Lecture 3: Baseball stats & Multivariate regression

Skidmore College, MA 276

Multivariate regression

Model:

$$y_i = \beta_0 + \beta_1 * x_{i1} + \beta_2 * x_{i2} + \ldots + \beta_{p-1} * x_{i,p-1} + \epsilon_i$$

Assumptions:

- $\epsilon_i \sim N(0, \sigma^2)$
- ϵ_i,$\epsilon_{i'}$ independent for all i, i'
- Linear relationship between y and x

Multivariate regression

Estimated model:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 * x_{i1} + \hat{\beta}_2 * x_{i2} + \ldots + \hat{\beta_{p-1}} * x_{i,p-1}$$

Interpretations:

- $\hat{\beta}_0$:
- $\hat{\beta}_1$:

Ex: Runs against (RA)

```
library(tidyverse)
library(Lahman)
Teams.1 <- Teams %>% filter(yearID >= 1970)
fit.pitcher <- lm(RA ~ HRA + BBA + SOA, data = Teams.1)
```

Write the multiple regression model:

Ex: Runs against (RA)

```
library(broom)
tidy(fit.pitcher)  ### alternatively, use summary(fit.pitcher)
```

```
## # A tibble: 4 x 5
##   term        estimate std.error statistic   p.value
##   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)  223.      11.3          19.8 1.28e- 76
## 2 HRA            1.97     0.0448       44.0 2.74e-263
## 3 BBA            0.583    0.0195       29.9 2.39e-151
## 4 SOA           -0.110    0.00759     -14.5 1.69e- 44
```

Write the estimated multiple regression model

Ex: Runs against (RA)

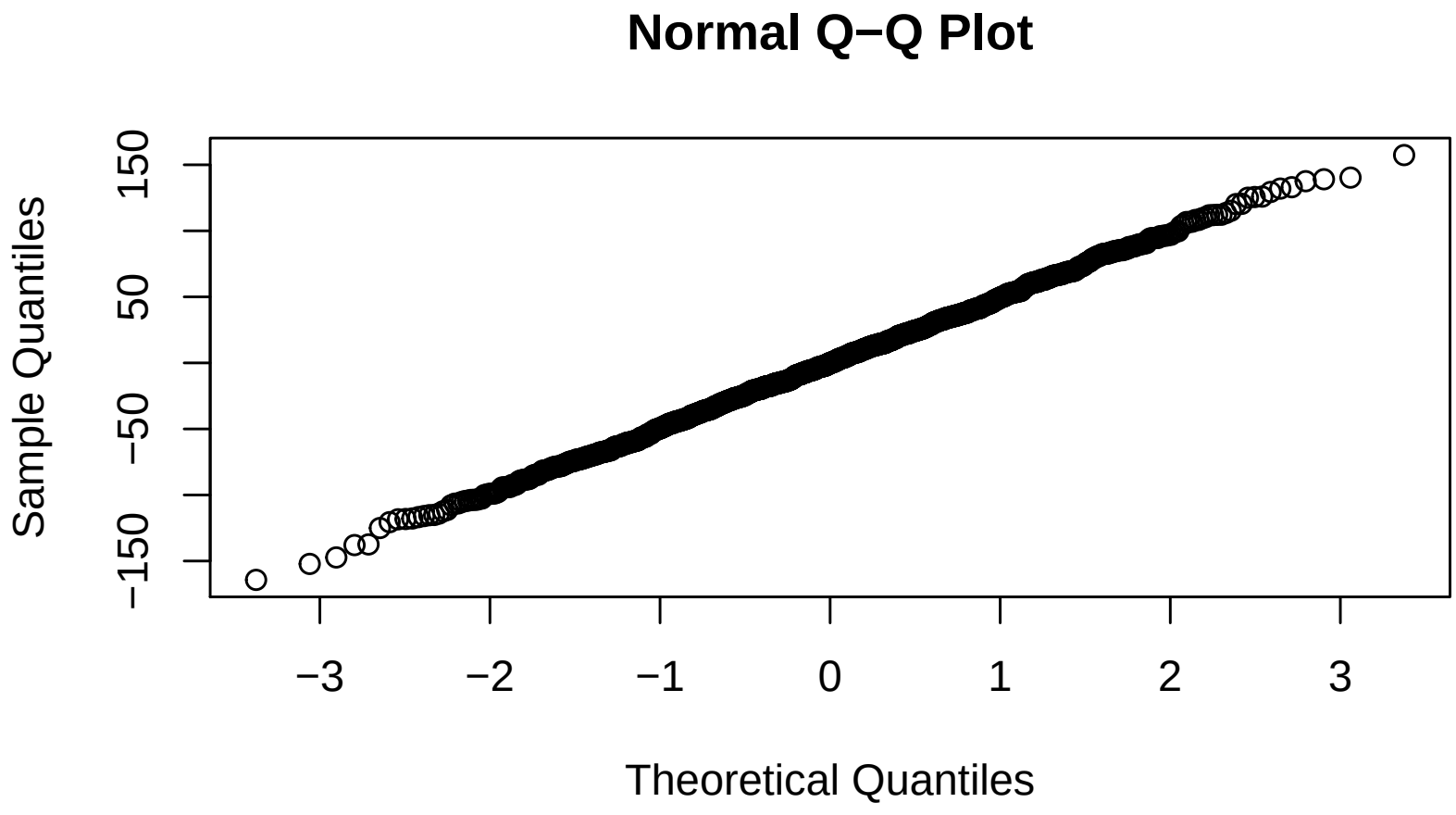
```
tidy(fit.pitcher)
```

```
## # A tibble: 4 x 5
##   term         estimate std.error statistic   p.value
##   <chr>           <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)  223.      11.3           19.8 1.28e- 76
## 2 HRA            1.97     0.0448        44.0 2.74e-263
## 3 BBA            0.583    0.0195        29.9 2.39e-151
## 4 SOA           -0.110    0.00759      -14.5 1.69e- 44
```

Interpret the slope for SOA. Interpret the intercept

Ex: Runs against (RA)

```
qqnorm(fit.pitcher$resid)
```



Conclusions from the model