04b Sample Examination Problems Chapter 13
SOLUTIONS

1. (a) Write down a sum of squares identity for a multiple regression model, and show how it implies that the solution of the least squares equations is a least squares estimator.

Model : p explanatory variables , p > 1

$$Y_i = \alpha + \beta_1 x_{i1} + \beta_1 x_{i2} + ... + \beta_1 x_{ip} + \varepsilon_i$$

Sum of squares identity :

$$\sum_1^n \varepsilon_i^2 = \sum_{i=1}^n (Y_i - \alpha - \beta_1 x_{i1} - \beta_1 x_{i2} - ..... - \beta_1 x_{ip})^2$$

$$= \sum_{i=1}^n (Y_i - \hat{Y}_i + \hat{Y}_i - \alpha - \beta_1 x_{i1} - \beta_1 x_{i2} - ..... - \beta_1 x_{ip})^2$$

$$\sum_{i=1}^n (Y_i - \hat{Y}_i)^2 + \sum_{i=1}^n (\hat{Y}_i - \alpha - \beta_1 x_{i1} - ... - \beta_1 x_{ip})^2 + 2\sum_{i=1}^n (Y_i - \hat{Y}_i)(\hat{Y}_i - \alpha - \beta_1 x_{i1} - ... - \beta_1 x_{ip})^2$$

$$= \sum_1^n \hat{\varepsilon}_i^2 + \sum_{i=1}^n (\hat{Y}_i - \alpha - \beta_1 x_{i1} - ... - \beta_1 x_{ip})^2 + 2(\hat{Y}_i - \alpha - \beta_1 x_{i1} - ... - \beta_1 x_{ip})^2 \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

$$= 0 + \sum_{i=1}^n (\hat{Y}_i - \alpha - \beta_1 x_{i1} - ... - \beta_1 x_{ip})^2 + 0$$

$$= \sum_{i=1}^n (\hat{Y}_i - \alpha - \beta_1 x_{i1} - ... - \beta_1 x_{ip})^2$$

(b) The following output is from a regression of record times in hours for Scottish Hill Races on the explanatory variables of distance run in miles and height climbed in feet. These were discussed by A.C. Atkinson in a paper in 'Statistical Science' in 1986.

Coefficients:

	Value	Std. Error	t value	Pr(>|t|)
(Intercept)	-8.9920	4.3027	-2.0898	0.0447
dist	6.2180	0.6011	10.3435	0.0000
climb	0.0110	0.0021	5.3869	0.0000

Residual standard error: 14.68 on 32 degrees of freedom Multiple R-Squared: 0.9191 F-statistic: 181.7 on 2 and 32 degrees of freedom, the p-value is 0

Analysis of Variance Table

Response: time

Terms added sequentially (first to last)

	Df	Sum of Sq	Mean Sq	F Value	Pr(F)
dist	1	71996.89	71996.89	334.2926	000000e+000
climb	1	6249.74	6249.74	29.0185	6.445183e-006
Residuals	32	6891.87	215.37		

i. What is the fitted model? Interpret the model.
ii. What is the estimated value of the record time in hours for The Goatfell Hill Race which has distance 8.0 miles and height climbed 2866 feet?
iii. How would you interpret the value of R^2?
iv. What diagnostic plots would you suggest for these data?

i. **Model:** $Y_i = \alpha + \beta_1 x_1 + \beta_1 x_2 + \varepsilon_i$,

Y_i : record time(house) , x_1 = dist(mile) , x_2 = climb(ft)

From the coefficients table :

$\hat{Y}_i = -8.9920 + 6.2180 x_1 + 0.0110 x_2$

Interpretation: controlling for climb ,each additional mile run adds 6.2180 hours to record time. Controlling for distance,each additional foot climbed adds 0.0110 hours to record time

ii. $\widehat{Y}_i = -8.9920 + 6.2180(8.0) + 0.0110(2866) = 72.278$ **hours**

ii. **R^2 = 0.9191 : coefficient of determination, The measure of the explanatory power of the regression model.**

$R^2 = \frac{RSS}{TSS}$ **, RSS : Regression sum of squares**

TSS : total sum of squares

Since we have two explanatory variables :
RSS = SS(dist)+SS(climb)= 71996.89 + 6249.74 = 78246.63
TSS = RSS + Residual SS = **78246.63 + 6891.87 = 85138.5**

$R^2 = \frac{RSS}{TSS}$ **=** $\frac{78246.63}{85138.5}$ **= 0.9191**

R2 provides the proportion(percentage) of the variation in the response variable explained by the model the above model has explained 91.91 % of the variation in the recorded time.

iii. Suggested diagnostic plots:

- **Inspect outliers or non linearity by plotting :**

  <u>You are NOT required to do so</u>

  1. **The fitted values $\widehat{Y}_i$ against the responses Y_i**
  2. **Residuals ε_i against the fitted $\widehat{Y}_i$**
  3. **Residuals against each explanatory variable**
  4. **Normal plot to see whether residuals are approximately Normal.**