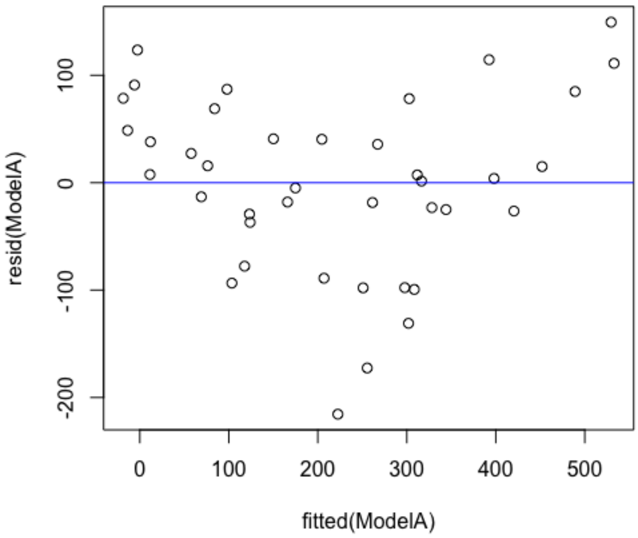
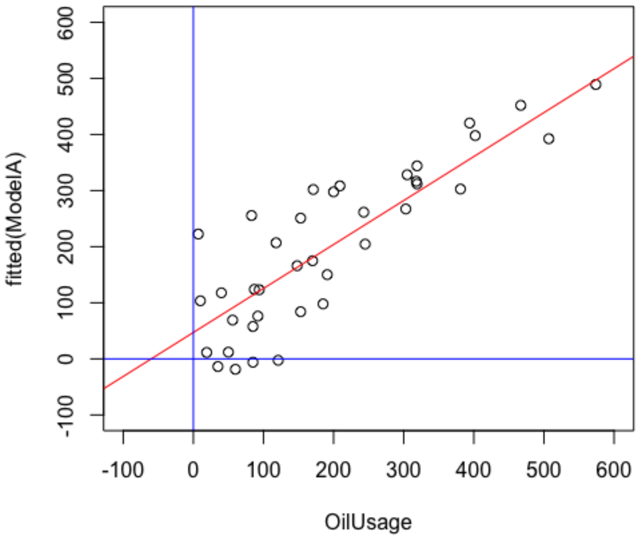
Model A – Linear Regression

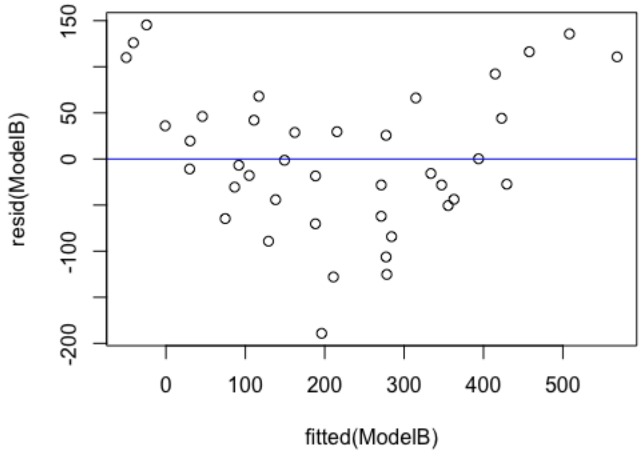
A regression model for OilUsage using all three variables (Degree Days, Home Factor, Number People) as the independent variables.

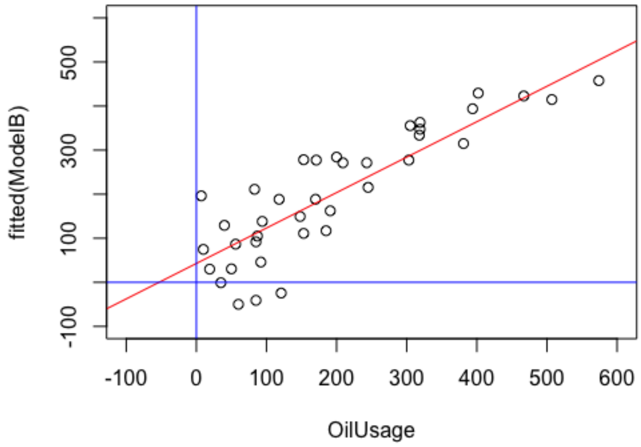




Model B – Adding Categorical Variables

Model A treats the HomeFactor variable as a numerical variable. Model B treats the HomeFactor variable as a categorical variable.





Q) Provide an economic interpretation of the coefficient of (HomeFactor level = 5).

Answer) If all other factors remain constant, if the coefficient of HomeFactor level = 5 increases by 1, the oil usage increases by 347.60906 gallons ~ 348 gallons.

Q) According to Model *B* estimated above, by how much higher/lower is the average oil consumption of customers in HomeFactor level 2 compared to the average oil consumption of customers in HomeFactor level 4, when DegreeDays and NumberPeople remain the same?

When DegreeDays and NumberPeople remain the same, the average oil consumption of customers in HomeFactor level 2 is 170.43542 ≈ 170 gallons lower than the average oil consumption of customers in HomeFactor level 4.

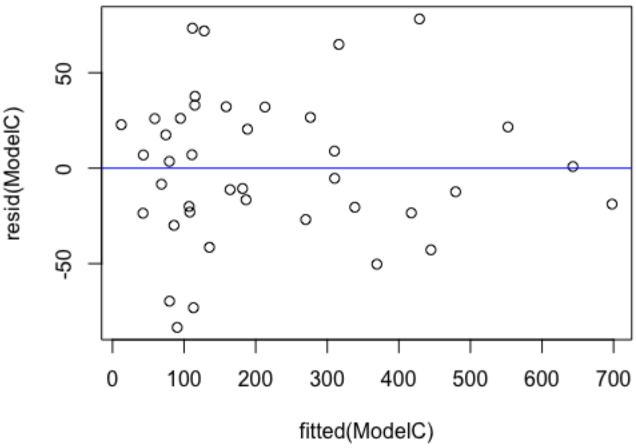
Q) Compare the performance of two models (Model A and Model B). Explain why use dummies for HomeFactor instead of the variable itself?

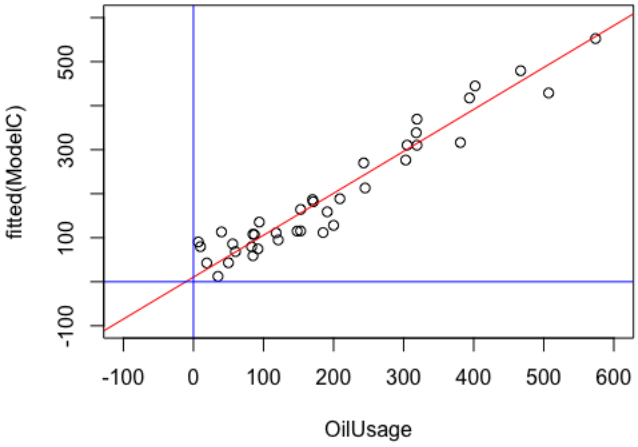
	Model A	Model B
R^2 value	0.784	0.804
Adjusted R^2 value	0.766	0.7684
Residual Standard Error	85.47	85.04
P value	4.547e-12	2.254e-10

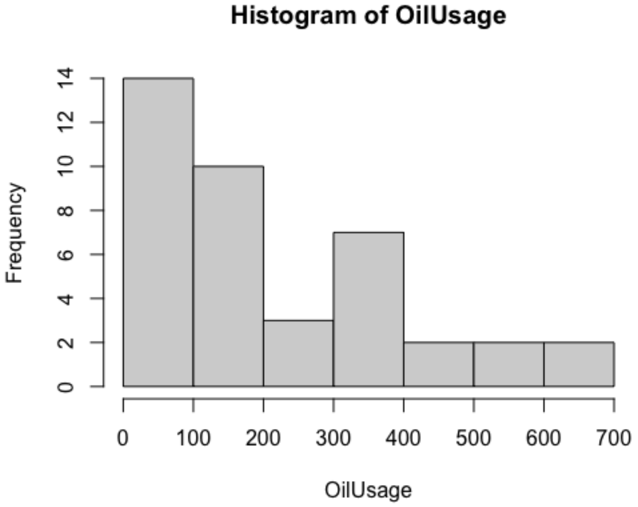
Looking at the above statistics, we can see that there is a very small improvement in model B over model A as the R2 and adjusted R2 values increase by a very small amount and residual standard error decreases very slightly. We introduce dummy variables because then we can use a single regression equation to model out different categories of HomeFactor instead of writing separate equations for each sub-category. If we used the variable itself, R may have not recognised it as a categorical variable and instead interpreted it a numerical variable which would not make sense as HomeFactor can't be compared based on factor value (home factor 2 is not twice of home factor 1)

Model C – Adding Interactions

Next, suppose it is conjectured that the *DegreeDays varies by HomeFactor*. To account for this conjecture, we augment Model *B* with interaction terms between DegreeDays and HomeFactor. Let us call this model Model *C*.







Q) Discuss the three plots. Is model C valid? Why or why not.

- Although the residual plot looks random, we notice that the scattered points are more concentrated to the left. Hence, we can assume that the residuals are not normally distributed and independent. However, this assumption is often questionable—the variation in Y often increases as X
increases.
- The scatter plot of fit vs. OilUsage seems to follow the regression line and hence we can assume that our model has a constant error variance. Our model is Homoscedastic- which means that the variability of Y values is constant, irrespective of the values of X.

- There is an obvious skewness in the histogram – it is positively skewed to the right. This indicates a violation of the normality assumption and hence we conclude our OilUsage data is not normally distributed.

  We can conclude that this model is not valid since the normality assumption is violated.

Q) Write out the estimated regression equations for each category of HomeFactor (five equations).

Equation for oil usage for HomeFactor1:

OilUsage = (0.05191*DegreeDays) + (12.74242*NumberPeople) -11.62366

Equation for oil usage for HomeFactor2:

OilUsage = (0.05191*DegreeDays) + (0.19301*DegreeDays*1) + (-27.61211*1) + (12.74242*NumberPeople) -11.62366

Equation for oil usage for HomeFactor3:

OilUsage = (0.05191*DegreeDays) + (0.25644*DegreeDays*1) + (15.74445*1) + (12.74242*NumberPeople) -11.62366

Equation for oil usage for HomeFactor4:

OilUsage = (0.05191*DegreeDays) + (0.47745*DegreeDays*1) + (-73.14821*1) + (12.74242*NumberPeople) -11.62366

Equation for oil usage for HomeFactor5:

OilUsage = (0.05191*DegreeDays) + (0.50518*DegreeDays*1) + (6.03819*1) + (12.74242*NumberPeople) -11.62366

Q) According to Model *C* estimated above, by how much higher/lower is the average oil consumption of customers in HomeFactor level 2 compared to the average oil consumption of customers in HomeFactor level 4 when DegreeDays = 1000 and NumberPeople is the same?

Answer) When DegreeDays = 1000,

HomeFactor level 2 oil usage

= 51.91 + 193.01 – 27.61211 -11.62366 + (12.74242*NumberPeople)

=205.68423 + (12.74242*NumberPeople)

HomeFactor level 4 oil usage

=51.91 + 477.45 – 73.14821 -11.62366 + (12.74242*NumberPeople)

=444.58813 + (12.74242*NumberPeople)

If NumberPeople is the same, the average oil consumption of customers in HomeFactor level 2 is 238.9039 ≈ 239 gallons lower than the average oil consumption of customers in HomeFactor level 4.

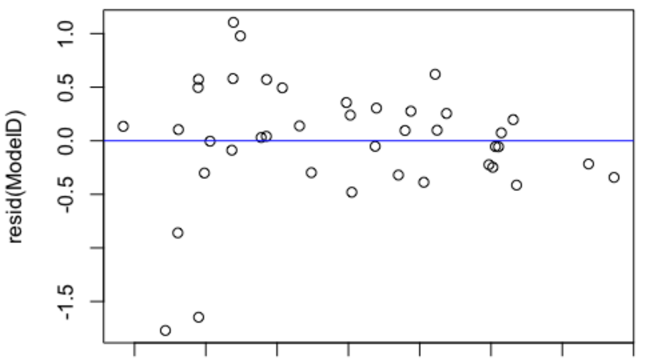
Q) Compare the performance of two models (Model B and Model C). Explain why use interaction terms between DegreeDays and HomeFactor?

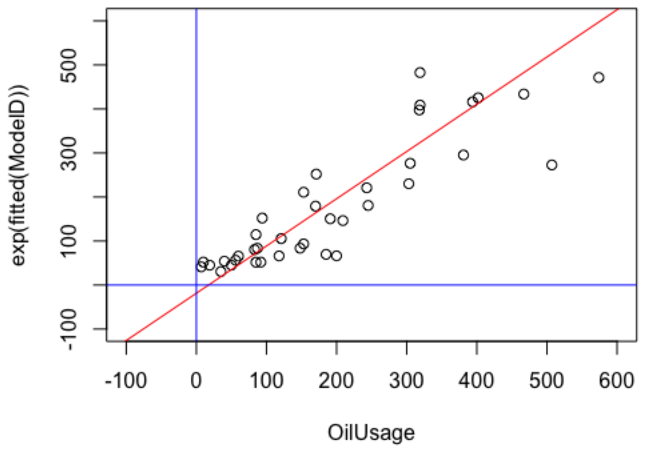
	Model B	Model C
R^2 value	0.804	0.9525
Adjusted R^2 value	0.7684	0.9361
Residual Standard Error	85.04	44.66
P value	2.254e-10	< 2.2e-16

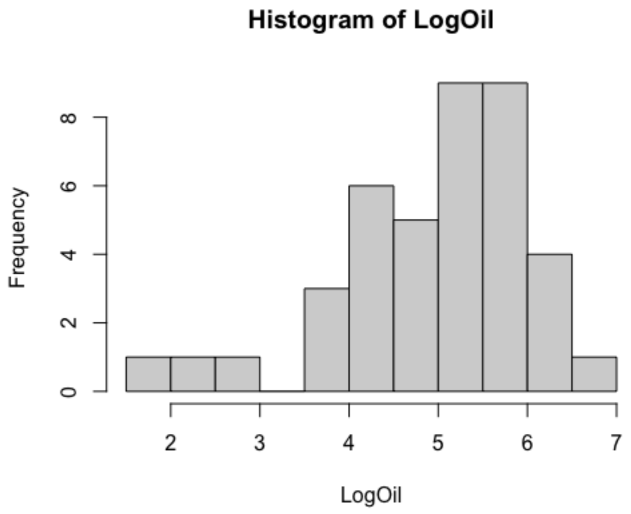
Based on the above statistics, model C outperforms model D because it has a higher R^2 and adjusted R^2 value and also a lower residual standard error. We should use interaction terms between DegreeDays and HomeFactor as this improvises the model (as indicated in the table above). This is because the effect of DegreeDays (independent variable) on Oil Usage depends on the level of HomeFactor which is another independent variable and hence we conclude they have interaction between them and go ahead and introduce an interaction term in the regression equation which significantly improves the model.

Model D – Nonlinear Regression

Model D is an exponential model by replacing OilUsage with the logarithmic transformation of OilUsage in Model C.







Q) Discuss the residual plot, the scatter plot of fit vs. OilUsage , and the histogram of log(OilUsage).

- The residual plot looks random, there is no noticeable pattern or trend. Hence, we can assume that the residuals are normally distributed and independent.
- The scatter plot of fit vs. OilUsage is partially concentrated around the regression line but we do have some outliers and hence we can assume that our model does not have a constant error variance. Our model is Heteroscedastic- which means that the variability of Y values is larger for some X values than for others.
- There is a obvious skewness in the histogram – it is negatively skewed to the left. This indicates a violation of the normality assumption and hence we conclude our OilUsage data is not normally distributed.

Q) Write out the estimated regression equations for each category of HomeFactor (five equations).

Equation for oil usage for HomeFactor1:

Log(OilUsage) = 2.993 + (9.731e-06*DegreeDays) + (2.373e-01*NumberPeople)

Equation for oil usage for HomeFactor2:

Log(OilUsage) = 2.993 + (9.731e-06*DegreeDays) + (2.013e-03*DegreeDays*1) + (-3.931e-01*1) + (2.373e-01*NumberPeople)

Equation for oil usage for HomeFactor3:

Log(OilUsage) = 2.993 + (9.731e-06*DegreeDays) + (1.928e-03*DegreeDays*1) + (2.622e-01*1) + (2.373e-01*NumberPeople)

Equation for oil usage for HomeFactor4:

Log(OilUsage) = 2.993 + (9.731e-06*DegreeDays) + (1.685e-03*DegreeDays*1) + (3.813e-01*1) + (2.373e-01*NumberPeople)

Equation for oil usage for HomeFactor5:

Log(OilUsage) = 2.993 + (9.731e-06*DegreeDays) + (2.072e-03*DegreeDays*1) + (5.086e-01*1) + (2.373e-01*NumberPeople)

Q) Estimate the oil consumption of a customer with DegreeDays =380, NumberPeople =4, HomeFactor = 1.

Equation for oil usage for HomeFactor1:

Log(OilUsage) = 2.993 + (9.731e-06*DegreeDays) + (2.373e-01*NumberPeople)

Log(OilUsage) = 3.9459

OilUsage = exp(3.9459) = 51.72 gallons ≈ 52 gallons.

Q) Compare the performance of two models (Model C and Model D). Which model will you use to estimate the oil consumption? Explain why or why not use the logarithmic transformation of OilUsage.

	Model C	Model D
R^2 value	0.9525	0.7264
Adjusted R^2 value	0.9361	0.632
Residual Standard Error	44.66	0.651
P value	< 2.2e-16	7.323e-06

Based on the above statistics, Model C clearly outperforms Model D. It has a higher R^2 value and adjusted R^2 value and also a lower residual standard error implying the regression model better fits our dataset.

We should not use the logarithmic transformation of OilUsage because it is causing the model to be less fitted to the data (indicated by a higher residual standard error and lower R^2 and adjusted R^2 values. Also, the logarithmic transformation is not helping normalize the data either as indicated by the histogram and hence we should not use it.