Lecture 3: Baseball stats & Multivariate regression

Skidmore College

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 + lgID, data = Teams.1)
```

Write the multiple regression model:

Ex: Runs against (RA)

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

```
## # A tibble: 5 x 5
##   term        estimate std.error statistic   p.value
##   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)  229.      11.1         20.6   1.73e- 82
## 2 HRA            1.93     0.0455      42.3   1.34e-251
## 3 BBA            0.591    0.0195      30.3   2.54e-155
## 4 SOA           -0.114    0.00773    -14.7   9.21e- 46
## 5 lgIDNL        -2.52     2.82        -0.893 3.72e-  1
```

Write the estimated multiple regression model

Ex: Runs against (RA)

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

```
## # A tibble: 5 x 5
##   term         estimate std.error statistic   p.value
##   <chr>           <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)  229.      11.1         20.6   1.73e- 82
## 2 HRA            1.93     0.0455      42.3   1.34e-251
## 3 BBA            0.591    0.0195      30.3   2.54e-155
## 4 SOA           -0.114    0.00773    -14.7   9.21e- 46
## 5 lgIDNL        -2.52     2.82        -0.893 3.72e-  1
```

Interpret the slope for SOA. Interpret the intercept

Ex: Runs against (RA)

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

```
## # A tibble: 5 x 5
##   term         estimate std.error statistic   p.value
##   <chr>           <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)  229.      11.1         20.6   1.73e- 82
## 2 HRA            1.93     0.0455      42.3   1.34e-251
## 3 BBA            0.591    0.0195      30.3   2.54e-155
## 4 SOA           -0.114    0.00773    -14.7   9.21e- 46
## 5 lgIDNL        -2.52     2.82        -0.893 3.72e-  1
```

Interpret the slope for `lgID`.

Assumptions

Model:

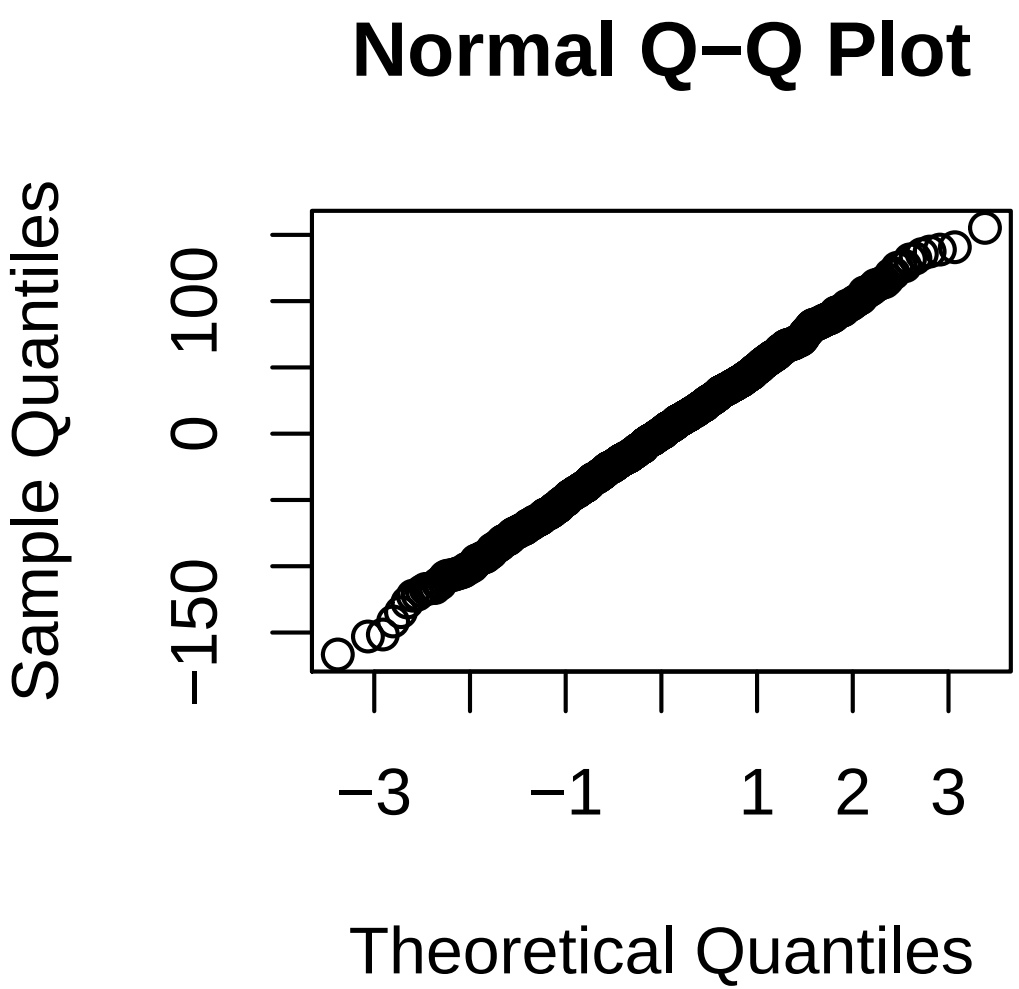
$$y_i = \beta_0 + \beta_1 * x_{i1} + \beta_2 * x_{i2} + \ldots + \beta_{p-1} * x_{i,p-1} + \epsilon_i$$

Assumptions:

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

Ex: Runs against (RA)

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



Conclusions from the model

Open ended question 1

Write the following model, and interpret the coefficients

```
fit.offense <- lm(R ~ X2B + X3B + lgID, data = Teams.1)
tidy(fit.offense)
```

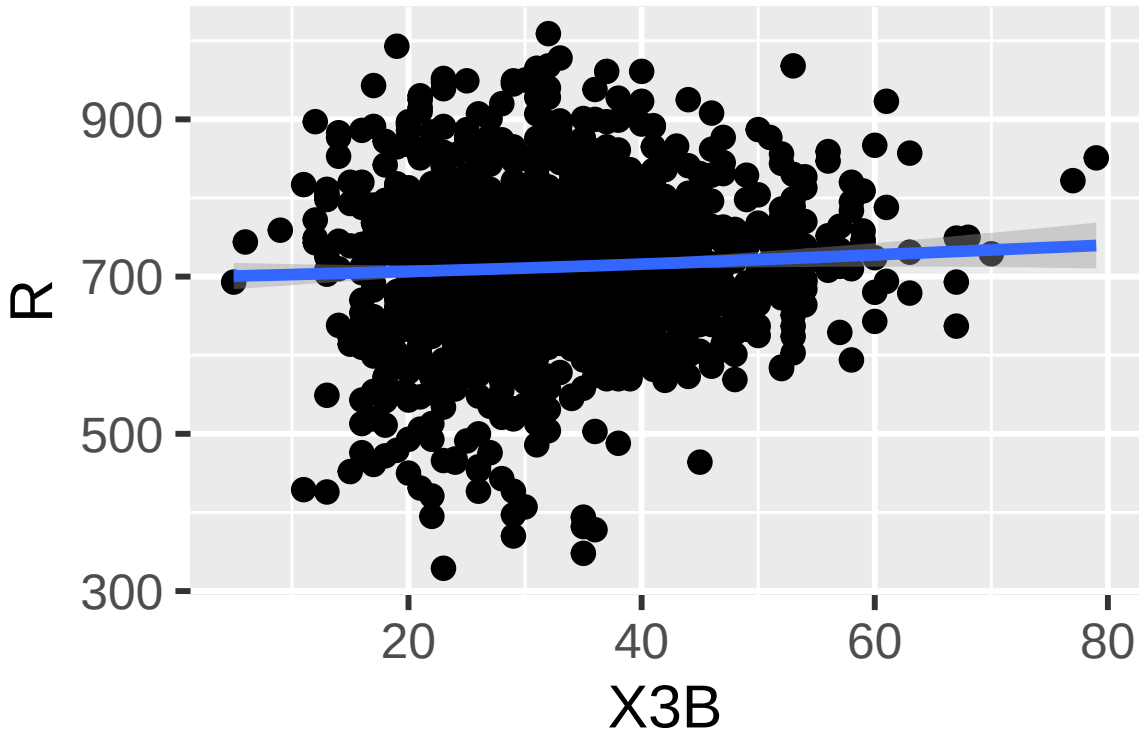
```
## # A tibble: 4 x 5
##   term        estimate std.error statistic   p.value
##   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)  259.       13.0        20.0  4.93e- 78
## 2 X2B            1.68      0.0437     38.6  2.00e-221
## 3 X3B            0.850     0.177       4.80 1.75e-  6
## 4 lgIDNL       -30.4       3.62       -8.39 1.20e- 16
```

Open ended question 2

What does the following plot say about the multiple regression model in Open Ended question 1?

```
ggplot(Teams.1, aes(X3B, R)) +
  geom_point() +
  geom_smooth()
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs
```

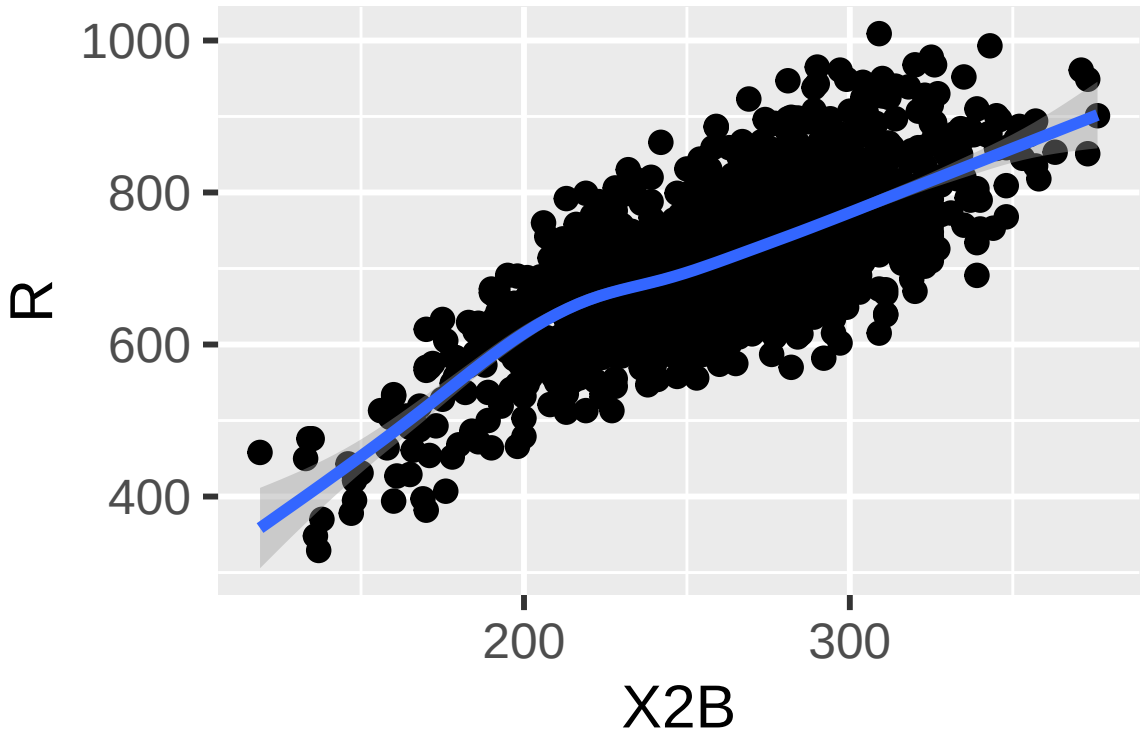


Open ended question 3

What does the following plot say about the multiple regression model in Open Ended question 1?

```
ggplot(Teams.1, aes(X2B, R)) +
  geom_point() +
  geom_smooth()
```

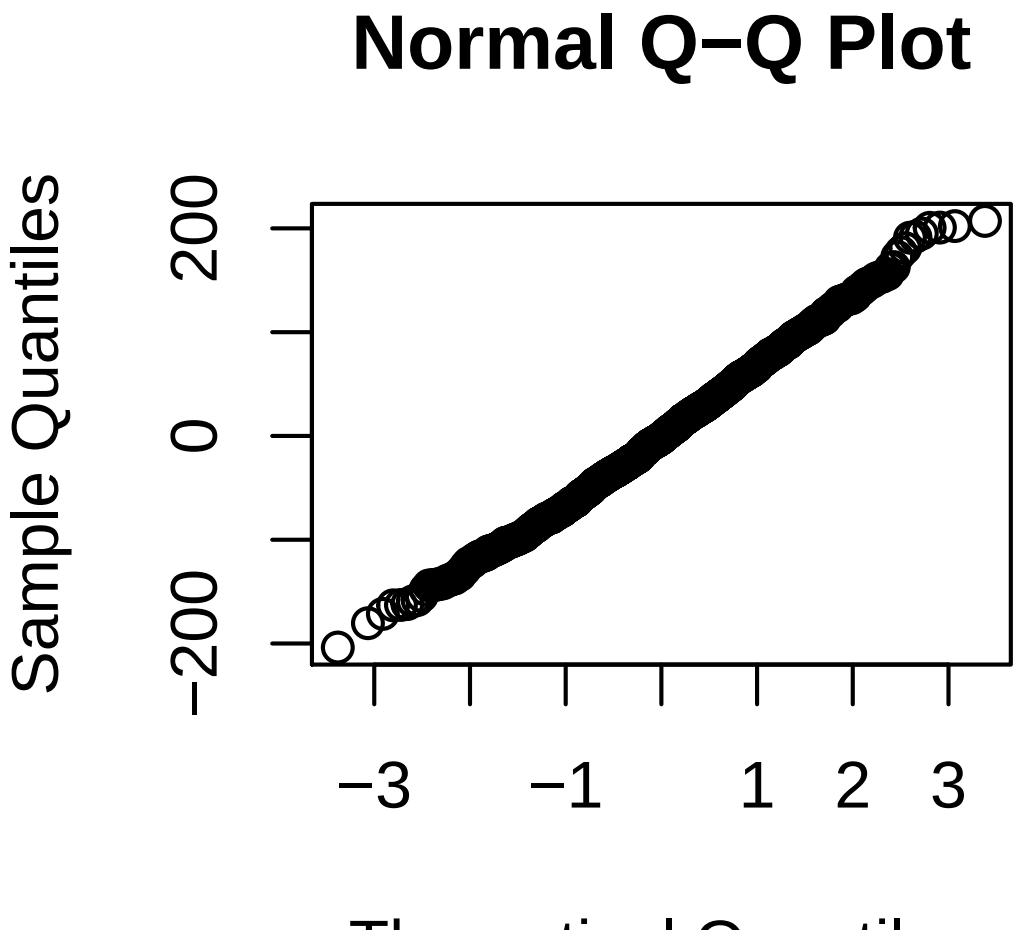
```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs
```



Open ended question 4

What does the following plot say about the multiple regression model in Open Ended question 1?

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



Open ended question 5

Find another variable that's a significant predictor of Runs – does it change the coefficients on the variables currently in the model?