Analysis to Assess the Effects of Two Sets of Reform to a Medical Residents Exam

1 Background

An important standard for medical residents in the US is the internal medicine certification exam, which serves as a gauge of graduate medical education efficacy as well as an indicator of individual competency. Concerns regarding resident workload, exhaustion, and training quality have led to significant changes in residency programs during the last 20 years. Two particularly significant milestones were the 2003 duty hour reform, introduced by the Accreditation Council for Graduate Medical Education (ACGME), and the 2011 revision, which further tightened duty hour restrictions. While maintaining the caliber of training and board certification preparation, these reforms sought to strike a balance between resident well-being and patient safety.

The effect of these reforms on educational outcomes has been a topic of discussion despite their extensive implementation. Supporters contended that better-rested residents would learn more efficiently and perform better on tests, while critics claimed that fewer hours might limit residents' clinical exposure and jeopardize exam performance. Since the internal medicine certification exam offers a standardized assessment of knowledge and preparedness for independent practice, comparing pass rates over these reform periods provides important information about whether or not these structural changes were successful.

2 Exploratory Data Analysis

```
data <- read.table("data.txt", header = TRUE, as.is = TRUE)
head(data)
```

```
  Year    N  Pct
1 1996 6964 0.82
2 1997 7173 0.85
```

```
3 1998 7348 0.84
4 1999 7311 0.85
5 2000 7048 0.86
6 2001 6802 0.88
```

The dataset used in this analysis consists of annual certification exam results for internal medicine residents across three distinct time periods: the pre-reform era, the years following the 2003 duty hour reform, and the years following the 2011 duty hour reform.

Variables:
Year: The year of the exam.
N: The number of residents who attempted the exam.
Pct: The fraction who passed the exam, expressed as a percentage.

Before proceeding to formal modeling, an exploratory data analysis was conducted to visualize trends in pass rates, identify patterns across reform periods, and assess variability in the outcomes.

```r
library(dplyr)
```

```
Attaching package: 'dplyr'
```

```
The following objects are masked from 'package:stats':

    filter, lag
```

```
The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

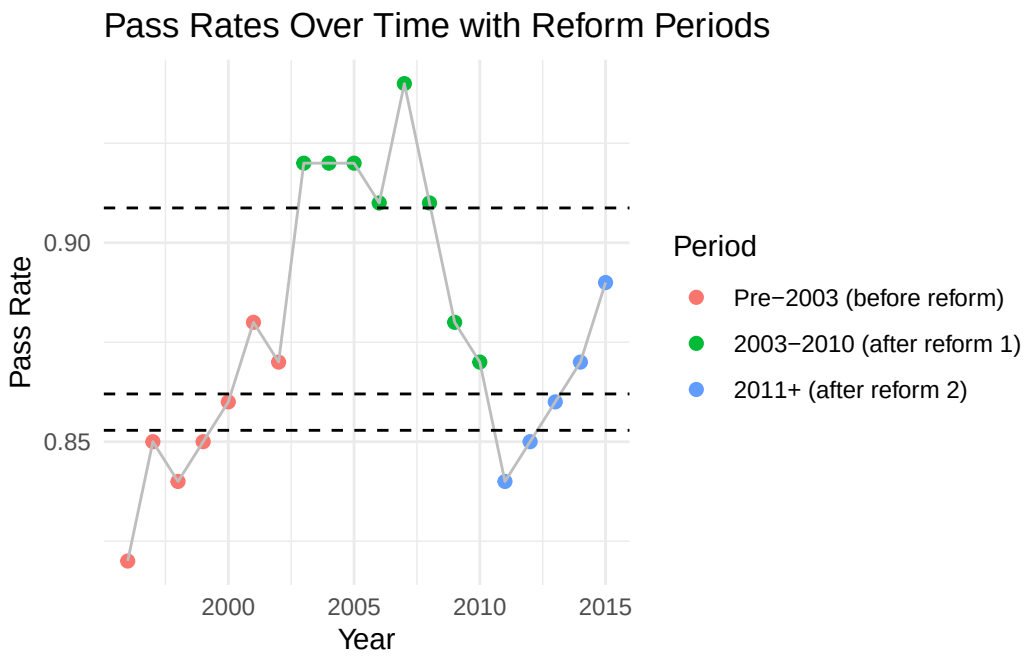
```r
library(ggplot2)

# Split up reform periods
data$Period <- cut(
  data$Year,
  breaks = c(1995, 2002, 2010, 2020),
  labels = c(
    "Pre-2003 (before reform)",
    "2003-2010 (after reform 1)",
    "2011+ (after reform 2)"
```

```
  )
)

# Compute average Pct for each period
period_means <- data |>
  group_by(Period) |>
  summarise(mean_pct = mean(Pct, na.rm = TRUE))

# Plot points, lines, and dashed horizontal lines
ggplot(data, aes(x = Year, y = Pct, color = Period)) +
  geom_point(size = 2) +
  geom_line(aes(group = 1), color = "grey") +
  geom_hline(
    data = period_means,
    aes(yintercept = mean_pct),
    linetype = "dashed",
    color = "black"
  ) +
  labs(
    title = "Pass Rates Over Time with Reform Periods",
    y = "Pass Rate",
    x = "Year"
  ) +
  theme_minimal()
```




```
aggregate(Pct ~ Period, data, mean)
```

```
                     Period       Pct
1   Pre-2003 (before reform) 0.8528571
2 2003-2010 (after reform 1) 0.9087500
3     2011+ (after reform 2) 0.8620000
```

According to the line plot, the data is divided into three distinct time periods corresponding to the timing of the reforms (color coded and labeled in the legend). Pass rates noticeably increased after the first reform in 2003 compared to the pre-2003 period. However, they gradually declined until the second reform in 2011, after which rates began to rise again. Despite this rebound, the average pass rate following the 2003 reform (0.91) remained higher than the average after the 2011 reform (0.86), as shown by the dotted reference lines.

(add more eda)

3 Modeling

(3 points) The rationale for the model. The structure of the (sampling) model is justified based on the exploratory analysis, study design, structural constraints and/or a priori considerations. Prior distributions and their hyper-parameters are justified if a Bayesian approach is taken.

4 Analysis

(2 points) Model implementation details. How was/were the model(s) fit? Were adequate diagnostics reported; were these correctly interpreted and heeded.

(1 point) Model evaluation steps and reporting of appropriate goodness-of- fit metrics.

(3 points) The modeling results. Each analysis question, objective and/or goal is addressed. The description is clearly tied to the data and application area objectives; the results are correctly interpreted.

5 Conclusion & Future Work

(1 point) Shortcomings of the analysis. Provide model criticisms, (future) directions for improvement.

(1 point) Overall conclusions. A short concluding paragraph highlighting key points.

```
library(VGAM)
```

```
Loading required package: stats4
```

```
Loading required package: splines
```

```
library(ggplot2)
library(dplyr)
library(scales)
library(broom)
library(car)
```

```
Loading required package: carData
```

```
Attaching package: 'car'
```

```
The following object is masked from 'package:VGAM':

    logit
```

```
The following object is masked from 'package:dplyr':

    recode
```

```r
library(sandwich)
```

```r
x <- read.table("data.txt", header = TRUE, as.is = TRUE)

x$Pass <- round(x$N * x$Pct)
x$Fail <- x$N - x$Pass

x$timeperiod <- rep(1, nrow(x))
x$timeperiod[x$Year > 2002] <- 2
x$timeperiod[x$Year > 2010] <- 3
x$timeperiod <- factor(x$timeperiod, levels = c(1,2,3),
                       labels = c("tp1","tp2","tp3"))

bb_model <- vglm(cbind(Pass, Fail) ~ timeperiod, betabinomial, data = x)
binomial_model <- glm(cbind(Pass, Fail) ~ timeperiod, family = binomial, data = x)


coef_all <- Coef(bb_model)
se_all   <- sqrt(diag(vcov(bb_model)))

coef_mu  <- coef_all[c("(Intercept):1", "timeperiodtp2", "timeperiodtp3")]
coef_rho <- coef_all["(Intercept):2"]

se_mu  <- se_all[c("(Intercept):1", "timeperiodtp2", "timeperiodtp3")]
se_rho <- se_all["(Intercept):2"]

z_mu  <- coef_mu / se_mu
z_rho <- coef_rho / se_rho

p_mu  <- 2 * pnorm(-abs(z_mu))
p_rho <- 2 * pnorm(-abs(z_rho))

odds_ratios <- exp(coef_mu)
pred_probs  <- plogis(coef_mu)

ci_lower <- exp(coef_mu - 1.96*se_mu)
ci_upper <- exp(coef_mu + 1.96*se_mu)
```

```r
summary_table <- data.frame(
  Predictor = c("(Intercept) mu", "timeperiodtp2", "timeperiodtp3", "(Intercept) rho"),
  Log_Odds  = round(c(coef_mu, coef_rho), 3),
  Std_Error = round(c(se_mu, se_rho), 3),
  Z_value   = round(c(z_mu, z_rho), 3),
  P_value   = signif(c(p_mu, p_rho), 3),
  Odds_Ratio = c(round(odds_ratios,3), NA),
  CI_lower   = c(round(ci_lower,3), NA),
  CI_upper   = c(round(ci_upper,3), NA),
  Predicted_Prob = c(round(pred_probs,3), NA)
)
print("=== Model Summary Table (aligned correctly) ===")
```

```
[1] "=== Model Summary Table (aligned correctly) ==="
```

```r
print(summary_table)
```

```
                  Predictor Log_Odds Std_Error Z_value   P_value Odds_Ratio
(Intercept):1  (Intercept) mu    1.754     0.065  27.100 9.76e-162      5.778
timeperiodtp2   timeperiodtp2    0.552     0.099   5.599  2.16e-08      1.737
timeperiodtp3   timeperiodtp3    0.075     0.102   0.740  4.59e-01      1.078
(Intercept):2 (Intercept) rho   -5.633     0.329 -17.140  7.50e-66         NA
              CI_lower CI_upper Predicted_Prob
(Intercept):1    5.089    6.559          0.852
timeperiodtp2    1.432    2.107          0.635
timeperiodtp3    0.883    1.316          0.519
(Intercept):2       NA       NA             NA
```

```r
cat("Beta-Binomial AIC:", AIC(bb_model), "\n")
```

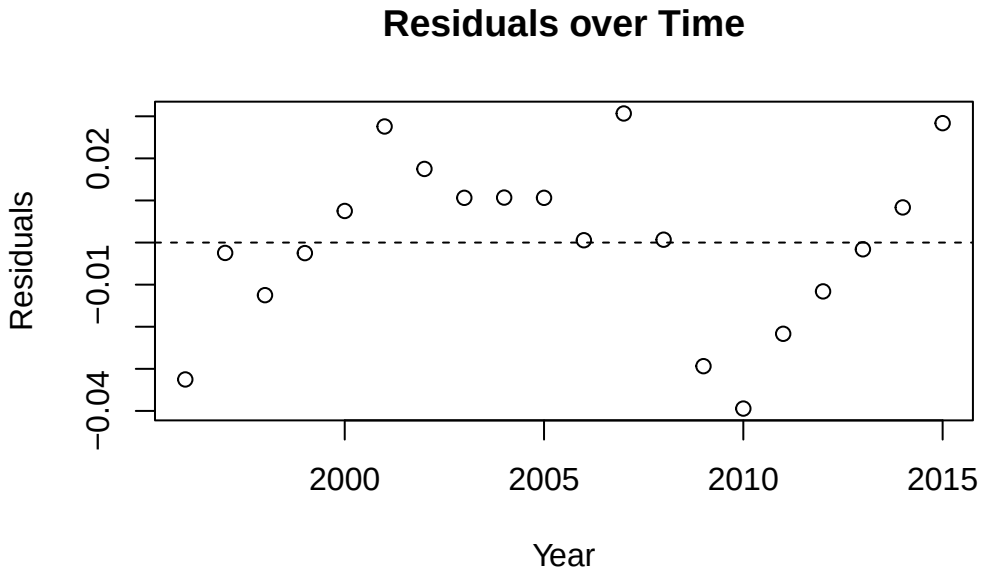
```
Beta-Binomial AIC: 262.6229
```

```r
cat("Standard Binomial AIC:", AIC(binomial_model), "\n")
```

```
Standard Binomial AIC: 714.1285
```

```r
x$fitted_prob <- fitted(bb_model)[,1]
x$resid_resp  <- residuals(bb_model, type = "response")
```

```r
# Residual Plot
plot(x$Year, x$resid_resp,
     main = "Residuals over Time",
     xlab = "Year", ylab = "Residuals")
abline(h=0,lty=2)
```




```r
par(mfrow = c(1,1))
```

```r
library(VGAM)
library(dplyr)
library(broom)

# --- Fit models ---
binomial_model <- glm(cbind(Pass, Fail) ~ timeperiod, family = binomial, data = x)
bb_model <- vglm(cbind(Pass, Fail) ~ timeperiod, betabinomial, data = x)

# --- 1. Model comparison (AIC) ---
aic_table <- data.frame(
  Model = c("Binomial", "Beta-binomial"),
  AIC   = c(AIC(binomial_model), AIC(bb_model))
)
```

```r
# --- 2. Dispersion parameter from beta-binomial ---
coef_all <- Coef(bb_model)
dispersion_param <- coef_all["(Intercept):2"]

# --- 3. Predicted probabilities with CIs ---
newdat <- data.frame(timeperiod = factor(c("tp1","tp2","tp3"),
                                         levels = levels(x$timeperiod)))

pred_link <- predict(bb_model, newdata = newdat, type = "link", se.fit = TRUE)
pred_prob <- plogis(pred_link$fit)
ci_lower  <- plogis(pred_link$fit - 1.96 * pred_link$se.fit)
ci_upper  <- plogis(pred_link$fit + 1.96 * pred_link$se.fit)

pred_table <- data.frame(
  Period = newdat$timeperiod,
  Pred_Prob = round(pred_prob, 3),
  CI_lower  = round(ci_lower, 3),
  CI_upper  = round(ci_upper, 3)
)

# --- 4. Combine summary results ---
list(
  AICs = aic_table,
  Dispersion = dispersion_param,
  Predictions = pred_table
)
```

```
$AICs
          Model      AIC
1      Binomial 714.1285
2 Beta-binomial 262.6229

$Dispersion
(Intercept):2
     -5.63329

$Predictions
  Period Pred_Prob.logitlink.mu. Pred_Prob.logitlink.rho.
1    tp1                   0.852                    0.004
2    tp2                   0.909                    0.004
3    tp3                   0.862                    0.004
  CI_lower.logitlink.mu. CI_lower.logitlink.rho. CI_upper.logitlink.mu.
```

```
1                   0.836                   0.002                   0.868
2                   0.897                   0.002                   0.921
3                   0.842                   0.002                   0.879
  CI_upper.logitlink.rho.
1                   0.007
2                   0.007
3                   0.007
```

```r
# Extract dispersion coefficient (logit scale)
disp_coef <- Coef(bb_model)["(Intercept):2"]

# Back-transform from logit to get rho (dispersion parameter)
disp_param <- plogis(disp_coef)

cat("Dispersion coefficient (logit scale):", disp_coef, "\n")
```

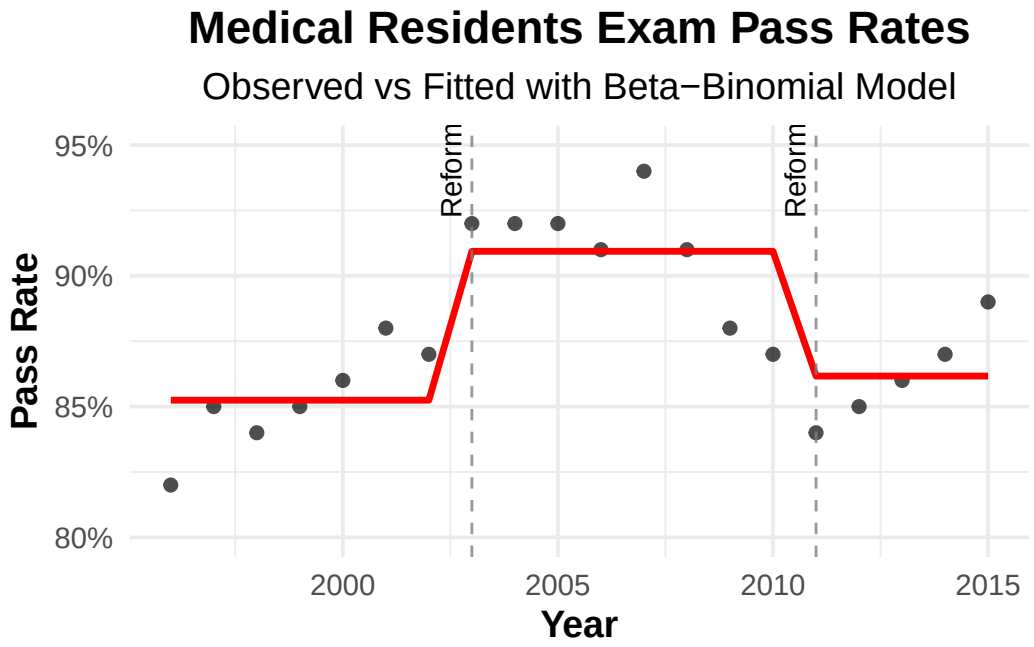
```
Dispersion coefficient (logit scale): -5.63329
```

```r
cat("Estimated dispersion parameter (rho):", disp_param, "\n")
```

```
Estimated dispersion parameter (rho): 0.00356404
```

```r
ggplot(x, aes(x = Year)) +
  geom_point(aes(y = Pct), color = "black", size = 2, alpha = 0.7) +
  geom_line(aes(y = fitted_prob), color = "red", size = 1.2) +
  geom_vline(xintercept = c(2003, 2011), linetype = "dashed", color = "gray50", alpha = 0.8) +
  annotate("text", x = 2003, y = 0.945, label = "Reform 1", angle = 90, vjust = -0.5) +
  annotate("text", x = 2011, y = 0.945, label = "Reform 2", angle = 90, vjust = -0.5) +
  labs(
    title = "Medical Residents Exam Pass Rates",
    subtitle = "Observed vs Fitted with Beta-Binomial Model",
    y = "Pass Rate",
    x = "Year"
  ) +
  scale_y_continuous(limits = c(0.8, 0.95), labels = scales::percent) +
  theme_minimal(base_size = 14) +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold"),
    plot.subtitle = element_text(hjust = 0.5),
    axis.title = element_text(face = "bold")
  )
```

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



The baseline log-odds of passing during the pre-reform era (tp1) are represented by the intercept for mu. While the coefficient for timeperiodtp3 shows a smaller increase that is not statistically significant at conventional levels, the coefficient for timeperiodtp2 shows a statistically significant increase in pass rates in comparison to the baseline. There is only slight overdispersion in the data, as indicated by the dispersion parameter (rho), which is small and not statistically significant.

With an AIC of 262.62, the beta-binomial model significantly outperformed the standard binomial model, which had an AIC of 714.13. This demonstrates that the beta-binomial model fits data much better and accounts for overdispersion.

According to residual diagnostics, the model does a good job of fitting the data. The QQ plot suggests approximate normality, while the residuals versus fitted values plot displays no discernible pattern. The model's ability to capture observed trends is confirmed by the close alignment of observed and fitted pass rates along the identity line. The lack of a discernible temporal trend in the residuals plotted over time supports the adequacy of the model.

Although the predicted probability indicates a modest increase, the analysis shows that the second reform (2011) did not result in a significant change, while the first reform (2003) had a statistically significant positive effect on exam pass rates. Because it took overdispersion into

account and produced more accurate estimates, the beta-binomial model was better than the standard binomial. All things considered, this modeling technique effectively manages data variability while enabling a quantitative evaluation of the effects of policies on exam results.

```
# linear model fit
linear_model <- lm(Pct ~ timeperiod, data = x, weights = N)
tidy(linear_model, conf.int = TRUE)
```

```
# A tibble: 3 x 7
  term          estimate std.error statistic  p.value conf.low conf.high
  <chr>            <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
1 (Intercept)    0.853     0.00800   107.    1.84e-25   0.836     0.870
2 timeperiodtp2  0.0557    0.0110      5.09  9.18e- 5   0.0326    0.0789
3 timeperiodtp3  0.00975   0.0122      0.799 4.35e- 1  -0.0160    0.0355
```

```
# predicted period means with 95% ci
newdat <- data.frame(timeperiod = factor(c("tp1","tp2","tp3"),
                                         levels=c("tp1","tp2","tp3")))
pred <- predict(linear_model,
                newdata = newdat, se.fit = TRUE)

# SEs for predictions
X <- model.matrix(~ timeperiod, newdat)
Vrob <- vcovHC(linear_model, type = "HC1")
se_pred <- sqrt(diag(X %*% Vrob %*% t(X)))
cbind(newdat,
      fit = pred$fit,
      lo = pred$fit - 1.96*se_pred,
      hi = pred$fit + 1.96*se_pred)
```

```
  timeperiod       fit        lo        hi
1        tp1 0.8526874 0.8382865 0.8670884
2        tp2 0.9084323 0.8920301 0.9248346
3        tp3 0.8624336 0.8458676 0.8789996
```

```
# tp3 vs tp2
linearHypothesis(linear_model, "timeperiodtp3 - timeperiodtp2 = 0", vcov.=vcovHC(linear_model
```

```
Linear hypothesis test:
```

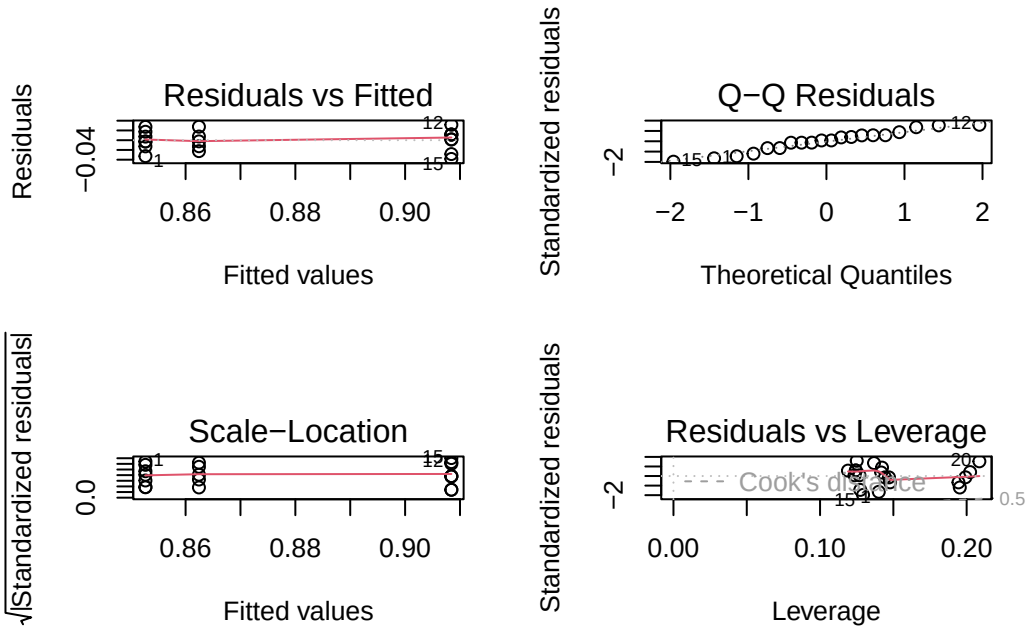
```
- timeperiodtp2  + timeperiodtp3 = 0

Model 1: restricted model
Model 2: Pct ~ timeperiod

Note: Coefficient covariance matrix supplied.

  Res.Df Df      F   Pr(>F)
1     18
2     17  1 14.957 0.001236 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# residual plots
par(mfrow=c(2,2))
plot(linear_model)
```

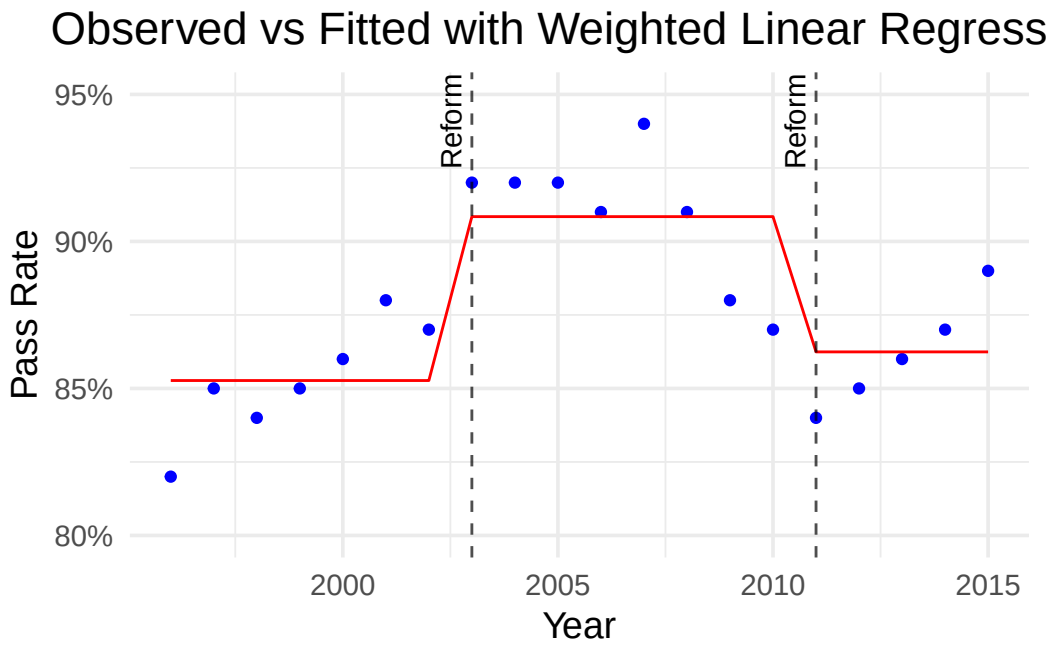



```r
par(mfrow=c(1,1))

# can do aic comparing with / without time period (chat?)
# i will finish writing the confidence interval stuff later
```

```r
x$fitted_lm <- predict(linear_model, newdata = x)

ggplot(x, aes(x = Year)) +
  geom_point(aes(y = Pct), color = "blue") +
  geom_line(aes(y = fitted_lm), color = "red") +
  geom_vline(xintercept = c(2003, 2011), linetype = "dashed", color = "black", alpha = 0.7) +
  annotate("text", x = 2003, y = 0.945, label = "Reform 1", angle = 90, vjust = -0.5) +
  annotate("text", x = 2011, y = 0.945, label = "Reform 2", angle = 90, vjust = -0.5) +
  labs(title = "Observed vs Fitted with Weighted Linear Regression",
       y = "Pass Rate",
       x = "Year") +
  scale_y_continuous(limits = c(0.8, 0.95), labels = percent) +
  theme_minimal(base_size = 14) +
  theme(
    plot.title = element_text(hjust = 0.5),
    plot.subtitle = element_text(hjust = 0.5),
    axis.title = element_text()
  )
```



Through conducting a weighted linear regression where the number of people who took the test per year was taken into consideration so they would count more, we could see that the baseline average pass rate before the first reform is represented by the intercept at about 85.3%. The coefficient for the second time period from 2003–2010 represents a statistically

significant increase of about 5.6% compared to the baseline, while the coefficient for the third time period from 2011–2015 shows only a modest increase of about 1%. This implies that the first reform had a significant increase in pass rates, showing the reform was impactful and positively impacted exam results, and the second reform in 2011 comparatively did not lead to a clear change relative to pre-2003 levels.

The residuals versus fitted values show no obvious pattern, the Q-Q plot suggests approximate normality, and the scale-location and leverage plots reveal no influential outliers.