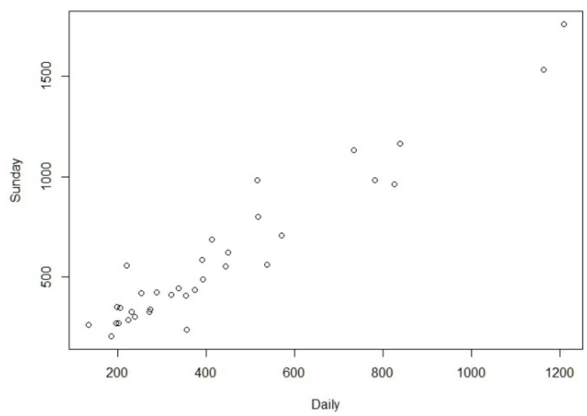
NORTHWESTERN UNIVERSITY | SCHOOL OF PROFESSIONAL STUDIES

R Lesson 12 - Solutions
MSPA 401 – Introduction to Statistical Analysis

Exercises: Problems 1 through 6 use the data listed in the data file **newsapers.csv**. The data are from the *Gale Directory of Publications,* 1994. A sample of 34 newspapers are listed along with their Daily and Sunday circulations (in thousands).

1) Plot Sunday circulation versus Daily circulation. Does the scatter plot suggest a linear relationship between the two variables? Calculate the Pearson product moment correlation coefficient between Sunday and Daily circulation.



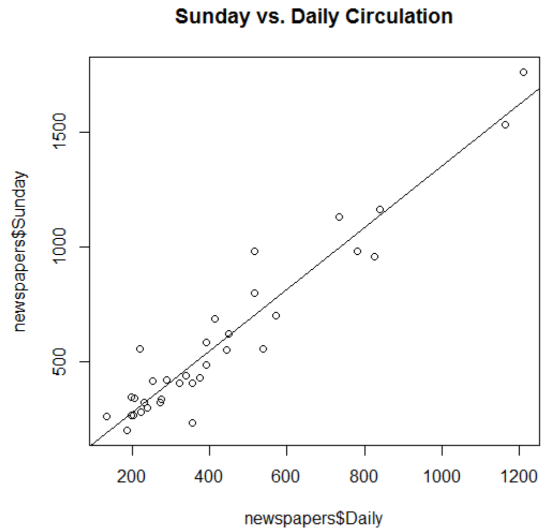
Scatterplot shows strong, positive relationship between Daily and Sunday circulation. Pearson product moment correlation coefficient: 0.9581543.

2) Fit a regression line with Sunday circulation as the dependent variable. Plot the regression line with the circulation data. (Use Lander pages 212-213 for reference.) Comment on the quality of the fit. What percent of the total variation in Sunday circulation is accounted for by the regression line?

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 13.83563   35.80401   0.386    0.702
Daily        1.33971    0.07075  18.935   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 109.4 on 32 degrees of freedom
Multiple R-squared:  0.9181,    Adjusted R-squared:  0.9155
F-statistic: 358.5 on 1 and 32 DF,  p-value: < 2.2e-16
```

The model appears to be well-fitted, with 91.81% of the variation in Sunday explained.



3) Obtain 95% confidence intervals for the coefficients in the regression model. Use confint().

```
> confint(my_model, level = 0.95)
                 2.5 %    97.5 %
(Intercept) -59.094743 86.766003
Daily         1.195594  1.483836
```

4) Determine a 95% prediction interval to predict Sunday circulation for all available values of Daily circulation. Use predict(model, interval="prediction", level=0.95). Then, define a new data frame using Daily = 500 and Sunday = NA. Predict an interval for Sunday circulation.

To predict Sunday circulation for all available values of Daily circulation use:

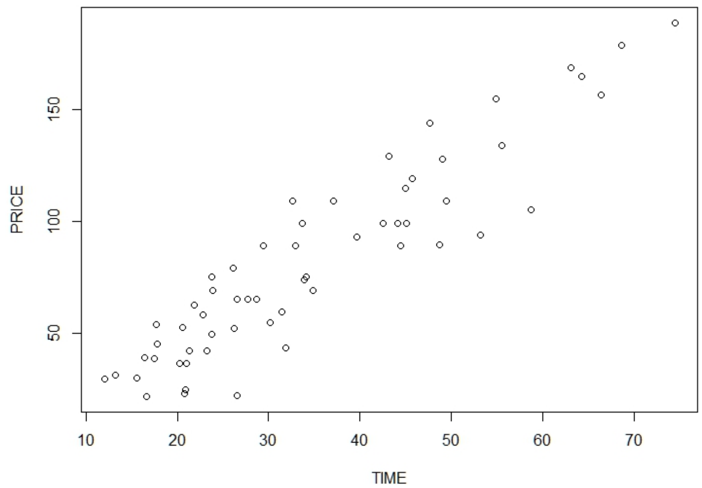
predict(my_model, interval = "prediction", level = 0.95)

```
> head(predict(my_model, interval = "prediction", level = 0.95))
       fit      lwr       upr
1 538.9395 312.7321  765.1469
2 706.4427 479.9656  932.9198
3 490.2757 263.8777  716.6737
4 333.4313 105.5999  561.2626
5 734.3074 507.6465  960.9683
6 996.8848 766.5747 1227.1950
```

```
> predict(my_model, newdata=new_data_frame, interval="prediction", level=0.95)
      fit      lwr      upr
1 683.693 457.3367 910.0493
```

Exercise: Refer to **tableware.cvs** described in Lesson 10. Solve the following problem.

5) Regress PRICE as a dependent variable against TIME. Comment on the quality of the fit. Is a simple linear regression model adequate or is something more needed?




```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -7.1891     5.4053   -1.33    0.189
TIME          2.5625     0.1421   18.03   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16.77 on 57 degrees of freedom
Multiple R-squared:  0.8508,    Adjusted R-squared:  0.8482
F-statistic: 325.1 on 1 and 57 DF,  p-value: < 2.2e-16
```

The model appears to fit well, explaining 85.08% of the PRICE variation.

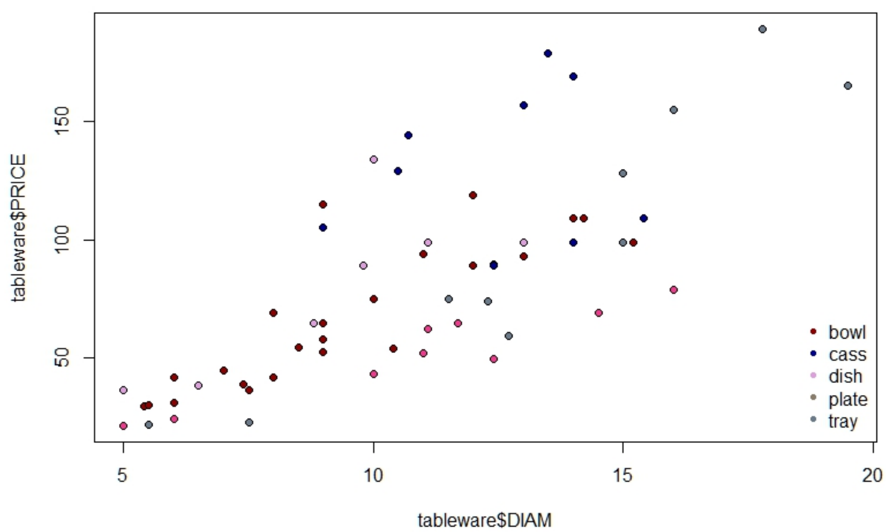
6) ANOVA can be accomplished using a regression model. Regress PRICE against the variables BOWL, CASS, DISH and TRAY as they are presented in the data file. What do the coefficients represent in this regression model? How is the effect of plate accounted for?


```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    51.83      12.11   4.281 7.68e-05 ***
BOWL           15.56      14.28   1.089  0.28086
CASS           75.09      16.69   4.499 3.67e-05 ***
DISH           28.31      18.31   1.546  0.12785
TRAY           47.12      16.69   2.823  0.00665 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The multiple R-squared value for this regression is 0.3367 and the adjusted R-squared is 0.2876. The estimated coefficients represent incremental costs associated with the types of tableware. The type plate is represented by all zeroes for the indicator variables included in the model with binary indicators. The intercept measures its average price. This can be demonstrated with the following statements:

```
> index <- tableware$TYPE == "plate"
> mean(tableware[index,8])
[1] 51.83333
```

7) Plot PRICE versus DIAM and calculate the Pearson product moment correlation coefficient. Include DIAM in the regression model in (6). Compare results between the two models. DIAM is referred to as a covariate. Does its inclusion improve upon the fit of the first model without DIAM?




```
> with(tableware, print(cor(DIAM, PRICE)))
[1] 0.7552496
```

```
> # First fit PRICE as a function of TYPE.
> Price_Type <- {PRICE ~ TYPE}
> Price_Type_fit <- lm(Price_Type, data = tableware)
> summary(Price_Type_fit)

Call:
lm(formula = Price_Type, data = tableware)

Residuals:
    Min      1Q  Median      3Q     Max 
-76.950 -26.362  -2.333  26.109  90.050 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   67.391      7.575   8.897 3.62e-12 ***
TYPEcass      59.529     13.760   4.326 6.59e-05 ***
TYPEdish      12.752     15.681   0.813   0.4197    
TYPEplate    -15.558     14.283  -1.089   0.2809    
TYPEtray      31.559     13.760   2.294   0.0257 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 36.33 on 54 degrees of freedom
Multiple R-squared:  0.3367, Adjusted R-squared:  0.2876 
F-statistic: 6.853 on 4 and 54 DF,  p-value: 0.0001548

> anova(Price_Type_fit)
Analysis of Variance Table

Response: PRICE
          Df Sum Sq Mean Sq F value    Pr(>F)    
TYPE       4  36174  9043.5  6.8532 0.0001548 ***
Residuals 54  71258  1319.6                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> # Then, expand the model to include DIAM
> bigger_model <- {PRICE ~ DIAM + TYPE}
> bigger_model_fit <- lm(bigger_model, data = tableware)
> summary(bigger_model_fit)

Call:
lm(formula = bigger_model, data = tableware)

Residuals:
    Min      1Q  Median      3Q     Max
-44.341 -14.426  -1.617  11.102  51.596

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -18.3107    10.4568  -1.751 0.085719 .
DIAM          9.0794     0.9872   9.197 1.44e-12 ***
TYPEcass     31.8285     9.1312   3.486 0.000994 ***
TYPEdish     15.1821     9.8272   1.545 0.128318
TYPEplate   -28.4183     9.0563  -3.138 0.002778 **
TYPEtray     -3.3142     9.4172  -0.352 0.726284
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 22.76 on 53 degrees of freedom
Multiple R-squared:  0.7445, Adjusted R-squared:  0.7204
F-statistic: 30.89 on 5 and 53 DF,  p-value: 1.422e-14

> anova(bigger_model_fit)  # both variables are significant
Analysis of Variance Table

Response: PRICE
          Df Sum Sq Mean Sq  F value    Pr(>F)
DIAM       1  61280   61280 118.3229 4.091e-15 ***
TYPE       4  18704    4676   9.0287 1.228e-05 ***
Residuals 53  27449     518
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Comparing the two models, it is apparent adding DIAM improves the fit based on a comparison of the adjusted R-squared values. Regardless, the model involving PRICE and TIME is better which indicates a more involved model should be considered.