HW1 George W Nakhla

Package installs:

```
#install.packages("MPV")
library("MPV")
```

```
## Loading required package: lattice
```

```
## Loading required package: KernSmooth
```

```
## KernSmooth 2.23 loaded
## Copyright M. P. Wand 1997-2009
```

```
## Loading required package: randomForest
```

```
## randomForest 4.7-1.1
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```
library("knitr")
```

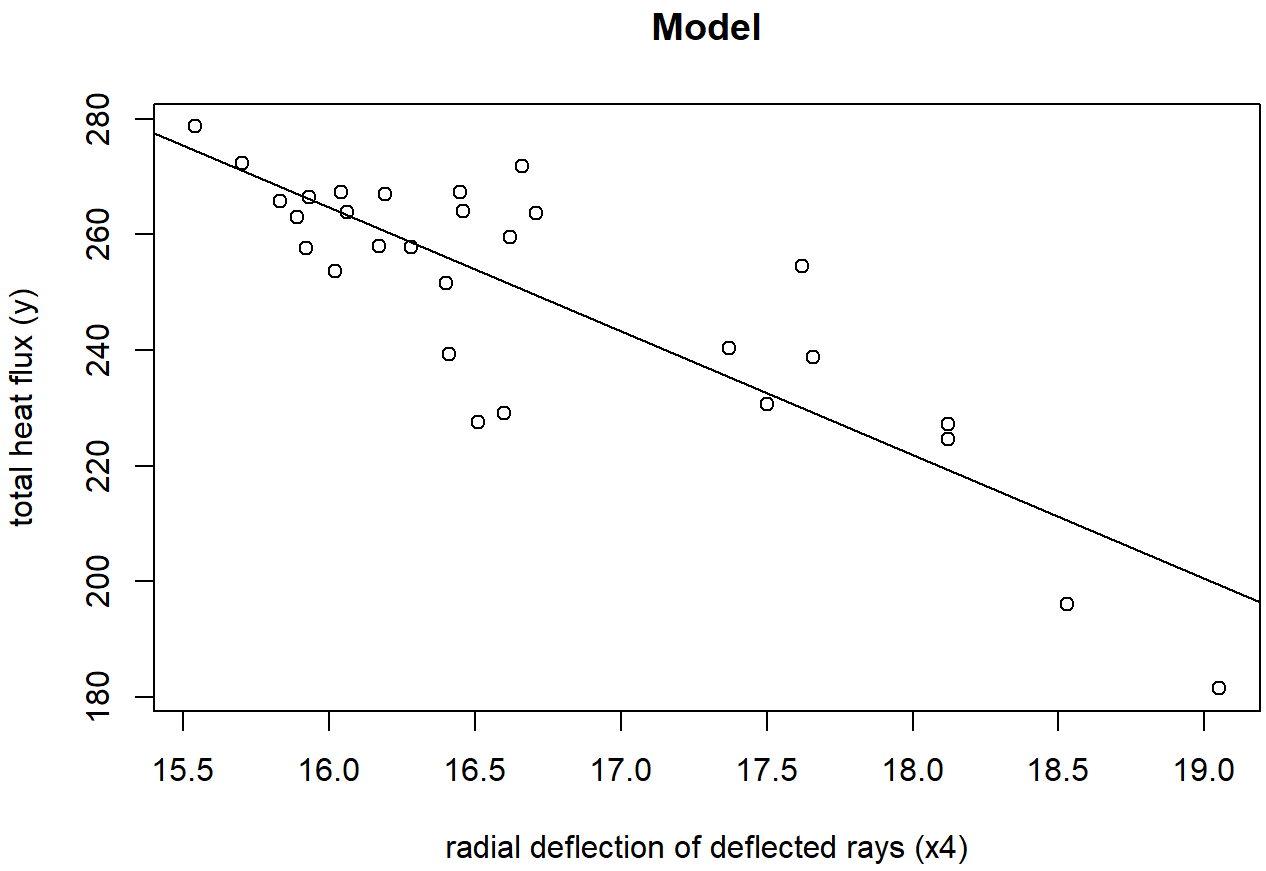
Q1: Question 2.3. Table B.2 presents data collected during a solar energy project at Georgia Tech

a) Fit a simple linear regression model relating total heat flux y (kilowatts) to the radial deflection of the deflected rays x4 (milliradians).

```
solarData <- table.b2
fit <- lm( y ~ x4, data = solarData)
lm( y ~ x4, data = solarData)
```

```
##
## Call:
## lm(formula = y ~ x4, data = solarData)
##
## Coefficients:
## (Intercept)           x4
##       607.1        -21.4
```

```
plot(solarData$x4, solarData$y,
     xlab = "radial deflection of deflected rays (x4)",
     ylab = "total heat flux (y)",
     main = "Model")
abline(fit)
```



So we have the equation y = 607.1 -21.4(x4) as plotted above. This is a negative relationship, and describes that for each one unit increase in x4, we can expect a 21.4 unit decrease of y. The 607.1 can be interpreted as a radial deflection of 0, we can expect a heat flux to be 607.1

b) Construct the analysis-of-variance table and test for significance of regression.

```
# here, we want to conduct a hypothesis test for our regression coefficient. Our H0 (null hypoth
esis) is that there is no relationship between x4 and y (B1 is 0). The alternative hypothesis is
that B1 is non 0. In this case, the previous graph more strongly suggests that B1 is less than
0, or an inverse relationship.
summary(fit)
```

```
##
## Call:
## lm(formula = y ~ x4, data = solarData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -26.2487  -4.5029   0.5202   7.9093  24.5080
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  607.103     42.906  14.150 5.24e-14 ***
## x4           -21.402      2.565  -8.343 5.94e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.33 on 27 degrees of freedom
## Multiple R-squared:  0.7205, Adjusted R-squared:  0.7102
## F-statistic: 69.61 on 1 and 27 DF,  p-value: 5.935e-09
```

```
anova(fit)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## x4         1 10578.7   10579  69.609 5.935e-09 ***
## Residuals 27  4103.2     152
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

72.05% of the variance in y is explained by x4
The p value of the F statistic is extremely low (5.935e-09), indicating that the overall model is significant and that the slope (B1) is not zero. We reject the initial hypothesis in favor of the alternative.

c) Find a 99% CI on the slope.

```
conf_interval <- confint(fit, level = 0.99)
print(conf_interval)
```

```
##                0.5 %    99.5 %
## (Intercept) 488.22411 725.98242
## x4          -28.50995 -14.29497
```

After running this and seeing the results, we can say that we are 99% certain that the true slope of the regression (beta1) falls between:
(-28.50995, -14.29497)

d) Calculate R^2

```
#previously printed the summary, but this will show only r^2
summary(fit)$r.squared
```

```
## [1] 0.7205242
```

Again, we can say that 72.05% of the variance in y is explained by x4 based on this r^2 value

e) Find a 95% CI on the mean heat flux when the radial deflection is 16.5 milliradians.

```
# Define the new data point
new_data <- data.frame(x4 = 16.5)

# Predict the mean heat flux and obtain confidence intervals
prediction <- predict(fit, newdata = new_data, interval = "confidence", level = 0.95)

# Print the prediction and confidence intervals
print(prediction)
```

```
##        fit      lwr      upr
## 1 253.9627 249.1468 258.7787
```

We see that based on the model, we get an expected heat flux of 253.9627 kilowatts. We are 95% certain that the true expected heat flux would be between:
(249.1468, 258.7787)

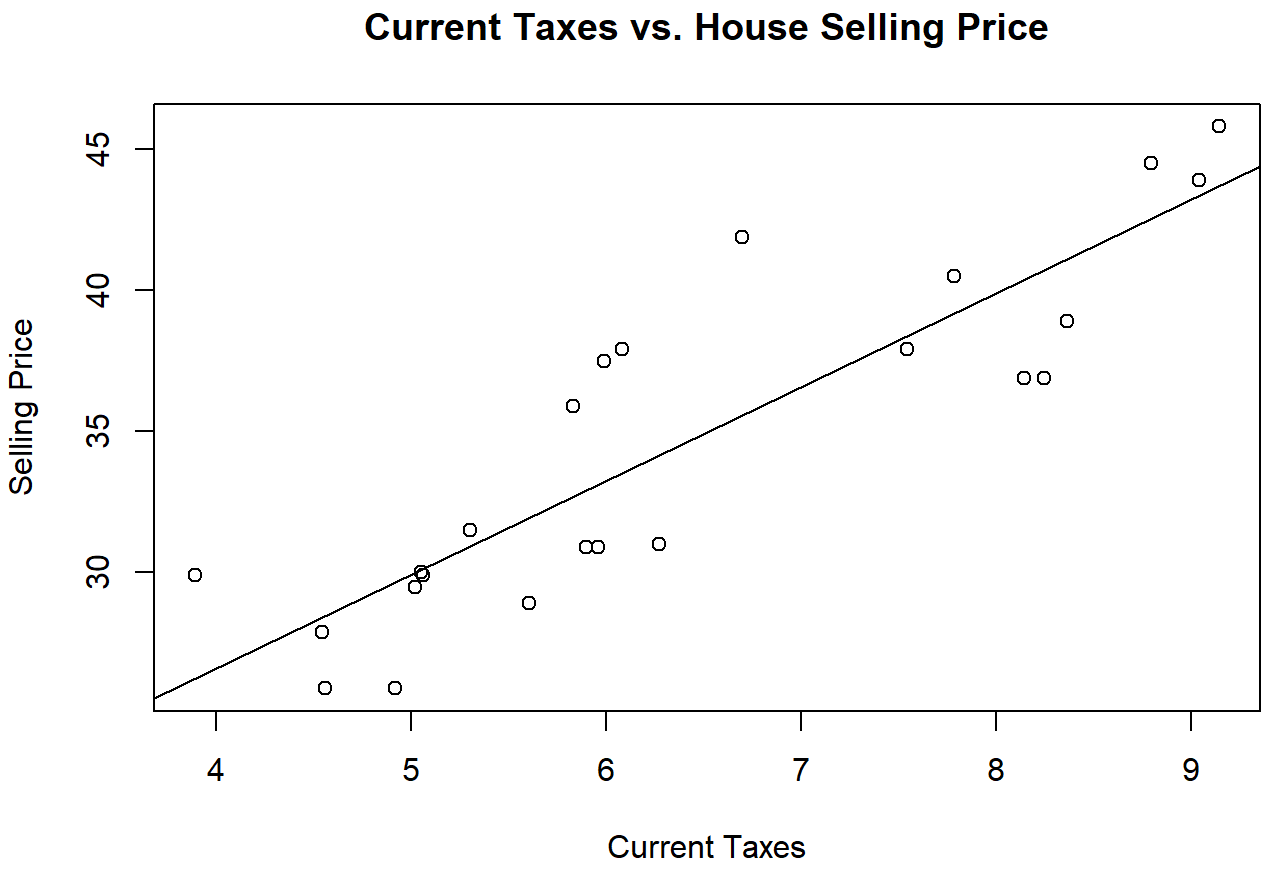
Q2: Question 2.6. Table B.4 presents data for 27 houses sold in Erie, Pennsylvania.

a) Fit a simple linear regression model relating selling price of the house to the current taxes (x1).

```
houseData <- table.b4
lm(y ~ x1, data = houseData)
```

```
##
## Call:
## lm(formula = y ~ x1, data = houseData)
##
## Coefficients:
## (Intercept)           x1
##      13.320        3.324
```

```
homeFit <- lm(y ~ x1, data = houseData)
plot(houseData$x1,houseData$y,
     xlab = "Current Taxes",
     ylab = "Selling Price",
     main = "Current Taxes vs. House Selling Price")
abline(homeFit)
```



Based on the regression, we see that the equation of the model plotted above is: y = 13.320 + 3.324(x1). This suggests that for each 1 unit increase in x1, we can expect a 3.324 increase in y. The interpretation of 13.320 is that at a tax cost of 0 (which in this case is unrealistic), we can expect a house selling price of 13.320.

b) Test for significance of regression.

```
# here, we want to conduct a hypothesis test for our regression coefficient. Our H0 (null hypoth
esis) is that there is no relationship between x1 and y (B1 is 0). The alternative hypothesis is
that B1 is non 0. In this case, the previous graph more strongly suggests that B1 is greater tha
n 0, or a positive relationship.
summary(homeFit)
```

```
##
## Call:
## lm(formula = y ~ x1, data = houseData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8343 -2.3157 -0.3669  1.9787  6.3168
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  13.3202     2.5717   5.179 3.42e-05 ***
## x1            3.3244     0.3903   8.518 2.05e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.961 on 22 degrees of freedom
## Multiple R-squared:  0.7673, Adjusted R-squared:  0.7568
## F-statistic: 72.56 on 1 and 22 DF,  p-value: 2.051e-08
```

At a p-value of 2.051e-08, we can confidently reject the null hypothesis in favor of the alternate. This is evidence that there is a non-zero relationship between tax cost and house selling price.

c) Calculate R^2

```
summary(homeFit)$r.squared
```

```
## [1] 0.7673344
```

We can say that 76.73344% of the variance in y is explained by x2 based on this r^2 value

d) Find a 95% CI on the slope.

```
confint(homeFit, level = 0.95)
```

```
##                2.5 %    97.5 %
## (Intercept) 7.986755 18.653604
## x1          2.514988  4.133754
```

We are 95% certain that the true slope of B1 lies between: (2.514988, 4.133754)

e) Find a 95% CI on the mean selling price of a house for which the current taxes are $750.

```
# NOTE: x1 is documented to be represented in thousands, so we will use 750/1000 = .75
new_data <- data.frame(x1 = .75)
predict(homeFit, newdata = new_data, interval = "confidence", level = 0.95)
```

```
##        fit      lwr      upr
## 1 15.81346 11.06792 20.55899
```

NOTE: All in thousands The prediction model expresses that we can expect a house selling price of 15.81346 for a tax cost of .75. We are 95% certain that the true selling price for a house with a .75 tax cost is: (11.06792, 20.55899)

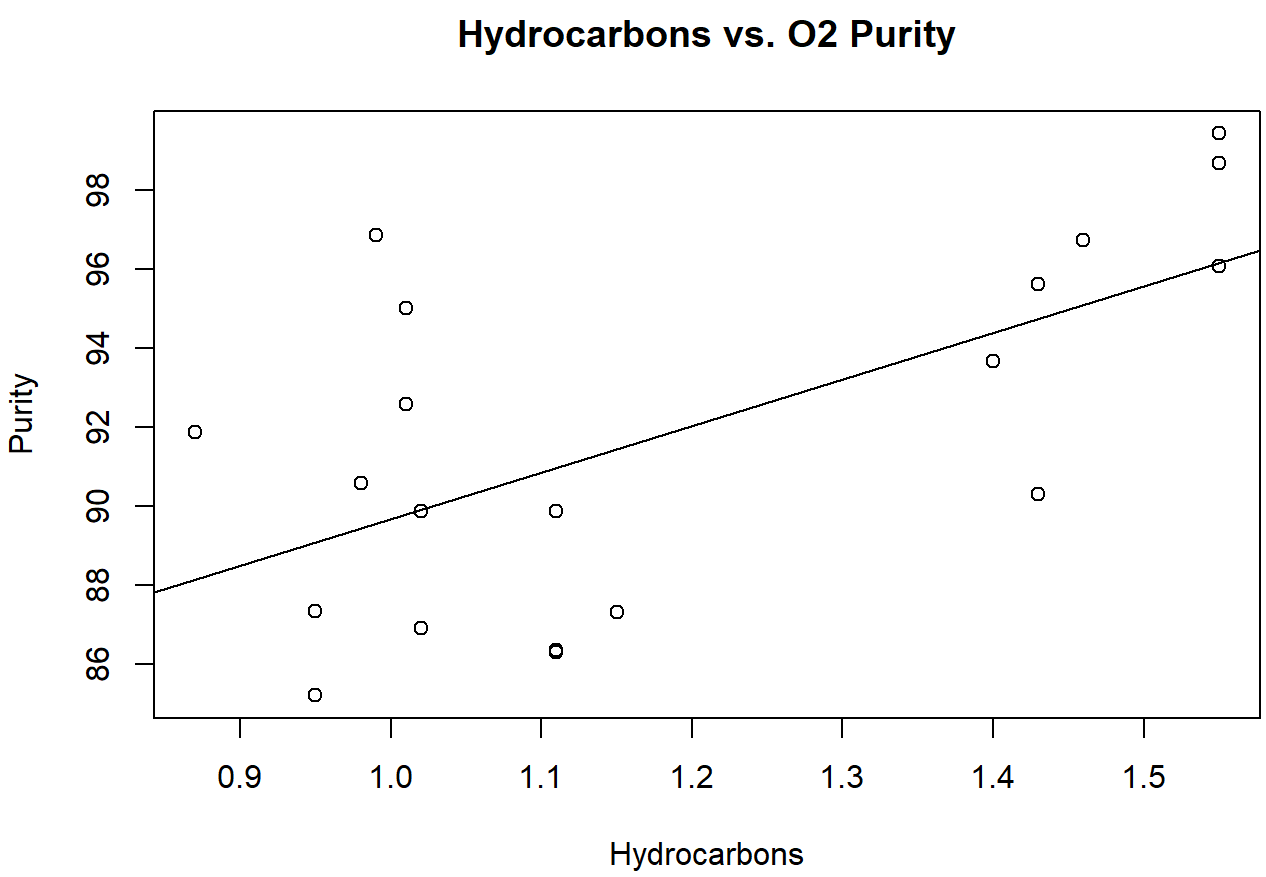
Q3: Question 2.7. The purity of oxygen produced by a fractional distillation process is thought to be related to the percentage of hydrocarbons in the main condenser of the processing unit.

a) Fit a simple linear regression model to the data.

```
chemData <- p2.7
chemFit = lm(purity ~ hydro, data = chemData)
print(chemFit)
```

```
##
## Call:
## lm(formula = purity ~ hydro, data = chemData)
##
## Coefficients:
## (Intercept)        hydro
##       77.86        11.80
```

```
plot(chemData$hydro, chemData$purity,
     xlab = "Hydrocarbons",
     ylab = "Purity",
     main = "Hydrocarbons vs. O2 Purity")
abline(chemFit)
```



The formula describing the model is y = 77.86 + 11.80β1, where 77.86 represents the oxygen purity at a 0 percent hydrocarbon concentration and 11.80 represents the expected units of increase of y for every unit of x.

b) Test the hypothesis H0:β1 = 0.

```
# we can set up our null hypothesis to be: H0:β1 = 0
# Our alternative hypothesis, H_a is β1 != 0
# since no alpha value is given, we can assume to have a two tailed test with an alpha of .1
# we can conduct this with a hypothesis t test. we can find the p value by calling:
summary(chemFit)
```

```
##
## Call:
## lm(formula = purity ~ hydro, data = chemData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.6724 -3.2113 -0.0626  2.5783  7.3037
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   77.863      4.199  18.544 3.54e-13 ***
## hydro         11.801      3.485   3.386  0.00329 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.597 on 18 degrees of freedom
## Multiple R-squared:  0.3891, Adjusted R-squared:  0.3552
## F-statistic: 11.47 on 1 and 18 DF,  p-value: 0.003291
```

Since our p-value is 0.003291 which is less than our one tailed alpha of .05, we can reject the null in favor of the alternate. We have sufficient evidence to claim that β1 != 0.

c) c. Calculate R^2

```
summary(chemFit)$r.squared
```

```
## [1] 0.3891224
```

We get an R^2 value of 0.3891224. This means that 38.91224% of the variability in the oxygen purity can be attributed to the percentage of hydrocarbons in the main condenser.

d) Find a 95% CI on the slope.

```
confint(chemFit, level = 0.95)
```

```
##                2.5 %   97.5 %
## (Intercept) 69.041747 86.68482
## hydro        4.479066 19.12299
```

We are 95% certain that the true slope (β1) falls between: (4.479066, 19.12299)

e) Find a 95% CI on the mean purity when the hydrocarbon percentage is 1.00.

```
new_data = data.frame(hydro = 1.00)
predict(chemFit, newdata = new_data, interval = "confidence", level = 0.95)
```

```
##        fit      lwr      upr
## 1 89.66431 87.51017 91.81845
```

The model predicts that at a hydrocarbon percentage of 1.00, the purity will be at 89.66%. We are 95% certain that the true purity will be between (in percent): (87.51017, 91.81845)

Q4: Question 2.8. Consider the oxygen plant data in Problem 2.7 and assume that purity and hydrocarbon percentage are jointly normally distributed random variables.

a) What is the correlation between oxygen purity and hydrocarbon percentage?

```
# we are using the same data as Q3

# two ways to calculate it!
cor(chemData$purity, chemData$hydro)
```

```
## [1] 0.6237968
```

```
sqrt(summary(chemFit)$r.squared)
```

```
## [1] 0.6237968
```

The correlation coefficient (r) is equal to 0.6237968. This indicates a moderately strong positive linear relationship between the hydrocarbon % and the purity of the oxygen.

b) Test the hypothesis that ρ = 0.

```
# can assume alpha of .1
# H0 :ρ = 0; our alternate hypothesis is that ρ != 0. We can conduct Pearson's product moment co
rrelation by calling:
cor.test(chemData$purity, chemData$hydro)
```

```
##
##  Pearson's product-moment correlation
##
## data:  chemData$purity and chemData$hydro
## t = 3.3861, df = 18, p-value = 0.003291
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2503961 0.8356439
## sample estimates:
##       cor
## 0.6237968
```

```
# and checking the p-value
```

We get a p value of .003 which is less than our alpha, and as such gives us sufficient evidence to reject H0 in favor of our alternate.

c) Construct a 95% CI for ρ.

```
# can call the same function as before as it also gives us a 95% CI
cor.test(chemData$purity, chemData$hydro)
```

```
##
##  Pearson's product-moment correlation
##
## data:  chemData$purity and chemData$hydro
## t = 3.3861, df = 18, p-value = 0.003291
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2503961 0.8356439
## sample estimates:
##       cor
## 0.6237968
```

We see that running this gives us a 95% CI of (0.2503961, 0.8356439). The interpretation of this is that we are 95% certain that the true value of ρ lies between:
(0.2503961, 0.8356439)