MA678 Homework 2

9/10/2020

11.5

Residuals and predictions: The folder Pyth contains outcome y and predictors x1, x2 for 40 data points, with a further 20 points with the predictors but no observed outcome. Save the file to your working directory, then read it into R using read.table().

(a)

Use R to fit a linear regression model predicting y from x1, x2, using the first 40 data points in the file. Summarize the inferences and check the fit of your model.

```
#pyth <- read.table(file = #"/users/zarawaheed/Documents/BostonUniversity/MA678/R/HW_2/pyth.txt", heade
```

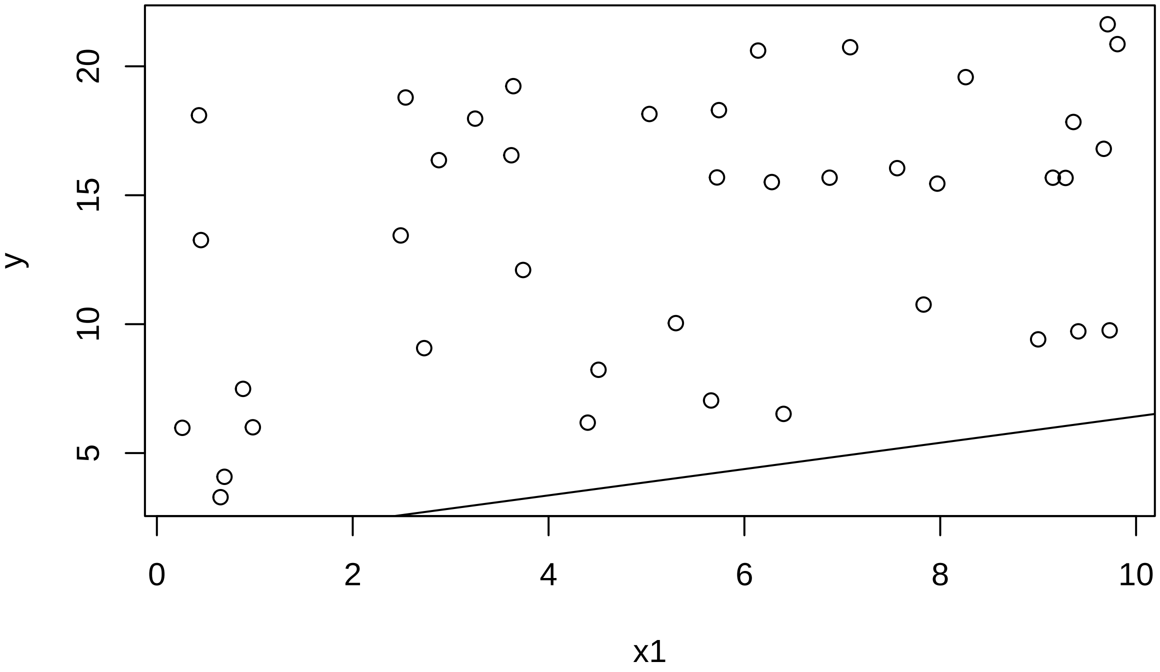
(b)

Display the estimated model graphically as in Figure 10.2

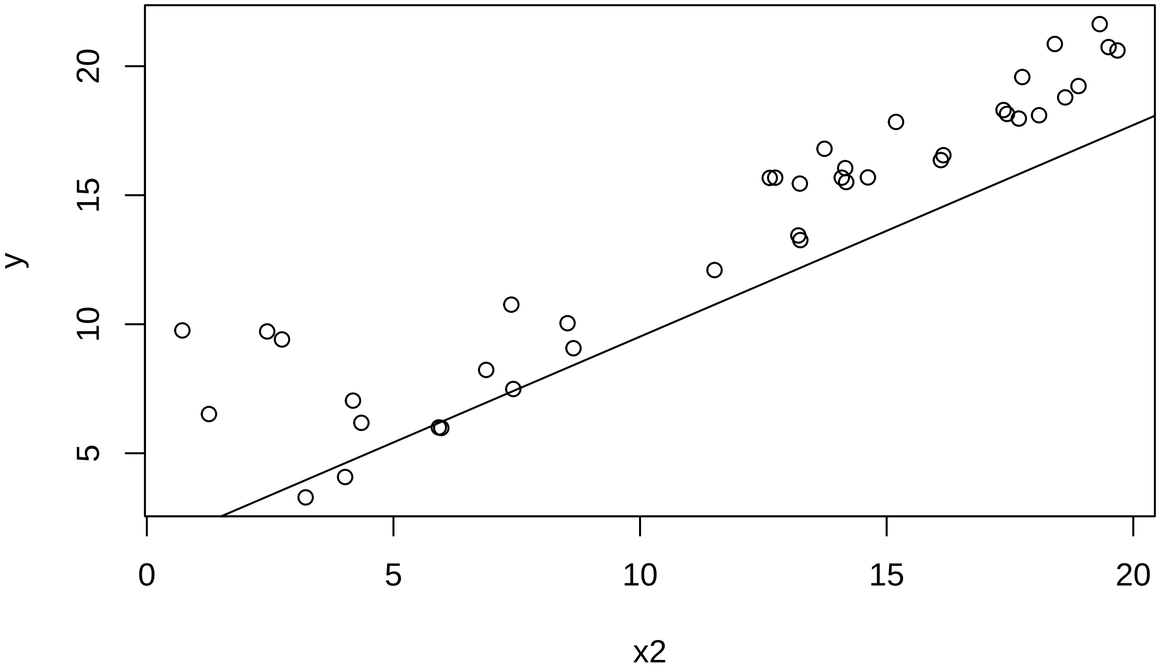
```
pyth <- read.table(url("https://raw.githubusercontent.com/avehtari/ROS-Examples/master/Pyth/pyth.txt"))

colnames(pyth) <- as.character(pyth[1,])
pyth <- pyth[-1,]
rownames(pyth) <- 1:nrow(pyth)

pyth1 <- pyth[1:40,]
pyth2 <- pyth[41:60,]

y <- as.numeric(pyth1$y)
x1 <- as.numeric(pyth1$x1)
x2 <- as.numeric(pyth1$x2)

fit_11.5b <- lm(y ~ x1 + x2)
summary(fit_11.5b)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9585 -0.5865 -0.3356  0.3973  2.8548
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.31513    0.38769   3.392  0.00166 **
## x1           0.51481    0.04590  11.216 1.84e-13 ***
## x2           0.80692    0.02434  33.148  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9 on 37 degrees of freedom
## Multiple R-squared:  0.9724, Adjusted R-squared:  0.9709
## F-statistic: 652.4 on 2 and 37 DF,  p-value: < 2.2e-16
```

```
plot(x1,y)
abline(1.32,0.51)
```



```
plot(x2,y)
abline(1.32,0.82)
```

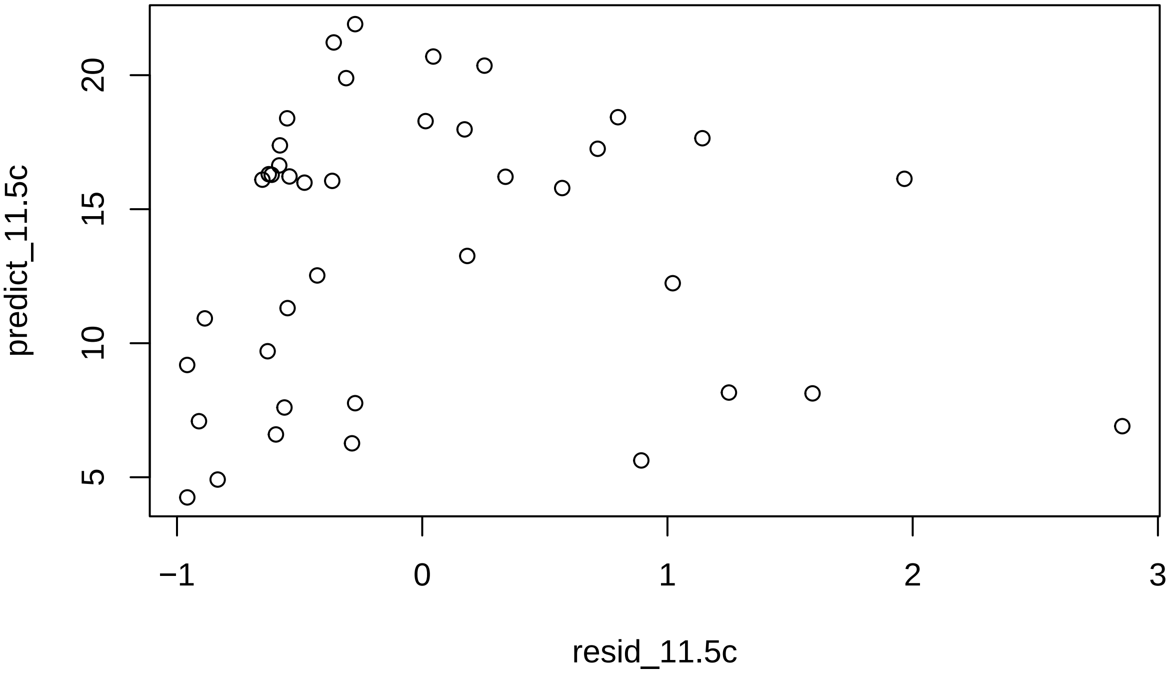


```
# y = 1.32+0.51*x1+0.81*x2+error
```

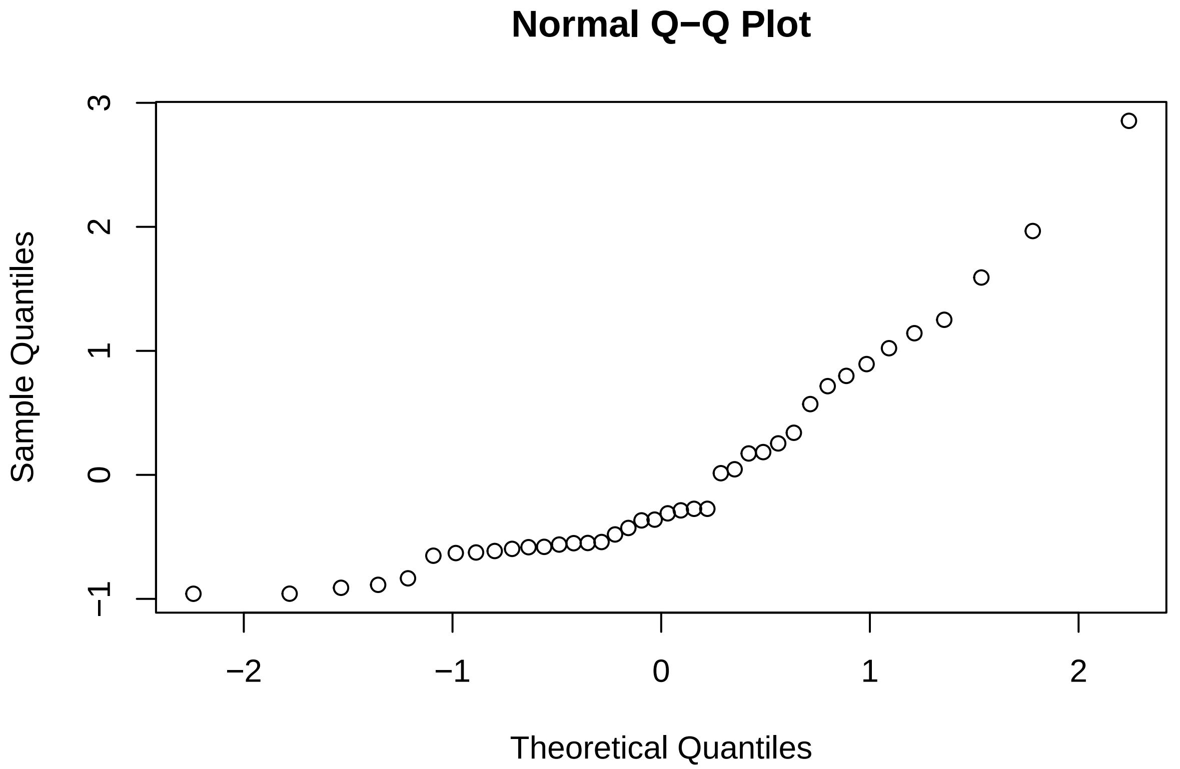
(c)

Make a residual plot for this model. Do the assumptions appear to be met?

```
resid_11.5c <- residuals(fit_11.5b)
predict_11.5c <- predict(fit_11.5b)

plot(resid_11.5c, predict_11.5c)
```



```
qqnorm(resid(fit_11.5b))
```



Homoscedascticity is observed. The normality of residuals is not met according to the QQ plot. The distribution slightly skewed to the right

(d) Make predictions for the remaining 20 data points in the file. How confident do you feel about these predictions?

pred_11.5d <- predict(fit_11.5b, pyth2)

12.5

Logarithmic transformation and regression: Consider the following regression: log(weight)=-3.8+2.1log(height)+error, with errors that have standard deviation 0.25. Weights are in pounds and heights are in inches.

(a)

Fill in the blanks: Approximately 68% of the people will have weights within a factor of -25% and 25% of their predicted values from the regression.

(b)

Using pen and paper, sketch the regression line and scatterplot of log(weight) versus log(height) that make sense and are consistent with the fitted model. Be sure to label the axes of your graph.

(Attached separately)

12.6

Logarithmic transformations: The folder Pollution contains mortality rates and various environmental factors from 60 US metropolitan areas. For this exercise we shall model mortality rate given nitric oxides, sulfur dioxide, and hydrocarbons as inputs. this model is an extreme oversimplication, as it combines all sources of mortality and does not adjust for crucial factors such as age and smoking. We use it to illustrate log transformation in regression.

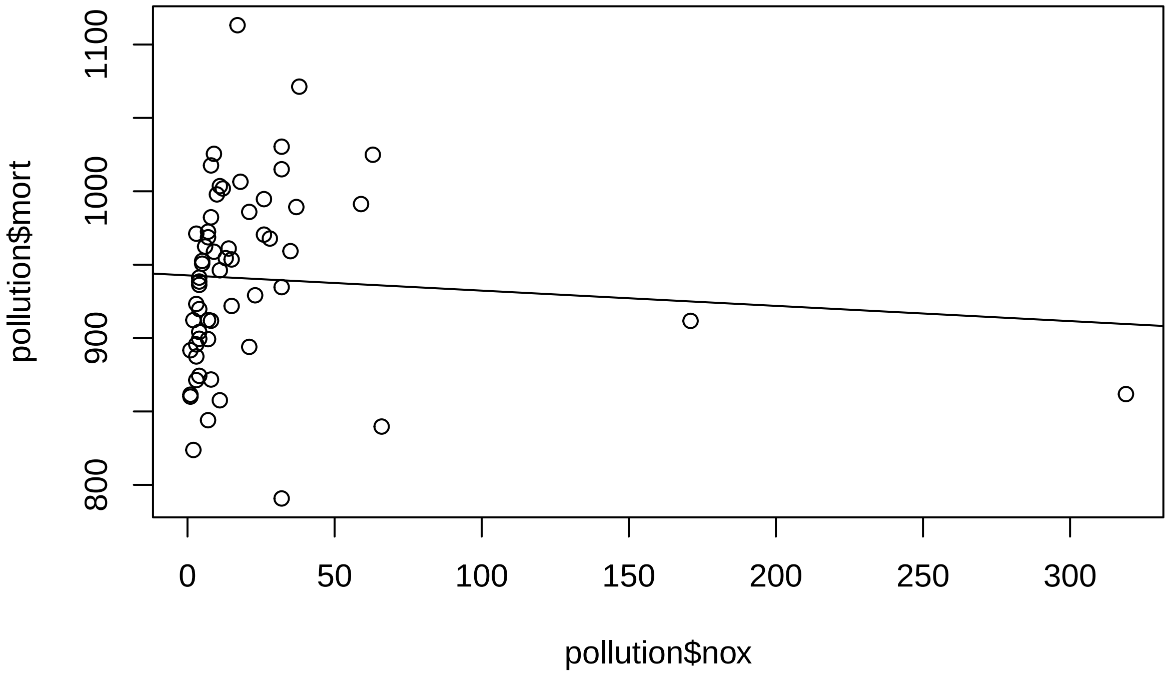
(a)

create a scatterplot of mortality rate versus level of nitric oxides. Do you think linear regression will fit these data well? Fit the regression and evaluate a residual plot from the regression.

```
pollution <- read.csv(file = "/users/zarawaheed/Documents/BostonUniversity/MA678/R/HW_2/pollution.csv",

plot(pollution$nox, pollution$mort)

fit_12.6a <- lm(mort ~ nox, data = pollution)
summary(fit_12.6a)
```

```
##
## Call:
## lm(formula = mort ~ nox, data = pollution)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -148.654  -43.710    1.751   41.663  172.211
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 942.7115     9.0034 104.706   <2e-16 ***
## nox          -0.1039     0.1758  -0.591    0.557
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 62.55 on 58 degrees of freedom
## Multiple R-squared:  0.005987,   Adjusted R-squared:  -0.01115
## F-statistic: 0.3494 on 1 and 58 DF,  p-value: 0.5568
```

```
abline(fit_12.6a)
```



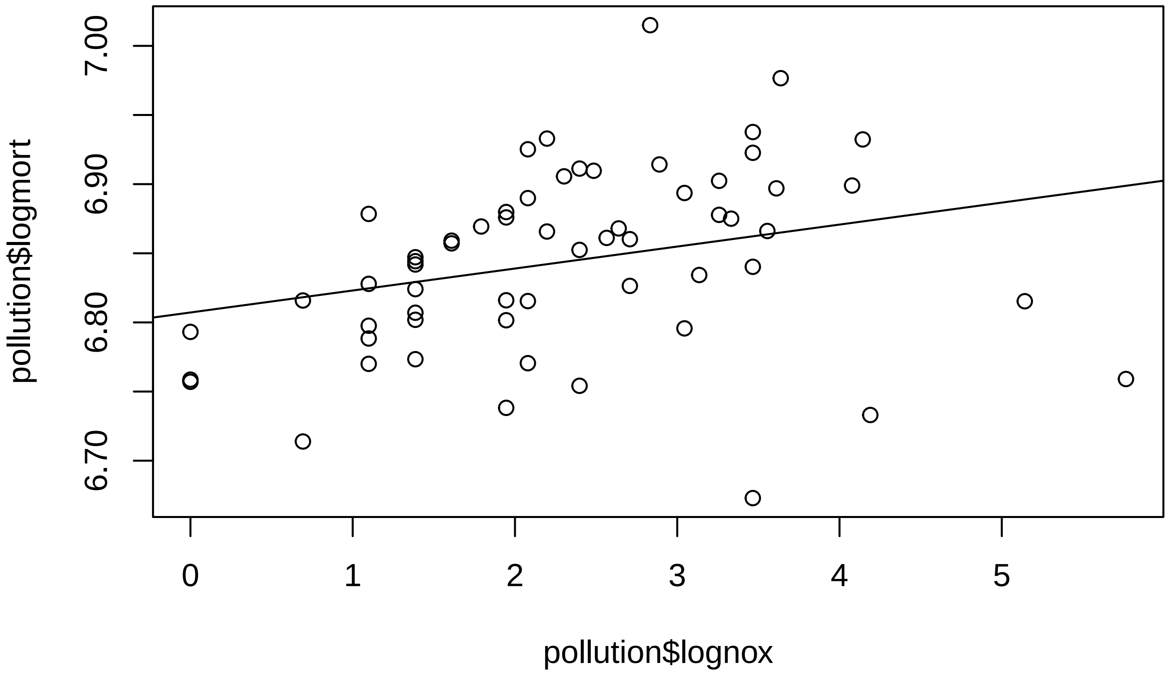
Some outliers in the data are present that could effect accuracy but a regression will still fit.

(b) Find an appropriate transformation that will result in data more appropriate for linear regression. Fit a regression to the transformed data and evaluate the new residual plot.

```
pollution$lognox <- log(pollution$nox)
pollution$logmort <- log(pollution$mort)
plot(pollution$lognox, pollution$logmort)

fit_12.6ba <- lm(logmort ~ lognox, data = pollution)
summary(fit_12.6ba)
```

```
##
## Call:
## lm(formula = logmort ~ lognox, data = pollution)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.18930 -0.02957  0.01132  0.03897  0.16275
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.807175   0.018349 370.975   <2e-16 ***
## lognox      0.015893   0.007048   2.255   0.0279 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06412 on 58 degrees of freedom
## Multiple R-squared:  0.08061,    Adjusted R-squared:  0.06476
## F-statistic: 5.085 on 1 and 58 DF,  p-value: 0.02792
```

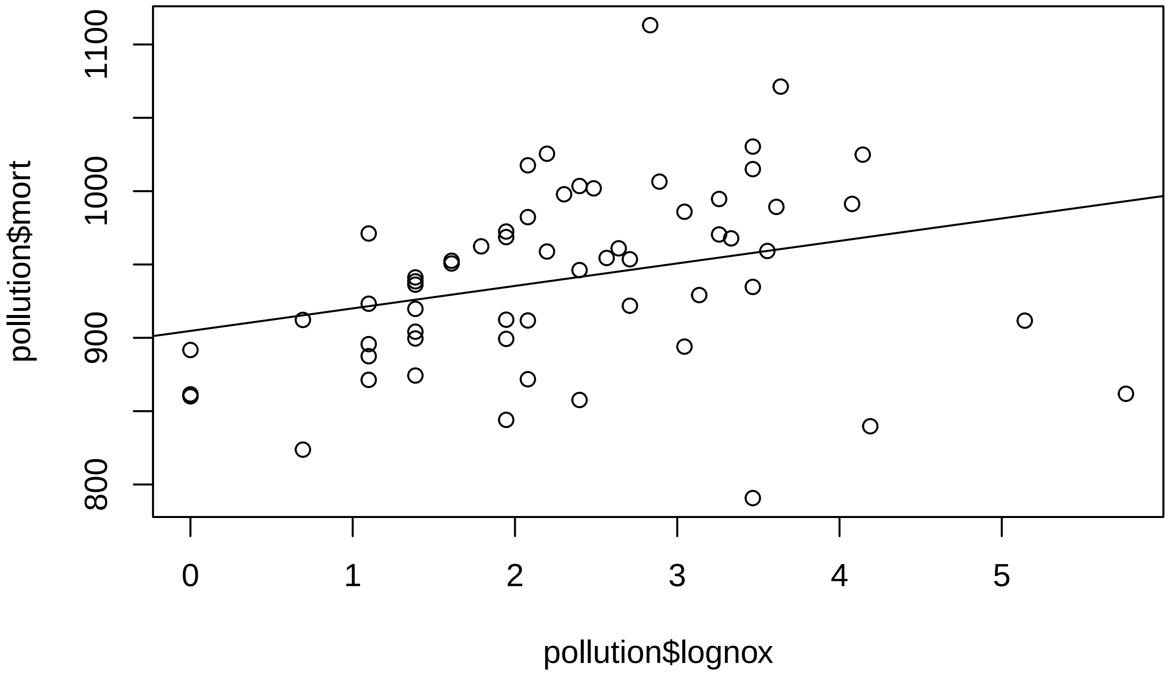
```
abline(fit_12.6ba)
```



```
plot(pollution$lognox, pollution$mort)
fit_12.6bb <- lm(mort ~ lognox, data = pollution)
summary(fit_12.6bb)
```

```
##
## Call:
## lm(formula = mort ~ lognox, data = pollution)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -167.140  -28.368    8.778   35.377  164.983
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  904.724     17.173  52.684   <2e-16 ***
## lognox        15.335      6.596   2.325   0.0236 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 60.01 on 58 degrees of freedom
## Multiple R-squared:  0.08526,    Adjusted R-squared:  0.06949
## F-statistic: 5.406 on 1 and 58 DF,  p-value: 0.02359
```

```
abline(fit_12.6bb)
```



(c)

Interpret the slope coefficient from the model you chose in (b)

The slope coefficient is 15.335, meaning that when the the value of nox increases by a factor of 1, the expected increase in mort is 15.335 units.

(d)

Now fit a model predicting mortality rate using levels of nitric oxides, sulfur dioxide, and hydrocarbons as inputs. Use appropriate transformation when helpful. Plot the fitted regression model and interpret the coefficients.

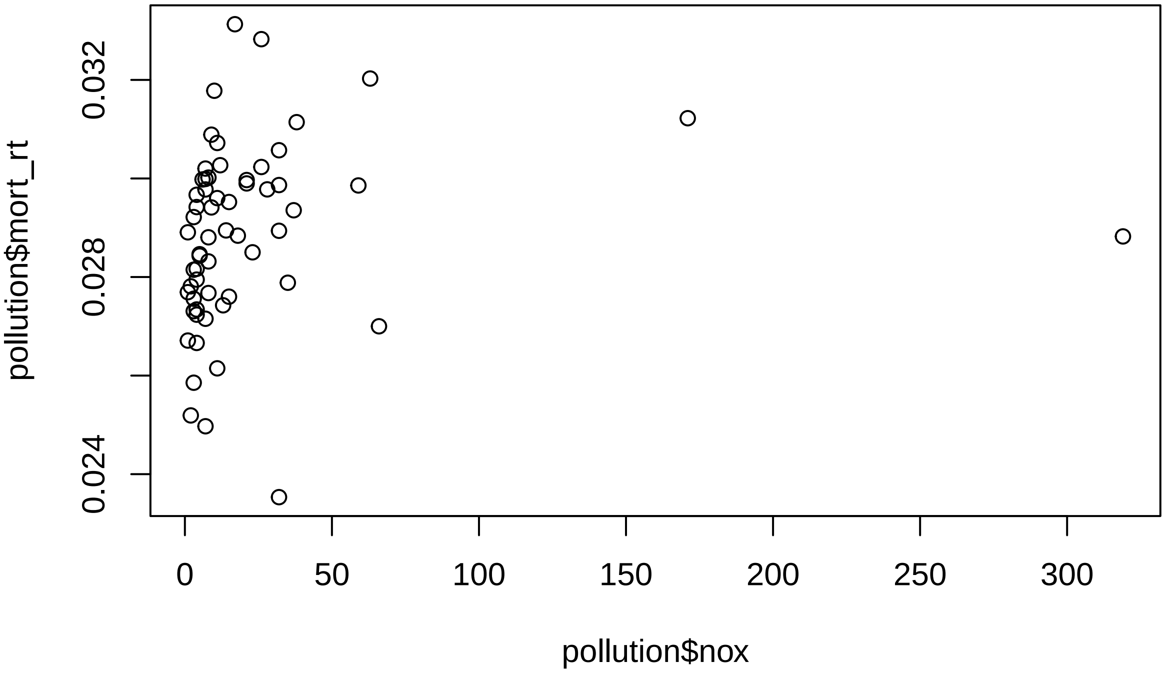
```r
pollution$rpopn <- pollution$popn*10000
pollution$mort_rt <- (pollution$mort/pollution$rpopn)
fit12.6d <- stan_glm(mort ~ nox + so2 + hc, data=pollution, refresh = 0)
fit12.6d
```

```
## stan_glm
##  family:       gaussian [identity]
##  formula:      mort ~ nox + so2 + hc
##  observations: 60
##  predictors:   4
## ------
##             Median MAD_SD
## (Intercept) 923.4    9.0
## nox           2.2    1.2
## so2           0.3    0.2
## hc           -1.3    0.6
##
## Auxiliary parameter(s):
##       Median MAD_SD
## sigma 52.1    5.0
##
```

```
## ------
## * For help interpreting the printed output see ?print.stanreg
## * For info on the priors used see ?prior_summary.stanreg
```

```
plot(pollution$nox, pollution$mort_rt)
abline(fit12.6d)
```

```
## Warning in abline(fit12.6d): only using the first two of 4 regression
## coefficients
```



We can see from the coefficients that if no emissions were considered, the mortalities would would be 923.4. With each unit of nitric oxide emmitted, the mortality score increases by 2.3. Similarly, with each unit of sulfur dioxide, mortalities increase by 0.3. With each unit of hydrocarbon, however, the score decreases by 1.3.

(e)

Cross validate: fit the model you chose above to the first half of the data and then predict for the second half. You used all the data to construct the model in (d), so this is not really cross validation, but it gives a sense of how the steps of cross validation can be implemented.

```
pollution_half <- head(pollution, 23)
fit12.6e <- stan_glm(mort ~ nox + so2 + hc, data=pollution_half, refresh = 0)
loo_1 <- loo(fit12.6e)
```

```
## Warning: Found 3 observation(s) with a pareto_k > 0.7. We recommend calling 'loo' again with argumen
```

```
loo_1
```

```
##
## Computed from 4000 by 23 log-likelihood matrix
##
##          Estimate  SE
## elpd_loo   -125.3 2.3
## p_loo         4.7 1.3
## looic       250.6 4.6
## ------
## Monte Carlo SE of elpd_loo is NA.
##
## Pareto k diagnostic values:
##                          Count Pct.    Min. n_eff
## (-Inf, 0.5]   (good)     19    82.6%   1451
##  (0.5, 0.7]   (ok)        1     4.3%   679
##    (0.7, 1]   (bad)       3    13.0%   27
##    (1, Inf)   (very bad)  0     0.0%   <NA>
## See help('pareto-k-diagnostic') for details.
```

```
kfold_1 <- kfold(fit12.6e, K=10)
```

```
## Fitting model 1 out of 10

## Fitting model 2 out of 10

## Fitting model 3 out of 10

## Fitting model 4 out of 10

## Fitting model 5 out of 10

## Fitting model 6 out of 10

## Fitting model 7 out of 10

## Fitting model 8 out of 10

## Fitting model 9 out of 10

## Fitting model 10 out of 10
```

```
kfold_1
```

```
##
## Based on 10-fold cross-validation
##
##            Estimate  SE
## elpd_kfold   -126.3 2.4
## p_kfold          NA  NA
## kfoldic       252.5 4.8
```

12.7

Cross validation comparison of models with different transformations of outcomes: when we compare models with transformed continuous outcomes, we must take into account how the nonlinear transformation warps the continuous outcomes. Follow the procedure used to compare models for the mesquite bushes example on page 202.

(a)

Compare models for earnings and for log(earnings) given height and sex as shown in page 84 and 192. Use earnk and log(earnk) as outcomes.

(b)

Compare models from other exercises in this chapter.

12.8

Log-log transformations: Suppose that, for a certain population of animals, we can predict log weight from log height as follows:

- An animal that is 50 centimeters tall is predicted to weigh 10 kg.
- Every increase of 1% in height corresponds to a predicted increase of 2% in weight.
- The weights of approximately 95% of the animals fall within a factor of 1.1 of predicted values.

(a)

Give the equation of the regression line and the residual standard deviation of the regression.

log(weight) = 10 + 1.05log(height)

(b)

Suppose the standard deviation of log weights is 20% in this population. What, then, is the R^2 of the regression model described here?

The R^2 will be the 1-0.04, so it will be 0.96.

12.9

Linear and logarithmic transformations: For a study of congressional elections, you would like a measure of the relative amount of money raised by each of the two major-party candidates in each district. Suppose that you know the amount of money raised by each candidate; label these dollar values Di and Ri. You would like to combine these into a single variable that can be included as an input variable into a model predicting vote share for the Democrats. Discuss the advantages and disadvantages of the following measures:

(a)

The simple difference, $D_i - R_i$ This will just give us the difference between the amount of money raised by the two candidates. This is useful because it gives us a general idea about which candidate is in the lead and by how much. However it does not give us much to compare the value to on its own. We will need more info. This might cause issues if there is correlation between D and R.

(b)

The ratio, D_i/R_i
This will give us a good idea about who is in the lead and by what ratio but similar to the difference function, it is not of much use on it's own.

(c)

The difference on the logarithmic scale, $log\ D_i - log\ R_i$
This will help us minimize the skewness of the data while also giving us an idea of the differnece between the two candidates. This would be a useful variable to include in the model.

(d)

The relative proportion, $D_i/(D_i + R_i)$. This will give us the proportion of one variable over another. This would be useful to determine how much our candidate needs to lead so it would be a useful variable to include. The values will be more normal and all candidates will be on a standard scale.

12.11

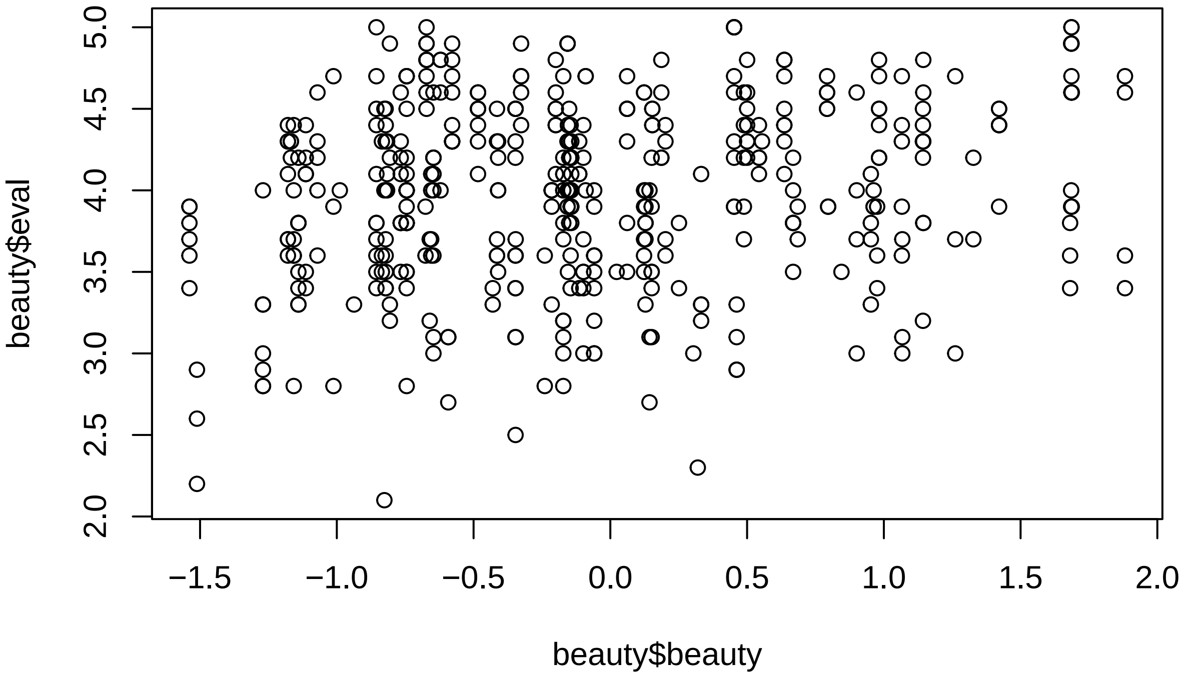
Elasticity: An economist runs a regression examining the relations between the average price of cigarettes, P, and the quantity purchased, Q, across a large sample of counties in the United States, assuming the functional form, $logQ = \alpha + \beta logP$. Suppose the estimate for β is 0.3. Interpret this coefficient.

0.3 is the coefficient for logP meaning that with every 1 unit increase in logP which would be 10 increase in P (logP=1 means P=10), the log of quantity purchased would increase by 0.3. That means that quantity purchased would increase by 2 (logQ=0.3 means Q=1.995 rounded to n2).

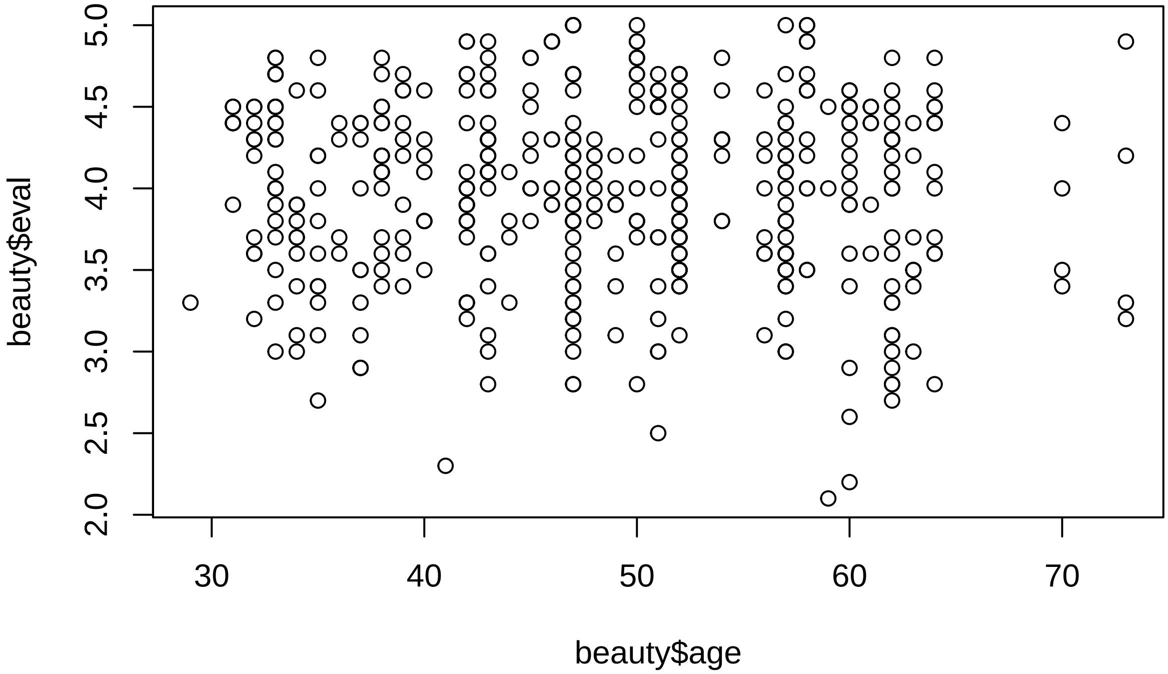
12.13

Building regression models: Return to the teaching evaluations data from Exercise 10.6. Fit regression models predicting evaluations given many of the inputs in the dataset. Consider interactions, combinations of predictors, and transformations, as appropriate. Consider several models, discuss in detail the final model that you choose, and also explain why you chose it rather than the others you had considered.

```
beauty <- read.csv(file = "/users/zarawaheed/Documents/BostonUniversity/MA678/R/Data/beautyy.csv")

plot(beauty$beauty, beauty$eval )
```



```
plot(beauty$age, beauty$eval )
```



```
fit_12.3a <- lm(eval ~ beauty + female *minority*nonenglish*lower, data = beauty)

fit_12.3b <- lm(eval ~ beauty*female*nonenglish*lower + age*female*nonenglish*lower, data = beauty)

summary(fit_12.3a)
```

```
##
## Call:
## lm(formula = eval ~ beauty + female * minority * nonenglish *
##     lower, data = beauty)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.82983 -0.32433  0.05491  0.36804  1.07494
##
## Coefficients: (2 not defined because of singularities)
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    4.06512    0.04280  94.969  < 2e-16 ***
## beauty                         0.16377    0.03179   5.151 3.89e-07 ***
## female                        -0.13113    0.06457  -2.031 0.042858 *
## minority                      -0.24085    0.18982  -1.269 0.205163
## nonenglish                    -0.48557    0.23476  -2.068 0.039175 *
## lower                          0.09880    0.07013   1.409 0.159578
## female:minority                0.11689    0.25694   0.455 0.649367
## female:nonenglish              0.48962    0.30898   1.585 0.113752
## minority:nonenglish            0.41012    0.34227   1.198 0.231458
## female:lower                  -0.04827    0.11979  -0.403 0.687168
## minority:lower                 0.96697    0.25483   3.795 0.000168 ***
## nonenglish:lower              -0.20795    0.56973  -0.365 0.715293
## female:minority:nonenglish    -0.95162    0.50162  -1.897 0.058460 .
## female:minority:lower         -1.15489    0.33694  -3.428 0.000665 ***
## female:nonenglish:lower        1.11838    0.83496   1.339 0.181108
## minority:nonenglish:lower           NA         NA      NA       NA
## female:minority:nonenglish:lower       NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5161 on 448 degrees of freedom
## Multiple R-squared:  0.161,  Adjusted R-squared:  0.1348
## F-statistic: 6.141 on 14 and 448 DF,  p-value: 3.242e-11
```

I chose fit_12.3a because it compares beauty score and age and also includes the interaction between the two to get a clearer picture of how the variables affect each other.

12.14

Prediction from a fitted regression: Consider one of the fitted models for mesquite leaves, for example fit_4, in Section 12.6. Suppose you wish to use this model to make inferences about the average mesquite yield in a new set of trees whose predictors are in data frame called new_trees. Give R code to obtain an estimate and standard error for this population average. You do not need to make the prediction; just give the code.

```
mesquite <- read.table(file = “/Users/zarawaheed/Documents/BostonUniversity/MA678/R/Data/mesquite.dat.txt”)
```

```
colnames(mesquite) <- as.character(mesquite[1,]) mesquite <- mesquite[-1,] rownames(mesquite) <- 1:nrow(mesquite)
```

```
fit_12.14 <- stan_glm(formula = log(weight) ~ log(canopy_volume) + log(canopy_area) + log(canopy_shape) + log(total_height) +log(density) + group, data = mesquite)
```

```
summary(fit_12.14)
```

```
y<- posterior_predict(fit_12.14, newdata = new_trees)
```