HW1 - STAT 627

Homayoon Fotros

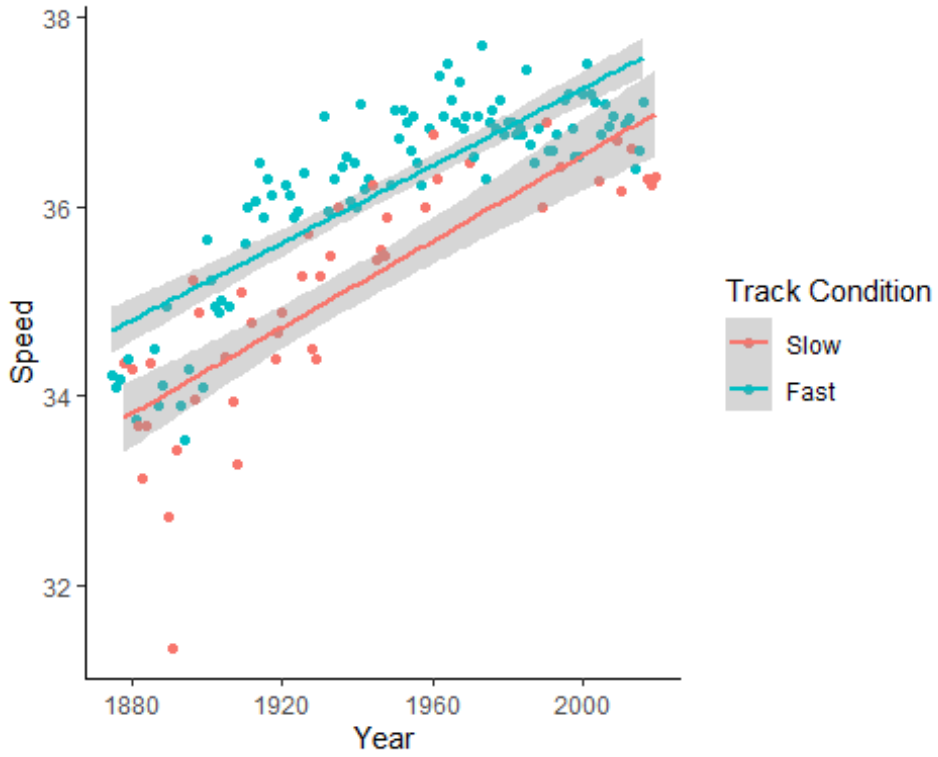
```
library(rio)
library(tidyverse)
library(lubridate)
library(stargazer)

KD_dat <- import('KentuckyDerby.csv')
```

Make a scatterplot of the winning speed vs. year, coding separately by track condition and discuss whether/how the association changes across the track conditions and any other features of the plot you find interesting.

```
KD_dat %>%
  ggplot(aes(Year,Speed, col=factor(Condition))) +
  geom_point() +
  geom_smooth(method = 'lm') +
  theme_classic() +
  scale_color_discrete(name='Track Condition',
                       labels = c('Slow', 'Fast'))

## `geom_smooth()` using formula 'y ~ x'
```

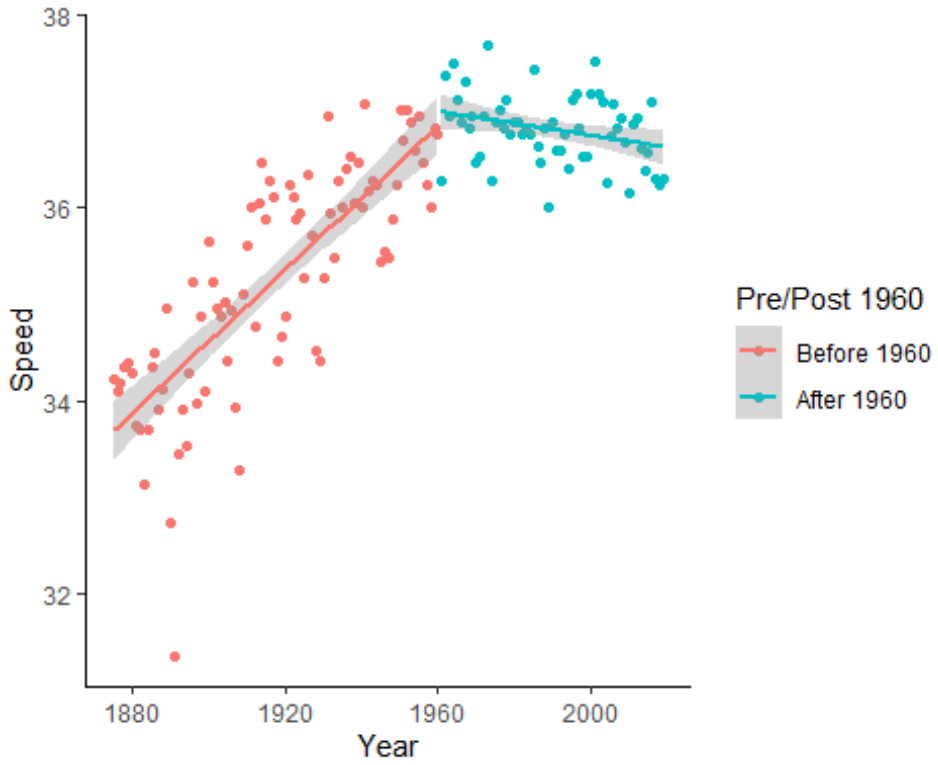


It seems that the winning speed has significantly increased since 1875. The data also suggests that the track condition (fast vs. slow) has a meaningful effect on the winning speed. Specifically, winning speeds in slow condition appears to be lower than those in fast conditions. Meanwhile, notice that for most of the data points between 1920 to 1960 in fast condition, the winning speed is above the approximated trendline.

A closer look into the data shows the increase in winning speeds has not been consistent over time. Since the 1960s, the increasing trend has been plateaued or reversed.

```
KD_dat %>%
  mutate(recent=ifelse(Year>1960, 1,0)) %>%
  ggplot(aes(Year,Speed, col=factor(recent))) +
  geom_point() +
  geom_smooth(method = 'lm', aes(group=recent)) +
  theme_classic() +
  scale_color_discrete(name='Pre/Post 1960',
                       labels = c('Before 1960', 'After 1960'))
```

```
## `geom_smooth()` using formula 'y ~ x'
```



b. Find an appropriate model that predicts Speed from Year taking into account track condition. Discuss your model choice. What criteria are you using to determine if the model is 'appropriate' (or not)? Give justification for your decisions about your model.

A linear regression model can be an appropriate choice for this data set. First, the number of observations (145) are quite large, and at the same time, there is a limited number of

features (inputs) in the model (Year and Condition). Least squared method (OLS) regression should be a good candidate for this set up because it can detect the overarching pattern without risking the overfitting issue. The model's outcome is also a continuous variable, so classification models are not appropriate. Finally, we are interested in the relationship between the inputs and outcome (inference), which we can obtain from the result of linear regression.

I use the following model:

$$Y_t = \beta_0 + \beta_1 Year_t + \beta_2 {Year_t}^2 + \beta_3 Condition + \epsilon$$

where y is the outcome variable ($Speed$) at year t, $Year$ and $Condition$ are race's year and track condition respectively. The inclusion of ${Year}^2$ in the model is based on the pattern that I mentioned in the previous question. Specifically, the squared value of year can help capture the curvilinear behavior of $Speed$-$Year$ relationship.

```
mdl1 <- lm(Speed ~ Year +  I(Year^2) + factor(Condition), data = KD_dat)
options(scipen = 999, digits = 4)
stargazer(mdl1, type = 'text')
```

```
##
## ================================================
##                         Dependent variable:
##                     ---------------------------
##                                Speed
## ------------------------------------------------
## Year                          0.950***
##                                (0.102)
##
## I(Year2)                     -0.0002***
##                                (0.00003)
##
## factor(Condition)1             0.750***
##                                (0.090)
##
## Constant                     -909.500***
##                                (99.620)
##
## ------------------------------------------------
## Observations                     145
## R2                              0.834
## Adjusted R2                     0.831
## Residual Std. Error      0.493 (df = 141)
## F Statistic          236.500*** (df = 3; 141)
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

The result shows given the track condition, the effect of one year increase on the winning speed peaks around the 1960s and then becomes slightly negative. These estimates are statistically significant at 0.01 level.

c. Is there evidence of a nonlinear trend in speeds over time? If so, does this time trend depend on the track condition. That is, is there a difference in the trend for fast vs. slow conditions?

The above model provides evidence that there is a nonlinear relationship between the winning speed and time (year). Compared to a linear relationship and a cubic model (including $Year^3$ in variables), the quadratic model yields a greater explanatory power (R^2) while maintaining the statistical significance of input variables.

The effect of track condition (Condition) is substantive and statistically significant. As shown in Question 1, the trendline for Slow condition is below the the Fast condition's trendline.

d. Describe the impact of track condition on winning speed.

The result indicates compared to a Slow condition, a Fast track condition improves the winning speed by 0.75 mph in a given year. This effect is statistically significant at 0.01 level.

e. What Speeds do you predict for the two most recent winners in 2020 and 2021? Note that both races were held under Fast conditions.

```
w_2020 <- data.frame(Year=2020,Condition = 1)
pred2020 <- predict(mdl1,newdata =w_2020)
pred2020_time <- duration(minutes = (1.25/pred2020) * 60 )

w_2021 <- data.frame(Year=2021,Condition = 1)
pred2021 <- predict(mdl1,newdata =w_2021)
pred2021_time <- duration(minutes = (1.25/pred2021) * 60 )
```

The predicted time for 2020 race is `2` minutes and `2.15` seconds.

The predicted time for 2021 race is `2` minutes and `2.2` seconds.

f. How do your predictions compare to the actual winning times?

2020 Authentic won in 2:00.61 (2 minutes, 0.61 seconds)

2021: Medina Spirit won in 2:01.36 (2 minutes, 1.36 seconds)

```
diff_2020 <- pred2020_time - duration(minutes=2, seconds=0.61)
diff_2021 <- pred2021_time - duration(minutes=2, seconds=1.36)
```

The predicted time **for 2020 is `1.54` seconds greater** than the actual winning time.

The predicted time **for 2021 is `0.84` seconds greater** than the actual winning time.

g. There is one clear outlier in the data set. Identify the outlier (give the Year, Speed, Condition). Is this outlying race influential on your model in (b)?

The outlier has a winning speed lower than 32 mph, and belongs to the race in 1891 highlighted below:

```
KD_dat %>%
  filter(Speed<32) %>%
  mutate(Duration = round(duration(minutes = (1.25/Speed) * 60 ),2)) %>%
select(Year, Winner, Duration, Condition)

##   Year  Winner                Duration Condition
## 1 1891 Kingman 143.54s (~2.39 minutes)         0
```

Excluding this data point from the data set does not affect the overall result and only changes the points estimates by a negligible amount.

```
mdl_ex <- lm(Speed ~ Year +  I(Year^2) + factor(Condition), data = KD_dat,
subset = (Speed>32))
stargazer(mdl_ex, type = 'text')

##
## ================================================
##                         Dependent variable:
##                     ---------------------------
##                                Speed
## ------------------------------------------------
## Year                          0.916***
##                                (0.093)
##
## I(Year2)                     -0.0002***
##                                (0.00002)
##
## factor(Condition)1            0.704***
##                                (0.081)
##
## Constant                     -875.700***
##                                (90.160)
##
## ------------------------------------------------
## Observations                     144
## R2                              0.851
## Adjusted R2                     0.847
## Residual Std. Error      0.445 (df = 140)
## F Statistic          265.900*** (df = 3; 140)
## ================================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```