Australian Homefield Advantage in Cricket

Eric Chen, Domenic Hoffman, Valeria Martinez, Sidhart Sathya, Mark Combs

2023-12-15

Executive Summary

This report will attempt to decide if there is a true 'homefield advantage' in Australian Cricket, or if it is simply a figure of speech that holds no competitive value. We will apply ideas and methods learned throughout the course of our Statistics 461 class taught by Professor Neil Hatfield during the Fall 2023 semester at The Pennsylvania State University.

Introduction and Background

One of the most well-known phenomena in sports that has been used for decades to 'predict' the result of a match has been home-field advantage. In its simplest form, home-field advantage is described as the benefit the home team has over the visiting team. Researchers have tried for decades to pinpoint whether this phenomenon can be proved statistically, mathematically, and psychologically.

Most people were introduced to the idea of using statistics in sports with the uber-famous baseball movie Moneyball (2011). In the film, a team manager takes a revolutionary approach to scouting by relying on sabermetrics (a form of game data in baseball). However, the origins of sports analytics can be traced back to the early 20th century. In recent years, the field of sports analytics has exploded. Thousands, if not millions, of websites are filled with data from almost every sport on the planet; it is easier than ever to analyze sports and players alike.

Cricket is no exception to the boom of interest in analytics. A historic sport with three formats and origins in England boasts a fan base of millions all over the world. Researchers and fans want to know what the numbers reveal about the game and perhaps how to predict winners and losers.

Literature Review

LIT REVIEW GOES HERE LIT REVIEW GOES HERE LIT REVIEW GOES HERE LIT REVIEW GOES HERE LIT REVIEW GOES HERE LIT REVIEW GOES HERE

Research Questions

We have several research questions that we wanted to answer in this study:

Does home field percentage influence win-loss percentage for Australia's cricket team?

Hypotheses:

Null hypothesis: Location does not have a statistically significant impact on win-loss percentage for Australia's cricket team.

Alternative hypothesis: Location does have a statistically significant impact on win-loss percentage for Australia's cricket team.

Does game format influence win-loss percentage for Australia's cricket team?

Hypotheses:

Null hypothesis: Game format does not have a statistically significant impact on win-loss percentage for Australia's cricket team.

Alternative hypothesis: Game format does have a statistically significant impact on win-loss percentage for Australia's cricket team.

Methods

The data was scraped from our sources, only including Australia as the team. After inputting the data and verifying any errors, we calculated the numeric values we wished to test: win percentage(as a proportion) and the proportion of games played at home. The home game percentage was calculated by grouping the data by year, then simply taking the number of home games that year divided by the number of total games that year. The win percentage was calculated by again grouping by year, but this time taking only wins, counting a tie as non-win, and dividing by the total number of games.

After we found these values, the standard assumptions of homoscedasticity, residuals follow a Gaussian distribution, and independence of observations. Assuming these assumptions can be met we will then continue with the analysis and fit a one-way ANCOVA model with a block. If the ANCOVA model produces any significant results, according to a Type-I error rate of 0.1, we will continue with Post-Hoc analysis, and discuss our results accordingly.

Appropriateness of ANCOVA

Since our response of interest is the win percentage of the Australian cricket team, we have a continuous response. Furthermore, our factor of interest, game format, is categorical with three levels (ODI, T20I, or Test). We also have a covariate (home game proportion) that could influence our response. Since our covariate is a proportion, we have a continuous attribute that we are interested in as well.

Figure (hasse_diagram) shows the Hasse diagram for our study. As seen in the diagram, we have a block, a fixed factor, and a covariate. In addition, with our sample size of 33, we have enough degrees of freedom to estimate all main effects and error terms.

These properties show that ANCOVA is the most appropriate model to use for our study since we are interested in the effect of a continuous attribute on our response. Since we also want to control a nuisance attribute (year the match took place), we decided to add a block. Thus, we can explore the effect of our factor and covariate on the response using our one-way ANCOVA with a block model.

$y_{ij} = \mu_{\cdots} + \alpha_i + \beta x_{ij} + \gamma_k + \epsilon_{ijk}$

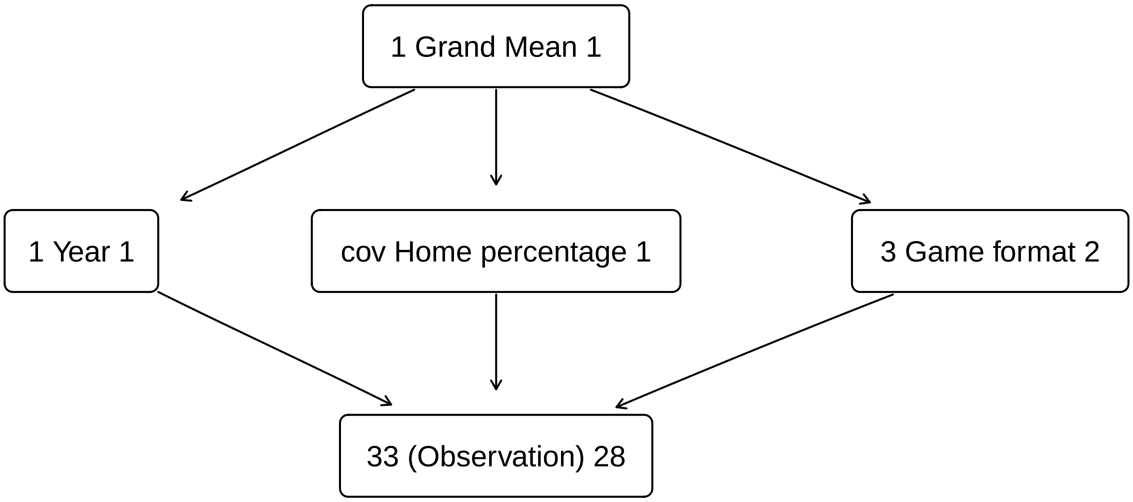

Figure 1: Hasse Diagram

The above equation and hasse diagram (Figure 1) both represent the underlying fixed-effects one-way ANCOVA model with a block used for our study. More precisely, $\mu_{\cdot\cdot\cdot}$ represents the baseline effect of the Australian cricket team's win percentage. α_i represents the main effect of game format, and βx_{ij} is the effect of the covariate (home game proportion). γ_k represents the effect of the block (year). Finally, ϵ_{ijk} represents the residual term, or the inherent variability and other sources not accounted for in the model.

Data Analysis

Table 1: Summary Statistics for Win Rate by Type of Match

	n	Min	Q1	Median	Q3	Max	MAD	SAM	SASD	Sample Skew	Sample Ex. Kurtosis
ODI	11	0.154	0.451	0.567	0.701	0.789	0.191	0.551	0.202	-0.512	-1.035
T20I	11	0.000	0.422	0.500	0.573	1.000	0.135	0.486	0.260	-0.007	-0.237
Test	11	0.000	0.406	0.556	0.626	0.667	0.120	0.491	0.198	-1.279	0.612

Looking at Table 1 we are looking at the Win Percent summary statistics grouped by each type of cricket match, most notably we see that the T20I match type has the largest range of win percentage going from 0% all the way to 100%. This will be important to keep in mind later on when it is time to check the assumptions.

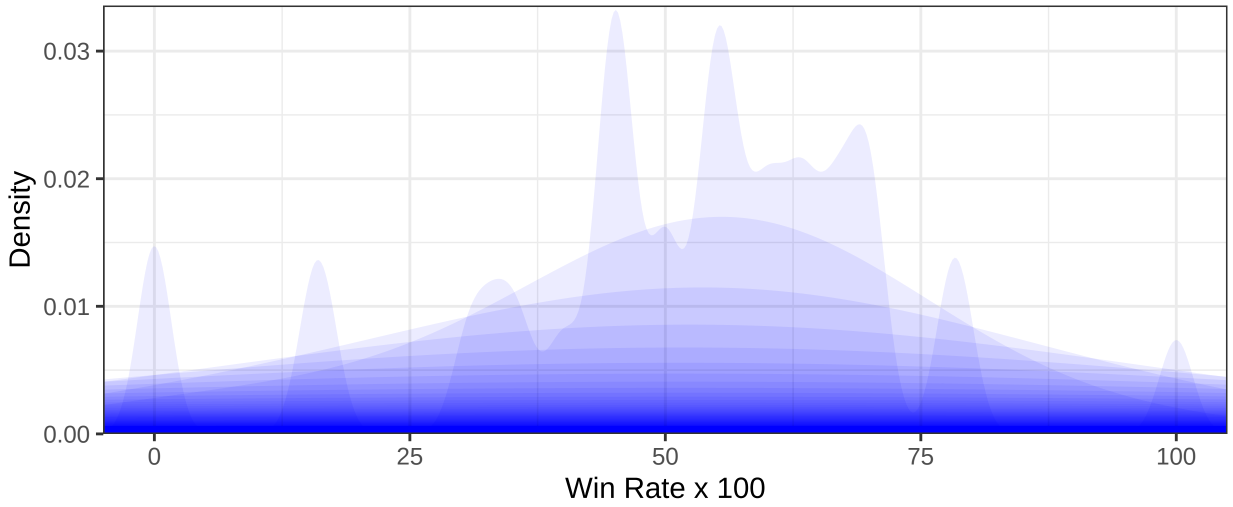

Figure 2: Shadowgram of Win Rate

Figure 2 provides the shadowgram for our 33 win rates. In examining the shadowgram, we can that there is one dominant modal clump (essentially the entire graph) that peaks around a 60% win rate with some faint separation in the background. While we know that we have three groups based upon the types of cricket matches, Figure 2 suggests that there may not be significant differences in terms of win rate between the type of match.

ANCOVA Assumptions

Gaussian Residuals

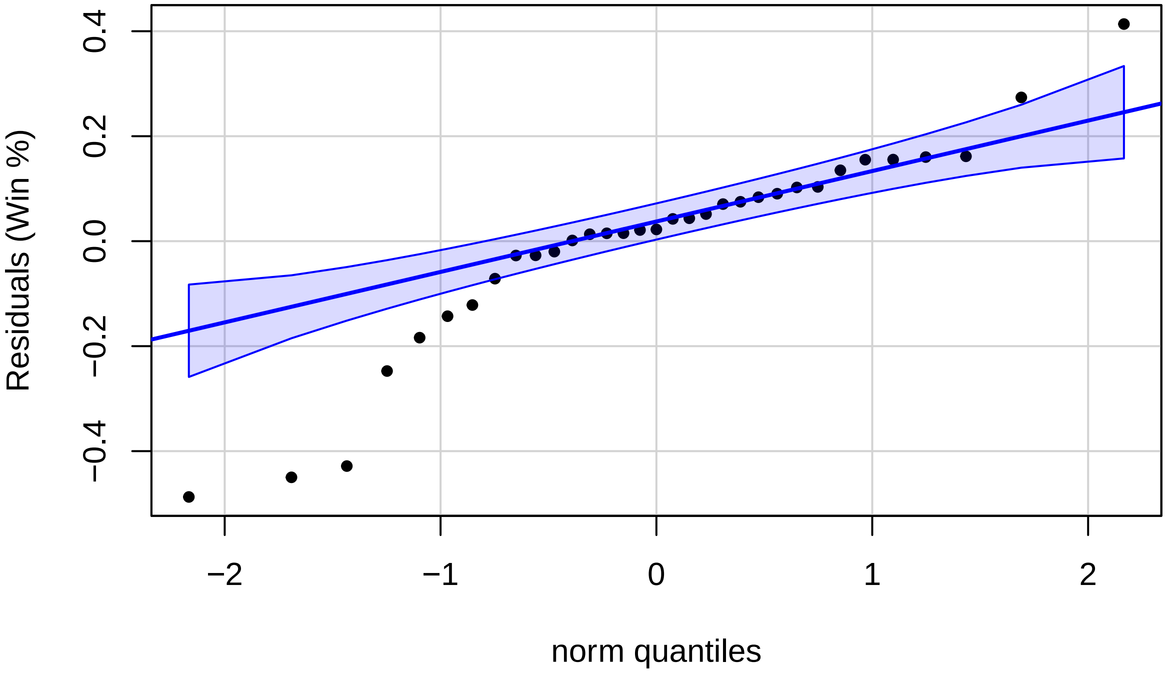

Figure 3: Gaussian Residuals

The first assumption that was tested was if the residuals followed a Gaussian distribution, and looking at Figure 3 we see about 10 points outside the 90% confidence envelope, therefore invalidating this assumption. Due to this error, we must transform the residuals, the method we will use will be to square the win percentage in order to help makeup for the imbalance we currently see.

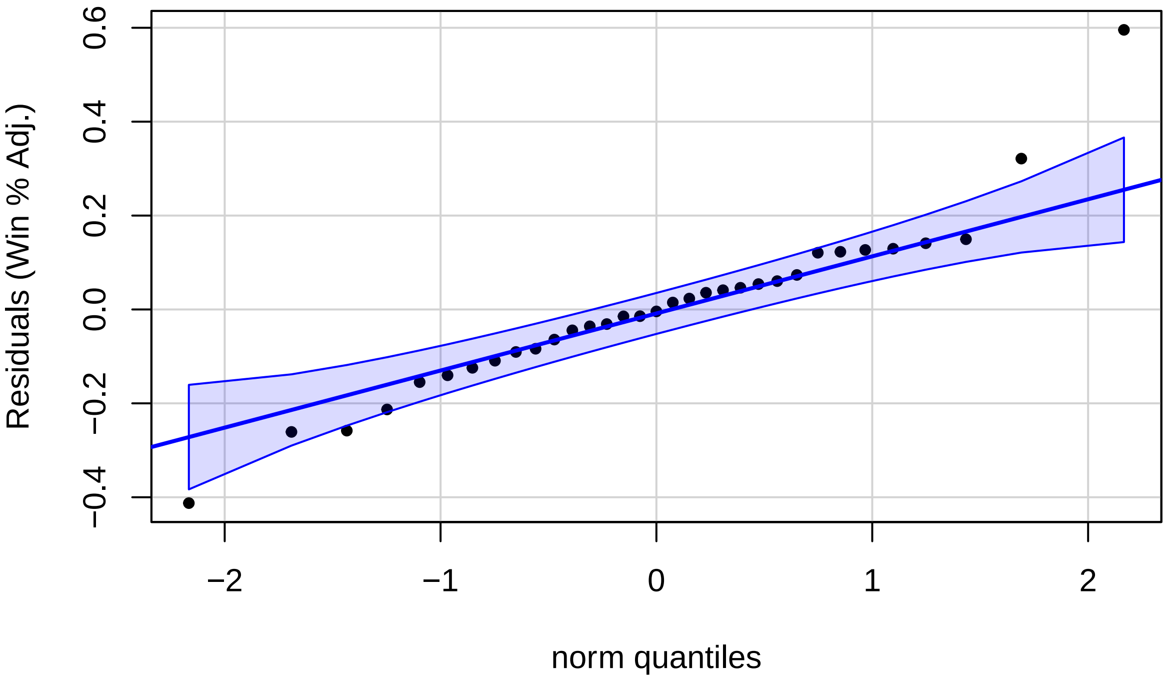

Figure 4: Transformed Gaussian Residuals

After performing the transformation, we see in Figure 4 only 3 points lie fully outside the envelope with one on the edge, this vastly improves the assumption, though due to the limited nature of the data we would ideally want to have less than 3.3 observations to fulfill the 90% envelope, with this is mind we will decide that the assumption is satisfied and continue on.

Linear Relationship

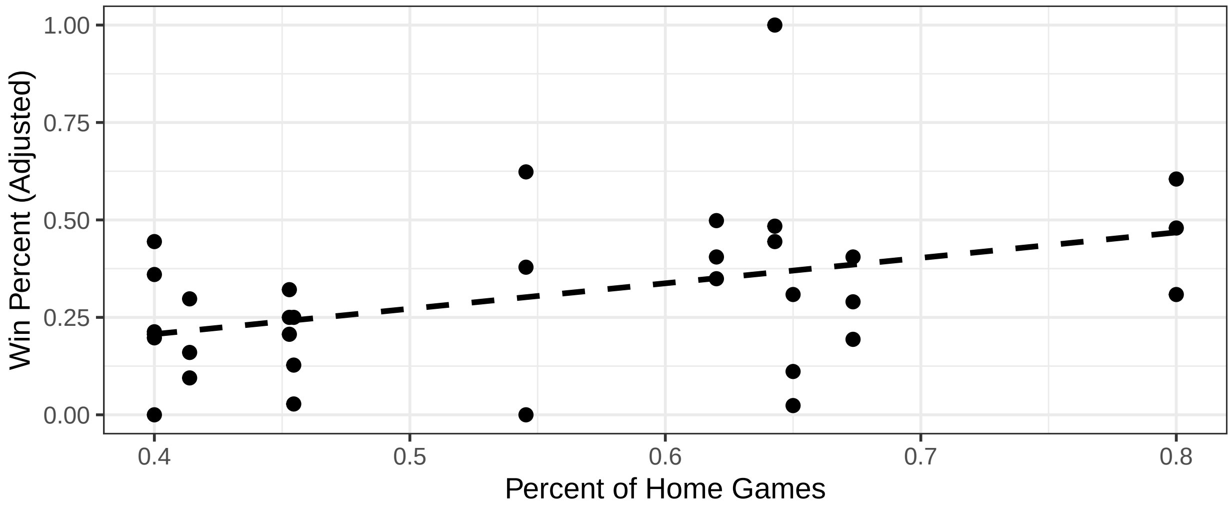

Figure 5: Linear Relationshop

While not much is easily seen in Figure 5, there is subtle positive linear relationship between the percent of home games and the adjusted win percent, thus satisfying the linear relationship bewteen the response and covariate assumption.

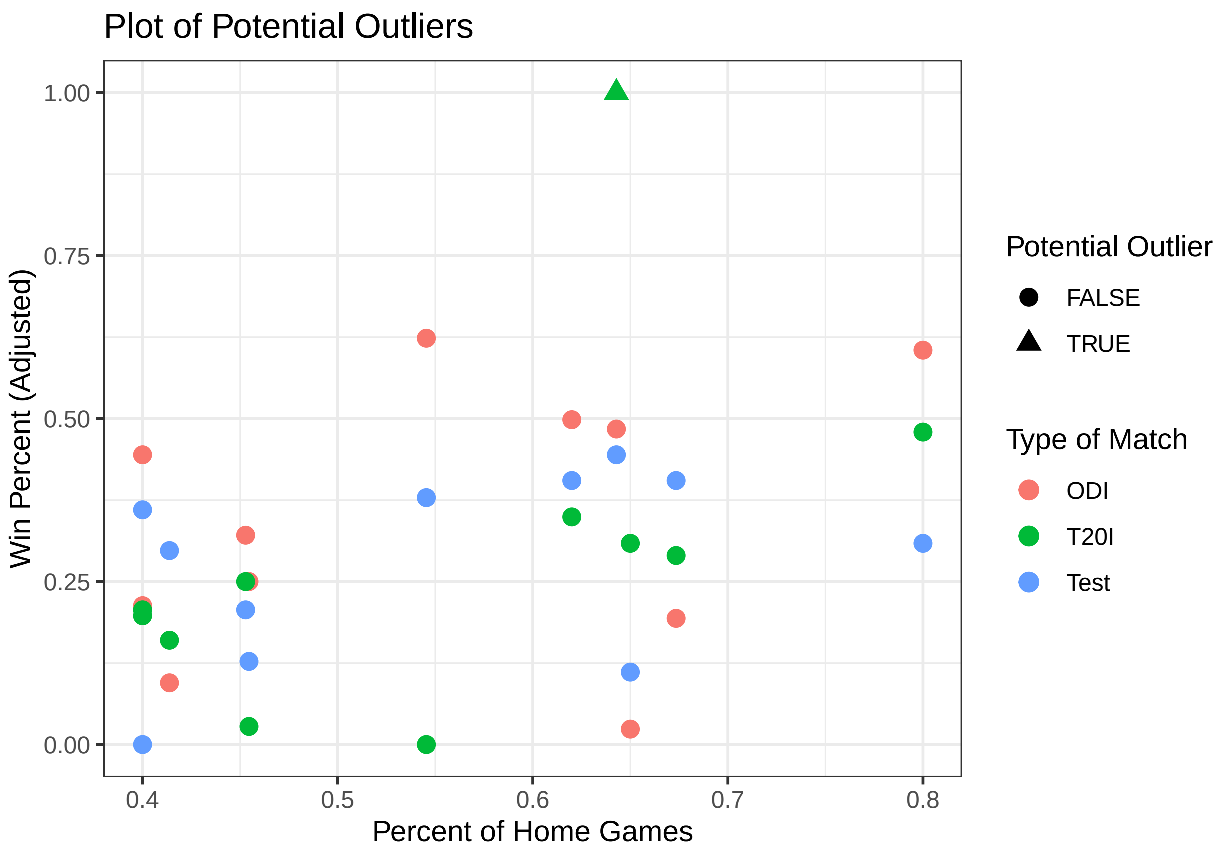

Figure 6: Outliers

Potential Outliers

While checking for outliers we see that we have only one point in Figure 6 that is a potential outlier. In order to determine if it is a true outlier or not we will have to think about why it was marked as one, if it was a data entry error, or a result of the transformation. First, to rule out the data entry, we re-ran the code, and re-gathered all the data used. Now, thinking about the transformation, this data point just so happened to be the only team that went undefeated in the entire sample we chose. Due to our choice to transform the data by squaring it, we in turn made every other value smaller, except for this one. Due to this scenario we will decide to keep it in, as we also have teams that failed to win even a single match.

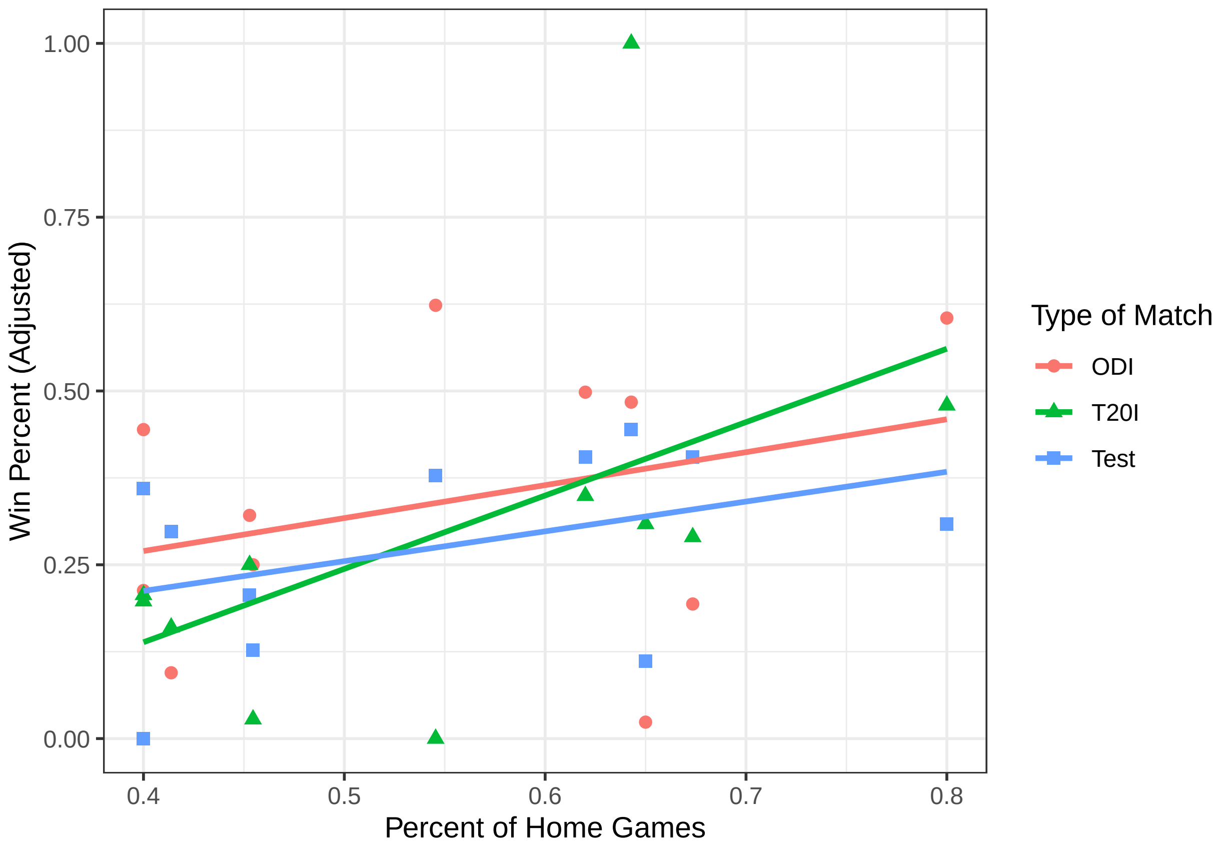

Figure 7: Homogeneity of Slopes

Homegeneity of Slopes

Continuing onto the homogeneity of slopes assumption, we see that there is a subtle violation of the slopes, this is due to two different things: The outlier as mentioned from Figure 6 and an unaccounted aspect of the type of match, the length of the match. The ODI and Test matches last roughly the same amount of time, whereas T20I take almost half as long from start to finish, this means that a team may have less time to adjust to a crowd, and therefore giving the away team a potential disadvantage.

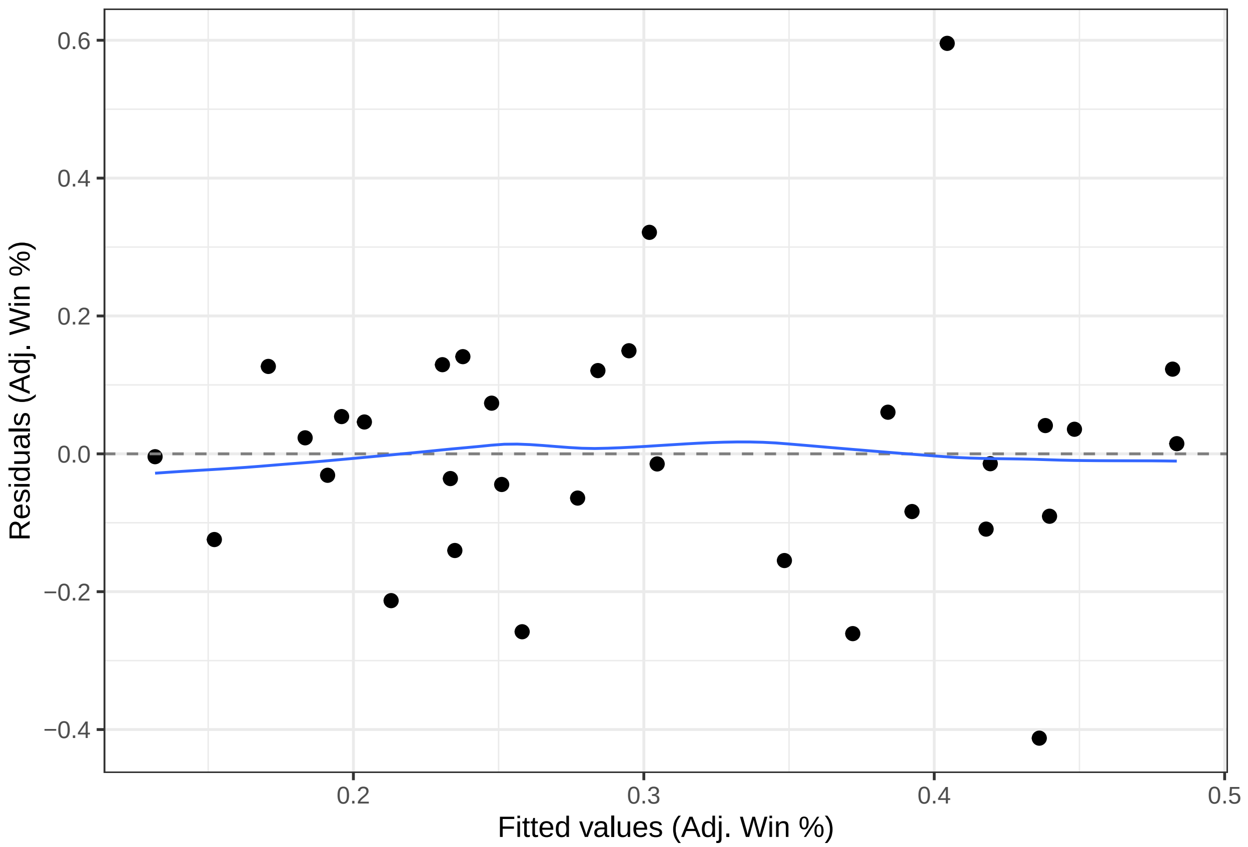

Figure 8: Tukey Plot

Homoscedasticity of Variance

There is really nothing questionable for the homoscedasticity assumption given the nature of the line (see Figure 8). There is also not really any sense of a pattern to the residuals. Given this information we will decide to that this assumption is satisfied.

Independent Observations

In our data we do not have any distinct measurement order aside from the years in which the games were played. This forces us to think about this assumption in the context of the data. Given that we are dealing with a professional sports team, each year we are able to assume that a team is very different from the team they were the year before, much like how an NFL team could win the Superbowl one year, then be the worst team in the league the following year. Thinking about our data in this context, we will decide that this assumption is met.

Results

Omnibus Results

Table 2: ANOVA Table for Cricket Home Games Study

Source	SS	df	MS	F	p-value	Partial Eta Sq.	Partial Omega Sq.	Partial Epsilon Sq.
Year	0.0240	1	0.0240	0.6550	0.4251	0.0229	0.0000	0.0000
Type of Cricket Match	0.0237	2	0.0118	0.3233	0.7264	0.0226	0.0000	0.0000
Percent of Games at Home	0.3047	1	0.3047	8.3191	0.0075	0.2291	0.1815	0.2015
Residuals	1.0257	28	0.0366					

The percent of games played at home accounts of about 8.3 times as much variation in win percent as our residuals/what's left unexplained (F (1, 28) = 11.64), even when we account for what year and what type of match was played. Under the null hypothesis that the proportion of home games has no impact on the proportion of games won, we would only anticipate such a result less than 1% of the time (p = 0.0075). The home game proportion accounts for around 18% of the variation in win proportion (see Table 2), which in the context of a professional sport is very substantial.

Table 3: Marginal Means-Tukey 90% Adjustment

Match Type	Marginal Mean	SE	DF	Lower Bound	Upper Bound
ODI	0.3410	0.0577	28	0.2428	0.4392
T20I	0.2972	0.0577	28	0.1990	0.3953
Test	0.2768	0.0577	28	0.1786	0.3749

After accounting for the impact of the proportion of home games, none of the match type treatment groups accumulated a total proportion of wins that was greater than the size of that treatment group (see Table 3). This means that the effect that match type had on win proportion was negligible at best.

Post-Hoc Analysis

Table 4: Marginal Means-Tukey

Comparison	Difference	SE	DF	t Statistic	p-value
ODI - T20I	0.0438	0.0816	28	0.5369	0.8539
ODI - Test	0.0642	0.0816	28	0.7869	0.7140
T20I - Test	0.0204	0.0816	28	0.2500	0.9662

Now looking at Table 4 we quickly notice that none of the contrasts are statistically different from one another in the context of win proportion from year to year. This is important considering we are interested in the effect that home proportion plays, and not the effect that type of match has.

Table 5: Effect Sizes for Match Type

Keyboard Comparison	Cohen's d	Probability of Superiority
(ODI - T20I)	0.229	0.564
(ODI - Test)	0.336	0.594
(T20I - Test)	0.107	0.530

Again very similar to what we saw in Table 4, we see that the different types of match are very similar to one another, which again only furthers our evidence that the variation in win proportion comes from the proportion of games played at home.

```r
# This template file is based off of a template created by Alex Hayes
# https://github.com/alexpghayes/rmarkdown_homework_template

# Setting Document Options
knitr::opts_chunk$set(
  cache = TRUE,
  echo = FALSE,
  warning = FALSE,
  message = FALSE,
  fig.align = "center"
)

# Add additional packages by name to the following list
packages <- c("tidyverse", "knitr", "rstatix","kableExtra")
lapply(
  X = packages,
  FUN = library,
  character.only = TRUE
)

# Loading Helper Files and Setting Global Options ----
options(knitr.kable.NA = "")
options("contrasts" = c("contr.sum", "contr.poly"))
source("https://raw.github.com/neilhatfield/STAT461/master/rScripts/ANOVATools.R")

source("https://raw.github.com/neilhatfield/STAT461/master/rScripts/shadowgram.R")


# Australia Only Cricket Data ---
cricketData <- read.csv('cricketData.csv')

# Change names for fucntionality
names(cricketData) <- c('Year', 'Location', 'Outcome', 'Type')

cricketData <- cricketData %>%
  mutate(Outcome = ifelse((Outcome=='Win'), "Win", "Non-win"))

# Mutate Home/Away Ratio
cricketData <- cricketData %>%
  group_by(Year) %>%
  mutate(homeRatio = (sum(Location == 'Home')/((sum(Location == 'Away')) + sum(Location == 'Home'))))

# Mutate Win percent ---
cricketData <- cricketData %>%
  group_by(Year, Type) %>%
  mutate(winPercent = (sum(Outcome == 'Win')/((sum(Outcome == 'Win')) + sum(Outcome == 'Non-win'))))



# Convert to Factor ---
cricketData$Location <- as.factor(cricketData$Location)
cricketData$Outcome <- as.factor(cricketData$Outcome)
cricketData$Type <- as.factor(cricketData$Type)
```

```
# Attempt Transformation ---
cricketData$winTransformed <- (cricketData$winPercent)^2

cricketData <- cricketData %>%
  dplyr::select(Year, Type, homeRatio, winPercent, winTransformed) %>%
  distinct()


modelLabels <- c("1 Grand Mean 1", "1 Year 1", "cov Home percentage 1", "3 Game format 2", "33 (Observa

modelMatrix <- matrix(

  data = c(FALSE, FALSE, FALSE, FALSE, FALSE, TRUE, FALSE, FALSE, FALSE, FALSE, TRUE, FALSE, FALSE, FALS

  nrow = 5,

  ncol = 5,

  byrow = FALSE

)

hasseDiagram::hasse(

 data = modelMatrix,

 labels = modelLabels

)
# Descriptive statistics on win % by Type ----
winStats <- psych::describeBy(
  x = cricketData$winPercent,
  group = cricketData$Type,
  na.rm = TRUE,
  skew = TRUE,
  ranges = TRUE,
  quant = c(0.25, 0.75),
  IQR = FALSE,
  mat = TRUE,
  digits = 4
)
winStats %>%
  tibble::remove_rownames() %>%
  tibble::column_to_rownames(
    var = "group1"
  ) %>%
  dplyr::select(
    n, min, Q0.25, median, Q0.75, max, mad, mean, sd, skew, kurtosis
  ) %>%
  knitr::kable(
    caption = "Summary Statistics for Win Rate by Type of Match",
    digits = 3,
    format.args = list(big.mark = ","),
    align = rep('c', 11),
    col.names = c("n", "Min", "Q1", "Median", "Q3", "Max", "MAD", "SAM", "SASD",
                  "Sample Skew", "Sample Ex. Kurtosis")
  )  %>%
  kableExtra::kable_styling(
    font_size = 12,
    latex_options = c("scale_down", "HOLD_position")
  )

# Create a shadowgram of the distances ----
shadowgram(
  dataVec = cricketData$winPercent *100,
  label = "Win Rate x 100",
  layers = 50,
  aStep = 4,
  color = "blue"
)
# Define AOV for cricketModel ---
cricketModel <- aov(
formula = winTransformed ~ Year + Type + homeRatio,
data = cricketData,
na.action = "na.omit"
)
# Model for homogeneity ---
cricketModel2 <- aov(
formula = winTransformed ~ Year + Type*homeRatio,
data = cricketData,
na.action = "na.omit"
)

cricketModel3 <- aov(
formula = winPercent ~ Year + Type + homeRatio,
data = cricketData,
na.action = "na.omit"
)

# QQ plot ----
car::qqPlot(
x = residuals(cricketModel3),
distribution = "norm",
envelope = 0.90,
id = FALSE,
pch = 20,
ylab = "Residuals (Win %)"
)
# QQ plot ----
car::qqPlot(
x = residuals(cricketModel),
distribution = "norm",
envelope = 0.90,
id = FALSE,
pch = 20,
```

```r
ylab = "Residuals (Win % Adj.)"
)
ggplot(
data = cricketData,
mapping = aes(
y = winTransformed,
x = homeRatio
)
) +
geom_point(size = 2) +
geom_smooth( # Adds a smoother function's graph
inherit.aes = FALSE,
mapping = aes(x = homeRatio, y = winTransformed),
method = "lm", # Fit a Linear Model
formula = y ~ x, # Specifies the form of the "linear" model
color = "black",
linetype = "dashed",
se = FALSE
) +
theme_bw() +
xlab("Percent of Home Games") +
ylab("Win Percent (Adjusted)")
## Step 1: send the data through the Mahalanobis function
outlierDetection <- rstatix::mahalanobis_distance(cricketData)
## Step 2: OPTIONAL--reattach the factor
outlierDetection <- cbind(
outlierDetection,
factor = cricketData$Type
)
## Step 3: Make a scatter plot
ggplot(
data = outlierDetection,
mapping = aes(
y = winTransformed,
x = homeRatio,
shape = is.outlier,
color = factor
)
) +
geom_point(size = 3) +
theme_bw() +
ggtitle("Plot of Potential Outliers")+
xlab("Percent of Home Games") +
ylab("Win Percent (Adjusted)") +
labs(
color = "Type of Match",
shape = "Potential Outlier"
)
# Demo Code for Assessing Homogeneity of Slopes in Keyboarding Pain Study ----
ggplot(
data = cricketData,
mapping = aes(
y = winTransformed,
x = homeRatio,
color = Type,
shape = Type
)
) +
geom_point(size = 2) +
geom_smooth(
method = "lm", # See notes below
mapping = aes(y = predict(cricketModel2)),
formula = y ~ x,
se = FALSE
) +
theme_bw() +
xlab("Percent of Home Games") +
ylab("Win Percent (Adjusted)") +
labs(
color = "Type of Match",
shape = "Type of Match"
)
# Generate the Tukey-Anscombe plot ----
ggplot(
  data = data.frame(
    residuals = residuals(cricketModel),
    fitted = fitted.values(cricketModel)
  ),
  mapping = aes(x = fitted, y = residuals)
) +
  geom_point(size = 2) +
  geom_hline(
    yintercept = 0,
    linetype = "dashed",
    color = "grey50"
  ) +
  geom_smooth(
    formula = y ~ x,
    method = stats::loess,
    method.args = list(degree = 1),
    se = FALSE,
    linewidth = 0.5
  ) +
  theme_bw() +
  xlab("Fitted values (Adj. Win %)") +
  ylab("Residuals (Adj. Win %)")
# Omnibus Test/Modern ANCOVA Table ----
parameters::model_parameters(
model = cricketModel,
effectsize_type = c("eta", "omega", "epsilon")
) %>%
dplyr::mutate( #Fixing the Parameter (Source) Column's values
Parameter = dplyr::case_when(
Parameter == "homeRatio" ~ "Percent of Games at Home",
Parameter == "Type" ~ "Type of Cricket Match",
Parameter == "Type:homeRatio" ~ "Type:Home Interaction",
TRUE ~ Parameter
)
) %>%
dplyr::mutate(
p = ifelse(
test = is.na(p),
yes = NA,
no = pvalRound(p)
)
) %>%
knitr::kable(
digits = 4,
col.names = c("Source", "SS", "df", "MS", "F", "p-value",
"Partial Eta Sq.", "Partial Omega Sq.", "Partial Epsilon Sq."),
caption = "ANOVA Table for Cricket Home Games Study",
align = c('l',rep('c',8)),
booktab = TRUE
) %>%
kableExtra::kable_styling(
bootstrap_options = c("striped", "condensed"),
font_size = 12,
latex_options = c("scale_down", "HOLD_position")
)
## Type
emmOutKey <- emmeans::emmeans(
object = cricketModel,
specs = pairwise ~ Type,
adjust = "tukey",
level = 0.9
)
## Point Estimates
as.data.frame(emmOutKey$emmeans) %>%
knitr::kable(
digits = 4,
col.names = c("Match Type", "Marginal Mean","SE", "DF",
"Lower Bound","Upper Bound"),
caption = "Marginal Means-Tukey 90\\% Adjustment",
align = c("l", rep("c", 5)),
booktabs = TRUE
) %>%
kableExtra::kable_styling(
bootstrap_options = c("striped", "condensed"),
font_size = 12,
latex_options = c("HOLD_position")
)
as.data.frame(emmOutKey$contrasts) %>%
knitr::kable(
digits = 4,
col.names = c("Comparison", "Difference","SE", "DF",
"t Statistic","p-value"),
caption = "Marginal Means-Tukey",
align = c("l", rep("c", 5)),
booktabs = TRUE
```

```
) %>%
kableExtra::kable_styling(
bootstrap_options = c("striped", "condensed"),
font_size = 12,
latex_options = c("HOLD_position")
)
as.data.frame(
emmeans::eff_size(
object = emmOutKey,
sigma = sigma(cricketModel),
edf = df.residual(cricketModel)
)
) %>%
dplyr::mutate(
ps = probSup(effect.size),
.after = effect.size
) %>%
dplyr::select(contrast, effect.size, ps) %>%
knitr::kable(
digits = 3,
col.names = c("Keyboard Comparison", "Cohen's d", "Probability of Superiority"),
align = "lccc",
caption = "Effect Sizes for Match Type",
booktab = TRUE
) %>%
kableExtra::kable_styling(
bootstrap_options = c("striped", "condensed"),
font_size = 12,
latex_options = "HOLD_position"
)
```