - \left(\tfrac{1}{n}\sum_{i=1}^{n}\gamma_i \sum_{i=1}^{n} x_i\right) = b\left(\sum_{i=1}^{n} x_i^2 - \left(\tfrac{1}{n}\sum_{i=1}^{n} x_i\right)^2\right)$$

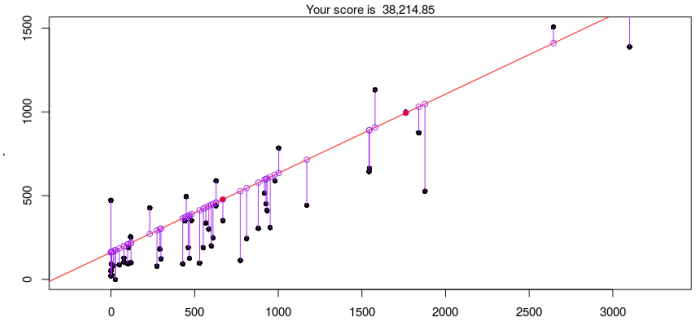
$$\frac{\sum_{i=1}^{n}(\gamma_i x_i) - \left(\frac{1}{n}\sum_{i=1}^{n}\gamma_i \sum_{i=1}^{n} x_i\right)}{\sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2} = b$$

$$\frac{\left[\sum_{i=1}^{n}(\gamma_i x_i) - \left(\frac{1}{n}\sum_{i=1}^{n}\gamma_i x_i\right)\right]}{\sum_{i=1}^{n} x_i^2 - \left(\left(\frac{1}{n}\right)\sum_{i=1}^{n} x_i\right)^2} = \frac{\left[\sum_{i=1}^{n}(x_i - \bar{x})(\gamma_i - \bar{\gamma})\right]}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = b$$

Exercise 2: Web App

Visit the following website https://htgaws.shinyapps.io/premonition/ and read the instructions. Interact with the scatterplot shown in the web page by drawing lines (click the plot on two distinct points to draw a line, click a third point to erase the previous line). The goal of the game is to minimize the score written above the plot after generating a line.

(a) Generate a line and see how the plot changes. Describe how the plot changes and write down the score you got. **[5pts]**



My score was 38,214.85.

(b) What are the purple points? What are the purple lines? **[5pts]**

The purple points represent the place where each observation is on the model. The lines are the residuals of the distances between the actual observations and the model.

(c) Try more than one line and try to minimize the score. How are you selecting the line to minimize the score? Which is your strategy? **[5pts]**

I picked a single point on the graph that is centered around a cluster of points. Then I picked the second point further away to get my first model. After which, I select the same first point, but I change the height of the second point to affect the slope. Then I move in between values until I can minimize the score.

(d) What is the statistical meaning of the score? *Hint: it's something that we studied in class.* **[5pts]**

The score represents the residual sum of the squares (RSS). Therefore, we are trying to minimize the residual sum of the squares.

Exercise 3: Least Squares Regression Models: Sample

For the following exercise, use the provided **sample** data on NBA 2023 season stats, nba.sample.data.csv, which includes the following variables:

- POS - Position (str): position of player (Center (C), Forward (F), Small Forward (SF), Power Forward (PF), Shooting Guard (SG), Point Guard (PG), Guard (G), Not Available (N/A)). • Team - Team Abbreviation (str): The team whose this player is playing for this season in abbreviated term.
- Age - Age (float): The Age of the player.
- GP - Games Played (float): The total games that the player has played in this season.
- W - Wins (float): Total of Games won while the player has played.
- L - Losses (float): Total of Games lost while the player has played.
- Min - Minutes Played (float): Total Minutes the player has played for this season.
- PTS - Points (float) : Total Points made by the player.

We are considering drafting a young player, aged 20, to play center for our team next season. We want to predict the number of points a 20-year-old center will score over a season.

(a) Data cleaning and overview: [5 points]

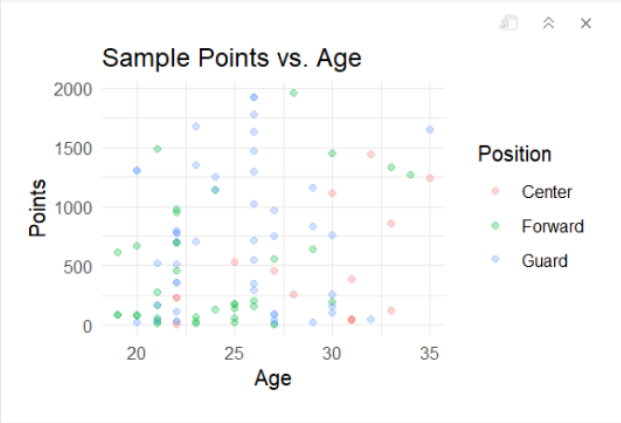
i. Make new variable to group positions to three categories: center, guard, and forward (such that, players labeled as shooting guards, point guards, or guards in the original data are all categorized as guards; players labeled as power forwards, small forwards, or forwards in the original data are all categorized as forwards.)

ii. Remove any rows from your data in which position is N/A

iii. Make a scatter plot with age on the X-axis, points on the Y-axis, and a different color for each position group (center, guard, forward)

```r
nba_2023 <- nba_2023 %>%
  mutate(Position = case_when(
    POS == 'C' ~ 'Center',
    POS %in% c('SG','PG','G') ~ 'Guard',
    POS %in% c('SF','PF','F') ~ 'Forward')) %>%
  na.omit()
```

```r
ggplot(nba_2023, aes(x = Age, y = PTS)) +
  geom_point(aes(x = Age, y = PTS,
                 color = factor(Position)),
             alpha = 0.3) +
  labs(
    title = "Sample Points vs. Age",
    x = "Age",
    y = "Points",
    color = "Position") +
  theme_minimal()
```



(b) Fit a parallel lines model with a different intercept for each position group. Plot your lines on a scatter plot with age on the X-axis, points on the Y-axis, and a different color line for each position group (center, guard, forward). What is the predicted number of points for a Center (C) aged 20? [5 points]

```r
mod.coarsen1 <- lm(PTS ~ Position + Age, data = nba_2023)
mod.coarsen1
```

```
Call:
lm(formula = PTS ~ Position + Age, data = nba_2023)

Coefficients:
    (Intercept)  PositionForward  PositionGuard              Age
        -496.84           160.07         375.72            33.44
```

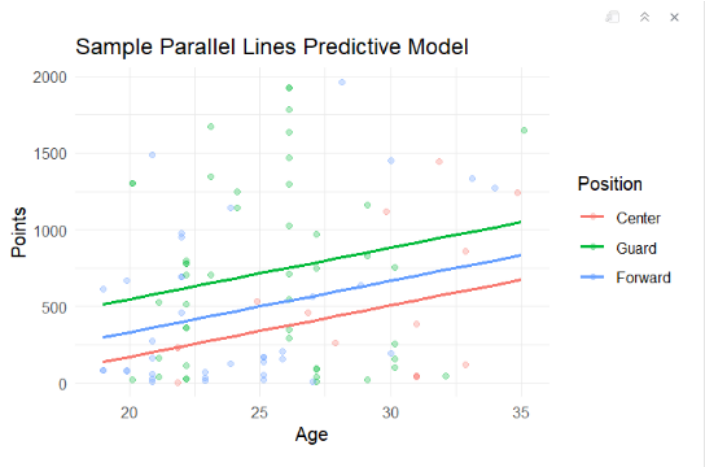
```r
{r}
age_seq <- seq(min(nba_2023$Age), max(nba_2023$Age), by = 1)

pred_data <- expand.grid(
  Age = age_seq,
  Position = c(0, 1, 2))

pred_data$Position <- factor(pred_data$Position,
                             levels = c(0, 1, 2),
                             labels = c("Center", "Guard", "Forward"
))

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

ggplot() +
  geom_point(data = nba_2023, aes(x = Age, y = PTS, color = Position
),
             alpha = 0.3, position=position_dodge(width=0.5)) +
  geom_line(data = pred_data, aes(x = Age, y = predicted_PTS,
                                  color = Position), size = 1) +
  labs(
    title = "Sample Parallel Lines Predictive Model",
    x = "Age",
    y = "Points",
    color = "Position"
  ) +
  theme_minimal()
```



The predicted points for a 20-year-old Center using the sample parallel lines predictive model is ~368 points.

(c) Use bootstrap estimated sampling distribution to make a 95% confidence interval for the predicted points for a 20-year-old Center using this model. Report and interpret your confidence interval. [5 points]

```r
{r}
set.seed(42)

B <- 10000

boot_preds <- numeric(B)

for (b in 1:B) {
  boot_sample <- nba_2023[sample(1:nrow(nba_2023), replace = TRUE), ]

  boot_mod <- lm(PTS ~ Position + Age, data = boot_sample)

  boot_preds[b] <- predict(boot_mod,
                           newdata = data.frame(Age = 20,
                                                Position = "Center"))
}

boot_ci <- quantile(boot_preds, probs = c(0.025, 0.975))
boot_ci
```

```
     2.5%     97.5% 
-115.0936  496.2464 
```

For the sample parallel lines predictive model, if we repeated this estimator many times

with a sufficiently large sample size, the mean of our population of 20-year-old Center players would be between –115.0936 and 496.2464 95% of the time.

(d) Fit a not-necessarily-parallel lines model with a different intercept and slope for each position group. Plot your lines on a scatter plot with age on the X-axis, points on the Y-axis, and a different color line for each position group (center, guard, forward). What is the predicted number of points for a Center (C) aged 20? [5 points]

```r
mod.coarsen2 <- lm(PTS ~ Age + Position + Age*Position, data =
nba_2023)
mod.coarsen2
```

```
Call:
lm(formula = PTS ~ Age + Position + Age * Position, data = nba_2023)

Coefficients:
        (Intercept)                 Age      PositionForward
PositionGuard  Age:PositionForward
           -875.077              46.538              13.653
1427.339               8.832
  Age:PositionGuard
            -39.879
```

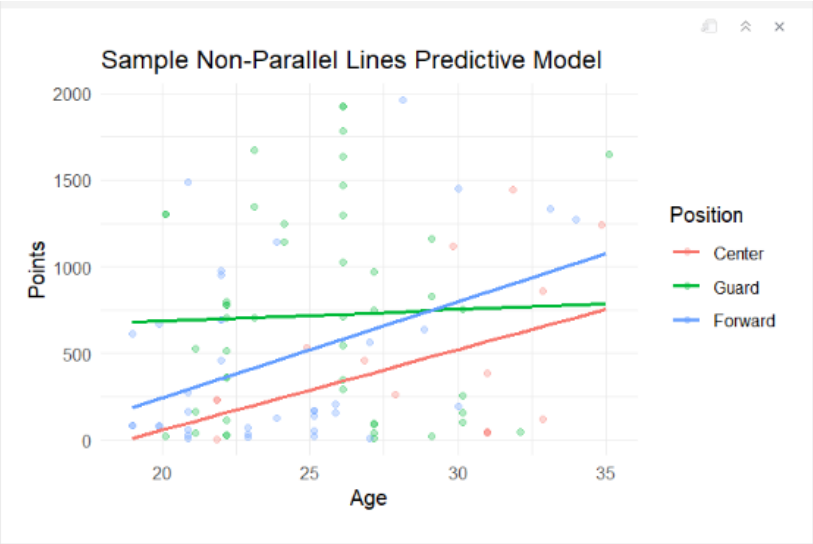
```r
age_seq <- seq(min(nba_2023$Age), max(nba_2023$Age), by = 1)

pred_data <- expand.grid(
  Age = age_seq,
  Position = c(0, 1, 2))

pred_data$Position <- factor(pred_data$Position,
                             levels = c(0, 1, 2),
                             labels = c("Center", "Guard", "Forward"
))

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

ggplot() +
  geom_point(data = nba_2023, aes(x = Age, y = PTS, color = Position
),
             alpha = 0.3, position=position_dodge(width=0.5)) +
  geom_line(data = pred_data, aes(x = Age, y = predicted_PTS,
                                  color = Position), size = 1) +
  labs(
    title = "Sample Non-Parallel Lines Predictive Model",
    x = "Age",
    y = "Points",
    color = "Position"
  ) +
  theme_minimal()
```



The predicted points for a 20-year-old Center using the sample non-parallel lines predictive model are ~250 points.

(e) Use bootstrap estimated sampling distribution to make a 95% confidence interval for the predicted points for a 20-year-old Center using this model. Report and interpret your confidence interval. [5 points]

```{r}
set.seed(42)

B <- 10000

boot_preds <- numeric(B)

for (b in 1:B) {
  boot_sample <- nba_2023[sample(1:nrow(nba_2023), replace = TRUE), ]

  boot_mod <- lm(PTS ~ Age + Position + Age*Position, data =
boot_sample)

  boot_preds[b] <- predict(boot_mod,
                           newdata = data.frame(Age = 20,
                                                Position = "Center"))
}

boot_ci <- quantile(boot_preds, probs = c(0.025, 0.975))
boot_ci
```

```
     2.5%     97.5%
-254.9305  368.3993
```

For the sample non-parallel lines predictive model, if we repeated this estimator many times with a sufficiently large sample size, the mean of our population of 20-year-old Center players would be between –254.9305 and 368.3993 95% of the time.

(f) Fit an additive model with a different intercept for each position group. Plot your curves on a scatter plot with age on the X-axis, points on the Y-axis, and a different color line for each position group (center, guard, forward). What is the predicted number of points for a Center (C) aged 20? [5 points]

```r
{r}
mod.coarsen3 <- lm(PTS ~ factor(Age) + factor(Position), data =
nba_2023)
mod.coarsen3
```

```
Call:
lm(formula = PTS ~ factor(Age) + factor(Position), data = nba_2023)

Coefficients:
          (Intercept)            factor(Age)20
factor(Age)21            factor(Age)22
                220.15                   239.47
-4.80                  146.81
          factor(Age)23            factor(Age)24
factor(Age)25            factor(Age)26
                304.30                   578.72
-70.97                 635.02
          factor(Age)27            factor(Age)28
factor(Age)29            factor(Age)30
                -26.36                   870.76
288.37                 234.32
          factor(Age)31            factor(Age)32
factor(Age)33            factor(Age)34
                -89.15                   430.56
536.79                1013.67
          factor(Age)35  factor(Position)Forward
factor(Position)Guard
               1127.06                    39.19
192.59
```

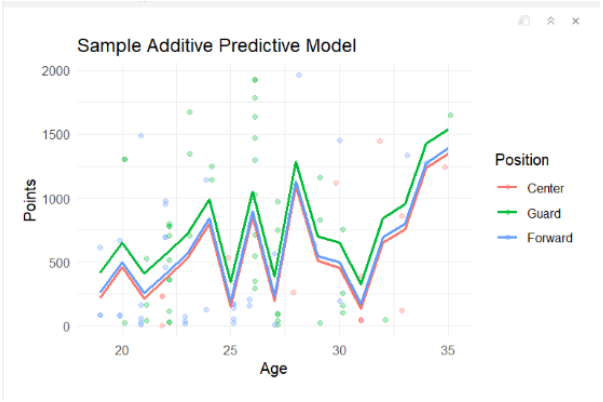
```r
{r}
age_seq <- seq(min(nba_2023$Age), max(nba_2023$Age), by = 1)

pred_data <- expand.grid(
  Age = age_seq,
  Position = c(0, 1, 2))

pred_data$Position <- factor(pred_data$Position,
                             levels = c(0, 1, 2),
                             labels = c("Center", "Guard", "Forward"
))

pred_data$predicted_PTS <- predict(mod.coarsen3, newdata = pred_data)

ggplot() +
  geom_point(data = nba_2023, aes(x = Age, y = PTS, color = Position
),
             alpha = 0.3, position=position_dodge(width=0.5)) +
  geom_line(data = pred_data, aes(x = Age, y = predicted_PTS,
                                  color = Position), size = 1) +
  labs(
    title = "Sample Additive Predictive Model",
    x = "Age",
    y = "Points",
    color = "Position"
  ) +
  theme_minimal()
```



The predicted points for a 20-year-old Center using the sample additive lines predictive model are ~500 points.

(g) Use bootstrap estimated sampling distribution to make a 95% confidence interval for the predicted points for a 20-year-old Center using this model. Report and interpret your confidence interval. [5 points]

```{r}
set.seed(42)

B <- 10000

boot_preds <- numeric(B)

for (b in 1:B) {

  boot_sample <- nba_2023[sample(1:nrow(nba_2023), replace = TRUE), ]

  if (any(boot_sample$Age == 20)) {
    boot_mod <- lm(PTS ~ factor(Age) + factor(Position), data =
boot_sample)

    boot_preds[b] <- predict(boot_mod,
            newdata = data.frame(Age = 20, Position = "Center"))
  } else {
    boot_preds[b] <- NA
  }
}

boot_ci <- quantile(boot_preds, probs = c(0.025, 0.975), na.rm = TRUE
)
boot_ci
```

```
     2.5%     97.5%
-190.2244 1119.2791
```

For the sample additive lines predictive model, if we repeated this estimator many times with a sufficiently large sample size, the mean of our population of 20-year-old Center players would be between –190.2244 and 1119.2791 95% of the time.

Exercise 4: Least Squares Regression Models: Population

For the following exercise, use the provided population data on NBA 2023 season stats, nba.data.csv, which includes the same variables as the sample data.

We want to compare the estimates from our sample to the predicted values we get from fitting the models on the population.

(a) Data cleaning and overview: [5 points]

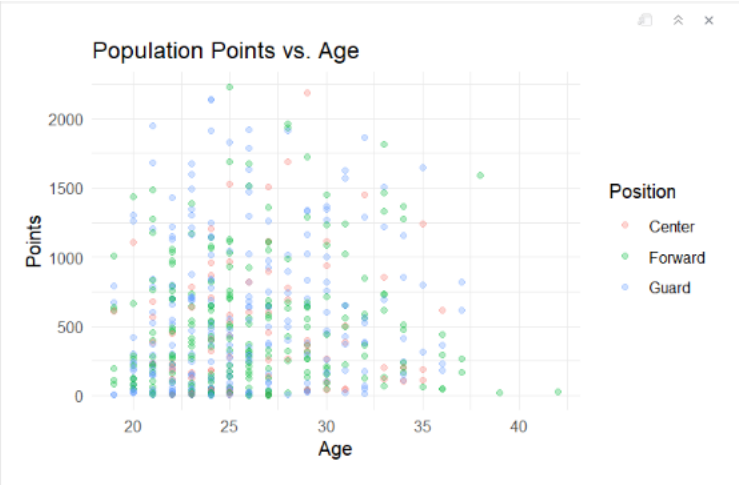
i. Make new variable to group positions to three categories: center, guard, and forward (such that, players labeled as shooting guards, point guards, or guards in the original data are all categorized as guards; players labeled as power forwards, small forwards, or forwards in the original data are all categorized as forwards.)

ii. Remove any rows from your data in which position is N/A

iii. Make a scatter plot with age on the X-axis, points on the Y-axis, and a different color for each position group (center, guard, forward)

```{r}
nba_pop <- nba_pop %>%
  mutate(Position = case_when(
    POS == 'C' ~ 'Center',
    POS %in% c('SG','PG','G') ~ 'Guard',
    POS %in% c('SF','PF','F') ~ 'Forward')) %>%
  na.omit()

ggplot(nba_pop, aes(x = Age, y = PTS)) +
  geom_point(aes(x = Age, y = PTS,
                 color = factor(Position)),
             alpha = 0.3) +
  labs(
    title = "Population Points vs. Age",
    x = "Age",
    y = "Points",
    color = "Position") +
  theme_minimal()
```



(b) Fit a parallel lines model with a different intercept for each position group on the population. Plot your lines on a scatter plot with age on the X-axis, points on the Y-axis, and a different color line for each position group (center, guard, forward). What is the predicted number of points for a Center aged 20 from this model? How does it compare to your estimate and confidence interval from the sample? What is the sub-population mean points for Centers aged 20? How does it compare to your predicted value from the population and your predicted value that you estimated in your sample? [5 points]

```r
mod.coarsen4 <- lm(PTS ~ Position + Age, data = nba_pop)
mod.coarsen4
```

```
Call:
lm(formula = PTS ~ Position + Age, data = nba_pop)

Coefficients:
    (Intercept)  PositionForward    PositionGuard              Age
         185.44            32.44            87.57            11.18
```

```r
age_seq <- seq(min(nba_pop$Age), max(nba_pop$Age), by = 1)

pred_data <- expand.grid(
  Age = age_seq,
  Position = c(0, 1, 2))

pred_data$Position <- factor(pred_data$Position,
                             levels = c(0, 1, 2),
                             labels = c("Center", "Guard", "Forward"
))

pred_data$predicted_PTS <- predict(mod.coarsen4, newdata = pred_data)

ggplot() +
  geom_point(data = nba_pop, aes(x = Age, y = PTS, color = Position),
             alpha = 0.3, position=position_dodge(width=0.5)) +
  geom_line(data = pred_data, aes(x = Age, y = predicted_PTS,
                                  color = Position), size = 1) +
  labs(
    title = "Population Parallel Lines Predictive Model",
    x = "Age",
    y = "Points",
    color = "Position"
  ) +
  theme_minimal()
```

The predicted points for a 20-year-old Center using the population parallel lines predictive model is ~460 points. It is larger than the estimate made by the sample parallel lines predictive model, but it is within the confidence interval from the sample.

The subpopulation mean for 20-year-old Centers is:

```{r}
mean(nba_pop$PTS[nba_pop$Position == "Center" & nba_pop$Age == 20],
na.rm = TRUE)
```

```
[1] 1109
```

The subpopulation mean is larger than both values predicted in the population and sample parallel lines predictive model.

(c) Fit a not-necessarily-parallel lines model with a different intercept and slope for each position group on the population. Plot your lines on a scatter plot with age on the X-axis, points on the Y-axis, and a different color line for each position group (center, guard, forward). What is the predicted number of points for a Center aged 20 from this model? How does it compare to your estimate and confidence interval from the sample? What is the sub-population mean points for Centers aged 20? How does it compare to your predicted value from the population and your predicted value that you estimated in your sample? [5 points]

```{r}
mod.coarsen5 <- lm(PTS ~ Age + Position + Age*Position, data =
nba_pop)
mod.coarsen5
```

```
Call:
lm(formula = PTS ~ Age + Position + Age * Position, data = nba_pop)

Coefficients:
        (Intercept)                 Age      PositionForward
PositionGuard  Age:PositionForward
            298.518               6.929             -18.376
-125.863               1.870
  Age:PositionGuard
              8.165
```

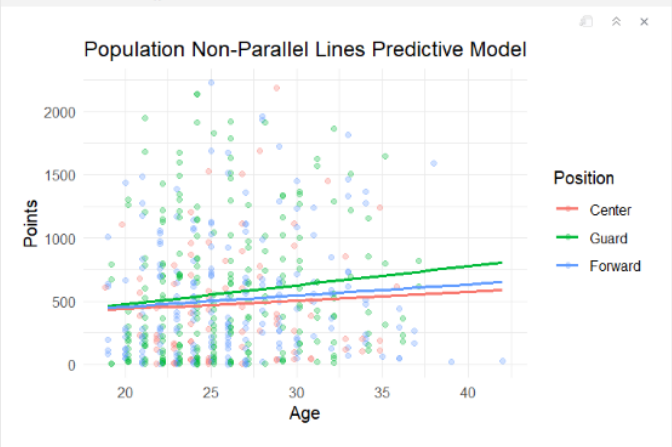
```r
{r}
age_seq <- seq(min(nba_pop$Age), max(nba_pop$Age), by = 1)

pred_data <- expand.grid(
  Age = age_seq,
  Position = c(0, 1, 2))

pred_data$Position <- factor(pred_data$Position,
                             levels = c(0, 1, 2),
                             labels = c("Center", "Guard", "Forward"
))

pred_data$predicted_PTS <- predict(mod.coarsen5, newdata = pred_data)

ggplot() +
  geom_point(data = nba_pop, aes(x = Age, y = PTS, color = Position),
             alpha = 0.3, position=position_dodge(width=0.5)) +
  geom_line(data = pred_data, aes(x = Age, y = predicted_PTS,
                                  color = Position), size = 1) +
  labs(
    title = "Population Non-Parallel Lines Predictive Model",
    x = "Age",
    y = "Points",
    color = "Position"
  ) +
  theme_minimal()
```



The predicted points for a 20-year-old Center using the population non-parallel lines predictive model is ~490 points. It is larger than the estimate made by the sample non-parallel lines predictive model, and it is not within the confidence interval from the sample.

The subpopulation mean for 20-year-old Centers is 1109 points. The subpopulation mean is larger than both values predicted in the population and sample non-parallel lines predictive model.

(d) Fit an additive model with a different intercept for each position group on the population. Plot your curves on a scatter plot with age on the X-axis, points on the Y-axis, and a different color line for each position group (center, guard, forward). What is the predicted number of points for a Center aged 20 from this model? How does it compare to your estimate and confidence interval from the sample? What is the sub-population mean points for Centers aged 20? How does it compare to your predicted value from the population and your predicted value that you estimated in your sample? [5 points]

```r
mod.coarsen6 <- lm(PTS ~ factor(Age) + factor(Position), data =
nba_pop)
mod.coarsen6
```

```
Call:
lm(formula = PTS ~ factor(Age) + factor(Position), data = nba_pop)

Coefficients:
          (Intercept)           factor(Age)20
factor(Age)21           factor(Age)22
                367.63                  -99.75
57.65                   11.34
          factor(Age)23           factor(Age)24
factor(Age)25           factor(Age)26
                 39.51                  101.95
74.47                  147.31
          factor(Age)27           factor(Age)28
factor(Age)29           factor(Age)30
                 89.37                  365.23
141.06                 168.93
          factor(Age)31           factor(Age)32
factor(Age)33           factor(Age)34
                163.21                  115.39
404.33                  67.99
          factor(Age)35           factor(Age)36
factor(Age)37           factor(Age)38
                204.29                 -151.59
24.54                 1176.65
          factor(Age)39           factor(Age)42
factor(Position)Forward    factor(Position)Guard
               -396.35                 -386.35
45.72                  103.94
```

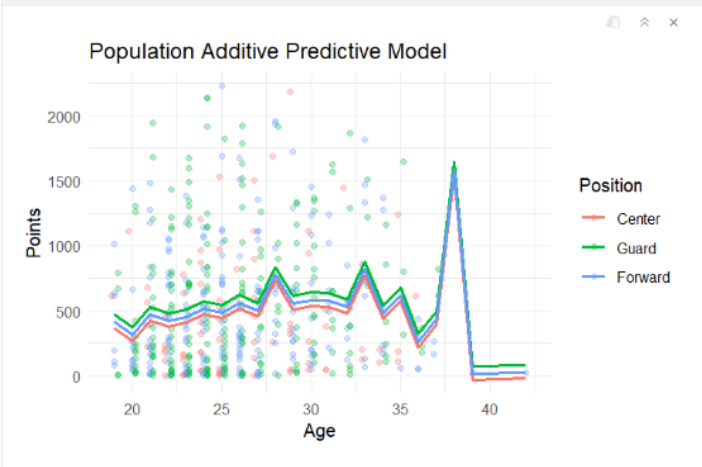
```r
age_seq <- unique(nba_pop$Age)

pred_data <- expand.grid(
  Age = age_seq,
  Position = c(0, 1, 2))

pred_data$Position <- factor(pred_data$Position,
                             levels = c(0, 1, 2),
                             labels = c("Center", "Guard", "Forward"
))

pred_data$predicted_PTS <- predict(mod.coarsen6, newdata = pred_data)

ggplot() +
  geom_point(data = nba_pop, aes(x = Age, y = PTS, color = Position),
             alpha = 0.3, position=position_dodge(width=0.5)) +
  geom_line(data = pred_data, aes(x = Age, y = predicted_PTS,
                                  color = Position), size = 1) +
  labs(
    title = "Population Additive Predictive Model",
    x = "Age",
    y = "Points",
    color = "Position"
  ) +
  theme_minimal()
```



The predicted points for a 20-year-old Center using the population parallel lines predictive model is ~280 points. It is smaller than the estimate made by the sample additive lines predictive model, but it is within the confidence interval from the sample.

The subpopulation mean for 20-year-old Centers is 1109 points. The subpopulation mean is larger than both values predicted in the population and sample non-parallel lines predictive model.

(e) Now that you have seen the population, if you had to choose one of these three models to predict the points for a 20-year-old Center, which would you prefer to use? Explain your reasoning. [5 points]

I would use the additive lines model. While the predictions from this model for both the sample and the population were much smaller than the actual value in the population, it was the only model to include the actual value in the sample confidence interval. Therefore, while the model features the highest bias (it is overfit), it has the largest variability. Therefore, it is the model that is the most likely to contain the estimand.

3 Submission Instructions

Please submit your completed exercises by **October 31** through **Gradescope**. Ensure that your solutions are well-organized, clearly written, and include all necessary calculations and explanations. Questions about submission should be directed to your TA.

4 Helpful Resources

To better assist you in the completion of this exercise sheet, we suggest you to review the following material:

- **Lecture 3** - bootstrapping estimated sampling distributions and confidence intervals
- **Lecture 6** – estimators ;
- **Lecture 11** - least squares regression;
- **Lecture 12** - bias in least squares regression;
- **Lecture 13** - misspecification in least squares regression;
- **Lab** - practicing all of the above