Chapter 1 Problems:

1) 1.21

a) The estimated regression function $E(Y) = 10.2 + 4X$, where $Y$ is the number of broken and $X$ is the number of transfers. According to the figure below, the linear regression supplies a good fit, and there is tight fit of the data around the line.

[SAS output]

```
The REG Procedure
Model: MODEL1
Dependent Variable: ampule
```

```
Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > |t| |
|----------|----|--------------------|----------------|---------|-------|
| Intercept| 1  | 10.20000           | 0.66332        | 15.38   | <.0001|
| route    | 1  | 4.00000            | 0.46904        | 8.53    | <.0001|
```

b) This is the fitted value at $X = 1$, which is $14.2(=10.2 + 4 \times 1)$.

c) This is just $\beta_1$ which is estimated by 4.

d) In general, the fitted line is $\hat{Y} = b_0 + b_1X$, with $b_0 = \bar{Y} - b_1\bar{X}$. So the fitted value at $\bar{X}$ is $b_0 + b_1\bar{X} = \bar{Y} - b_1\bar{X} + b_1\bar{X} = \bar{Y}$. Thus, the point $(\bar{X}, \bar{Y})$ is on the fitted line. In this example, $\bar{X} = 1$ and $\bar{Y} = 14.2$. Since the fitted value at $\bar{X} = 1$ is $10.2 + 4 \times 1 = 14.2$, the fitted line $\bar{Y} = 10.2 + 4X$ goes through $(\bar{X}, \bar{Y}) = (1, 14.2)$.

2) 1.29

[Ans] The regression function, $Y_i = \beta_1X_i + \epsilon_i$, will be a line through the origin with slope $\beta_1$.

3) 1.30

[Ans] The expected value of $Y$ is the same for any $X$ and the regression function, $Y_i = \beta_0 + \epsilon_i$, will be a flat line at height $\beta_0$. 

---

1
4) Given \( n \) pairs of observations, \((X_i, Y_i)\) where \( i = 1, \ldots, n \), we consider the following simple linear regression model, \( Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \) where \( X_1, \ldots, X_n \) are some fixed numbers, and \( \epsilon_1, \ldots, \epsilon_n \) are uncorrelated random errors, each with mean 0 and variance \( \sigma^2 \).

As explained in class, the least squares estimator for \( \beta_0 \) and \( \beta_1 \) are

\[
b_1 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \quad \text{and} \quad b_0 = \bar{Y} - b_1 \bar{X}
\]

and the following three equations hold:

\[
\sum_{i=1}^{n} e_i = 0, \quad \sum_{i=1}^{n} X_i e_i = 0, \quad \sum_{i=1}^{n} \hat{Y}_i e_i = 0
\]

where \( \hat{Y}_i = b_0 + b_1 X_i \) and \( e_i = Y_i - \hat{Y}_i \). Discuss the implications of the three equations above.

**[Ans]** \( \sum_{i=1}^{n} e_i = 0 \) indicates that the residuals are randomly scattered around zero in a scatter plot of \( e_i \) versus \( i \). \( \sum_{i=1}^{n} X_i e_i = 0 \) (\( \sum_{i=1}^{n} \hat{Y}_i e_i = 0 \)) indicates that the entries of residual vector \( \mathbf{e} = (e_1, \ldots, e_n) \) and \( \mathbf{X} = (X_1, \ldots, X_n)(\mathbf{Y} = (Y_1, \ldots, Y_n)) \) are uncorrelated and thus there should be no specific pattern in the plot of \( e_i \) versus \( X_i(Y_i) \).