c1-worksheet

1. For a given day, the number of heating degrees is 65 minus the mean temperature in degrees Fahrenheit. If the mean is above 65, the number of heating degrees is 0 for that day. **Heating degree days** is the sum of the heating degrees over a certain period (say a month). They are commonly used in calculations relating to the energy consumption required to heat buildings. We will use linear regression to predict kilowatt hour (kWh) energy use from heating degree days (hdd) (for a particular house in Scotland). Here are the data.

Month	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep
hdd	163	228	343	373	301	238	137	84	38	15	14	34
kWh	593	676	1335	1149	1127	892	538	289	172	131	134	134

```
# Write out data in temporary vectors/lists
mon <- c("Oct","Nov","Dec","Jan","Feb","Mar","Apr","May","Jun","Jul","Aug","Sep")
heating_degree_days <- c(163,228,343,373,301,238,137,84,38,15,14,34)
usage <- c(593,676,1335,1149,1127,892,538,289,172,131,134,134)

# Store it in a data frame
heating_data <- data.frame(month = mon, hdd = heating_degree_days, kWh = usage)

# Clean up the temporary variables
rm(heating_degree_days,usage,mon)

# summary(dataframe) gives a summary of each column of the data frame
summary(heating_data)
```

```
##     month                hdd             kWh
##  Length:12          Min.   : 14.0   Min.   : 131.0
##  Class :character   1st Qu.: 37.0   1st Qu.: 162.5
##  Mode  :character   Median :150.0   Median : 565.5
##                     Mean   :164.0   Mean   : 597.5
##                     3rd Qu.:253.8   3rd Qu.: 950.8
##                     Max.   :373.0   Max.   :1335.0
```

   a. Which is the explanatory variable and which is the response?

Heating degree days is the explanatory variable and the kilowatt hours is the response variable

   b. Display a scatterplot and describe any structure you see in the data. Give the equation of the regression line and place it on a scatterplot.

```
# Usage for plot is plot(x,y,xlab = "",ylab = "")
# To access vectors from your data set use
#   heating_data$hdd and heating_data$kWh
# set parameters xlab = "", ylab = "" to have nice labels

# uncomment the lines below and fill in the blanks
 plot(x = heating_data$hdd,
      y = heating_data$kWh,
      xlab = "Heating Degree Days",
      ylab = "Kilowatt Hours")
```

```r
# lm(formula,data = dataframe) # gives a linear model based on data
# An example formula might be y ~ x
# (R uses a tilda instead of an equals sign for formulas)
# You'll have to write it in terms of our variables
# the variable heating.lm will contain your linear model

heating.lm <- lm(kWh ~ hdd,data = heating_data)

summary(heating.lm) # can also accept a model as an input
```

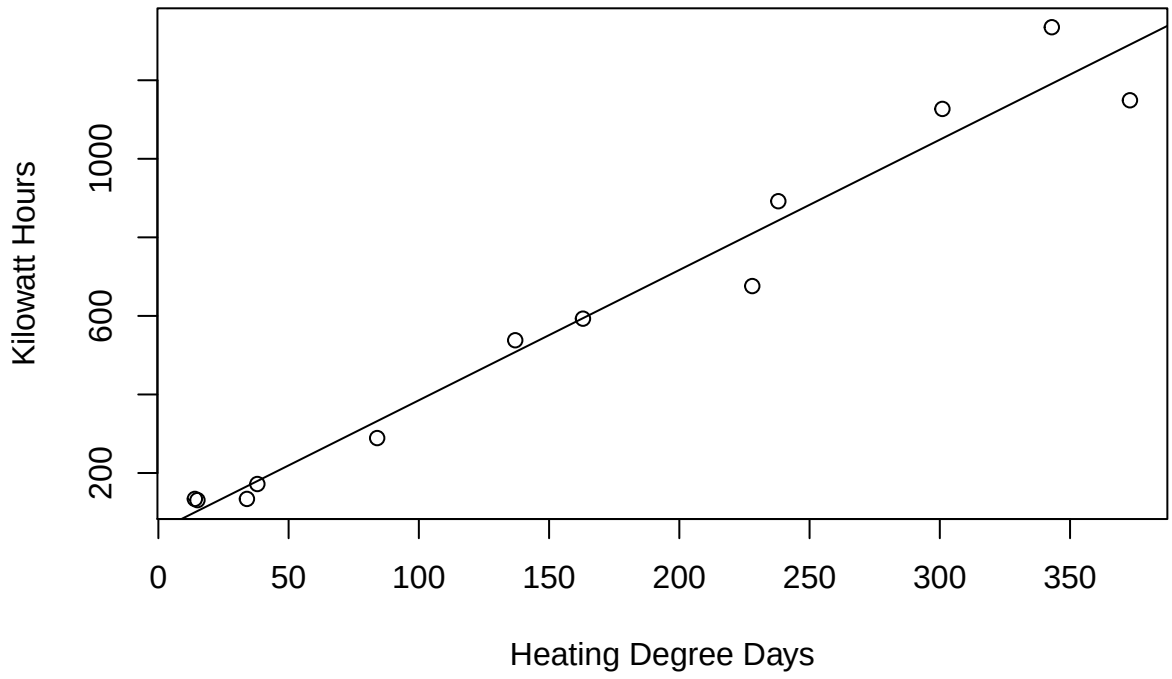
```
##
## Call:
## lm(formula = kWh ~ hdd, data = heating_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -141.76  -35.00   13.28   37.80  143.75
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   53.505     40.467   1.322    0.216
## hdd            3.317      0.196  16.922 1.09e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 85.14 on 10 degrees of freedom
## Multiple R-squared:  0.9663, Adjusted R-squared:  0.9629
## F-statistic: 286.4 on 1 and 10 DF,  p-value: 1.092e-08
```

```r
summary(heating.lm)
```

```
##
## Call:
## lm(formula = kWh ~ hdd, data = heating_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -141.76  -35.00   13.28   37.80  143.75
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   53.505     40.467   1.322    0.216
## hdd            3.317      0.196  16.922 1.09e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 85.14 on 10 degrees of freedom
## Multiple R-squared:  0.9663, Adjusted R-squared:  0.9629
## F-statistic: 286.4 on 1 and 10 DF,  p-value: 1.092e-08
```

```
# abline draws a line of the form y = a + bx
# abline can also accept an model of that form

 abline(heating.lm)
```



c. What is the predicted energy use for a month in which the average temperature is 50°F and no day has an average temperature above 65°F? The R code below shows how to do the calculation in R. Show how to do the calculation by hand either by adding in R code that shows the algebra or by typing it out as text.

```
# predict(linear model, data frame with new place to evaluate)
# you'll have to work out what number to replace XXX with
 predict(heating.lm,newdata = data.frame(hdd = 15*30))
```

```
##        1
## 1546.175
```

```
# put your calculation below either inside the chunk as R code or outside the chunk as text
(kWh_450 = 53.504 + 3.317*450)
```

```
## [1] 1546.154
```

d. What is the predicted average temperature for a month in which 1150 kWh of energy is used? Include details of your calculation.

```
# put your calculation below either inside the chunk as R code or outside the chunk as text
(hdd_pred = (1150-53.504)/3.317)
```

```
## [1] 330.5686
```

e. In a cooler climate, the baseload energy or non-weather-dependent consumption is the amount of energy used when none is devoted to heating. Estimate this using the regression line (no calculation needed if you're clever).

This question is simply asking us to find the value of power used when the days in which people switched on excess heating is 0. Therefore, this means that we need to substitute hdd = 0 which is simply the y-intercept of the graph. We already know that the y-intercept of our predicted line is 53.504.

f. Find the correlation of these two quantitative variables. What does this value for the correlation tell you?

We can see that the correlation is very close to one, meaning that the observations are strongly correlated and that they are almost linearly correlated. In other words, this means that the values of the observations tend to lie above or below the mean value in tandem most of the time.

```
# cor(x,y) gives the correlation between x & y
# access your variables as heating_data$hdd and heating_data$kWh

 cor(heating_data$hdd,heating_data$kWh)
```

```
## [1] 0.9829835
```

g. Which month has the largest negative residual? Use the R code and show how to calculate a single residual by hand. What is the value of that residual and what does it represent?

January has the largest negative residual. This means that the predicted value for the power consumption is much larger compared to the actual value recorded.

We can calculate it by hand using the predicted and actual recorded values of the power consumption for that month.

Residual = Predicted value - Actual value

Predicted power consumption in January(in kWh) = 53.504 + 3.317*373 = 1290.75 Actual power consumption recorded in January = 1149

Residual = 1149 - 1290.75 = -141.75

```
# we'll compute the residuals with the resid function
 heating_data$residual <- resid(heating.lm)
 heating_data
```

```
##    month hdd  kWh    residual
## 1    Oct 163  593   -1.182954
## 2    Nov 228  676 -133.790921
## 3    Dec 343 1335  143.748831
## 4    Jan 373 1149 -141.762538
```

```
## 5    Feb 301 1127   75.064748
## 6    Mar 238  892   49.038623
## 7    Apr 137  538   30.060232
## 8    May  84  289  -43.136349
## 9    Jun  38  172   -7.552250
## 10   Jul  15  131   27.739800
## 11   Aug  14  134   34.056845
## 12   Sep  34  134  -32.284067
```

```
# put your calculation below either inside the chunk as R code or outside the chunk as text
```