Linear regression continued
Outline for today

Better know a player: Ken Griffey Jr.

Review and continuation of linear regression
Better know a player

Ken Griffey Jr.
Thoughts on Big Data Baseball?

Please keep up with the reading and the quote and reactions

We will have a brief discussion of the book next class
Meetings on Friday

If anyone would like to meet with me on Friday to discuss anything related to the class, please sign up for a time here:

https://emeyers.youcanbook.me

Also, midterm self-evaluations were due yesterday so please complete them if you have not done so already.
Review
How to create z-score in R

Z-score formula is? \[ z_i = \frac{x_i - \bar{x}}{s} \]

Suppose we have two data frames
- Jeter.data  # Derek Jeter’s yearly data
- PA500.Jeter7th  # Data from all players in Jeter’s 7th season

How do we create a z-score for Jeter’s 7th season BA?
1. BA.Jeter.7th <- Jeter.data[7, ]$BA
2. mean.all.players.7th <- mean(PA500.Jeter7th$BA)
3. sd.all.players.7th <- sd(PA500.Jeter7th$BA)
4. zscore.Jeter <- (BA.Jeter.7th - mean.all.players.7th)/sd.all.players.7th
Regression

Regression is a method of using one variable to predict the value of a second variable.

In **linear regression** we fit a line to the data, called the **regression line**.

\[
\hat{y} = a + b \cdot x
\]

*Response = a + b \cdot Explanatory*
1. An additional win for ~11 additional runs scored
   • Now we know the value of a run! It’s worth $\frac{1}{11}{\text{th}}$ of a win!
2. There will be 14.47 wins if you score 0 runs all season

\[ \hat{y} = a + b \cdot x \]
\[ a = 14.47 \]
\[ b = 0.088 \]
\[ \hat{w} = 14.47 + 0.088 \cdot \text{Runs} \]
\[ \hat{y} = a + b \cdot x \]

\[ a = -529.8 \]

\[ b = 4755.7 \]

Write the equation for predicting runs as a function of BA

\[ \hat{r} = -529.8 + 4755.7 \cdot BA \]
Link to download data is on Moodle (see Feb 28):

```r
load('/home/shared/baseball_stats_2007/data/team_batting_stats.Rda')
```

We can build a linear model using:

```r
fit <- lm(y ~ x, data = my_df)
```

We can extract \( a \) (the intercept) and \( b \) (the slope) from the model using:

```r
coef(fit)
```

What are \( y \) and \( x \) for predicting runs from batting average?

\( y = R \) and \( x = BA \)

See if you can use R to get the linear regression coefficients for predicting runs from batting average and write down the linear regression equation.
Building the linear model:

```r
> fit <- lm(R ~ BA, data = team.batting.162)
```

Get the coefficients:

```r
> coef(fit)

(Intercept)    BA
-786.211  5797.657
```

Writing an equation using these coefficients

\[ \hat{r} = -786.2 + 5797.7 \cdot BA \]

Can you use this equation to predict the amount of run (\( \hat{r} \)) a team score if they have a particular BA of .300?
Residuals

The **residual** at a data value is the difference between the observed (y) and predicted value (ŷ) of the response variable

\[
\text{Residual} = \text{Observed} - \text{Predicted} = y - ŷ
\]
Run vs. batting average (2013)
Measuring goodness of fit

If the residuals are small, then the line does a good job describing the data.
We can measure how well the line fits the data using the equation:

\[ MSE = \frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 = \frac{1}{n} \sum_{i=1}^{n} (a + b \cdot x_i - y_i)^2 \]
Measuring goodness of fit

See if you can change the slope (a) and intercept (b) to find a line with a small MSE (link to the app is on Moodle)

https://emeyers.shinyapps.io/baseball_regression_app/
Calculating residuals for the runs as a function of batting average

<table>
<thead>
<tr>
<th>Team</th>
<th>BA</th>
<th>Runs obs (y) ((r \text{ in this case}))</th>
<th>Runs pred (ŷ) ((\hat{r} \text{ in this case}))</th>
<th>Residuals ((y - \hat{y})) ((r - \hat{r} \text{ in this case}))</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARI</td>
<td>.259</td>
<td>685</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ATL</td>
<td>.249</td>
<td>688</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BAL</td>
<td>.260</td>
<td>745</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BOS</td>
<td>.277</td>
<td>853</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CHA</td>
<td>.249</td>
<td>598</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[
\hat{r} = -529.8 + 4755.7 \cdot BA
\]
Residuals for the runs as a function of batting average

<table>
<thead>
<tr>
<th></th>
<th>BA</th>
<th>Runs obs (y) (r in this case)</th>
<th>Runs pred (ŷ) (r̂ in this case)</th>
<th>Residuals (y - ŷ) (r - ŷ in this case)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARI</td>
<td>.259</td>
<td>685</td>
<td>702.0</td>
<td></td>
</tr>
<tr>
<td>ATL</td>
<td>.249</td>
<td>688</td>
<td>702.0</td>
<td></td>
</tr>
<tr>
<td>BAL</td>
<td>.260</td>
<td>745</td>
<td>702.0</td>
<td></td>
</tr>
<tr>
<td>BOS</td>
<td>.277</td>
<td>853</td>
<td>702.0</td>
<td></td>
</tr>
<tr>
<td>CHA</td>
<td>.249</td>
<td>598</td>
<td>702.0</td>
<td></td>
</tr>
</tbody>
</table>

\[ \hat{r} = -529.8 + 4755.7 \cdot BA \]
Residuals for the runs as a function of batting average

<table>
<thead>
<tr>
<th></th>
<th>BA</th>
<th>Runs obs ( (y) ) ( (r \text{ in this case}) )</th>
<th>Runs pred ( (\hat{y}) ) ( (\hat{r} \text{ in this case}) )</th>
<th>Residuals ( (y - \hat{y}) ) ( (r - \hat{r} \text{ in this case}) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARI</td>
<td>.259</td>
<td>685</td>
<td>702.0</td>
<td>-17.0</td>
</tr>
<tr>
<td>ATL</td>
<td>.249</td>
<td>688</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BAL</td>
<td>.260</td>
<td>745</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BOS</td>
<td>.277</td>
<td>853</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CHA</td>
<td>.249</td>
<td>598</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[
\hat{r} = -529.8 + 4755.7 \cdot BA
\]
Residuals for the runs as a function of batting average

<table>
<thead>
<tr>
<th></th>
<th>BA</th>
<th>Runs obs (y) (r in this case)</th>
<th>Runs pred (ŷ) (r̂ in this case)</th>
<th>Residuals (y - ŷ) (r - r̂ in this case)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARI</td>
<td>.259</td>
<td>685</td>
<td>702.0</td>
<td>-17.0</td>
</tr>
<tr>
<td>ATL</td>
<td>.249</td>
<td>688</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BAL</td>
<td>.260</td>
<td>745</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BOS</td>
<td>.277</td>
<td>853</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CHA</td>
<td>.249</td>
<td>598</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Fill in the predicted values (ŷ) and the residuals (y – ŷ) for ATL, BAL, BOS and CHA

\[ \hat{r} = -529.8 + 4755.7 \cdot BA \]
Residuals for the runs as a function of batting average

<table>
<thead>
<tr>
<th>Team</th>
<th>BA</th>
<th>Runs obs (y) (r in this case)</th>
<th>Runs pred (ŷ) (r̂ in this case)</th>
<th>Residuals (y - ŷ) (r - r̂ in this case)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARI</td>
<td>.259</td>
<td>685</td>
<td>702.0</td>
<td>-17.0</td>
</tr>
<tr>
<td>ATL</td>
<td>.249</td>
<td>688</td>
<td>654.4</td>
<td>33.6</td>
</tr>
<tr>
<td>BAL</td>
<td>.260</td>
<td>745</td>
<td>706.7</td>
<td>38.3</td>
</tr>
<tr>
<td>BOS</td>
<td>.277</td>
<td>853</td>
<td>787.6</td>
<td>65.4</td>
</tr>
<tr>
<td>CHA</td>
<td>.249</td>
<td>598</td>
<td>654.4</td>
<td>-56.4</td>
</tr>
</tbody>
</table>

\[
\hat{r} = -529.8 + 4755.7 \cdot BA
\]
Sum of squared residuals

<table>
<thead>
<tr>
<th></th>
<th>Runs obs $(y)$</th>
<th>Runs pred $(\hat{y})$</th>
<th>Residuals $(y - \hat{y})$</th>
<th>Residuals$^2$ $(y - \hat{y})^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARI</td>
<td>685.0</td>
<td>702.0</td>
<td>-17.0</td>
<td>287.5</td>
</tr>
<tr>
<td>ATL</td>
<td>688.0</td>
<td>654.4</td>
<td>33.6</td>
<td>1129.0</td>
</tr>
<tr>
<td>BAL</td>
<td>745.0</td>
<td>706.7</td>
<td>38.3</td>
<td>1465.9</td>
</tr>
<tr>
<td>BOS</td>
<td>853.0</td>
<td>787.6</td>
<td>65.4</td>
<td>4282.4</td>
</tr>
<tr>
<td>CHA</td>
<td>598.0</td>
<td>654.4</td>
<td>-56.4</td>
<td>3181.0</td>
</tr>
</tbody>
</table>

Taking the average of these deviations yields the Mean Squared Error (MSE)

\[
MSE = \frac{1}{n} \sum_{i} (\hat{y}_i - y_i)^2
\]
Root Mean Squared Error

The square root of the MSE (RMSE) gives a sense of how much a points typically differ from the regression line.

RMSE for predicting runs from BA is 51.35
  • i.e., predictions typically off by 51 runs
Question

Where did the regression line come from?
The **least squares line**, also called "the line of best fit", is the line which **minimizes the sum of squared residuals** i.e., the least squares line is the line that minimizes the Mean Squared Error (MSE)

\[
MSE = \frac{1}{n} \sum_{i}^{n}(\hat{y}_i - y_i)^2
\]
Compare batting measures based on RMSE

Last class we found the ‘best’ statistic based on the correlation between each statistic and runs scored

$r = 0.74$

$r = 0.95$
Compare batting measures based on RMSE

We can also find the ‘best’ statistic based on the root mean squared error (RMSE)
- i.e., which statistic leads to a model with the minimal squared residuals
How to calculate the RMSE in R

\[ \text{MSE} = \frac{1}{n} \sum_{i} (\hat{y}_i - y_i)^2 \]

Let’s calculate the MSE when predicting a team’s Run totals from HR totals.

In this case, what are the \( y_i \)’s?

How would we get a vector of them using R?

\[
> \text{real.runs.y} \leftarrow \text{team.batting.162}\$R
\]

What is the equation for predicting runs from home runs?

\[
\hat{y} = \hat{r} = a + b \cdot \text{HR}
\]

How can we get the coefficients \( a \) and \( b \) to predict runs from HR in R?

\[
> \text{fit} \leftarrow \text{lm(}\text{team.batting.162}\$R \sim \text{team.batting.162}\$HR) \\
> \text{coef(fit)}
\]

How can we make predictions for each point \( y \)?

\[
> \text{predicted.runs.yhat} \leftarrow \text{coef(fit)[1]} + \text{coef(fit)[2]} * \text{team.batting.162}\$HR \\
> \text{predicted.runs.yhat} \leftarrow \text{predict(fit)} \quad \# \text{an alternative way}
\]
How to calculate the RMSE in R

\[ MSE = \frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 \]

Now we have:

- A vector of real runs \((y)\): \text{real.runs.y}
- A vector of predicted runs \((\hat{y})\): \text{predicted.runs.yhat}

How do we calculate the RMSE from these vectors?

1. create a vector of residuals by subtracting one vector from the other
   \[
   \text{residual.vec} \ <- \ \text{real.runs.y} - \text{predicted.runs.yhat}
   \]
2. Square the values in this vector of residuals
3. Sum these squared values
4. Divide by \(n\)
5. Take the square root
Compare batting measures based on RMSE

<table>
<thead>
<tr>
<th></th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>HR</td>
<td>60.76</td>
</tr>
<tr>
<td>BA</td>
<td></td>
</tr>
<tr>
<td>OBP</td>
<td></td>
</tr>
<tr>
<td>Slug</td>
<td></td>
</tr>
<tr>
<td>OPS</td>
<td></td>
</tr>
</tbody>
</table>
Compare batting measures based on RMSE

<table>
<thead>
<tr>
<th></th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>HR</td>
<td>60.76</td>
</tr>
<tr>
<td>BA</td>
<td>51.35</td>
</tr>
<tr>
<td>OBP</td>
<td></td>
</tr>
<tr>
<td>Slug</td>
<td></td>
</tr>
<tr>
<td>OPS</td>
<td></td>
</tr>
</tbody>
</table>
Compare batting measures based on RMSE

<table>
<thead>
<tr>
<th></th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>HR</td>
<td>60.76</td>
</tr>
<tr>
<td>BA</td>
<td>51.35</td>
</tr>
<tr>
<td>OBP</td>
<td>39.74</td>
</tr>
<tr>
<td>Slug</td>
<td>36.95</td>
</tr>
<tr>
<td>OPS</td>
<td>27.46</td>
</tr>
</tbody>
</table>
## Compare batting measures based on RMSE

\[
r^2 = 1 - \frac{\text{MSE}}{\text{var}(y)} \cdot \left[\frac{(n-1)}{n}\right]
\]

<table>
<thead>
<tr>
<th></th>
<th>RMSE</th>
<th>r</th>
</tr>
</thead>
<tbody>
<tr>
<td>HR</td>
<td>60.76</td>
<td>0.74</td>
</tr>
<tr>
<td>BA</td>
<td>51.35</td>
<td>0.82</td>
</tr>
<tr>
<td>OBP</td>
<td>39.74</td>
<td>0.90</td>
</tr>
<tr>
<td>Slug</td>
<td>36.95</td>
<td>0.91</td>
</tr>
<tr>
<td>OPS</td>
<td>27.46</td>
<td>0.95</td>
</tr>
</tbody>
</table>
Regression caution # 1

Avoid trying to apply the regression line to predict values far from those that were used to create the line. i.e., do not extrapolate too far
Regression caution # 2

Plot the data! Regression lines are only appropriate when there is a linear trend in the data.
Regression caution #3

Be aware of outliers – they can have an huge effect on the regression line.
Worksheet 4

source('/home/shared/baseball_stats_2017/baseball_class_functions.R')

get.worksheet(4)

As always, worksheet is due on midnight on Sunday
  • i.e., 11:59pm on March 5th

Start early and use Piazza and TA office hours!