Assignment Two - Emily Gill's Part

2024-10-11

```
# creating variables of the Hitters data base from ISLR
HittersAE <- Hitters%>%subset(League == "A")%>%subset(Division == "E")
HittersAW <- Hitters%>%subset(League == "A")%>%subset(Division == "W")
HittersNE <- Hitters%>%subset(League == "N")%>%subset(Division == "E")
HittersNW <- Hitters%>%subset(League == "N")%>%subset(Division == "W")
```

```
# cleaning the data
HittersAEclean<-HittersAE%>% subset(!is.na(Salary))
HittersAWclean<-HittersAW%>% subset(!is.na(Salary))
HittersNEclean<-HittersNE%>% subset(!is.na(Salary))
HittersNWclean<-HittersNW%>% subset(!is.na(Salary))
```

LARS predicting AE

```
HittersAEclean.lars<-lars(as.matrix(HittersAEclean[,-c(20,19,14,15)]),HittersAEclean$Salary)
I_AE<-HittersAEclean.lars$Cp==min(HittersAEclean.lars$Cp)
betalarsAE<-HittersAEclean.lars$beta[I_AE,]
betalarsAE
```

```
##       AtBat        Hits       HmRun        Runs         RBI       Walks       Years
## -3.3530002 10.0477043 -5.1870815  1.9252299  0.0000000  6.1607285  0.0000000
##      CAtBat       CHits      CHmRun       CRuns        CRBI      CWalks     PutOuts
## -0.3771068  0.0000000  0.8418551  2.9480375  1.0145354 -1.1760056  0.3408382
##     Assists      Errors
##  0.1259118  0.0000000
```

```
HittersAWclean.lars<-lars(as.matrix(HittersAWclean[,-c(20,19,14,15)]),HittersAWclean$Salary)
I_AW<-HittersAWclean.lars$Cp==min(HittersAWclean.lars$Cp)
betalarsAW<-HittersAWclean.lars$beta[I_AW,]
betalarsAW
```

```
##       AtBat        Hits       HmRun        Runs         RBI       Walks       Years
## 0.00000000 2.30243447 0.00000000 0.00000000 0.00000000 1.03768371 0.00000000
##      CAtBat       CHits      CHmRun       CRuns        CRBI      CWalks     PutOuts
## 0.00000000 0.33434397 0.00000000 0.00000000 0.00000000 0.00000000 0.08998272
##     Assists      Errors
## 0.00000000 0.00000000
```

```
HittersNEclean.lars<-lars(as.matrix(HittersNEclean[,-c(20,19,14,15)]),HittersNEclean$Salary)
I_NE<-HittersNEclean.lars$Cp==min(HittersNEclean.lars$Cp)
betalarsNE<-HittersNEclean.lars$beta[I_NE,]
betalarsNE
```

```
##       AtBat        Hits       HmRun        Runs         RBI       Walks       Years
## -0.2481822  0.0000000  0.0000000  0.0000000  0.0000000  2.5323541  0.0000000
##      CAtBat       CHits      CHmRun       CRuns        CRBI      CWalks     PutOuts
```

```
##  0.0000000  0.0000000  0.0000000  0.0000000  0.8941327  0.0000000  0.4312271
##    Assists     Errors
##  0.7716719 -8.4960887
```

```
HittersNWclean.lars<-lars(as.matrix(HittersNWclean[,-c(20,19,14,15)]),HittersNWclean$Salary)
I_NW<-HittersNWclean.lars$Cp==min(HittersNWclean.lars$Cp)
betalarsNW<-HittersNWclean.lars$beta[I_NW,]
betalarsNW
```

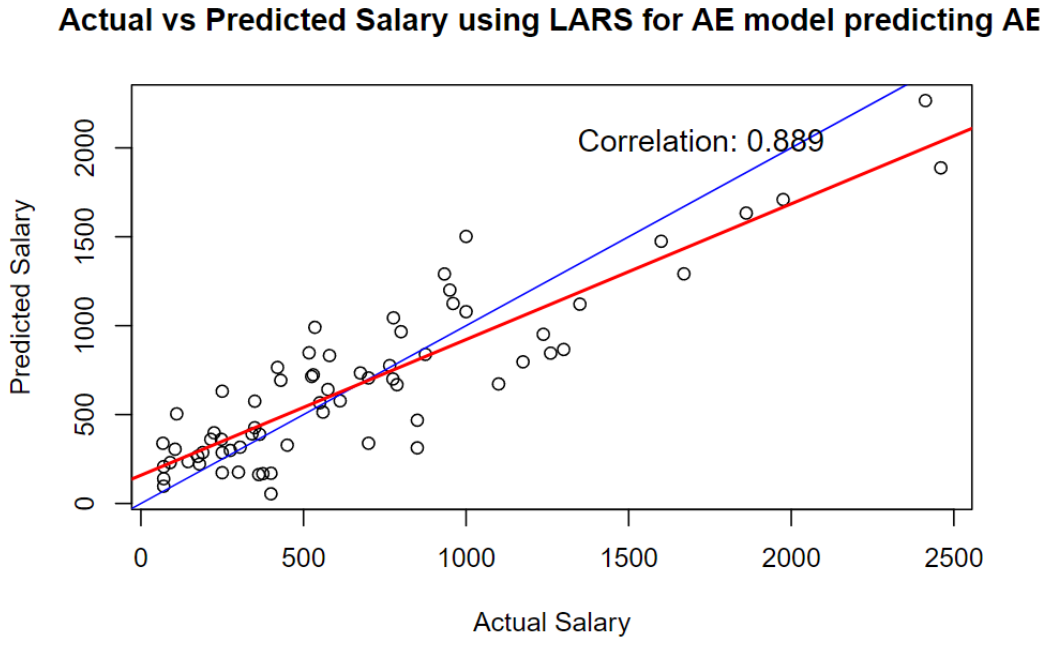
```
##       AtBat        Hits       HmRun        Runs         RBI       Walks
##   1.3713374   1.9961871  32.2358918  -9.7338778 -11.9613345   7.3627544
##       Years      CAtBat       CHits      CHmRun       CRuns        CRBI
##   7.7832856   0.0000000   0.4414019   0.4438961   0.0000000   0.0000000
##      CWalks     PutOuts     Assists      Errors
##  -0.9914020   0.0000000  -0.7387407   3.8378122
```

```
predict_and_plot <-
  function(train_data, model_betas, test_data, train_division, test_division) {
    # Predict salaries for the test data
    pred <- as.matrix(test_data[,-c(20,19,14,15)]) %*% model_betas
    # Adjust the predicted values using the mean adjustment
    pred <- pred + (mean(train_data$Salary) - mean(pred))

    # Plot actual vs predicted salaries with Actual Salary on the x-axis
    # abline (0,1) adds a reference line with an intercept of 0 and a slope of 1 to the plot.
    # It represents a perfect prediction where predicted salary = actual salary.
    # lm_fit variable will add a linear regression line to the data
    plot(test_data$Salary, pred,
         main = paste("Actual vs Predicted Salary using LARS for",
                      train_division,"model predicting", test_division),
         xlab = "Actual Salary", ylab = "Predicted Salary")
    abline(0, 1, col = "blue")
    lm_fit <- lm(pred ~ test_data$Salary)
    abline(lm_fit, col = "red", lwd = 2)

    # Calculate the correlation
    correlation <- cor(pred, test_data$Salary)

    # Print the correlation on the plot
    text(x = max(test_data$Salary) * 0.7, y = max(pred) * 0.9,
         labels = paste("Correlation:", round(correlation, 3)),
         col = "black", cex = 1.2)

    return(correlation)
  }
```

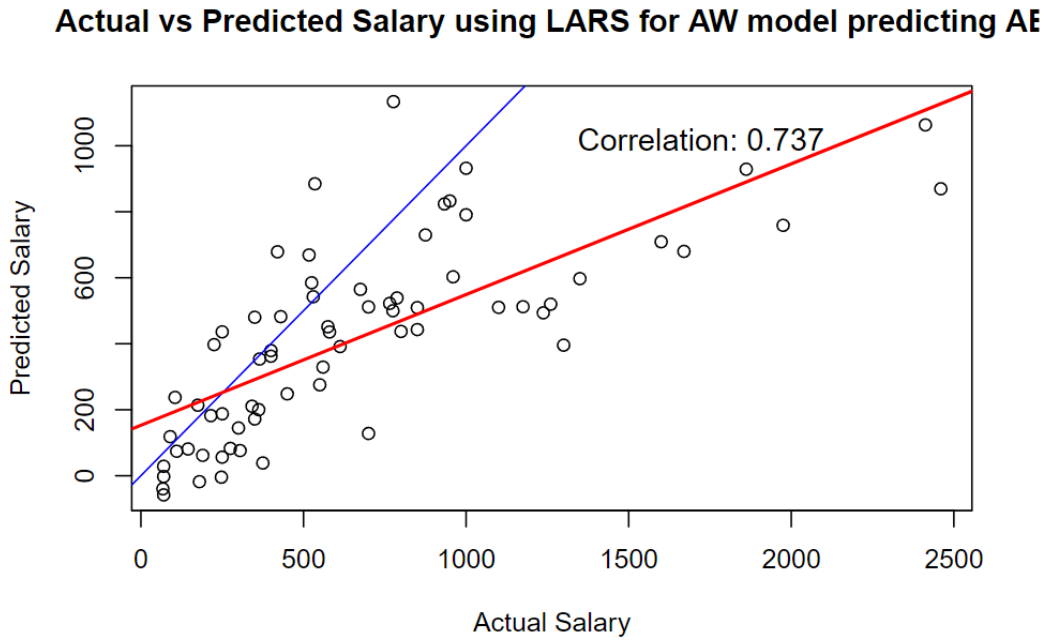
Red = Linear Regression Line

Blue = Perfect Representation Line

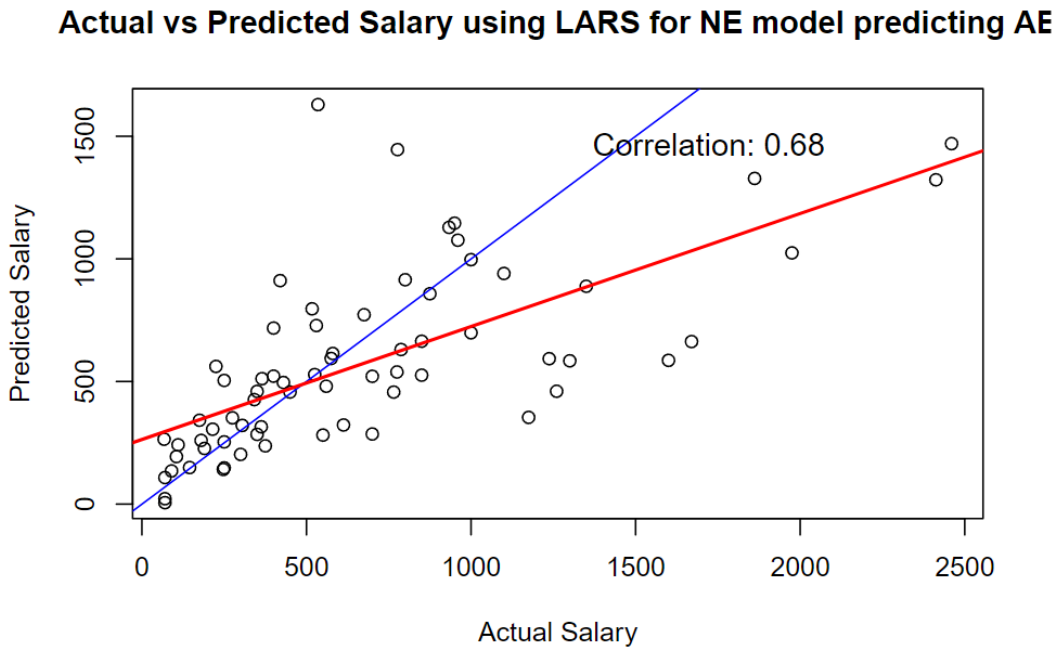
```
# Predict for AE using AE model
correlation_AE_AE <-
  predict_and_plot(HittersAEclean, betalarsAE, HittersAEclean, "AE", "AE")
```



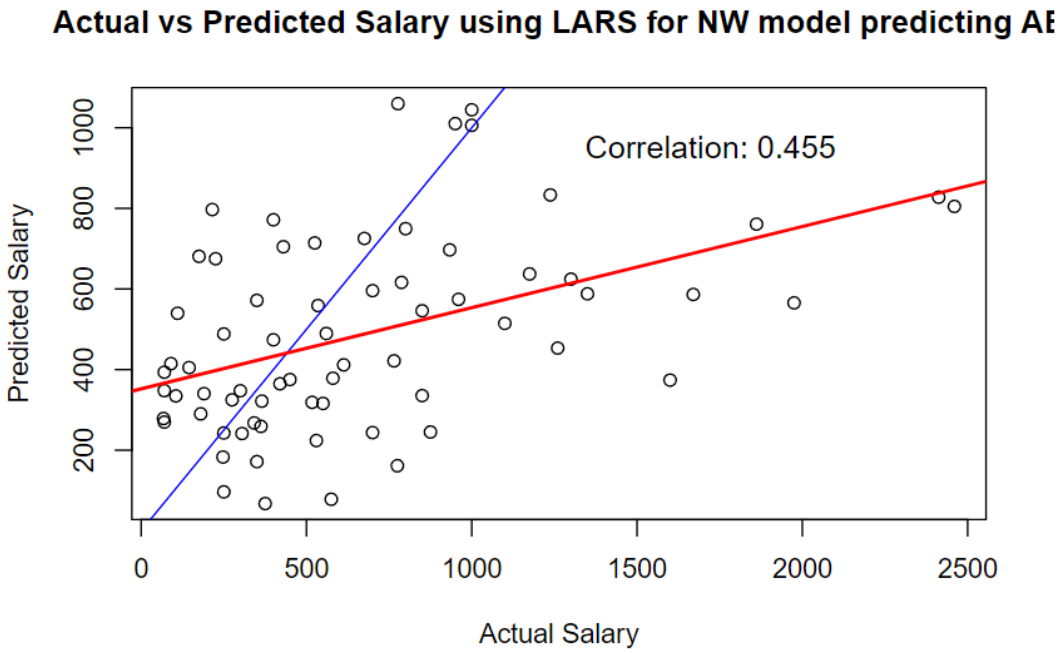

```
# Predict for AE using AW
correlation_AE_AW <-
  predict_and_plot(HittersAWclean, betalarsAW, HittersAEclean, "AW", "AE")
```



```r
# Predict for AE using NE
correlation_AE_NE <-
  predict_and_plot(HittersNEclean, betalarsNE, HittersAEclean, "NE", "AE")
```




```r
# Predict for AE using NW
correlation_AE_NW <-
  predict_and_plot(HittersNWclean, betalarsNW, HittersAEclean, "NW", "AE")
```



LM Predicting AE

```
# Fitting linear models with the selected variables (based on LARS)
lm_AE <- lm(Salary ~ .,
            data = HittersAEclean[, c(names(betalarsAE)[betalarsAE != 0], "Salary")])
lm_AW <- lm(Salary ~ .,
            data = HittersAWclean[, c(names(betalarsAW)[betalarsAW != 0], "Salary")])
lm_NE <- lm(Salary ~ .,
            data = HittersNEclean[, c(names(betalarsNE)[betalarsNE != 0], "Salary")])
lm_NW <- lm(Salary ~ .,
            data = HittersNWclean[, c(names(betalarsNW)[betalarsNW != 0], "Salary")])
```

```
# Predict salaries in the AE Division using each model
pred_AE_AE <-
  predict(lm_AE, HittersAEclean) # AE model on AE data
pred_AE_AW <-
  predict(lm_AW, HittersAEclean) # AW model on AE data
pred_AE_NE <-
  predict(lm_NE, HittersAEclean) # NE model on AE data
pred_AE_NW <-
  predict(lm_NW, HittersAEclean) # NW model on AE data
```

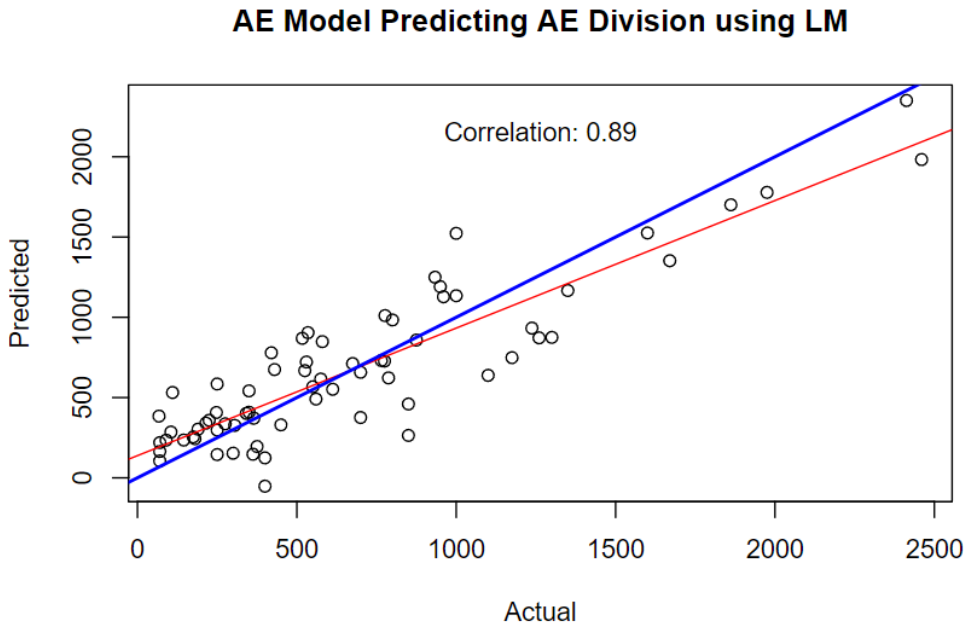
```
# Function to plot scatter plots and calculate correlations
plot_correlation <- function(actual, predicted, title) {
  # Calculate the correlation
  correlation <- cor(actual, predicted)

  # Create the scatter plot
```

```r
  plot(actual, predicted,
       main = title,
       xlab = "Actual",
       ylab = "Predicted",
       pch = 1,
       col = "black")

  # The slope and position of the regression line reflects the relationship
  # between the actual and predicted values.
  # If the regression line has a slope close to 1 and passes through the origin (0, 0),
  # it indicates a good fit and suggests that the predictions are accurate.
  abline(lm(predicted ~ actual), col = "red")
  # Adds a line that represents when actual = predicted
  abline(0, 1, col = "blue", lwd = 2)


  # Add a subtitle with the correlation value
  mtext(paste("Correlation:", round(correlation, 2)), side = 3, line = -2)
}
```

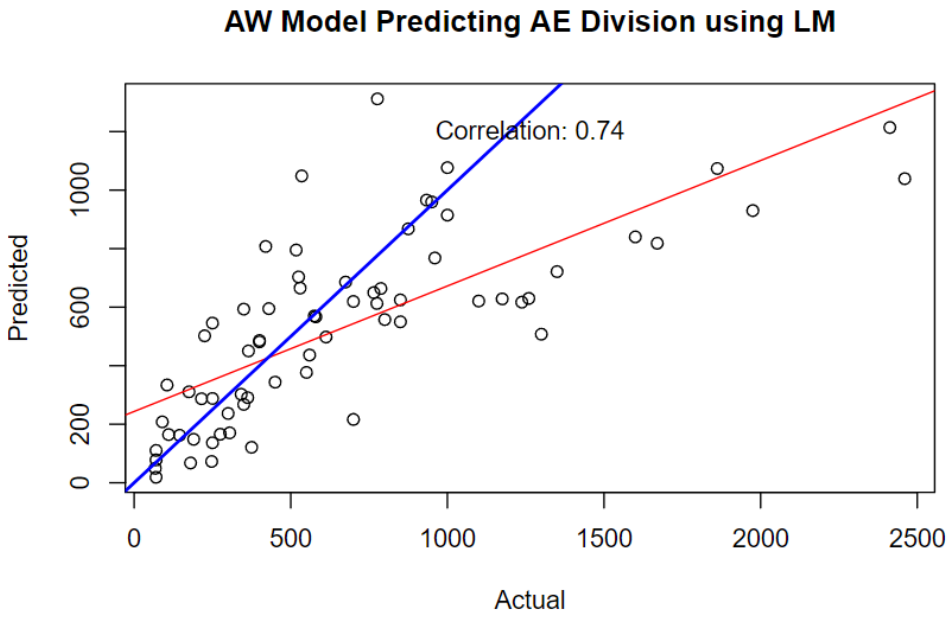
Red = Linear Regression Line

Blue = Perfect Representation Line

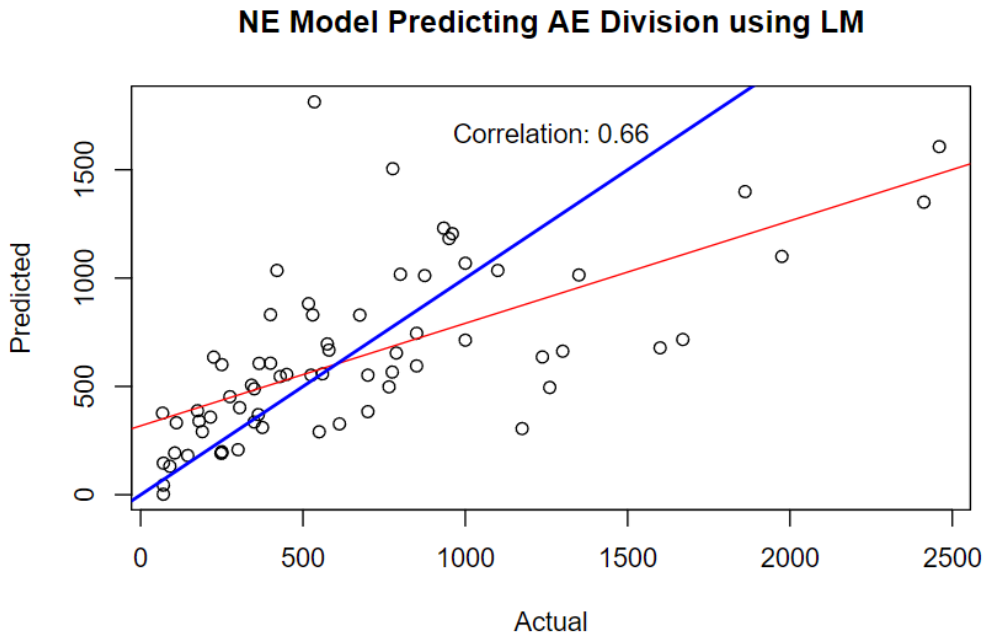
```r
# Scatterplots and correlations for AE Division
# AE predicting AE
plot_AE_AE <-
  plot_correlation(HittersAEclean$Salary, pred_AE_AE, "AE Model Predicting AE Division using LM")
```



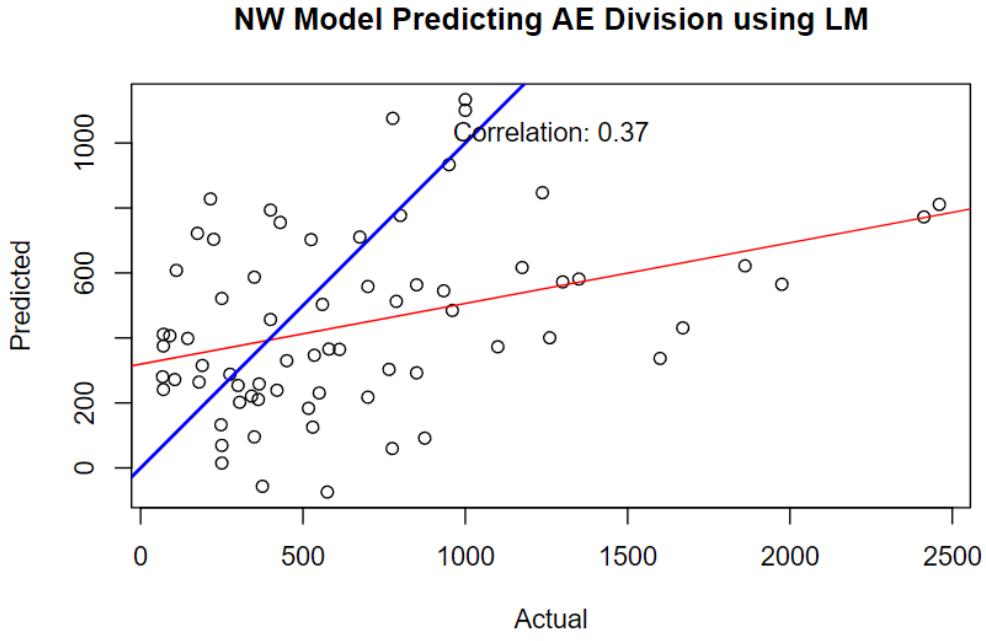
```r
# AW predicting AE
plot_AE_AW <-
  plot_correlation(HittersAEclean$Salary, pred_AE_AW, "AW Model Predicting AE Division using LM")
```




```r
# NE predicting AE
plot_AE_NE <-
  plot_correlation(HittersAEclean$Salary, pred_AE_NE, "NE Model Predicting AE Division using LM")
```




```
# NW predicting AE
plot_AE_NW <-
  plot_correlation(HittersAEclean$Salary, pred_AE_NW, "NW Model Predicting AE Division using LM")
```



Emily Gill - Analysis

AE Predicting AE:

When using LARS to create an AE model to predict AE salaries, the correlation is 0.889, which is very close to a perfect positive correlation of 1. This is apparent when viewing the linear regression line in red and the perfect prediction line in blue. The blue line represents the distribution of data if the predictions matched the actual salary exactly. The linear regression line is very close and similar to the blue line, indicating that LARS provides a very good correlation. Using LM to create an AE model to predict AE salaries has a correlation of 0.89, almost identical to the LARS model. The linear regression line and the perfect prediction line are also very close, indicating that LM also does a good job of predicting. The data points on both plots fall very close to the lines. Predicting “within” divisions is the most accurate and provides the highest correlation.

AW Predicting AE:

When using LARS to create an AW model to predict AE salaries, the correlation is 0.737, fairly close to a perfect 1. Though 0.737 may seem like a fairly good correlation, When viewing the data on the plot, a lot of it does not fall close to the blue perfect prediction line. A lot of it stretches too far to the right, indicating that actual salaries are really much higher than predicted. The LM plot showcases a very similar distribution with a correlation of 0.74, indicating that using the AW division to predict AE is not the best choice for accurate predictions.

NE Predicting AE:

When using LARS to create an NE model to predict AE salaries, the correlation is 0.68. This is not as close to a perfect 1 but does suggest some positive correlation. When looking at the data, actual salary extends further to the right, showcasing that actual salaries are much higher than we predicted. There are also a few outliers where we predicted the salary to be higher than they actually are (ex: predicted value = over 1500, actual value = close to 500). The LM model has a correlation of 0.66, which is very close to the correlation we got from the LARS model. This showcases that the LARS model does a better job of predicting salaries. The data points are very spread out from the red linear regression line, showing a correlation that is not very high.

NW Predicting AE:

When using LARS to create an NW model to predict AE salaries, the correlation is 0.455. The correlation using LM is much lower at 0.37. The LARS did a far better job predicting salaries, possibly because it is more complex than simple linear regression with LM, thus providing more accurate results. Both plots, however, have the lowest correlation compared to other divisions predicting AE, showcasing that predicting across divisions does not provide accurate predictions. A stark difference exists between the red linear regression lines and the blue perfect prediction lines; they form a visible diagonal cross on both plots.

1) There is a difference in distributions between the 4 divisions. This is apparent because using other divisions to predict AE salary leads to very inaccurate predictions. If the distributions were all similar, there would be much greater correlation when using other divisions to predict AE. Instead, predicting within divisions, using AE to predict AE, yields the most accurate predictions that are very close to the actual observed values.
2) There is a difference in how salary depends on performance between the 4 divisions. This is apparent as some values that we predicted to be low were actually higher. This indicates that division-specific factors affect distribution and the salary one receives. The same is true for when

we observe the opposite, and actual salary values are much larger than we predicted, which occurred frequently in our analysis.