WNBA Betting Data Manipulation

2025-11-06

Importing Data

```r
library(readr)
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(tidyr)
library(ggplot2)
library(janitor)
```

```
##
## Attaching package: ’janitor’

## The following objects are masked from ’package:stats’:
##
##     chisq.test, fisher.test
```

```r
library(scales)
```

```
##
## Attaching package: ’scales’

## The following object is masked from ’package:readr’:
##
##     col_factor
```

```r
library(broom)

wnba <- read_csv("WNBA Rookie Impact II - CSV Export.csv")
```

```
## Rows: 386 Columns: 12
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (2): Favorite, Underdog
## dbl (10): Period, Fav Odds, Fav Implied Odds, Dog Odds, Dog Implied Odds, Fa...
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

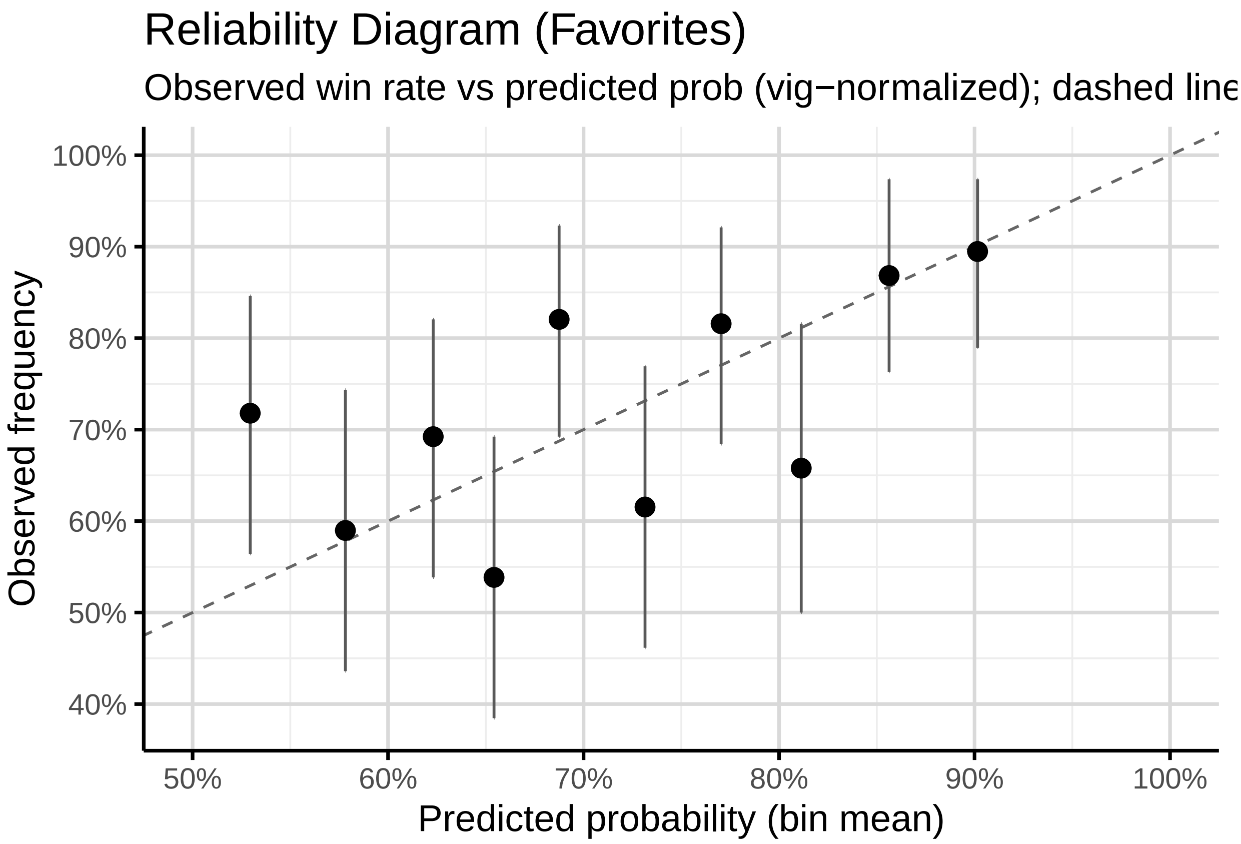
```r
#glimpse(wnba)
wnba <- wnba |> clean_names()
```

General Data Manipulation

```r
wnba <- wnba |>
  mutate(total_prob = fav_implied_odds + dog_implied_odds,
         fav_implied_norm = fav_implied_odds / total_prob,
         dog_implied_norm = dog_implied_odds / total_prob,
         inefficiency = fav_implied_norm - fav_win,
         fav_win = as.integer(fav_win %in% 1))


set.seed(1)

wnba_cal <- wnba |>
  filter(!is.na(fav_implied_norm), !is.na(fav_win)) |>
  mutate(bin = ntile(fav_implied_norm, 10)) |>
  group_by(bin) |>
  summarise(
    n = n(),
    pred_mean = mean(fav_implied_norm),
    obs_mean  = mean(fav_win),
    # Wilson 95% CI for binomial proportion
    ci_low  = qbinom(0.025, n, obs_mean) / n,
    ci_high = qbinom(0.975, n, obs_mean) / n
  ) |>
  ungroup()
```

Data Visualization and Conclusions

```r
p <- ggplot(wnba_cal, aes(x = pred_mean, y = obs_mean)) +
  geom_abline(slope = 1, intercept = 0, linetype = "dashed", color = "grey40") +
  geom_errorbar(aes(ymin = ci_low, ymax = ci_high), width = 0, color = "grey35") +
  geom_point(size = 3, color = "black") +
  scale_x_continuous(labels = percent_format(accuracy = 1),
                     breaks = seq(0, 1, 0.1), limits = c(0.5, 1)) +
  scale_y_continuous(labels = percent_format(accuracy = 1),
                     breaks = seq(0, 1, 0.1), limits = c(0.38, 1)) +
  labs(title = "Reliability Diagram (Favorites)",
       subtitle = "Observed win rate vs predicted prob (vig-normalized); dashed line = perfect calibrat
       x = "Predicted probability (bin mean)", y = "Observed frequency") +
  theme_minimal(base_size = 14) +
  theme(
    plot.background  = element_rect(fill = "white", color = NA),
    panel.background = element_rect(fill = "white", color = NA),
    panel.grid.major = element_line(color = "grey85"),
    panel.grid.minor = element_line(color = "grey93"),
    axis.line        = element_line(color = "black"),
    axis.ticks       = element_line(color = "black")
  )

p
```



```
ggsave("wnba_calibration_plot.png", p, width = 9, height = 6, dpi = 300, units = "in")


logit <- function(p) log(p/(1-p))
fit_slope <- glm(fav_win ~ logit(fav_implied_norm),
                 data = wnba, family = binomial())
coef(fit_slope)
```

```
##             (Intercept) logit(fav_implied_norm)
##               0.3686309               0.6120693
```

```
brier <- mean((wnba$fav_implied_norm - wnba$fav_win)^2, na.rm = TRUE)
base_rate <- mean(wnba$fav_win, na.rm = TRUE)
brier_baseline <- mean((base_rate - wnba$fav_win)^2, na.rm = TRUE)
brier
```

```
## [1] 0.1992337
```

```
brier_baseline
```

```
## [1] 0.2015088
```

The calibration slope = 0.61 The intercept = +0.37 The Brier score = 0.199 vs. the baseline which = 0.201 This shows that the WNBA betting market's probabilities were largely accurate and well-calibrated. Favorites slightly outperformed their implied odds, indicating that the market undervalued them marginally, but the effect was not statistically significant. Overall, the market's pricing was efficient. It correctly ranked team strength and produced near optimal predictive accuracy.