Smith Fules sells heating oil to residential customers and would like to build a model to predict its customer's oil consumption.

Oil customers are exposed to the risk of running out of fuel. Home heating oil suppliers therefore have to guarantee that the customer's oil tank will not be allowed to run dry. Home heating oil industry try uses the concept of a *degree-day*, equal to the difference between the average daily temperature and 68 degree Fahrenheit. So if the average temperature on a given day is 50, the degree-days for that day will be 18. If the degree-day calculation results in a negative number, the degree-day number is recorded as 0.

By keeping track of the number of degree-days since the customer's last oil fill, knowing the size of the customer's oil tank, and estimating the customer's oil consumptions as a function of the number of degree-days, the oil supplier can estimate when the customer is getting low on fuel and then resupply the customer. However, Smith has more than 2000 customers and computational burden of keeping track of all of these customers is enormous.

Smith wants to develop a consumption estimation model that is practical and reliable.

The file usage.xlsx contains recent oil usage of 40 customers recent with the following variables:

- OilUsage: The oil consumption amounts in gallons for 40 customers.
- DegreeDays: The number of degree-days since the last oil fill for 40 customers.
- HomeFactor: An assessment of the home type of each of the 40 customers (levels= {1,2,3,4,5}).
- NumberPeople: The number of people residing in the home of each of the 40 customers.

Use R to conduct the statistical analysis asked below. For questions that ask for an oil usage (or change in oil usage), use zero decimal places in your final numerical answer.

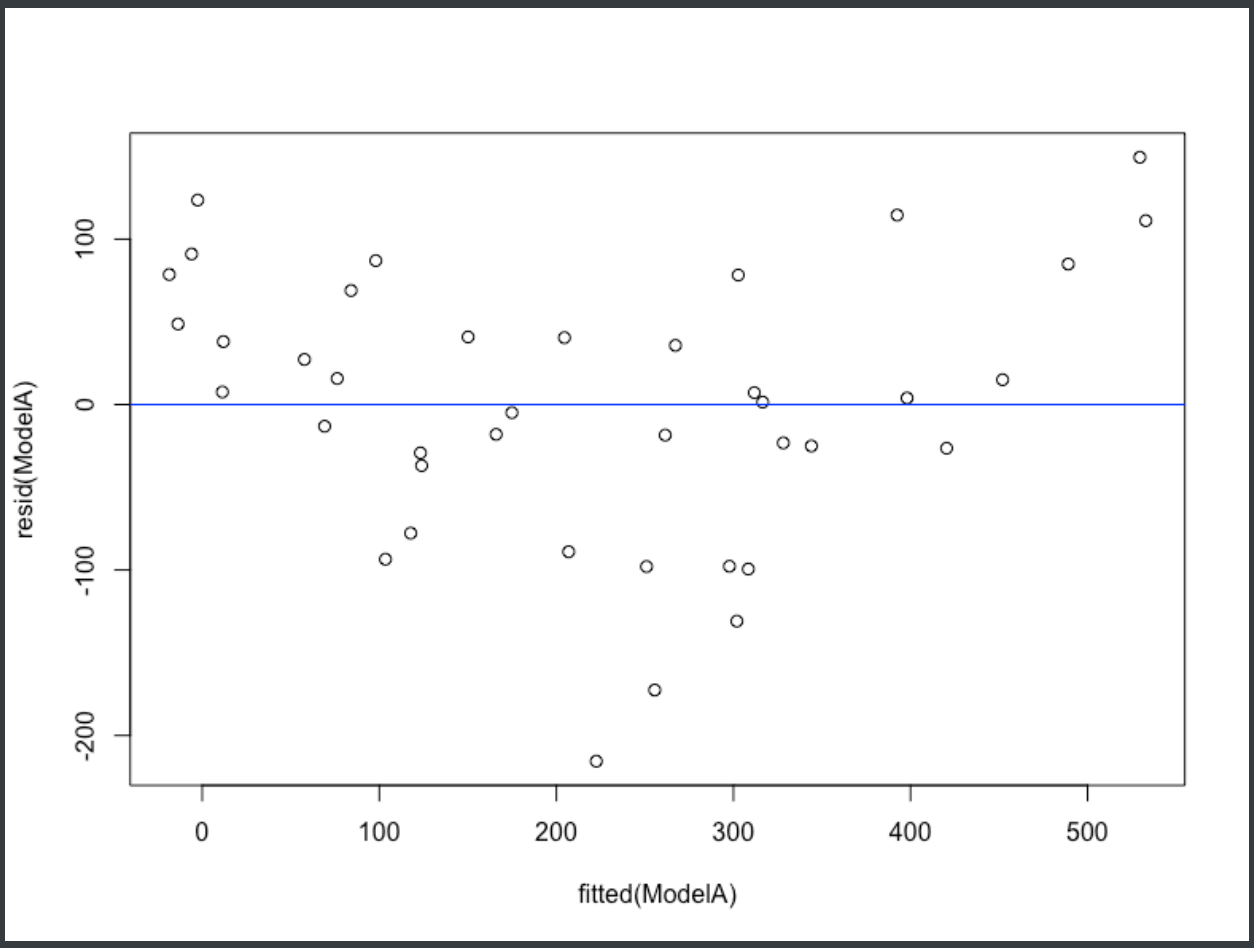
Part A – Linear Regression (6 points)

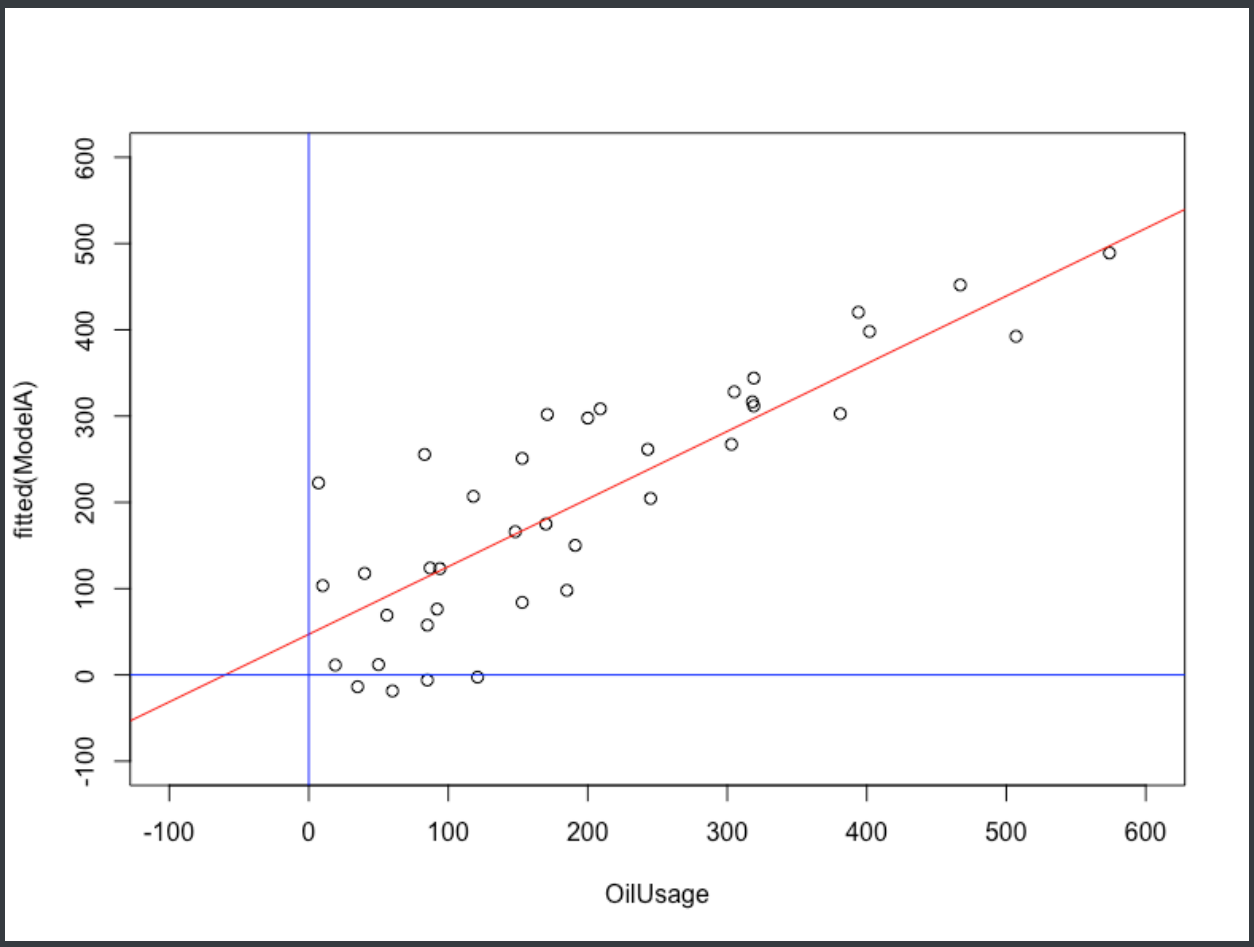
Create a regression model for OilUsage using all three variables (Degree Days, Home Factor, Number People) as the independent variables. Let us call this Model *A*. Create the residual plot and the scatter plot of fit vs. OilUsage. Use the following R-script to build the regression model and the plots:

1. (2 pts) Copy and paste the R regression output and the plots*.**

Solution:

```
> ModelA <- lm(OilUsage~DegreeDays+HomeFactor+NumberPeople)
> summary(ModelA)
"
Call:
lm(formula = OilUsage ~ DegreeDays + HomeFactor + NumberPeople)

Residuals:
     Min       1Q   Median       3Q      Max
-215.553  -31.148    5.583   53.743  149.461

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  -218.30987   63.95851  -3.413   0.0016 **
DegreeDays      0.27508    0.03633   7.571 5.94e-09 ***
HomeFactor     86.98875    9.63044   9.033 8.75e-11 ***
NumberPeople    5.26724   10.56179   0.499   0.6210
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 85.47 on 36 degrees of freedom
Multiple R-squared:  0.784, Adjusted R-squared:  0.766
F-statistic: 43.57 on 3 and 36 DF,  p-value: 4.547e-12
"
>
> plot(resid(ModelA)~fitted(ModelA))
> abline(h=0, col='blue')
>
> plot(OilUsage, fitted(ModelA), xlim=c(-100,600), ylim=c(-100,600))
> abline(lm(fitted(ModelA)~OilUsage), col='red')
> abline(h=0, col='blue')
> abline(v=0, col='blue')
```


resid(ModelA)
100
0
-100
-200
0
100
200
300
400
500
fitted(ModelA)


fitted(ModelA)
600
500
400
300
200
100
0
-100
-100
0
100
200
300
400
500
600
OilUsage

2. (2 pts) Write out the estimated regression equation.

   Solution:

   `OilUsage = 0.27508*DegreeDays+86.98875*HomeFactor+5.26724*NumberPeople-218.30987`

3. (2 pts) Provide an economic interpretation of the coefficient of NumberPeople.

   Solution:

   $b_{people} = 5.26724$:

   - When number of people increase by 1, on average, the oil usage **increases 5.26724**, while all the other factors remain the same.

Model B – Adding Categorical Variables (12 points)

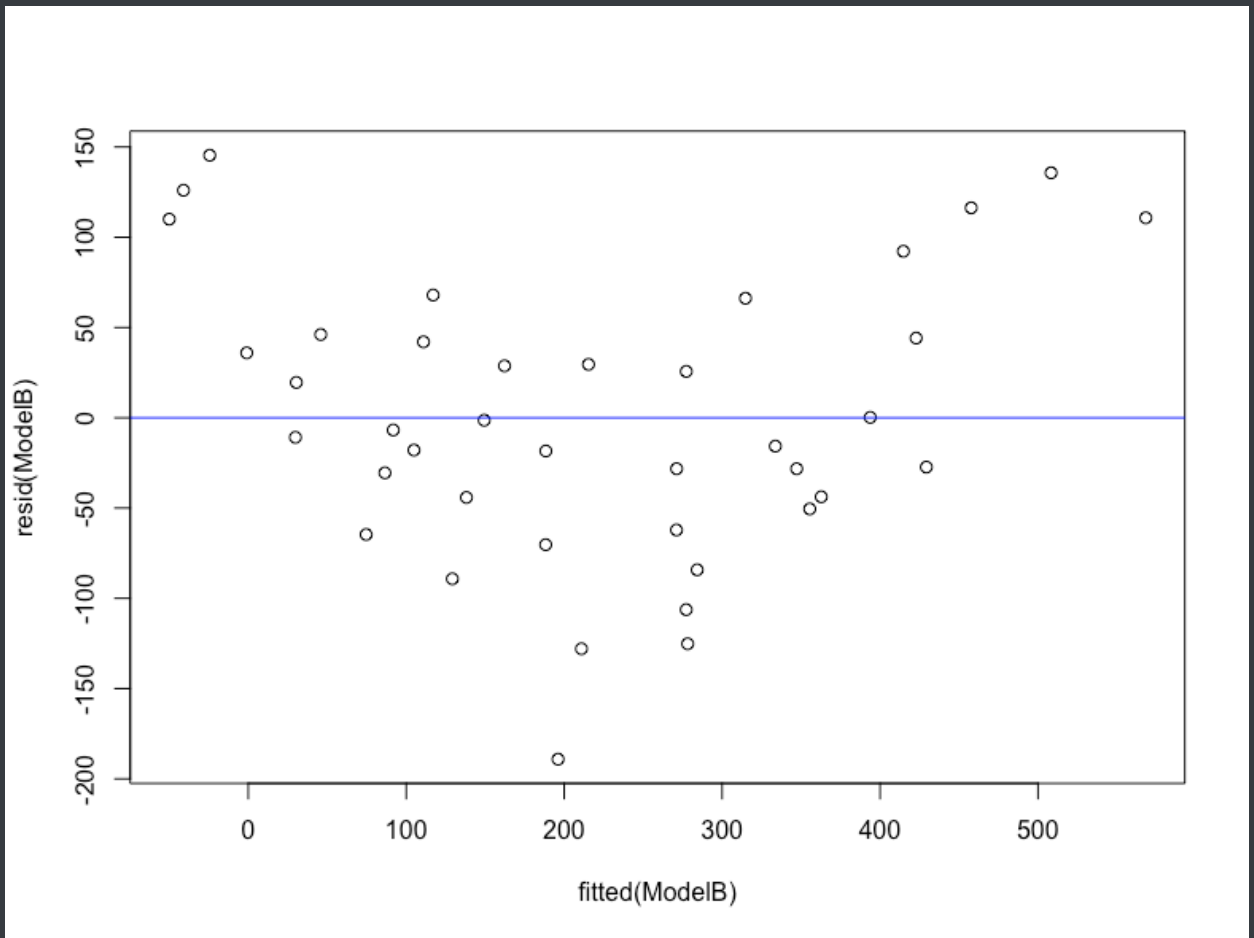
Model A treats the HomeFactor variable as a numerical variable. Build a model, which treats the HomeFactor variable as a categorical variable. Let us refer to this model as Model *B*. Create the residual plot and the scatter plot of fit vs. OilUsage.

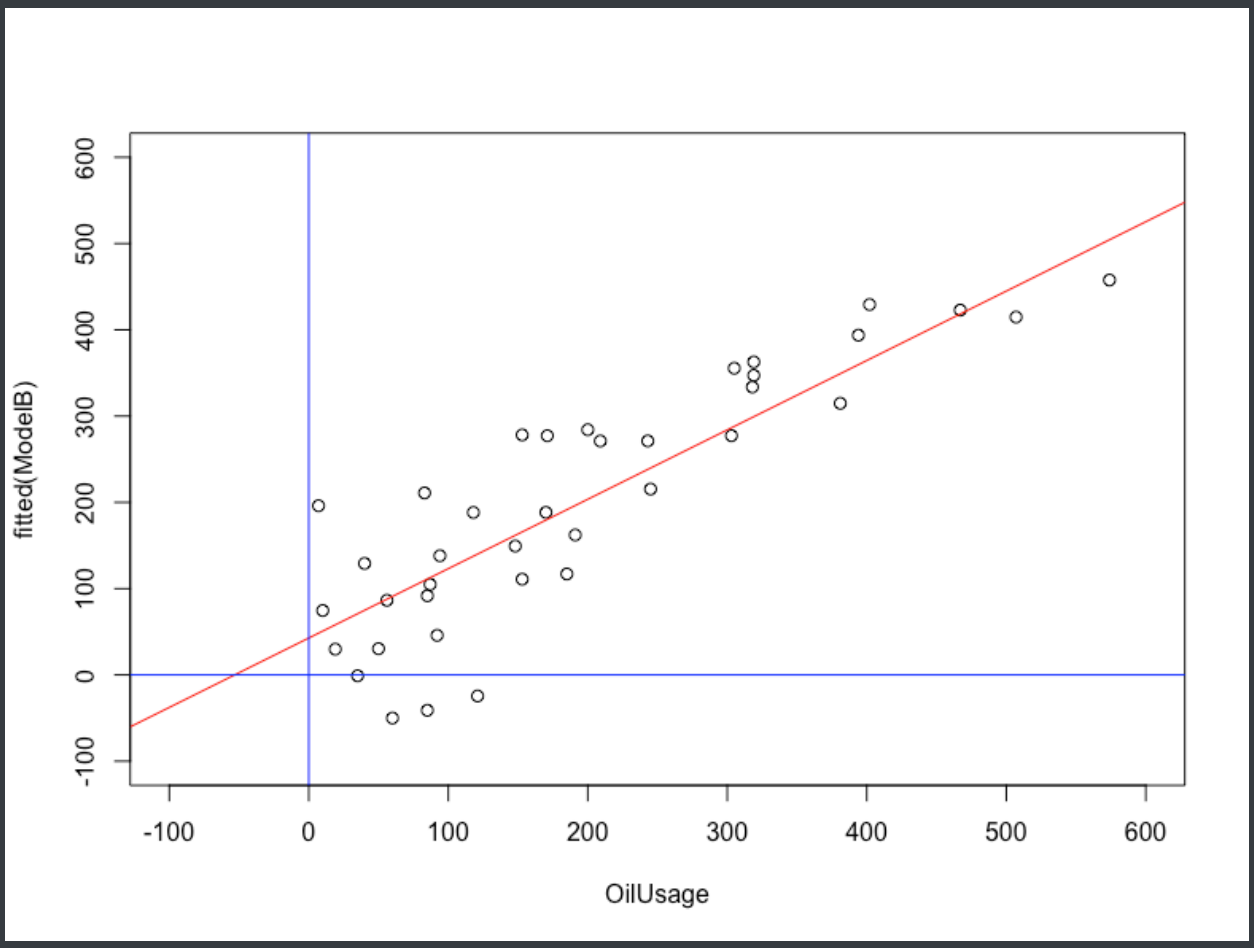
1. (3 pts) Copy and paste the R code, the regression output, and the plots.

   Solution:

```
> ModelB <- lm(OilUsage~DegreeDays+factor(HomeFactor)+NumberPeople)
> summary(ModelB)
"
Call:
lm(formula = OilUsage ~ DegreeDays + factor(HomeFactor) + NumberPeople)

Residuals:
     Min       1Q   Median       3Q      Max
-189.127  -45.724   -8.817   44.647  145.417

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         -190.74293   79.73080  -2.392  0.02259 *
DegreeDays             0.29063    0.04174   6.963 5.84e-08 ***
factor(HomeFactor)2  144.39634   46.34862   3.115  0.00379 **
factor(HomeFactor)3  217.78307   41.33502   5.269 8.35e-06 ***
```

```
17  factor(HomeFactor)4  314.83176   46.58965   6.758 1.06e-07 ***
18  factor(HomeFactor)5  347.60906   44.25996   7.854 4.71e-09 ***
19  NumberPeople           9.89345   11.63920   0.850  0.40144
20  ---
21  Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
22
23  Residual standard error: 85.04 on 33 degrees of freedom
24  Multiple R-squared:  0.804, Adjusted R-squared:  0.7684
25  F-statistic: 22.57 on 6 and 33 DF,  p-value: 2.254e-10
26  "
27  >
28  > plot(resid(ModelB)~fitted(ModelB))
29  > abline(h=0, col='blue')
30  >
31  > plot(OilUsage, fitted(ModelB), xlim=c(-100,600), ylim=c(-100,600))
32  > abline(lm(fitted(ModelB)~OilUsage), col='red')
33  > abline(h=0, col='blue')
34  > abline(v=0, col='blue')
```





2. (3 pts) Write out the estimated regression equations for each category of HomeFactor (five equations).

   Solution:

   - HomeFactor = 1:

     `OilUsage = 0.29063*DegreeDays+9.89345*NumberPeople-190.74293`

   - HomeFactor = 2:

     `OilUsage = 0.29063*DegreeDays+9.89345*NumberPeople-46.34659`

   - HomeFactor = 3:

     `OilUsage = 0.29063*DegreeDays+9.89345*NumberPeople+27.04014`

   - HomeFactor = 4:

     `OilUsage = 0.29063*DegreeDays+9.89345*NumberPeople+124.08883`

   - HomeFactor = 5:

     `OilUsage = 0.29063*DegreeDays+9.89345*NumberPeople+156.86613`

3. (2 pts) Provide an economic interpretation of the coefficient of (HomeFactor level = 5).

Solution:

For HomeFactor level = 5, as $b_5 = 347.60906$:

- On average, the oil usage is **347.60906 higher** if home level equals 5 (compare to HomeFactor level = 1), for the same DegreeDays and NumberPeople.

4. (2 pts) 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?

   Solution:

$$\begin{aligned}\Delta &= \text{OilUsage}(\text{HomeFactor} = 2) - \text{OilUsage}(\text{HomeFactor} = 4) \\ &= (0.29063 * \text{DegreeDays} + 9.89345 * \text{NumberPeople} - 190.74293 + 144.39634) \\ &\quad - (0.29063 * \text{DegreeDays} + 9.89345 * \text{NumberPeople} - 190.74293 + 314.83176) \\ &= 144.39634 - 314.83176 = -170.43542\end{aligned}$$

   Thus, the average oil consumption of customers in HomeFactor level 2 is **170.43542** lower, compared to the average oil consumption of customers in HomeFactor level 4.

5. (2 pts) Compare the performance of two models (Model A and Model B). Explain why use dummies for HomeFactor instead of the variable itself?

   Solution:

   To start with, we say that using dummies can always get a better model than using variable itself, since we are calculating intercept for each category seperately, instead of relating category number to intercept linearly. At the most optimal situation the both methods could give the same result, when category number is exactly linearly related with intercept, but through the above calculation, it's not. Thus, Model B performs better than A and we use dummies for HomeFactor.

Part C – Adding Interactions (17 points)

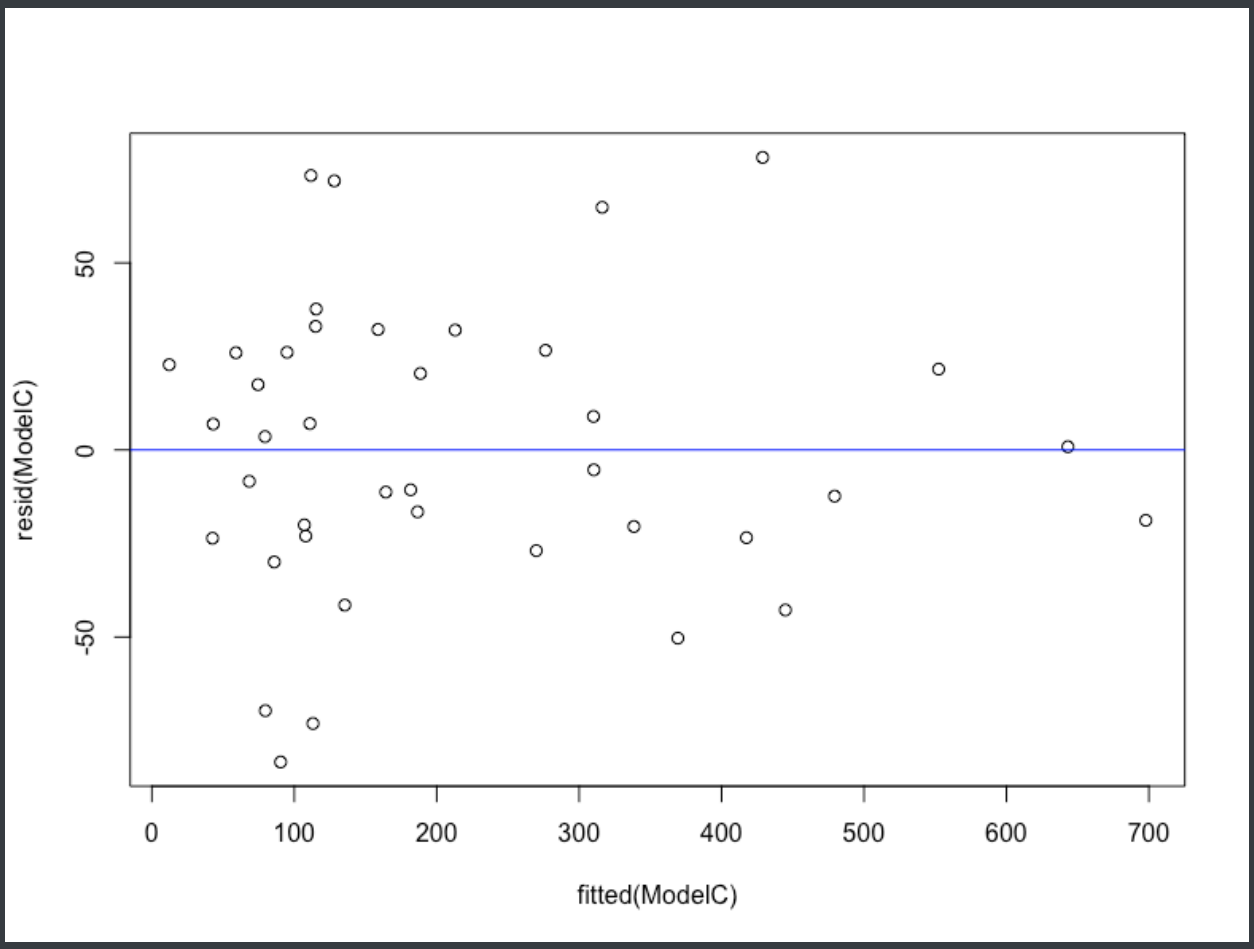
Next, suppose it is conjectured that the *DegreeDays varies by HomeFactor*. To account for this conjecture, we augment Model *B* with interaction terms between DegreeDays and HomeFactor. Let us call this model Model *C*. Create the residual plot and the scatter plot of fit vs. OilUsage. Also create a histogram of OilUsage.

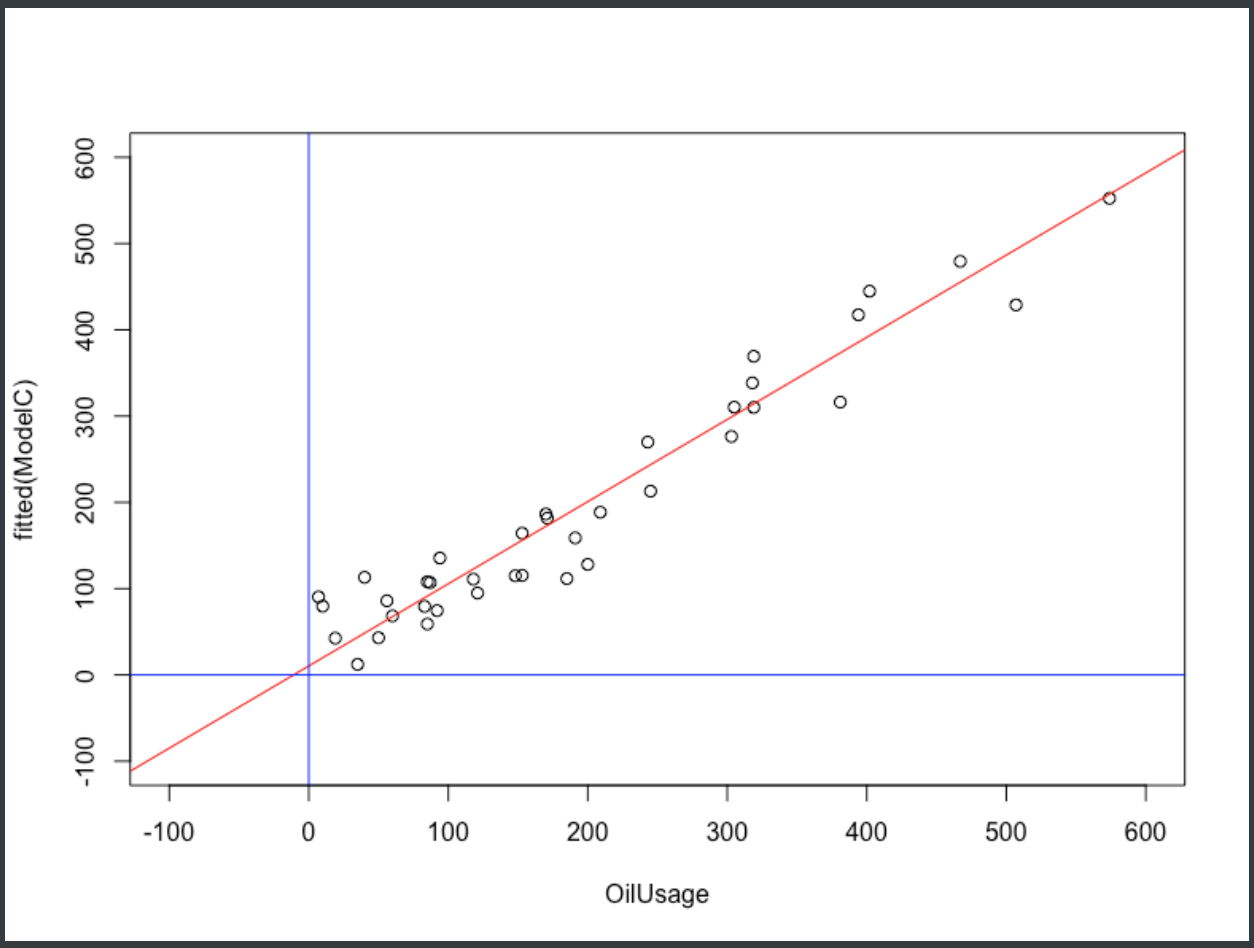
1. (3 pts) Copy and paste the R code, the regression output, and the plots.

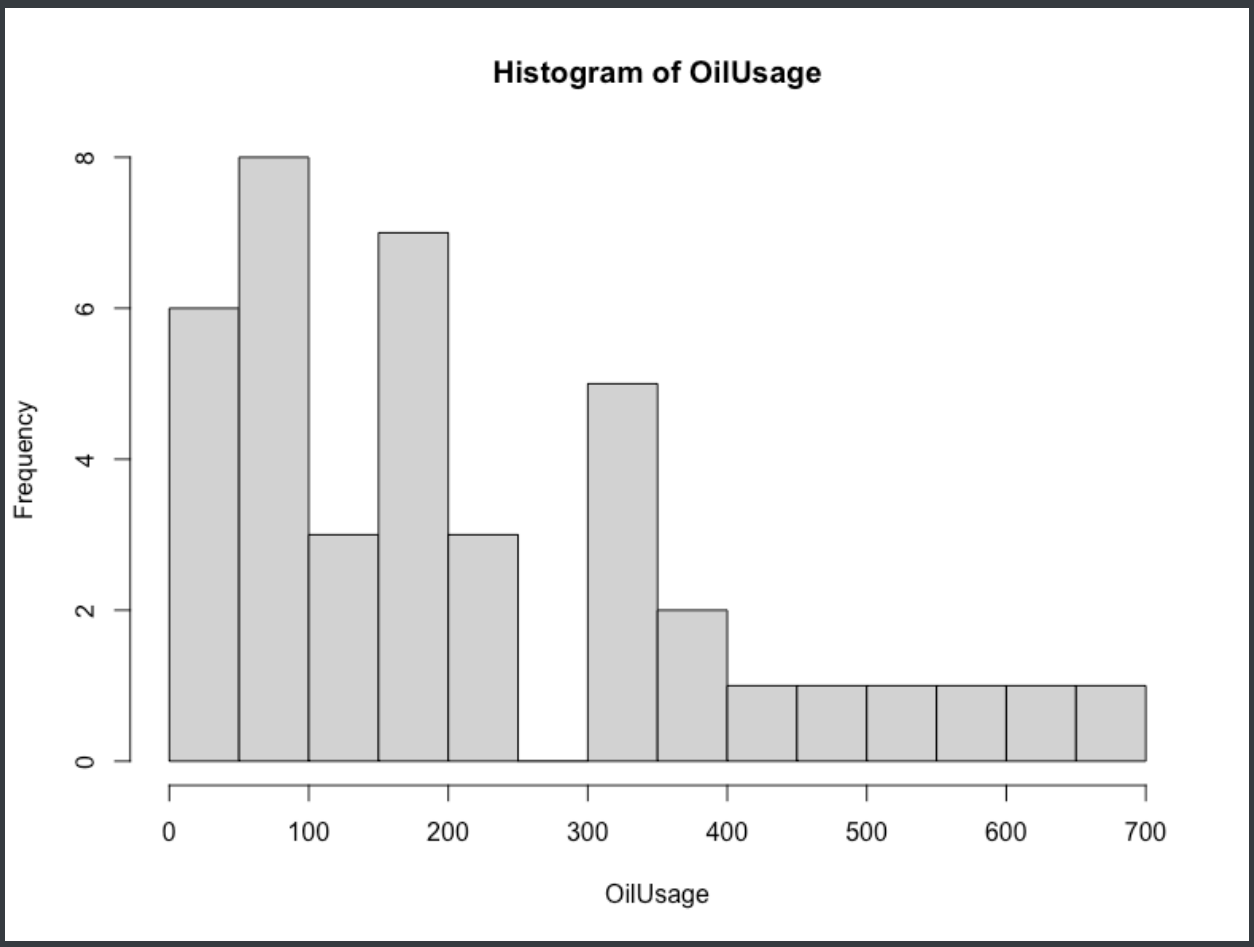
Solution:

```
> ModelC <-
lm(OilUsage~DegreeDays+factor(HomeFactor)+factor(HomeFactor)*DegreeDays+N
umberPeople)
> summary(ModelC)
"
Call:
lm(formula = OilUsage ~ DegreeDays + factor(HomeFactor) +
factor(HomeFactor) *
    DegreeDays + NumberPeople)

Residuals:
    Min      1Q  Median      3Q     Max
-83.403 -23.118  -2.231  26.007  78.175

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                    -11.62366   47.99418  -0.242 0.810338
DegreeDays                       0.05191    0.03789   1.370 0.181238
factor(HomeFactor)2            -27.61211   41.62147  -0.663 0.512307
factor(HomeFactor)3             15.74445   49.59466   0.317 0.753167
factor(HomeFactor)4            -73.14821   61.81585  -1.183 0.246291
factor(HomeFactor)5              6.03819   45.95053   0.131 0.896361
NumberPeople                    12.74242    6.24764   2.040 0.050598 .
DegreeDays:factor(HomeFactor)2   0.19301    0.05840   3.305 0.002535 **
DegreeDays:factor(HomeFactor)3   0.25644    0.06590   3.891 0.000537 ***
DegreeDays:factor(HomeFactor)4   0.47745    0.06967   6.853 1.58e-07 ***
DegreeDays:factor(HomeFactor)5   0.50518    0.05982   8.444 2.64e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 44.66 on 29 degrees of freedom
Multiple R-squared:  0.9525,  Adjusted R-squared:  0.9361
F-statistic: 58.16 on 10 and 29 DF,  p-value: < 2.2e-16
"
>
> plot(resid(ModelC)~fitted(ModelC))
> abline(h=0, col='blue')
```

```
35 >
36 > plot(OilUsage, fitted(ModelC), xlim=c(-100,600), ylim=c(-100,600))
37 > abline(lm(fitted(ModelC)~OilUsage), col='red')
38 > abline(h=0, col='blue')
39 > abline(v=0, col='blue')
40 > hist(OilUsage, breaks = 14)
41 > summary(OilUsage)
42 "
43    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
44     7.0    85.0   170.5   218.1   318.2   679.0
45 "
```




fitted(ModelC)
-100
0
100
200
300
400
500
600
OilUsage


Histogram of OilUsage
Frequency
0
2
4
6
8
0
100
200
300
400
500
600
700
OilUsage

2. (3 pts) Discuss the residual plot, the scatter plot of fit vs. OilUsage , and the histogram of OilUsage.

   Solution:

   The residual plot are skewed to the right, and the histogram shows the same trend, which means that OilUsage variable needs to be normalized through some non-linear transformation. From the scatter plot we can also see the trend that most points are distributed in the corner, because of the right-skewed distribution of OilUsage. And as the point are close to the regression line, which is close to a 45 degree line.

3. (3 pts) Write out the estimated regression equations for each category of HomeFactor (five equations).

   Solution:

   - HomeFactor = 1:

     `OilUsage = 0.05191*DegreeDays+12.74242*NumberPeople-11.62366`

   - HomeFactor = 2:

     `OilUsage = 0.24492*DegreeDays+12.74242*NumberPeople-39.23577`

   - HomeFactor = 3:

     `OilUsage = 0.30835*DegreeDays+12.74242*NumberPeople+4.12079`

   - HomeFactor = 4:

     `OilUsage = 0.52936*DegreeDays+12.74242*NumberPeople-84.77187`

   - HomeFactor = 5:

     `OilUsage = 0.55709*DegreeDays+12.74242*NumberPeople-5.58547`

4. (2 pts) Provide an economic interpretation of the coefficient of NumberPeople.

   Solution:

   $b_{people} = 12.74242$:

   - When number of people increase by 1, on average, the oil usage **increases 12.74242**, while all the other factors remain the same.

5. (2 pts) 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?

   Solution:

$$\begin{aligned}\Delta &= \text{OilUsage}(\text{HomeFactor} = 2) - \text{OilUsage}(\text{HomeFactor} = 4) \\ &= (0.24492 * \text{DegreeDays} + 12.74242 * \text{NumberPeople} - 39.23577) \\ &\quad - (0.52936 * \text{DegreeDays} + 12.74242 * \text{NumberPeople} - 84.77187) \\ &= (0.19301 - 0.47745) * \text{DegreeDays} + (-27.61211 + 73.14821) \\ &= -238.9039\end{aligned}$$

   Thus, the average oil consumption of customers in HomeFactor level 2 is **238.9039** lower, compared to the average oil consumption of customers in HomeFactor level 4.

6. (2 pts) Estimate the oil consumption of a customer with DegreeDays =380, NumberPeople =4, HomeFactor = 1.

   Solution:

$$\begin{aligned}&\text{OilUsage}(\text{DegreeDays} = 380, \text{NumberPeople} = 4, \text{HomeFactor} = 1) \\ &= 0.05191 * \text{DegreeDays} + 12.74242 * \text{NumberPeople} - 11.62366 \\ &= 0.05191 * 380 + 12.74242 * 4 - 11.62366 \\ &= 59.07182\end{aligned}$$

   Thus, the oil consumption of a customer with DegreeDays =380, NumberPeople =4, HomeFactor = 1 is **59.07182**.

7. (2 pts) Compare the performance of two models (Model B and Model C). Explain why use interaction terms between DegreeDays and HomeFactor?

   Solution:

   Similar to 5. in part B, here we also have Model C always greater or at least equal to Model B, since in Model C we take the conjection that the *DegreeDays varies by HomeFactor* into our consideration. At the most optimal situation the both methods could give the same result, when *DegreeDays and HomeFactor* are **independent**, but through the above calculation, it's not (in fact they are highly related). Thus, Model C performs better than B and we use interaction terms between *DegreeDays* and *HomeFactor*.

Part D – Nonlinear Regression (15 points)

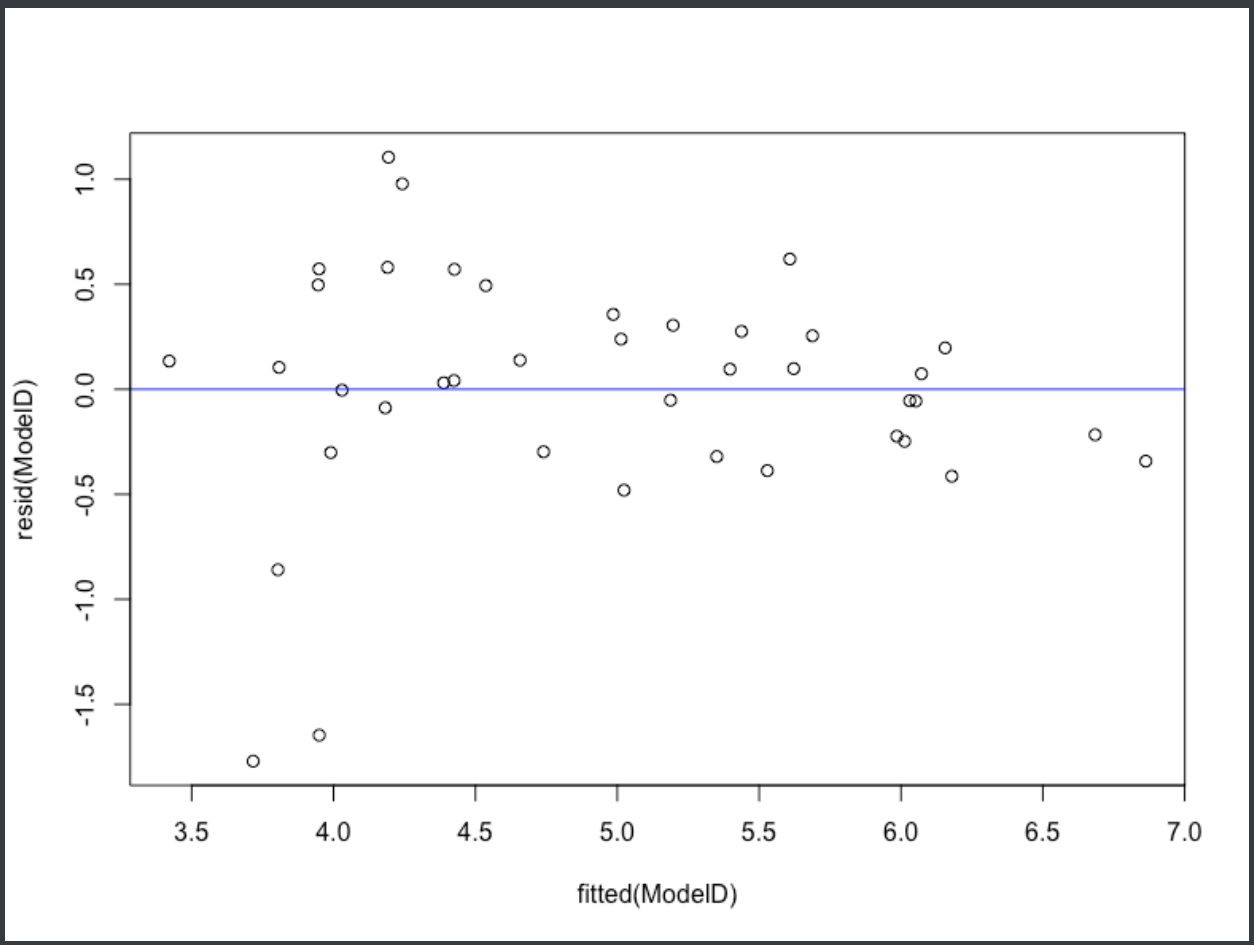
Build an exponential model by replacing OilUsage with the logarithmic transformation of OilUsage in Model C. Let us call this model Model *D*. Create the residual plot and the scatter plot of fit vs. OilUsage (not log(OilUsage)). Also create a histogram of log(OilUsage). Recall that for our purposes, “log” refers to natural logarithms.

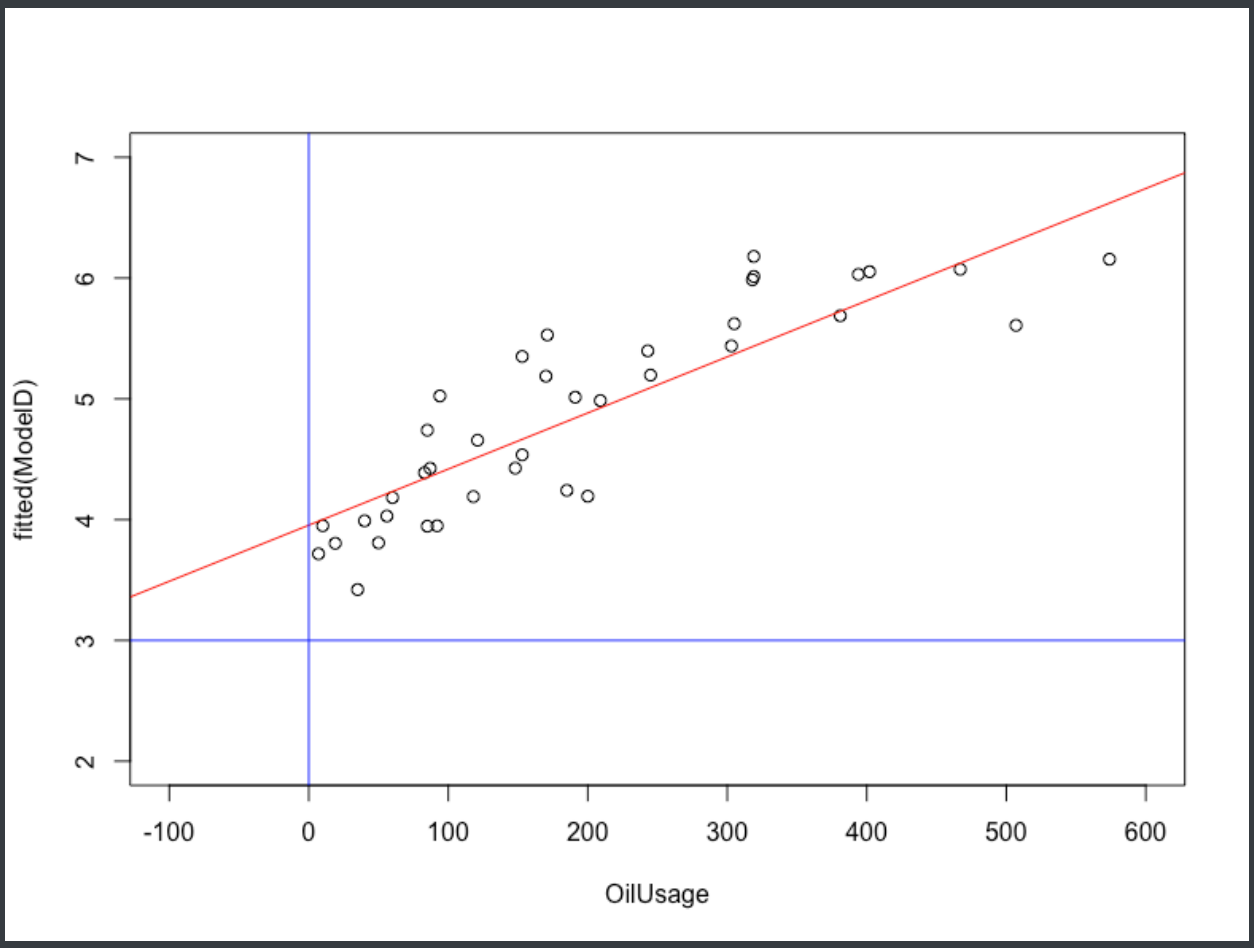
1. (3 pts) Copy and paste the R code, the regression output, and the plots.

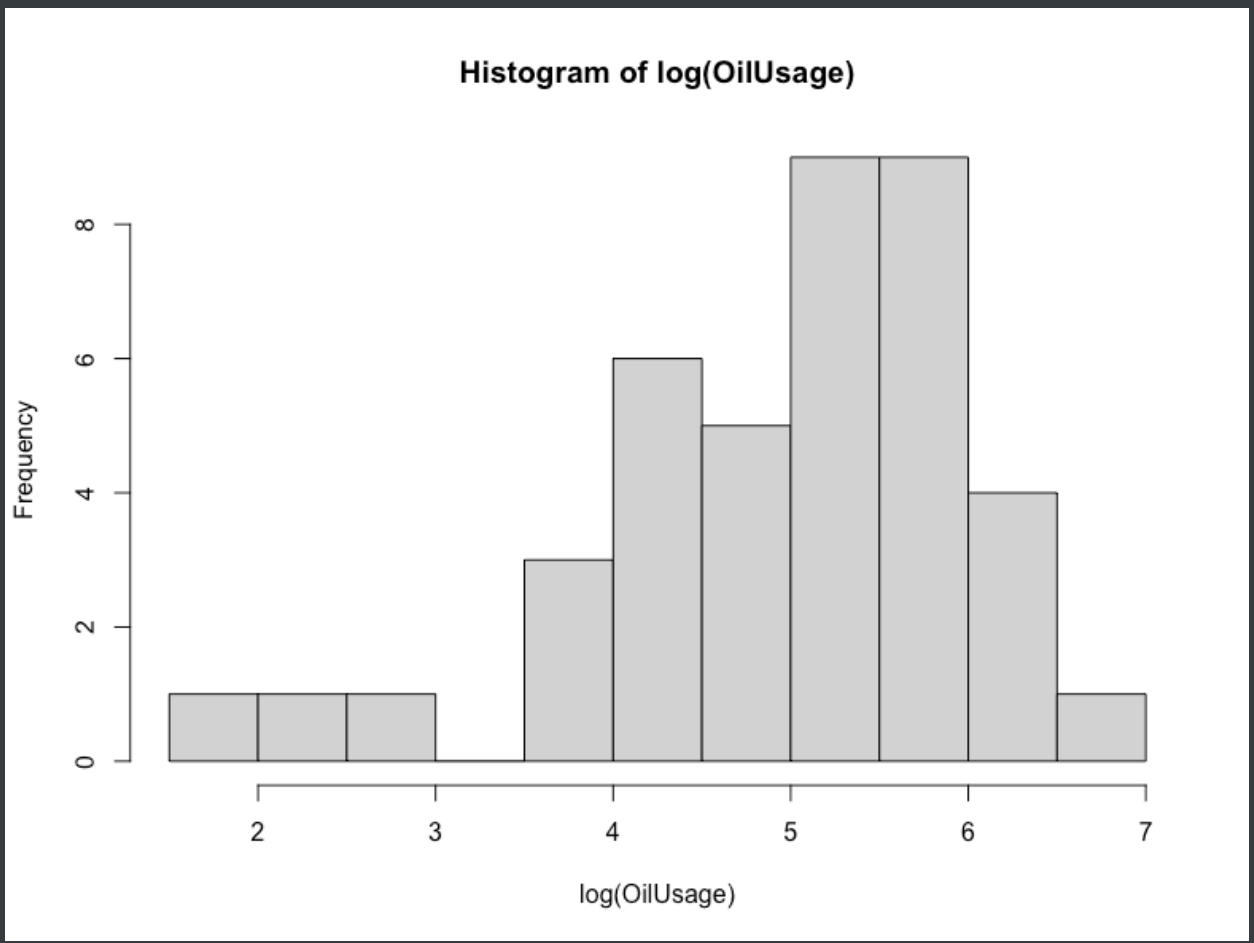
   Solution:

```
> ModelD <-
lm(log(OilUsage)~DegreeDays+factor(HomeFactor)+factor(HomeFactor)*DegreeD
ays+NumberPeople)
> summary(ModelD)
"
Call:
lm(formula = log(OilUsage) ~ DegreeDays + factor(HomeFactor) +
    factor(HomeFactor) * DegreeDays + NumberPeople)

Residuals:
     Min       1Q   Median       3Q      Max
-1.77083 -0.26011  0.05766  0.28296  1.10469

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                     2.993e+00  6.996e-01   4.278 0.000187 ***
DegreeDays                      9.731e-06  5.524e-04   0.018 0.986065
factor(HomeFactor)2            -3.931e-01  6.067e-01  -0.648 0.522152
factor(HomeFactor)3             2.622e-01  7.229e-01   0.363 0.719456
factor(HomeFactor)4             3.813e-01  9.011e-01   0.423 0.675272
factor(HomeFactor)5             5.086e-01  6.698e-01   0.759 0.453808
NumberPeople                    2.373e-01  9.107e-02   2.606 0.014324 *
DegreeDays:factor(HomeFactor)2  2.013e-03  8.513e-04   2.365 0.024958 *
DegreeDays:factor(HomeFactor)3  1.928e-03  9.606e-04   2.007 0.054109 .
DegreeDays:factor(HomeFactor)4  1.685e-03  1.016e-03   1.660 0.107789
DegreeDays:factor(HomeFactor)5  2.072e-03  8.721e-04   2.376 0.024325 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.651 on 29 degrees of freedom
```

```
Multiple R-squared:  0.7264,  Adjusted R-squared:  0.632
F-statistic: 7.698 on 10 and 29 DF,  p-value: 7.323e-06
"
>
> plot(resid(ModelD)~fitted(ModelD))
> abline(h=0, col='blue')
>
> plot(OilUsage, fitted(ModelD), xlim=c(-100,600), ylim=c(2,7))
> abline(lm(fitted(ModelD)~OilUsage), col='red')
> abline(h=3, col='blue')
> abline(v=0, col='blue')
>
> hist(log(OilUsage))
> summary(OilUsage)
"
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    7.0    85.0   170.5   218.1   318.2   679.0
"
```







2. (3 pts) Discuss the residual plot, the scatter plot of fit vs. OilUsage , and the histogram of log(OilUsage).

Solution:

From the residual plot, the residuals should randomly vary around zero, meaning that our transformation looks appropriate. The histogram, however, is a little left-skewed, which means that we may over-modified a little to the OilUsage. From the scatter plot we can also see that the points now distributed nearly randomly around the regression line, but not as close as the plot generated before transformation. It means that our transformation lower the R^2. The regression line is not close to a 45 degree line any more, because the two functions are not in the same scale.

3. (3 pts) Write out the estimated regression equations for each category of HomeFactor (five equations).

   Solution:

   - HomeFactor = 1:

     `log(OilUsage) = 9.731e-06*DegreeDays+0.2373*NumberPeople+2.993`

   - HomeFactor = 2:

     `log(OilUsage) = 2.003e-03*DegreeDays+0.2373*NumberPeople+2.5999`

   - HomeFactor = 3:

     `log(OilUsage) = 1.918e-03*DegreeDays+0.2373*NumberPeople+3.2552`

   - HomeFactor = 4:

     `log(OilUsage) = 1.675e-03*DegreeDays+0.2373*NumberPeople+3.3743`

   - HomeFactor = 5:

     `log(OilUsage) = 2.062e-03*DegreeDays+0.2373*NumberPeople+3.5016`

4. (2 pts) Provide an economic interpretation of the coefficient of NumberPeople.

   Solution:

   $b_{people} = 0.2373$:

   - When number of people increases by 1, the expected percentage change in oil usage is approximately **23.73%**, while all the other factors remain the same.

5. (2 pts) Estimate the oil consumption of a customer with DegreeDays =380, NumberPeople =4, HomeFactor = 1.

   Solution:

$$
\begin{aligned}
&log(\text{OilUsage}(\text{DegreeDays} = 380, \text{NumberPeople} = 4, \text{HomeFactor} = 1)) \\
&\quad = 9.731 * 10^{-6} * \text{DegreeDays} + 0.2373 * \text{NumberPeople} + 2.993 \\
&\quad = 9.731 * 10^{-6} * 380 + 0.2373 * 4 + 2.993 \\
&\quad = 3.94589778 \\
&\therefore \ \text{OilUsage}(\text{DegreeDays} = 380, \text{NumberPeople} = 4, \text{HomeFactor} = 1) \\
&\quad = e^{3.94589778} = 51.72275293
\end{aligned}
$$

Thus, the oil consumption of a customer with DegreeDays =380, NumberPeople =4, HomeFactor = 1 is **51.72275293**.

6. (2 pts) 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.

   Solution:

   I prefer Model D. Admittedly, Model C has higher R^2, as the logarithmic transformation will decrease the R^2. When we compare the histogram, we see that the logarithmic transformation changes C's right-skewed plot to a nearly normal distributed plot (a little bit left-skewed). Also, D's residual plot shows less trend and not as skewed as C's, and in D's scatter plot, the points are more close to the regression line the those in C's plot (not as obvious as Q-Q plot, which I also checked to verify), meaning that OilUsage are more linear related with fitted Model D than fitted Model C.