602Project_part1

Ze Yu

2024-02-09

Part1: ***Read data***:

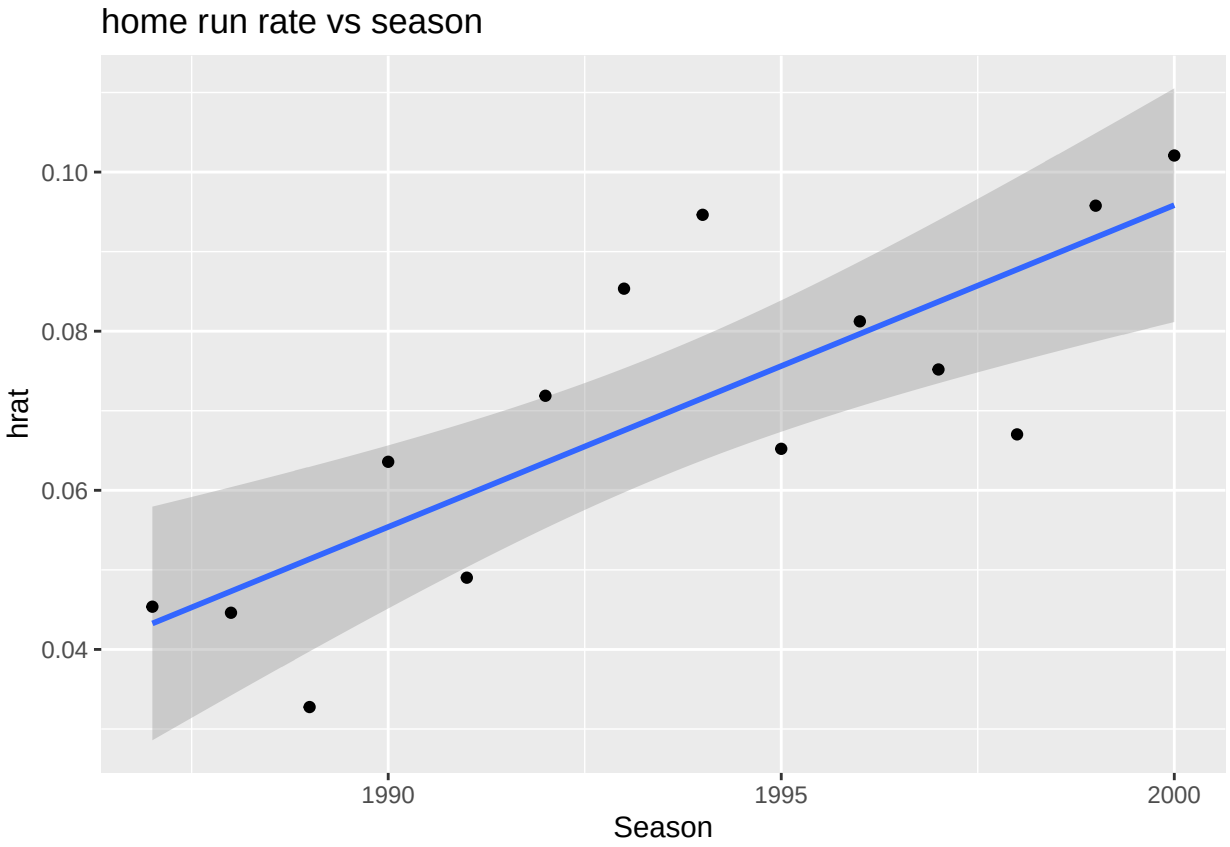
```
# Read data
data = read.csv("bondsdata.csv")
# Remove year 2001
data <- data[-c(15), ]
data
```

```
##    season     hrat
## 1    1987 0.045372
## 2    1988 0.044610
## 3    1989 0.032759
## 4    1990 0.063584
## 5    1991 0.049020
## 6    1992 0.071882
## 7    1993 0.085343
## 8    1994 0.094629
## 9    1995 0.065217
## 10   1996 0.081238
## 11   1997 0.075188
## 12   1998 0.067029
## 13   1999 0.095775
## 14   2000 0.102083
```

Build Model:

```
ggplot(data, aes(x = season, y = hrat)) + geom_smooth(method = "lm") + geom_point() + xlab("Season") + g
```

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




```
predicted_rate = lm(hrat ~ season, data=data)
predicted_rate$coef
```

```
##  (Intercept)       season
## -7.992499290  0.004044169
```

The linear regression equation representing the model is $\hat{HRAT_i} = -7.992499290 + 0.004044169 Year_i$.
Correlation Coefficient Check:

```
season <- data$season
hrat <- data$hrat
cor(season, hrat)
```

```
## [1] 0.7981544
```

Correlation Coefficient shows that there is a strong positive correlation between the season/year and home run rate of Barry Bonds.

Check significance of coefficient estimates Null hypothesis: $H_0 : \beta_1 = 0$, $H_0 : \beta_0 = 0$

Alternative hypothesis: $H_A : \beta_1 \neq 0$, $H_0 : \beta_0 \neq 0$

We will set the alpha value to 0.05.

We can use a t test to check our claim.

```r
summary(predicted_rate)
```

```
##
## Call:
## lm(formula = hrat ~ season, data = data)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.020722 -0.009931  0.001841  0.007701  0.023055
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -7.9924993  1.7566775  -4.550 0.000666 ***
## season       0.0040442  0.0008812   4.589 0.000622 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01329 on 12 degrees of freedom
## Multiple R-squared:  0.6371, Adjusted R-squared:  0.6068
## F-statistic: 21.06 on 1 and 12 DF,  p-value: 0.0006222
```

From our t-test, we get test statistics of β_0 is -4.549782 and β_1 is 4.589384. P-values of β_0 is 0.0006664296 and β_1 is 0.0006222474. The p-values are smaller than the set alpha value of 0.05, so we reject our null hypothesis that the linear regression coefficient and intercept are 0. Therefore, we can conclude that the home run rate of Bary Bonds can be expressed as a linear function of the season, and since $\beta_1 > 0$, we can say it is also positive.

We also get an R-squared value of 0.6371, meaning out independent variable explains approx. 63.71% of the variance in the dependent variable, which is quite good.

Residual Analysis:

There are two conditions that must be met for our linear regression model to be valid.

1. Normality of residuals: The dependent variable (hrat) must be normally distributed with a mean of μ and standard deviations of σ. To check this we will plot a stat_qq plot of the residuals since $e_i = y_i - \hat{y_i}$, if y is normally distributed, so will the residuals.

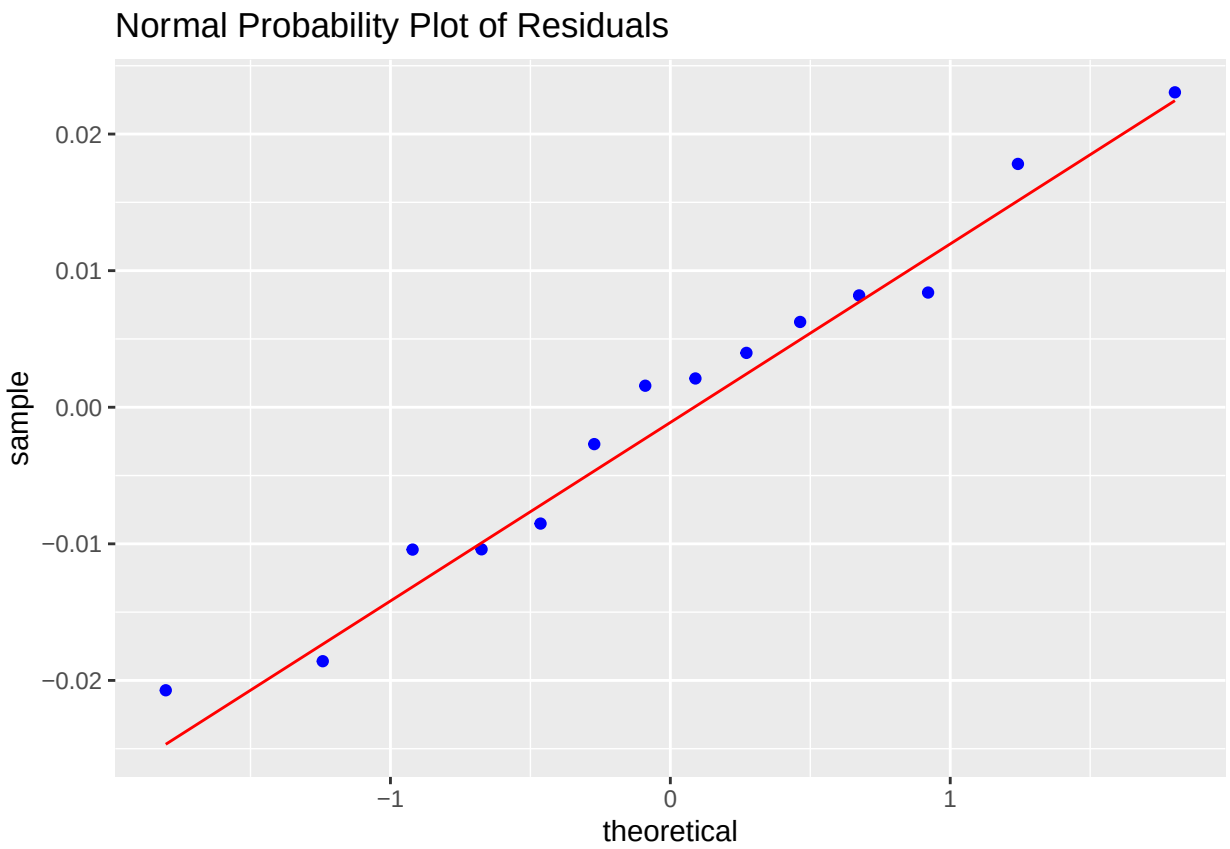
2. Homoscedasticity: For each distinct value of the independent variable (season), the dependent variable (hrat) has the same standard deviation σ. To check this, we will plot a scatter plot of the fitted values and the residuals.

```r
# Get the and residuals fitted values
predicted.rate = predicted_rate$fitted.values
ei_hrat = predicted_rate$residuals
data.df = data.frame(predicted.rate, ei_hrat)
```

Normality of Residuals Plot:

```r
ggplot(data.df, aes(sample = ei_hrat)) +  stat_qq(col='blue') + stat_qqline(col='red') + ggtitle("Norma
```

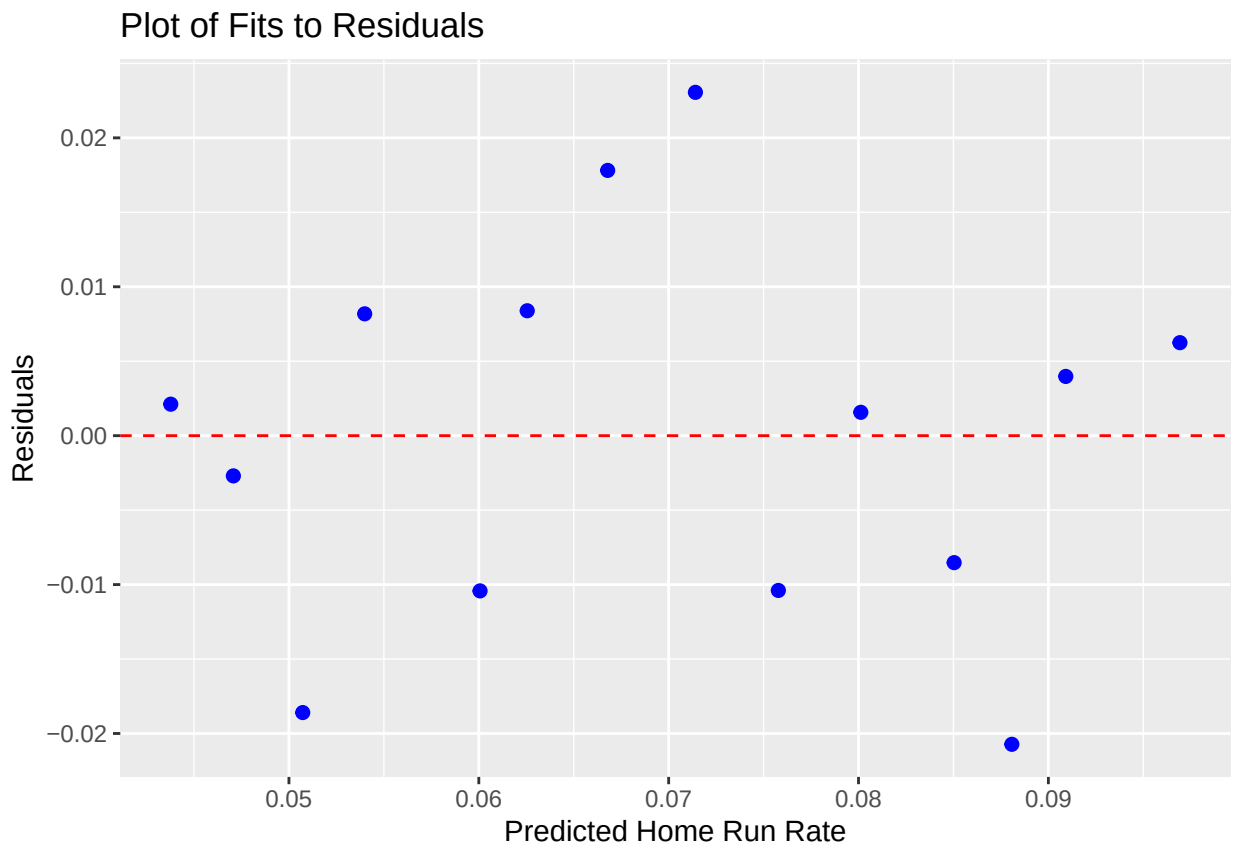
```
## Warning: The following aesthetics were dropped during statistical transformation: sample
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## i Did you forget to specify a 'group' aesthetic or to convert a numerical
##   variable into a factor?
```



Looking the normality probability plot, the residuals do seem to be approximately normally distributed and therefore so is the dependent variable (hrat). The normality of residuals condition holds.

Homoscedasticity:

```
ggplot(data.df, aes(x = predicted.rate, y = ei_hrat)) +  geom_point(size=2, col='blue', position="jitte
```



Looking the plot of fits to residuals, the residuals do seem to be evenly distributed over the home run rate. We can they say that the condition of homoscedasticity holds.

Since both conditions hold, our linear regression model is valid.

Predict home run rate in season 2001:

```
predict(predicted_rate, newdata=data.frame(season = 2001), interval="predict")
```

```
##          fit        lwr       upr
## 1 0.09988334 0.06662845 0.1331382
```

Using our linear regression model, we get a predicted HRAT of 0.099883 for Barry Bonds' 2001 season. Barry Bonds' actual HRAT for the 2001 season was 0.153400, so our predicted error is $0.153400 - 0.099883 = 0.053517$.

We also calculated a 95% prediction interval of HRAT for the 2001 season. This means we can say with 95% confidence that the value of Barry Bonds' HRAT for the 2001 season is between 0.06663 and 0.13314 based on our previous records. However, Barry Bonds' actual HRAT is not in this range.

In conclusion, we cannot say whether or not Barry Bonds was on steroids in the 2001 season, but it does seem like there was a factor that attributed to a much larger HRAT in the 2001 season than what was expected.