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.

Answer: The residual standard deviation is small(0.9) and the fraction of variance (R-Squared=0.97) is large, so this is a good fit.

```r
pyth <- read.table("pyth.txt", header = TRUE)
head(pyth)
```

```
##       y   x1    x2
## 1 15.68 6.87 14.09
## 2  6.18 4.40  4.35
## 3 18.10 0.43 18.09
## 4  9.07 2.73  8.65
## 5 17.97 3.25 17.68
## 6 10.04 5.30  8.53
```

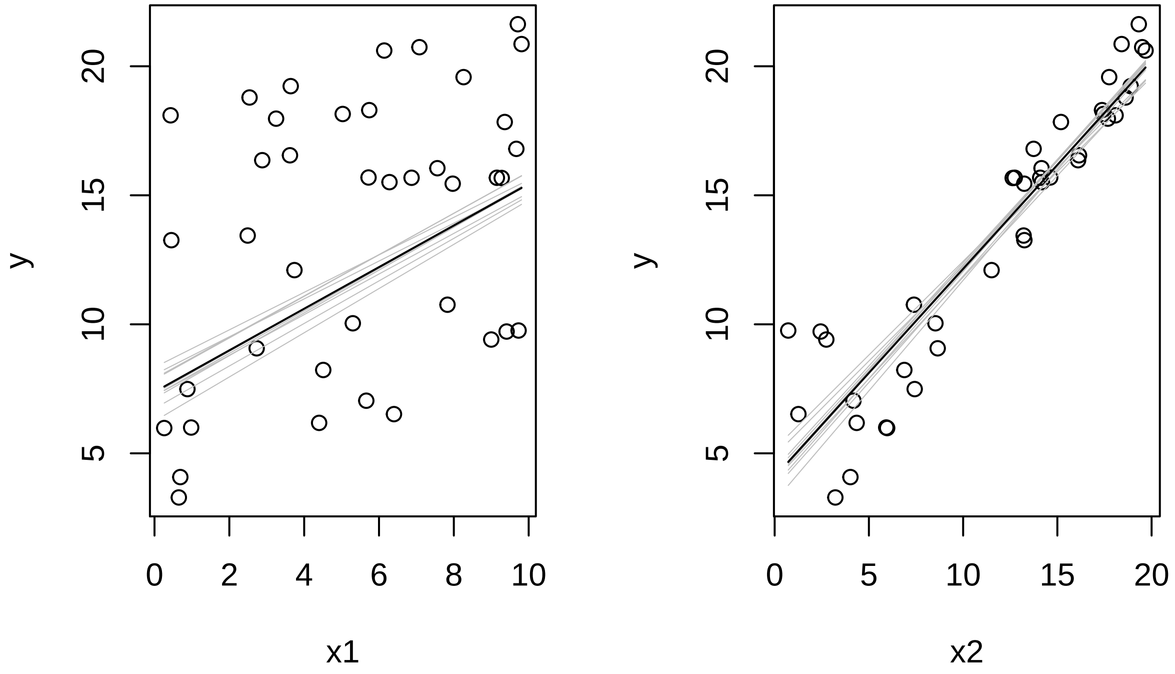
```r
p.fit <- pyth[1:40,]
p.predict <- pyth[41:60,]
fit1 <- stan_glm(y ~ x1 + x2, data = p.fit, refresh =0)
print(fit1)
```

```
## stan_glm
##  family:       gaussian [identity]
##  formula:      y ~ x1 + x2
##  observations: 40
##  predictors:   3
## ------
##             Median MAD_SD
## (Intercept) 1.3    0.4
## x1          0.5    0.0
## x2          0.8    0.0
##
## Auxiliary parameter(s):
##       Median MAD_SD
## sigma 0.9    0.1
##
## ------
## * For help interpreting the printed output see ?print.stanreg
## * For info on the priors used see ?prior_summary.stanreg
```

(b)

Display the estimated model graphically as in Figure 10.2

```
sim_1 <- as.matrix(fit1)

# sim_b <- sim(lm)
n_sim.b <- nrow(sim_1)
par(mfrow = c(1,2))

# Plot of fit against x1 with x2 held to it's average value
plot(p.fit$x1, p.fit$y, xlab = "x1", ylab = "y")
x2_bar <- mean(p.fit$x2)
sim_dis <- sample(n_sim.b, 10)

for (i in sim_dis) {
  curve(cbind(1, x2_bar, x)  %*% sim_1[i,1:3], lwd=0.5, col="gray", add=TRUE)
}
curve(cbind(1, x2_bar, x) %*% coef(fit1), col="black", add=TRUE)

plot(p.fit$x2, p.fit$y, xlab = "x2", ylab= "y")
x1_bar = mean(p.fit$x1)
for (i in sim_dis) {
  curve(cbind(1, x1_bar, x) %*% sim_1[i, 1:3], lwd=0.5, col="gray", add=TRUE)
}
curve(cbind(1, x1_bar, x) %*% coef(fit1), col="black", add=TRUE)
```

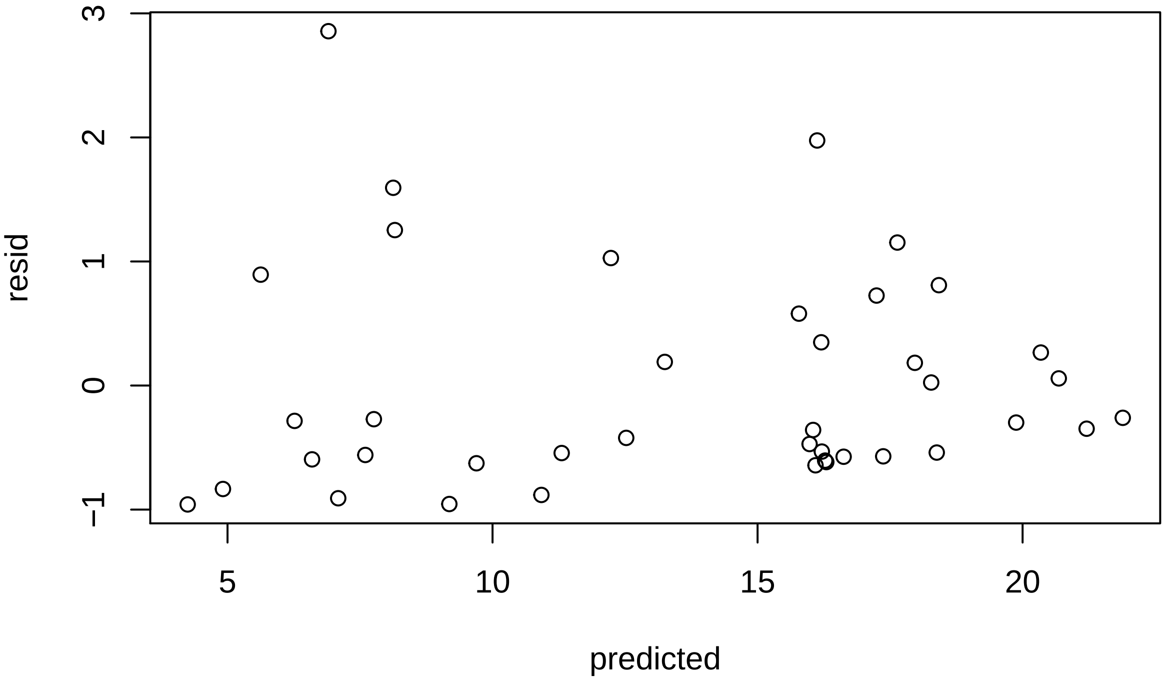


(c)

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

Answer: The residual plot appears odd with a lot of mass in the lower end.

```
sim1 <- as.matrix(fit1)
predicted <- predict(fit1)
resid <- p.fit$y - predicted
plot(predicted, resid)
```



(d)

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

Answer: The residual plot may indicate issues (specifically the linearity assumption) so not quite sure of the goodness of the predictions.

```
y_predict <- predict(fit1, p.predict, interval="prediction", level=0.95)
print(y_predict)
```

```
##         1         2         3         4         5         6         7         8
## 16.212431  7.088627 16.124554  9.696506 17.244698 10.921861 20.682423  6.903800
##         9        10        11        12        13        14        15        16
##  9.185097  5.626108 16.048154 15.981477 20.343981 19.878476  8.125977 15.780438
##        17        18        19        20        21        22        23        24
## 18.275705 12.232461 12.522043 17.966893 17.370663 16.201436 17.637347 16.296878
##        25        26        27        28        29        30        31        32
##  4.913988 16.092979 13.249665 21.207999 16.623834  6.595102  4.248388  8.156940
##        33        34        35        36        37        38        39        40
## 11.304235  6.265019 18.420925 16.275064  7.599753 21.890268 18.380289  7.761136
```

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 _________ and _________ of their predicted values from the regression. Answer: exp(0.25) = 1.28

(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.

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)

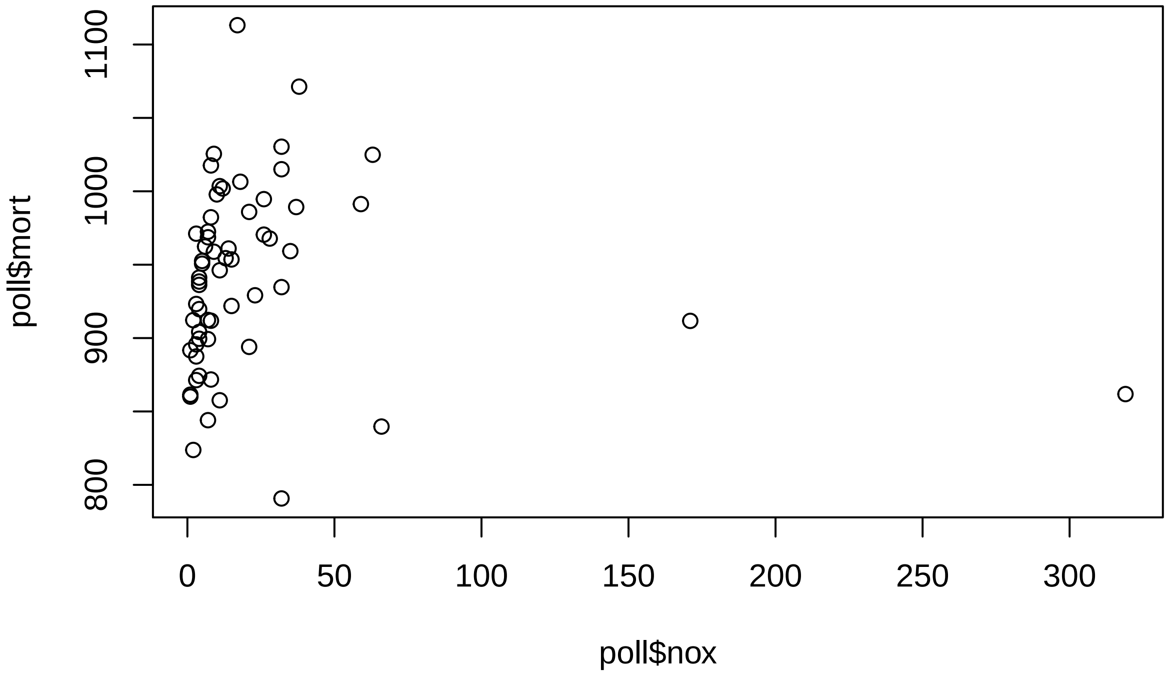
create a scatterplot of mortality rate versus level of nitric oxides. Do you think linear regression will fit these data well? Fit the regression and evaluate a residual plot from the regression.

Answer: Linear regression should be a good start to evaluate this relationship, but the residual standard deviation is large.

```r
poll <- read.csv("pollution.csv", header = TRUE)
head(poll)
```

```
##   prec jant jult ovr65 popn educ hous dens nonw wwdrk poor hc nox so2 humid
## 1   36   27   71   8.1 3.34 11.4 81.5 3243  8.8  42.6 11.7 21  15  59    59
## 2   35   23   72  11.1 3.14 11.0 78.8 4281  3.5  50.7 14.4  8  10  39    57
## 3   44   29   74  10.4 3.21  9.8 81.6 4260  0.8  39.4 12.4  6   6  33    54
## 4   47   45   79   6.5 3.41 11.1 77.5 3125 27.1  50.2 20.6 18   8  24    56
## 5   43   35   77   7.6 3.44  9.6 84.6 6441 24.4  43.7 14.3 43  38 206    55
## 6   53   45   80   7.7 3.45 10.2 66.8 3325 38.5  43.1 25.5 30  32  72    54
##       mort
## 1  921.870
## 2  997.875
## 3  962.354
## 4  982.291
## 5 1071.289
## 6 1030.380
```

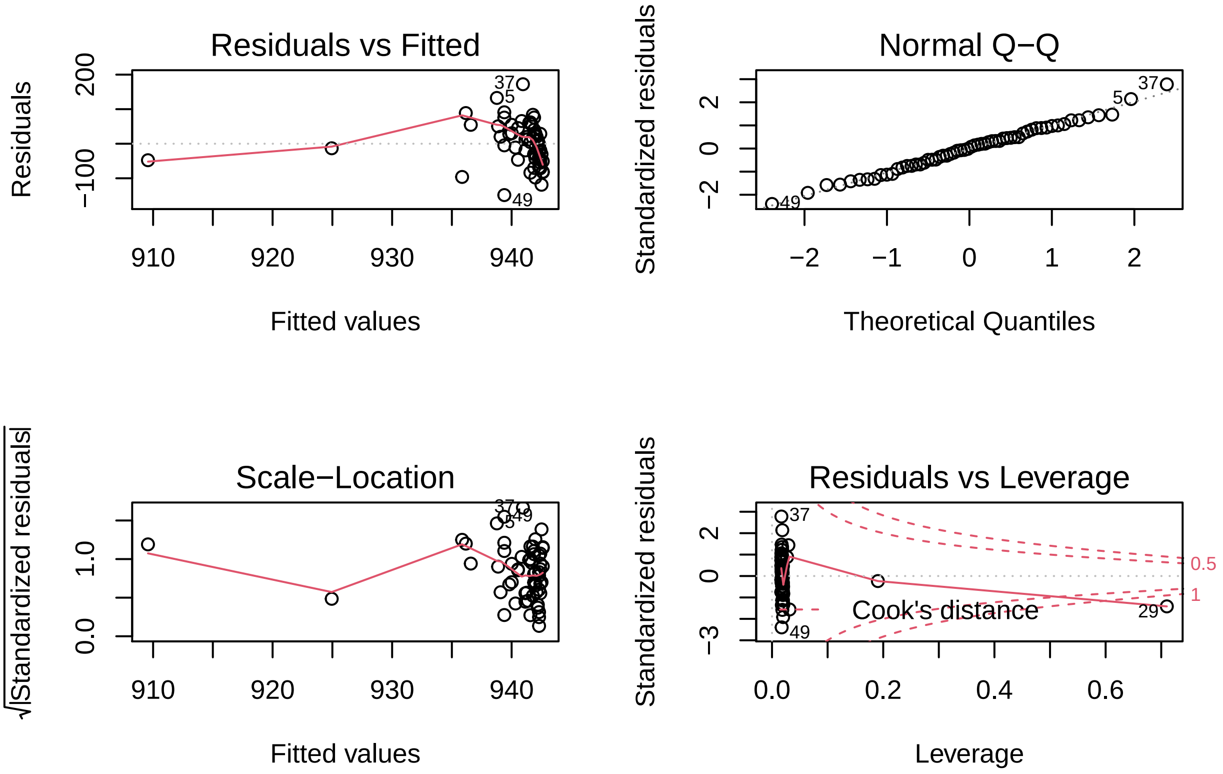
```r
plot(poll$nox, poll$mort)
```



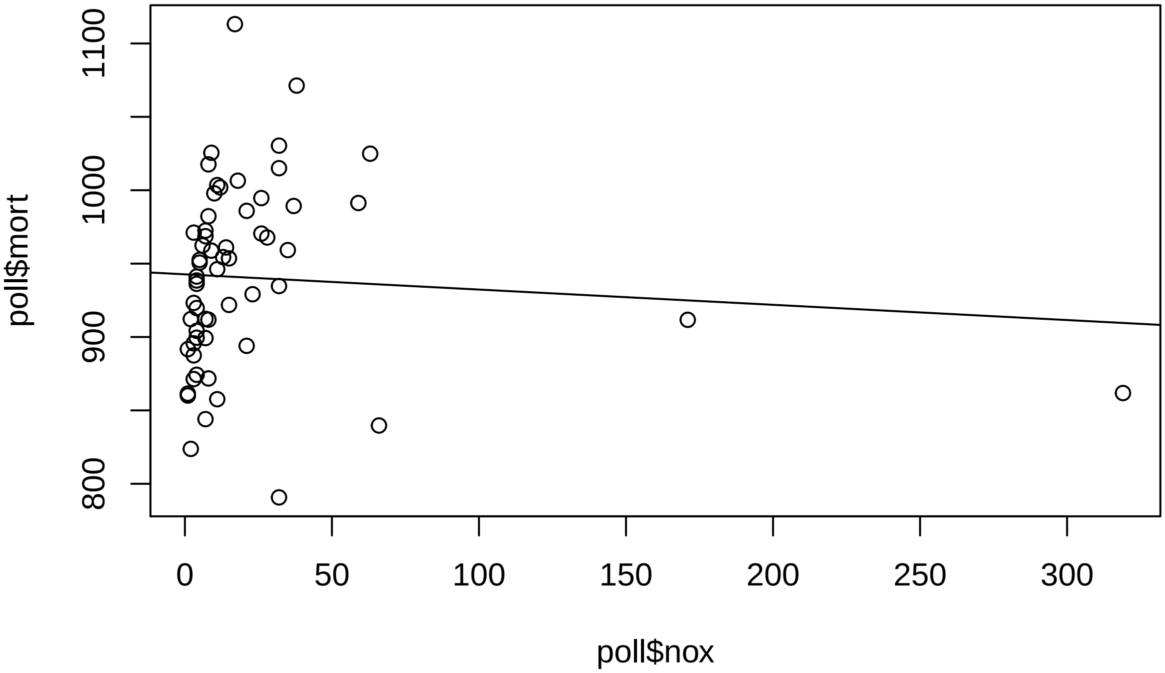
```
fit2 <- lm(mort ~ nox, data = poll)
display(fit2)
```

```
## lm(formula = mort ~ nox, data = poll)
##             coef.est coef.se
## (Intercept) 942.71     9.00
## nox          -0.10     0.18
## ---
## n = 60, k = 2
## residual sd = 62.55, R-Squared = 0.01
```

```
par(mfrow = c(2,2))
plot(fit2)
```



```r
par(mfrow = c(1,1))
plot(poll$nox, poll$mort)
abline(fit2)
```



(b)

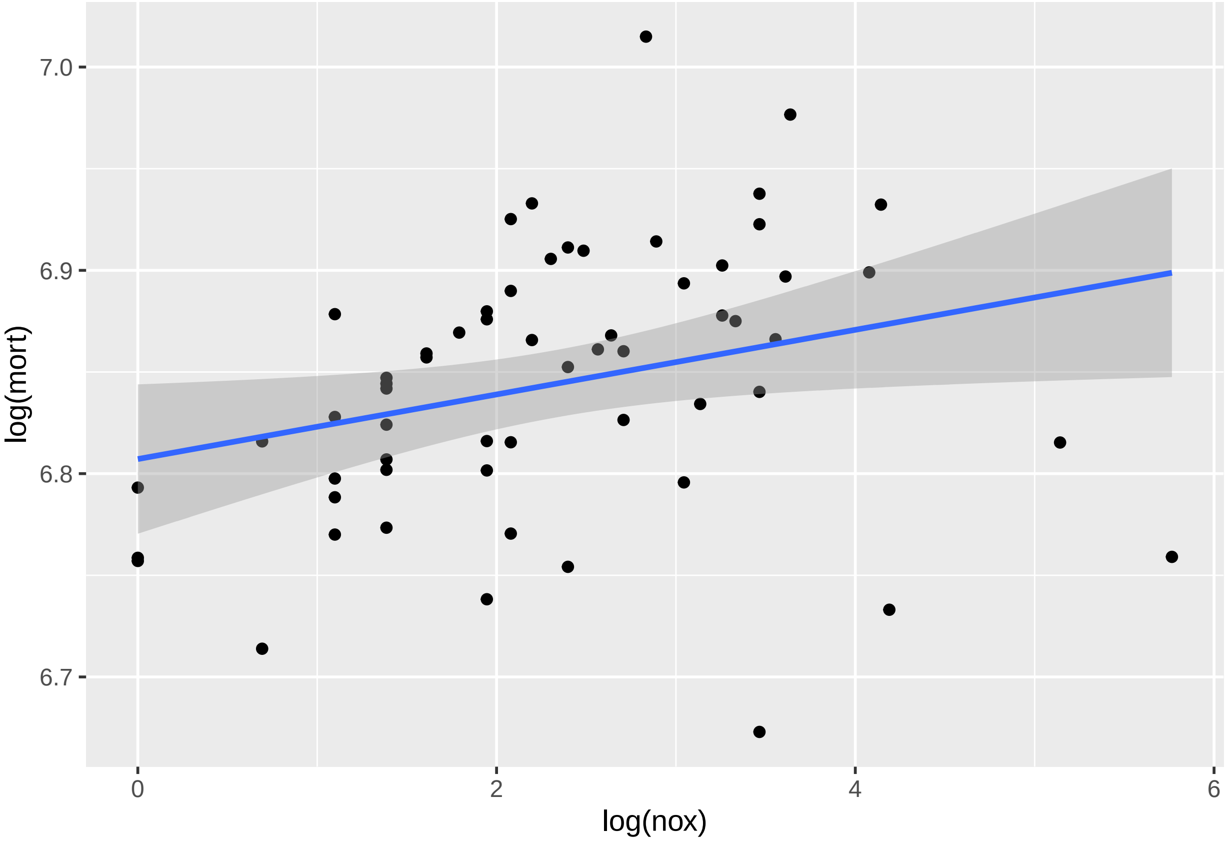
Find an appropriate reansformation that will result in data more appropriate for linear regression. Fit a regression to the transformed data and evaluate the new residual plot.

Answer: We can observe that using log transformation significantly improved out fitted model's residual sd and R-Squared.

```
fit3 <- lm(log(mort) ~ log(nox), data = poll)
display(fit3)
```

```
## lm(formula = log(mort) ~ log(nox), data = poll)
##             coef.est coef.se
## (Intercept) 6.81     0.02
## log(nox)    0.02     0.01
## ---
## n = 60, k = 2
## residual sd = 0.06, R-Squared = 0.08
```

```
ggplot(data = poll, aes(x=log(nox), y= log(mort))) +
  geom_point() +
  stat_smooth(method = "lm", formula = y~ x, se = TRUE)
```



(c)

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

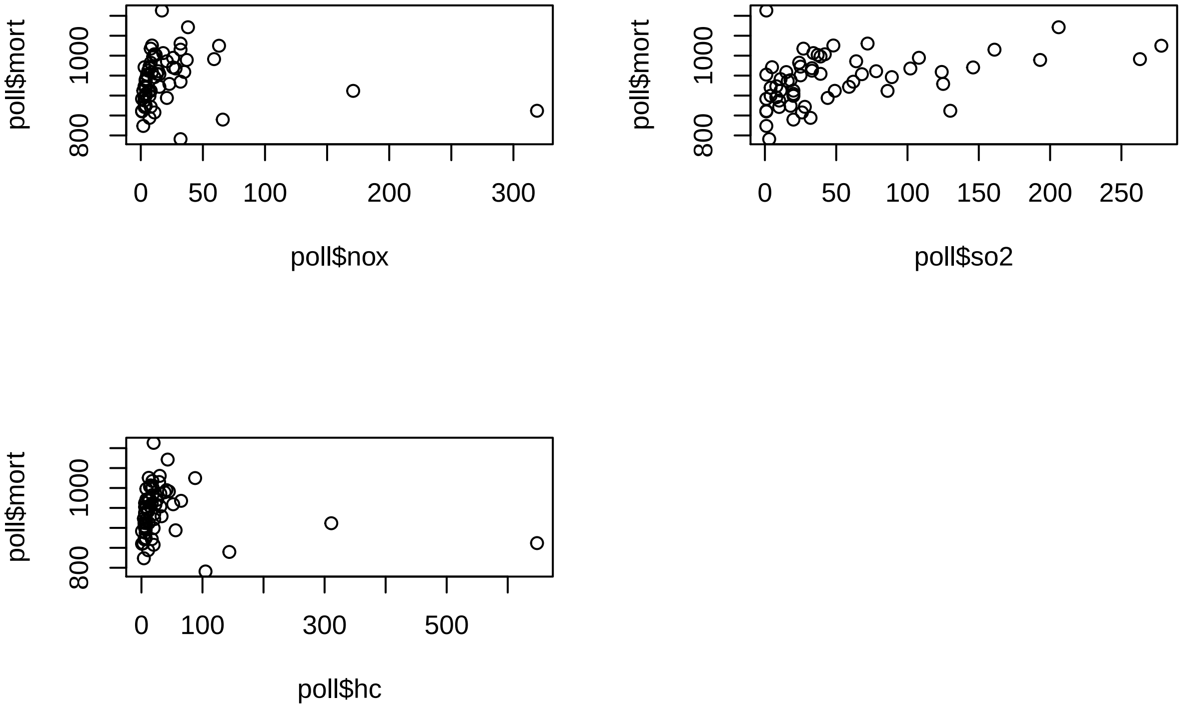
(d)

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.

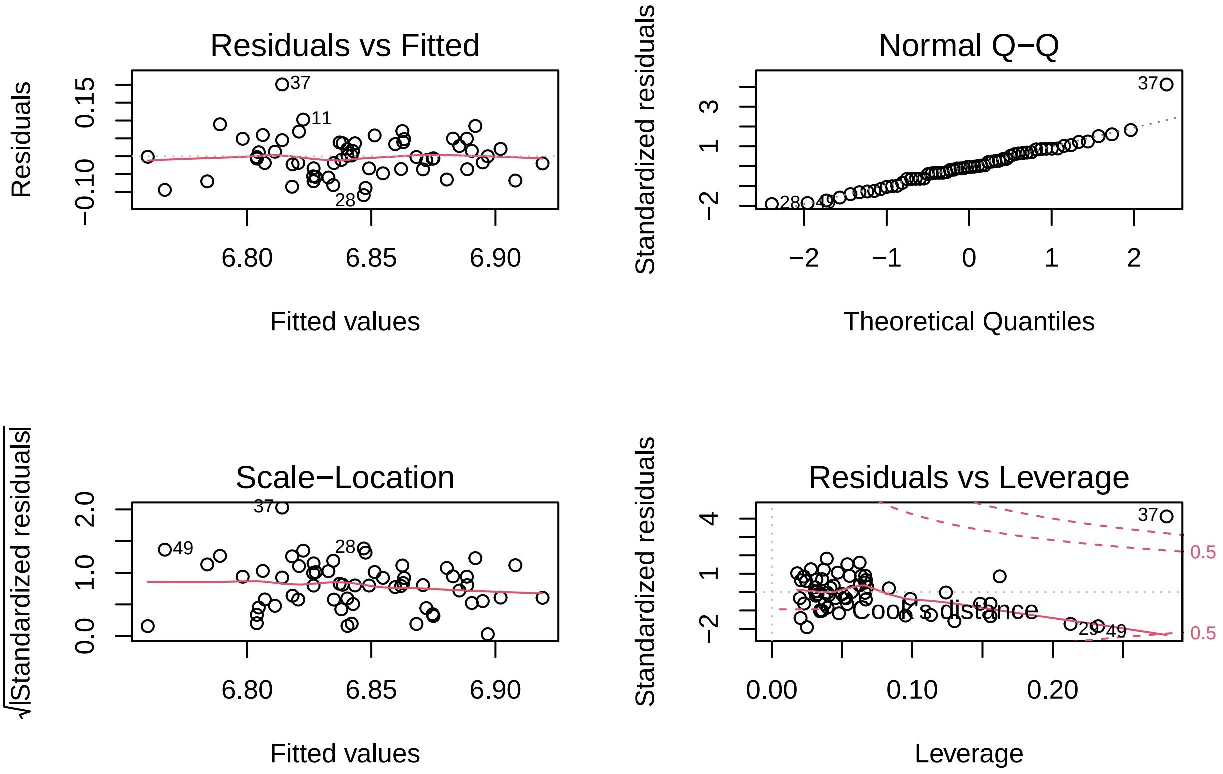
```
par(mfrow = c(2,2))
plot(poll$nox, poll$mort)
plot(poll$so2 , poll$mort)
plot(poll$hc, poll$mort)
lm0 <- lm(log(mort) ~ log(nox)+ log(so2) + log(hc), data = poll)
display(lm0)
```

```
## lm(formula = log(mort) ~ log(nox) + log(so2) + log(hc), data = poll)
##             coef.est coef.se
## (Intercept)  6.83     0.02
## log(nox)     0.06     0.02
## log(so2)     0.01     0.01
## log(hc)     -0.06     0.02
## ---
## n = 60, k = 4
## residual sd = 0.06, R-Squared = 0.29
```

```
par(mfrow=c(2,2))
```



```
plot(lm0)
```



(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.

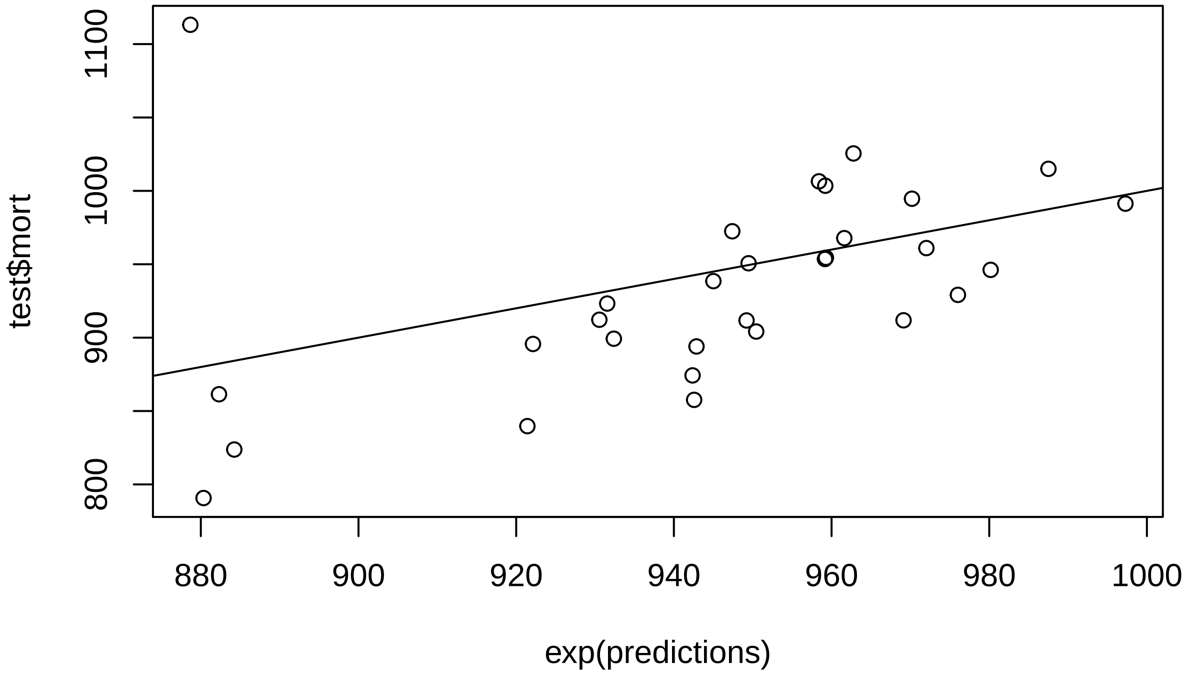
```
train <- poll[1:30, ]
test <- poll[31:60, ]

m1 <- lm(log(mort) ~ log(nox)+ log(so2) + log(hc), data = train)
display(m1)
```

```
## lm(formula = log(mort) ~ log(nox) + log(so2) + log(hc), data = train)
##             coef.est coef.se
## (Intercept)  6.80     0.03
## log(nox)     0.01     0.03
## log(so2)     0.02     0.01
## log(hc)     -0.02     0.03
## ---
## n = 30, k = 4
## residual sd = 0.06, R-Squared = 0.24
```

```
#predict
predictions <- predict(m1, test)
cbind(predictions = exp(predictions), observed = test$mort)
```

```
##    predictions observed
## 31    958.4017 1006.490
## 32    882.3047  861.439
## 33    976.0271  929.150
## 34    942.5715  857.622
## 35    972.0281  961.009
## 36    931.5476  923.234
## 37    878.6695 1113.156
## 38    970.2004  994.648
## 39    987.5079 1015.023
## 40    997.2779  991.290
## 41    942.8677  893.991
## 42    945.0112  938.500
## 43    980.1823  946.185
## 44    962.7805 1025.502
## 45    942.3703  874.281
## 46    959.1658  953.560
## 47    921.4196  839.709
## 48    949.2163  911.701
## 49    880.3459  790.733
## 50    932.3869  899.264
## 51    950.4423  904.155
## 52    949.4753  950.672
## 53    947.4063  972.464
## 54    930.5496  912.202
## 55    961.6173  967.803
## 56    884.2419  823.764
## 57    959.2060 1003.502
## 58    922.1316  895.696
## 59    969.1311  911.817
## 60    959.2870  954.442
```

```
plot(exp(predictions), test$mort)
abline(a=0, b =1)
```



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.
Answer: $log(weight) = 0.02 * log(height) + intercept$ and adding the first condition,we have: $log(10) = 0.02 * log(50) + intercept, sointercept = 0.966$ Thus the equation is:

$$log(weight) = 0.966 + 0.02 * log(height)$$

And the residual standard deviation is $1.1/2 = 0.55$

(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?
Answer: $R^2 = (1 - (\frac{\sigma^2}{sd^2}))$, taking in the values for σ and sd, we have: $R^2 = 1 - \frac{20\%^2}{0.55^2} = 0.132$

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, $D_i - R_i$
Advantage: Straightforward, easy to interpret. Disadvantage: Does not show the absolute amount of money raised, only the relative difference. ### (b) The ratio, D_i/R_i
Advantage: Simple relationship. Disadvantage: Hard to interpret relative increase. ### (c) The difference on the logarithmic scale, $log\ D_i - log\ R_i$

Advantage: Easy to interpret on the percentage change. Disadvantage: Not accounting the absolute amount of money. ### (d) The relative proportion, $D_i/(D_i + R_i)$. Advantage: Shows relative size between party donations and each party as a percentage of the whole amount. Disadvantage: Hard to interpret on the relative increase. ## 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.

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.

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.