Homework #2 (Solutions)
STAT-UB.0003: Regression and Forecasting Models

1. MBS, Ex. 7.34.

Solution:

(a) The null and alternative hypotheses are

$$H_0 : \mu = 85$$
$$H_a : \mu \neq 85,$$

where μ is the true mean Mach rating score of all purchasing managers.

(b) We reject the null hypothesis is the absolute value of the test statistic is larger than $z_{.050} = 1.645$. (It is also acceptable to give $t_{.050,121} \approx t_{.050,120} = 1.658$.)

(c) The test statistic is

$$\begin{aligned} t &= \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \\ &= \frac{(99.6) - (85)}{(12.6)/\sqrt{(122)}} \\ &= 12.80 \end{aligned}$$

(d) Since $|t| > 12.80$, we reject the null hypothesis.

2. MBS, Ex. 7.55, parts (a)-(e).

Solution:

(a) The null and alternative hypotheses are $H_0 : \mu = 2$ and $H_a : \mu \neq 2$, where μ is the mean surface roughness of coated interior pipe.

(b) The test statistic (from the Minitab output) is $t = -1.02$.

(c) We reject H_0 if $|t| > t_{.025,19} = 2.093$. You get full credit for this problem if you say $|t| > z_{.025} = 1.96$ or $|t| > 2$.

(d) We do not reject H_0.

(e) The p-value is $p = 0.322$. If the true mean surface roughness were 2 microns, then there would be a 32.2% chance of seeing data at least as extreme as observed. In other words, the data is consistent with the null hypothesis (it would be very typical if H_0 were true).

3. MBS, Ex. 11.7. Give one example where a probabilistic model is preferable, and one example where a deterministic model is preferable.

Solution:

Probabilistic models are generally preferable because they allow for approximate rather than exact linear relationships; they allow for deviations around the linear trend. One example where a deterministic relationship is more appropriate is a taxi fare:

$$\text{(total fare)} = \text{(flag drop fee)} + \beta_1\text{(number of miles driven)}$$

One example where a probabilistic model is more appropriate is a model relating a movie's box office gross to its advertising budget:

$$\text{(box office gross)} = \beta_0 + \beta_1\text{(advertising budget)} + \varepsilon;$$

there will not be an exact relationship between these quantities, so we need a probabilistic model

(Many other examples of deterministic and probabilistic linear models are valid.)

4. MBS, Ex. 11.9.

Solution: No, it does not. The random error will cause a deviation above or below the regression line (the mean).

5. MBS, Ex. 11.10.

Solution:

(a) Here is the completed table:

	x_i	y_i	x_i^2	x_iy_i
	7	2	49	14
	4	4	16	16
	6	2	36	12
	2	5	4	10
	1	7	1	7
	1	7	1	6
	3	6	9	18
Totals	$\sum x_i = 24$	$\sum y_i = 31$	$\sum x_i^2 = 116$	$\sum x_iy_i = 80$

(b)

$$\begin{aligned} SS_{xy} &= \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n} \\ &= (80) - \frac{(24)(31)}{7} \\ &= -26.28571. \end{aligned}$$

(c)

$$\begin{aligned} SS_{xx} &= \sum x_i^2 - \frac{(\sum x_i)^2}{n} \\ &= (116) - \frac{(24)^2}{7} \\ &= 33.71429 \end{aligned}$$

(d)

$$\begin{aligned} \hat{\beta}_1 &= \frac{SS_{xy}}{SS_{xx}} \\ &= \frac{(-26.28571)}{(33.71429)} \\ &= -0.77966 \end{aligned}$$

(e)

$$\begin{aligned} \bar{x} &= \frac{\sum x_i}{n} \\ &= \frac{(24)}{(7)} \\ &= 3.42857, \end{aligned}$$

$$\begin{aligned} \bar{y} &= \frac{\sum y_i}{n} \\ &= \frac{(31)}{(7)} \\ &= 4.42857 \end{aligned}$$

(f)

$$\begin{aligned} \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 - \bar{x} \\ &= (4.42857) - (-0.77966)(3.42857) \\ &= 7.10169 \end{aligned}$$

(g) The least squares line is

$$\begin{aligned} \hat{y} &= \hat{\beta}_0 + \hat{\beta}_1 + x \\ &= 7.10169 - 0.77966\,x \end{aligned}$$

6. MBS, Ex. 11.15. Here is the Minitab output from fitting the model described in the problem; use this output instead of the SPSS output given in the textbook:

```
Model Summary

      S    R-sq  R-sq(adj)  R-sq(pred)
11.9119  92.06%     91.89%      91.44%

Coefficients

Term        Coef  SE Coef  T-Value  P-Value   VIF
Constant   -97.4     26.7    -3.65    0.001
MATH2001  1.1882   0.0499    23.83    0.000  1.00

Regression Equation

MATH2011 = -97.4 + 1.1882 MATH2001
```

Solution:

(a) $E(Y \mid x) = \beta_0 + \beta_1 x$. (The book writes this as $E(y) = \beta_0 + \beta_1 x$).

You get full credit for a deterministic model ($y = \beta_0 + \beta_1 x$), but you should understand why a probabilisitic model is better.

(b) $\hat{y} = -97.4 + 1.1882x$

(c) No practical interpretation; it doesn't make sense to have an SAT score of 0.

(d) For every additional point on a state's avearge 2001 Math SAT score, a state's expected average Math 2011 SAT score increases by 1.1882 points. The interepretation is valid for the range of the x values in the data (between 474 and 603; this information is not given in the problem, but can be gotten by looking at the raw data).