phase6b_ancova_modeling.R

abdulbasir

2025-12-09

```
# phase 6b: ancova modeling with expedited program covariates
#
# disentangling therapeutic area effect from expedited program usage
# to determine if oncology's speed advantage is due to expedited programs or
# an independent regulatory effect

# sourcing configuration and utilities
source("config.R")
source("utils.R")

# loading required libraries
suppressPackageStartupMessages({
  library(dplyr)
  library(readr)
  library(car)  # for Type III SS (Marginal) on unbalanced designs
  library(ggplot2)
})

print_section_header("Phase 6b: ANCOVA Modeling With Expedited Program Covariates")
```

```
##
## ========================================================================
## Phase 6b: ANCOVA Modeling With Expedited Program Covariates
## ========================================================================
```

```
# 1. loading analysis-ready data from phase 5
input_file = file.path(RESULTS_DIR, "analysis_ready_dataset.csv")
analysis_data = load_csv(input_file)

cat(paste("Analysis-ready sample: n =", nrow(analysis_data), "\n"))
```

```
## Analysis-ready sample: n = 1038
```

```
# 2. creating binary covariates for expedited programs
analysis_data = analysis_data %>%
  mutate(
    accelerated_approval_binary = if_else(`Accelerated Approval` == "Yes", 1, 0),
    fast_track_binary = if_else(`Fast Track Designation` == "Yes", 1, 0),
    orphan_binary = if_else(`Orphan Drug Designation` %in% c("Yes", "yes"), 1, 0)
  )
```

```r
cat("\nExpedited program prevalence:\n")
```

```
##
## Expedited program prevalence:
```

```r
cat(sprintf("  Accelerated Approval: %d (%.1f%%)\n",
            sum(analysis_data$accelerated_approval_binary),
            mean(analysis_data$accelerated_approval_binary) * 100))
```

```
##   Accelerated Approval: 137 (13.2%)
```

```r
cat(sprintf("  Fast Track: %d (%.1f%%)\n",
            sum(analysis_data$fast_track_binary),
            mean(analysis_data$fast_track_binary) * 100))
```

```
##   Fast Track: 223 (21.5%)
```

```r
cat(sprintf("  Orphan Drug: %d (%.1f%%)\n",
            sum(analysis_data$orphan_binary),
            mean(analysis_data$orphan_binary) * 100))
```

```
##   Orphan Drug: 438 (42.2%)
```

```r
# 3. baseline model (from Phase 6 Model 3) - no covariates
model_baseline = lm(
  log_review_time_days_response ~ therapeutic_area_factor * review_type_factor + regulatory_era_factor,
  data = analysis_data
)

anova_baseline = car::Anova(model_baseline, type = 3)

cat("\nBaseline model (no covariates):\n")
```

```
##
## Baseline model (no covariates):
```

```r
print(anova_baseline)
```

```
## Anova Table (Type III tests)
##
## Response: log_review_time_days_response
##                                          Sum Sq   Df    F value    Pr(>F)
## (Intercept)                              3724.7    1 11171.9454 < 2.2e-16 ***
## therapeutic_area_factor                     6.2    1    18.6996 1.679e-05 ***
## review_type_factor                         25.8    1    77.5170 < 2.2e-16 ***
## regulatory_era_factor                      88.7    3    88.6370 < 2.2e-16 ***
## therapeutic_area_factor:review_type_factor  0.8    1     2.4333    0.1191
## Residuals                                 343.7 1031
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# extracting baseline therapeutic area F-statistic
f_area_baseline = anova_baseline["therapeutic_area_factor", "F value"]
p_area_baseline = anova_baseline["therapeutic_area_factor", "Pr(>F)"]
f_interaction_baseline = anova_baseline["therapeutic_area_factor:review_type_factor", "F value"]
p_interaction_baseline = anova_baseline["therapeutic_area_factor:review_type_factor", "Pr(>F)"]

cat("\nBaseline key statistics:\n")
```

```
##
## Baseline key statistics:
```

```r
cat(sprintf("  Therapeutic area: F=%.2f, p=%.2e\n", f_area_baseline, p_area_baseline))
```

```
##   Therapeutic area: F=18.70, p=1.68e-05
```

```r
cat(sprintf("  Interaction: F=%.2f, p=%.2e\n", f_interaction_baseline, p_interaction_baseline))
```

```
##   Interaction: F=2.43, p=1.19e-01
```

```r
# 4. ancova model - with expedited program covariates
model_ancova = lm(
  log_review_time_days_response ~
    therapeutic_area_factor * review_type_factor +
    regulatory_era_factor +
    accelerated_approval_binary +
    fast_track_binary +
    orphan_binary,
  data = analysis_data
)

anova_ancova = car::Anova(model_ancova, type = 3)

cat("\nANCOVA model (with expedited program covariates):\n")
```

```
##
## ANCOVA model (with expedited program covariates):
```

```r
print(anova_ancova)
```

```
## Anova Table (Type III tests)
##
## Response: log_review_time_days_response
##                                          Sum Sq   Df   F value    Pr(>F)
## (Intercept)                              3170.9    1 9655.5949 < 2.2e-16 ***
## therapeutic_area_factor                     3.4    1   10.2771 0.0013885 **
## review_type_factor                         19.0    1   57.9311 6.126e-14 ***
## regulatory_era_factor                      65.1    3   66.0967 < 2.2e-16 ***
## accelerated_approval_binary                 3.9    1   11.9171 0.0005788 ***
## fast_track_binary                           1.7    1    5.3132 0.0213629 *
## orphan_binary                               0.2    1    0.5176 0.4720204
```

```
## therapeutic_area_factor:review_type_factor    0.4    1    1.1269 0.2886850
## Residuals                                   337.6 1028
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# extracting ancova therapeutic area F-statistic
f_area_ancova = anova_ancova["therapeutic_area_factor", "F value"]
p_area_ancova = anova_ancova["therapeutic_area_factor", "Pr(>F)"]
f_interaction_ancova = anova_ancova["therapeutic_area_factor:review_type_factor", "F value"]
p_interaction_ancova = anova_ancova["therapeutic_area_factor:review_type_factor", "Pr(>F)"]

# extracting covariate effects
f_acc_approval = anova_ancova["accelerated_approval_binary", "F value"]
p_acc_approval = anova_ancova["accelerated_approval_binary", "Pr(>F)"]
f_fast_track = anova_ancova["fast_track_binary", "F value"]
p_fast_track = anova_ancova["fast_track_binary", "Pr(>F)"]
f_orphan = anova_ancova["orphan_binary", "F value"]
p_orphan = anova_ancova["orphan_binary", "Pr(>F)"]

cat("\nANCOVA key statistics:\n")
```

```
##
## ANCOVA key statistics:
```

```
cat(sprintf("  Therapeutic area: F=%.2f, p=%.2e\n", f_area_ancova, p_area_ancova))
```

```
##   Therapeutic area: F=10.28, p=1.39e-03
```

```
cat(sprintf("  Interaction: F=%.2f, p=%.2e\n", f_interaction_ancova, p_interaction_ancova))
```

```
##   Interaction: F=1.13, p=2.89e-01
```

```
cat("\nExpedited program covariate effects:\n")
```

```
##
## Expedited program covariate effects:
```

```
cat(sprintf("  Accelerated Approval: F=%.2f, p=%.2e\n", f_acc_approval, p_acc_approval))
```

```
##   Accelerated Approval: F=11.92, p=5.79e-04
```

```
cat(sprintf("  Fast Track: F=%.2f, p=%.2e\n", f_fast_track, p_fast_track))
```

```
##   Fast Track: F=5.31, p=2.14e-02
```

```
cat(sprintf("  Orphan Drug: F=%.2f, p=%.2e\n", f_orphan, p_orphan))
```

```
##   Orphan Drug: F=0.52, p=4.72e-01
```

```r
# 5. comparison and interpretation
f_area_change = f_area_ancova - f_area_baseline
f_area_pct_change = (f_area_change / f_area_baseline) * 100
f_interaction_change = f_interaction_ancova - f_interaction_baseline
f_interaction_pct_change = (f_interaction_change / f_interaction_baseline) * 100

comparison_table = data.frame(
  effect = c("Therapeutic Area", "Interaction"),
  f_baseline = c(f_area_baseline, f_interaction_baseline),
  f_ancova = c(f_area_ancova, f_interaction_ancova),
  f_change = c(f_area_change, f_interaction_change),
  pct_change = c(f_area_pct_change, f_interaction_pct_change)
)

cat("\nF-statistic comparison:\n")
```

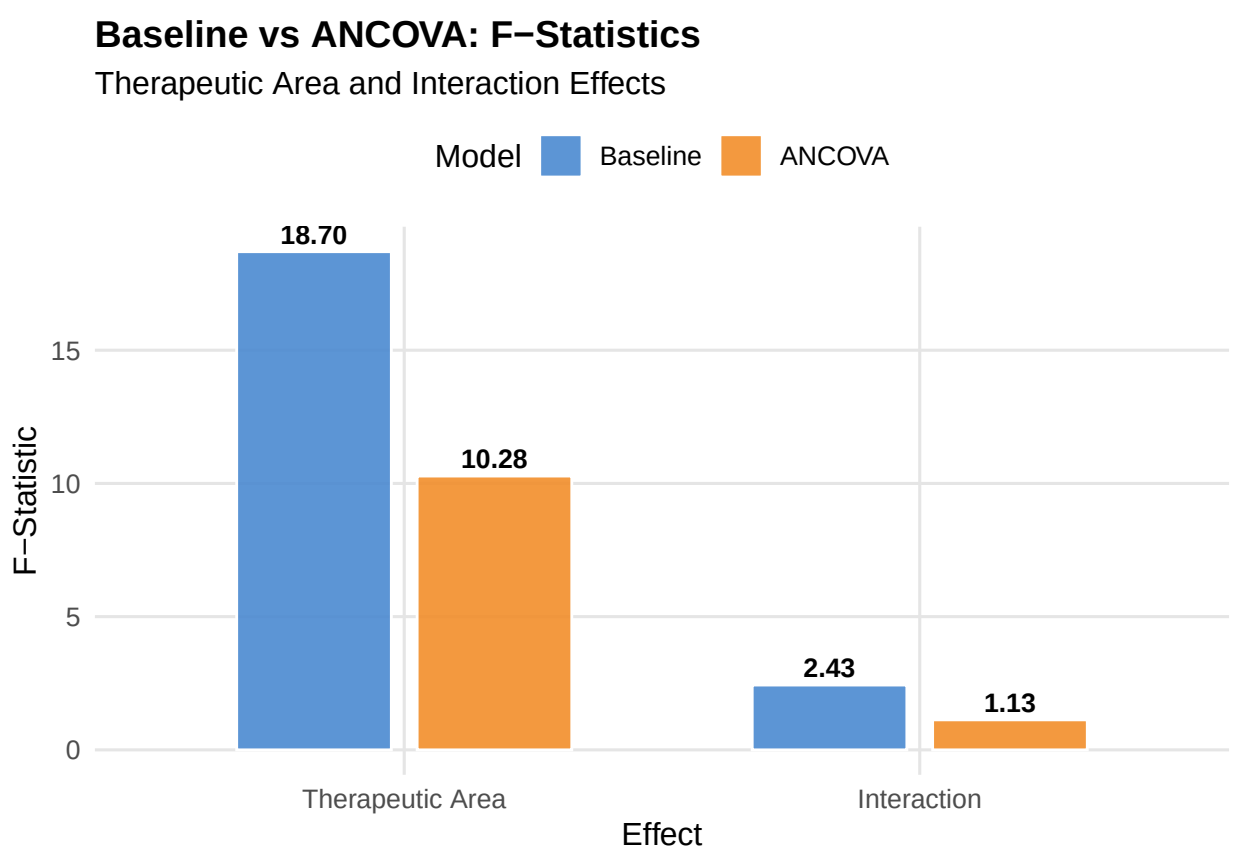
```
##
## F-statistic comparison:
```

```r
print(comparison_table)
```

```
##             effect f_baseline  f_ancova  f_change pct_change
## 1 Therapeutic Area  18.699579 10.277050 -8.422529  -45.04127
## 2      Interaction   2.433274  1.126909 -1.306365  -53.68755
```

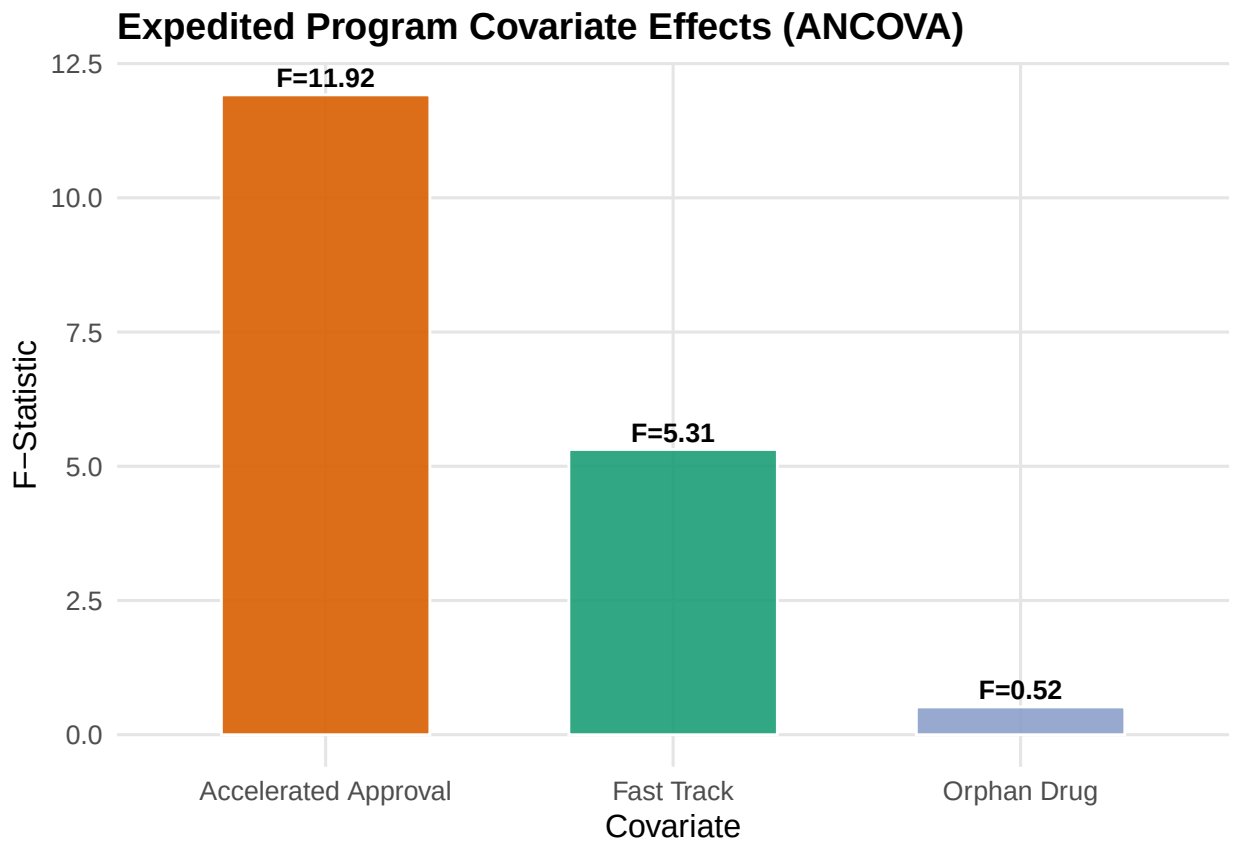
```r
# 6. visualizing baseline vs ancova effects
effect_compare = comparison_table %>%
  mutate(effect = factor(effect, levels = c("Therapeutic Area", "Interaction")))

effect_long = data.frame(
  effect = rep(effect_compare$effect, times = 2),
  model = factor(rep(c("Baseline", "ANCOVA"), each = nrow(effect_compare)), levels = c("Baseline", "ANC
  f_value = c(effect_compare$f_baseline, effect_compare$f_ancova)
)

effect_palette = c("Baseline" = "#4b8ad1", "ANCOVA" = "#f28e2b")

effect_plot = ggplot(effect_long, aes(x = effect, y = f_value, fill = model)) +
  geom_col(position = position_dodge(width = 0.7), width = 0.6, alpha = 0.9, color = "white") +
  geom_text(
    aes(label = sprintf("%.2f", f_value)),
    position = position_dodge(width = 0.7),
    vjust = -0.4,
    size = 3.5,
    fontface = "bold"
  ) +
  scale_fill_manual(values = effect_palette) +
  labs(
    title = "Baseline vs ANCOVA: F-Statistics",
    subtitle = "Therapeutic Area and Interaction Effects",
    x = "Effect",
    y = "F-Statistic",
```

```
    fill = "Model"
  ) +
  theme_minimal(base_size = 12) +
  theme(
    plot.title = element_text(face = "bold", size = 14),
    plot.background = element_rect(fill = "white", color = NA),
    panel.background = element_rect(fill = "white", color = NA),
    panel.grid.major = element_line(color = "gray90"),
    panel.grid.minor = element_blank(),
    legend.position = "top",
    axis.text = element_text(size = 10)
  )

print(effect_plot)
```




```
ggsave(
  file.path(FIGURES_DIR, "ancova_effects_baseline_vs_ancova.png"),
  plot = effect_plot,
  width = 8,
  height = 6,
  dpi = DPI
)

# 7. visualizing expedited covariate effects
covariate_effects = data.frame(
```

```
  program = factor(c("Accelerated Approval", "Fast Track", "Orphan Drug"),
                   levels = c("Accelerated Approval", "Fast Track", "Orphan Drug")),
  f_value = c(f_acc_approval, f_fast_track, f_orphan)
) %>%
  mutate(label = sprintf("F=%.2f", f_value))

covar_palette = c(
  "Accelerated Approval" = "#d95f02",
  "Fast Track" = "#1b9e77",
  "Orphan Drug" = "#8da0cb"
)

covar_plot = ggplot(covariate_effects, aes(x = program, y = f_value, fill = program)) +
  geom_col(width = 0.6, alpha = 0.9, color = "white") +
  geom_text(
    aes(label = label),
    vjust = -0.4,
    size = 3.5,
    fontface = "bold"
  ) +
  scale_fill_manual(values = covar_palette) +
  labs(
    title = "Expedited Program Covariate Effects (ANCOVA)",
    x = "Covariate",
    y = "F-Statistic",
    fill = "Covariate"
  ) +
  theme_minimal(base_size = 12) +
  theme(
    plot.title = element_text(face = "bold", size = 14),
    plot.background = element_rect(fill = "white", color = NA),
    panel.background = element_rect(fill = "white", color = NA),
    panel.grid.major = element_line(color = "gray90"),
    panel.grid.minor = element_blank(),
    legend.position = "none",
    axis.text = element_text(size = 10)
  )

print(covar_plot)
```




```
ggsave(
  file.path(FIGURES_DIR, "ancova_expedited_covariate_effects.png"),
  plot = covar_plot,
  width = 8,
  height = 6,
  dpi = DPI
)

cat("\nInterpretation:\n")
```

```
##
## Interpretation:
```

```
if (f_area_ancova < 10) {
  cat(sprintf("Therapeutic area F-statistic dropped from %.2f to %.2f (%.1f%% reduction)\n",
              f_area_baseline, f_area_ancova, abs(f_area_pct_change)))
  cat("→ Expedited programs explain most of therapeutic area effect\n")
  cat("  Oncology's speed advantage is primarily due to greater utilization of expedited pathways\n")
} else if (f_area_ancova > 30) {
  cat(sprintf("Therapeutic area F-statistic: %.2f → %.2f (%.1f%% change)\n",
              f_area_baseline, f_area_ancova, f_area_pct_change))
  cat("→ Therapeutic area has independent effect beyond expedited programs\n")
  cat("  Even after accounting for expedited pathways, oncology still has significantly faster review t
} else {
  cat(sprintf("Therapeutic area F-statistic: %.2f → %.2f (%.1f%% change)\n",
```

```
              f_area_baseline, f_area_ancova, f_area_pct_change))
  cat("→ Partial explanation by expedited programs\n")
  cat("  Expedited programs account for some, but not all, of oncology's speed advantage\n")
}
```

```
## Therapeutic area F-statistic: 18.70 → 10.28 (-45.0% change)
## → Partial explanation by expedited programs
##   Expedited programs account for some, but not all, of oncology's speed advantage
```

```
if (abs(f_interaction_pct_change) > 50) {
  cat(sprintf("\nInteraction F-statistic: %.2f → %.2f (%.1f%% change)\n",
              f_interaction_baseline, f_interaction_ancova, f_interaction_pct_change))
  cat("→ Differential Priority Review benefit substantially affected by expedited program covariates\n")
} else {
  cat(sprintf("\nInteraction F-statistic: %.2f → %.2f (%.1f%% change)\n",
              f_interaction_baseline, f_interaction_ancova, f_interaction_pct_change))
  cat("→ Interaction effect persists after adjusting for expedited programs\n")
}
```

```
##
## Interaction F-statistic: 2.43 → 1.13 (-53.7% change)
## → Differential Priority Review benefit substantially affected by expedited program covariates
```

```
# 8. expedited program contribution ranking
covariate_effects = data.frame(
  program = c("Accelerated Approval", "Fast Track", "Orphan Drug"),
  f_value = c(f_acc_approval, f_fast_track, f_orphan),
  p_value = c(p_acc_approval, p_fast_track, p_orphan)
) %>%
  mutate(significant = p_value < ALPHA) %>%
  arrange(desc(f_value))

cat("\nExpedited program effects (ranked by F-statistic):\n")
```

```
##
## Expedited program effects (ranked by F-statistic):
```

```
print(covariate_effects)
```

```
##                program    f_value      p_value significant
## 1 Accelerated Approval 11.9171457 0.0005788336        TRUE
## 2           Fast Track  5.3132004 0.0213629085        TRUE
## 3          Orphan Drug  0.5176243 0.4720203834       FALSE
```

```
strongest_program = covariate_effects$program[1]
cat(sprintf("\nStrongest contributor: %s (F=%.2f)\n", strongest_program, covariate_effects$f_value[1]))
```

```
##
## Strongest contributor: Accelerated Approval (F=11.92)
```

```r
# 9. saving results
output_file = file.path(RESULTS_DIR, "ancova_comparison.csv")
save_csv(comparison_table, output_file)
```

```
## saving results to: /Users/abdulbasir/Downloads/Experimental AI/fda-oncology-approval-analysis/results
```

```r
cat("\nPhase 6b complete\n")
```

```
##
## Phase 6b complete
```