STAT 420 Fall 2015

Practice Problems 2a

1. Listed below are the price quotations of used cars along with their age and odometer mileage. A multiple linear regression analysis was performed using R, the output is given below.

	Age (years) X1	Mileage (thousand miles) X2	Price (thousand dollars) Y
1	1	8.1	9.45
2	2	17	8.4
3	2	12.6	8.6
4	3	18.4	6.8
5	3	19.5	6.5
6	4	29.2	5.6
7	6	40.4	4.75
8	7	51.6	3.89
9	8	62.6	2.7
10	10	80.1	1.47

```
> autos.fit <- lm(Y ~ X1 + X2)
> summary(autos.fit)

Call:
lm(formula = Y ~ X1 + X2)

Residuals:
    Min      1Q  Median      3Q     Max
-0.7390 -0.2545  0.1114  0.3066  0.5674

Coefficients:
            Estimate Std. Error t value  Pr(>|t|)
(Intercept)  9.98543    0.34289  29.121  1.45e-08 ***
X1          -1.38474    [ 7 ]    [ 8 ]   [      ]
X2           0.06481    [ 9 ]    [ 10 ]  [      ]
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

Residual standard error: [ 1 ] on [ 2 ] degrees of freedom
Multiple R-Squared: [ 3 ],      Adjusted R-squared: [ 4 ]
F-statistic: [ 5 ] on [ 6₁ ] and [ 6₂ ] DF,  p-value: [      ]
```

1. (continued)

```
> X <- cbind(c(rep(1,10)), X1, X2)
> X
      [,1] [,2]  [,3]
 [1,]    1    1   8.1
 [2,]    1    2  17.0
 [3,]    1    2  12.6
 [4,]    1    3  18.4
 [5,]    1    3  19.5
 [6,]    1    4  29.2
 [7,]    1    6  40.4
 [8,]    1    7  51.6
 [9,]    1    8  62.6
[10,]    1   10  80.1

> solve(t(X) %*% X)
            [,1]       [,2]        [,3]
[1,]  0.47166883 -0.3716589  0.03940979
[2,] -0.37165893  0.9238623 -0.11422997
[3,]  0.03940979 -0.1142300  0.01431659

> sum(autos.fit$residuals^2)                    # SSResid
[1] 1.744894

> sum((Y-mean(Y))^2)                            # SYY
[1] 62.55944
```

a) Fill in ① and ②.

b) Fill in ③ and ④.

c) Fill in ⑤ and ⑥. Is regression significant at a 1% level of significance?

d) Fill in ⑦ and ⑧. Test $H_0: \beta_1 = 0$ vs. $H_a: \beta_1 \neq 0$ at a 5% level of significance.

e) Fill in ⑨ and ⑩. Test $H_0: \beta_2 = 0$ vs. $H_a: \beta_2 \neq 0$ at a 5% level of significance.

f) Test $H_0: \beta_1 = -2$ vs. $H_a: \beta_1 > -2$ at a 5% level of significance.

2. Consider the population of high school graduates who were admitted to a particular university during a ten-year time period and who completed at least the first year of coursework after being admitted. We are interested in investigating how well Y, the first year grade point average (GPA), can be predicted by using the following quantities with $n = 20$ students:

X_1 = the score on the mathematics part of the SAT (SATmath)
X_2 = the score on the verbal part of the SAT (SATverbal)
X_3 = the grade point average of all high school mathematics courses (HSmath)
X_4 = the grade point average of all high school English courses (HSenglish)

Consider the model of the form:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \varepsilon_i, \quad i = 1, 2, \dots, 20,$$

where ε_i's are independent $N(0, \sigma^2)$ random variables.

```
> fit = lm(GPA ~ SATmath + SATverbal + HSmath + HSenglish)
> summary(fit)

Call:
lm(formula = GPA ~ SATmath + SATverbal + HSmath + HSenglish)

Residuals:
      Min        1Q    Median        3Q       Max
-0.443283 -0.128374  0.002571  0.133996  0.538996

Coefficients:
               Estimate   Std. Error   t value   Pr(>|t|)
(Intercept)   0.1615496    0.4375321     0.369    0.71712
SATmath       0.0020102    0.0005844     3.439    0.00365  **
SATverbal     0.0012522    0.0005515     2.270    0.03835  *
HSmath        0.1894402    0.0918680     2.062    0.05697  .
HSenglish     0.0875637    0.1764963     0.496    0.62700
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2685 on 15 degrees of freedom
Multiple R-squared: 0.8528,     Adjusted R-squared: ________
F-statistic: 21.72 on 4 and 15 DF,  p-value: 4.255e-06
```

2. (continued)

a) Find the value of the Adjusted *R*-squared.

b) If the Backward Elimination variable selection procedure and $\alpha_{crit} = 0.05$ is used, can the model be improved? If so, how (what is the next step)? Explain.

c) If the AIC model selection criteria is used, can the model be improved? If so, how (what is the next step)? Explain.

```
> drop1(fit)
Single term deletions

Model:
GPA ~ SATmath + SATverbal + HSmath + HSenglish
          Df Sum of Sq     RSS      AIC
<none>                   1.081  -48.348
SATmath    1     0.853   1.934  -38.718
SATverbal  1     0.372   1.453  -44.440
HSmath     1     0.307   1.388  -45.356
HSenglish  1     0.018   1.099  -50.022
```

3. Suppose the interaction model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$$

was fit to $n = 14$ data points, and the following results were obtained:

$SYY = \sum (y - \bar{y})^2 = 100$ $SSResid = \sum (y - \hat{y})^2 = 30$

```
Coefficients:
                  Estimate     Std. Error
(Intercept)             30              6
x1                      -4              2
x2                       3              2
x1x2                     7            2.5
```

3. (continued)

a) Is the model adequate for predicting Y? That is, perform the significance of the regression test at $\alpha = 0.05$.

b) Find the values of Multiple *R*-squared and Adjusted *R*-squared.

c) Do x_1 and x_2 interact? That is, test $H_0: \beta_3 = 0$ vs $H_1: \beta_3 \neq 0$. Use $\alpha = 0.05$.

d) Estimate the change in $E(Y)$ for every 1-unit increase in x_1, when $x_2 = 5$.

e) Estimate the change in $E(Y)$ for every 1-unit increase in x_2, when $x_1 = 2$.

4. Suppose the model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$$

was fit to $n = 20$ data points, R command `drop1` was applied, and the following results were obtained:

```
> fit = lm(Y ~ X1 + X2 + X3 + X4)
> drop1(fit)
Single term deletions

Model:
Y ~ X1 + X2 + X3 + X4
             Df   Sum of Sq      RSS     AIC
<none>                         25.681   _____
X1            1       5.685    _____   _____
X2            1       7.293    _____   _____
X3            1       1.316    _____   _____
X4            1       8.984    _____   _____
```

a) Fill in the missing AIC values.

b) If the AIC variable selection criteria is used, can the model be improved? If so, how (what is the next step)? Explain.

5. Consider the following data set:

x	0	2	4	6	8
y	110	123	119	86	62

a) Construct a scatter plot. Does the plot suggest that a linear relationship is appropriate?

Consider the model

$$\mathrm{Y}_i = \beta_0 + \beta_1 x_i + e_i, \qquad i = 1, 2, 3, 4, 5,$$

where e_i's are i.i.d. $\mathrm{N}(0, \sigma^2)$.

$\Sigma x = 20$, $\Sigma y = 500$, $\Sigma x^2 = 120$, $\Sigma y^2 = 52{,}630$, $\Sigma x y = 1{,}734$,

$\Sigma (x - \overline{x})^2 = 40$, $\Sigma (y - \overline{y})^2 = 2{,}630$, $\Sigma (x - \overline{x})(y - \overline{y}) = \Sigma (x - \overline{x}) y = -266$.

OR

$$\mathbf{X}^{\mathrm{T}}\mathbf{X} = \begin{bmatrix} 5 & 20 \\ 20 & 120 \end{bmatrix} \qquad (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} = \begin{bmatrix} 0.6 & -0.1 \\ -0.1 & 0.025 \end{bmatrix} \qquad \mathbf{X}^{\mathrm{T}}\mathrm{Y} = \begin{bmatrix} 500 \\ 1{,}734 \end{bmatrix}$$

b) Find the equation of the least-squares regression line. Add the regression line to the scatter plot.

$\Sigma (y - \hat{y})^2 = 861.1$.

c) What proportion of the observed variation in y values is explained by a straight-line relationship with x?

d) Is regression significant at a 5% level of significance?

e) Test $\mathrm{H}_0: \beta_1 = 0$ vs. $\mathrm{H}_1: \beta_1 < 0$ at a 5% level of significance.

f) Test $\mathrm{H}_0: \beta_0 = 100$ vs. $\mathrm{H}_1: \beta_0 > 100$ at a 5% and at a 10% level of significance.

g) Construct a 90% prediction interval for the future value of y corresponding to $x = 10$.

4. (continued)

Consider the model

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + e_i, \qquad i = 1, 2, 3, 4, 5,$$

where e_i's are i.i.d. $N(0, \sigma^2)$.

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} 5 & 20 & 120 \\ 20 & 120 & 800 \\ 120 & 800 & 5{,}664 \end{bmatrix} \qquad (\mathbf{X}^T\mathbf{X})^{-1} = \begin{bmatrix} 0.8857 & -0.3857 & 0.0357 \\ -0.3857 & 0.3107 & -0.0357 \\ 0.0357 & -0.0357 & 0.0045 \end{bmatrix}$$

$$\mathbf{X}^T\mathbf{Y} = \begin{bmatrix} 500 \\ 1{,}734 \\ 9{,}460 \end{bmatrix}, \qquad \hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat\beta_0 \\ \hat\beta_1 \\ \hat\beta_2 \end{bmatrix} = \begin{bmatrix} 111.8857 \\ 8.0643 \\ -1.8393 \end{bmatrix}.$$

h) Add the quadratic regression curve to the scatter plot. Does a quadratic regression function provide a better fit to the data than a straight line does?

$\Sigma(y - \hat{y})^2 = 103.3143.$

i) What proportion of the observed variation in y values is explained by a quadratic relationship with x?

j) Is regression significant at a 5% level of significance?

k) Test $H_0: \beta_2 = 0$ vs. $H_1: \beta_2 \neq 0$ at a 5% level of significance. Is β_2 significant at a 5% level of significance?

l) Test $H_0: \beta_0 = 100$ vs. $H_1: \beta_0 > 100$ at a 10% level of significance.

m) Construct a 90% prediction interval for the future value of y corresponding to $x = 10$.

n) Find the values of $R^2_{Adjusted}$ for both models. Which model is "better"?

o) Find the values of the Akaike's Information Criterion (AIC) for both models. Which model is "better"?

Answers:

1. a) ②. $= n - p = 10 - 3 = \mathbf{7}$.

$$s^2 = \frac{\text{SSResid}}{n-p} = \frac{1.744894}{7} = 0.24927.$$

① $= s = \sqrt{0.24927} = \mathbf{0.49927}$.

b) ③ $= R^2 = 1 - \frac{\text{SSResid}}{\text{SYY}} = 1 - \frac{1.744894}{62.55944} = \mathbf{0.9721}$.

④ $= R^2_{Adjusted} = 1 - \left(\frac{n-1}{n-p}\right) \cdot \left(1 - R^2\right) = 1 - \frac{9}{7} \cdot (1 - 0.9721) = \mathbf{0.96414}$.

c)

Source	*SS*	*df*	*MS*	*F*
Regression	60.814546	⑥$_1$ = 2	30.407273	⑤ = **121.985**
Residuals	1.744894	⑥$_2$ = 7	0.24927	
Total	62.55944	9		

$F_{0.01}(2, 7) = 9.55$. **Reject** $H_0: \beta_1 = \beta_2 = 0$ at $\alpha = 0.01$.

d) $H_0: \beta_1 = 0$ vs. $H_a: \beta_1 \neq 0$.

$\hat{\text{Var}}(\hat{\beta}_1) = 0.24927 \times 0.9238623 = 0.2303$.

⑦ $= \text{S.E.}(\hat{\beta}_1) = \sqrt{0.2303} = \mathbf{0.47989}$.

Test Statistic: ⑧ $= t = \frac{-1.38474 - 0}{0.47989} = \mathbf{-2.8855}$.

Rejection Region: $t < -t_{0.025}(7) = -2.365$ or $t > t_{0.025}(7) = 2.365$.

Reject H_0 at $\alpha = 0.05$.

e) $H_0: \beta_2 = 0$ vs. $H_a: \beta_2 \neq 0$.

$\hat{\text{Var}}(\hat{\beta}_2) = 0.24927 \times 0.01431659 = 0.0035687$.

⑨ $= \text{S.E.}(\hat{\beta}_2) = \sqrt{0.0035687} = \mathbf{0.05974}$.

Test Statistic: ⑩ $= t = \frac{0.06481 - 0}{0.05974} = \mathbf{1.0849}$.

Rejection Region: $t < -t_{0.025}(7) = -2.365$ or $t > t_{0.025}(7) = 2.365$.

Do NOT Reject H_0 at $\alpha = 0.05$.

f) $H_0: \beta_1 = -2$ vs. $H_a: \beta_1 > -2$.

Test Statistic: $t = \dfrac{-1.38474 - (-2)}{0.47989} = \mathbf{1.282}$.

Rejection Region: $t > t_{0.05}(7) = 1.895$.

Do NOT Reject H_0 at $\alpha = 0.05$.

```
> autos.dat
   X1   X2    Y
1   1  8.1 9.45
2   2 17.0 8.40
3   2 12.6 8.60
4   3 18.4 6.80
5   3 19.5 6.50
6   4 29.2 5.60
7   6 40.4 4.75
8   7 51.6 3.89
9   8 62.6 2.70
10 10 80.1 1.47

> autos.fit = lm(Y ~ X1 + X2, data=autos.dat)

> summary(autos.fit)

Call:
lm(formula = Y ~ X1 + X2, data = autos.dat)

Residuals:
    Min       1Q  Median      3Q     Max
-0.7390 -0.2545  0.1114  0.3066  0.5674

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.98543    0.34289  29.121 1.45e-08 ***
X1          -1.38474    0.47989  -2.886   0.0235 *
X2           0.06481    0.05974   1.085   0.3139
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

Residual standard error: 0.4993 on 7 degrees of freedom
Multiple R-Squared: 0.9721,     Adjusted R-squared: 0.9641
F-statistic:   122 on 2 and 7 DF,  p-value: 3.624e-06
```

2.

a) Adjusted R-squared = $1 - \frac{n-1}{n-p} \cdot \left(1 - R^2\right) = 1 - \frac{19}{15} \cdot (1 - 0.8528) \approx \mathbf{0.813547}$.

b) We remove the predictor (does not include `(Intercept)`) with highest p-value greater than α_{crit}. Therefore, the next step is to **remove `X4` `(HSenglish)`**:

$$\texttt{GPA} = \beta_0 + \beta_1\,\texttt{SATmath} + \beta_2\,\texttt{SATverbal} + \beta_3\,\texttt{HSmath} + \varepsilon.$$

We will then refit the model and remove the remaining least significant predictor provided its p-value is greater than α_{crit}.

c)

```
> drop1(GPA.fit)
Single term deletions

Model:
GPA ~ SATmath + SATverbal + HSmath + HSenglish
          Df Sum of Sq    RSS     AIC
<none>                  1.081 -48.348
SATmath    1     0.853  1.934 -38.718
SATverbal  1     0.372  1.453 -44.440
HSmath     1     0.307  1.388 -45.356
HSenglish  1     0.018  1.099 -50.022      <-- lowest
```

If the AIC model selection criteria is used, can the model be improved?
If so, how? Explain.

Want a model with lowest AIC value. Therefore, we can improve the model by **dropping `X4` `(HSenglish)`**.

$$\texttt{GPA} = \beta_0 + \beta_1\,\texttt{SATmath} + \beta_2\,\texttt{SATverbal} + \beta_3\,\texttt{HSmath} + \varepsilon.$$

3. $n = 14$ $p = 4$

a)

Source	*SS*	*DF*	*MS*	*F*
Regression	70	3	23.33333	7.77778
Residuals	30	10	3	
Total	100	13		

$\mathrm{F}_{0.05}(3, 10) = 3.71$

Reject $\mathbf{H_0}$ at $\alpha = 0.05$

b) $\mathrm{R}^2 = 1 - \frac{30}{100} = \mathbf{0.70}$

$\mathrm{R}^2_{Adj} = 1 - \frac{13}{10} \cdot (1 - 0.70) = \mathbf{0.61}$

c) $\mathrm{t} = \frac{7}{2.5} = 2.8$

$\pm t_{0.025}(10) = \pm 2.228$

Reject $\mathbf{H_0}$ at $\alpha = 0.05$

d) $-4 + 7 \cdot 5 = \mathbf{31}$

e) $3 + 7 \cdot 2 = \mathbf{17}$

4. $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$ $n = 20$

```
> fit = lm(Y ~ X1 + X2 + X3 + X4)
> drop1(fit)
Single term deletions

Model:
Y ~ X1 + X2 + X3 + X4
```

	Df	Sum of Sq		RSS	AIC
<none>				25.681	____
X1	1	5.685	25.681+5.685 =	**31.366**	____
X2	1	7.293	25.681+7.293 =	**32.974**	____
X3	1	1.316	25.681+1.316 =	**26.997**	____
X4	1	8.984	25.681+8.984 =	**34.665**	____

a) R: $\text{AIC} = n \ln\left({}^{\text{RSS}}\!/_{n} \right) + 2p.$

<none> 25.681 _____

None of the variables have been dropped. Y ~ X1 + X2 + X3 + X4

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$$

$p = 5.$

$$\text{AIC} = 20 \ln\left({}^{25.681}\!/_{20} \right) + 2 \times 5 = \mathbf{15}.$$

X1 1 5.685 **31.366** _____

X1 has been dropped. Y ~ X2 + X3 + X4

$$Y = \beta_0 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$$

$p = 4.$

$$\text{AIC} = 20 \ln\left({}^{31.366}\!/_{20} \right) + 2 \times 4 = \mathbf{17}.$$

X2 1 7.293 **32.974** _____

X2 has been dropped. Y ~ X1 + X3 + X4

$$Y = \beta_0 + \beta_1 x_1 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$$

$p = 4.$

$$\text{AIC} = 20 \ln\left({}^{32.974}\!/_{20}\right) + 2 \times 4 = \mathbf{18}.$$

```
X3                1           1.316    26.997    _____
```

`X3` has been dropped. `Y ~ X1 + X2 + X4`

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_4 x_4 + \varepsilon$$

$p = 4.$

$$\text{AIC} = 20 \ln\left({}^{26.997}\!/_{20}\right) + 2 \times 4 = \mathbf{14}.$$

```
X4                1           8.984    34.665    _____
```

`X4` has been dropped. `Y ~ X1 + X2 + X3`

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

$p = 4.$

$$\text{AIC} = 20 \ln\left({}^{34.665}\!/_{20}\right) + 2 \times 4 = \mathbf{19}.$$

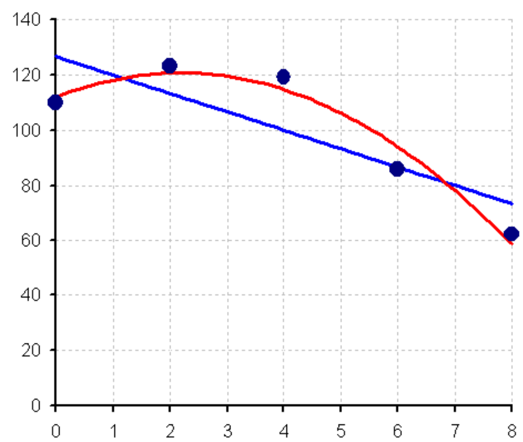
```
> drop1(fit)
Single term deletions

Model:
Y ~ X1 + X2 + X3 + X4
               Df   Sum of Sq      RSS    AIC
<none>                          25.681     15
X1              1       5.685   31.366     17
X2              1       7.293   32.974     18
X3              1       1.316   26.997     14     <- lowest
X4              1       8.984   34.665     19
```

b) Want a model with **lowest** AIC value. Therefore, we can improve the model by **dropping X3**:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_4 x_4 + \varepsilon$$

5.

a)
h)



b) $\hat{y} = 126.6 - 6.65\,x$

c) $R^2 = 0.6726$.

d) $H_0: \beta_1 = 0$ vs. $H_1: \beta_1 \neq 0$

$T = -2.4825$ 3 d.f.

OR

$F = 6.1627$ $(1, 3)$ d.f.

Accept H_0 at $\alpha = 0.05$.

e) $H_0: \beta_1 = 0$ vs. $H_1: \beta_1 < 0$ $T = -2.4825$ 3 d.f.

Reject H_0 at $\alpha = 0.05$.

f) $H_0: \beta_0 = 100$ vs. $H_1: \beta_0 > 100$ $T = 2.027$ 3 d.f.

Accept H_0 at $\alpha = 0.05$. Reject H_0 at $\alpha = 0.10$.

g) $60.1 \pm 2.353 \cdot 16.942 \cdot \sqrt{1 + \frac{1}{5} + \frac{(10-4)^2}{40}}$ 60.1 ± 57.77

i) $R^2 = 0.9607$.

j) $H_0: \beta_1 = \beta_2 = 0$ vs. $H_1:$ not H_0

$F = 24.4563$ $(2, 2)$ d.f. Reject H_0 at $\alpha = 0.05$.

k) $H_0: \beta_2 = 0$ vs. $H_1: \beta_2 \neq 0$

$T = -3.81488$ $(-3.83008$ without rounding$)$ 2 d.f.

Accept H_0 at $\alpha = 0.05$.

l) $H_0 : \beta_0 = 100$ vs. $H_1 : \beta_0 > 100$

$T = 1.757$ 2 d.f. Accept H_0 at $\alpha = 0.10$.

m) $\mathbf{X_0} = (1, 10, 100)^T$

$8.6 \pm 2.920 \cdot 7.1873 \cdot \sqrt{1 + 4.9817}$ 8.6 ± 51.3287

Without rounding $(\mathbf{X}^T \mathbf{X})^{-1}$:

$8.6 \pm 2.920 \cdot 7.1873 \cdot \sqrt{1 + 4.6}$ 8.6 ± 49.664

n)

	linear	**quadratic**
$R^2_{Adjusted}$	0.56345	**0.92143**

o)

	linear	**quadratic**
AIC	43.933	**35.331**
AIC (R)	29.744	**21.142**