```
#Multiple linear regression analysis:  Harris Trust discrimination data

#In 1970s, Harris Trust and Savings Bank was sued for discrimination on the basis of sex.
#Analysis of salaries of employees of one type (skilled, entry-level clerical) presented as
evidence by the defense.
#Did female employees tend to receive lower starting salaries than similarly qualified and
experienced male employees?

#read in data
wages = read.csv("wagediscrim.txt", header= T)

#let's get a sense of the data: how many rows, how many columns?
dim(wages)

#quick summaries of each variable to know the range of data we are working with
summary(wages)

#comparison of males and females on beginning salary (bsal)
boxplot(bsal~sex, data = wages, xlab = "Sex", ylab = "Beginning Salary", main = "Beginning
salaries for male and female employees")

#2-sample inferences -- suggest significant differences in average bsal for men and women in
this bank
t.test(bsal~sex, data = wages)

#look at plots of variables with bsal, one at a time
plot(bsal~ senior + age + educ + exper, data = wages, ask=T)

#you can look at all plots at once, if you want, although plot can be hard to interpret
pairs(wages)

#correlations among all variables, excluding the sex variable since it is a character
variable
cor(wages[,c(1,2, 4:8)])

#see whether there might be differences across men and women in the predictors.
#could use box plots with sex as the X variable.  We'll do summaries to get quick results.

#summary statistics for women
summary(wages[wages$fsex==1,])

#summary statistics for men
summary(wages[wages$fsex==0,])

#it does appear that there are differences in distributions of other variables
#for men and women. Since those other variables are associated with salary, we can't
#simply compare average salaries for men and women.

#let's do a multiple regression! Start with default linear specification of model.

regwage = lm(bsal~ fsex + senior + age + educ + exper, data= wages)

summary(regwage)

#here is the output from R
#Call:
#lm(formula = bsal ~ fsex + senior + age + educ + exper, data = wages)

#Residuals:
#     Min       1Q   Median       3Q      Max
#-1217.36  -342.83   -55.61   297.10  1575.53
#
#Coefficients:
#             Estimate Std. Error t value Pr(>|t|)
```

```
#(Intercept) 6277.8934   652.2713   9.625 2.36e-15 ***
#fsex        -767.9127   128.9700  -5.954 5.39e-08 ***
#senior       -22.5823     5.2957  -4.264 5.08e-05 ***
#age            0.6310     0.7207   0.876 0.383692
#educ          92.3060    24.8635   3.713 0.000361 ***
#exper          0.5006     1.0553   0.474 0.636388
#---
#Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1
#
#Residual standard error: 508.1 on 87 degrees of freedom
#Multiple R-squared: 0.5152,     Adjusted R-squared: 0.4873
#F-statistic: 18.49 on 5 and 87 DF,  p-value: 1.811e-12


#confidence interval for coefficients
confint(regwage)

#                    2.5 %      97.5 %
#(Intercept)  4981.4335106 7574.353262
#fsex        -1024.2545333 -511.570844
#senior        -33.1081429  -12.056463
#age            -0.8014178    2.063338
#educ           42.8870441  141.725002
#exper          -1.5968086    2.598088

## check model assumptions: residual plots versus each predictor (all need to show no
pattern)

plot(y=regwage$residual, x = wages$senior, xlab = "Seniority", ylab = "Residual")
abline(0,0)

plot(y=regwage$residual, x = wages$age, xlab = "Age", ylab = "Residual")
abline(0,0)

plot(y=regwage$residual, x = wages$educ, xlab = "Education", ylab = "Residual")
abline(0,0)

plot(y=regwage$residual, x = wages$exper, xlab = "Experience", ylab = "Residual")
abline(0,0)

boxplot(regwage$residual~ wages$sex, ylab = "Residual")

#results show an apparent quadratic trend and possible nonconstant variance. we should
consider transformations to try to improve
#the reasonableness of the regression assumptions

### let's mean-center the continuous predictor to improve interpretation of outputs
### this has nothing to do with improving the model fit -- it is just a recentering of
results.

#mean centering the variables
wages$agec = wages$age - mean(wages$age)
wages$seniorc = wages$senior - mean(wages$senior)
wages$experc = wages$exper - mean(wages$exper)
wages$educc = wages$educ - mean(wages$educ)

#now let's fit the model with the mean-centered predictors
regwagec = lm(bsal~ fsex + seniorc + agec + educc + experc, data= wages)

summary(regwagec)

#Coefficients:
#             Estimate Std. Error t value Pr(>|t|)
#(Intercept) 5924.0072    99.6588  59.443  < 2e-16 ***
#fsex        -767.9127   128.9700  -5.954 5.39e-08 ***
```

```
#seniorc      -22.5823     5.2957  -4.264 5.08e-05 ***
#agec           0.6310     0.7207   0.876 0.383692
#educc         92.3060    24.8635   3.713 0.000361 ***
#experc         0.5006     1.0553   0.474 0.636388
#---
#Signif. codes:  0 �***� 0.001 �**� 0.01 �*� 0.05 �.� 0.1 � � 1

#Residual standard error: 508.1 on 87 degrees of freedom
#Multiple R-squared:  0.5152,    Adjusted R-squared:  0.4873
#F-statistic: 18.49 on 5 and 87 DF,  p-value: 1.811e-12

#notice that the coefficients for the predictors have not changed.
#but the intercept has changed.  we interpret the intercept as the
#average bsal for male employees who are 474 months old, have 82 months
#of seniority, 12.5 years of education, and 101 months of experience.

### back to model diagnostics and refinement....
#now let's add the squared terms of the centered age and centered experience predictors
#first, let's add them to the dataset.

wages$agec2 = wages$agec^2
wages$experc2 = wages$experc^2

regwagecsquares = lm(bsal~ fsex + seniorc + agec + agec2 + educc + experc + experc2, data=
wages)

summary(regwagecsquares)

#(Intercept)  6.098e+03  1.123e+02  54.313
#fsex        -7.684e+02  1.211e+02  -6.343
#seniorc     -1.764e+01  5.265e+00  -3.351
#agec        -3.473e-01  7.814e-01  -0.444
#agec2        7.195e-04  4.045e-03   0.178
#educc        7.561e+01  2.406e+01   3.142
#experc       4.035e+00  1.479e+00   2.729
#experc2     -2.298e-02  7.592e-03  -3.027
#            Pr(>|t|)
#(Intercept)  < 2e-16 ***
#fsex        1.04e-08 ***
#seniorc      0.00120 **
#agec         0.65783
#agec2        0.85925
#educc        0.00231 **
#experc       0.00772 **
#experc2      0.00326 **

#Here is the best way to interpret the effect of changing experience.  This code
#assumes you are using mean centered variables.  First, make a new dataset with
#however many values of experience you want to examine, say 20, and all other predictors
#equal to zero.  You can do this as follows.

#first, make the 20 values of experience that you want to examine
newexper = c(20, 30, 40, 50, 60, 70, 80, 90, 100, 110, 120, 130, 140, 150, 160, 170, 180,
190, 200, 210)

#now mean center it, since we use a mean centered value in the regression.  Subtract the mean
from whole wages dataset, not the 20 new values
newexperc = newexper - mean(wages$exper)

#now create the squared values, since we use those in the regression as well.
newexperc2 = newexperc^2

#we need to get these into a new dataset with 20 rows and 7 columns (one for each non-
intercept coefficient in the regression)
#we set all the entries equal to 0 when making this matrix. since we use mean-centered
```

```
predictors, the rows in the new
#dataset correspond to male people with average values of seniority, age, and education.

newdata = matrix(0, nrow = 20, ncol = 7)

#now we make it a data frame with the same names as wages for all the predictors in the model
newdata = data.frame(newdata)
names(newdata) = names(wages)[8:14]

#now we replace the 4th and 7th columns of the matrix, i.e., those corresponding to the mean-
centered experience variable, with new mean-centered experience values
newdata[,4] = newexperc
newdata[,7] = newexperc2

#note that here we made sure that the experience values are in the columns with the
experience names

#we set all values other than experience to zero.  this means that we are evaluating the
#predicted values for someone with fsex = 0 and mean centered scores of the other values at
zero.  In other
#words, we predict for a man with average values of age, seniority, and education.

#let's get the intervals for the average wages for the different experience levels for this
"average" man
preds = predict.lm(regwagecsquares, newdata, interval = "confidence")
preds

#you can plot the predicted values versus the experience
plot(y = preds[,1], x = newexper, xlab = "Experience", ylab = "Predicted Wages")
title("Expected Change in Wages with Experience (Male with Average Values of Other
Predictors)")

#if you want to get the 95% confidence bands on the plot as well, you can do the following

#stack the upper and lower limits and the predicted values in one vector
tempy = c(preds[,1], preds[,2], preds[,3])

#stack the newexper values three times, corresponding to each of the 3 preds columns
tempx = c(newexper, newexper, newexper)

#make the plot without the points
plot(y = tempy, x = tempx, type = "n", xlab = "Experience", ylab = "Predicted Wages")

#now add the points, with different plotting symbol for the limits
points(y = preds[,1], x = newexper, pch = 1)
points(y = preds[,2], x = newexper, pch = 2)
points(y = preds[,3], x = newexper, pch = 2)
title("Expected Change in Wages with Experience (Male with Average Values of Other
Predictors)")

## here is a bit of code for making dummy variable for males = 1 and females = 0

wages$msex = rep(0, nrow(wages))
wages$msex[wages$sex == "Male"] = 1

#check to make sure we did what we set out to do
cbind(wages$sex, wages$msex)

#let's run the regression with msex instead of fsex

regwagecmale = lm(bsal~ msex + seniorc + agec + educc + experc, data= wages)

summary(regwagecmale)

#Coefficients:
```

```
#              Estimate Std. Error t value Pr(>|t|)
#(Intercept) 5156.0946    68.8852  74.851  < 2e-16 ***
#msex         767.9127   128.9700   5.954 5.39e-08 ***
#seniorc      -22.5823     5.2957  -4.264 5.08e-05 ***
#agec           0.6310     0.7207   0.876 0.383692
#educc         92.3060    24.8635   3.713 0.000361 ***
#experc         0.5006     1.0553   0.474 0.636388

#Residual standard error: 508.1 on 87 degrees of freedom
#Multiple R-squared:  0.5152,    Adjusted R-squared:  0.4873
#F-statistic: 18.49 on 5 and 87 DF,  p-value: 1.811e-12

#compared to results of regwagec, output is same except for sign of msex coefficient
(positive instead of negative)
```