August 2021 MS Comprehensive Exam

Part IV: Report

GID9475

8/15/2021

Part IV:

The long jump is a sport is where athletes attempt the longest jump from a pre-specified starting point. The dataset `longjump_final` contains a subset of longest jumps recorded by female athletes. The dataset contains all jumps of at least 6.55 meters.

Using this dataset, explore the impact of windspeed, recorded in meters / second, on jump distance. To facilitate this analysis, the dataset is filtered to only include athletes with multiple jumps of at least 6.55 meters.

Present your findings as a statistical report comprised of the following four sections:

1. Introduction
2. Data
3. Statistical Procedures
4. Results

All items are graded on a 1 - 4 scale using the following scheme, but note some elements are multiples of four:

1. **No Credit**: Criterion was not addressed or was written in a way that was not understandable.
2. **Beginning**: Ideas are not clear and supporting ideas are not presented.
3. **Developing**: Ideas are identified but not well supported and developed or are minimally supported and developed.
4. **Advanced**: Ideas are clearly identified and are adequately supported and developed.

Report generalities	**Points**
Spelling, grammar, writing clarity	12
Paragraphs, section labels	4
Length, Double spaced	4
Appendix with complete code	4
Figures (reporting and labeling) included in text	4
Tables, if any, (reporting and labeling) included in text	4
Citations for papers and packages used	4
Acknowledgments for other resources	4
Proper use of statistical methods	12

Introduction	**Points**
Report motivation	4
Sample size(s)	4
Data source and study design	4

Introduction	Points
Research question	4

Data	Points
Variables with units and descriptive statistics	4
Data Visualization	8

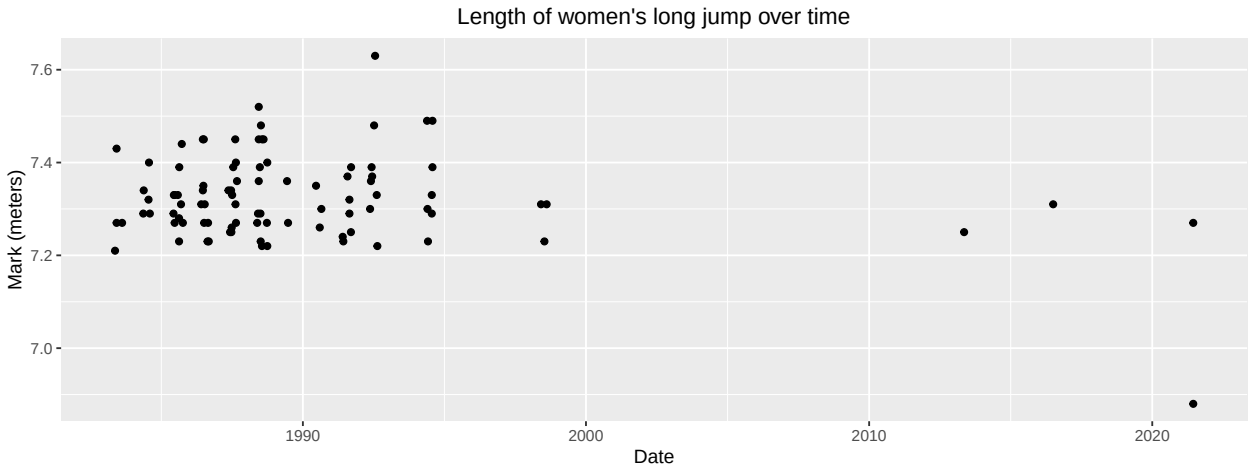
Statistical Procedures	Points
Describe model to fit and include justification	8
Discuss/assess model assumptions	8
Repeated measurements on individuals	8
Residuals: visual exploration	4

Results	Points
Summarize and interpret "size" of parameter estimates	4
Interpret intervals for parameter estimates	4
Interpret prediction intervals for at least one athlete across a range of windspeeds	4
Scope of Inference: how can the results be generalized?	4

Introduction

The women's long jump records have stood for decades, and hasn't even been approached in recent competitions. The record 7.63 meters was set in 1992 by Heike Drechsler in Italy. She also has 39 more of the longest women's long jumps ever recorded, more than double anyone else. That record was set on a day when the wind was among the highest it's ever been during a competition - 2.3 meters per second. How much does the wind affect the distance of in top women's long jump competitions? To answer this, we have observational data on 91 of the longest women's long jumps ever recorded, filtered to only include jumps over 6.55 meters and only competitors that appear at least twice in the data set.

The plot below shows some of the history of the women's long jump. Over the past 40 years, there has not been a strong trend towards increasing distance, unlike some other track and field events like the marathon.



The women's long jump records seem to be stagnant over the past 40 years.

Data

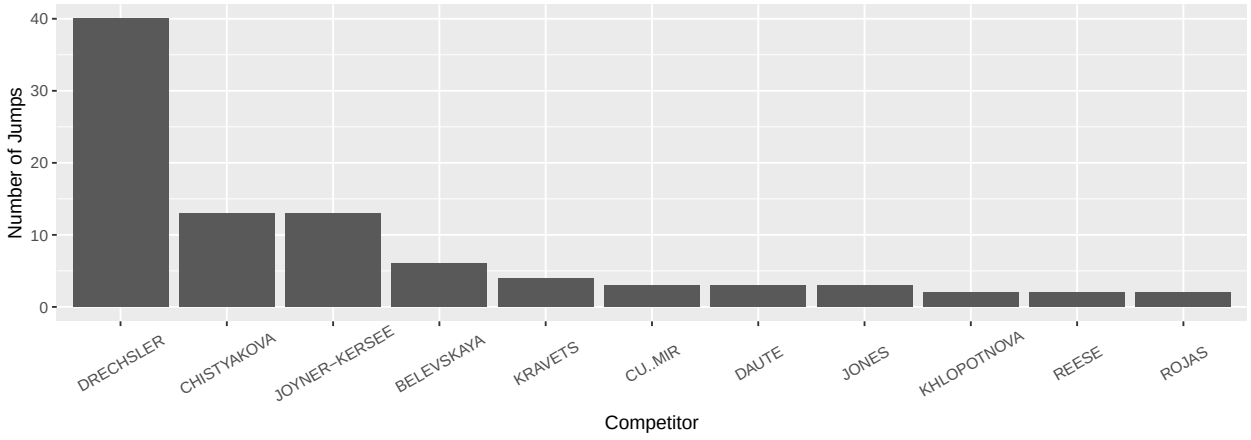
Due to the top competitors having many of the longest jumps ever, over 2/3rds of the observations were performed by the same 3 women - Heike Drechsler, Jackie Joyner-Kersee, and Galina Chistyakova. This repeated measurements design necessitates something more complicated than a simple linear regression to estimate the effect of the wind on top long jump competitors.

The two tables below summarize the means and standard deviations of relevant variables overall and by competitor.

	$Mean \pm SD$
Mark (meters)	7.32 ± 0.09
Wind speed (m/s)	1.19 ± 0.99

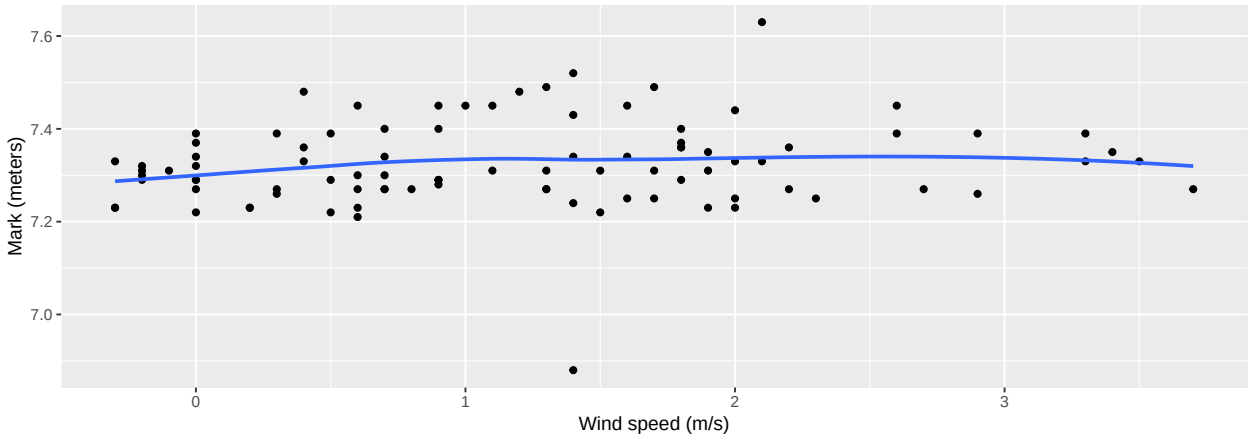
Competitor	**Mark (m)**	**# Jumps**
Joyner-Kersee	7.36 ± 0.09	13
Chistyakova	7.34 ± 0.08	13
Drechsler	7.33 ± 0.09	40
Belevskya	7.34 ± 0.05	6
Cusmir	7.30 ± 0.11	3
Daute	7.30 ± 0.04	3
Khlopotnova	7.30 ± 0.01	2
Jones	7.28 ± 0.05	3
Reese	7.28 ± 0.04	2
Kravets	7.27 ± 0.07	4
Rojas	7.07 ± 0.27	2

Figure 1: Number of long jumps over 6.55 meters



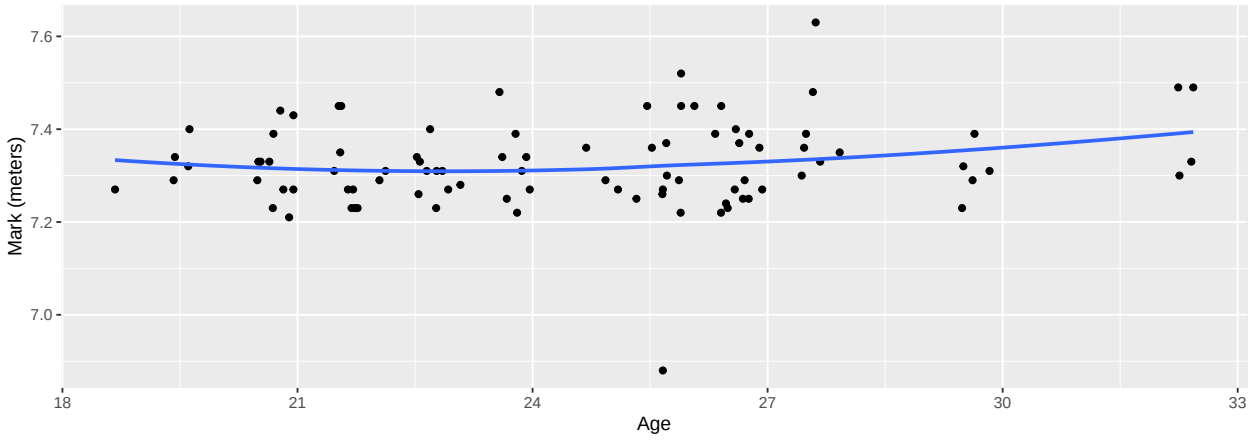
Apparently Heike Drechsler was alright at the long jump.

Figure 2: Effect of wind on women's long jump distance



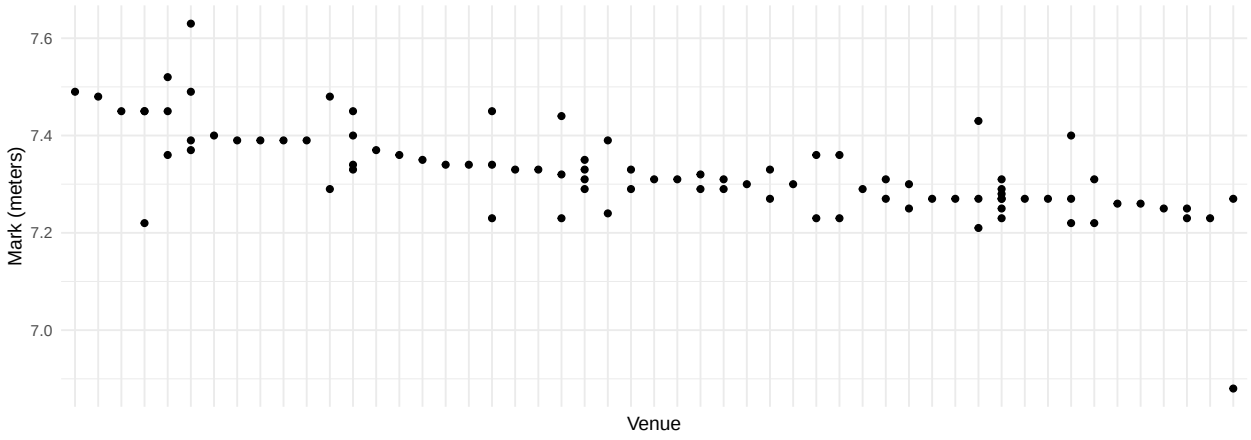
Not a strong effect due to the wind, on the scale of 1/20th of a meter.

Figure 3: Women's long jump distances by age



There is a slightly increasing trend in long jump distances as age increases.

Figure 4: Women's long jump distances by venue



Venues don't seem to be re−used often enough to know if there's a strong association between venue and distance.

The four plots above show some of the potential variables to explore as covariates. Although the goal is to measure the effect of wind on the distance, being able to control for some of these factors is necessary to measuring that effect accurately. Figure 1 shows the number of competitors in the data and how often each of them appear. It implies that the model should account for the repeated measurements on the same person - we can't assume that the observations are independent. Figure 2 shows the relationship of interest. There is not much effect of the wind on the jump distance before controlling for some of the other covariates. Figure 3 show a possible covariate of age, since it would be expected that a competitor isn't jumping the

same distances over their enitre careers. Figure 4 shows a possible confounding variable in venue, because some venues could be windier than others, and some competitors could prefer some venues to others. Some venues are re-used, which also should be accounted for in the model.

Statistical Procedures

These data should be modeled with a multilevel model to account for the repeated measurements of the same competitor and venue. The multilevel model structure will allow each of the competitors and venues to each have their own intercept term as a random effect, while keeping the wind speed effect constant for each competitor, assuming that the wind affects each competitor the same fixed amount. The model can be specified as:

$y_i = \alpha_{j[i]} + \gamma_{j[i]} + \beta_0 + \beta_1 * Wind\ Speed + \epsilon_i, \epsilon_i \sim N(0, \sigma_y^2)$ where y_i is the i^{th} long jump observation in the data and j refers to the subject-level grouping.

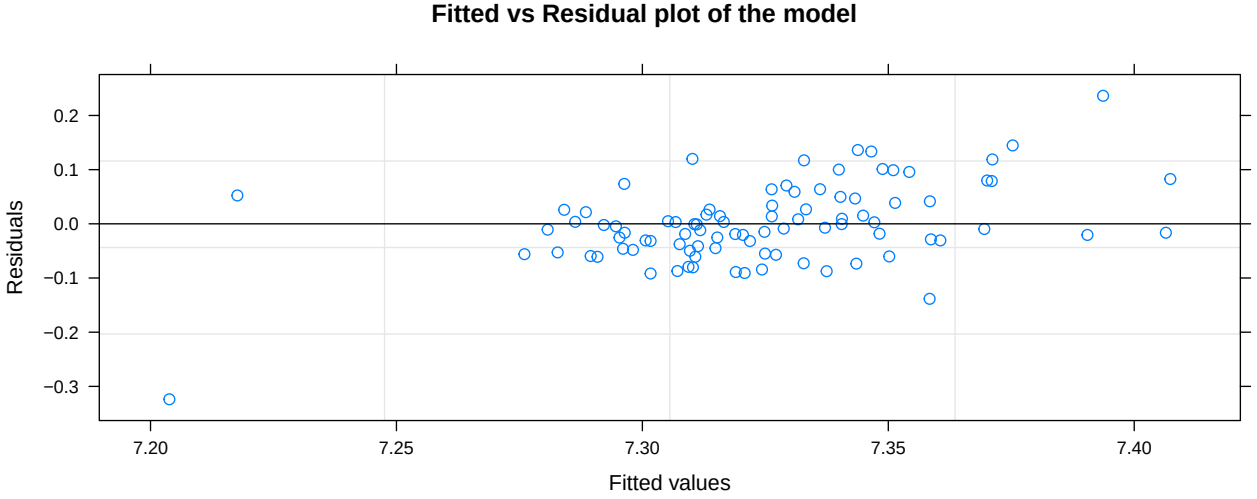
$\alpha_j \sim N(\mu_0, \sigma_\alpha^2)$ where α_j is the estimated intercept of the long jump distance for each competitor j with mean μ_0.

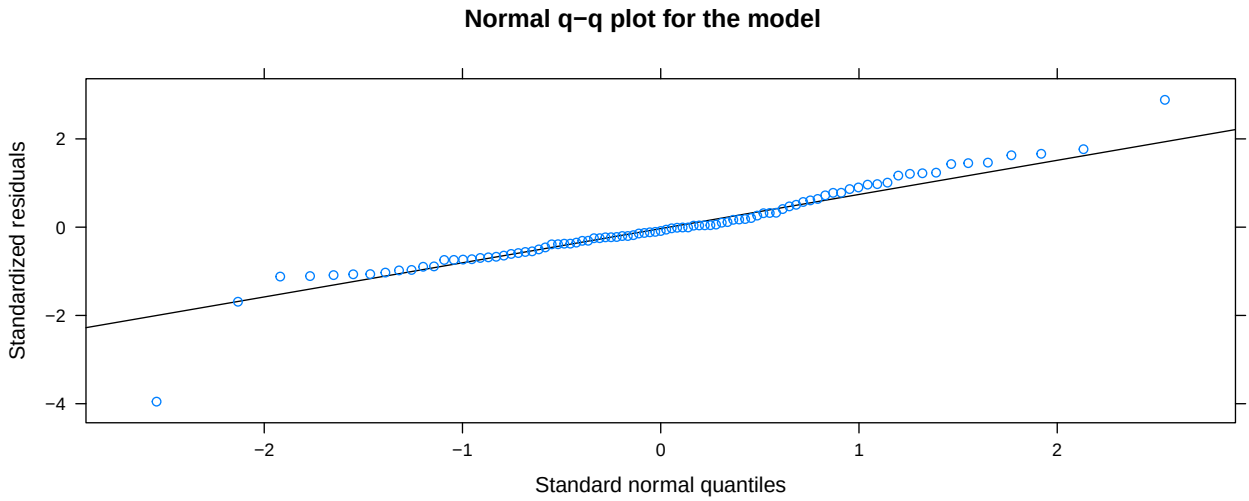
$\gamma_j \sim N(\nu_0, \sigma_\gamma^2)$ where γ_j is the estimated intercept for each venue j with average long jump distance of ν_0.

β_0 is the estimated fixed intercept coefficient of the jump distance.

β_1 is the estimated fixed slope coefficient of the wind speed on the jump distance. β_0 and β_1 are the same for all competitors and venues.

We are assuming with this model that there are underlying normally distributed effects for each competitor and venue, each with a different mean and variance. We assume that the wind speed affects every competitor the same fixed amount. We also assume normality and homoskedastic variances of the residuals and linearity of the model. The plots below show the fitted vs residuals plot and the normal q-q plot of the residuals. To evaluate if there is homoskedasticity and linearity in the model, we look at the fitted vs residuals plot to look for curvature or a fanning pattern. There is not any evidence of a lack of heteroskedasticity or non-linearity based on the plot. The normal q-q plot plots the residuals of the observed data against the quantiles of a normal distribution to see how well the residuals are normally distributed. In the second plot below, the residuals have a right skew, and are not necessarily well approximated by a normal distribution. Modeling the log or square root of the distance doesn't help much, so I left the model in the standard scale. The estimates could be affected, and future work could find a better functional form for the data to model it better.

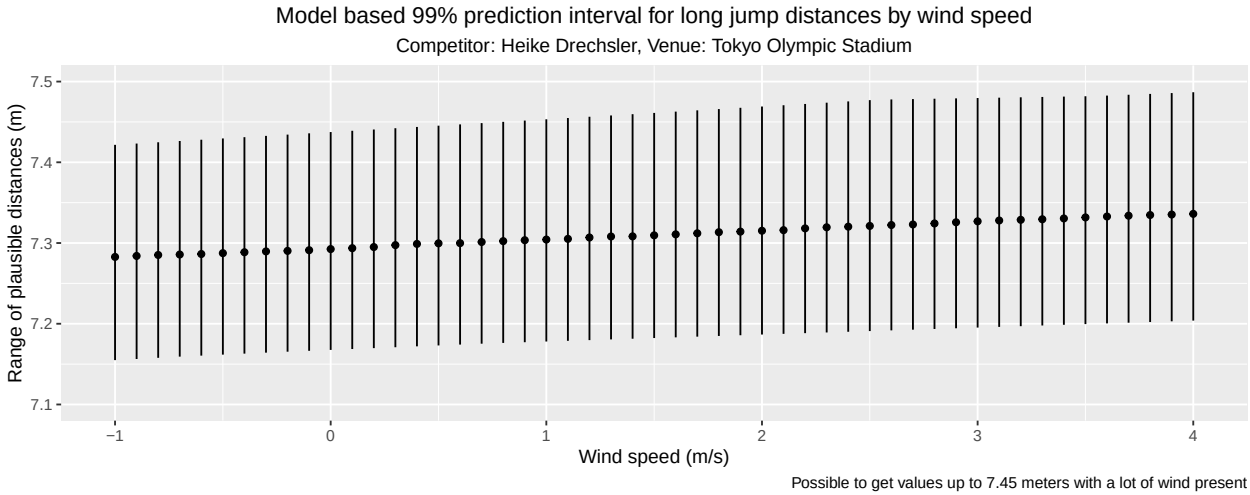




Results

The estimated average jump distance when controlling for the venue and competitor is 7.30 ± 0.020. This would be the best estimate for a new competitor at a new venue on a day with no wind based on these data, for a similar observational unit to the ones in the sample. There is 95% confidence that the interval (7.26, 7.34) contains the true unknown value for this intercept term.

The estimate for the effect of a 1 m/s increase in wind speed on the distance of the women's long jump is 0.011 ± 0.0096 meters when accounting for the venue and competitor. This estimate controls for the the repeated measurement design where up to 40 of the observations were of the same competitor, and several of the same venues were used for multiple jumps. This gives an approximate 95% confidence interval for the effect of wind speed of (-0.0087, 0.030). We have 95% confidence that the true unknown value of the slope is somewhere in that interval. If we could go recollect similar but new data and do the same modeling process many times, 95% of the calculated confidence intervals from that process would contain the true unknown value for the slope, but we don't know if this one does or not. This gives a range of plausible values that the effect of wind on the jumping distance could be.



The figure above shows a 99% prediction interval for Heike Drechsler at Tokyo Olympic Stadium across a range of wind speeds. If there is no wind, the 99% prediction interval contains values between 7.19 and 7.40 meters. If there is 4 m/s of wind, the prediction interval contains values between 7.22 and 7.47 meters. These prediction intervals give a range of plausible values based on the model. If we repeated getting measurements on this competitor with these wind speeds and at this venue many times, 99% of the calculated prediction

intervals would contain the observed data values between the upper and lower bounds.

This model does not account for any of the factors that would adjust a particular competitor's average jump distance - one that was examined in the exploratory data analysis is the competitor's age. Future work could include this sort of analysis, but more data would be necessary. These models struggle with overspecification for any of the more complicated versions, even when not including venue as a random effect. When trying to include the age of the competitor as a random effect with the data provided, the results are not reliable.

We shouldn't really care about any of the random effect parameter estimates, since they're not of interest for this experiment, but they're summarized in Appendix A. None of the random effects are larger than 1/18th of a meter in size.

We can apply the results of this analysis only to similar world-class women's long jump competitors. If a new observation is one of the 11 competitors in the data set, the best estimate will use the random effects for that participant in addition to the intercept and wind speed, and will use the venue estimate if they're using one of the 51 venues in the data set. If any of the random effects information is not available, only the intercept and wind speed will be used to estimate the jump distance.

Appendix A: Supplementary tables and plots

Table 8: Random effects estimate for each competitor

	Intercept
Jackie JOYNER-KERSEE	0.037
Galina CHISTYAKOVA	0.022
Heike DRECHSLER	0.020
Yelena BELEVSKAYA	0.012
Anișoara CUȘMIR	-0.001
Yelena KHLOPOTNOVA	-0.001
Heike DAUTE	-0.005
Marion JONES	-0.006
Brittney REESE	-0.009
Inessa KRAVETS	-0.015
Yulimar ROJAS	-0.055

Table 9: 10 largest random effects estimates of venue

	Intercept
Sestriere (ITA)	0.055
La Nucia (ESP)	-0.053
Leningrad (URS)	0.042
Moskva (URS)	-0.032
Stade Olympique de la Pontaise, Lausanne (SUI)	0.026
Bruxelles (BEL)	-0.024
Dresden (GER)	0.023
New York, NY (USA)	0.023
San José (USA)	-0.021
Budapest (HUN)	0.019

Appendix B: Citations

Papers and textbooks: Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. Journal of memory and language, 68(3), 255-278.

Gelman, A., & Hill, J. (2006). Data Analysis Using Regression and Multilevel/Hierarchical Models (Analytical Methods for Social Research). Cambridge: Cambridge University Press. doi:10.1017/CBO9780511790942

Gelman, A., Hill, J., & Vehtari, A. (2020). Regression and Other Stories (Analytical Methods for Social Research). Cambridge: Cambridge University Press. doi:10.1017/9781139161879

R packages: Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686,
https://doi.org/10.21105/joss.01686

Yihui Xie (2020). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.30.

Yihui Xie (2015) Dynamic Documents with R and knitr. 2nd edition. Chapman and Hall/CRC. ISBN 978-1498716963

Yihui Xie (2014) knitr: A Comprehensive Tool for Reproducible Research in R. In Victoria Stodden, Friedrich Leisch and Roger D. Peng, editors, Implementing Reproducible Computational Research. Chapman and Hall/CRC. ISBN 978-1466561595

Garrett Grolemund, Hadley Wickham (2011). Dates and Times Made Easy with lubridate. Journal of Statistical Software, 40(3), 1-25. URL https://www.jstatsoft.org/v40/i03/.

Hadley Wickham (2007). Reshaping Data with the reshape Package. Journal of Statistical Software, 21(12), 1-20. URL http://www.jstatsoft.org/v21/i12/.

Douglas Bates, Martin Maechler, Ben Bolker, Steve Walker (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1-48. doi:10.18637/jss.v067.i01.

Websites: R's lmer cheat sheet: https://stats.stackexchange.com/questions/13166/rs-lmer-cheat-sheet

Creating bootstrap prediction intervals in lme4: https://cran.r-project.org/web/packages/merTools/vignettes/Using_predictInterval.html

Appendix C: Full R code

```
knitr::opts_chunk$set(echo = FALSE, fig.height=4, fig.width=10)
options(show.signif.stars = FALSE)
library(tidyverse)
library(knitr)
library(lubridate)
library(reshape2)
library(lme4)
options(mc.cores = parallel::detectCores())

load('longjump.RData')
tibble(longjump_final) %>% mutate(DOB = lubridate::dmy(DOB), Date = dmy(Date)) -> longjump

ggplot(data = longjump) +
  geom_point(aes(x = Date, y = Mark)) +
  labs(title = "Length of women's long jump over time",
       caption = "The women's long jump records seem to be stagnant over the past 40 years.",
```

```
       y = "Mark (meters)") +
  theme(plot.title = element_text(hjust = 0.5))

Overall_mean = longjump %>% summarize(meanMark = mean(Mark) %>% round(2),
                                      sdMark = sd(Mark) %>% round(2),
                                      meanWind = mean(Wind) %>% round(2),
                                      sdWind = sd(Wind) %>% round(2))

desc_stats = longjump %>%
  group_by(Competitor) %>%
  summarize(meanMark = mean(Mark) %>% round(2),
            sdMark = sd(Mark) %>% round(2),
            numJumps = n())
longjump %>%
  count(Competitor) %>%
  mutate(Competitor = stringr::word(Competitor, -1)) %>%
  mutate(Competitor = fct_reorder(Competitor, n, .desc = TRUE)) %>%
  ggplot(data = ., aes(x = Competitor, y = n)) +
    geom_bar(stat = "identity") +
    labs(y = "Number of Jumps",
         title = "Figure 1: Number of long jumps over 6.55 meters",
         caption = "Apparently Heike Drechsler was alright at the long jump.") +
    theme(axis.text.x = element_text(angle = 30, vjust = 0.6),
          plot.title = element_text(hjust = 0.5))

longjump %>%
  ggplot(data = ., aes(x = Wind, y = Mark)) +
  geom_point() +
  geom_smooth(se = F, formula = y~x, method = "loess") +
  labs(x = "Wind speed (m/s)",
       y = "Mark (meters)",
       title = "Figure 2: Effect of wind on women's long jump distance",
       caption = "Not a strong effect due to the wind, on the scale of 1/20th of a meter.") +
  theme(plot.title = element_text(hjust = 0.5))

longjump %>%
  mutate(Age = as.numeric(Date - DOB)/365) %>%
  ggplot(data = ., aes(x = Age, y = Mark)) +
  geom_point() +
  geom_smooth(se = F, formula = y~x, method = "loess", span = 1) +
  labs(title = "Figure 3: Women's long jump distances by age",
       x = "Age",
       y = "Mark (meters)",
       caption = "There is a slightly increasing trend in long jump distances as age increases.") +
  theme(plot.title = element_text(hjust = 0.5))

longjump %>%
  mutate(Venue = fct_reorder(Venue, Mark, .fun = "median", .desc = TRUE)) %>%
  ggplot(data = ., aes(x = Venue, y = Mark)) +
  geom_point() +
  labs(y = "Mark (meters)",
       title = "Figure 4: Women's long jump distances by venue",
       caption = "Venues don't seem to be re-used often enough to know if there's a strong association
```

```r
  theme_minimal() +
  theme(axis.text.x = element_blank(),
        axis.ticks.x = element_blank(),
        plot.title = element_text(hjust = 0.5))

model = lmer(formula = Mark ~ (1|Competitor) + (1|Venue) + Wind, data = longjump, REML = T)
#summary(model)
#ranef(model)
plot(model, xlab = "Fitted values", ylab = "Residuals", main = "Fitted vs Residual plot of the model")
lattice::qqmath(model, main = "Normal q-q plot for the model")

# prediction intervals
preds = tibble(Competitor = "Heike DRECHSLER", Venue = "Olympic Stadium, Tokyo (JPN)", Wind = seq(-1, 4

# Adapted from https://cran.r-project.org/web/packages/merTools/vignettes/Using_predictInterval.html
mySumm <- function(.) {
  predict(., newdata=preds, re.form=NULL)
}
#### Collapse bootstrap into median, 95% PI
sumBoot <- function(merBoot) {
  return(
    data.frame(fit = apply(merBoot$t, 2, function(x) as.numeric(quantile(x, probs=.5, na.rm=TRUE))),
               lwr = apply(merBoot$t, 2, function(x) as.numeric(quantile(x, probs=.005, na.rm=TRUE))),
               upr = apply(merBoot$t, 2, function(x) as.numeric(quantile(x, probs=.995, na.rm=TRUE)))
    )
  )
}

boot1 <- lme4::bootMer(model, mySumm, nsim=250, use.u=FALSE, type="parametric")
PI.boot1 <- sumBoot(boot1)

ggplot(data = PI.boot1,
       aes(x = preds$Wind,
           y = fit,
           ymin = lwr,
           ymax = upr)) +
  geom_point() +
  geom_linerange() +
  labs(y = "Range of plausible distances (m)",
       x = "Wind speed (m/s)",
       title = "Model based 99% prediction interval for long jump distances by wind speed",
       subtitle = "Competitor: Heike Drechsler, Venue: Tokyo Olympic Stadium",
       caption = "Possible to get values up to 7.45 meters with a lot of wind present") +
  ylim(7.15, 7.5) +
  theme(plot.title = element_text(hjust = 0.5),
        plot.subtitle = element_text(hjust = 0.5))
```