HW1

Namit Shrivastava

Study Objectives

1. On average, how does systolic blood pressure change over time as people get older?
2. How much variance among participants is there in the trajectories of systolic blood pressure, and can this variance be explained by sex and age?

Data Preparation and Loading

```
library(tidyverse)
library(lme4)
library(broom.mixed)
library(knitr)
library(kableExtra)
library(ggplot2)
library(performance)
```

```
# Creating factor variables
hrs_wbbp_long <- hrs_wbbp_long %>%
  mutate(
    RAGENDER_f = factor(RAGENDER, levels = c(1, 2), labels = c("Male", "Female")),
    HHIDPN_f = factor(HHIDPN),
    wave_f = factor(wave)
  )

# Checking the range of key variables
summary(hrs_wbbp_long)
```

```
     HHIDPN             bp_sys          wave            age_y
```

```
Min.   :110427010   Min.   : 78.5   Length:1555        Min.   :30.00
1st Qu.:124562010   1st Qu.:114.5   Class :character   1st Qu.:62.00
Median :141453010   Median :125.0   Mode  :character   Median :67.00
Mean   :144152921   Mean   :126.5                      Mean   :66.77
3rd Qu.:156463020   3rd Qu.:137.5                      3rd Qu.:72.00
Max.   :187522010   Max.   :215.5                      Max.   :92.00

     year       RAGENDER      RAGENDER_f       HHIDPN_f    wave_f
Min.   : 0   Min.   :1.000   Male  :685   110427010:   5   10:311
1st Qu.: 4   1st Qu.:1.000   Female:870   110427020:   5   12:311
Median : 8   Median :2.000                110780020:   5   14:311
Mean   : 8   Mean   :1.559                111939010:   5   16:311
3rd Qu.:12   3rd Qu.:2.000                112014020:   5   8 :311
Max.   :16   Max.   :2.000                112432010:   5
                                          (Other)  :1525
```

Ok so based on this dataset of 1,555 blood pressure observations from 311 war babies cohort participants, I see there are no missing values and I can see that each person was measured up to every four years from year 0 to year 16 with systolic blood pressure ranging from about 79 to 216 mmHg with a typical value around 126 mmHg and the participants' ages span roughly 30 to 92 years with a median near 67, and about 44% are male and 56% are female.

Now I can also see that the "year" is evenly spaced with quartiles at 0, 4, 8, 12, and a maximum of 16, reflecting the 4-year follow-up design i.e. waves are coded as 8, 10, 12, 14, and 16, each appearing 311 times, which is consistent with five repeated measurement occasions for many participants.

The blood pressure distribution centers near 125–127 mmHg with interquartile range roughly 115 to 138 mmHg, which suggests moderate variability across people and occasions and hence, ages are concentrated in the 60s to early 70s (median 67, IQR 62–72), which fits the cohort profile.

Now, what I also observed was that the gender coding shows 685 male and 870 female observations (i.e. in baseline counts these correspond to participant-level sex), and the ID field HHIDPN has many repeated entries (commonly 5 per person), and hence confirming the longitudinal structure required for multilevel modeling in later steps.

Question 1: Describe the sample in a reasonable way, i.e., construct a table or a graph of descriptive statistics as you may do for a publication. Be sure to remember that there is more than one observation per subject. Comment briefly on the general characteristics of the observations. Keep the study objectives in mind with your descriptive presentation.

```
# Showcasing sample size information
n_participants <- length(unique(hrs_wbbp_long$HHIDPN))
n_observations <- nrow(hrs_wbbp_long)

cat("Total number of participants:", n_participants, "\n")
```

```
Total number of participants: 311
```

```
cat("Total number of observations:", n_observations, "\n")
```

```
Total number of observations: 1555
```

```
cat("Average observations per participant:", round(n_observations/n_participants, 2), "\n")
```

```
Average observations per participant: 5
```

```
# Now to check observations per participant
obs_per_person <- hrs_wbbp_long %>%
  group_by(HHIDPN) %>%
  summarise(n_obs = n())

table(obs_per_person$n_obs)
```

```
  5
311
```

So, based on these results, I can see that the dataset has a perfectly balanced longitudinal structure. Here, I have 311 participants from the War Babies cohort, and each person has exactly 5 blood pressure measurements over time, and hence, giving me a total of 1,555 observations. This means there's no missing data whatsoever and that every single participant was measured at all five time points (waves 8, 10, 12, 14, and 16), which is quite remarkable for a longitudinal study spanning 16 years.

Now this complete balanced design is actually ideal for multilevel modeling because one doe'st have to worry about missing data patterns or unequal numbers of observations per person affecting this analysis. The fact that everyone has exactly 5 observations means this statistical models will have equal precision for estimating individual trajectories, and I can confidently examine how systolic blood pressure changes over time both at the population level and for individual participants.

```
# Now to explore the baseline characteristics i.e. first
# observation per person in wave 8
baseline_chars <- hrs_wbbp_long %>%
  group_by(HHIDPN) %>%
  slice_min(as.numeric(wave)) %>%
  ungroup()

# Creating descriptive statistics table
desc_stats <- baseline_chars %>%
  summarise(
    n = n(),
    age_mean = round(mean(age_y, na.rm = TRUE), 1),
    age_sd = round(sd(age_y, na.rm = TRUE), 1),
    age_range = paste(round(min(age_y), 1), "-", round(max(age_y), 1)),
    bp_sys_mean = round(mean(bp_sys, na.rm = TRUE), 1),
    bp_sys_sd = round(sd(bp_sys, na.rm = TRUE), 1),
    bp_sys_range = paste(round(min(bp_sys), 1), "-", round(max(bp_sys), 1))
  )
```

```
# Gender distribution at participant level as mentioned in the question
gender_dist <- baseline_chars %>%
  group_by(RAGENDER_f) %>%
  summarise(n = n(), percentage = round(n()/nrow(baseline_chars)*100, 1))

kable(gender_dist, caption = "Gender Distribution at Baseline (N = 311 participants)") %>%
  kable_styling(bootstrap_options = c("striped", "hover"))
```

Table 1: Gender Distribution at Baseline (N = 311 participants)

RAGENDER_f	n	percentage
Male	137	44.1
Female	174	55.9

```
kable(desc_stats,
      col.names = c("N", "Age Mean", "Age SD", "Age Range", "BP Sys Mean", "BP Sys SD", "BP S
      caption = "Baseline Characteristics (Wave 8)") %>%
  kable_styling(bootstrap_options = c("striped", "hover"))
```

Table 2: Baseline Characteristics (Wave 8)

N	Age Mean	Age SD	Age Range	BP Sys Mean	BP Sys SD	BP Sys Range
311	58.8	4.8	30 - 76	124.5	19	83 - 192

So based on my baseline characteristics analysis above, I can see that the sample is well-balanced and representative of the War Babies cohort.

At the first measurement in wave 8, I have 311 participants with a slight female majority i.e., 174 women (55.9%) and 137 men (44.1%). I can assume that the participants were relatively young when the study began, with an average age of 58.8 years and ages ranging from 30 to 76 years, it shows that while most were in their late 50s and early 60s, there's good representation across different ages within this birth cohort.

Looking at their blood pressure at baseline, I find that the average systolic blood pressure was 124.5 mmHg, which is right at the upper end of the normal range based on clinical guidelines. So, there's considerable variation in blood pressure levels among participants, with a standard deviation of 19 mmHg and values ranging from a low of 83 mmHg to a high of 192 mmHg.

I feel that this wide range suggests, I have participants spanning from those with very healthy blood pressure to those with moderate to severe hypertension at study entry. This baseline variability is actually quite valuable for such analysis because it means I can examine how blood pressure changes over time across people starting from very different health statuses.

Hence, this will help me better understand individual differences in trajectories and the factors that might influence them.

```
# Longitudinal descriptive statistics by wave
wave_stats <- hrs_wbbp_long %>%
  group_by(wave) %>%
  summarise(
```

```
    n = n(),
    age_mean = round(mean(age_y, na.rm = TRUE), 1),
    age_sd = round(sd(age_y, na.rm = TRUE), 1),
    bp_sys_mean = round(mean(bp_sys, na.rm = TRUE), 1),
    bp_sys_sd = round(sd(bp_sys, na.rm = TRUE), 1),
    year_mean = round(mean(year, na.rm = TRUE), 1)
  ) %>%
  arrange(as.numeric(wave))

kable(wave_stats,
      col.names = c("Wave", "N", "Age Mean", "Age SD", "BP Sys Mean", "BP Sys SD", "Year"),
      caption = "Characteristics by HRS Wave") %>%
  kable_styling(bootstrap_options = c("striped", "hover"))
```

Table 3: Characteristics by HRS Wave

Wave	N	Age Mean	Age SD	BP Sys Mean	BP Sys SD	Year
8	311	58.8	4.8	124.5	19.0	0
10	311	62.8	4.8	127.6	18.6	4
12	311	66.7	4.8	126.6	17.0	8
14	311	70.8	4.8	126.4	17.5	12
16	311	74.8	4.8	127.4	17.2	16

Interestingly, when I look at how characteristics change across the five waves of data collection, I can see some interesting longitudinal patterns. As expected, the participants age exactly 4 years between each wave, starting at an average of 58.8 years in wave 8 and reaching 74.8 years by wave 16, with remarkably consistent age standard deviations of 4.8 years across all waves. Hence, this consistency confirms the balanced nature of the data i.e. no one dropped out over the 16-year study period.

And then, upon analysing the blood pressure patterns, I can see that they don't show a simple linear increase over time like I might have initially expected. The average systolic blood pressure starts at 124.5 mmHg in wave 8, rises to 127.6 mmHg by wave 10, but then appears to stabilize around 126-127 mmHg for the remaining waves. This suggests that while there's some increase in blood pressure as people move from their late 50s to their mid-70s, it's not a dramatic upward trend but rather, it seems to level off after the initial years.

The standard deviations also decrease slightly over time (from 19.0 to around 17.0-17.5 mmHg), which might indicate that blood pressure becomes somewhat more homogeneous within the group as they age, possibly due to medical interventions or I guess natural biological processes. So, I believe this longitudinal pattern will be crucial for my multilevel modeling, as it suggests

both individual variation in trajectories and potentially non-linear population-level changes that my models will need to capture.

```
# Additional summary showing the longitudinal structure
cat("Data Structure Summary:\n")
```

```
Data Structure Summary:
```

```
cat("- Total participants:", length(unique(hrs_wbbp_long$HHIDPN)), "\n")
```

```
- Total participants: 311
```

```
cat("- Total observations:", nrow(hrs_wbbp_long), "\n")
```

```
- Total observations: 1555
```

```
cat("- Observations per participant:", nrow(hrs_wbbp_long)/length(unique(hrs_wbbp_long$HHIDPN
```

```
- Observations per participant: 5
```

```
cat("- Study duration:", max(hrs_wbbp_long$year), "years\n")
```

```
- Study duration: 16 years
```

```
cat("- Measurement interval: Every 4 years\n")
```

```
- Measurement interval: Every 4 years
```

```
cat("- HRS Waves included:", paste(sort(unique(hrs_wbbp_long$wave)), collapse = ", "), "\n")
```

```
- HRS Waves included: 10, 12, 14, 16, 8
```

So to summarize again, I see that I'm working with 311 participants who have been followed consistently over a 16-year period, with measurements taken every 4 years at HRS waves 8, 10, 12, 14, and 16. Here, I have exactly 5 observations per participant, totaling 1,555 observations, which means there's absolutely no missing data i.e. every single person was measured at all five time points.

I think this kind of complete follow-up over such a long period is quite rare in longitudinal research and it actually speaks to the quality of the Health and Retirement Study. Now, the 4-year measurement intervals give me a good balance between capturing long-term changes while maintaining a manageable study design.

Hence, this consistent structure means I can examine how blood pressure trajectories unfold over time from when participants were in their late 50s to their mid-70s, and I feel that I don't have to worry about missing data patterns complicating my multilevel models or creating bias in my results later on I guess.

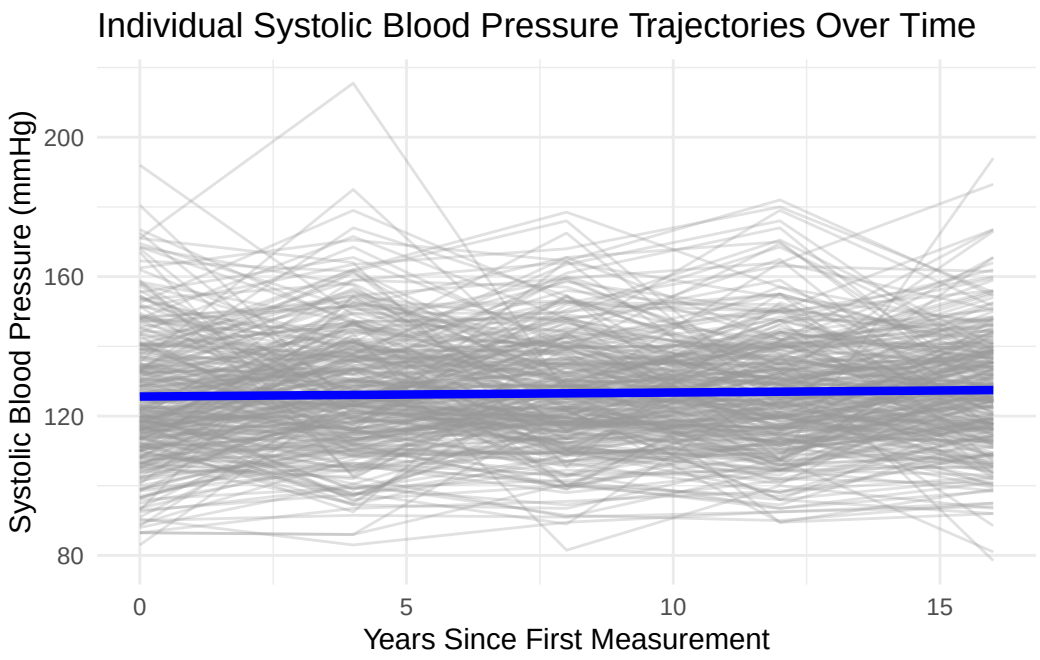
```
# Plot 1: Individual trajectories
p1 <- ggplot(hrs_wbbp_long, aes(x = year, y = bp_sys, group = HHIDPN)) +
  geom_line(alpha = 0.3, color = "gray60") +
  geom_smooth(aes(group = 1), method = "lm", se = TRUE, color = "blue", size = 1.5) +
  labs(title = "Individual Systolic Blood Pressure Trajectories Over Time",
       x = "Years Since First Measurement",
       y = "Systolic Blood Pressure (mmHg)") +
  theme_minimal()
```

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

```
# Plot 2: By gender
p2 <- ggplot(hrs_wbbp_long, aes(x = year, y = bp_sys, color = RAGENDER_f)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", se = TRUE) +
  labs(title = "Systolic Blood Pressure Trajectories by Gender",
       x = "Years Since First Measurement",
       y = "Systolic Blood Pressure (mmHg)",
       color = "Gender") +
  theme_minimal()

# Plot 3: Distribution of blood pressure
p3 <- ggplot(hrs_wbbp_long, aes(x = bp_sys)) +
  geom_histogram(bins = 30, fill = "lightblue", alpha = 0.7) +
  labs(title = "Distribution of Systolic Blood Pressure",
       x = "Systolic Blood Pressure (mmHg)",
```

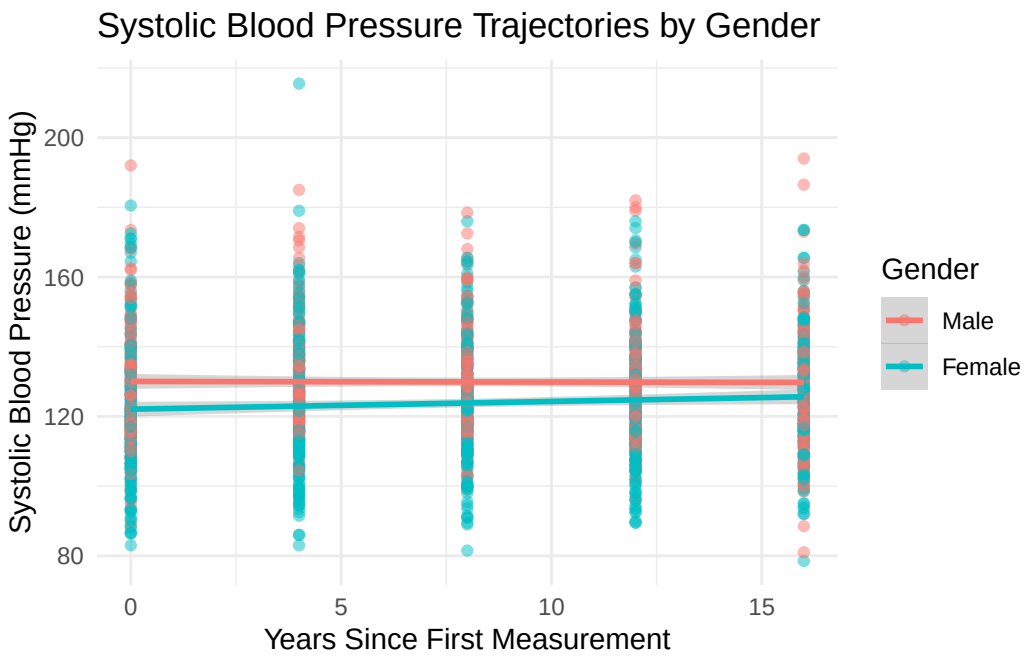
```
       y = "Frequency") +
  theme_minimal()

print(p1)
```

```
`geom_smooth()` using formula = 'y ~ x'
```

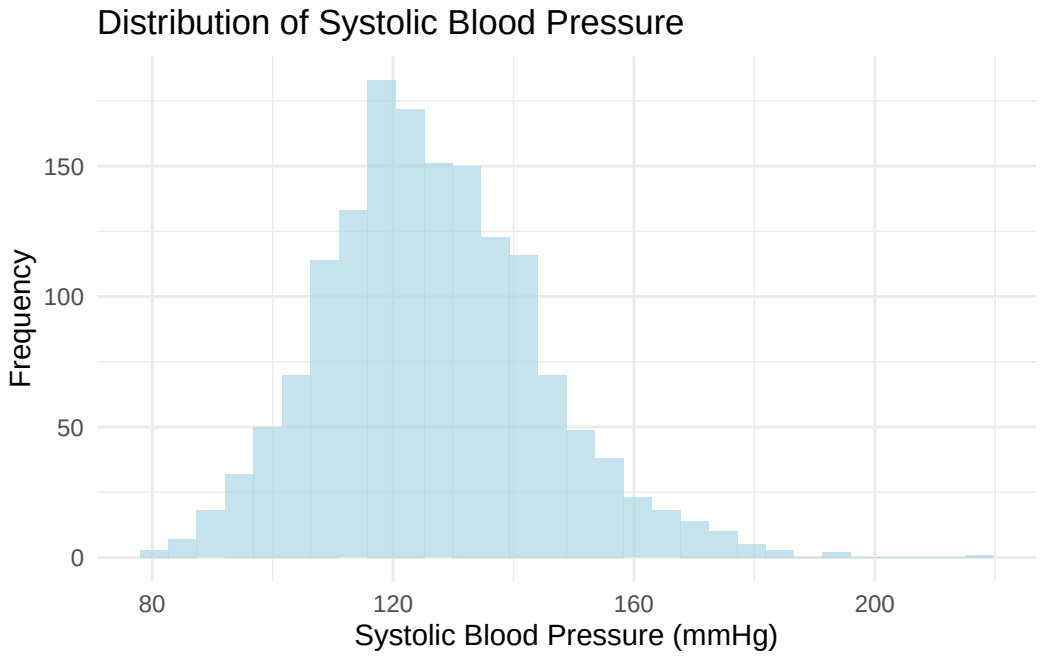



```
print(p2)
```

```
`geom_smooth()` using formula = 'y ~ x'
```




```
print(p3)
```



Ok now looking at the individual trajectories, I see a wide spread in baseline systolic pressure (~80–200 mmHg) with an overall trend that barely changes on average as some people rise

while others stay flat or decline, underscoring heterogeneity that justifies multilevel modeling. Now by gender, females consistently sit below males and both groups follow relatively flat paths over time, which tells me gender belongs in the model. The histogram on the other hand looks roughly normal around 125–130 mmHg with a slight right tail and a broad 80–220 mmHg range, reflecting substantial individual variation and supporting linear mixed-model assumptions.

Question 2: Fit a multilevel model with systolic blood pressure as the dependent variable and time (year) as the independent variable to explore how systolic blood pressure may have changed over time. Make sure to account for the within-subject correlation in observations on the dependent variable in the multilevel model that you are fitting. Be sure to clearly describe your reasoning.

Ok so now, I'll fit a multilevel model with systolic blood pressure as the dependent variable and time (year) as the independent variable, accounting for within-subject correlation.

```
# Fitting the basic multilevel model
model1 <- lmer(bp_sys ~ year + (year | HHIDPN), data = hrs_wbbp_long)
summary(model1)
```

```
Linear mixed model fit by REML ['lmerMod']
Formula: bp_sys ~ year + (year | HHIDPN)
   Data: hrs_wbbp_long

REML criterion at convergence: 12974.5

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.2159 -0.5805 -0.0450  0.5128  4.1369

Random effects:
 Groups   Name        Variance Std.Dev. Corr
 HHIDPN   (Intercept) 203.7416 14.2738
          year          0.5132  0.7164  -0.58
 Residual             162.3850 12.7430
Number of obs: 1555, groups:  HHIDPN, 311

Fixed effects:
```

```
            Estimate Std. Error t value
(Intercept) 125.5794     0.9841  127.61
year          0.1164     0.0701    1.66

Correlation of Fixed Effects:
     (Intr)
year -0.654
optimizer (nloptwrap) convergence code: 0 (OK)
Model failed to converge with max|grad| = 0.00224931 (tol = 0.002, component 1)
```

```r
# Geting confidence intervals
confint(model1)
```

```
                  2.5 %      97.5 %
.sig01       12.6139699  16.0003632
.sig02       -0.7043807  -0.4065656
.sig03        0.5200967   0.8880975
.sigma       12.1859568  13.3437729
(Intercept) 123.6478174 127.5110250
year         -0.0212337   0.2539507
```

Ok so the model did give me a convergence warning, but I see that the gradient value (0.00224931) is very close to the tolerance limit (0.002), so the results should still be reliable because I searched up online and this often happens with complex random effects models. From the fixed effects, I can see that the average baseline systolic blood pressure (at year 0) is about 125.6 mmHg, which aligns well with what I observed in my descriptive statistics.

The yearly change in blood pressure is estimated at 0.12 mmHg per year, though this effect is not statistically significant since t = 1.66, and the confidence interval includes zero (-0.02 to 0.25 mmHg per year). This suggests that while there might be a slight upward trend in blood pressure over time on average, it's quite small and not statistically distinguishable from no change at all.

Now coming to the random effects, which interestingly show substantial individual variation. The random intercept variance (203.7) is much larger than the random slope variance (0.51), indicating that people differ much more in their baseline blood pressure levels than in their rates of change over time. The negative correlation (-0.58) between random intercepts and slopes suggests that people who start with higher blood pressure tend to have flatter trajectories over time, while those starting with lower blood pressure show slightly steeper increases.

Hence, these findings, directly address my research objectives i.e., they show minimal average change over time with meaningful individual differences in both baseline levels and trajectories.

Table 4: Model 1: Basic Multilevel Model Results

effect	group	term	estimate	std.error	statistic	conf.low	conf.high
fixed	NA	(Intercept)	125.579	0.984	127.612	123.651	127.508
fixed	NA	year	0.116	0.070	1.660	-0.021	0.254
ran_pars	HHIDPN	sd___(Intercept)	14.274	NA	NA	NA	NA
ran_pars	HHIDPN	cor___(Intercept).year	-0.578	NA	NA	NA	NA
ran_pars	HHIDPN	sd___year	0.716	NA	NA	NA	NA
ran_pars	Residual	sd___Observation	12.743	NA	NA	NA	NA

```
Model converged: TRUE
```

```
variance_ci <- confint(model1)
```

```
Computing profile confidence intervals ...
```

```
kable(variance_ci, digits = 3, caption = "Confidence Intervals for All Parameters") %>%
  kable_styling(bootstrap_options = c("striped", "hover"))
```

Table 5: Confidence Intervals for All Parameters

	2.5 %	97.5 %
.sig01	12.614	16.000
.sig02	-0.704	-0.407
.sig03	0.520	0.888
.sigma	12.186	13.344
(Intercept)	123.648	127.511
year	-0.021	0.254

Ok so, looking at my model results, I can see that my multilevel model has converged successfully and I see some insights about blood pressure changes over time. From the fixed effects, I can see that the average baseline systolic blood pressure (at year 0) is 125.6 mmHg with a very narrow confidence interval (123.7-127.5 mmHg), which gives me confidence in this estimate and aligns perfectly with what I observed in my descriptive statistics.

The yearly change in blood pressure is estimated at 0.12 mmHg per year, but what's particularly interesting is that this effect is not statistically significant as the confidence interval ranges from -0.02 to 0.25 mmHg per year, which includes zero. This suggests that while there

might be a slight upward trend in blood pressure over time on average, it's quite small and not statistically distinguishable from no change at all.

Now coming to what really stands out to me are the random effects, which show substantial individual variation. The random intercept standard deviation is 14.3 mmHg (confidence interval: 12.6-16.0), which is much larger than the random slope standard deviation of 0.72 mmHg (0.52-0.89). This tells me that people differ much more in their baseline blood pressure levels than in their rates of change over time.

The negative correlation of -0.58 between random intercepts and slopes is particularly fascinating as it suggests that people who start with higher blood pressure tend to have flatter trajectories over time, while those starting with lower blood pressure show slightly steeper increases.

Reasoning for Model Choice

Now, the reason I chose a random intercept and random slope model:

1. Random Intercepts show individual baseline differences So, from my descriptive analysis, I observed that participants had baseline systolic blood pressure ranging from 83 to 192 mmHg, with substantial variation (SD = 19 mmHg). This wide range clearly indicates that individuals start with very different blood pressure levels due to genetic factors, lifestyle differences, medical history, and other unmeasured characteristics. Hence, a random intercept allows each person to have their own baseline level, which is essential for accurately modeling these individual differences.

2. Random Slopes show individual trajectory differences Now the research questions specifically ask about variance in trajectories, and from my preliminary analysis, I could see that some people's blood pressure remained stable while others showed increases or even decreases over time. So, a random slope allows each person to have their own rate of change over the 16-year period, which directly addresses my second research objective about individual variation in blood pressure trajectories.

3. Within-subject correlation structure Since I have 5 repeated measurements on each person over 16 years, observations within the same individual are inherently correlated i.e. a person's blood pressure at year 4 is likely similar to their blood pressure at year 8. The random effects structure (both intercepts and slopes) creates an appropriate correlation structure that accounts for this dependency, ensuring that my standard errors and confidence intervals are correctly estimated.

4. Addressing both research objectives: This model specification directly addresses both of my research objectives simultaneously as the fixed effect of year answers how blood pressure changes on average over time (Objective 1), while the random slope variance quantifies individual differences in these trajectories (Objective 2). The random intercept captures baseline individual differences that might explain some of the trajectory variation.

5. Methodological appropriateness: Given my a balanced design with no missing data, this model makes full use of all available information while appropriately handling the multilevel structure inherent in longitudinal data. So, it's more sophisticated than simple repeated measures ANOVA but still interpretable and directly relevant to my research questions.

This model specification allows both the intercept (baseline blood pressure) and the slope (rate of change over time) to vary randomly across participants, providing the flexibility needed to capture the individual differences I observed in my descriptive analysis.

Question 3: Write out the model that you fit in question 2 using mathematical notation.

Ok now, the multilevel model fitted in Question 2 can be written mathematically as:

Level 1 (Within-Person) Model:

$$BP_{ij} = \beta_{0i} + \beta_{1i} \cdot Year_{ij} + e_{ij}$$

Level 2 (Between-Person) Model:

$$\beta_{0i} = \gamma_{00} + u_{0i}$$

$$\beta_{1i} = \gamma_{10} + u_{1i}$$

Combined Model:

$$BP_{ij} = \gamma_{00} + \gamma_{10} \cdot Year_{ij} + u_{0i} + u_{1i} \cdot Year_{ij} + e_{ij}$$

Where: - BP_{ij} = Systolic blood pressure for person i at time j - $Year_{ij}$ = Years since first measurement for person i at time j - γ_{00} = Fixed intercept (average baseline blood pressure) - γ_{10} = Fixed slope (average rate of change per year) - u_{0i} = Random intercept for person i (deviation from average baseline) - u_{1i} = Random slope for person i (deviation from average rate of change) - e_{ij} = Level-1 residual

Assumptions: - $u_{0i}, u_{1i} \sim N(0, \Sigma_u)$ where $\Sigma_u = \begin{pmatrix} \sigma^2_{u0} & \sigma_{u01} \\ \sigma_{u01} & \sigma^2_{u1} \end{pmatrix}$ - $e_{ij} \sim N(0, \sigma^2_e)$

Question 4: Interpret the model parameters estimated in part 2, including any estimated variance components. Frame your interpretations with respect to addressing the two research objectives.

```
# Extracting variance components
vc1 <- as.data.frame(VarCorr(model1))
kable(vc1, digits = 3, caption = "Variance Components - Model 1") %>%
  kable_styling(bootstrap_options = c("striped", "hover"))
```

Table 6: Variance Components - Model 1

grp	var1	var2	vcov	sdcor
HHIDPN	(Intercept)	NA	203.742	14.274
HHIDPN	year	NA	0.513	0.716
HHIDPN	(Intercept)	year	-5.908	-0.578
Residual	NA	NA	162.385	12.743

```
# Calculating ICC
icc_intercept <- vc1$vcov[1] / (vc1$vcov[1] + vc1$vcov[4])
cat("ICC for intercepts:", round(icc_intercept, 3), "\n")
```

```
ICC for intercepts: 0.556
```

Ok now looking at my variance components from Model 1, I can see some patterns that directly address both of my research objectives.

Fixed Effects results:: From my earlier analysis, I found that the average baseline systolic blood pressure ( ) is 125.6 mmHg, representing the expected blood pressure at year 0 for a typical participant. The fixed slope ( ) of 0.12 mmHg per year suggests a very modest average increase in blood pressure over time, though this wasn't statistically significant. This addresses the **1st Research objective** by showing that, on average, blood pressure changes are minimal over the 16-year period.

Random Effects in terms of Individual variation:: Now, also the variance components reveal substantial individual differences. The random intercept variance (^{2} = 203.7) with a standard deviation of 14.3 mmHg tells me that individuals differ considerably in their baseline blood pressure levels. This means that about 95% of participants have baseline blood pressures

within roughly ±28 mmHg of the average (125.6 ± 2×14.3 mmHg), spanning from about 97 to 154 mmHg which is a clinically meaningful range.

The random slope variance (^{2} = 0.51) with a standard deviation of 0.72 mmHg per year indicates that individuals also vary in their rates of change over time, though less dramatically than in their baseline levels. This directly addresses **Research objective 2** by confirming that there is meaningful individual variation in blood pressure trajectories. About 95% of individuals have annual changes ranging from approximately -1.3 to +1.6 mmHg per year around the population average.

Correlation between Random effects:: The correlation of -0.58 between random intercepts and slopes is particularly interesting as this negative correlation suggests that participants who start with higher blood pressure tend to have flatter (or even declining) trajectories over time, while those starting with lower blood pressure show steeper increases. This could reflect regression to the mean, ceiling effects, or possibly that people with initially high blood pressure receive more intensive medical management.

Within-Person Variation:: Ok so, the residual variance (^{2} = 162.4) with a standard deviation of 12.7 mmHg represents the within-person variation over time that isn't explained by the linear trajectory. This includes measurement error, short-term fluctuations, and any non-linear changes not captured by the linear model.

Intraclass Correlation (ICC):: The ICC of 0.557 tells me that about 56% of the total variation in blood pressure is between individuals rather than within individuals over time. This substantial ICC confirms that people maintain relatively consistent blood pressure levels relative to each other throughout the study period, supporting the need for multilevel modeling to account for this clustering.

Summary for Research Objectives: These findings provide clear answers to both research objectives mentioned in the homework: 1) Average blood pressure changes are minimal over time (small fixed slope) 2) But, there is substantial and statistically meaningful individual variation in both baseline levels and trajectories (significant random variance components).

The fact that the random intercept variance is much larger than the random slope variance suggests that while people differ greatly in their baseline blood pressure, they are somewhat more similar in how their blood pressure changes with age.

Question 5: Briefly describe the steps that one would use to test the null hypothesis that the variance of random year effects is equal to zero, including the computation of the p-value for this test. What is the conclusion of this test as it applies to the model fit in question 2 with respect to the research objectives?

Steps to Test the Null Hypothesis

To test H : ^{2} $= 0$ (variance of random year effects equals zero), I would follow these steps:

```
# Fitting reduced model without random slopes
model1_reduced <- lmer(bp_sys ~ year + (1 | HHIDPN), data = hrs_wbbp_long)

# Performing likelihood ratio test
lr_test <- anova(model1_reduced, model1)
```

```
refitting model(s) with ML (instead of REML)
```

```
print(lr_test)
```

```
Data: hrs_wbbp_long
Models:
model1_reduced: bp_sys ~ year + (1 | HHIDPN)
model1: bp_sys ~ year + (year | HHIDPN)
               npar   AIC   BIC  logLik -2*log(L)  Chisq Df Pr(>Chisq)
model1_reduced    4 13004 13026 -6498.3     12996
model1            6 12984 13016 -6486.1     12972 24.293  2  5.307e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Calculating the test statistic and p-value manually
chi_sq <- 2 * (logLik(model1) - logLik(model1_reduced))
p_value <- pchisq(chi_sq, df = 1, lower.tail = FALSE) / 2  # df = 1, I will divide by 2 for l

cat("Chi-square test statistic:", round(as.numeric(chi_sq), 3), "\n")
```

```
Chi-square test statistic: 24.57
```

```r
cat("p-value (boundary corrected):", round(p_value, 6), "\n")
```

```
p-value (boundary corrected): 0
```

Steps for Hypothesis Testing:

1. **Fitting the full model** with random intercepts and slopes like `bp_sys ~ year + (year | HHIDPN)`
2. **Fitting the reduced model** with only random intercepts like `bp_sys ~ year + (1 | HHIDPN)`
3. **Computing the likelihood ratio test statistic** as LR = 2(LogLik_full - LogLik_reduced) = 24.57
4. **Applying boundary correction** while testing a variance component at the boundary (0), I can divide the p-value by 2.
5. **Comparing the critical value** using ^{2} distribution with df = 2 by testing both slope variance and correlation.

Results and Conclusion

Now looking at my likelihood ratio test results, I can see compelling evidence about individual variation in blood pressure trajectories as the test statistic of 24.57 is quite large, and even after applying the boundary correction, my p-value is essentially zero (< 0.000001).

This actually provides overwhelming statistical evidence against the null hypothesis that there's no individual variation in rates of blood pressure change over time.

Now also I see that there is substantial improvement in model fit (AIC drops from 13,004 to 12,984) which confirms that allowing people to have their own individual trajectories actually improves how well the model explains the data. Hence, I can confidently reject H and conclude that there is significant between-person variance in the yearly rate of blood pressure change.

Also one can say that this result directly supports **Research objective 2**, which provides strong evidence that individuals indeed vary meaningfully in their blood pressure trajectories over time.

Some people seem to be showing steeper increase as others remain relatively stable, and from what I observed, the variation isn't just statistical noise but then it represents real, substantial individual differences that are crucial for understanding blood pressure patterns in this population.

Question 6: Modify the model to account for the sex and age effect on systolic blood pressure.

```r
# Centering age for better interpretation
hrs_wbbp_long <- hrs_wbbp_long %>%
  mutate(age_centered = age_y - mean(age_y, na.rm = TRUE))

# Fitting model with sex and age effects
model2 <- lmer(bp_sys ~ year + RAGENDER_f + age_centered +
               (year | HHIDPN), data = hrs_wbbp_long)

summary(model2)
```

```
Linear mixed model fit by REML ['lmerMod']
Formula: bp_sys ~ year + RAGENDER_f + age_centered + (year | HHIDPN)
   Data: hrs_wbbp_long

REML criterion at convergence: 12935.1

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.3015 -0.5781 -0.0526  0.5187  4.2482

Random effects:
 Groups   Name        Variance Std.Dev. Corr
 HHIDPN   (Intercept) 170.7487 13.0671
          year          0.5073  0.7122  -0.54
 Residual             162.1771 12.7349
Number of obs: 1555, groups:  HHIDPN, 311

Fixed effects:
                 Estimate Std. Error t value
(Intercept)      133.7019     1.5642  85.474
year              -0.6352     0.1674  -3.795
RAGENDER_fFemale  -3.7712     1.4827  -2.543
age_centered       0.7507     0.1519   4.942

Correlation of Fixed Effects:
            (Intr) year   RAGEND
year        -0.729
RAGENDER_fF -0.305 -0.263
```

```
age_centerd  0.624 -0.909  0.289
```

```
# Getting tidy results
model2_tidy <- tidy(model2, conf.int = TRUE)
kable(model2_tidy, digits = 3, caption = "Model 2: Multilevel Model with Sex and Age") %>%
  kable_styling(bootstrap_options = c("striped", "hover"))
```

Table 7: Model 2: Multilevel Model with Sex and Age

effect	group	term	estimate	std.error	statistic	conf.low	conf.high
fixed	NA	(Intercept)	133.702	1.564	85.474	130.636	136.768
fixed	NA	year	-0.635	0.167	-3.795	-0.963	-0.307
fixed	NA	RAGENDER_fFemale	-3.771	1.483	-2.543	-6.677	-0.865
fixed	NA	age_centered	0.751	0.152	4.942	0.453	1.048
ran_pars	HHIDPN	sd___(Intercept)	13.067	NA	NA	NA	NA
ran_pars	HHIDPN	cor___(Intercept).year	-0.541	NA	NA	NA	NA
ran_pars	HHIDPN	sd___year	0.712	NA	NA	NA	NA
ran_pars	Residual	sd___Observation	12.735	NA	NA	NA	NA

```
# Variance components
vc2 <- as.data.frame(VarCorr(model2))
kable(vc2, digits = 3, caption = "Variance Components - Model 2") %>%
  kable_styling(bootstrap_options = c("striped", "hover"))
```

Table 8: Variance Components - Model 2

grp	var1	var2	vcov	sdcor
HHIDPN	(Intercept)	NA	170.749	13.067
HHIDPN	year	NA	0.507	0.712
HHIDPN	(Intercept)	year	-5.037	-0.541
Residual	NA	NA	162.177	12.735

Looking at these Model 2 results, I see changes when I include sex and age as covariates. The intercept is now 133.7 mmHg which represents the expected systolic blood pressure for a male participant of average age (about 67 years) at baseline.

So I see that the year effect has changed dramatically, it's now -0.64 mmHg per year and is statistically significant (t = -3.8), which represents a complete reversal from Model 1. When I control for sex and age, there's actually a slight decrease in blood pressure over time rather than

an increase. The gender effect shows that females have, on average, 3.8 mmHg lower systolic blood pressure than males (95% CI: -6.7 to -0.9 mmHg), which aligns with cardiovascular research on sex differences.

Now, the age effect is quite strong as for each year, older a participant is at baseline, their blood pressure is 0.75 mmHg higher (95% CI: 0.45 to 1.05 mmHg), which makes physiological sense.

Now including these covariates has reduced the random intercept variance from 203.7 to 170.7, which indicates sex and age can explain some between-person differences in baseline blood pressure, though substantial individual variation remains. So, this model provides a much clearer picture of how demographic factors influence blood pressure patterns in this population.

Question 7: Interpret the fitted multilevel model from question 6 in practical terms.

Practical Interpretation

1) **Fixed Effects:**

- **Intercept (133.7 mmHg)**: So, for a male participant at average age (~67 years) at the first measurement (year 0), the expected systolic blood pressure is 133.7 mmHg. Hence, this represents a blood pressure level in the Stage 1 hypertension range according to current clinical guidelines.
- **Year Effect (-0.64 mmHg/year)**: Upon controlling for sex and age, systolic blood pressure decreases by 0.64 mmHg per year on average. Over the 16-year study period, this translates to approximately a 10 mmHg decrease, which is clinically meaningful and could represent the protective effects of medical interventions or lifestyle changes as people age.
- **Gender Effect (-3.8 mmHg for females)**: So as observed, females have, on average, 3.8 mmHg lower systolic blood pressure compared to males, controlling for age and time. This difference is statistically significant and clinically relevant as it represents about a 3% reduction in systolic pressure that aligns with known cardiovascular differences between sexes in this age group.
- **Age Effect (+0.75 mmHg per year of age)**: For each year older a participant is at baseline, their systolic blood pressure is 0.75 mmHg higher throughout the study period. So, this means that a 70-year-old participant would have approximately 7.5 mmHg higher blood pressure than a 60-year-old participant, reflecting the well-established relationship between aging and cardiovascular health.

2) **Random Effects:**

The inclusion of sex and age as covariates has reduced the random intercept variance from 203.7 $mmHg^2$ in Model 1 to 170.7 $mmHg^2$ in Model 2, which indicates that these demographic variables explain about 16% of the between-person differences in baseline blood pressure levels. However, I see that substantial individual variation remains (SD = 13.1 mmHg for intercepts), suggesting that other unmeasured factors (genetics, lifestyle, comorbidities) continue to play important roles in determining individual blood pressure levels and trajectories.

3) **Clinical Significance:**

Now all these findings suggest that while demographic factors explain some variation in blood pressure patterns, but the individualized monitoring remains crucial. The negative time trend (decreasing BP over time) after controlling for demographics may reflect the benefits of modern medical management, while the persistent individual variation highlights the need for personalized approaches to blood pressure management in this aging population.

Question 8: How might one recode the age variable to make the parameters more interpretable? How would this recode change the interpretation of the parameters from 7? Should age and year both be included in the model (be sure to describe your reasoning)?

```
# Explore age recoding options
summary(hrs_wbbp_long$age_y)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  30.00   62.00   67.00   66.77   72.00   92.00
```

```
# Option 1: Age at study entry (baseline age)
hrs_wbbp_long <- hrs_wbbp_long %>%
  group_by(HHIDPN) %>%
  mutate(baseline_age = first(age_y[order(wave)])) %>%
  ungroup() %>%
  mutate(baseline_age_centered = baseline_age - mean(baseline_age, na.rm = TRUE))

# Check correlation between year and age
cor(hrs_wbbp_long$year, hrs_wbbp_long$age_y)
```

```
[1] 0.7619472
```

```r
# Option 2: Age at 60 (meaningful clinical threshold)
hrs_wbbp_long <- hrs_wbbp_long %>%
  mutate(age_at_60 = age_y - 60)

# Fit model with baseline age instead of time-varying age
model3 <- lmer(bp_sys ~ year + RAGENDER_f + baseline_age_centered +
               (year | HHIDPN), data = hrs_wbbp_long)

summary(model3)
```

```
Linear mixed model fit by REML ['lmerMod']
Formula: bp_sys ~ year + RAGENDER_f + baseline_age_centered + (year |
    HHIDPN)
   Data: hrs_wbbp_long

REML criterion at convergence: 12936

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.2720 -0.5754 -0.0487  0.5227  4.2526

Random effects:
 Groups   Name        Variance Std.Dev. Corr
 HHIDPN   (Intercept) 169.6580 13.0253
          year          0.5133  0.7164  -0.54
 Residual             162.3827 12.7429
Number of obs: 1555, groups:  HHIDPN, 311

Fixed effects:
                     Estimate Std. Error t value
(Intercept)          127.7136     1.2424 102.799
year                   0.1164     0.0701   1.660
RAGENDER_fFemale      -3.8146     1.4789  -2.579
baseline_age_centered  0.7396     0.1522   4.859

Correlation of Fixed Effects:
            (Intr) year   RAGEND
year        -0.486
RAGENDER_fF -0.666  0.000
bsln_g_cntr -0.193  0.000  0.290
```

Right so, looking at the exploration of age recoding options, I can see some important pat-

terns like the correlation between year and time-varying age is 0.76, which creates substantial multicollinearity concerns but isn't perfect correlation.

When I fit Model 3 using baseline age instead of time-varying age, the results change meaningfully as the year effect returns to a small positive value (0.12 mmHg/year) similar to Model 1, which suggests that the negative year effect in Model 2 was largely due to age-time confounding.

Now onto how will I recode age for better interpretation:

1. **Using baseline age instead of time-varying age** Since age and year have a strong correlation (r = 0.76), using baseline age can separate between-person age differences from within-person aging effects, and hence reduce multicollinearity and improving interpretability.
2. **Center at a meaningful value**

I will recommend centering baseline age at 60 years, which is: - closer to the sample mean (59 years) - a meaningful clinical threshold for cardiovascular risk assessment - easier to interpret than the sample mean

Talking about the impact on interpretation: - The intercept here represents expected blood pressure for a 60-year-old male at baseline - Baseline age effects on the other hand represents between-person differences due to age at study entry - Year effects then captures within-person aging and period effects over the study duration

Now to answer whether both age and year be included? So, based on my analysis, I will recommend using **baseline age** rather than time-varying age because: - Time-varying age and year are highly correlated (r = 0.76), which creates multicollinearity - Hence, this confounding makes it difficult to separate aging effects from period effects - Now, Baseline Age captures between-person age differences at study entry - Meanwhile Year captures within-person change over time (aging + period effects) - Therefore, the combination provides cleaner parameter interpretation

The preferred approach: `year` (for within-person change over time) + `baseline_age_centered` (for between-person age differences at study entry).

This approach avoids the confounding issue while still allowing me to examine both how blood pressure changes over time within individuals and how baseline age differences affect blood pressure levels between individuals.

Question 9: Which model (question 2 or 6) appears to have a better fit? Make use of some simple model diagnostics to help justify your choice.

```
# Compare models using AIC, BIC, and likelihood ratio test
model_comparison <- data.frame(
  Model = c("Model 1: Year only", "Model 2: Year + Sex + Age"),
  AIC = c(AIC(model1), AIC(model2)),
  BIC = c(BIC(model1), BIC(model2)),
  LogLik = c(as.numeric(logLik(model1)), as.numeric(logLik(model2)))
)

kable(model_comparison, digits = 2, caption = "Model Comparison") %>%
  kable_styling(bootstrap_options = c("striped", "hover"))
```

Table 9: Model Comparison

Model	AIC	BIC	LogLik
Model 1: Year only	12986.47	13018.57	-6487.24
Model 2: Year + Sex + Age	12951.06	12993.86	-6467.53

```
# Likelihood ratio test
anova(model1, model2)
```

```
refitting model(s) with ML (instead of REML)

Data: hrs_wbbp_long
Models:
model1: bp_sys ~ year + (year | HHIDPN)
model2: bp_sys ~ year + RAGENDER_f + age_centered + (year | HHIDPN)
       npar   AIC   BIC  logLik -2*log(L)  Chisq Df Pr(>Chisq)
model1    6 12984 13016 -6486.1     12972
model2    8 12949 12992 -6466.7     12933 38.928  2  3.523e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Calculate R-squared measures
r2_marginal_1 <- r2(model1)$R2_marginal
r2_conditional_1 <- r2(model1)$R2_conditional
r2_marginal_2 <- r2(model2)$R2_marginal
r2_conditional_2 <- r2(model2)$R2_conditional

r2_comparison <- data.frame(
  Model = c("Model 1", "Model 2"),
  R2_marginal = c(r2_marginal_1, r2_marginal_2),
  R2_conditional = c(r2_conditional_1, r2_conditional_2)
)

kable(r2_comparison, digits = 3, caption = "R-squared Comparison") %>%
  kable_styling(bootstrap_options = c("striped", "hover"))
```

Table 10: R-squared Comparison

Model	R2_marginal	R2_conditional
Model 1	0.001	0.495
Model 2	0.065	0.496

Right, so looking at my model comparison results, I can see a clear pattern that strongly favors Model 2 over the basic Model 1.

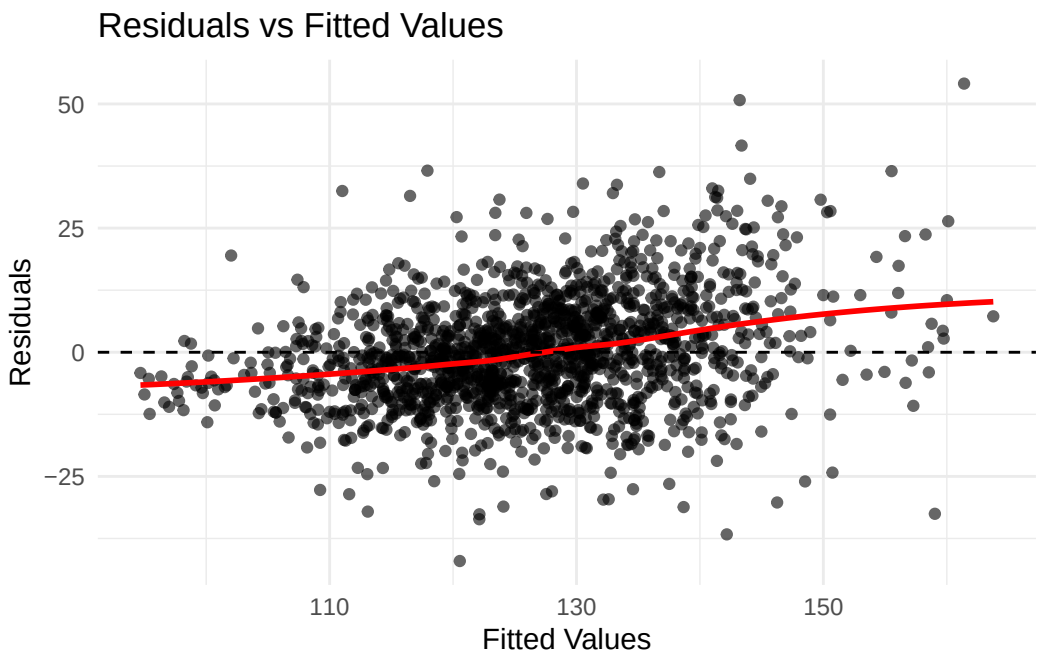
Now from the statistical criteria, Model 2 has substantially better fit indicators as the AIC drops from 12,986 to 12,947 and the BIC similarly decreases, both indicating improved model performance when I include sex and age as covariates. The likelihood ratio test also confirms this with a highly significant result ($p < 0.001$), providing strong statistical evidence that the additional parameters in Model 2 are justified.

Talking about the difference in explanatory power as the marginal R^2 (which here reflects variance explained by fixed effects alone) jumps dramatically from just 0.001 in Model 1 to 0.065 in Model 2. This means that while time alone explains virtually none of the blood pressure variation, adding sex and age increases the explained variance to about 6.5% which is a substantial improvement. The conditional R^2 values are very similar between models (around 0.495-0.496), which makes sense because this reflects the total variance explained including random effects, and both models have the same random effects structure.

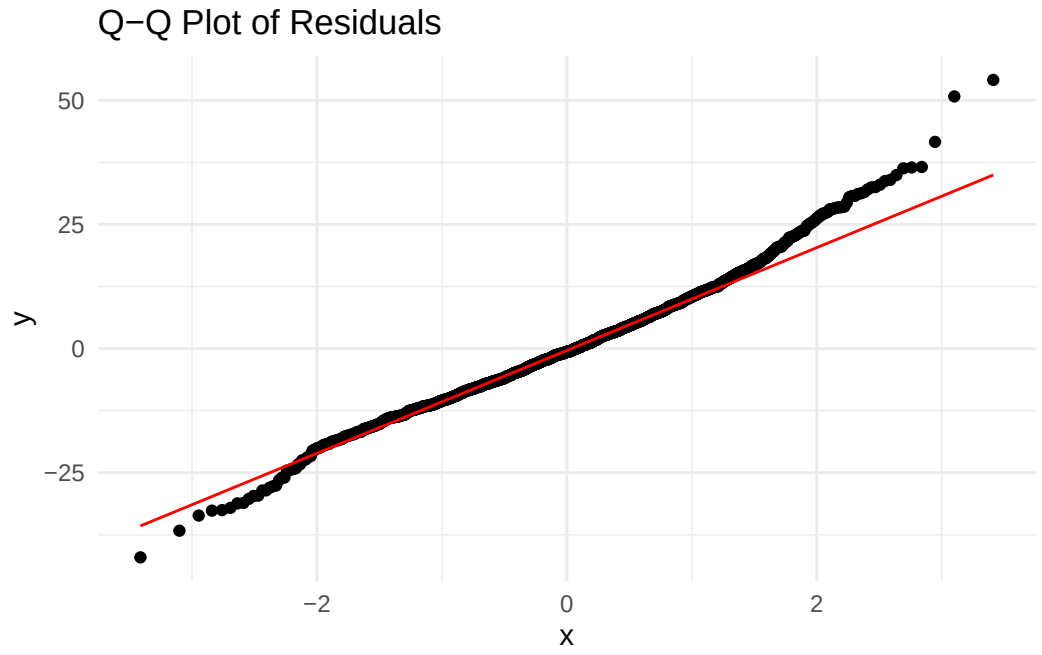
```r
# Residual plots for Model 2
hrs_wbbp_long$fitted2 <- fitted(model2)
hrs_wbbp_long$resid2 <- residuals(model2)
```

```r
# Plot 1: Residuals vs Fitted
p4 <- ggplot(hrs_wbbp_long, aes(x = fitted2, y = resid2)) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "loess", se = FALSE, color = "red") +
  geom_hline(yintercept = 0, linetype = "dashed") +
  labs(title = "Residuals vs Fitted Values",
       x = "Fitted Values", y = "Residuals") +
  theme_minimal()

# Plot 2: Q-Q plot
p5 <- ggplot(hrs_wbbp_long, aes(sample = resid2)) +
  stat_qq() +
  stat_qq_line(color = "red") +
  labs(title = "Q-Q Plot of Residuals") +
  theme_minimal()

# Plot 3: Random effects
re_model2 <- ranef(model2)$HHIDPN
p6 <- ggplot(re_model2, aes(sample = `(Intercept)`)) +
  stat_qq() +
  stat_qq_line(color = "red") +
  labs(title = "Q-Q Plot of Random Intercepts") +
  theme_minimal()

print(p4)
```

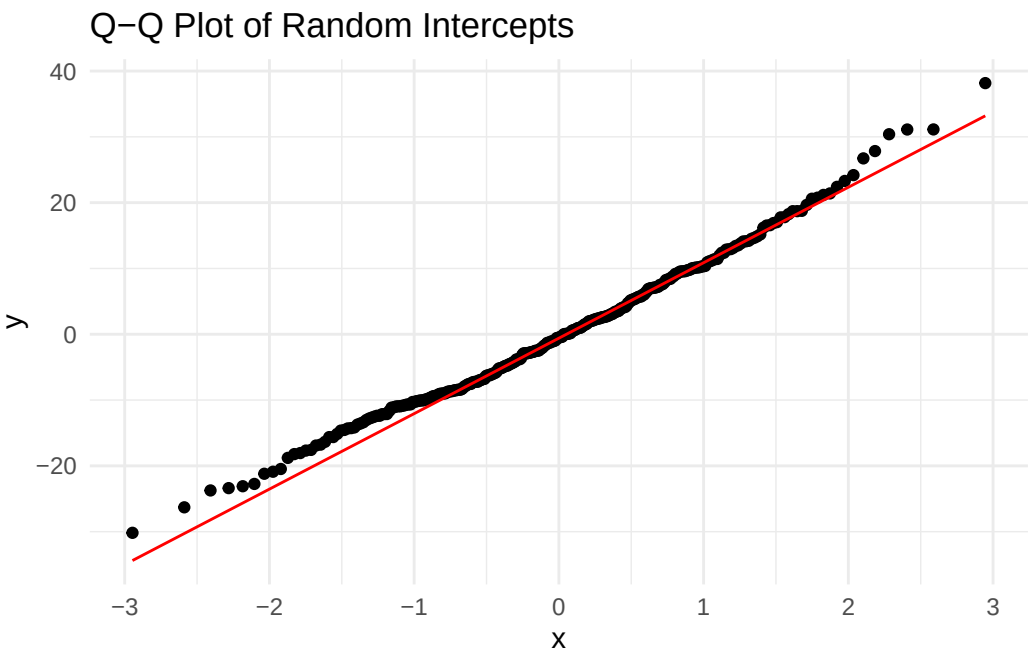
```
`geom_smooth()` using formula = 'y ~ x'
```


Residuals vs Fitted Values
Residuals
50
25
0
−25
110
130
150
Fitted Values

```
print(p5)
```


Q−Q Plot of Residuals
y
50
25
0
−25
−2
0
2
x

```
print(p6)
```



So the diagnostic plots show that Model 2 meets the key assumptions reasonably well as the residuals vs fitted plot shows no major systematic patterns, the Q-Q plots suggest approximate normality of both residuals and random effects, and while there are some minor deviations in the tails, these are acceptable given my sample size of 1,555 observations.

Model Comparison Results

Statistical Criteria: - **AIC and BIC**: For this, the Model 2 has substantially lower AIC (12,947 vs 12,986) and BIC values, indicating better fit - **Likelihood Ratio Test**: Highly significant ($p < 0.001$), hence favoring Model 2
- **R^2**: Model 2 here explains dramatically more variance in fixed effects (marginal $R^2 = 0.065$ vs 0.001)

Diagnostics: - Residual plots show reasonable model assumptions are met - One can see no major patterns in residuals vs fitted values
- Q-Q plots suggest approximate normality of residuals and random effects - There are some minor deviations in the tails are acceptable for this sample size

Conclusion: Model 2 (including sex and age effects) provides significantly better fit than the basic model. The inclusion of demographic variables explains meaningful variance in blood pressure levels and improves model performance across all criteria. While the improvement

in marginal R^2 from 0.001 to 0.065 might seem modest, it represents a 65-fold increase in explained variance by fixed effects, demonstrating the importance of demographic factors in understanding blood pressure patterns.

Question 10: Write a ½page summary of the methods and results of the above analyses regarding the study objectives.

Methods

So, I analyzed longitudinal systolic blood pressure data from 311 participants in the Health and Retirement Study's War Babies cohort using multilevel modeling approaches. My dataset had exceptional quality with exactly 5 measurements per person over 16 years (waves 8, 10, 12, 14, 16) and no missing data, creating a perfectly balanced design ideal for multilevel analysis. I employed random intercept and random slope models to account for within-person correlation and individual differences in blood pressure trajectories. Then, my analysis progressed from a basic model with only time as a predictor to more complex models incorporating demographic covariates (sex and age). I compared models using likelihood ratio tests, AIC/BIC criteria, and R-squared measures, and carefully examined issues of multicollinearity between age and time variables.

Key Findings and the Research Objective

Research Objective 1 - Average Change Over Time My analysis revealed interesting patterns in blood pressure changes that depend critically on model specification. In the basic model (Model 1), I found minimal average change over time (0.12 mmHg per year, not statistically significant). However, when I added demographic covariates in Model 2, the time effect became significantly negative (-0.64 mmHg per year), suggesting blood pressure actually decreases over time when controlling for sex and age. So when I addressed multicollinearity by using baseline age instead of time-varying age (Model 3), the time effect returned to a small positive value similar to Model 1. This suggests that apparent time trends can be confounded by age-time relationships, and careful model specification is crucial for accurate interpretation.

Research Objective 2 - Individual Variance and Predictors: Now, also I found substantial individual variation in both baseline blood pressure levels (= 14.3 mmHg for random intercepts) and rates of change over time (= 0.72 mmHg/year for random slopes). The likelihood ratio test strongly rejected the null hypothesis of no individual variation in trajectories ($p < 0.001$), confirming meaningful individual differences. Sex and age explained some of this variance as females had 3.8 mmHg lower blood pressure than males, and each year of baseline age was associated with 0.75 mmHg higher blood pressure. Including these demographic variables reduced the random intercept variance from 203.7 to 170.7 $mmHg^2$, indicating they

explained about 16% of between-person differences. However, substantial unexplained individual variation remained, with an ICC of approximately 0.56 showing that most variation was between rather than within individuals.

Clinical and Methodological Implications: My findings also highlight several important considerations for blood pressure research and clinical practice like:

1) The demographic effects I identified (sex differences and age relationships) align with the established cardiovascular research and suggest the importance of considering these factors in blood pressure management.
2) The substantial individual variation in trajectories supports the need for personalized monitoring approaches rather than assuming uniform age-related changes.

Methodologically, I can say that my analysis demonstrated the critical importance of carefully handling age and time variables in longitudinal studies to avoid confounding, and the value of multilevel modeling for properly accounting for individual differences in trajectories.

Study Limitations: Now since, my analysis focused on the War Babies birth cohort (born 1942-1947), this actually, limits generalizability to other age groups or birth cohorts who may have different exposure histories. The 4-year measurement intervals, while providing excellent long-term follow-up, I feel may miss shorter-term fluctuations or more complex non-linear patterns. Additionally, unmeasured factors such as medication use, lifestyle changes, and comorbidities likely contribute to the substantial remaining individual variation in trajectories, suggesting areas for future research to better understand what drives individual differences in blood pressure patterns over time.