Week 4: Regression analysis in-class example (R version)

A large national grocery retailer tracks productivity and costs of its facilities closely. Data were obtained from a single distribution center for a one-year period. Each data point for each variable represents one week of activity. The variables included are

- **Cases**: the number of cases shipped,
- **Costs**: the indirect costs of the total labor hours as percentage X2,
- **Holiday**: a qualitative predictor that is coded as 1 if the week has a holiday and 0 otherwise
- **Hours**: and the total labor hours

1) Analyze the interrelationships between Hours and the other three predictors

3) Two models were fitted to predict hours:

    Model M1 includes all three variables and Model M2 includes only cases and holiday as predictors (see R output)

Select the best model to predict Total labor hours, and write the model expression.

4) Which of the predictors have a significant effect on total labor hours?

5) Analyze residuals to check model assumptions:

    Linearity:

    Constant Variance:

    Normality of errors :

    Are there any outliers?

6) Analyze the Coefficient of Determination R^2 and the goodness of fit test. What can you conclude about the predictive power of the model?

7) You are asked to estimate the **average** weekly labor hours for **shipments of 302,000 cases** with 7.20% indirect costs in a non-holiday week. Find appropriate estimates and 95% confidence intervals in the computer output.

8) How does the result change, if you need to predict the weekly labor hours for a specific shipment of 302,000 cases with 7.20% indirect costs in a non-holiday week? The actual handling time will be compared with the predicted time for quality control. Find appropriate estimates and prediction intervals in the R output.

9) You are asked to estimate the **average** weekly labor hours with the following conditions:

1. Cases = 302,000 with 7.20% indirect costs in a holiday week
2. Cases = 295,000 with 6.95% indirect costs in a holiday week
3. Cases = 230,000 in a non-holiday week

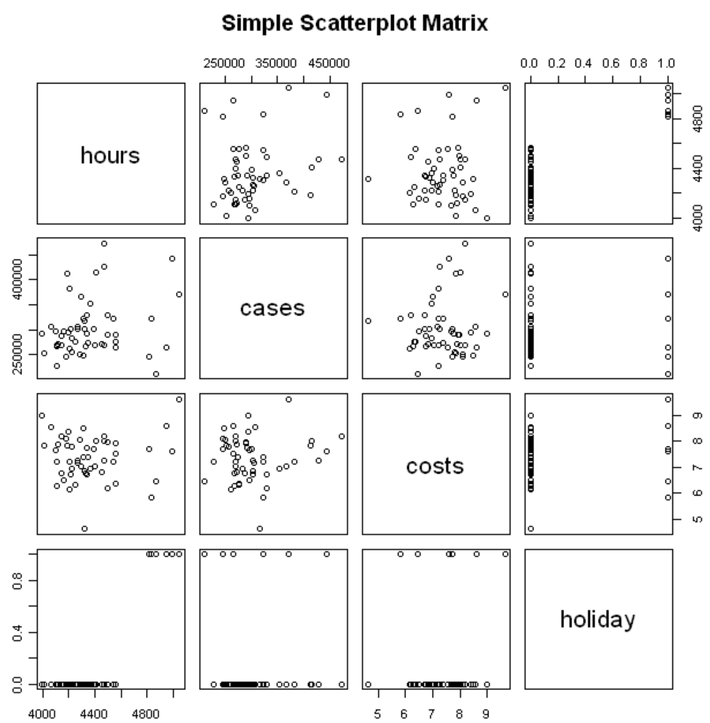
Find appropriate estimates and 95% confidence intervals in the R output

10) Estimate the difference in the numbers of total hours needed to handle a certain shipment during a holiday week and a non-holiday week.

R code

```
> #Data on large national grocery retailer obtained
> #from a single distribution center for a one-year period.
> #Each data point represents one week of activity.
> #VARIABLES: cases = number of cases shipped,
> #costs = the indirect costs of the total labor hours as percentage,
> #holiday = 1 if the week has a holiday and 0 otherwise,
> #hours = total labor hours */# IP costs and margin profits.
>
> mydata = read.table("groceryretailer.txt", header=T)
> # header = T since first row contains variables names
> # skip= N where N is the number of lines of the data file
> # to skip before beginning to read data
```

```
>
> #define variables from dataset
> hours = mydata[, 1]
> cases = mydata[,2]
> costs = mydata[,3]
> holiday = mydata[,4]
>
```
> #compute correlation values
```
> cor(mydata)
            hours      cases      costs    holiday
hours   1.0000000 0.20766494 0.06002960 0.81057940
cases   0.2076649 1.00000000 0.08489639 0.04565698
costs   0.0600296 0.08489639 1.00000000 0.11337076
holiday 0.8105794 0.04565698 0.11337076 1.00000000
>
> # Basic Scatterplot Matrix
> pairs(~hours + cases + costs + holiday, data=mydata, main="Simple
Scatterplot Matrix")
```




```
>
> #Scatterplots
> plot(hours, cases, xlab="number of employees", ylab="salary")
>
```
> # MULTIPLE LINEAR REGRESSION
> # FITTING MODEL M1
```
> fit <- lm(hours ~ cases + costs + holiday, data=mydata)
> summary(fit) # show results

Call:
```

```
lm(formula = hours ~ cases + costs + holiday, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max 
-264.05 -110.73  -22.52   79.29  295.75 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  4.150e+03  1.956e+02  21.220  < 2e-16 ***
cases        7.871e-04  3.646e-04   2.159   0.0359 *  
costs       -1.317e+01  2.309e+01  -0.570   0.5712    
holiday      6.236e+02  6.264e+01   9.954 2.94e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 143.3 on 48 degrees of freedom
Multiple R-squared: 0.6883,     Adjusted R-squared: 0.6689 
F-statistic: 35.34 on 3 and 48 DF,  p-value: 3.316e-12

> 
> # FITTING MODEL M2
> fit <- lm(hours ~ cases + holiday, data=mydata)
> summary(fit) # show results

Call:
lm(formula = hours ~ cases + holiday, data = mydata)

Residuals:
     Min       1Q   Median       3Q      Max 
-286.249  -99.650   -9.251   70.746  292.311 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 4.058e+03  1.109e+02  36.592  < 2e-16 ***
cases       7.704e-04  3.609e-04   2.135   0.0378 *  
holiday     6.196e+02  6.183e+01  10.021 1.88e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 142.3 on 49 degrees of freedom
Multiple R-squared: 0.6862,     Adjusted R-squared: 0.6734 
F-statistic: 53.58 on 2 and 49 DF,  p-value: 4.647e-13
> 
> # USEFUL FUNCTIONS
> #analysis of variance table to display MSE/SSE values
> anova(fit)
Analysis of Variance Table

Response: hours
          Df  Sum Sq Mean Sq  F value    Pr(>F)    
```

```
cases       1  136366  136366   6.7344   0.01244 *
holiday     1 2033565 2033565 100.4276 1.875e-13 ***
Residuals 49  992204   20249
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> # CIs for model parameters
> confint(fit, level=0.95)
                   2.5 %       97.5 %
(Intercept) 3.835475e+03 4.281231e+03
cases       4.520001e-05 1.495570e-03
holiday     4.953725e+02 7.438786e+02
>
> # FUNCIONS FOR PREDICTONS (QUESTIONS 7 and 8)
>
> #create a new dataset containing values for predictions
> # Example of prediction for one data point.
> new <- data.frame(cases=c(302000), costs=c(7.20),holiday=c(0))
> # compute predicted value and standard error
> predict(fit, new, se.fit = T)
$fit
       1
4291.009
$se.fit
[1] 20.98101
$df
[1] 49
$residual.scale
[1] 142.2992

> # compute predicted value and prediction interval
> predict(fit, new, interval="prediction", level=0.95)
       fit      lwr      upr
1 4291.009 4001.957 4580.062

> # compute average value and confidence interval
> predict(fit, new, interval="confidence",level=0.95)
       fit      lwr      upr
1 4291.009 4248.846 4333.172
>
> # Compute predicted hours for several data points:
> #  CASES COSTS  HOLIDAYS          HOURS
> # 302000  7.20  1 (holiday)         ?
> # 295000  6.95  1                   ?
> # 230000    NA  0                   ?
> #
> # enter data values into separate variable using the c() function that have
> cases = c(302000, 295000, 230000)
> costs = c(7.20, 6.95, NA )
> holiday=c(1,1,0)
```

```
> newPred <- data.frame(cases, costs, holiday) # define new dataframe
>
> # compute predicted values and prediction intervals
> predict(fit, newPred, se.fit = T, interval="prediction", level=0.95)
$fit
       fit      lwr      upr
1 4910.635 4601.712 5219.557
2 4905.242 4596.187 5214.297
3 4235.542 3941.838 4529.245
$se.fit
       1        2        3
58.15834 58.33227 33.33695
$df
[1] 49
$residual.scale
[1] 142.2992

> # compute average response values and confidence intervals
> predict(fit, newPred, se.fit = T, interval="confidence",level=0.95)
$fit
       fit      lwr      upr
1 4910.635 4793.761 5027.508
2 4905.242 4788.019 5022.465
3 4235.542 4168.548 4302.535
$se.fit
       1        2        3
58.15834 58.33227 33.33695
$df
[1] 49
$residual.scale
[1] 142.2992

> # DIAGNOSTICS METHODS
> #Diagnostics plots
> layout(matrix(c(1,2,3,4),2,2)) # optional 4 graphs/page
> plot(fit)
> layout(matrix(c(1),1,1)) # reset
>
> #residuals vs fitted values plot
> plot( fitted(fit), rstandard(fit), main="Predicted vs residuals plot")
> abline(a=0, b=0, col='red')
>
> #residuals vs independent variables
> plot(mydata$cases, rstandard(fit), main="Margin vs residuals plot")
> abline(a=0, b=0,col='red')
>
> #normal probability plot of residuals
> qqnorm(rstandard(fit))
> qqline(rstandard(fit), col = 2)
```