BIOST540HW1

Bryan Ng, 2427348

2025-04-17

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.2     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(reshape2)
```

```
##
## Attaching package: 'reshape2'
##
## The following object is masked from 'package:tidyr':
##
##     smiths
```

Problem 1

```
aug <- read.csv("C:/Users/ncwbr/Desktop/augmentation.csv")
aug <- aug[, c("id", "Treatment_Group", "HD_t0", "HD_t1", "HD_t2",
"HD_t3", "HD_t4", "HD_t5", "HD_t6")]
aug_long <- melt(aug, id=c("id","Treatment_Group"))
aug_long$week <- as.numeric(gsub("HD_t","",aug_long$variable))
```

(a)

We summarized the distribution of HDRS scores at baseline and weeks 1–6 in both groups by computing the mean, standard deviation, and median.

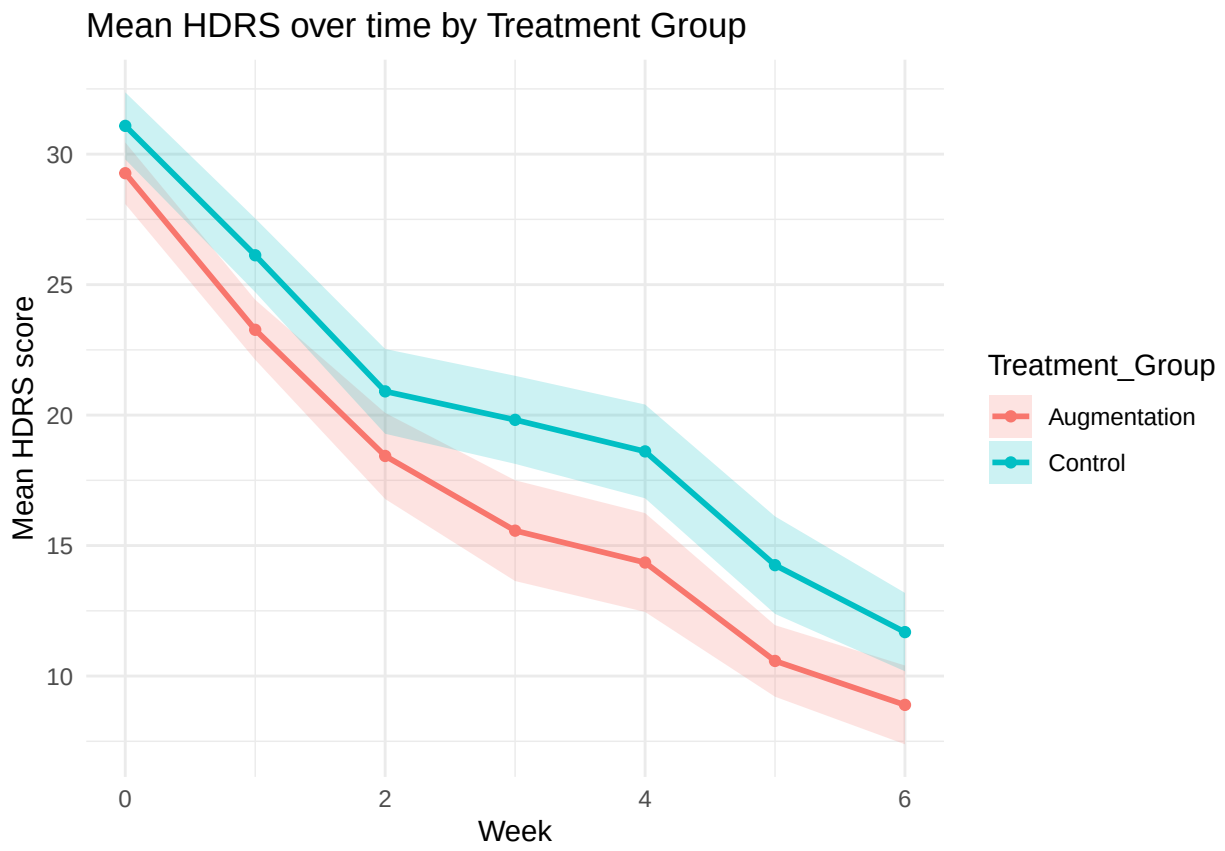
```r
hdrs_summary <- aug_long %>%
  group_by(Treatment_Group, week) %>%
  summarise(
    mean_HDRS = mean(value, na.rm=TRUE),
    sd_HDRS = sd(value,   na.rm=TRUE),
    median_HDRS = median(value, na.rm=TRUE),
    n = sum(!is.na(value))
  )
```

```
## `summarise()` has grouped output by 'Treatment_Group'. You can override using
## the `.groups` argument.
```

```r
hdrs_summary
```

```
## # A tibble: 14 x 6
## # Groups:   Treatment_Group [2]
##    Treatment_Group  week mean_HDRS sd_HDRS median_HDRS     n
##    <chr>           <dbl>     <dbl>   <dbl>       <dbl> <int>
##  1 Augmentation        0      29.3    6.02        28.5    26
##  2 Augmentation        1      23.3    5.86        22      26
##  3 Augmentation        2      18.4    7.91        17      23
##  4 Augmentation        3      15.6    8.85        14      21
##  5 Augmentation        4      14.4    8.47        13      20
##  6 Augmentation        5      10.6    5.98        10      19
##  7 Augmentation        6       8.89   6.58         8      19
##  8 Control             0      31.1    6.30        30      24
##  9 Control             1      26.1    6.92        26      24
## 10 Control             2      20.9    7.62        20.5    22
## 11 Control             3      19.8    7.91        18.5    22
## 12 Control             4      18.6    8.62        17      23
## 13 Control             5      14.2    8.35        13.5    20
## 14 Control             6      11.7    6.55        10      19
```

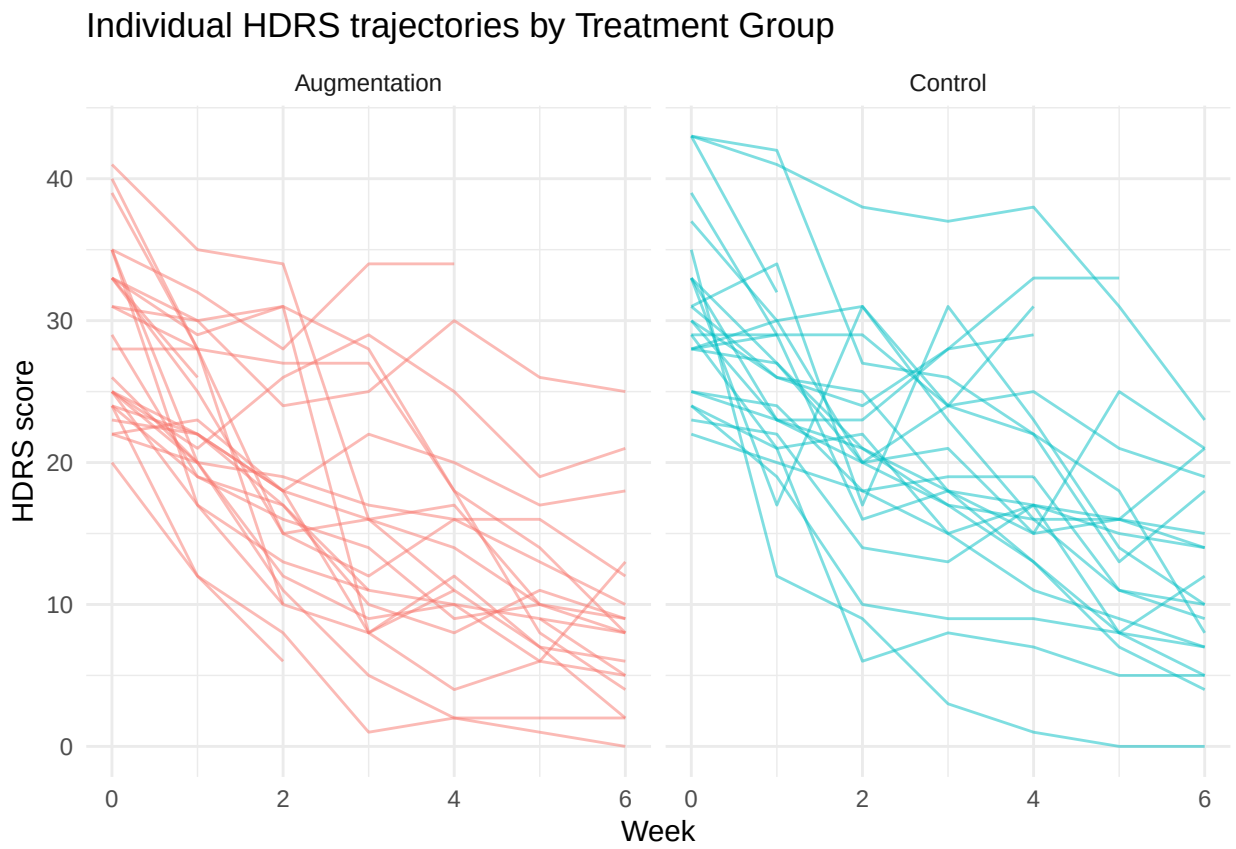
```r
hdrs_summary %>%
  mutate(se = sd_HDRS / sqrt(n)) %>%
  ggplot(aes(x = week, y = mean_HDRS, color = Treatment_Group)) +
  geom_line(linewidth = 1) +
  geom_point() +
  geom_ribbon(aes(ymin = mean_HDRS - se, ymax = mean_HDRS + se,
                  fill = Treatment_Group),
              alpha = 0.2, colour = NA) +
  labs(
    x = "Week",
    y = "Mean HDRS score",
    title = "Mean HDRS over time by Treatment Group"
  ) +
  theme_minimal()
```



We observed a decreasing trend in HDRS scores over time in both the augmentation and control groups, with the control group's mean slightly higher than that of the augmentation group.

(b)

```
aug_long %>%
  ggplot(aes(x = week, y = value, group = id, color = Treatment_Group)) +
  geom_line(alpha = 0.5) +
  facet_wrap(~ Treatment_Group) +
  labs(
    x = "Week",
    y = "HDRS score",
    title = "Individual HDRS trajectories by Treatment Group"
  ) +
  theme_minimal() +
  theme(legend.position = "none")
```



In the augmentation group, the lines become tighter than those in the control group over time, indicating substantially less variability in HDRS scores among augmentation group.

(c)

```
hdrs_wide <- aug_long %>%
  select(id, Treatment_Group, week, value) %>%
  pivot_wider(names_from = week, values_from = value, names_prefix = "W")

cor_overall <- hdrs_wide %>%
  select(starts_with("W")) %>%
  cor(use = "pairwise.complete.obs")
print(cor_overall)
```

```
##            W0        W1        W2        W3        W4         W5         W6
## W0 1.00000000 0.6447416 0.3558085 0.2166642 0.1563784 0.09970292 0.06143658
## W1 0.64474159 1.0000000 0.6748607 0.6381052 0.4910270 0.48272785 0.37953236
## W2 0.35580847 0.6748607 1.0000000 0.7199093 0.6249584 0.51444805 0.43006761
## W3 0.21666418 0.6381052 0.7199093 1.0000000 0.9005854 0.73887416 0.62437769
## W4 0.15637836 0.4910270 0.6249584 0.9005854 1.0000000 0.87663075 0.75796870
## W5 0.09970292 0.4827279 0.5144480 0.7388742 0.8766308 1.00000000 0.87672210
## W6 0.06143658 0.3795324 0.4300676 0.6243777 0.7579687 0.87672210 1.00000000
```

```r
cor_by_group <- hdrs_wide %>%
  group_by(Treatment_Group) %>%
  summarise(
    cor_mat = list(
      cor(
        across(starts_with("W")),
        use = "pairwise.complete.obs"
      )
    ),
    .groups = "drop"
  )
cor_by_group$cor_mat
```

```
## [[1]]
##             W0         W1        W2         W3         W4          W5
## W0  1.00000000 0.68092743 0.3228371 0.02998406 0.02511078 -0.05327108
## W1  0.68092743 1.00000000 0.7811287 0.52414083 0.47051107  0.35893453
## W2  0.32283713 0.78112868 1.0000000 0.66727887 0.45809453  0.27729528
## W3  0.02998406 0.52414083 0.6672789 1.00000000 0.89402917  0.71516192
## W4  0.02511078 0.47051107 0.4580945 0.89402917 1.00000000  0.90094026
## W5 -0.05327108 0.35893453 0.2772953 0.71516192 0.90094026  1.00000000
## W6 -0.14917525 0.06768359 0.1333135 0.60206817 0.81456297  0.89662223
##             W6
## W0 -0.14917525
## W1  0.06768359
## W2  0.13331354
## W3  0.60206817
## W4  0.81456297
## W5  0.89662223
## W6  1.00000000
##
## [[2]]
##           W0        W1        W2        W3        W4        W5        W6
## W0 1.0000000 0.5956827 0.3667421 0.3996550 0.2339445 0.1166280 0.1970558
## W1 0.5956827 1.0000000 0.5612090 0.7336095 0.4814409 0.4896895 0.5639022
## W2 0.3667421 0.5612090 1.0000000 0.7813042 0.7702279 0.6822212 0.7151093
## W3 0.3996550 0.7336095 0.7813042 1.0000000 0.9015276 0.7399557 0.6049846
## W4 0.2339445 0.4814409 0.7702279 0.9015276 1.0000000 0.8557903 0.6874626
## W5 0.1166280 0.4896895 0.6822212 0.7399557 0.8557903 1.0000000 0.8586438
## W6 0.1970558 0.5639022 0.7151093 0.6049846 0.6874626 0.8586438 1.0000000
```

We found that the correlation between baseline and week t decreases over time($\rho_{05} = 0.099$, $\rho_{06} = 0.06$), correlations between follow-up weeks remain relatively high overall and within the augmentation group, while the control group's correlations show less variability over time.

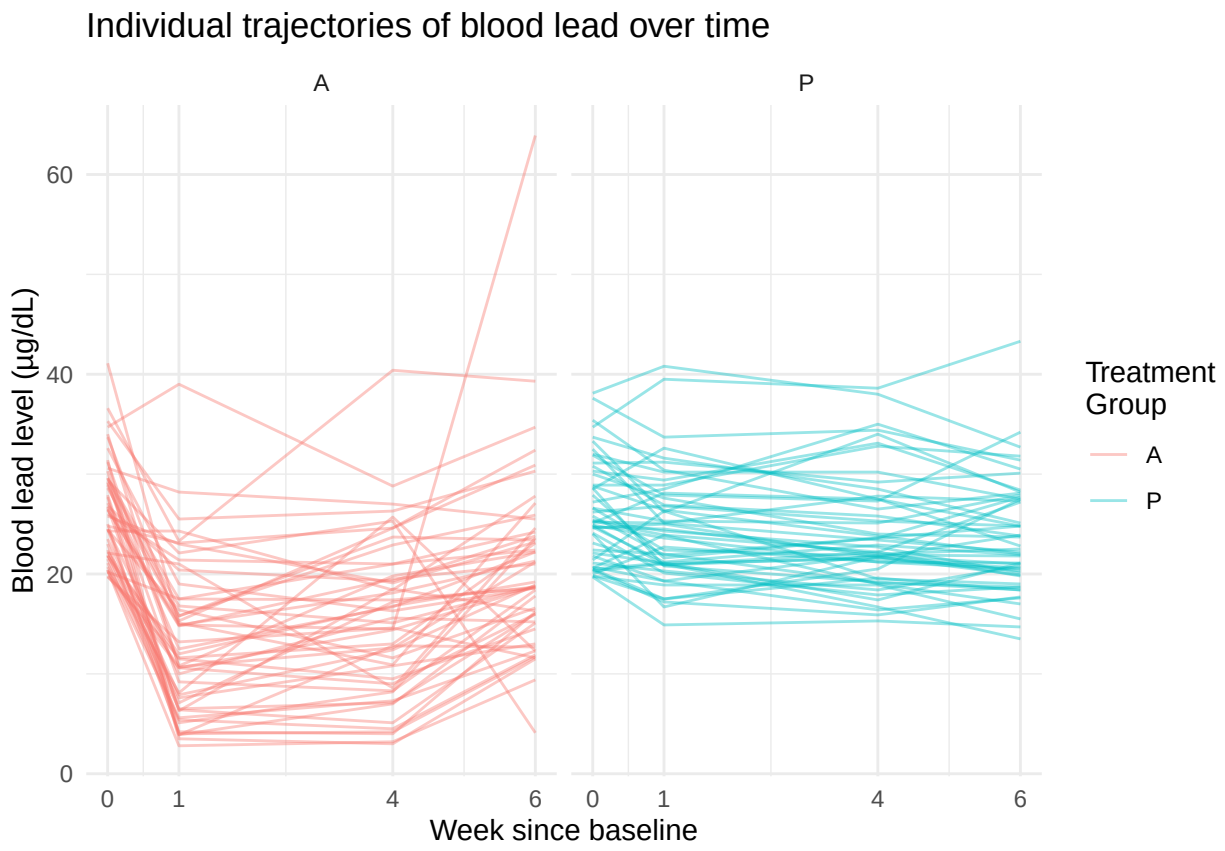
Problem 2

```r
tlc <- read.csv("C:/Users/ncwbr/Desktop/tlc.csv")
```

(a)

```
tlc_long <- tlc %>%
  pivot_longer(
    cols = starts_with("y"),
    names_to = "week",
    names_prefix = "y",
    values_to = "lead_level"
  ) %>%
  mutate(
    week = as.numeric(week)
  )

ggplot(tlc_long, aes(x = week, y = lead_level, group = id, color = tx)) +
  geom_line(alpha = 0.4) +
  facet_wrap(~ tx) +
  scale_x_continuous(breaks = c(0,1,4,6)) +
  labs(
    x = "Week since baseline",
    y = "Blood lead level (µg/dL)",
    title = "Individual trajectories of blood lead over time",
    color = "Treatment\nGroup"
  ) +
  theme_minimal()
```



(b)

```
lead_wide <- tlc %>% select(id, tx, y0, y1, y4, y6)

cor_overall <- cor(lead_wide %>% select(-id, -tx),
                   use = "pairwise.complete.obs")
print(cor_overall)
```

```
##           y0        y1        y4        y6
## y0 1.0000000 0.4188669 0.4681009 0.5617933
## y1 0.4188669 1.0000000 0.8446982 0.5572482
## y4 0.4681009 0.8446982 1.0000000 0.5826476
## y6 0.5617933 0.5572482 0.5826476 1.0000000
```

```
cor_by_tx <- lead_wide %>%
  group_by(tx) %>%
  summarise(
    cor_mat = list(
      cor(across(y0:y6), use = "pairwise.complete.obs")
    )
  )

cor_by_tx$cor_mat
```

```
## [[1]]
##           y0        y1        y4        y6
## y0 1.0000000 0.4014589 0.3839654 0.4951063
## y1 0.4014589 1.0000000 0.7308221 0.5069743
## y4 0.3839654 0.7308221 1.0000000 0.4548224
## y6 0.4951063 0.5069743 0.4548224 1.0000000
##
## [[2]]
##           y0        y1        y4        y6
## y0 1.0000000 0.8291362 0.8393547 0.7558796
## y1 0.8291362 1.0000000 0.8606685 0.7592246
## y4 0.8393547 0.8606685 1.0000000 0.8697065
## y6 0.7558796 0.7592246 0.8697065 1.0000000
```

(c)

```
tlc$tx <- factor(tlc$tx, levels = c("P", "A"))
mod_w6 <- lm(y6 ~ tx, data = tlc)
summary(mod_w6)
```

```
##
## Call:
## lm(formula = y6 ~ tx, data = tlc)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -16.662  -4.675  -1.604   3.700  43.138
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   23.646      1.083  21.833   <2e-16 ***
## txA           -2.884      1.532  -1.883   0.0627 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.658 on 98 degrees of freedom
## Multiple R-squared:  0.03491,    Adjusted R-squared:  0.02507
## F-statistic: 3.545 on 1 and 98 DF,  p-value: 0.06268
```

Using a linear model at the six-week time point, we estimated the difference in week 6 blood lead levels between group A (succimer) and group P (placebo) as txA = -2.884, the standard deviation is 1.532 and p-value is equal to 0.0627. At significance level $\alpha = 0.05$, we can conclude that treatment does not have a significant effect on the Week 6 lead levels.

(d)

```
mod_d <- lm(y6 - y0 ~ tx, data = tlc)
sum_d <- summary(mod_d)$coefficients
sum_d
```

```
##             Estimate Std. Error   t value    Pr(>|t|)
## (Intercept)   -2.626  0.8885211 -2.955473 0.003910403
## txA           -3.152  1.2565586 -2.508438 0.013768101
```

The estimated difference txA = -3.152 with the standard deviation is equal to 1.257, the p-value is 0.013, which is less than $\alpha = 0.05$, indicating a significant treatment effect on week 6 lead levels.

(e)

```
mod_e <- lm(y6 ~ tx + y0, data = tlc)
sum_e <- summary(mod_e)$coefficients
sum_e
```

```
##               Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)  0.5235145  3.4384107  0.1522548 8.793022e-01
## txA         -3.1198719  1.2576624 -2.4806911 1.483474e-02
## y0           0.8801190  0.1264275  6.9614518 4.007088e-10
```

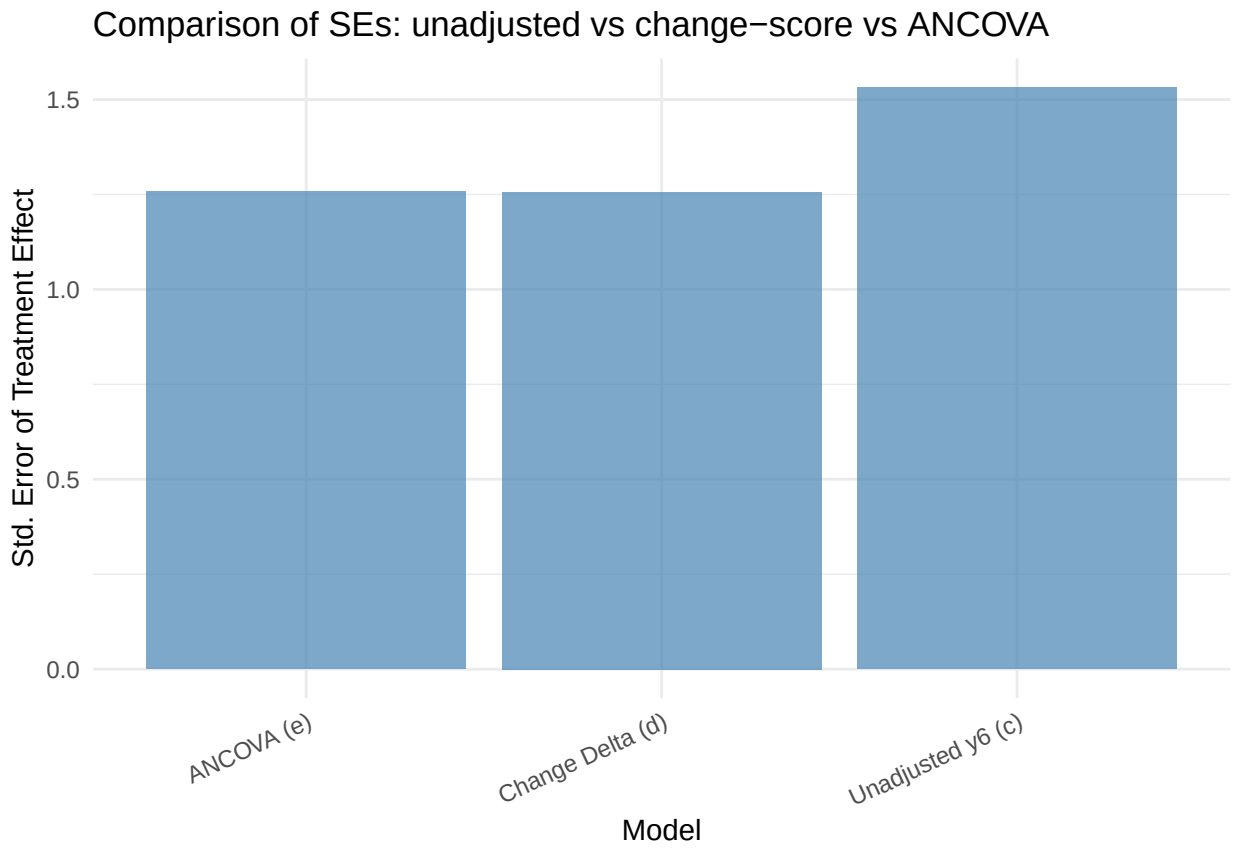
Applying an ANCOVA model, we derive that he estimated difference txA = -3.120 with the standard deviation is equal to 1.258, the p-value is 0.0148, since p-value is less than 0.05, the treatment effect on week 6 lead levels is significant.

(f)

```r
se_c <- summary(mod_w6)$coefficients["txA", "Std. Error"]
se_d <- sum_d   ["txA", "Std. Error"]
se_e <- sum_e   ["txA", "Std. Error"]

se_df <- tibble(
  model = c("Unadjusted y6 (c)", "Change Delta (d)", "ANCOVA (e)"),
  se    = c(se_c, se_d, se_e)
)

ggplot(se_df, aes(x = model, y = se)) +
  geom_col(fill = "steelblue", alpha = 0.7) +
  labs(
    x = "Model",
    y = "Std. Error of Treatment Effect",
    title = "Comparison of SEs: unadjusted vs change-score vs ANCOVA"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 25, hjust = 1))
```



Adjusting for baseline—whether via change scores or ANCOVA—accounts for baseline variability in the residual error and thus reduces the standard error of the treatment effect.