Lab 11 Multiple Regression

Estimated Multiple Regression Equation

If we choose the parameters α and β_k $(k = 1, 2, ..., p)$ in the multiple linear regression model so as to minimize the sum of squares of the error term ϵ, we will have the so called estimated multiple regression equation. It allows us to compute fitted values of y based on a set of values of x_k $(k = 1, 2, ..., p)$.

$$\boxed{\hat{y} = a + \sum_{k=1}^{n} b_k x_k}$$

Load file `bball.txt` into RStudio

Problem

Apply the multiple linear regression model for the data set `bball`, and predict the average points per game if the height is 6.8, weight is 250, field goal percent is 0.55 and free throw percent is 0.75.

Solution

We apply the `lm` function to a formula that describes the variable `av_pts` by the variables `Height`, `Weight`, `perc_fg`, and `perc_ft`. And we save the linear regression model in a new variable `av_pts.lm`.

```
> av_pts.lm = lm(av_pts ~
+ Height + Weight + perc_fg + perc_ft, data=bball)
```

We also wrap the parameters inside a new data frame named newdata.

```
> newdata = data.frame(Height=6.8,  # wrap the parameters
+     Weight=250, perc_fg=0.55, perc_ft=0.75)
```

Lastly, we apply the predict function to `av_pts.lm` and `newdata`.

```
> predict(av_pts.lm, newdata)
1
16.3133
```

Answer

Based on the multiple linear regression model and the given parameters, the predicted average points per game is 16.3133.

Multiple Coefficient of Determination

The coefficient of determination of a multiple linear regression model is the quotient of the variances of the fitted values and observed values of the dependent variable. If we denote y_i as the observed values of the dependent variable, $\bar{y}$ as its mean, and $\hat{y}_i$ as the fitted value, then the coefficient of determination is:

$$R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$

Problem
Find the coefficient of determination for the multiple linear regression model of the data set `bball`.

Solution
We apply the `lm` function to a formula that describes the variable `av_pts` by the variables `Height`, `Weight`, `perc_fg`, and `perc_ft`. And we save the linear regression model in a new variable `av_pts.lm` .

```
> av_pts.lm = lm(av_pts ~
+ Height + Weight + perc_fg + perc_ft, data=bball)
```

Then we extract the coefficient of determination from the `r.squared` attribute of its `summary`.

```
> summary(av_pts.lm)$r.squared
[1] 0.2222506
```

Answer
The coefficient of determination of the multiple linear regression model for the data set bball is 0.222

Significance Test for MLR

Assume that the error term ϵ in the [multiple linear regression (MLR) model](#) is independent of x_k (k = 1, 2, ..., p), and is [normally distributed](#), with zero [mean](#) and constant [variance](#). We can decide whether there is any **significant relationship** between the dependent variable y and any of the independent variables x_k (k = 1, 2, ..., p).

Problem

Decide which of the independent variables in the multiple linear regression model of the data set `bball` are statistically significant at $\alpha = 0.05$ significance level.

Solution

We apply the `lm` function to a formula that describes the variable `av_pts` by the variables `Height`, `Weight`, `perc_fg`, and `perc_ft`.. And we save the linear regression model in a new variable `av_pts.lm`.

```
> summary(av_pts.lm)

Call:
lm(formula = av_pts ~ Height + Weight + perc_fg + perc_ft, data = bball)

Residuals:
   Min     1Q Median     3Q    Max 
-8.966 -3.545 -1.187  2.613 15.211 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)   
(Intercept)  4.148707  14.855006   0.279  0.78121   
Height      -3.690499   2.970780  -1.242  0.22005   
Weight       0.009458   0.046297   0.204  0.83897   
perc_fg     47.940199  15.709131   3.052  0.00367 **
perc_ft     11.371019   7.868536   1.445  0.15479   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.411 on 49 degrees of freedom
Multiple R-squared:  0.2223,   Adjusted R-squared:  0.1588 
F-statistic: 3.501 on 4 and 49 DF,  p-value: 0.01364
```

Answer

As the *p-value* of perc_fg is less than 0.05, it is statistically significant in the multiple linear regression model of av_pts.

```
> av_pts.lm = lm(av_pts ~ perc_fg + perc_ft, data=bball)
> summary(av_pts.lm)$r.squared
[1] 0.1778529
> summary(av_pts.lm)

Call:
lm(formula = av_pts ~ perc_fg + perc_ft, data = bball)

Residuals:
   Min     1Q Median     3Q    Max 
-8.486 -3.267 -1.178  3.281 15.694 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept)  -15.277      8.244  -1.853  0.06966 . 
perc_fg       35.825     13.247   2.704  0.00928 **
```

```
perc_ft        14.799      7.480   1.978  0.05330 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.453 on 51 degrees of freedom
Multiple R-squared:  0.1779,   Adjusted R-squared:  0.1456
F-statistic: 5.516 on 2 and 51 DF,  p-value: 0.00678
```

```
> av_pts.lm = lm(av_pts ~ perc_ft, data=bball)
> summary(av_pts.lm)

Call:
lm(formula = av_pts ~ perc_ft, data = bball)

Residuals:
   Min     1Q Median     3Q    Max
-9.512 -3.338 -1.735  2.188 15.059

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.091      5.928   0.184   0.8547
perc_ft       14.423      7.920   1.821   0.0743 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.774 on 52 degrees of freedom
Multiple R-squared:  0.05995,  Adjusted R-squared:  0.04187
F-statistic: 3.316 on 1 and 52 DF,  p-value: 0.07435
```

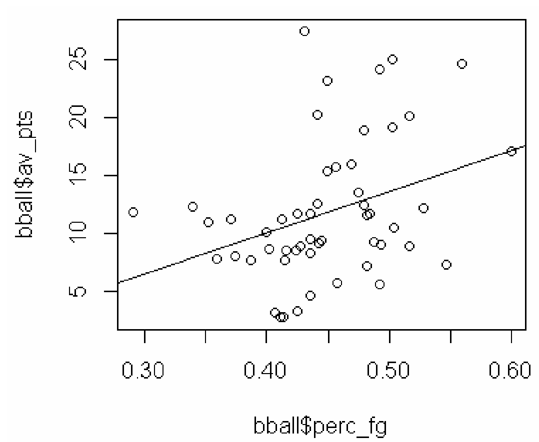
```
> av_pts.lm = lm(av_pts ~ perc_fg, data=bball)
> summary(av_pts.lm)

Call:
lm(formula = av_pts ~ perc_fg, data = bball)

Residuals:
   Min     1Q Median     3Q    Max
-7.815 -3.171 -1.180  3.290 16.249

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    -4.08       6.16  -0.662   0.5107
perc_fg        35.34      13.61   2.596   0.0122 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05
‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.604 on 52 degrees of freedom
Multiple R-squared:  0.1148,   Adjusted R-squared:  0.09773
F-statistic: 6.741 on 1 and 52 DF,  p-value: 0.01222
```



Pairwise Interactions

Regression With Interaction Variables

Interaction variables introduce an additional level of regression analysis by allowing researchers to explore the synergistic effects of combined predictors. This tutorial will explore how interaction models can be created in R.

Load `Icecream.txt` dataset into RStudio

Variables

- `consume`: Ice cream consumption in pints per capita
- `price`: Per pint price of ice cream in dollars
- `income`: Weekly family income in dollars
- `temp`: Mean temperature in degrees F

Planning The Model

Suppose that our research question is "how much of the variance in ice cream consumption can be predicted by per pint price, weekly family income, mean temperature, and *the interaction between per pint price and weekly family income*?" The italicized interaction term is the new addition to our typical multiple regression modeling procedure. This variable is relatively simple to incorporate, but it does require a few preparations. Before we look at the interaction between variables, let's build some models.

Compare Pairwise Models

First we build a model with all three explanatory variables

```
> model1 <- lm(consume ~ price + income + temp, Icecream)
> summary(model1)
```

```
Call:
lm(formula = consume ~ price + income + temp, data = Icecream)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.059405 -0.015665  0.005229  0.017157  0.070515 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.0877445  0.2447400   0.359   0.7230    
price       -0.3863577  0.7830856  -0.493   0.6261    
income       0.0026176  0.0010765   2.432   0.0225 *  
temp         0.0031191  0.0004168   7.483 7.78e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.03291 on 25 degrees of freedom
Multiple R-squared:  0.6948,	Adjusted R-squared:  0.6582 
F-statistic: 18.97 on 3 and 25 DF,  p-value: 1.256e-06
```

RSquared for Model1 is 0.6948 but notice the *p-value* for `price` is quite high, so let's look at a model without the `temp` variable and see the effect of the price variable.

```
> model2 <- lm(consume ~ price + income, Icecream)
> summary(model2)
```

```
Call:
lm(formula = consume ~ price + income, data = Icecream)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.098772 -0.034685 -0.009381  0.034351  0.117713 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.5777990  0.4161919   1.388    0.177
price       -0.6669640  1.3804937  -0.483    0.633
income      -0.0004845  0.0017534  -0.276    0.784

Residual standard error: 0.05809 on 26 degrees of freedom
Multiple R-squared:  0.01126,	Adjusted R-squared:  -0.0648 
F-statistic: 0.148 on 2 and 26 DF,  p-value: 0.8632
```

Again the *p-value* for `price` is quite high, and RSquared is very low, so the effect of `temp` in our model is significant. Let's build another model, this time with income and temp.

```
> model3 <- lm(consume ~ income + temp, Icecream)
> summary(model)

Call:
lm(formula = consume ~ income + temp, data = Icecream)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.059121 -0.021892  0.003275  0.020605  0.073075 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.0224003  0.0988245  -0.227   0.8225    
income       0.0026544  0.0010582   2.508   0.0187 *  
temp         0.0031289  0.0004103   7.627 4.28e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.03243 on 26 degrees of freedom
Multiple R-squared:  0.6918,	Adjusted R-squared:  0.6681 
F-statistic: 29.18 on 2 and 26 DF,  p-value: 2.262e-07
```

RSquared for model3 is high at 0.6918, and similar to model1,except we are only using 2 variables, so this confirms price of icecream has little to do with icecream consumption.

Question: Can we bump up the RSquared without measuring any new variables?

Creating The Interaction Variable

A two step process can be followed to create an interaction variable in R. First, the input variables must be centered to mitigate multicollinearity. Second, these variables must be multiplied to create the interaction variable.

Step 1: Centering

To center a variable, simply subtract its mean from each data point and save the result into a new R variable, as demonstrated below.

```
> #center the input variables
> pricec <- price - mean(price)
> incomec <- income - mean(income)
```

Step 2: Multiplication
Once the input variables have been centered, the interaction term can be created. Since an interaction is formed by the product of two or more predictors, we can simply multiply our centered terms from step one and save the result into a new R variable, as demonstrated below.

```
> #create the interaction variable
> priceincomei <- pricec * incomec
```

Creating The Model
Now we have all of the pieces necessary to assemble our complete interaction model.

```
> #create the interaction model using lm(formula, icecream)
> #predict ice cream consumption by its per pint price, weekly family income,
mean temperature, and the interaction between per pint price and weekly
family income
> interactionmodel <- lm(consume ~ price + income + temp + priceincomei,
Icecream)
> #display summary information about the model
> summary(interactionmodel)
```

A summary of our interaction model is displayed below.

```
Call:
lm(formula = consume ~ price + income + temp + priceincomei,
    data = Icecream)

Residuals:
      Min        1Q    Median        3Q       Max
-0.057528 -0.016359 -0.000848  0.016866  0.071892

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.1570203  0.2324673   0.675   0.5058
price        -0.1636906  0.7438870  -0.220   0.8277
income        0.0012301  0.0012133   1.014   0.3208
temp          0.0028231  0.0004171   6.769 5.31e-07 ***
priceincomei -0.2786003  0.1344397  -2.072   0.0491 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.03094 on 24 degrees of freedom
Multiple R-squared:  0.7411,   Adjusted R-squared:   0.698
F-statistic: 17.18 on 4 and 24 DF,  p-value: 8.968e-07
```

At this point we have a complete interaction model with an RSquared of 0.7411. Naturally, if this were a full research analysis, we would likely compare this model to others and assess the value of each predictor.   As follows:

```
> tempc <- temp - mean(temp)
> incometempi <- incomec * tempc
> interactionmodel <- lm(consume ~ income + temp + incometempi, Icecream)
> summary(interactionmodel)

Call:
lm(formula = consume ~ income + temp + incometempi, data = Icecream)

Residuals:
```

```
      Min        1Q    Median        3Q       Max 
-0.058731 -0.020999  0.004282  0.020877  0.074117 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -7.636e-03  1.074e-01  -0.071   0.9439    
income       2.493e-03  1.153e-03   2.163   0.0403 *  
temp         3.083e-03  4.334e-04   7.114 1.86e-07 ***
incometempi -2.853e-05  7.314e-05  -0.390   0.6998    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 

Residual standard error: 0.03297 on 25 degrees of freedom
Multiple R-squared:  0.6937,	Adjusted R-squared:  0.6569 
F-statistic: 18.87 on 3 and 25 DF,  p-value: 1.313e-06
```

RSquared is lower, so maybe we should try a full model with all interactions.

```
> interactionmodel <- lm(consume ~ price + income + temp + incometempi
+ + pricetempi + priceincomei, Icecream)
> summary(interactionmodel)

Call:
lm(formula = consume ~ price + income + temp + incometempi + 
    pricetempi + priceincomei, data = Icecream)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.055945 -0.015069 -0.001735  0.013128  0.074092 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)   2.390e-01  2.673e-01   0.894   0.3810    
price        -2.406e-01  7.682e-01  -0.313   0.7571    
income        5.752e-04  1.529e-03   0.376   0.7104    
temp          2.640e-03  4.670e-04   5.652  1.1e-05 ***
incometempi  -8.833e-05  7.461e-05  -1.184   0.2491    
pricetempi    1.603e-02  6.770e-02   0.237   0.8150    
priceincomei -3.149e-01  1.657e-01  -1.900   0.0706 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 

Residual standard error: 0.03123 on 22 degrees of freedom
Multiple R-squared:  0.7581,	Adjusted R-squared:  0.6922 
F-statistic: 11.49 on 6 and 22 DF,  p-value: 7.834e-06
```

RSquared is 0.7581

This looks like the best model. Our original model with all three variables and no interaction variables gave RSquared as 0.6948.

Our interaction model with only income and temp variables gave RSquared of 0.6937.

Answer: We could use three variables with all pairwise interactions, or two variables (income and temp) with one pairwise interaction, and get as good or better results than our original model with 3 variables.

Conclusion: A model containing interactions between variables and measuring data collected from only two variables may be cheaper and easier to obtain, and provide results as good or better than a model using more variables.

References

Kadiyala, K. (1970). Ice Cream [Data File]. Retrieved December 14, 2009 from http://lib.stat.cmu.edu/DASL/Datafiles/IceCream.html

Lab 11: On Your Own **Name** ________________________ **Score** _______

Load `HeartData.txt` dataset. It gives 2009 data that measures 7 variables that might contribute to the likelihood of US heart attacks.

The variables in the dataset are

`State`	includes 50 states and DC plus territories of Guam, Puert Rico, US Virgin Islands
`HeartAttack`	percent of population experiencing a heart attack
`ColGrad`	percent of population graduating from college
`HSDrop`	percent of population dropping out of high school
`FruitVeg`	percent of population eating lots of fruits and vegetables
`OBESE`	percent of population considered obese
`BingeDrink`	percent of population that binge drinks
`SmokesDay`	percent of population that smokes daily
`CholestChkd`	percent of population that gets their cholesterol checked regularly (at least once a year)

1. Show lsr model and plot of `HeartAttack` vs `OBESE`. What percent of the variability in `HeartAttack` can be explained by the variability in `OBESE`?

2. Show a multiple regression model that contains the 7 quantitative variables that influence `HeartAttack`. What does the multiple R Squared for this model indicate?

3. Come up with a multiple regression model using only two of the above 7 explanatory variables that accounts for the highest proportion of variability in the response variable.

4. Show a correlation matrix for `HeartData` and explain its meaning

5. Show pairwise interactions for the 7 explanatory variables. Using only two of the 7 explanatory variables,which model with interaction accounts for the most variation in the response variable (`HeartAttack`)?

6. Using dataset `bball`, come up with a model using `perc_fg` and `perc_ft` and include their interaction. How much of the variability in `av_pts` does this model explain?