HW2-DATA 627

Homayoon Fotros

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

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

Part (a)

Assessment of a non-linear relationship between Speed and Year

In Homework-1, I applied a regression model with Speed and Year having a quadratic relationship. Here, I first evaluate the extent to which this model is superior to a linear fit by Anova test.

```
quad_model <- lm(Speed ~ poly(Year,2, raw=TRUE) + factor(Condition), data = KD_dat)

ln_model <- lm(Speed ~ Year + factor(Condition), data = KD_dat)

anova(quad_model, ln_model)
```

```
## Analysis of Variance Table
##
## Model 1: Speed ~ poly(Year, 2, raw = TRUE) + factor(Condition)
## Model 2: Speed ~ Year + factor(Condition)
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    141 34.265
## 2    142 54.264 -1   -19.999 82.295 9.166e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The null hypothesis in this Anova test is that the two models are equally fit. The result shows, however, that the F-statistic is 82.3 and the p-value is almost zero, which means that the quadratic model is clearly superior to the other one.

We can also test whether including a cubic and quartic terms would improve the model:

```
options(scipen = 999)

cube_model <- lm(Speed ~ poly(Year, 3, raw = TRUE) + factor(Condition), data = KD_dat)

quart_model <- lm(Speed ~ poly(Year, 4, raw = TRUE) + factor(Condition), data = KD_dat)

stargazer(quad_model, cube_model,quart_model, header = FALSE, type = 'latex',
          column.labels = c( 'Quadratic', 'Cubic', 'Quartic'), title = 'Regression Results')
```

Unlike the cubic model, applying a quartic model (fourth-degree polynomial) retains the features' statistical significance. So we might test whether this quartic model is superior to the quadratic model:

Table 1: Regression Results

	Dependent variable:		
	Speed		
	Quadratic	Cubic	Quartic
	(1)	(2)	(3)
poly(Year, 2, raw = TRUE)1	0.950^{***}		
	(0.102)		
poly(Year, 2, raw = TRUE)2	-0.0002^{***}		
	(0.00003)		
poly(Year, 3, raw = TRUE)1		−5.077	
		(8.277)	
poly(Year, 3, raw = TRUE)2		0.003	
		(0.004)	
poly(Year, 3, raw = TRUE)3		−0.00000	
		(0.00000)	
poly(Year, 4, raw = TRUE)1			$-1{,}639.840^{***}$
			(563.403)
poly(Year, 4, raw = TRUE)2			1.263^{***}
			(0.434)
poly(Year, 4, raw = TRUE)3			-0.0004^{***}
			(0.0001)
poly(Year, 4, raw = TRUE)4			0.00000^{***}
			(0.00000)
factor(Condition)1	0.750^{***}	0.736^{***}	0.763^{***}
	(0.090)	(0.092)	(0.090)
Constant	-909.461^{***}	2,998.814	$798{,}166.100^{***}$
	(99.617)	(5,368.645)	(274,067.400)
Observations	145	145	145
R^2	0.834	0.835	0.844
Adjusted R^2	0.831	0.830	0.839
Residual Std. Error	0.493 (df = 141)	0.494 (df = 140)	0.481 (df = 139)
F Statistic	236.498^{***} (df = 3; 141)	176.915^{***} (df = 4; 140)	150.718^{***} (df = 5; 139)
Note:			*p<0.1; **p<0.05; ***p<0.01

```
anova(quad_model, quart_model)
```

```
## Analysis of Variance Table
##
## Model 1: Speed ~ poly(Year, 2, raw = TRUE) + factor(Condition)
## Model 2: Speed ~ poly(Year, 4, raw = TRUE) + factor(Condition)
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    141 34.265
## 2    139 32.186  2    2.0791 4.4896 0.0129 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The F-statistic is 4.49 and the p-value is 0.13. This result can indicate that the 4th degree polynomial model has slightly improved the quadratic model. However, we must be cautious about the potential overfitting issue of polynomial models as we increase the degree. It is also important to keep the model's complexity at a reasonable level. For these reasons, it appears that a quadratic equation is a better choice for our model.

We may also test the log transformation of time (Year) and see how it performs compared to our quadratic model:

```
log_model <- lm(Speed ~ log(Year) + factor(Condition), data = KD_dat)
# summary(log_model)

anova(log_model, quad_model)
```

```
## Analysis of Variance Table
##
## Model 1: Speed ~ log(Year) + factor(Condition)
## Model 2: Speed ~ poly(Year, 2, raw = TRUE) + factor(Condition)
##   Res.Df    RSS Df Sum of Sq      F               Pr(>F)
## 1    142 53.360
## 2    141 34.265  1    19.095 78.575 0.00000000000000304 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

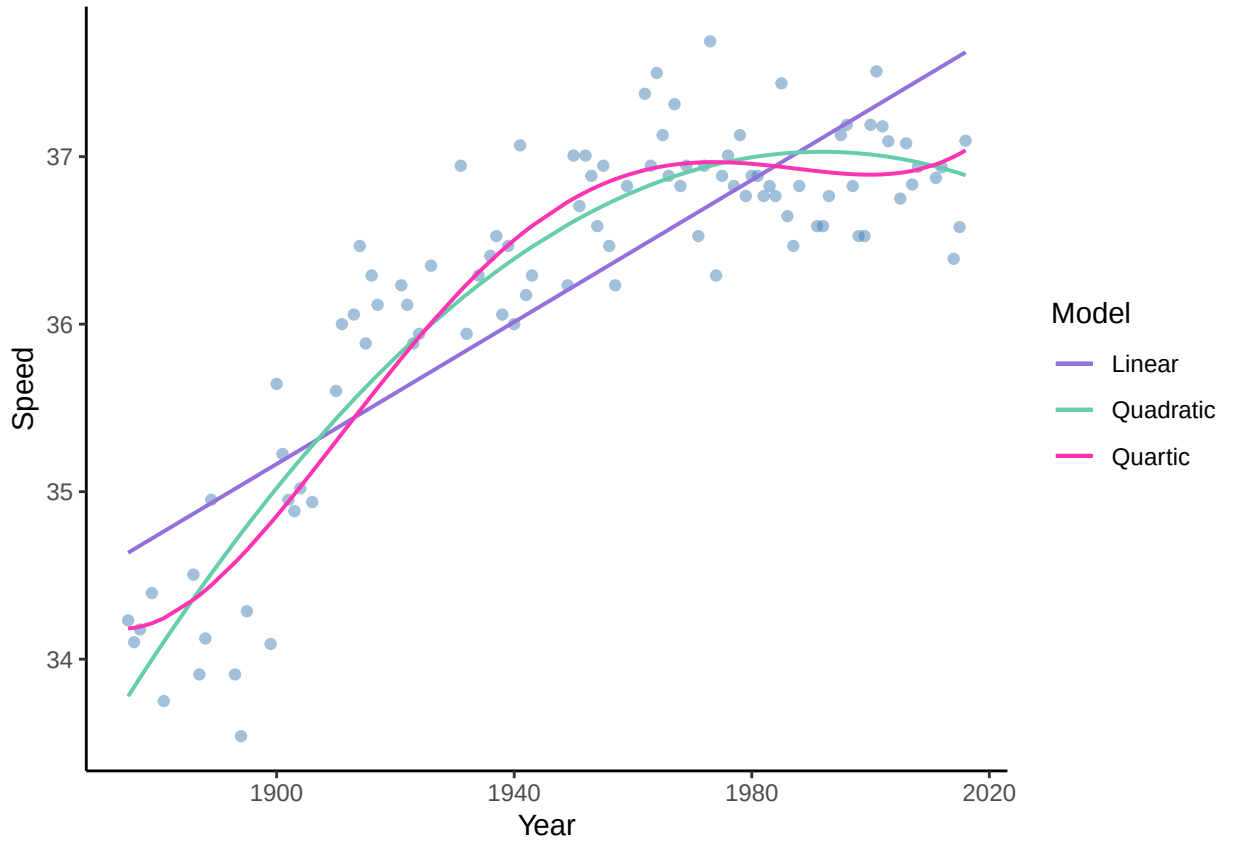
The Anova test shows that the quadratic model provides a better fit to our data.

Finally, we can plot the linear, quadratic, and fourth degree polynomial models to visually inspect their fit onto our data set (The following plots only for the Fast Condition).

```
fit_ln <-ln_model$fitted.values[KD_dat$Condition==1]
fit_quad <- quad_model$fitted.values[KD_dat$Condition==1]
fit_quart <- quart_model$fitted.values[KD_dat$Condition==1]
fit_log <- log_model$fitted.values[KD_dat$Condition==1]

cols <- c("Linear" = "mediumpurple", "Quadratic" = "mediumaquamarine", "Quartic" = "maroon1")

KD_dat %>%
  filter(Condition==1) %>%
  ggplot() +
  geom_point(aes(Year,Speed), col='steelblue', alpha=0.5) +
  geom_line(aes(Year, fit_ln, col='Linear'), size=0.7) +
  geom_line(aes(Year, fit_quad, col='Quadratic'),size=0.7) +
  geom_line(aes(Year, fit_quart, col='Quartic'), size = 0.7 ) +
  scale_color_manual(name = 'Model',  values = cols) +
  theme_classic()
```



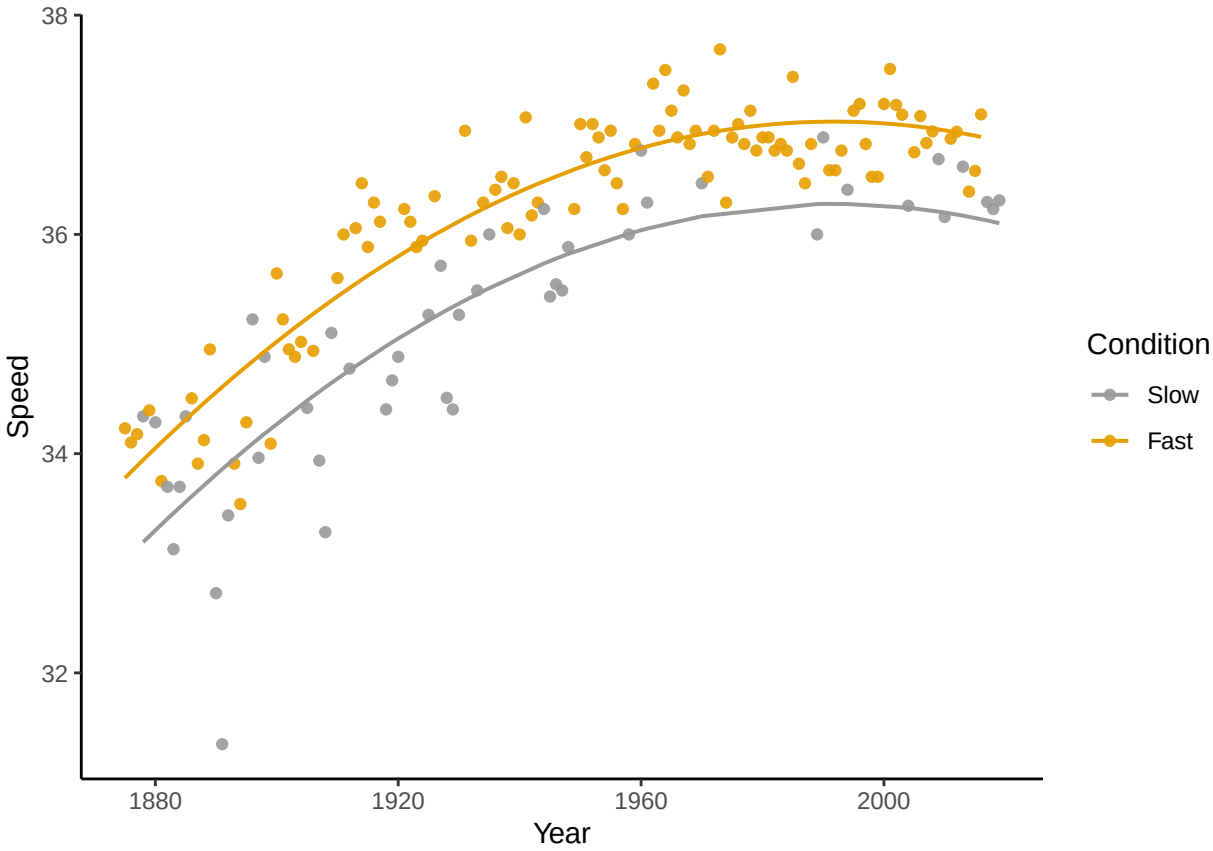
Part (b & c)

Impact of track conditions (Fast vs. Slow)

The result shows the track condition (Condition) has a significant effect on the winning speed (Speed). Concretely, the coefficient for Condition is 0.75. This means if everything else is equal, changing the track condition from Slow to Fast would increase the Speed by 0.75 mph.

The following plot demonstrates how the model fits differently based on the track condition:

```
KD_dat %>%
  ggplot() +
  geom_point(aes(Year,Speed, col=factor(Condition)), alpha=0.9) +
  geom_line(aes(Year, quad_model$fitted.values, col=factor(Condition)),size=0.7) +
  scale_color_manual(name = 'Condition',labels = c('Slow', 'Fast'),
                     values=c("#999999", "#E69F00")) +
  theme_classic()
```



As this plot shows, the Slow condition negatively affects the winning speed.

Part (d)

Prediction and comparison with 2020 and 2021 speeds

I answered this part in HW1 using the same quadratic model.