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

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 (Appendix 1).

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.

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 (Appendix 2). 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 (Appendix 2).

3 Modeling

Exam pass rate variability across cohorts was found to be greater than what could be explained by a standard binomial model, as evidenced by a residual deviance to degrees-of-freedom ratio of 31.71 (Appendix 7). This overdispersion implies that there are more underlying variations in pass probability among cohorts than can be explained by a straightforward binomial assumption. We chose a beta-binomial model to address this, which accounts for extra-binomial variability and appropriately inflates the variance by introducing a dispersion parameter ($\rho \approx 0.0036$). This is the logit of the dispersion parameterρ in the beta-binomial model.

$$\rho = \frac{\exp(-5.633)}{1 + \exp(-5.633)} \approx 0.0036$$

Therefore, the beta-binomial model gives us more realistic confidence intervals around the estimated pass rates and enables us to model the inherent heterogeneity among cohorts.

The model's structural division of students into three time periods, pre-2003, 2003–2010, and post-2011, reflects the training program's primary reforms and is in line with significant adjustments to the curriculum, evaluation standards, and policy. This grouping captures significant changes in pass rates linked to reforms while honoring the study design and mitigating noise that could result from year-to-year variations. Future work could include a lag term to account for potential delayed impacts of reforms on cohorts that underwent multiple policy changes during their training, even though our current analysis does not formally include lagged reform effects.

Lastly, the comparison of fitted probabilities and 95% CI bands with observed pass rates and their binomial error bars provides evidence for the validity of the model (Appendix 6). The observed rates show that the model fits the data better because they are mostly contained within the beta-binomial intervals but often fall outside the binomial confidence intervals.

4 Analysis

We compared a Binomial GLM model and a Beta Binomial Regression model to see which one would best fit our data. For both models, the time periods (pre-2003, 2003–2010, 2011+) based on when the reforms occurred were used as the predictor. The residual deviance and pearson chi-square tests for the Binomial model show the lack of fit due to overdispersion. The pearson chi-square statistic (545.25 on 17 df) and residual deviance (539.15 on 17 df) are both larger than their observed values (Appendix 3). The p-values are also very small, which rejects the null hypothesis of the binomial model being a good fit for the data.

For the beta binomial model, the residuals plot shows that the points are randomly scattered around the zero line with no discernable pattern (Appendix 5). This shows that the beta binomial model captures a reasonable fit to the data. Furthermore, AIC values were calculated to evaluate the goodness of fit. The AIC for the beta binomial model (262.6) was smaller than that of the binomial model (714.1), which shows that the beta-binomial provided a better fit for the given data (Appendix 3).

P-values were also calculated for each of the two reform periods to see whether they improved the exam pass rates. The 2003 reform was associated with a statistically significant improvement in pass rates compared to the pre-2003 baseline (p 2.16e-08, which is < 0.05) (Appendix 3). The 2011 reform did not show a statistically significant change in pass rates because the p-value (p 0.46) was not less than the alpha level of 0.05. Overall, the results showed that the 2003 reform did indicate an increase in pass rates, the 2011 reform had no measurable effect, and the beta-binomial model best fit the data that was provided.

5 Conclusion & Future Work

Some shortcomings of our analysis came from the limitations of the dataset, which only provided the pass rate per year overall and did not show any individual-level variation, such as demographics, residency rotation, and program resources, limiting control for confounding. A potential for improvement would come from obtaining program or resident-level data.

Additionally, our model, which splits the time periods distinctly into three groups, assumes immediate changes and jumps instantly to a new level the moment a reform starts, rather than the possibility of phasing-in effects, which is more realistic as residents are trained across three academic years and gradually would increase exposure to the reform effects. In the future, it could be possible to use interrupted time series to analyze outcomes measured over time, using the pre-2003 years as the baseline and estimating the jump and new slope after the 2003 and 2011 reforms.

Using a beta-binomial model with the reform periods (pre-2003, 2003-2010, 2011 onward), we find strong evidence that the reform implemented in 2003 is related to a statistically significant increase in the medical certification exam pass rates in a year compared to pre-2003 (p-value « 0.05), while the reform implemented in 2011 shows no statistically significant change compared

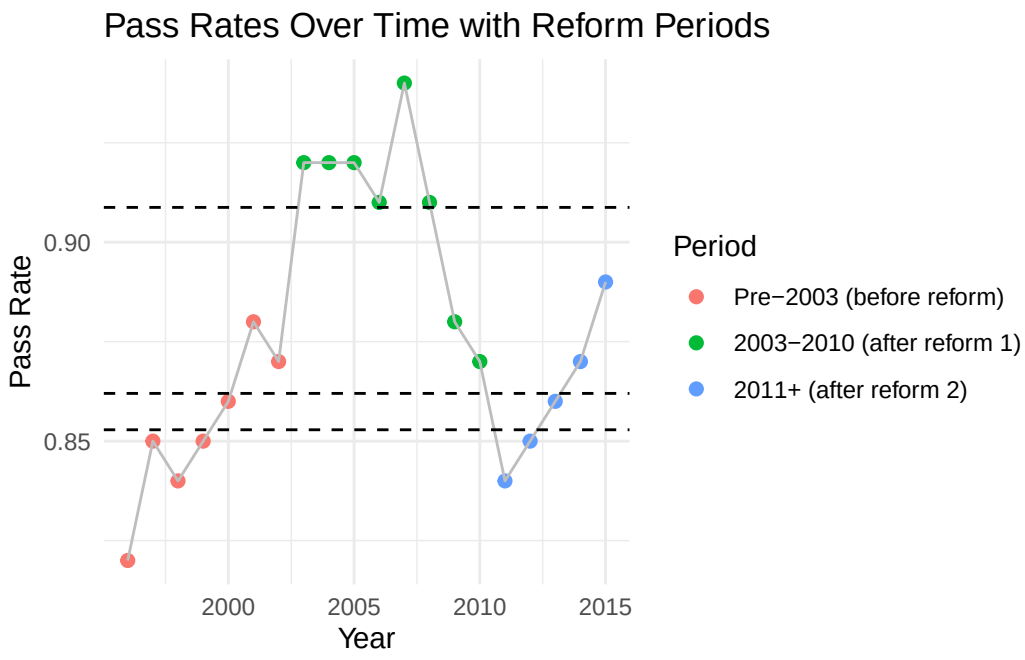
to pre-2003 and has lower average rates than 2003-2010. The beta-binomial model was favored compared to the binomial due to its lower AIC, and the residuals show no systematic pattern. These conclusions are consistent with prior evidence in medical education, where it is likely that the 80-hour cap on duty hours implemented in 2003 likely improved rest and study time, supporting higher pass rates, while the cap on first-year resident shift times introduced 2011 reduced clinical exposure.

6 Appendix

1. Table Output

```
  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
```

2. EDA Graph and Average




```
Period       Pct
```

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

3. Beta-Binomial Summary Table

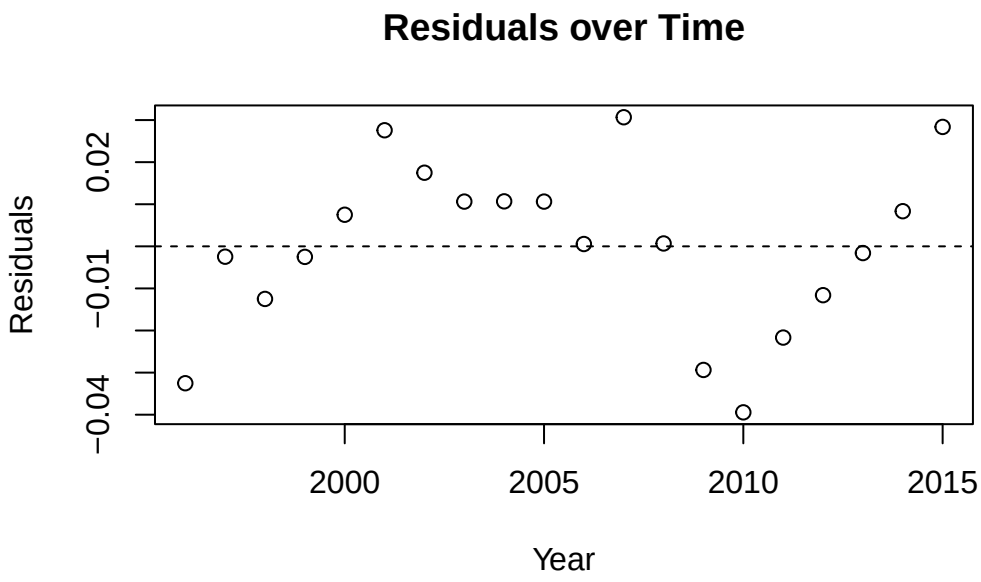
	Predictor	Log_Odds	Std_Error	Z_value	P_value	Odds_Ratio	CI_lower	CI_upper	Predicted_Prob
(Intercept):1	(Intercept) mu	1.754	0.065	27.100	9.76e-162	5.778	5.089	6.559	0.852
timeperiodtp2	timeperiodtp2	0.552	0.099	5.599	2.16e-08	1.737	1.432	2.107	0.635
timeperiodtp3	timeperiodtp3	0.075	0.102	0.740	4.59e-01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

4. Beta-Binomial Analysis Table

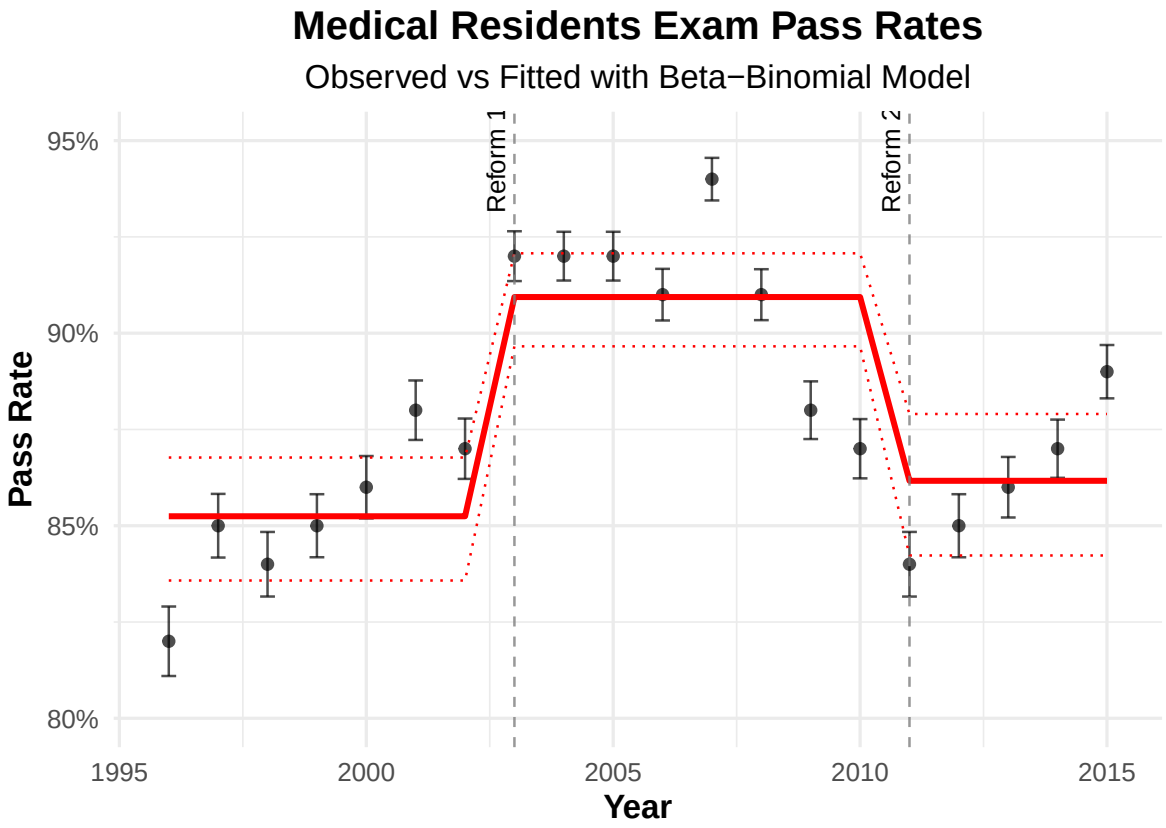
```
=== Pearson Residuals ===

Pearson Chi-square: 545.2477

Residual df: 17

Goodness-of-fit p-value: 5.40229e-105

=== Deviance Residuals ===

Residual Deviance: 539.1505

Residual df: 17

Lack-of-fit p-value: 1.0473e-103
```

```
$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

Dispersion coefficient (logit scale): -5.63329


Estimated dispersion parameter (rho): 0.00356404
```

5. Residual Plot



6. Beta-Binomial Model Fit



7. Binomial Model Dispersion Number

```
binomial_model <- glm(cbind(Pass, Fail) ~ timeperiod,
                      family = binomial, data = x)
res_dev <- deviance(binomial_model)
df_res  <- df.residual(binomial_model)

dispersion_ratio <- res_dev / df_res
dispersion_ratio
```

```
[1] 31.71473
```

Extra:

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

```r
# --- 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))
)

# --- 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
```

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
```

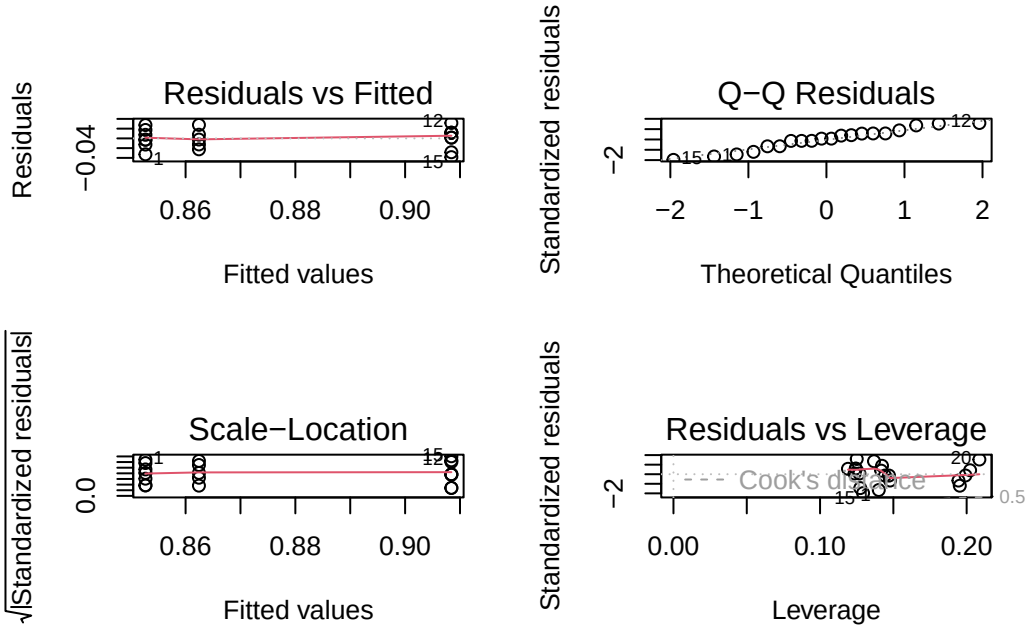
```
# predicted period means with 95% ci
newdat <- data.frame(timeperiod = factor(c("tp1","tp2","tp3"),
                                         levels=c("tp1","tp2","tp3")))
pred <- predict(linear_model,
                newdata = newdat, se.fit = TRUE)

# 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
```

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

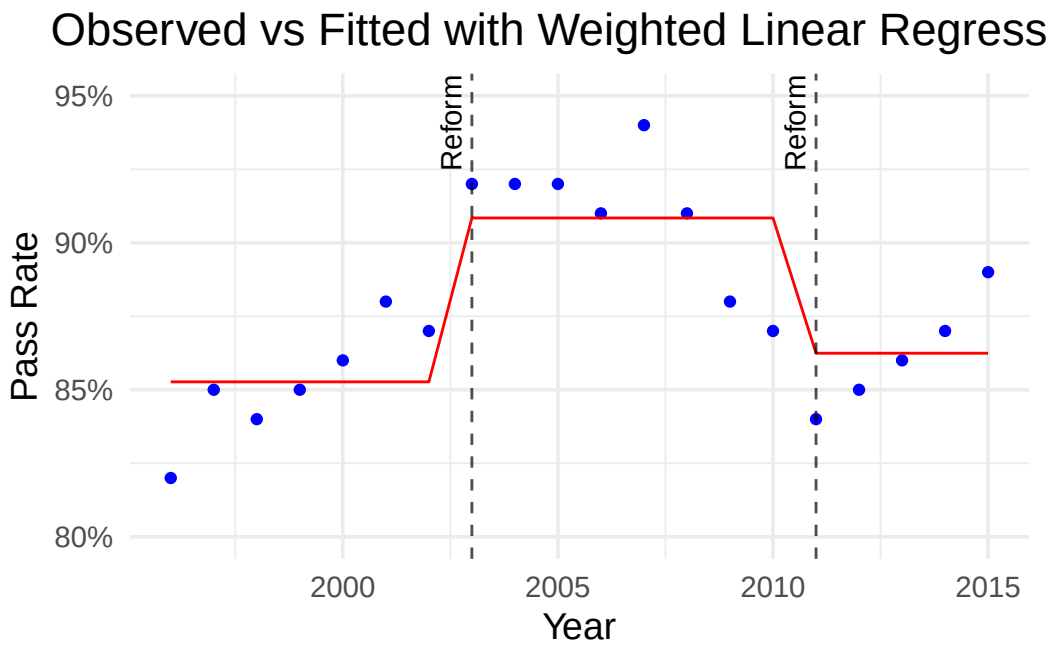
```
Linear hypothesis test

Hypothesis:
- 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)
```




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

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

```
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.