Simple Linear Regression

Modern Data Mining

1. Objectives
2. Case Study: Baseball
   - Data
   - EDA
3. Simple linear regression
   - Model specification
   - OLS estimates and properties
   - R-squared and RSE
   - Confidence intervals for coefficients
   - Prediction intervals
   - Model diagnoses
4. Summary

Objectives

- Data Science is a field of science. We try to extract useful information from data. In order to use the data efficiently and correctly we must understand the data first. According to the goal of the study, combining the domain knowledge, we then design the study. In this lecture we first go through some basic explore data analysis to understand the nature of the data, some plausible relationship among the variables.
- Data mining tools have been expanded dramatically in the past 20 years. Linear model as a building block for data science is simple and powerful. We introduce/review simple linear model. The focus is to understand what is it we are modeling; how to apply the data to get the information; to understand the intrinsic variability that statistics have.

Table of Content

- Suggested readings:
  - The full lecture: `Regression_1_simple_regression.html`
  - Data set: `MLPayData_Total.csv`
- Case Study
- EDA
- Simple regression
  - Model specification
  - OLS estimates and properties
  - R-squared and RSE
  - Confidence intervals for coefficients
  - Prediction intervals
  - Model diagnoses
- R-functions
  - lm()

Case Study: Baseball

- Baseball is one of the most popular sports in US. The highest level of baseball is Major League Baseball which includes American League and National League with total of 30 teams. New York Yankees, Boston Red Sox, Philadelphia Phillies and recently rising star team Oakland Athletics are among the top teams. Oakland A's is a low budget team. But the team has been moving itself up mainly due to its General Manager (GM) Billy Beane who is well known to apply statistics in his coaching.
- Questions of interests for us:
  - **Q1:** Will a team perform better when they are paid more?
  - **Q2:** Is Billy Beane (Oakland A's GM) worth 12.5 million dollars for a period of 5 years, as offered by the Red Sox?
- Read an article: Billion dollar Billy Beane.

Data

`MLPayData_Total.csv`, consists of winning records and the payroll of all 30 ML teams from 1998 to 2014 (17 years). There are 162 games in each season. We will create the following two exaggerated variables:

- `payroll`: total pay from 1998 to 2014 in **billion dollars**
- `win`: average winning percentage for the span of 1998 to 2014

To answer **Q1**: relationship $Y =$ win $X =$ payroll

1. How does `payroll` relate to the performance measured by `win`?
2. Given payroll= .84,

- on average what would be the **mean** winning percentage
- is Oakland A's performance super unusual?

(Oakland A's: payroll=.84, win=.54 and Red Sox: payroll=1.97, win=.55)

Data Preparation

We would normally do a thorough EDA. We skip that portion of the data analysis and get to the regression problem directly. Before that, let us take a quick look at the data and extract aggregated variables.

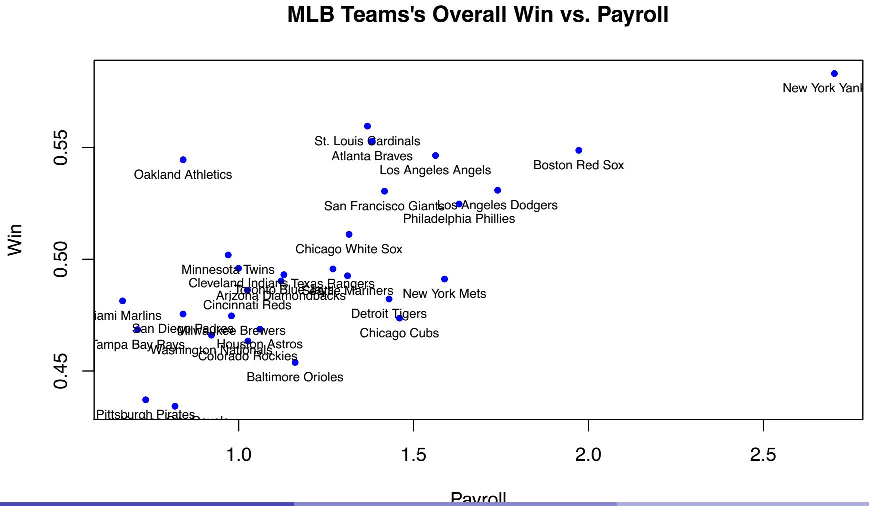
```
baseball <- read.csv("MLPayData_Total.csv")
# names(baseball)
datapay <- baseball %>%
  rename(team = "Team.name.2014",
         win = avgwin) %>%
  select(team, payroll, win)
```

Scatter Plot: Explore the relationship between ‘payroll‘, and ‘win‘

```
plot(x = datapay$payroll,
     y = datapay$win,
     pch  = 16,     # "point character": shape/character of points
     cex  = 0.8,    # size
     col  = "blue", # color
     xlab = "Payroll",  # x-axis
     ylab = "Win",  # y-axis
     main = "MLB Teams's Overall Win vs. Payroll") # title
# label all points
text(datapay$payroll, datapay$win,
     labels=datapay$team, cex=0.7,
     pos=1) # position 1,2,3,4: below, left, above, right of (x,y)
```

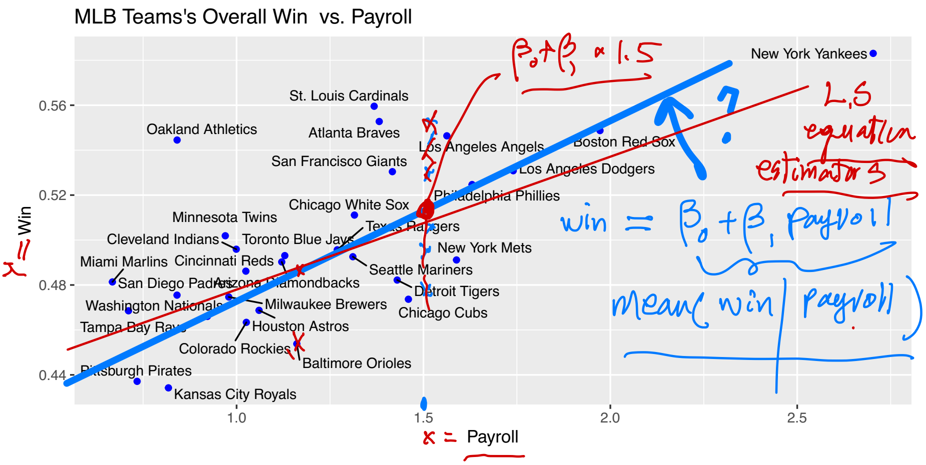
Scatter Plot: Explore the relationship between 'payroll', and 'win'

We notice the positive association: when `payroll` increases, so does `win`.



Plotting using ggplot

```
ggplot(datapay) +
  geom_point(aes(x = payroll, y = win), color = "blue") +
  geom_text_repel(aes(x = payroll, y = win, label=team), size=3) +
  labs(title = "MLB Teams's Overall Win  vs. Payroll", x = "Payroll", y = "Win")
```



Notes: Q1: function format ? $\underline{y = b_0 + b_1 X}$

why linear: a) appears like a linear

b) simple: suppose $\underline{\text{win} = .4 + .065 \text{ payroll}}$

interpretation slope = .065 informative

Q2: what does this unknown func. model

$$\underline{\text{win} = \beta_0 + \beta_1 \cdot \text{Payroll}}$$

$$\text{mean}(\text{win} \mid \text{payroll}) = \beta_0 + \underbrace{\beta_1 \cdot \text{Payroll}}_{\text{mean}}$$

Q3: $y_i = \beta_0 + \beta_1 \cdot x_i + \varepsilon_i$ Linear model !!!

win Payroll error

1 Objectives

2 Case Study: Baseball

- Data
- EDA

3 Simple linear regression

- Model specification
- OLS estimates and properties
- R-squared and RSE
- Confidence intervals for coefficients
- Prediction intervals
- Model diagnoses

4 Summary

Summary: Idealize linear model

Assume.

$$E(y \mid x) = \beta_0 + \beta_1 \cdot x$$

Next: Estimate unknown parameters β_0, β_1 !!

Simple Linear Regression

Often we would like to explore the relationship between two variables. Will a team perform better when they are paid more? The simplest model is a linear model. Let the response y_i be the `win` and the explanatory variable x_i be `payroll` ($i = 1, \ldots, n = 30$).

Assume there is linear relationship between `win` and `payroll`, i.e.

$$y_i = \underbrace{\beta_0 + \beta_1 x_i}_{\text{mean}(y_i \mid x_i)} + \epsilon_i$$

Model interpretation

- We assume that given `payroll`, on average the `win` is a linear function
- For each team the `win` is the average plus an error term
- Unknown parameters of interest
  - intercept: β_0
  - slope: β_1
  - both are unknown

***

use data estimate them

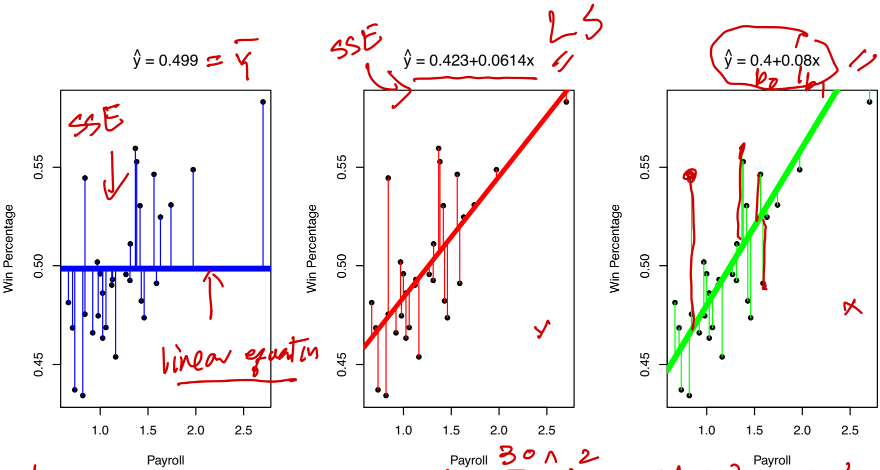
$\Rightarrow$ $\underbrace{\hat{\beta}_0, \hat{\beta}_1}$

numbers / stat

Estimation

How to estimate the parameters using the data we have? For example, how would you decide on the following three equations?

Estimation



$y = .499 + 0 \times \text{payroll}$

$$SSE = \sum_{i=1}^{30} (e_i)^2 = (\text{dis}_1)^2 + (\text{dis}_2)^2 \cdots (\text{dis}_{30})^2$$

b_0, b_1

Green

Ordinary least squares (OLS) estimates

Given an estimate (b_0, b_1), we first define residuals as the differences between actual and predicted values of the response y, i.e.

$$\hat{\epsilon}_i = \hat{y}_i - b_0 - b_1 x_i.$$

In previous plots, the residuals are the vertical lines between observations and the fitted line. Now we are ready to define the OLS estimate.

The OLS estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are obtained by minimizing sum of squared errors (RSS):

$$\cancel{(\hat{\beta}_0, \hat{\beta}_1) = \arg} \min_{b_0,\, b_1} \boxed{\sum_{i=1}^{n=30} (y_i - b_0 - b_1 x_i)^2. \quad (x_i, y_i) \text{ data}} = \text{quadratic equ. for } (b_0, b) \quad : \text{numbers}$$

Note: This is a quadratic equation of b_0 and b_1.

Ordinary least squares (OLS) estimates

We can derive the solution $\hat{\beta}_0$ and $\hat{\beta}_1$.

$$\hat{\beta}_1 = r_{xy} \cdot \frac{s_y}{s_x}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

formulae

where

- $\bar{x}$ = sample mean of x's (payroll)
- $\bar{y}$ = sample mean of y's (win)
- s_x = sample standard deviation of x's (payroll)
- s_y = sample standard deviation of y's (win)
- r_{xy} = sample correlation between x and y.

‘lm()‘

The function `lm()` will be used extensively. This function solves the minimization problem that we defined above. Below we use `win` as the dependent y variable and `payroll` as our x.

As we can see from the below output, this function outputs a list of many statistics. We will define these statistics later

```
myfit0 <- lm(win ~ payroll, data=datapay)
names(myfit0)
```

```
##  [1] "coefficients"  "residuals"     "effects"       "rank"
##  [5] "fitted.values" "assign"        "qr"            "df.residual"
##  [9] "xlevels"       "call"          "terms"         "model"
```

‘lm()‘

We can also view a summary of the `lm()` output by using the `summary()` command.

```
summary(myfit0)   # it is another object that is often used
results <- summary(myfit0)
names(results)
```

```
##
## Call:
## lm(formula = win ~ payroll, data = datapay)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.04003 -0.01749  0.00094  0.01095  0.07030
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.4226     0.0153   27.56  < 2e-16 ***
## payroll       0.0614     0.0117    5.23  1.5e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.027 on 28 degrees of freedom
## Multiple R-squared:  0.494,  Adjusted R-squared:  0.476
## F-statistic: 27.4 on 1 and 28 DF,  p-value: 1.47e-05
##
##  [1] "call"          "terms"         "residuals"     "coefficients"
##  [5] "aliased"       "sigma"         "df"            "r.squared"
##  [9] "adj.r.squared" "fstatistic"    "cov.unscaled"
```

2.3:

$\widehat{win} = .4226 + .0614\ payroll$

$\hat{\beta}_1 = .06$

β_1

Stat

'lm()'

Day 5: Recap: y = win Assume
x = pay ① $\text{Ave}(\text{win} \mid \text{pay}) = \beta_0 + \beta_1 \text{ pay}$

To summarize the OLS estimate, $\hat{\beta}_0 = 0.4$ and $\hat{\beta}_1 = 0.08$, we have the following estimator:

② $y_i \mid x_i = \beta_0 + \beta_1 \text{ pay}_i + \varepsilon_i$

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = 0.423 + 0.061 \cdot x_i$$

Notice that the outputs of `myfit0` and `summary(myfit0)` are different

③ Est. β_0, β_1 (ε_i)

LS: $\hat{\beta}_0, \hat{\beta}_1$

Solve $\min_{b_0, b_1} \sum_{i=1}^{30} \left(y_i - (b_0 + b_1 x_i) \right)^2$

```
myfit0
b0 <- myfit0$coefficients[1]
b1 <- myfit0$coefficients[2]
```

```
##
## Call:
## lm(formula = win ~ payroll, data = datapay)
##
## Coefficients:
## (Intercept)      payroll
##      0.4226       0.0614
```

④ lm(win ~ pay, data)

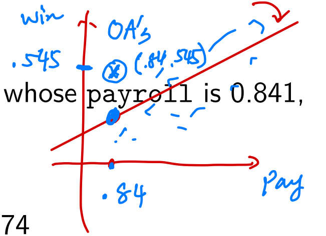
$\Longrightarrow \hat{\beta}_1 = .061, \ \hat{\beta}_0 = .423$

Interpretation of the slope

- When `payroll` increases by 1 unit (1 billion), we expect, on average the `win` will increase about 0.061. $= \hat{\beta}_1$
- When `payroll` increases by .5 unit (500 million), we expect, on average the `win` will increase about 0.031.

Mean estimation

$\hat{y}_{|x} = .423 + .061 \times x$



Data: OA's: pay = .841, win = .545

- For all the team similar to Oakland Athletics whose payroll is 0.841, we estimate on average the win to be

$$\text{Ave } y\Big|_{Pay=.84} = 0.423 + 0.061 \times 0.841 = \underline{.474}$$

- **Residuals**: For Oakland Athletics, the real win is 0.841. So the residual for team Oakland Athletics is

model:

$$y_{OA} = \underbrace{.423 + .061 \times .84}_{.474} + \varepsilon_{OA}$$

.545

ε_{OA} = unknown

$$\hat{\epsilon}_{\text{Oakland}} = .545 - .474 = .071$$

Fitted Values

Here are a few rows that show the fitted values from our model $= \hat{y}_1, \hat{y}_2 \ldots \hat{y}_{30}$

```
data.frame(datapay$team, datapay$payroll, datapay$win, myfit0$fitted,
           myfit0$res)[15:25, ] # show a few rows
```

Data

$\hat{y}$: mean est on the line

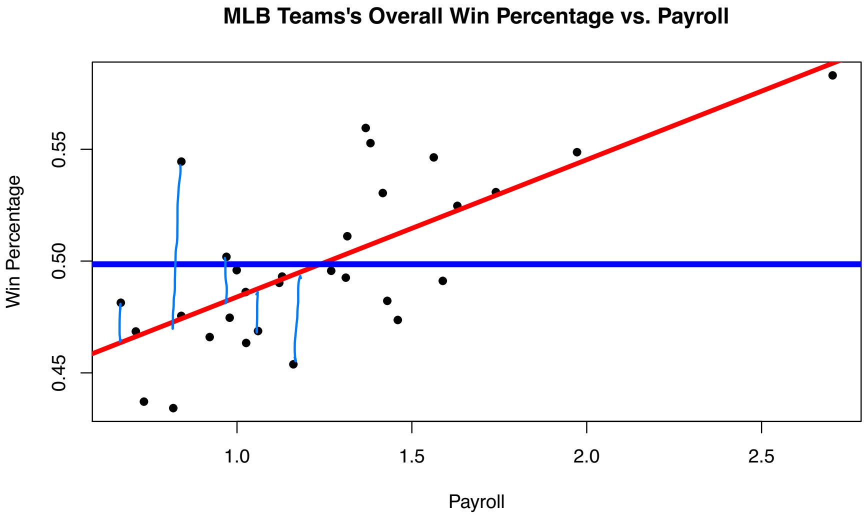
$\hat{e}$

```
##           datapay.team datapay.payroll datapay.win myfit0.fitted
## 15        Miami Marlins           0.668       0.481         0.464
## 16    Milwaukee Brewers           0.979       0.475         0.483
## 17      Minnesota Twins           0.970       0.502         0.482
## 18        New York Mets           1.588       0.491         0.520
## 19     New York Yankees           2.703       0.583         0.588
## 20    Oakland Athletics           0.841       0.545         0.474
## 21 Philadelphia Phillies          1.630       0.525         0.523
## 22    Pittsburgh Pirates          0.734       0.437         0.468
## 23      San Diego Padres          0.841       0.475         0.474
## 24  San Francisco Giants          1.417       0.530         0.510
## 25      Seattle Mariners          1.311       0.493         0.503
##    myfit0.res
## 15    0.01778
## 16   -0.00803
## 17    0.01979
## 18   -0.02894
## 19   -0.00542
## 20    0.07030
## 21    0.00207
## 22   -0.03051
## 23    0.00130
## 24    0.02089
## 25   -0.01048
```

Scatter plot with the LS line added

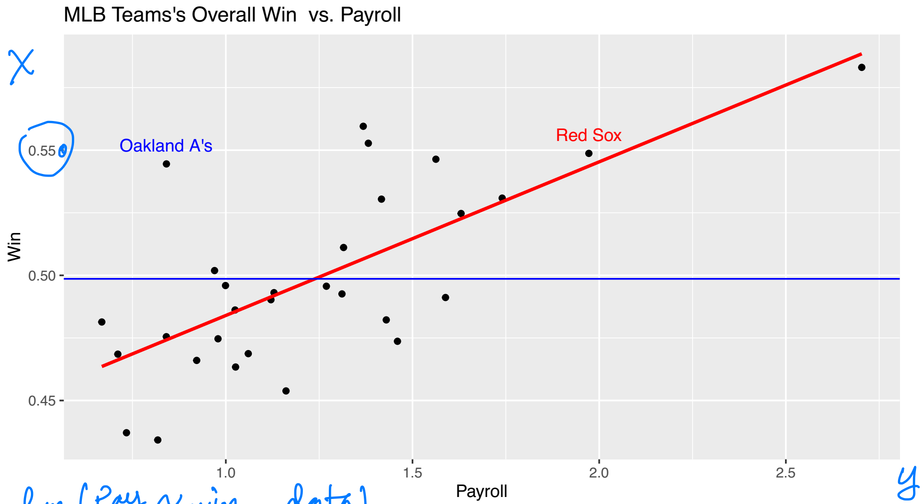
Base R

$\hat{e} = y - \hat{y}$



Scatter plot with the LS line added

ggplot



lm(Pay ~ win, data)

HERE is how the article concludes that Beane is worth as much as the GM in Red Sox. By looking at the above plot, Oakland A's win pct is more or less same as that of Red Sox, so based on the LS equation, the team should have paid 2 billion.

Do you agree on this statement? What is the right analysis?

Goodness of Fit

Q1: Is the LS line "good enough"

Q2: Confidence interval for β_1

How well does the linear model fit the data? A common, popular notion is through R^2.

Residual Sum of Squares (RSS): "Inference"

The least squares approach chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize the RSS. RSS is defined as:

Q3: CI for mean(y|x)

$$RSS = \sum_{i=1}^{n} \hat{\epsilon}_i^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

Q4: Prediction interval

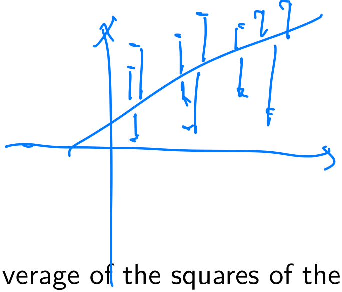
```
myfit0 <- lm(win~payroll, data=datapay)
RSS <- sum((myfit0$res)^2) # residual sum of squares
RSS
```

```
## [1] 0.0204
```

Q1: Good (line is)?

LS: $\hat{y} = .423 + .061 \cdot x$ how good is this??

Goodness of Fit



Mean Squared Error (MSE):

Mean Squared Error (MSE) is the average of the squares of the errors, i.e. the average squared difference between the estimated values and the actual values. For simple linear regression, MSE is defined as:

$$MSE = \frac{RSS}{n-2}.$$

Goodness of Fit

Residual Standard Error (RSE)/Root-Mean-Square-Error (RMSE):

Residual Standard Error (RSE) is the square root of MSE. For simple linear regression, RSE is defined as:

$$RSE = \sqrt{MSE} = \sqrt{\frac{RSS}{n-2}}.$$

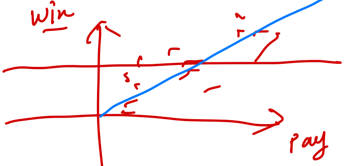
```
sqrt(RSS/myfit0$df)
summary(myfit0)$sigma
```

```
## [1] 0.027
## [1] 0.027
```

Can we use RSS, MSE or RMSE to check how good the LS equation works?

Goodness of Fit

$\hat{Y} = \bar{Y}$



Total Sum of Squares (TSS):

TSS measures the total variance in the response Y, and can be thought of as the amount of variability inherent in the response before the regression is performed. In contrast, RSS measures the amount of variability that is left unexplained after performing the regression.

$$TSS = \sum_{i=1}^{n}(y_i - \bar{y}_i)^2$$

TSS = 0
stop here

Nothing needs to be done

```
TSS <- sum((datapay$win-mean(datapay$win))^2) # total sum of sqs
TSS
```

```
## [1] 0.0403
```

Goodness of Fit

R^2:

R^2 measures the proportion of variability in Y that can be explained using X. An R^2 statistic that is close to 1 indicates that a large proportion of the variability in the response has been explained by the regression.

$$R^2 = \frac{TSS - RSS}{TSS} = \frac{TSS = .04 - RSS = .02}{.04} = .02$$

$= .02 / .04 = .5$

def

```
(TSS-RSS)/TSS    # Percentage reduction of the total errors
(cor(datapay$win, myfit0$fit))^2 # Square of the cor between response and fitted values
summary(myfit0)$r.squared
```

```
## [1] 0.494
## [1] 0.494
## [1] 0.494
```

Interpretation: Total variation in y explained by "this" LS line

Remarks

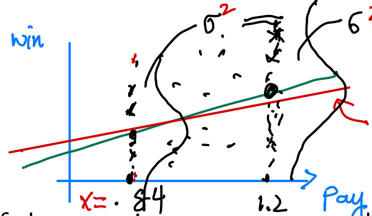
- How large R^2 needs to be so that you are comfortable to use the linear model?
- Though R^2 is a very popular notion of goodness of fit, but it has its limitation.

1. Objectives
2. Case Study: Baseball
   - Data
   - EDA
3. Simple linear regression
   - Model specification
   - OLS estimates and properties
   - R-squared and RSE
   - **Confidence intervals for coefficients**
   - Prediction intervals
   - Model diagnoses
4. Summary

σ^2 : Var y from the line

$$\underline{MSE} = \hat{\sigma}^2 = \frac{(\hat{e}_1^2 + \hat{e}_2^2 + \dots + \hat{e}_{30}^2)}{30 - 2} =$$

Inference



Assumptions:

1. linearity $win = \beta_0 + \beta_1 \, pay$
2. dist'n $y|x$
3. equal variances

true line $win = \beta_0 + \beta_1 \cdot Pay$

LS line $win = \hat\beta_0 + \hat\beta_1 \cdot Pay$

One $\hat\beta_1 = .061$ $\quad \hat\beta_0 = .423$

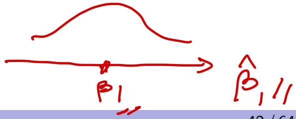
- One of the most important aspect about statistics is to realize the estimators or statistics we propose, such as the least squared estimators for the slope and the intercept, they change as a function of data. $y_i \mid x = .84 \sim$ Normal with $\beta_0 + \beta_1 \cdot .84$, σ^2 unknown
- Understanding the variability of the statistics, providing the accuracy of the estimators are one of the focus as statisticians.
- ~~In order to assess the accuracy of the OLS estimator, we need~~ assumptions!

We use $\hat\beta_1$ to est β_1

Have: 95% CI for β_1: $\hat\beta_1 \pm 2 * SD(\hat\beta_1)$

Can we construct CI's for β_1 based on $\hat\beta_1$???

Need to $\hat\beta_1$ dis'n ?



Linear model assumptions

Recall

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i.$$

We did not impose assumptions on ϵ_i when using OLS. In order to provide some desired statistical properties and guarantees to our OLS estimate $(\hat{\beta}_0, \hat{\beta}_1)$, we need to impose assumptions.

- Linearity:

$$\mathbf{E}(y_i | x_i) = \beta_0 + \beta_1 x_i$$

- Homoscedasticity:

$$\mathbf{Var}(y_i | x_i) = \sigma^2$$

- Normality:

$$\epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$$

or

$$y_i \overset{iid}{\sim} \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2)$$

Second i is wrong

iid : independent

y_i

identical dis^n,

Inference for the coefficients: β_0 and β_1

Under the model assumptions:

1. y_i independently and identically normally distributed
2. The mean of y given x is linear
3. The variance of y does not depend on x

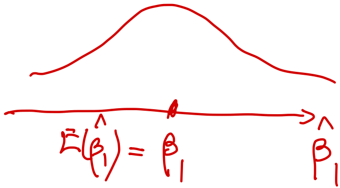
Inference for the coefficients: β_0 and β_1

Facts: Mathematics

The OLS estimates $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1)$ has the following properties:

1. Unbiasedness

$$\mathbf{E}(\hat{\beta}) = \beta$$



2. Normality

How to est σ^2 ???

$$\hat{\beta}_1 \sim \mathcal{N}(\beta_1, \mathbf{Var}(\hat{\beta}_1))$$

where

Est. $\mathrm{Var}(\hat{\beta}_1)$

$$\frac{\hat{\sigma}^2 = MSE}{\sum_{i=1} (X_i - \bar{X})^2}$$

unknown

$$\mathbf{Var}(\hat{\beta}_1) = \frac{\sigma^2}{x_x^2} = \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}.$$

data: Given

Inference for the coefficients: β_0 and β_1

(ignore this)

$$X = \begin{array}{cc} & \text{Pay} \\ \left(\begin{array}{cc} 1 & x_1 \\ \vdots & x_2 \\ 1 & \vdots \\ \vdots & \\ 1 & x_{30} \end{array}\right) \end{array} \qquad Y = \begin{pmatrix} y_1 \\ \vdots \\ y_{30} \end{pmatrix}$$

In general,

$$\hat{\beta} \sim \mathcal{N}(\beta,\ \sigma^2(X^T X)^{-1})$$

Here $\underline{X}$ is the design matrix where the first column is 1's and the second column is the values of x, in our case it is the column of `payroll1`.

2 : df

Two β's to est, β_0, β_1

$$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \qquad Var\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \sigma^2 (X^T X)^{-1}$$

$$\underline{MSE} = \frac{RSS}{n-2} = \frac{\hat{e}_1^2 + \cdots + \hat{e}_{30}^2}{28} = \hat{\sigma}^2 \text{ est } \sigma^2$$

Theorem: $E(MSE) \approx \sigma^2$ ✓

Confidence intervals for the coefficients

t-interval and t-test can be constructed using the above results.

For example, the 95% confidence interval for β approximately takes the form

95% CI: β_1

$$\hat{\beta} \pm 2 \cdot SE(\hat{\beta}).$$

$\hat{\beta}_1 \pm 2 \cdot SE(\hat{\beta}_1)$

stat

standard error $(\hat{\beta}_1)$

$\because$ est σ^2 in the $Va(\hat{\beta}_1)$

We can also perform hypothesis test on the coefficients. To be specific, we have the following test.

$$H_0 : \beta_1 = 0 \text{ v.s. } H_1 : \beta_1 \neq 0$$

Remark: t-distribution $\approx$ Normal distribution

Tests to see if the slope is 0

To test the null hypothesis, we need to decide whether $\hat{\beta}_1$ is far away from 0, which depends on $SE(\hat{\beta}_1)$. We now define the test statistics as follows.

$$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$$

Under the null hypothesis $\beta_1 = 0$, t will have a t-distribution with $(n-2)$ degrees of freedom. Now we can compute the probability of $T \sim t_{n-2}$ equal to or larger than $|t|$, which is termed $p - value$. Roughly speaking, a small p-value means the odd of $\beta_1 = 0$ is small, then we can reject the null hypothesis.

Tests to see if the slope is 0

$SE(\hat{\beta})$, t-value and p-value are included in the summary output.

```
summary(myfit0)
```

myfit0 = lm(   )

```
##
## Call:
## lm(formula = win ~ payroll, data = datapay)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.04003 -0.01749  0.00094  0.01095  0.07030
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.4226     0.0153   27.56  < 2e-16 ***
## payroll       0.0614     0.0117    5.23  1.5e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.027 on 28 degrees of freedom
## Multiple R-squared:  0.494,  Adjusted R-squared:  0.476
## F-statistic: 27.4 on 1 and 28 DF,  p-value: 1.47e-05
```

$\hat{\sigma}$

① $\widehat{win} = .4226 + .0614$ pay

② 95% CI for β_1

$.0614 \pm 2 \times .0117$

$= (.04, .08)$ //

β_1 is a number .04, .08

$\beta_1 \neq 0$

③ $\begin{cases} H_0: \beta_1 = 0 \\ H_1: \beta_1 \neq 0 \end{cases}$

step 1: $z = \frac{.0614 - 0}{.0117} = 5.23$

step 2: p-value $= 2 \times P(Z > 5.23) = .00015$

Tests to see if the slope is 0

conclusion: reject H_0 at $\alpha = .01$

(4) $R^2 = .494$ (5) $\hat{\sigma}$: RSE = .027

The `confint()` function returns the confident interval for us (95% confident interval by default).

```
confint(myfit0)
confint(myfit0, level = 0.99)
```

```
##              2.5 % 97.5 %
## (Intercept) 0.3912 0.4540
## payroll     0.0373 0.0854
##             0.5 % 99.5 %
## (Intercept) 0.380 0.4650
## payroll     0.029 0.0938
```

Confidence for the mean response

We use a confidence interval to quantify the uncertainty surrounding the mean of the response (`win`). For example, for teams like Oakland A's whose `payroll=.841`, a 95% Confidence Interval for the mean of response `win` is

95% $\quad x_i$

$$\hat{y}_{|x=.841} \pm t_{(\alpha/2, n-2)} \times \sqrt{MSE \times \left(\frac{1}{n} + \frac{(.841 - \bar{x})^2}{\sum (x_i - \bar{x})^2}\right)}.$$

$2 \qquad SE(\hat{y}_{x=.841})$

$\hat{y} \pm 2 \cdot SE(\hat{y})$

Stat $\qquad$ SE(Stat)

Confidence for the mean response

unbiased estimators.

$$\begin{cases} E(MSE) = \sigma^2 \\ E(\hat\beta_1) = \beta_1 \\ E(\hat\beta_0) = \beta_0 \end{cases}$$

The `predict()` provides prediction with confidence interval using the argument `interval="confidence"`.

$\hat\sigma$

```
new <- data.frame(payroll=c(.841))   #new <- data.frame(payroll=c(1.24))
CImean <- predict(myfit0, new, interval="confidence", se.fit=TRUE)
CImean
```

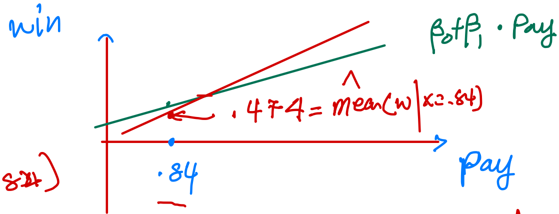
mean CI

mean estimate with CI

```
## $fit
##     fit  lwr   upr
## 1 0.474 0.46 0.488
##
## $se.fit
## [1] 0.00678
##
## $df
## [1] 28
##
## $residual.scale
## [1] 0.027
```

$= \hat\sigma$

$= SE(\hat y_{x=.84})$



Q1:

95% CI for mean win $| X=.84$:

$= \beta_0+\beta_1 \cdot .84$

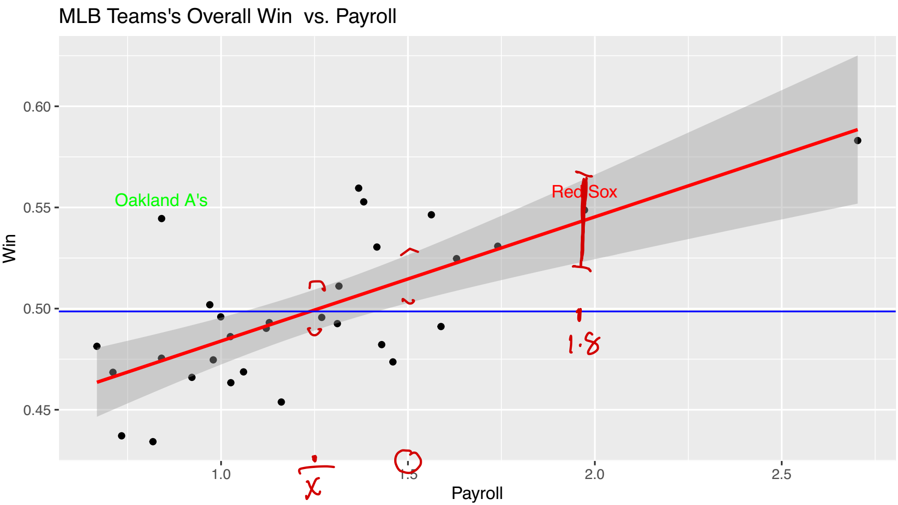
$.474 \pm 2 \cdot SE(\hat\beta_0+\hat\beta_1 \times .84)$

$= .474 \pm 2 \cdot X .027$

Confidence for the mean response

Band

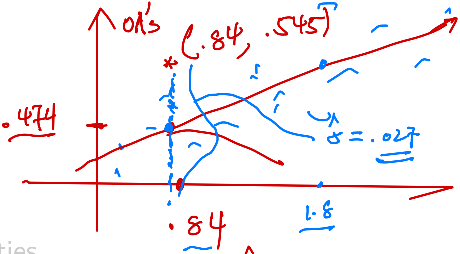
We can show the confidence interval for the mean response using `ggplot()` with `geom_smooth()` using the argument `se=TRUE`.



1. Objectives
2. Case Study: Baseball
   - Data
   - EDA
3. Simple linear regression
   - Model specification
   - OLS estimates and properties
   - R-squared and RSE
   - Confidence intervals for coefficients
   - **Prediction intervals**
   - Model diagnoses
4. Summary

Prediction interval:

Q: Is OA's unusually high in win



Recall: $\hat{y}_{x=.84} = .423 + .061 \times .84 = .474$

95%

$.474 \pm 2 \times .027$

prediction interval for one y given $x = .84$

Prediction interval for a response

A prediction interval can be used to quantify the uncertainty surrounding `win` for a **particular** team.

The prediction interval is approximately

$$\hat{y}_{|x} \pm 2\sqrt{MSE}$$

We now produce 95% & 99% PI for a future y given $x = .841$ using `predict()` again but with the argument `interval="prediction"`.

Notes: the exact equation for prediction interval is

$$\hat{y}_{|x} \pm t_{(\alpha/2, n-2)} \times \sqrt{MSE \times \left(1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right)}$$

Prediction interval for a response

```
new <- data.frame(payroll=c(.841))
CIpred <- predict(myfit0, new, interval="prediction", se.fit=TRUE)
CIpred
```

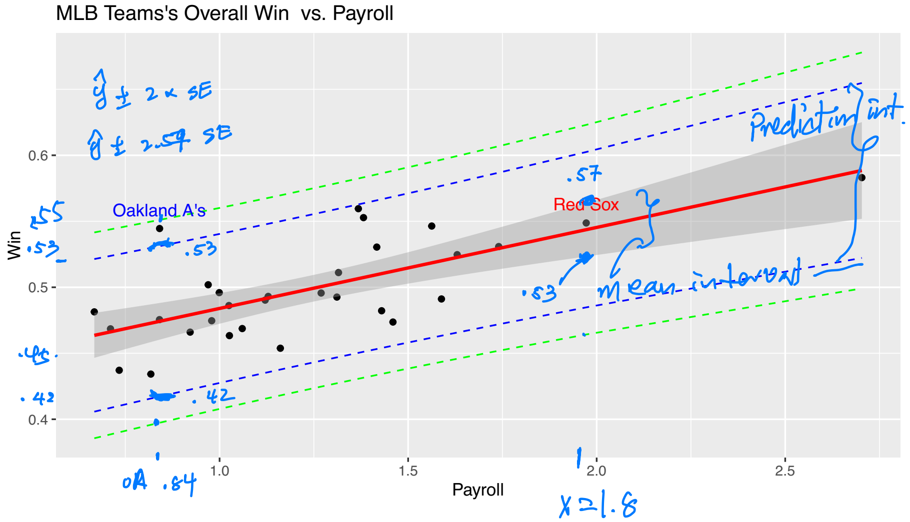
```
## $fit
##     fit   lwr   upr
## 1 0.474 0.417 0.531
##
## $se.fit
## [1] 0.00678
##
## $df
## [1] 28
##
## $residual.scale
## [1] 0.027
```

```
CIpred_99 <- predict(myfit0, new, interval="prediction", se.fit=TRUE, level=.99)
CIpred_99
```

```
## $fit
##     fit   lwr   upr
## 1 0.474 0.397 0.551
##
## $se.fit
## [1] 0.00678
##
## $df
## [1] 28
##
## $residual.scale
## [1] 0.027
```

Prediction interval for a response

Now we plot the confidence interval (shaded) along with the 95% prediction interval in blue and 99% prediction interval in green.



Model diagnoses

How reliable our confidence intervals and the tests are? We will need to check the model assumptions in the following steps:

1. Check **linearity** first; if linearity is satisfied, then
2. Check **homoscedasticity**; if homoscedasticity is satisfied, then
3. Check **normality**.

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \varepsilon_i$$

Model assumption: $\varepsilon_i \overset{iid}{\sim} N(0, \underline{\sigma^2})$

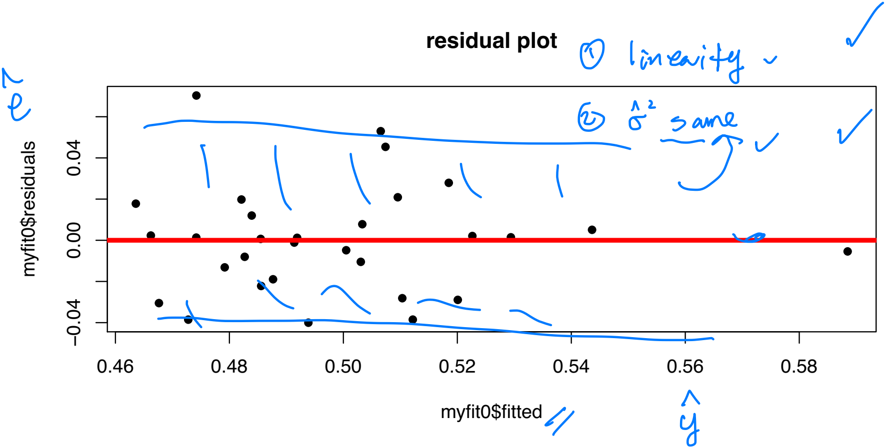
Residual plot

We plot the residuals against the fitted values to

- check **linearity** by checking whether the residuals follow a symmetric pattern with respect to $h = 0$.
- check **homoscedasticity** by checking whether the residuals are evenly distributed within a band.

Residual plot

```
plot(myfit0$fitted, myfit0$residuals,
     pch  = 16,
     main = "residual plot")
abline(h=0, lwd=4, col="red")
```



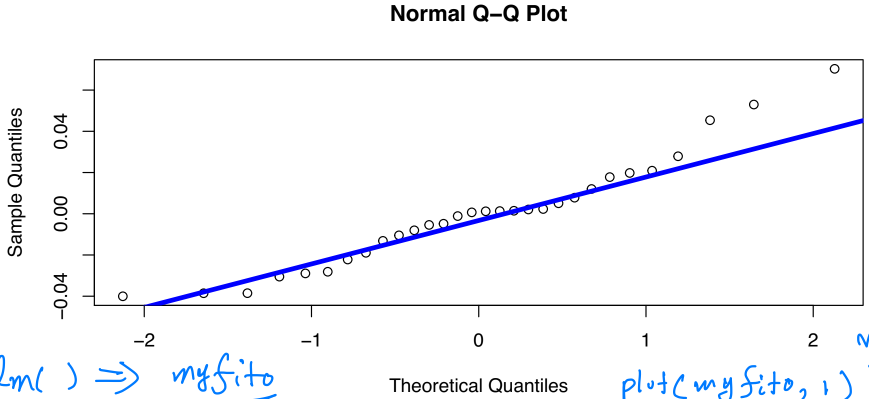
This can be also done with

```
plot(myfit0, 1)
```

Check normality

We look at the qqplot of residuals to check normality.

```
qqnorm(myfit0$residuals)
qqline(myfit0$residuals, lwd=4, col="blue")
```



lm( ) ⇒ myfit0

plot(myfit0, 1) resi.

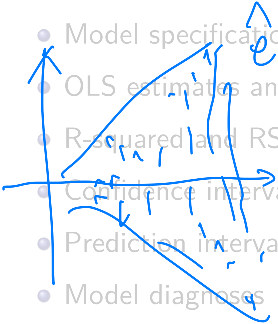
plot(myfit0, 2) normal

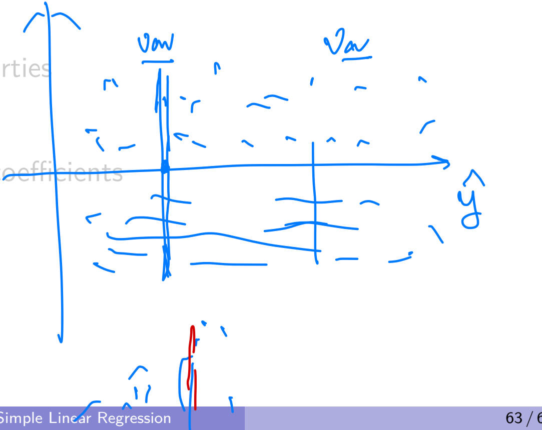
This can be also done with

```
plot(myfit0, 2)
```

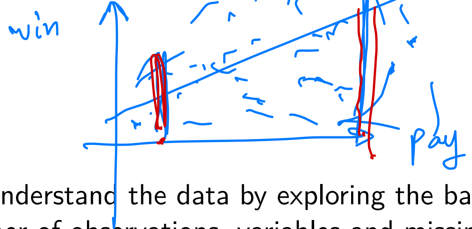
1. Objectives
2. Case Study: Baseball
   - Data
   - EDA
3. Simple linear regression
   - Model specification
   - OLS estimates and properties
   - R-squared and RSE
   - Confidence intervals for coefficients
   - Prediction intervals
   - Model diagnoses
4. **Summary**

Con: diagnoses
- linearity ok
- Normality ok
- Equal var ok





Summary



- EDA: We understand the data by exploring the basic structure of the data (number of observations, variables and missing data), the descriptive statistics and the relationship between variables. Visualization is a crucial step for EDA. Both graphical tools in base `R` an `ggplot2` come in handy.
- OLS: Simple linear regression is introduced. We study the OLS estimate with its interpretation and properties. We evaluate the OLS estimate and provide inference. It is important to perform model diagnoses before coming to any conclusion. The `lm()` function is one of the most important tools for statisticians.

Summary

- EDA: We understand the data by exploring the basic structure of the data (number of observations, variables and missing data), the descriptive statistics and the relationship between variables. Visualization is a crucial step for EDA. Both graphical tools in base `R` an `ggplot2` come in handy.
- OLS: Simple linear regression is introduced. We study the OLS estimate with its interpretation and properties. We evaluate the OLS estimate and provide inference. It is important to perform model diagnoses before coming to any conclusion. The `lm()` function is one of the most important tools for statisticians.

Summary

- EDA: We understand the data by exploring the basic structure of the data (number of observations, variables and missing data), the descriptive statistics and the relationship between variables. Visualization is a crucial step for EDA. Both graphical tools in base `R` an `ggplot2` come in handy.
- OLS: Simple linear regression is introduced. We study the OLS estimate with its interpretation and properties. We evaluate the OLS estimate and provide inference. It is important to perform model diagnoses before coming to any conclusion. The `lm()` function is one of the most important tools for statisticians.