Employment example

Jouni Helske and Santtu Tikka

2024-02-01

Load some packages and create the data:

```r
library(dplyr)
library(march)
library(dynamite)
library(ggplot2)
library(RColorBrewer)
N <- Employment.2@N
#T <- Employment.2@T[1]
d <- data.frame(
  employment = factor(c(Employment.2@yRaw), labels = c("Full-time", "Other")),
  gender = factor(Employment.2@cov[,,1], labels = c("Woman", "Man")),
  id = 1:N, age = rep(seq(20, 44, by = 2), each = N)
) |>
  mutate(fulltime = as.integer(employment == "Full-time"))
```

Define the model in dynamite:

```r
set.seed(1)
model_formula <-
  obs(
    fulltime ~ -1 + gender:lag(ever) + varying(~ -1 + gender + gender:lag(fulltime)),
    family = "bernoulli"
  ) +
  aux(numeric(ever) ~ fulltime == 1 | lag(ever) == 1 | init(0)) +
  splines(df = 6, noncentered = TRUE)
priors <- get_priors(model_formula, data = d, group = "id", time = "age")
priors$prior[priors$type == "tau"] <- "normal(0, 1)"
fit <- dynamite(
  model_formula, data = d, group = "id", time = "age", priors = priors,
  chains = 4, cores = 4, iter = 5000, refresh = 0,
  save_warmup = FALSE
)
saveRDS(fit, file = "fit_employment.rds")
```

Test different values of D:

```r
set.seed(1)
model_formula <-
  obs(
    fulltime ~ -1 + gender:lag(ever) + varying(~ -1 + gender + gender:lag(fulltime)),
    family = "bernoulli"
  ) +
  aux(numeric(ever) ~ fulltime == 1 | lag(ever) == 1 | init(0)) +
```

```r
  splines(df = 4, noncentered = TRUE)
priors <- get_priors(model_formula, data = d, group = "id", time = "age")
priors$prior[priors$type == "tau"] <- "normal(0, 1)"
fit4 <- dynamite(
  model_formula, data = d, group = "id", time = "age", priors = priors,
  chains = 4, cores = 4, iter = 5000, refresh = 0
)
saveRDS(fit4, file = "fit_employment_D4.rds")

set.seed(1)
model_formula <-
  obs(
    fulltime ~ -1 + gender:lag(ever) + varying(~ -1 + gender + gender:lag(fulltime)),
    family = "bernoulli"
  ) +
  aux(numeric(ever) ~ fulltime == 1 | lag(ever) == 1 | init(0)) +
  splines(df = 8, noncentered = TRUE)
priors <- get_priors(model_formula, data = d, group = "id", time = "age")
priors$prior[priors$type == "tau"] <- "normal(0, 1)"
fit8 <- dynamite(
  model_formula, data = d, group = "id", time = "age", priors = priors,
  chains = 4, cores = 4, iter = 5000, refresh = 0
)
saveRDS(fit8, file = "fit_employment_D8.rds")

set.seed(1)
model_formula <-
  obs(
    fulltime ~ -1 + gender:lag(ever) + varying(~ -1 + gender + gender:lag(fulltime)),
    family = "bernoulli"
  ) +
  aux(numeric(ever) ~ fulltime == 1 | lag(ever) == 1 | init(0)) +
  splines(df = 15, noncentered = TRUE)
priors <- get_priors(model_formula, data = d, group = "id", time = "age")
priors$prior[priors$type == "tau"] <- "normal(0, 1)"
fit15 <- dynamite(
  model_formula, data = d, group = "id", time = "age", priors = priors,
  chains = 4, cores = 4, iter = 5000, refresh = 0
)
saveRDS(fit15, file = "fit_employment_D15.rds")
```

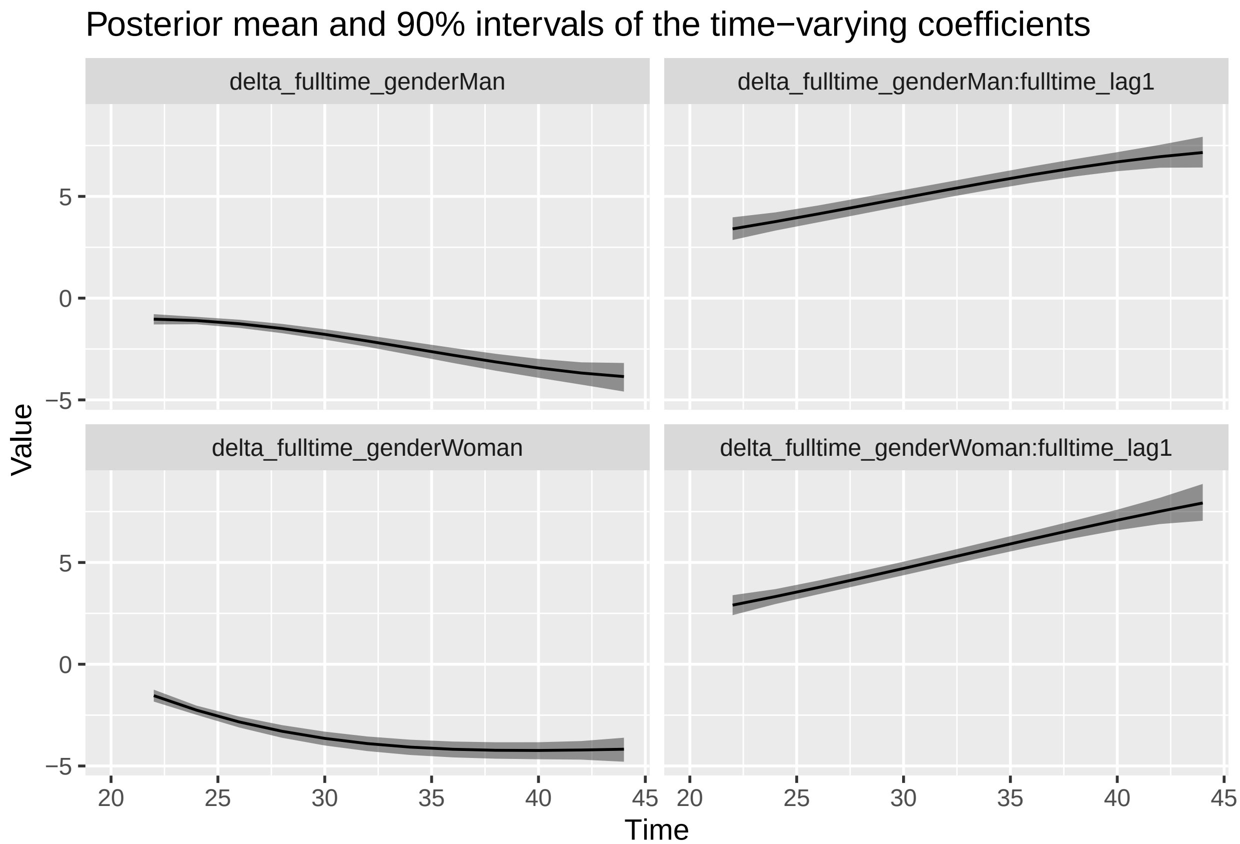
Compare models with different D values using leave-one-out cross-validation:

```r
l6 <- loo(fit)
l4 <- loo(fit4)
l8 <- loo(fit8)
l15 <- loo(fit15)
loo::loo_compare(l4, l6, l8, l15)
```

```
##        elpd_diff se_diff
## model3  0.0       0.0
## model2 -0.2       1.7
## model4 -0.7       1.7
## model1 -1.5       3.4
```

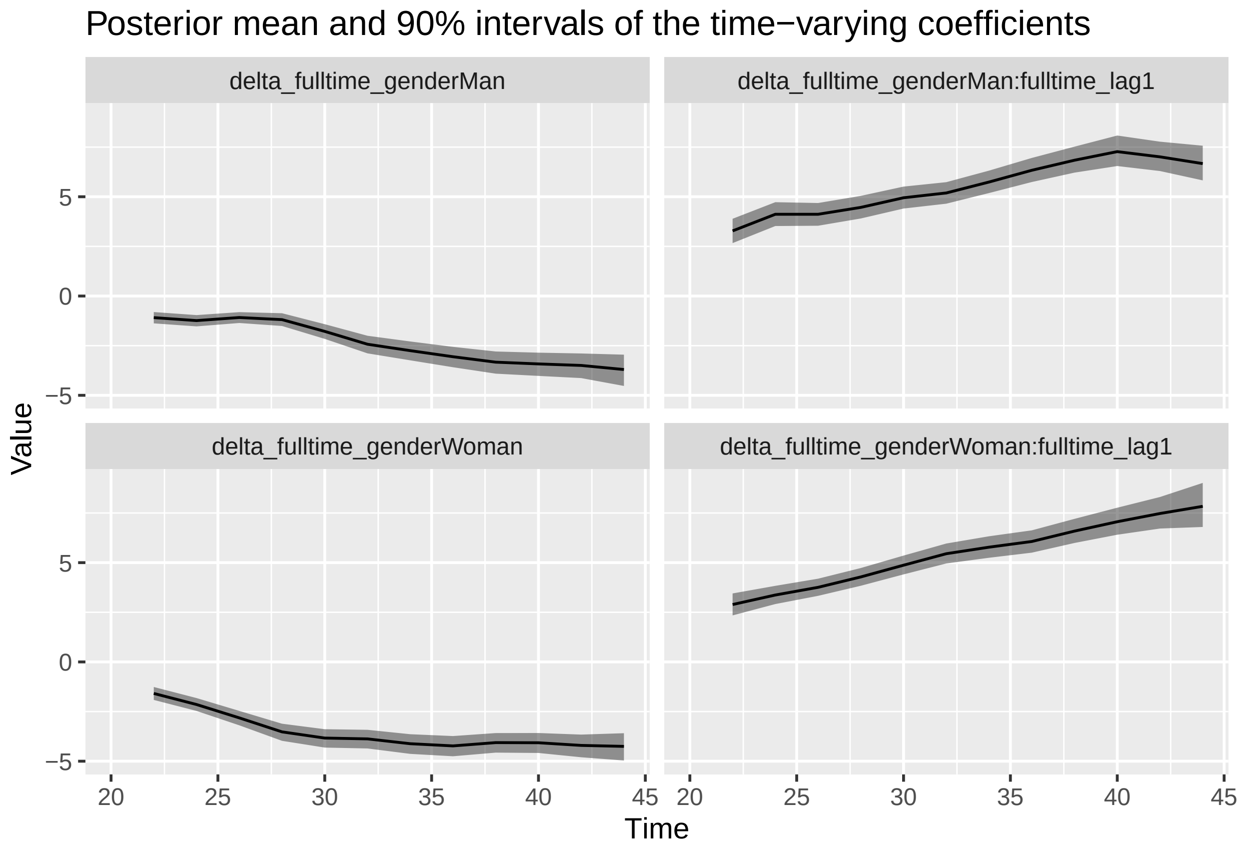
Deltas for $D = 4$:

```
plot_deltas(fit4)
```



Deltas for $D = 15$:

```
plot_deltas(fit15)
```



While there is some differences in the smoothness of the time-varying effects, from a predictive pespective the models are essentially indistinguishable. Leave-future-out cross-validation (LFO-CV) is not very informative here as we have only 13 time points and we need to condition on some initial data to make the LFO-CV stable. Nevertheless, it could be performed with `lfo` function, e.g. `lfo(fit, L = 6)` where `L = 6` defines that first 6 time points are used for the initial fit.

Check MCMC diagnostics:

```r
mcmc_diagnostics(fit)
```

```
## NUTS sampler diagnostics:
##
## No divergences, saturated max treedepths or low E-BFMIs.
##
## Smallest bulk-ESS values:
##
## tau_fulltime_genderWoman             6898
## tau_fulltime_genderMan               6954
## tau_fulltime_genderMan:fulltime_lag1 7739
##
## Smallest tail-ESS values:
##
## tau_fulltime_genderWoman:fulltime_lag1 5691
## tau_fulltime_genderMan                 7094
## tau_fulltime_genderWoman               7183
##
## Largest Rhat values:
```

```
##
## delta_fulltime_genderWoman[32] 1
## delta_fulltime_genderWoman[30] 1
## delta_fulltime_genderWoman[24] 1
```

Time-invariant parameters:

```
as_draws(fit, types = c("beta", "tau")) |>
  posterior::summarise_draws(
    "mean",
    "sd",
    ~quantile(.x, probs = c(0.025, 0.975)),
     "rhat", "ess_bulk", "ess_tail")
```

```
## # A tibble: 6 x 8
##   variable                      mean    sd `2.5%` `97.5%`  rhat ess_bulk ess_tail
##   <chr>                        <dbl> <dbl>  <dbl>   <dbl> <dbl>    <dbl>    <dbl>
## 1 beta_fulltime_genderMan:e~ 0.575  0.221  0.135   1.01   1.00   12463.    8182.
## 2 beta_fulltime_genderWoman~ 0.0677 0.230 -0.377   0.519  1.00   11623.    7958.
## 3 tau_fulltime_genderMan     1.13   0.388  0.558   2.06   1.00    6954.    7094.
## 4 tau_fulltime_genderMan:fu~ 1.32   0.414  0.681   2.26   1.00    7739.    7218.
## 5 tau_fulltime_genderWoman   1.10   0.375  0.557   2.00   1.00    6898.    7183.
## 6 tau_fulltime_genderWoman:~ 1.31   0.382  0.729   2.22   1.00    8086.    5691.
```

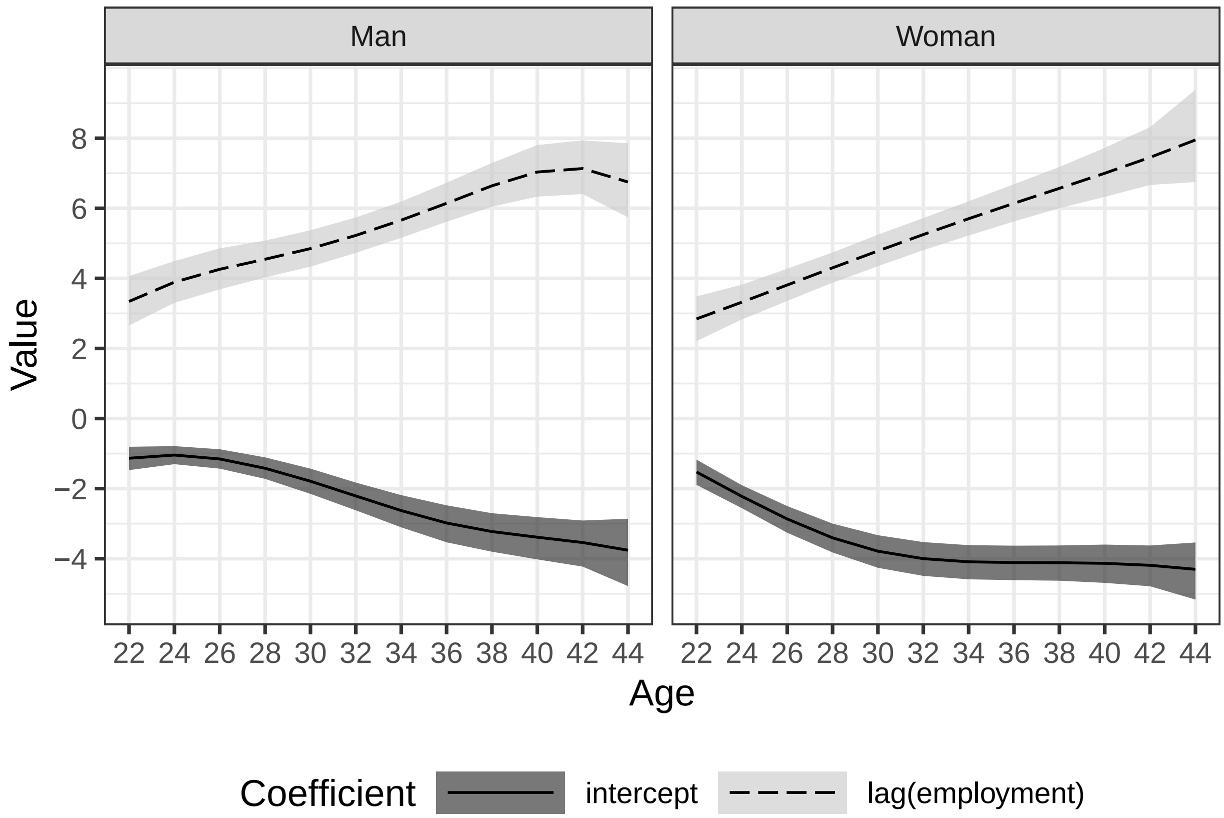
```
coef(fit, probs = c(0.025, 0.975))[, 1:5]
```

```
## # A tibble: 2 x 5
##   parameter                            mean    sd   q2.5 q97.5
##   <chr>                               <dbl> <dbl>  <dbl> <dbl>
## 1 beta_fulltime_genderWoman:ever_lag1 0.0677 0.230 -0.377 0.519
## 2 beta_fulltime_genderMan:ever_lag1   0.575  0.221  0.135 1.01
```

Draw figure of time-varying parameters:

```
coefs <- coef(fit, type = "delta", probs = c(0.025, 0.1, 0.9, 0.975)) |>
  filter(time > 20) |>
    mutate(
      gender = recode(parameter,
      delta_fulltime_genderMan = "Man",
      delta_fulltime_genderWoman = "Woman",
      "delta_fulltime_genderWoman:fulltime_lag1" = "Woman",
      "delta_fulltime_genderMan:fulltime_lag1" = "Man"
    ),
    Coefficient = recode(parameter,
        delta_fulltime_genderMan = "intercept",
        delta_fulltime_genderWoman = "intercept",
        "delta_fulltime_genderWoman:fulltime_lag1" = "lag(employment)",
        "delta_fulltime_genderMan:fulltime_lag1" = "lag(employment)"
    ))
p <- ggplot(coefs, aes(time, mean)) +
    geom_ribbon(aes(ymin = q2.5, ymax = q97.5, fill = Coefficient), alpha = 0.66) +
    geom_line(aes(linetype = Coefficient)) +
    scale_x_continuous("Age", seq(22, 44, by = 2)) +
    scale_y_continuous("Value", seq(-6, 8, by = 2)) +
    theme_bw(base_size = 14) +
    scale_linetype_manual(values = c("solid", "longdash")) +
    scale_fill_grey() +
```

```
    facet_wrap(~ gender, scales = "fixed") +
    theme(
      legend.position = "bottom",
      legend.key.width = unit(1.75, "cm"),
      panel.grid.minor.x = element_blank()
    )
p
```




```
ggsave(p, file = "../deltas_employment.png", width = 7, height = 4)
```

Estimate the causal effects:

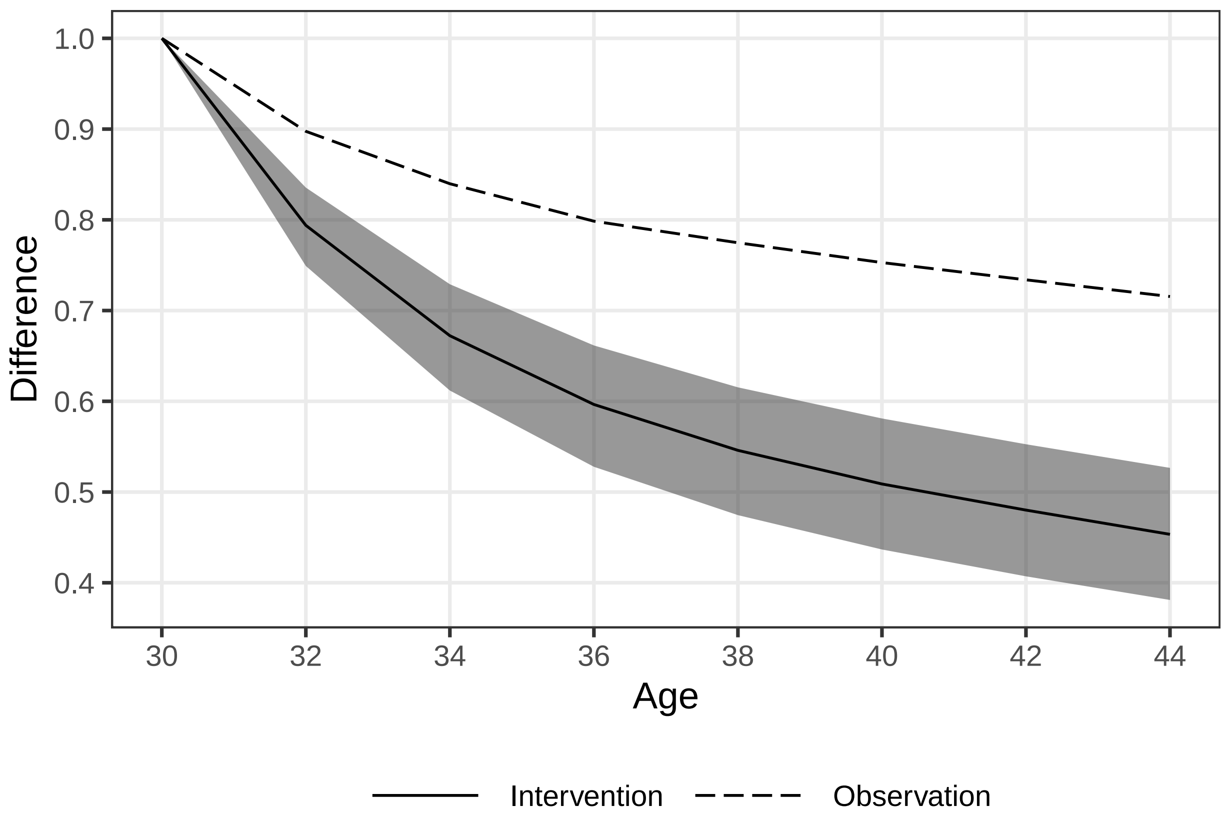
```
# No full time employment at age 30
newdata0 <- d |> filter(age >= 28)
newdata0$fulltime[newdata0$age == 30] <- 0
newdata0$fulltime[newdata0$age > 30] <- NA
# Full time employment at age 30
newdata1 <- d |> filter(age >= 28)
newdata1$fulltime[newdata1$age == 30] <- 1
newdata1$fulltime[newdata1$age > 30] <- NA

pred <-
  bind_rows(
    no = predict(
      fit, newdata = newdata0,
      funs = list(fulltime = list(mean = mean))
    )$simulated,
```

```
    yes = predict(
      fit, newdata = newdata1,
      funs = list(fulltime = list(mean = mean))
    )$simulated,
    .id = "fulltime_30"
  ) |>
  filter(age > 28) |>
  reframe(
    difference = mean_fulltime[fulltime_30 == "yes"] -
      mean_fulltime[fulltime_30 == "no"],
    .by = "age"
  ) |>
  group_by(age) |>
  summarise(
    mean = mean(difference),
    q2.5 = quantile(difference, 0.025),
    q97.5 = quantile(difference, 0.975),
    q10 = quantile(difference, 0.1),
    q90 = quantile(difference, 0.9)
  )
saveRDS(pred, file = "predictions_employment.rds")
```

Draw the figure:

```
pred <- readRDS("predictions_employment.rds")
obs_sumr <- d |> filter(age > 28) |>
  group_by(id) |>
  mutate(fulltime_30 = ifelse(fulltime[age == 30], "yes", "no")) |>
  group_by(age, fulltime_30) |>
  summarise(mean_fulltime = mean(fulltime)) |>
  group_by(age) |>
  summarise(
    mean = mean(mean_fulltime[fulltime_30 == "yes"] - mean_fulltime[fulltime_30 == "no"]),
    .groups = "keep"
  )
comb <- bind_rows(
  intervention = pred,
  observation = obs_sumr,
  .id = "Type"
)

p <- comb |>
  ggplot(aes(age, mean)) +
  geom_ribbon(
    data = comb |> filter(Type == "intervention"),
    aes(ymin = q2.5, ymax = q97.5, fill = Type),
    alpha = 0.50,
    show.legend = FALSE
  ) +
  geom_line(aes(linetype = Type)) +
  scale_x_continuous("Age", seq(30, 44, by = 2)) +
  scale_y_continuous("Difference", seq(0.2, 1, by = 0.1)) +
  theme_bw(base_size = 14) +
  scale_linetype_manual(
```

```
    name = NULL,
    values = c("solid", "longdash"),
    labels = c("Intervention", "Observation")
  )  +
  scale_fill_grey()  +
  theme(
    legend.position = "bottom",
    legend.key.width = unit(1.75, "cm"),
    panel.grid.minor.x = element_blank(),
    panel.grid.minor.y = element_blank()
  )
p
```




```
ggsave(p, file = "../causaleffect_employment.png", width = 7, height = 4)
```

Repeat the causal effect estimation for $D = 15$ for comparative purposes:

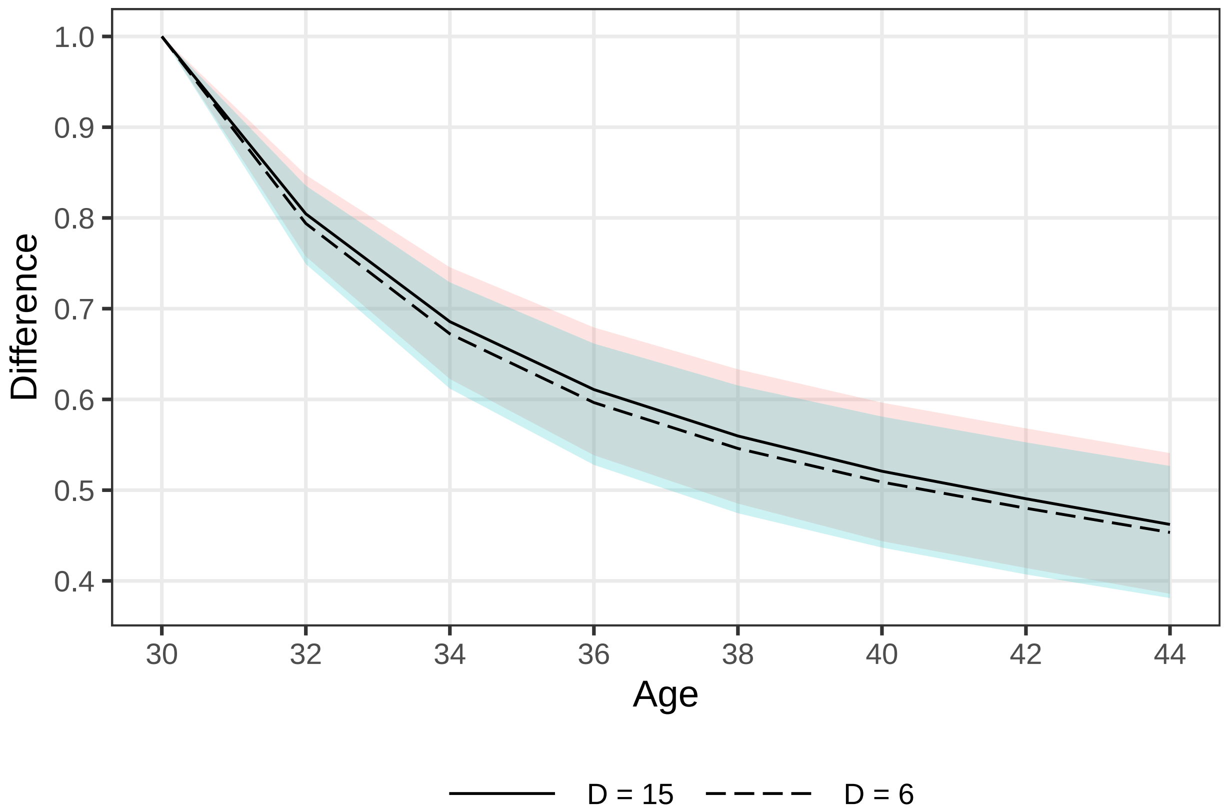
```
# No full time employment at age 30
newdata0 <- d |> filter(age >= 28)
newdata0$fulltime[newdata0$age == 30] <- 0
newdata0$fulltime[newdata0$age > 30] <- NA
# Full time employment at age 30
newdata1 <- d |> filter(age >= 28)
newdata1$fulltime[newdata1$age == 30] <- 1
newdata1$fulltime[newdata1$age > 30] <- NA

pred <-
```

```r
  bind_rows(
    no = predict(
      fit15, newdata = newdata0,
      funs = list(fulltime = list(mean = mean))
    )$simulated,
    yes = predict(
      fit, newdata = newdata1,
      funs = list(fulltime = list(mean = mean))
    )$simulated,
    .id = "fulltime_30"
  ) |>
  filter(age > 28) |>
  reframe(
    difference = mean_fulltime[fulltime_30 == "yes"] -
      mean_fulltime[fulltime_30 == "no"],
    .by = "age"
  ) |>
  group_by(age) |>
  summarise(
    mean = mean(difference),
    q2.5 = quantile(difference, 0.025),
    q97.5 = quantile(difference, 0.975),
    q10 = quantile(difference, 0.1),
    q90 = quantile(difference, 0.9)
  )
saveRDS(pred, file = "predictions_employment_D15.rds")
```

Differences in causal effect estimates with $D = 6$ and $D = 15$:

```r
pred15 <- readRDS("predictions_employment_D15.rds")
p <- bind_rows(
  d6 = pred,
  d15 = pred15,
  .id = "D"
) |>
  ggplot(aes(age, mean)) +
  geom_ribbon(
    aes(ymin = q2.5, ymax = q97.5, fill = D),
    alpha = 0.20,
    show.legend = FALSE
  ) +
  geom_line(aes(linetype = D)) +
  scale_x_continuous("Age", seq(30, 44, by = 2)) +
  scale_y_continuous("Difference", seq(0.2, 1, by = 0.1)) +
  theme_bw(base_size = 14) +
  scale_linetype_manual(
    name = NULL,
    values = c("solid", "longdash"),
    labels = c("D = 15", "D = 6")
  )  +
  theme(
    legend.position = "bottom",
    legend.key.width = unit(1.75, "cm"),
    panel.grid.minor.x = element_blank(),
```

```
    panel.grid.minor.y = element_blank()
  )
p
```




```
ggsave(p, file = "../causaleffect_employment_D6_vs_D15.png", width = 7, height = 4)
```