Biost 540: Homework 1

Department of Biostatistics @ University of Washington

Alejandro Hernandez

Due Thursday 11, April 2024

Problem 1

The Iris dataset, which Ronald Fisher presented in his 1936 paper "The use of multiple measurements in taxonomic problems," comprises three species of plants (setosa, virginica, and versicolor) with four measured features for each sample.

```r
# load Iris data
data(iris)

# dim(iris)
# names(iris)
# head(iris)
# table(iris$Species)
```

1a)

Please calculate the number, average petal length, and average petal width for each species.

```r
# summarize average petal measurements for each species
iris %>%
  group_by(Species) %>%
  summarise(
    n = length(Species),
    Average.petal.length = mean(Petal.Length),
    Average.petal.width = mean(Petal.Width)) %>%
  knitr::kable()
```

Species	n	Average.petal.length	Average.petal.width
setosa	50	1.462	0.246
versicolor	50	4.260	1.326
virginica	50	5.552	2.026

1b)

Please draw a scatterplot using ggplot for `Sepal.Length` and `Sepal.Width`. And please assign different colors for different species.

```
# color selection
colors <- c("#FC600A", # dark orange
            "#C21460", # dark pink
            "#3F0000") # darker red

# plot sepal measurements for each species
iris %>%
  ggplot(aes(x=Sepal.Length, y=Sepal.Width, color=Species, fill=Species)) +
    geom_point(size=2) +
    geom_smooth(alpha=0.15, lwd=.7) +
    xlab("Sepal Length") + ylab("Sepal Width") +
    scale_color_manual(values=colors) +
    scale_fill_manual(values=colors) +
    theme_bw()
```

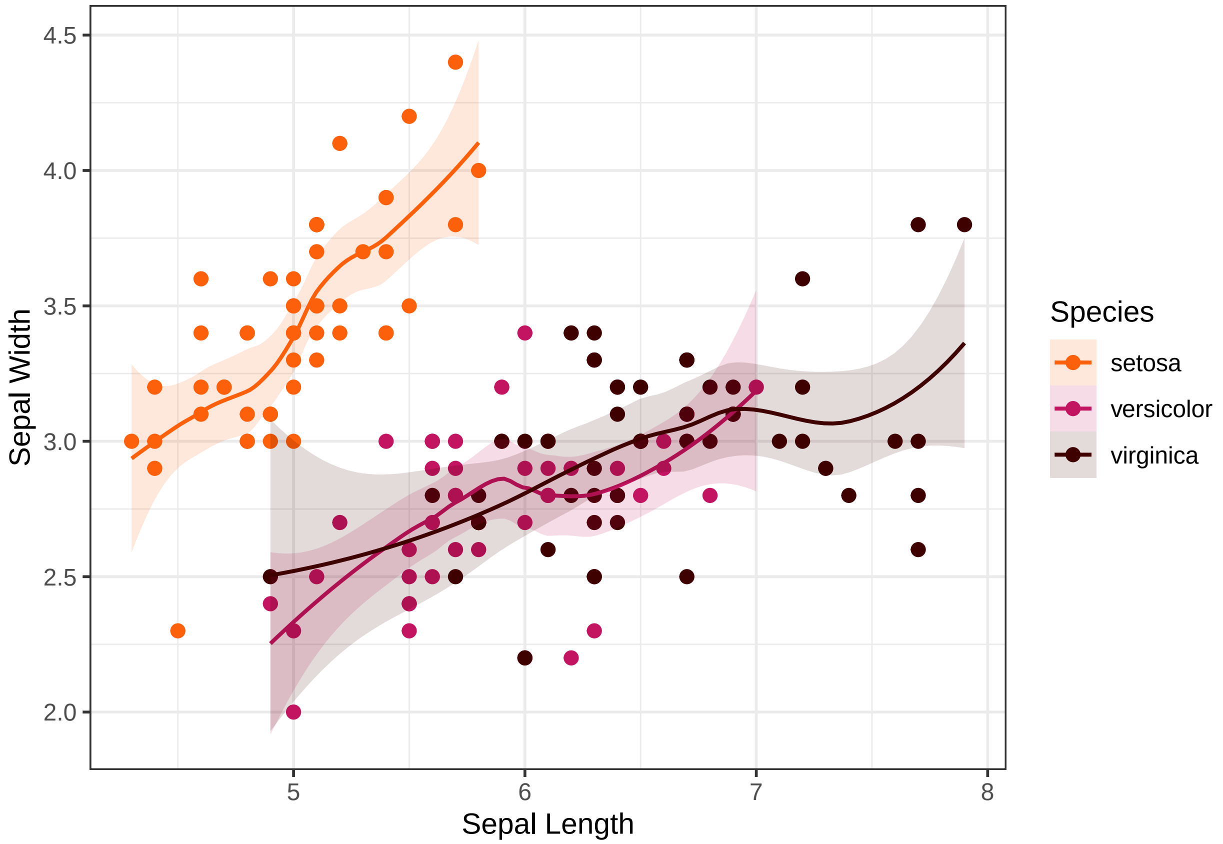


Figure 1: Scatterplot of Sepal Measures by Species

Problem 2

In this problem, we focused on a dataset called `WorldPhones` which summarizes the number of telephones in various regions of the world (in thousands). The data matrix has 7 rows and 8 columns. The columns of the matrix give the figures for a given region, and the rows the figures for a year.

```r
data(WorldPhones)

# dim(WorldPhones)
```

2a)

Please create a new column called `year` using the rownames of the data. Convert the data into long table such that there are 3 columns in total including year, region and number.

```r
# define new `year` column
WorldPhones <- WorldPhones %>%
  data.frame(.) %>%
  mutate(year = rownames(.))

WorldPhones %>%
  pivot_longer(cols = c("N.Amer", "Europe", "Asia", "S.Amer", "Oceania",
                        "Africa", "Mid.Amer"),
               names_to = "region",
               values_to = "n.telephones") %>%
  head(10) %>% knitr::kable()
```

year	region	n.telephones
1951	N.Amer	45939
1951	Europe	21574
1951	Asia	2876
1951	S.Amer	1815
1951	Oceania	1646
1951	Africa	89
1951	Mid.Amer	555
1956	N.Amer	60423
1956	Europe	29990
1956	Asia	4708

Problem 3

The Treatment of Lead-Exposed Children (TLC) trial was a placebo-controlled, randomized study of succimer (a chelating agent) in children with bloodlead levels of 20-44 micrograms/dL. These data consist of four repeated measurements of blood lead levels obtained at baseline (or week 0), week 1, week 4, and week 6 on 100 children who were randomly assigned to chelation treatment with succimer or placebo.

Each row of the data set contains the following 6 variables: ID, Treatment Group, Lead Level Week 0, Lead Level Week 1, Lead Level Week 4, Lead Level Week 6.

```r
# load lead exposure study data
tlc <- read.csv("data/tlc.csv")[,-1]

tlc <- tlc %>%
  mutate(tx = as.factor(tx)) %>%
  rename(week.0=y0,
         week.1=y1,
         week.4=y4,
         week.6=y6)

levels(tlc$tx) <- c("Treatment", "Placebo")
tlc$tx <- factor(tlc$tx, levels=c("Placebo", "Treatment"))
```

3a)

Transform the data into long format and produce spaghetti plots illustrating the progression in lead level. Facet the plot by treatment group; be sure to label axes appropriately with units. Hint: the functions `pivot_longer` (with argument names_prefix) and `facet_wrap` will be useful to you.

```
# transform data to long format
tlc_long <- tlc %>%
  pivot_longer(cols = starts_with("week."),
               names_to = "week",
               names_prefix = "week.",
               values_to = "lead.level") %>%
  mutate_at(c("id", "week"), as.integer)

# plot of lead levels and averages over time, by treatment
gg <- tlc_long %>%
  ggplot(aes(x=week, y=lead.level)) +
  geom_line(aes(group=id, color=tx), alpha=0.4) +
  geom_point(data = tlc_long %>%
               group_by(tx, week) %>%
               summarise(mean=mean(lead.level)),
             aes(x=week, y=mean, color=tx), size=2.5) +
  xlab("Week") + ylab("Lead Level (ug/dL)") +
  scale_color_manual(values=colors[c(1,3)]) +
  theme_bw()

gg                        # treatment groups overlayed
```

```
gg + facet_wrap(vars(tx)) # treatment group separated
```

3b)

Compute the Pearson correlation matrix between outcomes at different time points.

```
# correlation matrix
tlc %>%
  select(starts_with("week.")) %>%
  cor %>%
  knitr::kable(digits = 2)
```

	week.0	week.1	week.4	week.6
week.0	1.00	0.42	0.47	0.56

	week.0	week.1	week.4	week.6
week.1	0.42	1.00	0.84	0.56
week.4	0.47	0.84	1.00	0.58
week.6	0.56	0.56	0.58	1.00

3c)

Suppose you were interested in comparing the mean difference in lead levels between treatment and control groups 6 weeks after treatment provision. Using a linear model, compare the lead levels between treatment groups at the 6 week time point. Provide and describe a point estimate for the treatment effect along with the standard error. Does treatment have a significant effect on the Week 6 lead levels at significance level 0.05?

```r
# fit linear model of week 6 level from treatment group
mod_post <- rigr::regress("mean", week.6 ~ tx, data = tlc)

# coef(mod_post)[,c("Estimate","Robust SE","Pr(>|t|)")] %>% round(., 3)
coef(mod_post)[,c("Estimate","Robust SE","Pr(>|t|)")] %>%
  knitr::kable(digits = 3)
```

	Estimate	Robust SE	Pr(>|t|)
(Intercept)	23.646	0.798	0.000
txTreatment	-2.884	1.532	0.063

From a simple linear model we estimate the difference in average lead levels between the treated and control groups 6 weeks after treatment provision to be 2.884 $\mu g/dL$, with the treated group having lower average lead levels (standard error of 1.532 $\mu g/dL$). With a significance level of $\alpha = 0.05$, we conclude that this difference is not statistically significant (pval = 0.063).

3d)

Suppose we adopt an approach to analyze the change scores from week 0 to week 6. Using a linear model, compare the change in lead levels from week 0 to week 6 between treatment groups. Provide and describe a point estimate for the treatment effect along with the standard error. Does your conclusion from part C change?

```r
# fit linear model of change in lead level from treatment group
mod_change <- tlc %>%
  mutate(change = abs(week.6 - week.0)) %>%
  rigr::regress("mean", change ~ tx, data = .)

coef(mod_change)[,c("Estimate","Robust SE","Pr(>|t|)")] %>%
  knitr::kable(digits = 3)
```

	Estimate	Robust SE	Pr(>|t|)
(Intercept)	3.834	0.354	0
txTreatment	4.116	0.901	0

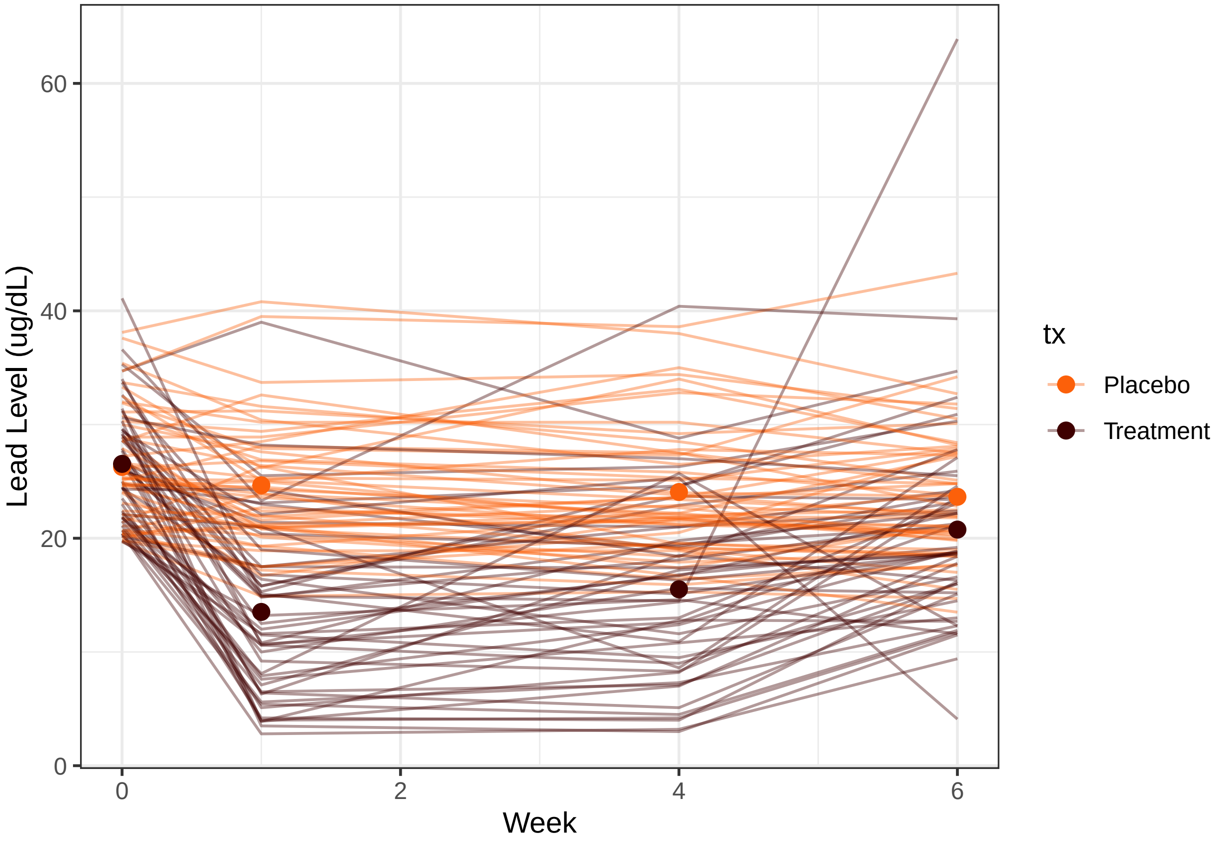


Figure 2: Lead levels and averages by week and treatment group

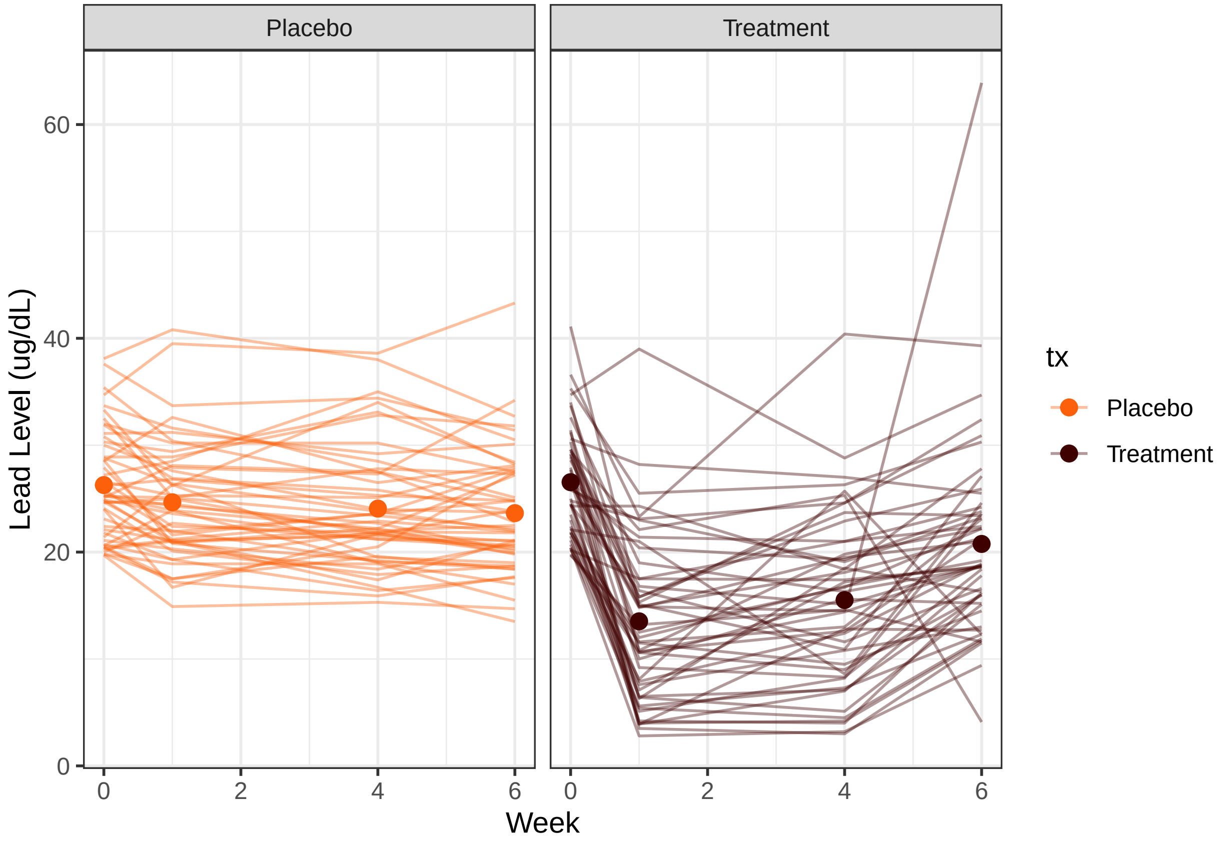


Figure 3: Lead levels and averages by week and treatment group

From a simple linear model we estimate the (absolute) change in lead levels from week 0 to week 6 to differ between the treated and control group by 4.116 $\mu g/dL$, on average, with a standard error of 0.901 $\mu g/dL$, and with the treated group experiencing greater absolute change. With a significance level of $\alpha = 0.05$, we conclude that this difference is statistically significant (pval < 0.001).

Compared to results from (c), this model suggests that treatment effect was meaningful, when measuring effect between week 0 and week 6.

3e)

Suppose we wish to use an ANCOVA model to evaluate the effect of treatment on Week 6 lead levels adjusting for the baseline lead level. Reinterpret the treatment effect and provide an estimate of the standard error. Does your conclusion differ from part C?

```r
# fit linear model of week 6 level from treatment group and week 0 level
mod_ancova <- rigr::regress("mean", week.6 ~ tx + week.0, data = tlc)

coef(mod_ancova)[,c("Estimate","Robust SE","Pr(>|t|)")] %>%
  knitr::kable(digits = 3)
```

	Estimate	Robust SE	Pr(>|t|)
(Intercept)	0.524	3.937	0.894
txTreatment	-3.120	1.238	0.013
week.0	0.880	0.151	0.000

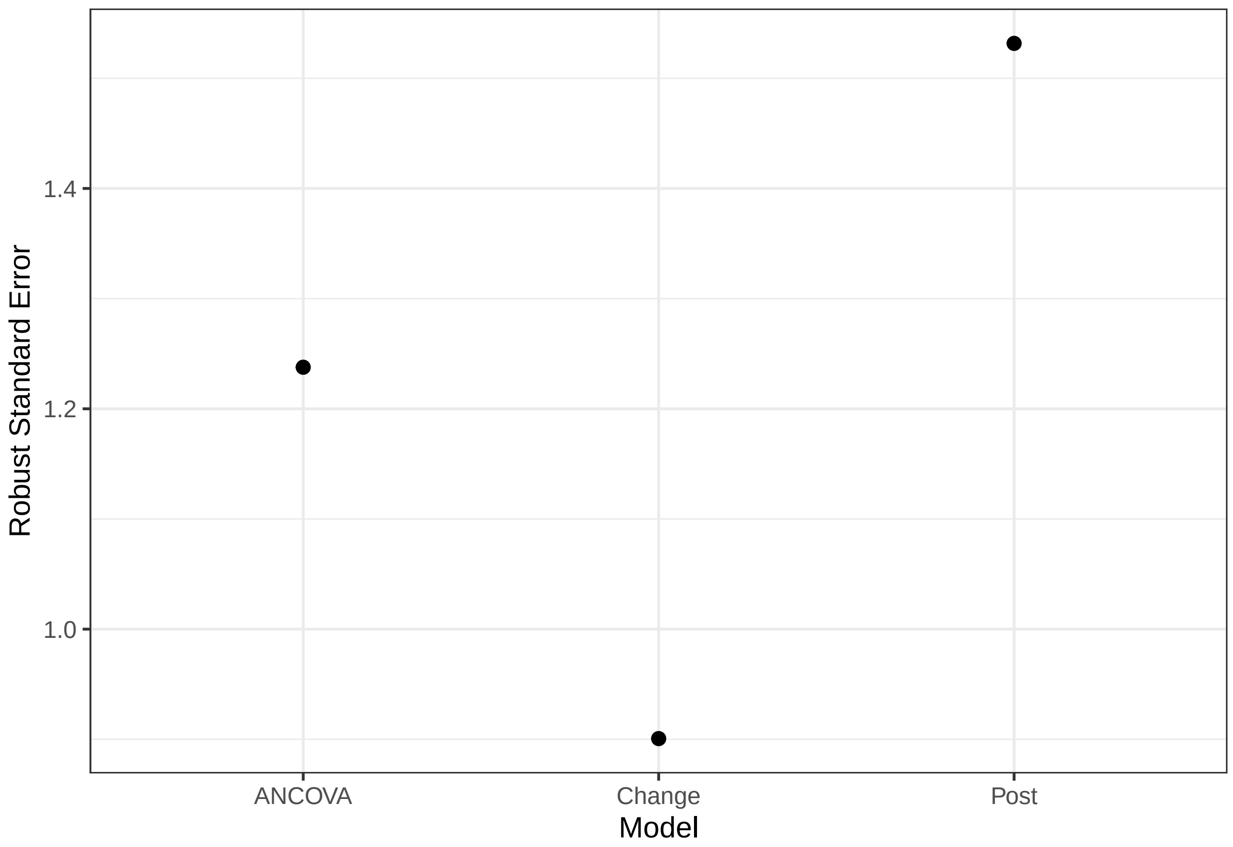
From a simple linear model we estimate the difference in average lead levels between the treated and control groups 6 weeks after treatment provision, given baseline lead levels, to be 3.120 $\mu g/dL$, with the treated group having lower average lead levels (standard error of 1.238 $\mu g/dL$). With a significance level of $\alpha = 0.05$, we conclude that this difference is statistically significant (pval $= 0.013$).

Compared to results from (c), this model suggests that treatment effect on week 6 lead levels was meaningful, when controlling for baseline levels.

3f)

Plot the standard errors of each of the models fit in parts C-E. How does incorporating knowledge of the baseline response impact the precision or power of your inference on the treatment effect at 6 weeks?

```r
# plot standard errors of each model
data.frame(
  model = c("Post", "Change", "ANCOVA"),
  robust.sd = c(coef(mod_post)["txTreatment","Robust SE"],
                coef(mod_change)["txTreatment","Robust SE"],
                coef(mod_ancova)["txTreatment","Robust SE"])) %>%
  ggplot(data = ., aes(x = model, y = robust.sd)) +
    geom_point(size = 2) +
    xlab("Model") + ylab("Robust Standard Error") + theme_bw()
```



Comparing the Post to the ANCOVA model, which measure treatment effect at six weeks, knowledge of the baseline response strengthens the precision our inference.

Problem 4

The Six Cities Study of Air Pollution and Health was a longitudinal study designed to characterize lung growth as measured by changes in pulmonary function in children and adolescents, and the factors that influence lung function growth. A cohort of 13,379 children born on or after 1967 was enrolled in six communities across the U.S.: Watertown (Massachusetts), Kingston and Harriman (Tennessee), a section of St. Louis (Missouri), Steubenville (Ohio), Portage (Wisconsin), and Topeka (Kansas). Most children were enrolled in the first or second grade (between the ages of six and seven) and measurements of study participants were obtained annually until graduation from high school or loss to follow-up. At each annual examination, spirometry, the measurement of pulmonary function, was performed and a respiratory health questionnaire was completed by a parent or guardian.

The dataset contains a subset of the pulmonary function data collected in the Six Cities Study. The data consist of all measurements of FEV1, height and age obtained from a randomly selected subset of the female participants living in Topeka, Kansas. The random sample consists of 300 girls, with a minimum of one and a maximum of twelve observations over time. The variables included in the dataset are `Subject ID`, `Height`, `Age`, `Initial Height`, `Initial Age`, and `Log(FEV1)` (a spirometry measure of lung function).

```r
# load FEV data
fev <- read.csv("data/Topeka-2.csv")
```

4a)

Produce a summary (e.g., mean) of the initial height, initial age, and initial log(FEV1). Hint: you will create the initial logFEV1 variable using mutate, selecting the logFEV1 variable when age==age0. You will also need to use the function select and distinct functions to select a single entry per participant, such that the mean is not calculated over repeated measures.

```r
# summarize number of times an individual was measured and their average
fev %>%
  group_by(id) %>%
  summarise(n.measures = length(logFEV1),
            mean.logFEV = min(logFEV1)) %>%
  head %>% knitr::kable()
```

id	n.measures	mean.logFEV
1	7	0.21511
2	8	0.30748
3	9	0.38526
4	10	0.05827
5	7	0.02956
6	11	0.26236

```r
# calculate averages of initial measures
# gather initial measures
initial.measures <- fev %>%
  group_by(id, height0, age0) %>%
  summarize(n = length(height0),
            logFEV1.0 = logFEV1[1]) %>%
  select(id, n, everything())

# average initial measures
initial.measures %>%
  ungroup(id) %>%
  select(height0, age0, logFEV1.0) %>%
  colMeans %>%
  knitr::kable(digits = 3, col.names = c("", "Average"))
```

	Average
height0	1.289
age0	8.297
logFEV1.0	0.383

End of document.