**R code for getting sales data for Product #1**

monthly.sales1 <- dbGetQuery(con,

"SELECT convert(date, DATEADD

(DAY, -1 \* DATEPART(DAY, soh.OrderDate)

+ 1, soh.OrderDate)) AS SalesMonth,

SUM(sod.LineTotal) AS MonthlySales,

'A' AS ActorPred

FROM Sales.SalesOrderDetail AS sod

JOIN Sales.SalesOrderHeader AS soh

on soh.SalesOrderID = sod.SalesOrderID

WHERE ProductID = 790

AND Year(soh.OrderDate) = 2013

GROUP BY DATEADD(DAY, -1 \* DATEPART(DAY, soh.OrderDate)

+ 1, soh.OrderDate)

ORDER BY SalesMonth")

**R code for date conversion**

monthly.sales1$SalesMonth <- as.Date(monthly.sales1$SalesMonth)

**R code for building linear regression models**

lm.fit <- lm(MonthlySales ~ SalesMonth, data = monthly.sales1)

**R code for plotting the model**

ggplot(monthly.sales1, aes(x=SalesMonth, y=MonthlySales)) + geom\_point() + geom\_smooth(method="lm") + labs(x="Month", y="Sales")

**R code predicting 2014 sales**

predicted.sales1 <- data.frame(SalesMonth = seq(from=as.Date('2014-01-01'), to=as.Date('2014-06-01'), by="month"), MonthlySales = 0, ActorPred="P")

predicted.sales1$MonthlySales <- predict(lm.fit, newdata = predicted.sales1)

all.months <- rbind(monthly.sales1, predicted.sales1)

**R code for getting sales data for Product #2**

monthly.sales2 <- dbGetQuery(con,

"SELECT convert(date, DATEADD

(DAY, -1 \* DATEPART(DAY, soh.OrderDate)

+ 1, soh.OrderDate)) AS SalesMonth2,

SUM(sod.LineTotal) AS MonthlySales2,

'A' AS ActorPred

FROM Sales.SalesOrderDetail AS sod

JOIN Sales.SalesOrderHeader AS soh

on soh.SalesOrderID = sod.SalesOrderID

WHERE ProductID = 792

AND Year(soh.OrderDate) = 2013

GROUP BY DATEADD(DAY, -1 \* DATEPART(DAY, soh.OrderDate)

+ 1, soh.OrderDate)

ORDER BY SalesMonth2")

**R code for date conversion**

monthly.sales2$SalesMonth2 <- as.Date(monthly.sales2$SalesMonth2)

**R code for building linear regression models**

lm.fit <- lm(MonthlySales2 ~ SalesMonth2, data = monthly.sales2)

**R code for plotting the model**

ggplot(monthly.sales2, aes(x=SalesMonth2, y=MonthlySales2)) + geom\_point() + geom\_smooth(method="lm") + labs(x="Month", y="Sales")

**R code predicting 2014 sales**

predicted.sales2 <- data.frame(SalesMonth2 = seq(from=as.Date('2014-01-01'), to=as.Date('2014-06-01'), by="month"), MonthlySales2 = 0, ActorPred="P")

predicted.sales2$MonthlySales2 <- predict(lm.fit, newdata = predicted.sales2)

all.months <- rbind(monthly.sales2, predicted.sales2)

View(all.months)

**R code for getting sales data for Product #3**

monthly.sales3 <- dbGetQuery(con,

"SELECT convert(date, DATEADD

(DAY, -1 \* DATEPART(DAY, soh.OrderDate)

+ 1, soh.OrderDate)) AS SalesMonth3,

SUM(sod.LineTotal) AS MonthlySales3,

'A' AS ActorPred

FROM Sales.SalesOrderDetail AS sod

JOIN Sales.SalesOrderHeader AS soh

on soh.SalesOrderID = sod.SalesOrderID

WHERE ProductID = 794

AND Year(soh.OrderDate) = 2013

GROUP BY DATEADD(DAY, -1 \* DATEPART(DAY, soh.OrderDate)

+ 1, soh.OrderDate)

ORDER BY SalesMonth3")

**R code for date conversion**

monthly.sales3$SalesMonth3 <- as.Date(monthly.sales3$SalesMonth3)

**R code for building linear regression models**

lm.fit <- lm(MonthlySales3 ~ SalesMonth3, data = monthly.sales3)

**R code for plotting the model**

ggplot(monthly.sales3, aes(x=SalesMonth3, y=MonthlySales3)) + geom\_point() + geom\_smooth(method="lm") + labs(x="Month", y="Sales")

**R code predicting 2014 sales**

predicted.sales3 <- data.frame(SalesMonth3 = seq(from=as.Date('2014-01-01'), to=as.Date('2014-06-01'), by="month"), MonthlySales3 = 0, ActorPred="P")

predicted.sales3$MonthlySales3 <- predict(lm.fit, newdata = predicted.sales3)

all.months <- rbind(monthly.sales3, predicted.sales3)

View(all.months)

**Component Questions.**

**Why might linear regression be a bad idea for forecasting sales when we only have one-year worth of data?**

With one year’s worth of data, it might be hard to test the model for homoscedasticity i.e., when you are looking for a constant deviation of the points from the zero-line. Without sufficient data (i.e., if the data points on a scatter plot do not form a cluster), the results of the model could be misleading.

**Do you believe these models could be used for any future date or is there likely a limit to how far out you can reliably predict?**

Using regression to make predictions doesn’t necessarily involve predicting the future. Instead, you can also predict the mean of the dependent variable given specific values of the independent variable(s).

**List the linear model equations for each of the three products. Provide an example of how to generate a prediction using these equations by hand**

From y = mx + b

Product #1

Y = (-327.1)x + 5239118.6

Product #2

Y = (41.58)x + -596779.96

Product #3

Y = (37.23)x + -494700.83

To predict Y (dependent variable) from x (independent variable) by hand, we would just plug any given values of x into the equations above.

For example, Y = (41.58)x + -596779.96 means that our line starts at **-596779.96**, and the y-values increase by **41.58** percentage points for every 1 order month.

**References**

Frost, J. (2021, April 1). Making predictions with regression analysis. Statistics By Jim. <https://statisticsbyjim.com/regression/predictions-regression/>

Korstanje, J. (2021, June 20). Assumptions of linear regression. Medium. <https://towardsdatascience.com/assumptions-of-linear-regression-fdb71ebeaa8b#:~:text=The%20sixth%20assumption%20of%20linear,residuals%20against%20the%20dependent%20variable>

Peterson, R. (2020, May 23). Linear regression by hand. Medium. <https://towardsdatascience.com/linear-regression-by-hand-ee7fe5a751bf>