# KEY Open-ended Practice Test (BaseballHits)

Here is the name of the dataset in DataCamp: BaseballHits

This is a dataset for 30 Major League Baseball teams from the 2010 season

1. What are the categorical variables in this data set? Which of these R functions could help you figure that out? Check all that apply: favstats(), tally(), select(), **head(), str()**

Team and League

1. Arrange the data from teams with the least to most wins, and save this arrangement as a new data frame cSEalled **WinView**. Write the R code here.

WinView <- arrange(BaseballHits, Wins)

1. Find out which team had the second-fewest wins in Major League Baseball in 2010.

SEA

Hint: can run head(WinView)

1. Are the contents of this dataset sample data (that is, a short run result of the Data Generating Process) or the population (that is, the long run result of the Data Generating Process)?

Sample; the teams will perform differently in the future and that future data is not in this data set. This is a short run result of the DGP.

1. Make histogram of the variable **Wins**. Write the R code here.

gf\_histogram(~Wins, data = BaseballHits)

Hint: They can use either BaseballHits or WinView. All the results will be the same. Why? Because WinView has the same data (same numbers) just sorted in a different order.

1. The default number of bins in a histogram is 30. What is the R code to see only 8 bins in your histogram?

gf\_histogram(~Wins, data = BaseballHits, bins = 8)

1. How would you describe the frequency distribution of the variable **Wins**? What is the shape, center, and spread? Do you notice any weird things?

A little bit skewed, center is probably between 80 and 90, we can eyeball the standard deviation as around 10.

I guess one weird thing I would ask myself is -- why would Wins be a little skewed? Are there only a few teams that really don’t win a lot but most of the teams are clustered together?

If they want, they can run favstats and see that indeed, the mean for Wins is 81 games and the standard deviation is 11.

1. Could we model the distribution of **Wins** with a normal distribution? Why or why not?

Sure, it could have come from a normal distribution. This may not look very normal because it’s a sample of 30 teams in one season. The DGP might very well be normal even if the sample doesn’t look exactly normal.

1. Wins constitute an important outcome variable for analyzing baseball data. What would the empty model predict for SEA (Seattle Mariners)? How about for NYY (New York Yankees)?

81

Hint: they could just use the favstats or they can run lm(Wins ~ NULL, data = BaseballHits)

1. There are data from teams from two leagues (AL and NL) in this data frame. Is **League** a good explanatory variable for **Wins**? Make a visualization to explore this idea. Also write this as a word equation.

Potential visualizations:

gf\_histogram(~Wins, data = BaseballHits) %>%

gf\_facet\_grid(League ~ .)

gf\_boxplot(Wins ~ League, data = BaseballHits)

gf\_point(Wins ~ League, data = BaseballHits)

gf\_jitter(Wins ~ League, data = BaseballHits)

Just from exploring the data with visualizations, League does NOT seem like a good explanatory variable (we wouldn’t be able to make a better guess about Wins if we knew what League the team was in).

Word equation: Wins = League + other stuff

1. Write R code to find the average Wins for AL (American League) and NL (National League). On average, which League wins more games?

favstats(Wins ~ League, data = BaseballHits)

On average, AL wins more games than NL (not by much though, just 1 game difference).

1. Does this prove that being in one of the leagues causes a team to win more games?

No, this is not data from an experiment.

Hint: Some people might say, this is not a big enough difference to suggest causation. This is true that this difference isn’t very big (and we’ll figure out whether we should even pay attention to it later) but that is not the reason why we can’t figure out causation.

1. Why is the grand mean a good model for the distribution of **Wins**?

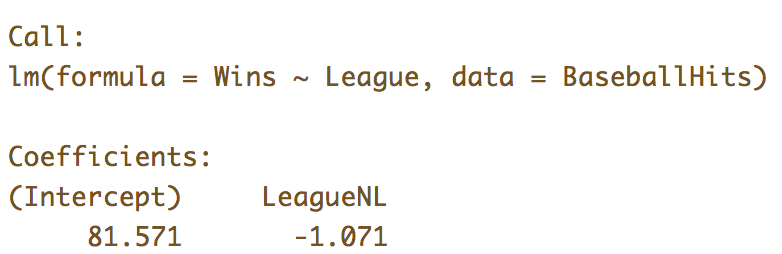
Because it is the one number that minimizes SS (more specifically, SS Total). The grand mean (81) is better than any other number we could use as the empty model (e.g., 82, 80, 81.5).

1. Fit a complex model that predicts **Wins** based on **League**. Interpret the resulting numbers.

lm(Wins ~ League, data = BaseballHits)

Alternatively they can save the complex model.

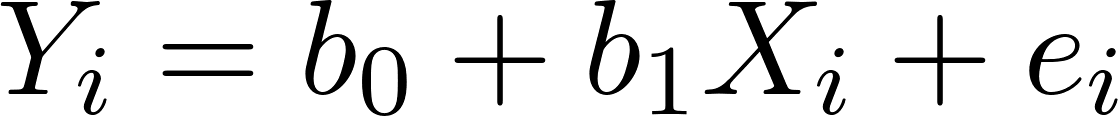
League.model <- lm(Wins ~ League, data = BaseballHits)

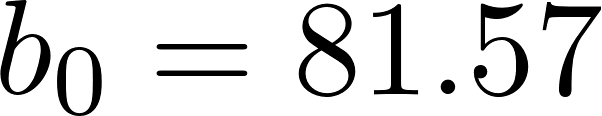


81.57 is the mean Wins for AL teams

-1.07 is what you add on to 81.57 to get the mean Wins for NL teams

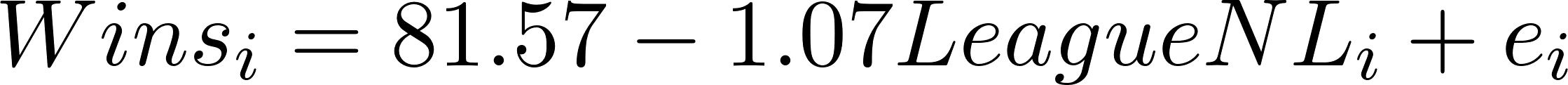
1. What would be the GLM notation used to represent the **League** model? Which numbers would go with which part of the GLM notation? What does the Y and X mean?

[](https://www.codecogs.com/eqnedit.php?latex=Y_i%20%3D%20b_0%20%2B%20b_1X_i%20%2B%20e_i%0)

[](https://www.codecogs.com/eqnedit.php?latex=b_0%20%3D%2081.57%0)

[](https://www.codecogs.com/eqnedit.php?latex=b_1%20%3D%20-1.07%0)

Another way you can think of the GLM notation is:

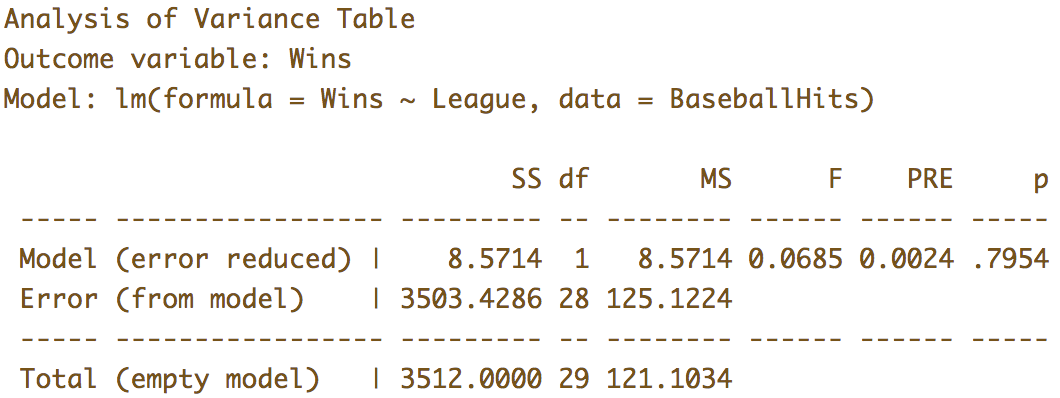
[](https://www.codecogs.com/eqnedit.php?latex=Wins_i%20%3D%2081.57%20-%201.07LeagueNL_i%20%2B%20e_i%0)

LeagueNL would be 0 for AL teams and 1 for NL teams

1. Which has a bigger SS: the empty model or the **League** model?

Empty model (3512) has bigger SS than League model (3503).

Hint: supernova(League.model)



1. Which R function would you use to compare the League model to the empty model?

supernova()

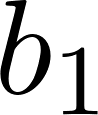
1. The PRE from the **League** model is .0024. What does the PRE mean?

Less than 1% of the leftover error (SS Total) has been reduced by the League model.

OR

.0024 of the leftover error (SS Total) has been explained by the League model.

1. Why is the degrees of freedom for the **League** model equal to 1?

Because we used up (or spent) one more degree of freedom to estimate [](https://www.codecogs.com/eqnedit.php?latex=b_1%0).

1. Let’s try to explain variation in **Wins** using **Walks**. Make a visualization to explore this idea of the DGP. What would be the word equation for this idea?

gf\_point(Wins ~ Walks, data = BaseballHits)

Wins = Walks + other stuff

1. Create a variable that splits the teams up into two groups, those with few walks and those with more walks. Call it **Walks2group** and make it a factor.

BaseballHits$Walks2group <- ntile(BaseballHits$Walks, 2)

BaseballHits$Walks2group <- factor(BaseballHits$Walks2group, levels = c(1,2), labels = c("few walks", "many walks"))

OR

BaseballHits$Walks2group <- ntile(BaseballHits$Walks, 2)

BaseballHits$Walks2group <- as.factor(BaseballHits$Walks2group)

OR

BaseballHits$Walks2group <- as.factor(ntile(BaseballHits$Walks, 2))

1. Create a visualization to explore whether **Walks2group** could explain some of the variation in **Wins**. Just from exploring the visualizations, which explanatory variable is better at predicting **Wins**: **Walks2group** or **League**?

Potential visualizations:

gf\_histogram(~Wins, data = BaseballHits) %>%

gf\_facet\_grid(Walks2group ~ .)

gf\_boxplot(Wins ~ Walks2group, data = BaseballHits)

gf\_point(Wins ~ Walks2group, data = BaseballHits)

gf\_jitter(Wins ~ Walks2group, data = BaseballHits)

**Walks2group** looks like knowing whether a team was a few or many walks team would help us make a better guess about their **Wins** because in the many walks group, the distribution is shifted up relative to the few walks group. **League** doesn’t seem as helpful because the distributions overlap more.

1. Create a model of **Walks2group**: specify the model in GLM notation, fit the model using R, and then interpret the numbers.

lm(Wins ~ Walks2group, data = BaseballHits)

OR

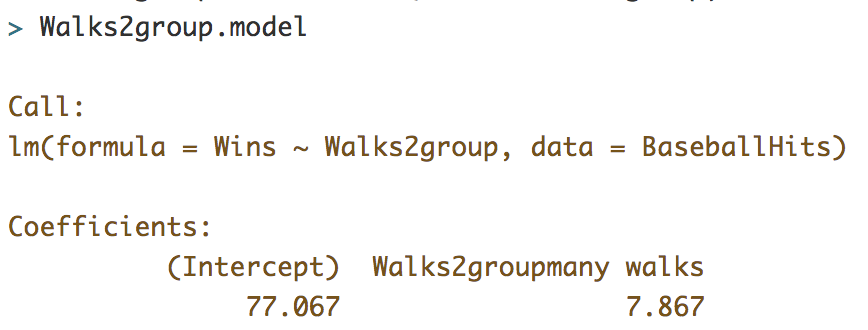
Walks2group.model <- lm(Wins ~ Walks2group, data = BaseballHits)

1. What would the **Walks2group** model predict for NYY’s wins? What would it predict for SEA’s wins?

We can just look at whether NYY is in the few or many walks group and then figure out which group mean would be the **Walks2group** model’s prediction.

select(BaseballHits, Team, Walks2group)

We then see that NYY is many walks and SEA is few walks.



We could look at the **Walks2group** model (above) and predict 77.07 for the few walks group and 77.07+7.87 for the many walks group.

OR

We can actually make the predictions from **Walks2group.model** and look at them.

Walks2group.model <- lm(Wins ~ Walks2group, data = BaseballHits)

BaseballHits$predicted <- predict(Walks2group.model)

select(BaseballHits, Team, Walks2group, predicted)

1. Let’s say we run supernova on both the League model and the **Walks2group** model. They will both have the same SS Total. Why?

Because the SS Total is calculated from the empty model (the mean of **Wins**) and that empty model is the same in both supernova tables.

1. Let’s say we run supernova on both the **League** model and the **Walks2group** model. Which one has a bigger PRE? Why?

**Walks2group** has a bigger PRE because it does a better job reducing the leftover error than **League** model. Another way of saying it is that **Walks2group** explains more variation than the **League** model. The **League** model explains less than .01 (.0024 to be exact) of the variation in **Wins** while **Walks2group** model explains almost .05 of the variation.