Testing for Gender Bias in COMPAS scores measuring Recidivism

We analyzed whether there was a gender bias in how COMPAS decile scores measured risk of recidivism by modeling by first defining what we considered to be gender bias. While there were various definitions that were taken into account, we decided to consider whether COMPAS itself adds extra gender bias in evaluating risk of recidivism compared to how contemporary society was evaluating. There may be gender bias in how courts evaluate recidivism risk that was measured by COMPAS rather than COMPAS creating further bias in deciding decile score. COMPAS itself may also be affecting court decisions and recidivism as a result. All these possibilities created a tangle of possible causalilties in analyzing the relationship between sex, COMPAS decile score, and recidivism. Since sex and COMPAS score could both have a causal influence on recidivism, modeling the effect of sex and score on recividism may have introduced a relationship between sex and score that does not exist, an effect known as collider bias.

We were careful in choosing the proper variables for modeling as a result. From the COMPAS data set available on ProPublica, we removed data points that showed clerical error and missing pertinent data, leaving 10,977 observations. Since we are uncertain if COMPAS and recidivism are influencing each other, we structured the analysis to find the effect of decile score on recidivism rather than vice versa. We also measured recidivism by the risk of recidivism rather than by whether that person recidivated; this removed possible variance due to differing periods of observing for recidivism.

With risk of recidivism as our dependent variable, we modeled a variety of independent variables to find the model that would be most appropriate for gender bias. The interaction between decile score and gender was the best in measuring for bias since we could interpret recidivism risk based on a fixed decile score. We also considered age and race as both individual and interacted control variables, but all resulting models were insignificant in determining gender bias. The best models included only the interaction between decile score and gender with one using numeric decile score as is, the score model, and the other with decile score as low-medium-high categories, the category model.

The p-values and 95% confidence intervals, shown in Table 1 and Table 2, suggested that, in the score model, there was no significance in gender bias, while the category model detected significant gender bias mainly with high decile scores. This was better illustrated in the predicted recidivism risk plots in Figure A. The differences in predicted risk for both models were smaller for lower scores and greater for higher scores, but the differences for the score model

were not significantly different since this difference could reasonably result from no gender bias. While these models seemed to contradict, the score model cared more for whether there was a difference between individual scores, so that a score of 4 would not be counted as the same as a score of 5 despite how similar the interpretations of the scores would be, and averaged gender differences for the entire spectrum of decile scores. We decided that the category model better displayed the gender bias because similarity between scores were considered and the effect of gender not averaged over a smaller range of scores.

From the category model, the coefficient of proportional recidivism risk for high scored men compared to high scored women was close to 2, meaning that high scored men were two-times more at risk of recidivism compared to high scored women. Since both are effectively receiving the same COMPAS score, we would expect insignificant difference in recidivism risk between the groups, but the existence of a significant difference suggested that men are more likely to recidivate given the same COMPAS score category. If men of the same COMPAS score were more at risk to recidivate, COMPAS should have calculated a higher COMPAS score for those men or a lower score for the women, as the women were less at risk to recidivate.

We concluded that COMPAS was giving women unfairly higher scores, but only for high scores. The difference in risk by gender is insignificant at low and medium scores but seems to increase as the score increased. There were still limitations that needed to be considered. The category model only worked with the assumption that the difference between each consecutive score would be linearly consistent where the difference between a score of 1 and 2 would be the same as that between a score of 8 and 9. We also did not consider weighting each category of sex and decile score, so some categories may be better represented than others. Assuming that this data set was representative of all COMPAS data sets, there was evidence of gender bias in high decile scores accounting for variance in recidivism observation period and possible collider bias in causal relationships.

Table 1: Proportional Risk of Recidivism compared to Female with 95% Conf. Int (Using Decile Score)

Variables	Prop. Risk	P-value
v_decile_score	1.238 (1.146, 1.337)	0.000**
sexMale	1.382 (0.938, 2.035)	0.102
v_decile_score:sexMale	1.035 (0.954, 1.123)	0.408

P-value: significant at 0.05(*) or 0.01(**) level

Table 2: Proportional Risk of Recidivism compared to High Scored Female with 95% Conf. Int. (Using Decile Score Category)

Variables	Prop. Risk	P-value
v_score_textLow	0.262 (0.134, 0.509)	0.000**
v_score_textMedium	0.587 (0.287, 1.201)	0.144
v_score_textN/A	NA (NA, NA)	NA
sexMale	1.965 (1.038, 3.722)	0.038*
v_score_textLow:sexMale	0.799 (0.400, 1.595)	0.524
v_score_textMedium:sexMale	0.859 (0.409, 1.806)	0.689
v_score_textN/A:sexMale	NA (NA, NA)	NA

P-value: significant at 0.05(*) or 0.01(**) level

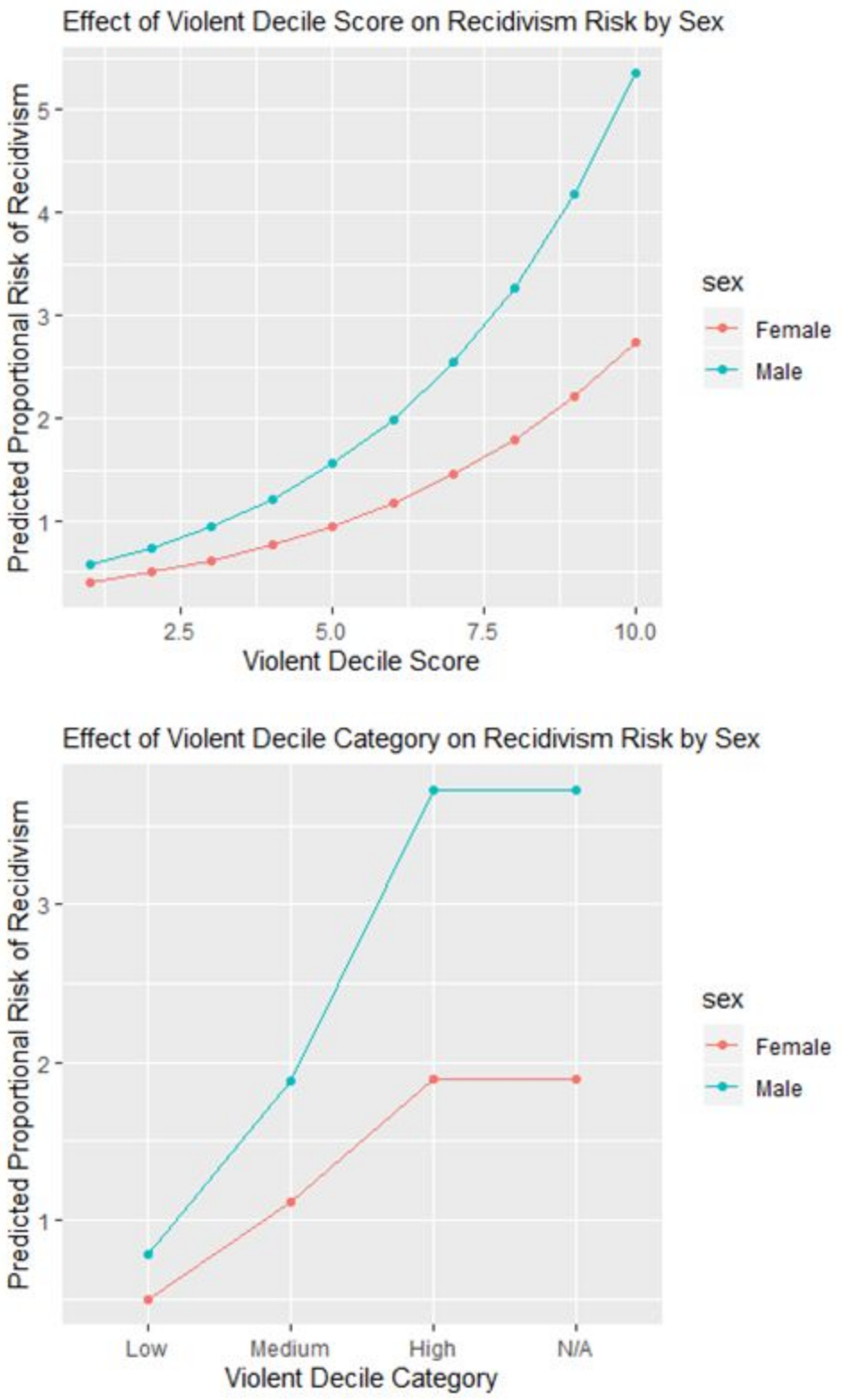

Figure A: Predicted Proportional Risk by Sex for Decile Score (top) and Decile Category (bottom)

```r
data <- read.csv("https://raw.githubusercontent.com/propublica/compas-analysis/master/cox-violent-parsed.csv")
data <- as.data.frame(data)
dat <- data %>%
  group_by(id) %>%
  filter(row_number()==1)%>%
  ungroup() %>%
  mutate(time_risk = end-start) %>%
  filter(time_risk > 0) %>%
  filter(v_decile_score >= 1) %>%
  select(id, sex, is_violent_recid,
         v_decile_score, v_score_text, time_risk)
dat <- dat[complete.cases(dat),]
```

```r
surv_obj <- with(dat, Surv(time_risk, is_violent_recid))

score_model <- coxph(surv_obj~v_decile_score*sex, data=dat)

cat_model <- coxph(surv_obj~v_score_text*sex, data=dat)

score_table <- data.frame(risk_coef = summary(score_model)$coef[,2],
                          lower_95 = exp(confint(score_model))[,1],
                          upper_95 = exp(confint(score_model))[,2],
                          p_value = summary(score_model)$coef[,5]) %>%
  transmute(variables = row.names(summary(score_model)$coef),
            risk_coef = sprintf("%0.3f (%0.3f, %0.3f)", risk_coef, lower_95, upper_95),
            p_value = sprintf("%0.3f", p_value))
knitr::kable(score_table,
             col.names=c("Variables", "Prop. Risk", "P-value"),
             caption="Proportional Risk of Recidivism compared to Female with 95% Conf. Int (Using Decile Score)")

cat_table <- data.frame(risk_coef = summary(cat_model)$coef[,2],
                          lower_95 = exp(confint(cat_model))[,1],
                          upper_95 = exp(confint(cat_model))[,2],
                          p_value = summary(cat_model)$coef[,5]) %>%
  transmute(variables = row.names(summary(cat_model)$coef),
            risk_coef = sprintf("%0.3f (%0.3f, %0.3f)", risk_coef, lower_95, upper_95),
            p_value = sprintf("%0.3f", p_value))
knitr::kable(cat_table,
             col.names=c("Variables", "Prop. Risk", "P-value"),
             caption="Proportional Risk of Recidivism compared to High Scored Female with 95% Conf. Int. (Using Decile Score Category)")

score_new_values <- data.frame(v_decile_score = rep(c(1:10), 2),
                         sex = rep(c("Male", "Female"), each=10))
score_new_values$sex <- as.factor(score_new_values$sex)
score_new_values$pred <- predict(score_model, newdata=score_new_values)
ggplot(data=score_new_values, aes(x=v_decile_score, y=exp(pred), group=sex)) +
  labs(title = "Effect of Violent Decile Score on Recidivism Risk by Sex",
       x = "Violent Decile Score",
       y = "Predicted Proportional Risk of Recidivism") +
  geom_line(aes(color=sex)) +
  geom_point(aes(color=sex)) + theme(plot.title = element_text(size=11))

cat_new_values <- data.frame(v_score_text = rep(c("Low", "Medium", "High", "N/A"), 2),
                         sex = rep(c("Male", "Female"), each=4))
cat_new_values$v_score_text <- as.factor(cat_new_values$v_score_text)
cat_new_values$sex <- as.factor(cat_new_values$sex)
```

```r
cat_new_values$pred <- predict(cat_model, newdata=cat_new_values)
ggplot(data=cat_new_values, aes(x=v_score_text, y=exp(pred), group=sex)) +
  labs(title = "Effect of Violent Decile Category on Recidivism Risk by Sex",
       y = "Predicted Proportional Risk of Recidivism") +
  geom_point(aes(color=sex)) +
  geom_line(aes(color=sex)) +
  scale_x_discrete(name ="Violent Decile Category",
                   limits = c("Low","Medium","High", "N/A")) +
  theme(plot.title = element_text(size=11))
```

Figure B: R code for Analysis and Graphs