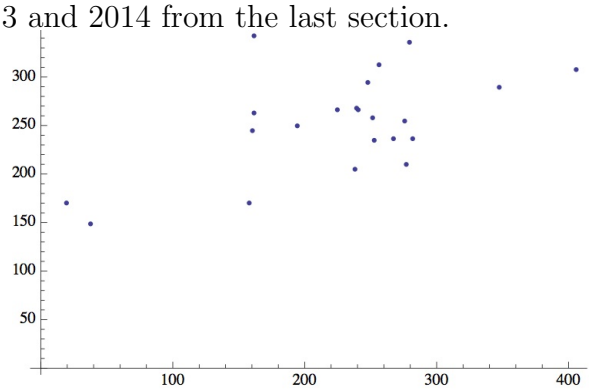
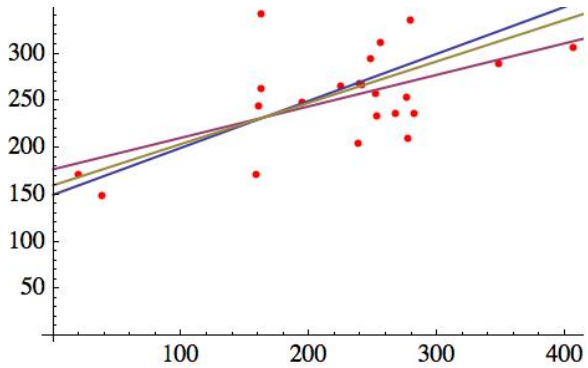
Learning Goals

1. What is the least squares line?
2. Finding the least squares line in R
3. Interpreting results of a linear regression in R

Topic 3: The Least Squares Line, Linear Regression

Linear Regression Recall the scatterplot of the data for fantasy football points for a set of quarterbacks in 2013 and 2014 from the last section.





Fitting a line to the data : The Least squares line. Given a set of data points in the xy-plane, $\{(x_1, y_1), (x_2, y_2), (x_3, y_3), \ \ldots \ , (x_n, y_n)\}$ such as those shown above, there are many lines that we could fit to the data to help us make predictions for the future. To find an equation for the line which best fits the data we use the method of least squares. This line minimizes the squares of the difference between the y values on the line and the y values for the points in the data. This equation gives us a linear formula which estimates the relationship between the variable x and the variable y which we can use for predictions.

Recall that the equation of a line is of the form $y = \beta_0 + \beta_1 x$ where β_0 and β_1 are constants. The idea is to find values of β_0 and β_1 so that the sum

$$\text{SSE} = ((y_1 - y(x_1))^2 + (y_2 - y(x_2))^2 + \cdots + (y_n - y(x_n))^2)$$

is minimal where $y(x_i)$ is the value corresponding to x_i from the formula $y = \beta_0 + \beta_1 x$ and y_i is the value corresponding to x_i in the datapoint (x_i, y_i). (SSE stands for the sum of the squared errors.)

We see from this interactive demonstration on Wolfram Alpha Demonstrations that for some lines the sum of the squared errors is larger than for others.

Quarterback Example Consider the data for our quarterbacks above. Let x_i be the number of fantasy points scored by quarterback i in 2013 and let y_i denote the number of fantasy points scored by quarterback i is 2014. Lets calculate the sum of the squared error (SSE) for a particular line $y = 150 + (0.5)x$.

points 2013	points 2014	predicted values	Errors	Squared Errors
x_i	y_i	$y(x_i) = 150 + (0.5)x_i$	$y_i - y(x_i)$	$(y_i - y(x_i))^2$
162	342	231	111	12321
279	336	289.5	46.5	2162.25
256	312	278	34	1156
406	307	353	−46	2116
248	295	274	21	441
348	290	324	−34	1156
239	268	269.5	−1.5	2.25
241	267	270.5	−3.5	12.25
225	266	262.5	3.5	12.25
162	263	231	32	1024
252	258	276	−18	324
276	254	288	−34	1156
194	249	247	2	4
160	244	230	14	196
267	237	283.5	−46.5	2162.25
282	237	291	−54	2916
253	234	276.5	−42.5	1806.25
277	210	288.5	−78.5	6162.25
238	205	269	−64	4096
19	171	159.5	11.5	132.25
158	171	229	−58	3364
38	149	169	−20	400
			$SE = -225$	$SSE = 43122.25$

From the table above , we see that for the line $y = 150 + (0.5)x$, the sum of the errors is an unreliable statistic in measuring how well the line fits the data due to cancellation. We avoid this problem by squaring the error and use the Sum of Squares of The Error (SSE) to measure how well the line fits the data. Naturally a smaller SSE will indicate that a line is a better fit for the data.

There is a unique line for which SSE is at a minimum. This line is called the Least Squares Line. The methodology used to obtain the equation of this line is called the method of least squares.

We can solve for the coefficients β_0 and β_1 of such a line, for a particular set of data, by using calculus to find the minimum of the function

$$SSE = \Sigma[y_i - (\beta_0 + \beta_1 x_i)]^2,$$

for the variable β_0 and β_1.

Definition The least squares line $y = \beta_0 + \beta_1 x$, for a set of data, is the unique line with the following properties:

1. The sum of errors equals 0; SE = 0

2. The sum of squared errors (SSE) is smaller than that for any other straight line model.

The values of β_0 and β_1 for the least squares line are given by the following formulas (where $\bar{x}$ and $\bar{y}$ demote the means of the data sets $\{x_1, x_2, \ldots, x_n\}$ and $\{y_1, y_2, \ldots, y_n\}$ respectively):

$$\beta_1 = \frac{SS_{xy}}{SS_{xx}}, \qquad \beta_0 = \bar{y} - \beta_1 \bar{x}$$

where

$$SS_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}$$

and

$$SS_{xx} = \sum (x_i - \bar{x})^2 = \sum x_i^2 - \frac{(\sum x_i)^2}{n}$$

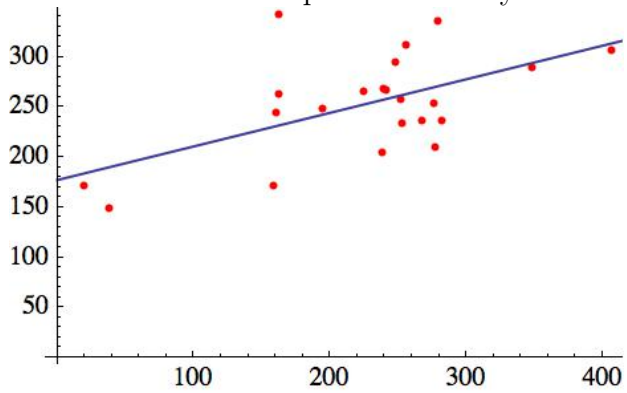
We have

$$SSE = \sum (y_i - (\beta_0 + \beta_1 x_i))^2 = SS_{yy} - \beta_1 SS_{xy}$$

where

$$SS_{yy} = \sum y_i^2 - \frac{\sum (y_i)^2}{n}.$$

Example For the above data, the least squares line is given by $y = 177.06 + 0.33x$ and the sum of the squared errors for this line is SSE = 36487.3. This is less than the sum of the squared errors for the line shown above, in fact it is the minimum such sum possible for any line that we might fit to the data.



We could use this line to estimate the fantasy points that a quarterback will score in 2020 given the number of points he scored in 2019. For example if a quarterback scored 250 points in fantasy football in 2019, we might expect the number of points he scores in 2020 to be roughly $177.06 + 0.33(250) \approx 259.56$.

Calculating the Least Squares Line in R We can use R to calculate the coefficients of the least squares line with a single command `lm()` as shown below:

```
Y2013<-c(162,279,256,406,248,348,239,241,225,162,252,
         276,194,160,267,282,253,277,238,19,158,38)
Y2014<-c(342,336,312,307,295,290,268,267,266,263,258,
         254,249,244,237,237,234,210,205,171,171,149)
fit<-lm(Y2014~Y2013)
fit

##
## Call:
## lm(formula = Y2014 ~ Y2013)
##
## Coefficients:
## (Intercept)        Y2013
##    177.0583       0.3353
```

How well does the line fit the data? We can always fit a least squares line to a set of data using the `lm()` (which stands for `Linear Model`) command. However, as we saw when exploring the correlation, a linear model for a set of data is not always appropriate. Therefore, we need some measure of how well the line we get, by applying the `lm()` function to the data, fits the data. We can find some measure of how well the model fits the data by looking at the summary (of the statistics related to the model) of the model we created and called `fit`:

```
summary(fit)
```

```
##
## Call:
## lm(formula = Y2014 ~ Y2013)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -59.932 -29.156  -3.644  13.451 110.626
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 177.0583    25.8494   6.850 1.17e-06 ***
## Y2013         0.3353     0.1069   3.137  0.00519 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 42.71 on 20 degrees of freedom
## Multiple R-squared:  0.3298, Adjusted R-squared:  0.2963
## F-statistic: 9.842 on 1 and 20 DF,  p-value: 0.005188
```

p-values Whenever we run a regression, we should always check the p-values next to the independent variables (in this case the intercept and Y2013). Each independent variable has a p-value between 0 and 1. Roughly it gives the probability that that variable (in the presence of all other variables in the model) does not enhance our predictive ability. An independent variable with a p-value less than 0.05 is considered a useful predictor of the dependent variable (in this case Y2014). We see that R makes it easy to identify significant variables with p-values with the use of asterisks. In our model we have that the variable Y2013 and the intercept are both significant in our model.

The Coefficient of Determination R^2 The multiple R^2 statistic also gives us information about the accuracy of our model for prediction. It is called the coefficient of determination and is the square of the correlation coefficient $r = 0.5742875$ that we found between the variables $Y2013, Y2014$ in the previous section. It can also be calculated as

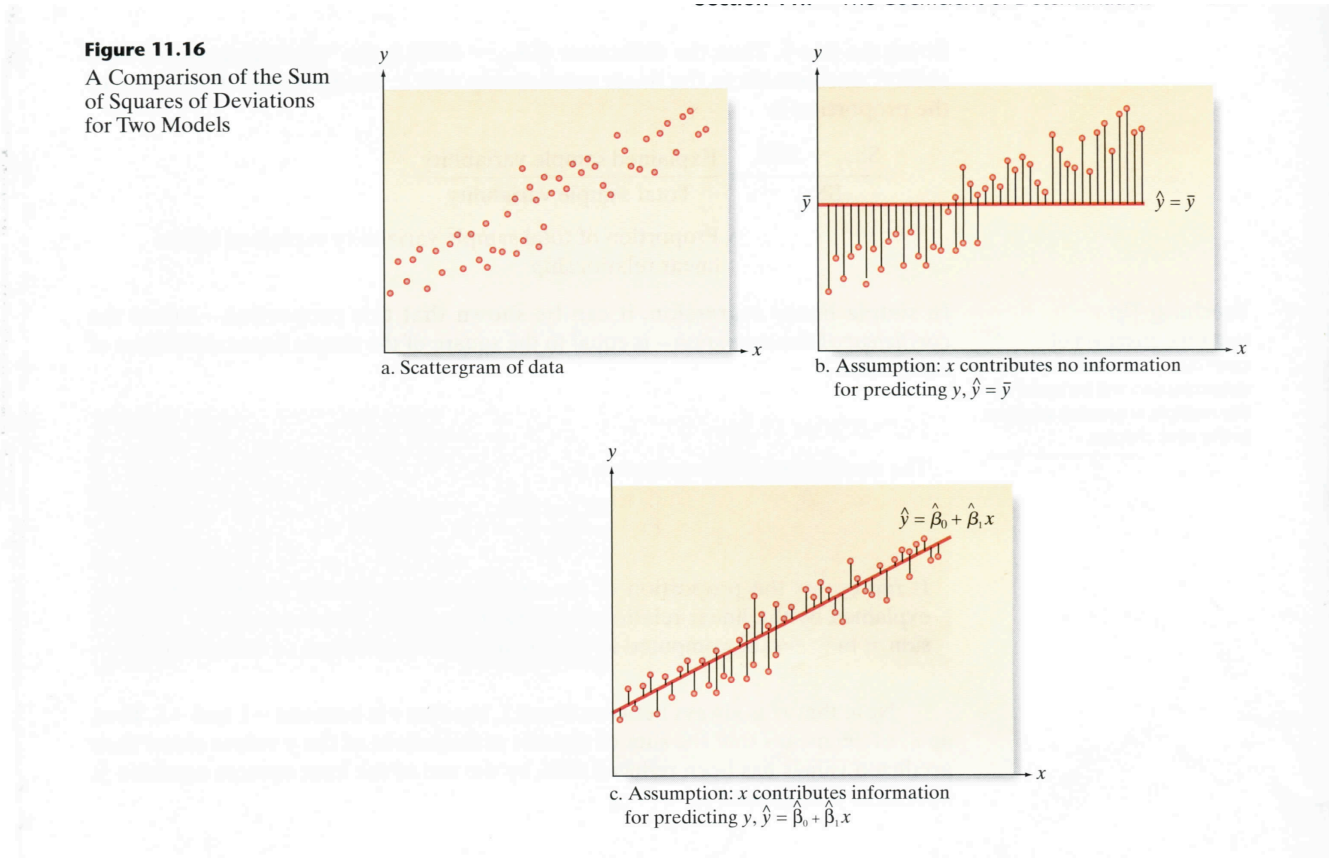
$$R^2 = \frac{SS_{yy} - SSE}{SS_{yy}}$$

and gives a measure of the total sampling variability that is explained by the linear relationship between x and y. The value of $R^2 = 0.3298$ in the example above means that the sum of squares of the deviations for this example are reduced by about 33% when we use the least squares line for prediction instead of the line $y = \bar{y}$, since

$$SS_{yy} = \sum (y_i - \bar{y})^2 \quad \text{and} \quad SSE = \sum (y_i - \hat{y})^2.$$

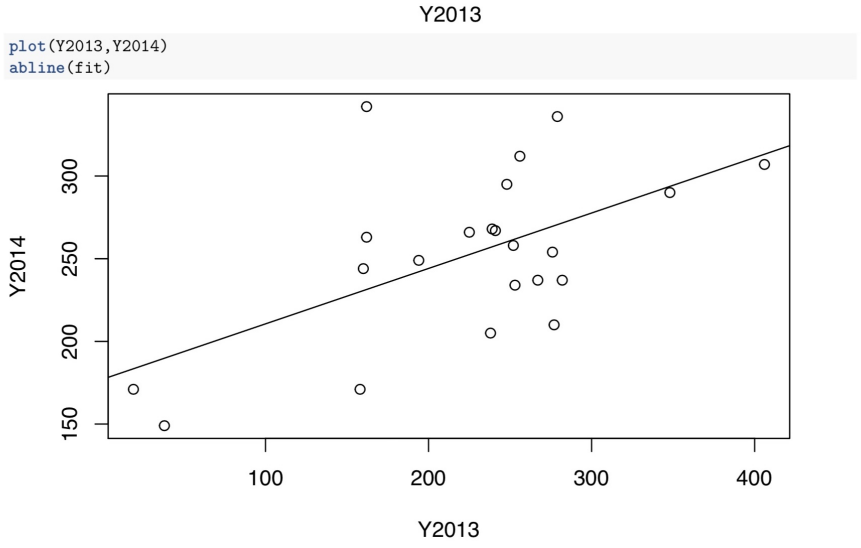
In other words, 33% of the error is explained away by using the least squares line for prediction. There are various underlying statistical conditions on the distribution of errors etc... that one needs to check

in order to use the linear model. You will learn about these conditions in more advanced courses. We see that our software R actually adjusts this statistic to take into account some of the abnormalities of the data. After this adjustment, we see that our data most like explains a little less (29.6%) of the error.



Plotting The Line With our scatterplot We add the line into the scatterplot with the command `abline(fit)`.

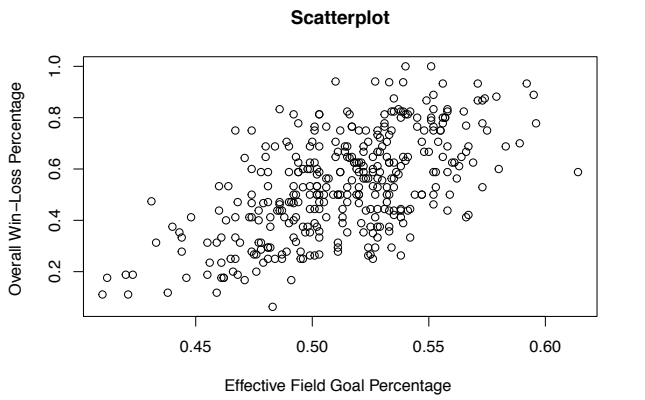
```
plot(Y2013,Y2014)
abline(fit)
```



Example Lets return to our imported data `BBdata` and find least squares lines that best fit the data for some team statistics and the Overall Win-Loss Percentage OWL.P We had the following scatterplots of

the paired data `OWL.P` vs. `eFG.P` and `OWL.P` vs. `Th.Par` resp.

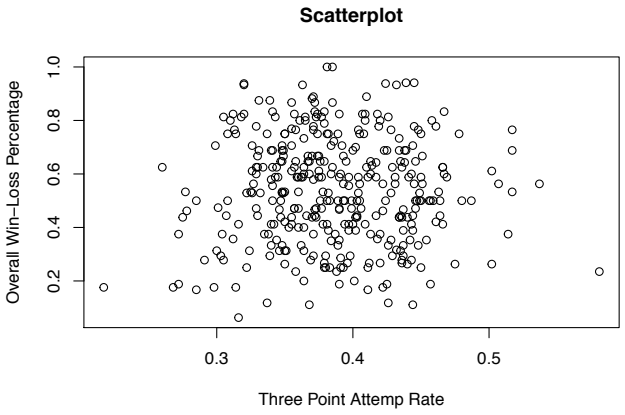
```
plot(BBdata$eFG.P, BBdata$OWL.P,  main="Scatterplot",
    xlab="Effective Field Goal Percentage ", ylab="Overall Win-Loss Percentage ")
```




```
cor(BBdata$eFG.P, BBdata$OWL.P)
```

```
## [1] 0.6002358
```

```
plot(BBdata$Th.Par, BBdata$OWL.P,  main="Scatterplot",
    xlab="Three Point Attemp Rate", ylab="Overall Win-Loss Percentage")
```




```
cor(BBdata$Th.Par, BBdata$OWL.P)
```

```
## [1] 0.004512129
```

(a) Find the Least squares line for Win-Loss Percentage `OWL.P` (dependent variable) and Effective Field Goal Percentage `eFG.P` (independent variable) for the teams in our file `BBdata` using the basic command `lm(y x)`.

(b) Print a summary of the model in R, find the p-values of the coefficients and the value of the coefficient of determination.

(c) Assess the validity of the model by taking into account the p-values and the amount of the error explained by the model.

(d) Find the Least squares line for Win-Loss Percentage `OWL.P` (dependent variable) and 3-Point Attempt Rate `Th.Par` f(independent variable) for the teams in our file `BBdata` using the basic command `lm(y x)`.

(e) Print a summary of the model in R, find the p-values of the coefficients and the value of the coefficient of determination.

(f) Assess the validity of the model by taking into account the p-values and the amount of the error explained by the model.

The results for parts (a)-(c) are as follows:

```
BBdata <- read.csv(file="BBStats.csv", header=TRUE, sep=",")
fit1<-lm(BBdata$OWL.P~BBdata$eFG.P)
fit1
```

```
##
## Call:
## lm(formula = BBdata$OWL.P ~ BBdata$eFG.P)
##
## Coefficients:
##  (Intercept)  BBdata$eFG.P
##       -1.231         3.445
```

```
summary(fit1)
```

```
##
## Call:
## lm(formula = BBdata$OWL.P ~ BBdata$eFG.P)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.3699 -0.1278  0.0002  0.1156  0.4151
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -1.2309     0.1261  -9.763   <2e-16 ***
## BBdata$eFG.P   3.4448     0.2450  14.060   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1585 on 351 degrees of freedom
## Multiple R-squared:  0.3603, Adjusted R-squared:  0.3585
## F-statistic: 197.7 on 1 and 351 DF,  p-value: < 2.2e-16
```

(a) We see that the best fit line for the data is

$$\texttt{OWL.P} = 3.445\texttt{eFG.P} - 1.231$$

(b) We see that both coefficients are significant and should be included in the model. The adjusted coefficient of determination is $R^2 = 0.3585$ which means that roughly 36% of the error is explained by the model.

(c) Clearly other factors influence the overall win-loss percentage but about 36% of the variability in the win-loss percentage is explained by this model. Roughly an increase of 1 unit in effective field goal percentage for a team leads to an increase of 3.4 units in overall win-loss percentage. For example an increase of 0.1 if eFG.P roughly leads to an increase of .34 in OWL.P.

The results for parts (d)-(e) are shown below:

```
fit2<-lm(BBdata$OWL.P~ BBdata$Th.Par)
fit2
```

```
##
## Call:
## lm(formula = BBdata$OWL.P ~ BBdata$Th.Par)
##
## Coefficients:
##   (Intercept)  BBdata$Th.Par
##       0.53114        0.01707
```

```
summary(fit2)
```

```
##
## Call:
## lm(formula = BBdata$OWL.P ~ BBdata$Th.Par)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.47353 -0.13735 -0.00604  0.14945  0.46236
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.53114    0.07873   6.746 6.27e-11 ***
## BBdata$Th.Par  0.01707    0.20187   0.085    0.933
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1981 on 351 degrees of freedom
## Multiple R-squared:  2.036e-05,  Adjusted R-squared:  -0.002829
## F-statistic: 0.007146 on 1 and 351 DF,  p-value: 0.9327
```

(d) The best fit line for the data is

$$\texttt{OWL.P} = 3.445\texttt{Th.Par} - 1.231$$

(e-f) We see that the 3-Point Attempt Rate `Th.Par` is not a significant factor in explaining overall win-loss percentage. The p-value is 0.933, which means that there is roughly a 93% chance that this statistic does not enhance our predictive ability. As we might have expected from our calculation of the correlation coefficient in the last section, the coefficient of determination is very small here, it is approximately 0.00002, which says that if we just used the average overall win-loss percentage to predict the overall win-loss percentage for all teams, our predictions would be essentially as good as the predictions we would get by using this least squares line for prediction, since it explains only 0.002% of the error we get by doing so. Therefor our conclusion is that this is not a good model for predicting overall win-loss percentage and the variable `Th.Par` should not be used for such predictions.

R commands

1. `fit<-lm(Y X)` : creates best fit linear model(line) for data $\{(x_1, y_1), (x_2, y_2), (x_3, y_3), \quad \dots \quad , (x_n, y_n)\}$, where $X < -c(x_1, \dots, x_n)$ and $Y < -c(y_1, \dots, y_n)$. To print details type `fit`, the name of your model.
2. `summary(fit)` : shows the p-values and R^2 stats associated to your model called `fit`.
3. `plot(X,Y)` followed by `abline(fit)`: shows the scatterplot and best fit line for the data.

References

[1] Kubatko, J., Oliver, D., Pelton K., and Rosenbaum, D. *A Starting Point For Analyzing Basketball Statistics*. Journal of Quantitative Analysis in Sport, **Vol 3, Issue 3, 2007, Article 1.**

[2] Shea, Stephen M., and Baker, Christopher E. *Basketball Analytics*. Advanced Metrics, LLC, Lake St. Louis, MO, 2013.

[3] Winston, Wayne L. *Mathletics* Princeton University Press.