Fall, 2023 *Monday, Sept. 25*

Stat 414 - Day 1

Review Linear Regression Models

Last Time:

- Multilevel data is when the *structure* of the data is characterized by "observational units" at different levels, often from clustering or nesting in the data (e.g., students nested in classrooms)
- Multilevel data needs to be analyzed differently from single level data

Example 1: Kentucky Derby winners

The Kentucky Derby is an annual horse race run at Churchill Downs in Louisville, KY, USA, on the first Saturday in May (2020 is the first year since 1945 that it wasn't run in May). The race is known as the "Most Exciting Two Minutes in Sports," and is the first leg of racing's Triple Crown. The dataset KYDerby23.txt contains information on each running of the Kentucky Derby since 1875.

First, load in the data:

```
KYDerby23 = read.table("https://www.rossmanchance.com/KYDerby23.txt", header=TRUE)
#You may want to comment this next line out before knitting, especially on a mac
View(KYDerby23)
```

Step 1 - Start with a graph!

```
hist(KYDerby23$Time)
```

a) Examine the distribution of times, what is the first thing you notice? Why is it called the most exciting **two minutes** in sports?

The two clumps in the data are caused by a change in track length. Let's change the variable of interest (the "response variable") to speed, taking the track length into account.

```
speed = (.25*(KYDerby23$Year<1896)+1.25)/(KYDerby23$Time/3600)
hist(speed)
with(KYDerby23, summary(speed)); sd(KYDerby23$speed)
qqnorm(KYDerby23$speed)
```

b) Interpret the mean and the standard deviation (in context).

c) Is the distribution of the response variable normally distributed? Is this a problem? What are some steps we can take if we think this is a problem?

Bivariate graph

```
with(KYDerby23, plot(speed ~ Year))
```

d) Summarize how the speeds have changed over time.

e) Is the association (time trend) linear? Is this a problem? What are some things we can do if we think this is a problem?

Fit and interpret a model

A *least squares regression* model fits the best fitting line by minimizing the sum of the squared residuals.

```
model1 = lm(speed~ Year, data=KYDerby23)
model1
##
## Call:
## lm(formula = speed ~ Year, data = KYDerby23)
##
## Coefficients:
## (Intercept)         Year
##     -7.1619       0.0221
```

f) Write out the least squares regression equation, using appropriate statistical notation, and interpret the coefficients in context.

Validate the model

g) Before we look at p-values and confidence intervals, what are the primary "assumptions" that need to be satisfied for inference in regression models? How do we "check" these assumptions? What can we do if any assumptions are not met?

With more complicated models, an important diagnostic tool is residual plots. The two to start with are a graph of residuals vs. fitted values (aka predicted values) and a histogram and/or normal probability plot of the residuals.

```
#residuals vs. fitted values, with a smoother
scatter.smooth(model1$residuals ~ model1$fitted.values)
#normal probability plot of residuals
qqnorm(model1$residuals)
```

For tomorrow: Summarize what you learn from these graphs.