Stat 615/415 (Regression) HW # 6 – Solutions

Multivariate Regression (chap. 6)

1. (**6.3**) A student stated: "Adding predictor variables to a regression model can never reduce R^2, so we should include all available predictor variables in the model." Comment.

   SOLUTION. *R^2 is a fair measure of comparison for models with the same number of variables. For models of different ranks p, there are other measures of comparison such as adjusted R^2 because R^2 can only increase when variables are added to the model, even if they are completely irrelevant.*

2. (**6.4**) Why is it not meaningful to attach a sign to the coefficient of multiple correlation R, although we do so for the coefficient of simple correlation r_{12}?

   SOLUTION. *Coefficient of multiple correlation measures the strength of linear relationship among several variables. In a space of dimension more than 1, there are many directions, and not just negative or positive. Thus, R shows how strong the mutual relationship is, but does not indicate any direction.*

3. (**6.27**) In a small-scale regression study, the following data were obtained,

Y	$X1$	$X2$
42.0	7.0	33.0
33.0	4.0	41.0
75.0	16.0	7.0
28.0	3.0	49.0
91.0	21.0	5.0
55.0	8.0	31.0

   Assume the standard multiple regression model with independent normal error terms. Compute **b, e, H,** SSErr, R^2, $s^2_{\mathbf{b}}$, $\hat{Y}$ for $X_1 = 10, X_2 = 30$. You can do the computations using software or by hand, although it would be lengthy to do them by hand.

   SOLUTION. *These answers are based on the R code and output in the end of these solutions.*

$$\boldsymbol{b} = \begin{pmatrix} 33.93 \\ 2.78 \\ -0.25 \end{pmatrix}, \quad \boldsymbol{e} = \begin{pmatrix} -2.70 \\ -1.23 \\ -1.64 \\ -1.33 \\ -0.90 \\ 6.99 \end{pmatrix}, \quad H = \begin{pmatrix} 0.23 & 0.25 & 0.21 & 0.15 & -0.05 & 0.21 \\ 0.25 & 0.31 & 0.09 & 0.27 & -0.15 & 0.22 \\ 0.21 & 0.09 & 0.70 & -0.32 & 0.10 & 0.20 \\ 0.15 & 0.27 & -0.32 & 0.61 & 0.14 & 0.15 \\ -0.05 & -0.15 & 0.10 & 0.14 & 0.94 & 0.02 \\ 0.21 & 0.22 & 0.20 & 0.15 & 0.02 & 0.20 \end{pmatrix},$$

$$SSErr = 62.07, \qquad R^2 = 0.98, \qquad s^2_{\mathbf{b}} = \begin{pmatrix} 715.47 & -34.16 & -13.59 \\ -34.16 & 1.66 & 0.64 \\ -13.59 & 0.64 & 0.26 \end{pmatrix}, \qquad \widehat{Y} = 53.85.$$

4. (Computer project, **#6.5—#6.8**) Dataset "Brand preference" is available on our Blackboard, on http://statweb.lsu.edu/EXSTWeb/StatLab/DataSets/NKNWData/CH06PR05.txt, and here:

Y_i	64	73	61	76	72	80	71	83	83	89	86	93	88	95	94	100
X_{i1}	4	4	4	4	6	6	6	6	8	8	8	8	10	10	10	10
X_{i2}	2	4	2	4	2	4	2	4	2	4	2	4	2	4	2	4

   It was collected to study the relation between degree of brand liking (Y) and moisture content (X_1) and sweetness (X_2) of the product.

(a) Fit a regression model to these data and state the estimated regression function. Interpret b_1.

(b) Obtain residual plots. What information do they provide? Plot residuals against fitted values, against each predictor, and against the product of predictors.

(c) Test homoscedasticity.

(d) Conduct a formal lack of fit test.

(e) Test whether the proposed linear regression model is significant. What do the results of the ANOVA F-test imply about the slopes?

(f) Estimate both slopes simultaneously using the Bonferroni procedure with at least a 99 percent confidence level.

(g) Report R^2 and adjusted R^2. How are they interpreted here?

(h) Calculate the squared correlation coefficient between Y_i and $\hat{Y}_i$. Compare with part (g).

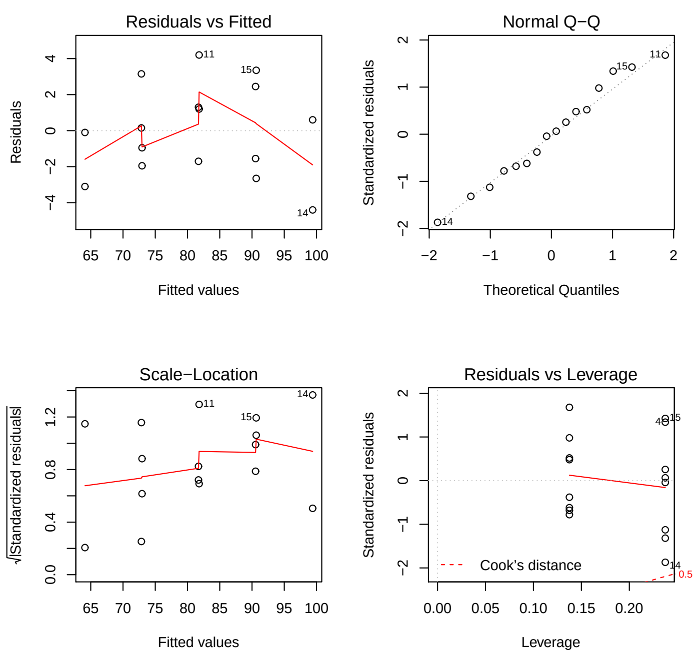
(i) Obtain a 99% confidence interval for $\mathbf{E}(Y)$ when $X_1 = 5$ and $X_2 = 4$. Interpret it.

(j) Obtain a 99% prediction interval for a new observation Y when $X_1 = 5$ and $X_2 = 4$. Interpret it.

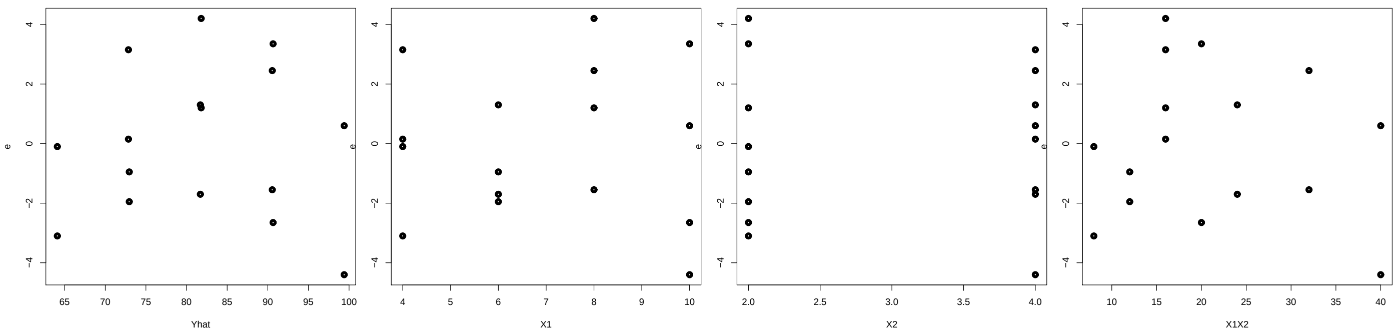
Solution. *These answers are based on the R code and output in the end of these solutions.*

(a) $\widehat{Y} = 37.65 + 4.425X_1 + 4.375X_2$. *The slope* $b_1 = 4.425$ *means that the brand liking is expected to increase by 4.425 when the product moisture content increases by 1 while sweetness is unchanged.*

(b) *Looking at the standard residual plots, there is some indication of a nonlinear trend; the Q-Q plot looks fairly straight, so probably, no problem with Normality; the variance of responses does not seem to change with the increase of their mean.*



Looking at residual e_i *plotted against fitted values* $\hat{Y}$, *predictors* X_1 *and* X_2, *and against the product of predictors* X_1X_2 *(the 4 plots below), there may be a concave nonlinear trend as a function of* X_1 *and no visible nonlinear relation with* X_2 *or* X_1X_2.



(c) *There is no significant evidence against the hypothesis of a constant variance $H_0 : \sigma^2 = const$, with the test statistic $\chi^2 = 0.626$ and p-value $p = 0.4288$.*

(d) *There is no significant evidence of a nonlinear trend, with the test statistic $F = 1.045$ and p-value $p = 0.384$.*

(e) *Significance of the whole model is tested by $H_0 : \beta_1 = \beta_2 = 0$ vs $H_1 : \beta_1 \neq 0$ or $\beta_2 \neq 0$. The ANOVA F-test shows that the model is significant, with the test statistic $F = 129.1$ and p-value $p = 2.66 \cdot 10^{-9}$. This means a significant evidence that at least one of the slopes is not 0.*

(f) *Using the Bonferroni adjustment for two simultaneous confidence intervals, the alpha level 0.01 is divided by 2. We obtain confidence intervals*

$$[3.41, 5.44] \text{ for } \beta_1 \quad \text{and} \quad [2.10, 6.65] \text{ for } \beta_2.$$

(g) *$R^2 = 0.9521$ is the proportion of the total variation SSTot explained by the two variables X_1 and X_2 combined. It measures goodness of fit, but it can only be used to compare models of the same rank p.*

$R^2_{adj} = 0.9447$ is the measure of a goodness of fit that can be used to compare models of different ranks, that is, different numbers of X-variables.

(h) *$r_{Y_i \hat{Y}_i} = 0.9521 = R^2$. Apparently, this is a general result, see exercise #5.*

(i) *A 99% confidence interval for $\mathbf{E}\{Y \mid X_1 = 6, X_2 = 4\}$ is $[73.88, 80.67]$. There is a 99% confidence that this interval covers the mean of responses with these values of X_1 and X_2. That is, in a long run of intervals computed from different samples, 99% of these intervals contain $\mathbf{E}\{Y \mid X_1 = 6, X_2 = 4\}$.*

(j) *A 99% prediction interval for Y when $X_1 = 6, X_2 = 4$ is $[68.48, 86.07]$. There is a 99% confidence that this interval covers the actual responses $X_1 = 6$ and $X_2 = 4$. That is, in a long run of intervals computed from different samples and random responses Y, 99% of these intervals will cover this response.*

5. (**# 6.26, Stat-615 only**) Show that the squared sample correlation coefficient between Y and $\hat{Y}$ equals R^2.
Remark. Now you can check if you did #3h correctly.
Hints. First, show that the sample averages of Y_i and $\widehat{Y}_i$ are the same. Then, write the sample correlation coefficient between Y and $\hat{Y}$ as

$$r_{Y\widehat{Y}} = \frac{\sum (Y_i - \overline{Y})(\widehat{Y}_i - \overline{Y})}{\sqrt{\sum (Y_i - \overline{Y})^2 \sum (\widehat{Y}_i - \overline{Y})^2}} = \frac{\sum (\widehat{Y}_i - \overline{Y} + Y_i - \widehat{Y}_i)(\widehat{Y}_i - \overline{Y})}{\sqrt{\sum (Y_i - \overline{Y})^2 \sum (\widehat{Y}_i - \overline{Y})^2}}$$

and use known properties of residuals $\sum e_i = 0$, $\sum X_{ij} e_i = 0$, $\sum \widehat{Y}_i e_i = 0$.

SOLUTION. *Following Hint 1, "First, show that the sample averages of Y_i and $\widehat{Y}_i$ are the same".*

We already know that $\sum e_i = \sum Y_i - \sum \widehat{Y}_i = 0$. *Therefore,* $\sum Y_i = \sum \widehat{Y}_i$*, and dividing by* n, $\overline{Y} = \overline{\widehat{Y}}$.
Following Hint 2, "write the sample correlation coefficient between Y *and* $\hat{Y}$ *as ...",*

$$
\begin{aligned}
r_{Y\widehat{Y}} &= \frac{\sum(Y_i - \overline{Y})(\widehat{Y}_i - \overline{Y})}{\sqrt{\sum(Y_i - \overline{Y})^2 \sum(\widehat{Y}_i - \overline{Y})^2}} = \frac{\sum(\widehat{Y}_i - \overline{Y} + Y_i - \widehat{Y}_i)(\widehat{Y}_i - \overline{Y})}{\sqrt{\sum(Y_i - \overline{Y})^2 \sum(\widehat{Y}_i - \overline{Y})^2}} \\
&= \frac{\sum(\widehat{Y}_i - \overline{Y})^2 + \sum(Y_i - \widehat{Y}_i)(\widehat{Y}_i - \overline{Y})}{\sqrt{\sum(Y_i - \overline{Y})^2 \sum(\widehat{Y}_i - \overline{Y})^2}} \\
&= \frac{SSReg + \sum e_i(\widehat{Y}_i - \overline{Y})}{\sqrt{SSTot \cdot SSReg}} = \frac{SSReg + \sum \widehat{Y}_i e_i - \overline{Y}\sum e_i}{\sqrt{SSTot \cdot SSReg}} = \frac{SSReg + 0 - 0}{\sqrt{SSTot \cdot SSReg}} = \sqrt{\frac{SSReg}{SSTot}} = \sqrt{R^2}
\end{aligned}
$$

So, $r^2_{Y\widehat{Y}} = R^2$.

R Code and Output for Problem #3

```
 # Enter the data
> Y = c(42,33,75,28,91,55)
> X1 = c(7,4,16,3,21,8)
> X2 = c(33,41,7,49,5,31)
> install.packages("matlib")
> library(matlib)

 # Define the design matrix X
> X = matrix(c(1,1,1,1,1,1,X1,X2),6,3)
> X
     [,1] [,2] [,3]
[1,]    1    7   33
[2,]    1    4   41
[3,]    1   16    7
[4,]    1    3   49
[5,]    1   21    5
[6,]    1    8   31

 # Compute the regression slope b
> b = inv(t(X) %*% X) %*% t(X) %*% Y
> b
           [,1]
[1,] 33.9321020
[2,]  2.7847707
[3,] -0.2643979

 # Fitted values, residuals, and error sum of squares
> Yhat = X %*% b
> e = Y - Yhat
> e
            [,1]
[1,] -2.70036663
[2,] -1.23087135
```

```
[3,] -1.63764825
[4,] -1.33091751
[5,] -0.09029763
[6,]  6.98606687
> SSErr = sum(e^2)
> SSErr
         [,1]
[1,] 62.07354

 # Hat matrix
> H = X%*%inv(t(X)%*%X)%*%t(X)
> H
            [,1]        [,2]        [,3]       [,4]        [,5]       [,6]
[1,]  0.23143639  0.25168006  0.21178834  0.1488734 -0.05475455 0.21099418
[2,]  0.25168006  0.31240977  0.09437951  0.2662835 -0.14787196 0.22314063
[3,]  0.21178834  0.09437951  0.70442097 -0.3191731  0.10446756 0.20412257
[4,]  0.14887339  0.26628346 -0.31917314  0.6142637  0.14143589 0.14834214
[5,] -0.05475455 -0.14787196  0.10446756  0.1414359  0.94040059 0.01632796
[6,]  0.21099418  0.22314063  0.20412257  0.1483421  0.01632796 0.19708945

 # Compute $R^2$
> SSTot = sum((Y - mean(Y))^2)
> SSReg = SSTot - SSErr
> Rsq = SSReg/SSTot
> Rsq
          [,1]
[1,] 0.9797938

 # Estimate VAR(b)
> s2 = SSErr/(6-3); # Estimated Var(Y)
> sb2 = s2*inv(t(X)%*%X); # Estimated VAR(b)
> sb2
          [,1]        [,2]        [,3]
[1,] 715.47117 -34.1589184 -13.5949378
[2,] -34.15892   1.6616665   0.6440674
[3,] -13.59494   0.6440674   0.2624678

 # Prediction for the given $X_1$ and $X_2$
> X0 = c(1,10,30)
> Y0hat = X0%*%b
> Y0hat
         [,1]
[1,] 53.84787
```

R Code and Output for Problem #4

```
 # Enter the data and rename variables

> attach(A)
> Y=V1; X1=V2; X2=V3;
```

```
 # Least squares estimation of regression slopes
> reg = lm( Y ~ X1 + X2 )
> reg
(Intercept)           X1           X2
     37.650        4.425        4.375

 # Residual plots
> par(mfrow=c(2,2))
> plot(reg)
> e = residuals(reg); Yhat = predict(reg); X1X2 = X1*X2;
> par(mfrow=c(1,1))
> plot(Yhat,e,lwd=5)
> plot(X1,e,lwd=5)
> plot(X2,e,lwd=5)
> plot(X1X2,e,lwd=5)

 # Testing for constant variance
> install.packages("car")
> library("car")
> ncvTest(reg)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 0.6261627, Df = 1, p = 0.42877

 # Lack of fit test
> full.model = lm( Y ~ as.factor(X1) + as.factor(X2) )
> anova(reg, full.model)
Model 1: Y ~ X1 + X2
Model 2: Y ~ as.factor(X1) + as.factor(X2)
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1     13 94.30
2     11 79.25  2     15.05 1.0445 0.3843

 # ANOVA F-test, R-square, and adjusted R-square
> summary(reg)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  37.6500     2.9961  12.566 1.20e-08 ***
X1            4.4250     0.3011  14.695 1.78e-09 ***
X2            4.3750     0.6733   6.498 2.01e-05 ***
---
Residual standard error: 2.693 on 13 degrees of freedom
Multiple R-squared:  0.9521,    Adjusted R-squared:  0.9447
F-statistic: 129.1 on 2 and 13 DF,  p-value: 2.658e-09

 # Prediction. Confidence and prediction intervals
> confint(reg, level=0.995)
               0.25 %   99.75 %
(Intercept) 27.545738 47.754262
X1           3.409483  5.440517
X2           2.104236  6.645764

> (cor(Y,Yhat))^2
```

```
[1] 0.952059

> predict(reg, data.frame(X1=5,X2=4), interval="confidence", level=0.99)
     fit      lwr      upr
1 77.275 73.88111 80.66889

> predict(reg, data.frame(X1=5,X2=4), interval="prediction", level=0.99)
     fit      lwr      upr
1 77.275 68.48077 86.06923
```