a. Fit the above two models using least squares to all data. Compute AIC and adjusted R2 for two models. Which one is a better model?

Adjusted R^2 and AIC for the models

	M1	M2
Adjusted R^2	0.875	0.876
AIC	1551.17	1549.57

Table - 1

Based on the Adjusted R^2 and AIC of model one (M1) and model two (M2) shown as above, M2 has larger Adjusted R^2 and smaller AIC, which means M2 is the better model.

b. Write out each model in equation form, being careful to handle the qualitative variables properly.

M1: Units = $\begin{cases} 0.51205\text{Hours} + 1.67220\text{Lines} + 0.96296\text{Workers} - 7.53347 & \text{Region} = \text{South} \\ 0.51205\text{Hours} + 1.67220\text{Lines} + 0.96296\text{Workers} - 5.31524 & \text{Region} = \text{North} \end{cases}$

M2: Units = $\begin{cases} 0.512\text{Hours} + 1.66854\text{Lines} + 0.87474\text{Workers} - 6.7345 & \text{Region} = \text{South} \\ 0.512\text{Hours} + 1.66854\text{Lines} + 1.01639\text{Workers} - 5.78603 & \text{Region} = \text{North} \end{cases}$

c. Use the sample() function to split the original data into one training set with 70% of the original observations and one testing set with the rest of observations. Compute the prediction MSE associated with each model. Which model is a better one in terms of the prediction MSE?

At first, we used the set.seed function to fix the data to make sure we got the same result each time we ran this model. Then we split data to training data set and testing data set by the ratio of 7:3. We mapped out the data as follows.

By comparing the MSE ($MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$) of M1 and M2:

MSE for the models

	M1	M2
MSE	11.82	11.62

Table - 2

Since MSE is a measure of the overall model fit, we want to choose the one better representing the data. Thus, M2 is better.

d. Compute the ten-fold cross-validation error (MSE) associated with each model. Which model is a better one?

By comparing the MeanMSE of M1 and M2:

MeanMSE for the models

	M1	M2
MeanMSE	10.48	10.51

Table - 3

With the ten-fold cross-validation, M1 has smaller MSE equals 10.48, and M2 has larger MSE equals 10.51. So, the M2 is the better model.

e. Select the "better" model as the final model. Which predictors appear to have a statistically significant relationship to the response (Units)? How does each predictor affect the response?

According to the output from the code, the second model has a lower AIC and a higher Adjusted R^2 which indicate the first model does a better job in explaining the dataset. However, according to the MSE in hold-out and cross-validation, we notice some divergence between the outputs. The second model is better according to the hold-out method while the first model is better according to cross-validation method.

The dataset only has 300 data points which couldn't be considered big, hold-out method can be problematic (the two datasets might not be equally representative; they are not large enough). On the other hand, cross-validation method utilizes the data better and thus is more convincing.

Also, this model is developed to predict the Units, so we should take into consideration of the potential problem of overfitting. The cross-validation gives us a better sense of how the model will perform in prediction. As a result, we want to put more weight on the result of cross-validation.

In conclusion, we choose the first model according to the MSE from cross-validation.

P-values of the predictors

	Intercept	Hours	Lines	Workers	Region (South)	Workers * Region
p-value	<2e^-16	<2e^-16	<2e^-16	<2e^-16	0.2233	0.0606

Table - 4

According to the p-value of the predictors, the following predictors have significant relationships with Units: Hours, Lines and Workers.

Coefficient of the predictors

	Intercept	Hours	Lines	Workers	Region	Workers * Region
Coefficient	-5.79	0.51	1.67	1.02	-0.95	-0.14

Table - 5

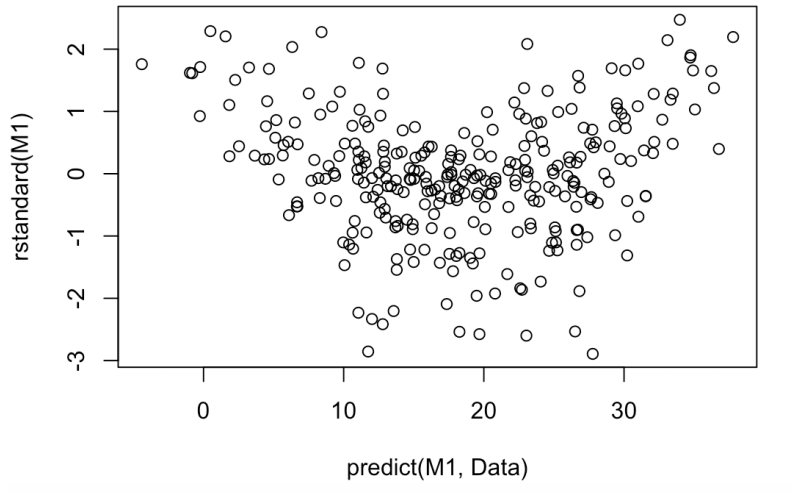
According to the coefficient of the predictors:

- 1 unit change in Hours will result in 0.512 unit change in Units
- 1 unit change in Lines will result in 1.67 unit change in Units
- 1 unit change in Workers will result in 1.02 unit change in Units
- 1 unit change in Region will result in (-0.95) unit change in Units
- 1 unit change in Worker*Region will result in (-0.14) unit change in Units

f. Is there evidence of outliers in the model selected from (e)? Please justify your answer.

In order to see if there is any outliers in the M1 model we have selected, we generated plots to have a better understanding.

Normalized Residual of M1



From the normalized residual of model 1, we can see none of the rstandard(M1) is outside of the range of the -3 and 3. Therefore, there is no outlier. But we can see some pattern in the plot which should be discussed further in the future.