STATISTICS 608
Homework 608 S17 01
Due: *11:59 PM, May 31, 2017*

Question 1 [2+2+2=6 Marks]

A random variable X is normally distributed with mean μ and variance $\sigma^2 > 0$.

(i) Find the *density function* of the random variable $exp(X)$ and show a plot of it. (Note: $exp(X)$ is said to have a *lognormal* distribution.)

(ii) Is it true that $E[exp(X)] = exp(E[X])$, where the symbol E denotes expected value? Substantiate your response with an appropriate calculation.

(iii) Denote the *median* of an arbitrary random variable Y by $m[Y]$. Is it true that $m[exp(X)] = exp(m[X])$? Substantiate your response with an appropriate calculation.

Question 2 [2 Marks]

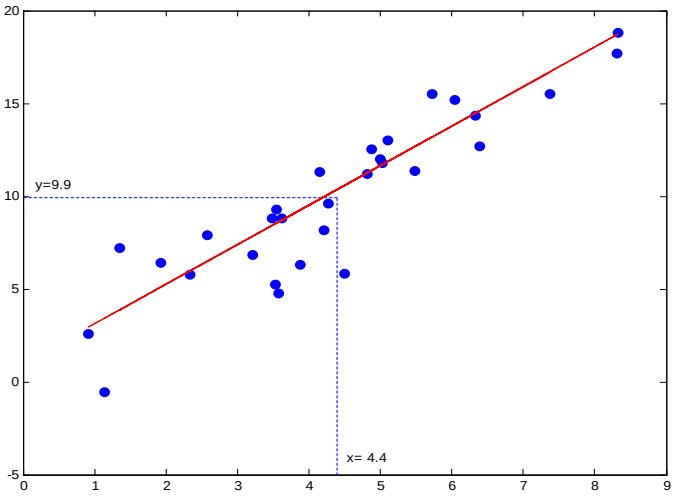
A packaging company obtained data on the size (X_1) of a lot and the cost (Y) of assembling the lot. A scatterplot of the data suggested a broken straight line regression with a breakpoint at lot size 250. The following linear model was formulated:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + e$$

where $X_2 = 1$ or 0 depending on whether the lot size was ≥ 250 or < 250 and where $X_3 = X_1 X_2$.
Which one of the following hypotheses is equivalent to the statement: "The two regression lines have the same *slope*."?
(a) $H_0 : \beta_0 = 0$ (b) $H_0 : \beta_1 = 0$ (c) $H_0 : \beta_2 = 0$ (d) $H_0 : \beta_3 = 0$.
Substantiate your answer by exhibiting appropriate algebraic manipulations.

Question 3 [2 marks]

The figure below shows a scatterplot of some data together with a line that purports to have been fitted by least squares. The averages of the x and y values are 4.4 and 9.9 respectively. The line in the figure cannot be the least squares line. Say why not AND provide a justification for your answer.



Question 4 [**2 marks**]

Show that

$$Var(y_i) = Var(e_i)$$

in the simple linear regression model.

Question 5 [**2+2+2+1=7 marks**] Work Exercise 4, page 40 in our textbook.

Question 6 [**2 marks**]

Show that the least squares criterion applied to the "intercept-only" model

$$y_i = \beta_0 + e_i, \ i = 1, \ldots, n$$

results in the least squares estimator $\hat{\beta}_0 = \bar{y}$ of β_0.

Question 7 [**2 marks**] Work Exercise 7, page 42 in our textbook.

Question 8 [**1+1+1+1+2+2+1+2+2+2+2=17 marks**]

This problem is based on Exercise 1 page 38 in our textbook. See that problem for a description of the data and the model.
The sales data are given in dollars. Give your answers below to 3 decimal places.

(a) The value of $R - Square$ is:

(b) The F value for testing $Ho : \beta_1 = 0$ is:

(c) What is your best estimate of β_0 ?

(d) What is your best estimate of β_1 ?

(e) The lower and upper confidence limits for a 95% confidence interval for β_1 are:

(f) Is $\beta_1 = 1$ a plausible value for β_1? Yes or No. *Give a reason to support your answer.*

(g) The estimated gross box office receipts for the current week for a production with \$ 369,000 in gross box office receipts the previous week is:

(h) A 95% **confidence interval** for the expected gross box office receipts for the current week for a production with \$ 369,000 in gross box office receipts the previous week is:

(i) A 95% **prediction interval** for the gross box office receipts for the current week for a production with \$ 369,000 in gross box office receipts the previous week is:

(j) Is \$ 497,000 a plausible value for the gross box office receipts for the current week for a production with \$ 369,000 in gross box office receipts the previous week? Yes or No. *Give a reason to support your answer.*

(k) Some promoters of Broadway plays use the prediction rule that next week's gross box office receipts will be equal to this week's gross box office receipts. *Comment on the appropriateness of this rule.*

2.8 Exercises

1. The web site www.playbill.com provides weekly reports on the box office ticket sales for plays on Broadway in New York. We shall consider the data for the week October 11–17, 2004 (referred to below as the current week). The data are in the form of the gross box office results for the current week and the gross box office results for the previous week (i.e., October 3–10, 2004). The data, plotted in Figure 2.6, are available on the book web site in the file playbill.csv.

   Fit the following model to the data: $Y = \beta_0 + \beta_1 x + e$ where Y is the gross box office results for the current week (in \$) and x is the gross box office results for the previous week (in \$). Complete the following tasks:

   (a) Find a 95% confidence interval for the slope of the regression model, β_1. Is 1 a plausible value for β_1? Give a reason to support your answer.
   (b) Test the null hypothesis $H_0 : \beta_0 = 10000$ against a two-sided alternative. Interpret your result.
   (c) Use the fitted regression model to estimate the gross box office results for the current week (in \$) for a production with \$400,000 in gross box office the previous week. Find a 95% prediction interval for the gross box office results for the current week (in \$) for a production with \$400,000 in gross box office the previous week. Is \$450,000 a feasible value for the gross box office results in the current week, for a production with \$400,000 in gross box office the previous week? Give a reason to support your answer.
   (d) Some promoters of Broadway plays use the prediction rule that next week's gross box office results will be equal to this week's gross box office results. Comment on the appropriateness of this rule.

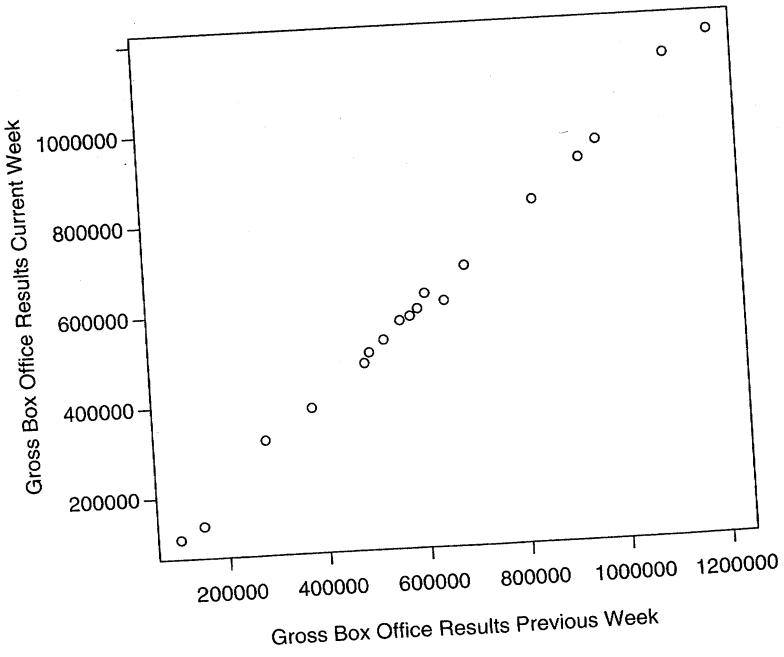


Figure 2.6 Scatter plot of gross box office results from Broadway

2. A story by James R. Hagerty entitled *With Buyers Sidelined, Home Prices Slide* published in the Thursday October 25, 2007 edition of the *Wall Street Journal* contained data on so-called fundamental housing indicators in major real estate markets across the US. The author argues that... *prices are generally falling and overdue loan payments are piling up*. Thus, we shall consider data presented in the article on

   Y = Percentage change in average price from July 2006 to July 2007 (based on the S&P/Case-Shiller national housing index); and
   x = Percentage of mortgage loans 30 days or more overdue in latest quarter (based on data from Equifax and Moody's).

   The data are available on the book web site in the file indicators.txt. Fit the following model to the data: $Y = \beta_0 + \beta_1 x + e$. Complete the following tasks:

   (a) Find a 95% confidence interval for the slope of the regression model, β_1. On the basis of this confidence interval decide whether there is evidence of a significant negative linear association.
   (b) Use the fitted regression model to estimate $E(Y|X{=}4)$. Find a 95% confidence interval for $E(Y|X{=}4)$. Is 0% a feasible value for $E(Y|X{=}4)$? Give a reason to support your answer.

3. The manager of the purchasing department of a large company would like to develop a regression model to predict the average amount of time it takes to process a given number of invoices. Over a 30-day period, data are collected on the number of invoices processed and the total time taken (in hours). The data are available on the book web site in the file invoices.txt. The following model was fit to the data: $Y = \beta_0 + \beta_1 x + e$ where Y is the processing time and x is the number of invoices. A plot of the data and the fitted model can be found in Figure 2.7. Utilizing the output from the fit of this model provided below, complete the following tasks.

   (a) Find a 95% confidence interval for the start-up time, i.e., β_0.
   (b) Suppose that a best practice benchmark for the average processing time for an additional invoice is 0.01 hours (or 0.6 minutes). Test the null hypothesis $H_0 : \beta_1 = 0.01$ against a two-sided alternative. Interpret your result.
   (c) Find a point estimate and a 95% prediction interval for the time taken to process 130 invoices.

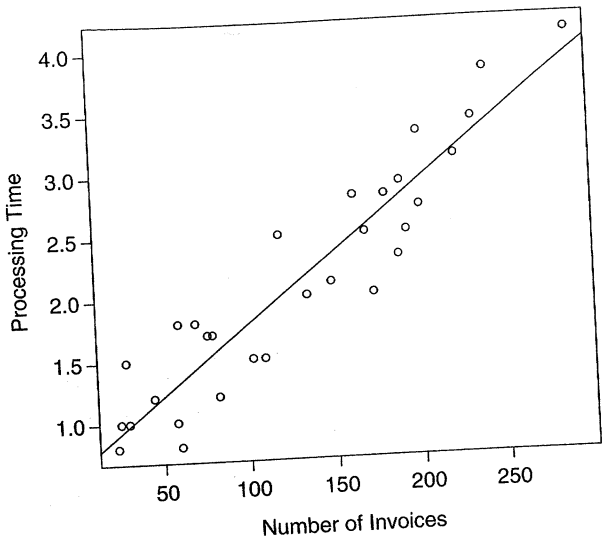


Figure 2.7 Scatter plot of the invoice data

Regression output from R for the invoice data

```
Call:
lm(formula = Time ~ Invoices)

Coefficients:
             Estimate  Std. Error  t value    Pr(>|t|)
(Intercept) 0.6417099   0.1222707    5.248    1.41e-05  ***
Invoices    0.0112916   0.0008184   13.797    5.17e-14  ***
---
Residual standard error: 0.3298 on 28 degrees of freedom
Multiple R-Squared: 0.8718, Adjusted R-squared: 0.8672
F-statistic: 190.4 on 1 and 28 DF, p-value: 5.175e-14

mean(Time)
2.1
median(Time)
2
mean(Invoices)
130.0
median(Invoices)
127.5
```

4. Straight-line regression through the origin:
In this question we shall make the following assumptions:

(1) Y is related to x by the simple linear regression model $Y_i = \beta x_i + e_i$ $(i = 1,2,...,n)$, i.e., $\mathrm{E}(Y \mid X = x_i) = \beta x_i$

(2) The errors $e_1, e_2, ..., e_n$ are independent of each other

(3) The errors $e_1, e_2, ..., e_n$ have a common variance σ^2

(4) The errors are normally distributed with a mean of 0 and variance σ^2 (especially when the sample size is small), i.e., $e \mid X \sim N(0, \sigma^2)$

In addition, since the regression model is conditional on X we can assume that the values of the predictor variable, $x_1, x_2, \ldots, x_n$ are known fixed constants.

(a) Show that the least squares estimate of β is given by

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

(b) Under the above assumptions show that

(i) $\mathrm{E}(\hat{\beta} \mid X) = \beta$

(ii) $\mathrm{Var}(\hat{\beta} \mid X) = \dfrac{\sigma^2}{\sum_{i=1}^{n} x_i^2}$

(iii) $\hat{\beta} \mid X \sim N\left(\beta, \dfrac{\sigma^2}{\sum_{i=1}^{n} x_i^2}\right)$

5. Two alternative straight line regression models have been proposed for Y. In the first model, Y is a linear function of x_1, while in the second model Y is a linear function of x_2. The plot in the first column of Figure 2.8 is that of Y against x_1, while the plot in the second column below is that of Y against x_2. These plots also show the least squares regression lines. In the following statements RSS stands for residual sum of squares, while SSreg stands for regression sum of squares. Which one of the following statements is true?

(a) RSS for model 1 is greater than RSS for model 2, while SSreg for model 1 is greater than SSreg for model 2.

(b) RSS for model 1 is less than RSS for model 2, while SSreg for model 1 is less than SSreg for model 2.

(c) RSS for model 1 is greater than RSS for model 2, while SSreg for model 1 is less than SSreg for model 2.

(d) RSS for model 1 is less than RSS for model 2, while SSreg for model 1 is greater than SSreg for model 2.

Give a detailed reason to support your choice.

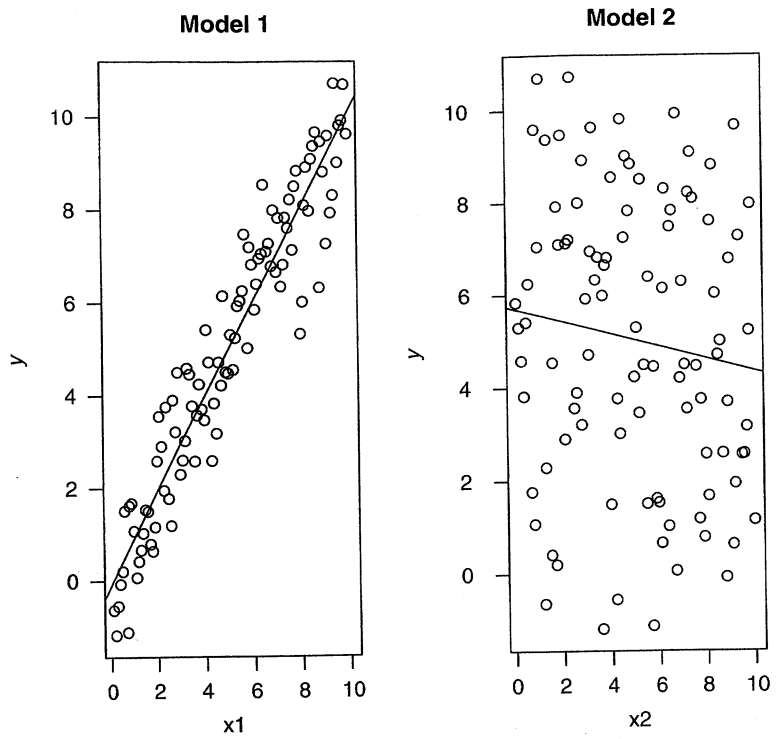


Figure 2.8 Scatter plots and least squares lines

6. In this problem we will show that SST=SSreg+RSS . To do this we will show that $\sum_{i=1}^{n}(y_i-\hat{y}_i)\,(\hat{y}_i-\overline{y})=0.$

   (a) Show that $(y_i-\hat{y}_i)=(y_i-\overline{y})-\hat{\beta}_1(x_i-\overline{x})$.

   (b) Show that $(\hat{y}_i-\overline{y})=\hat{\beta}_1(x_i-\overline{x})$.

   (c) Utilizing the fact that $\hat{\beta}_1=\dfrac{SXY}{SXX}$, show that $\sum_{i=1}^{n}(y_i-\hat{y}_i)\,(\hat{y}_i-\overline{y})=0.$

7. A statistics professor has been involved in a collaborative research project with two entomologists. The statistics part of the project involves fitting regression models to large data sets. Together they have written and submitted a manuscript to an entomology journal. The manuscript contains a number of scatter plots with each showing an estimated regression line (based on a valid model) and associated individual 95% confidence intervals for the regression function at each x value, as well as the observed data. A referee has asked the following question:

   I don't understand how 95% of the observations fall outside the 95% CI as depicted in the figures.

   Briefly explain how it is entirely possible that 95% of the observations fall outside the 95% CI as depicted in the figures.