1. The OLS estimators are derived by choosing the values for $b_0, b_1, \ldots, b_K$ so that the sum of squared vertical distances between the regression line (or hyperplane) is minimized.

What is being described is a minimization problem solvable by calculus:

$$\min_{b_0,\ldots,b_K} \sum_{i=1}^{n} e_i^2$$

The formulas for $b_0, b_1, \ldots, b_K$ are found by taking $(K+1)$ derivatives, setting them equal to 0, and solving the system of equations.

2. $Var[b_1]$ is smaller when:

i) the sample size, n, is larger
ii) the variance of ϵ is smaller
iii) the variance of X is larger

Smaller variance in b_1 is good because it makes us more certain of our results (for example, narrower confidence intervals).

3. ϵ contains all of the omitted factors (or variables) that determine Y_i. ϵ is the random component of the model.

4. Two extreme (limiting) situations bound R^2 between 0 and 1. These are "no fit" and "perfect fit."

To explain this, you can draw diagrams, or talk about how ESS = 0 in "no fit" and ESS = TSS in "perfect fit."

5. R^2 should not be used in the multiple regression model because it always increases when a variable is added to the model. $\bar{R}^2$ should be used instead.

6. The "dummy variable trap" is when too many dummy variables are included in the regression model. For example, if the dummy variable M = 1 if individual is male (0 otherwise) and if F = 1 if individual is female (0 otherwise), then there is a perfect linear relationship between the two dummies:

$$F = 1 - M$$

Including both F and M in the regression model would be a situtation of perfect multicollinearity. OLS will not work; the estimator on one of the dummies is undefined.

7. a) $\bar{Y} = 1$, $\bar{X} = 0$

$$b_1 = \frac{\Sigma(Y_i - \bar{Y})(X_i - \bar{X})}{\Sigma(X_i - \bar{X})^2}$$

$$= \frac{(2-1)(0) + (0-1)(1) + (1-1)(-1)}{0^2 + 1^2 + (-1)^2}$$

$$= \frac{-1}{2}$$

$$b_0 = \bar{Y} - b_1\bar{X} = 1 - (-\tfrac{1}{2})0 = 1$$

b) $\hat{Y} = b_0 + b_1(1) = 1 - \frac{1}{2}(1) = \frac{1}{2}$

$$e = Y - \hat{Y} = 0 - \tfrac{1}{2} = -\tfrac{1}{2}$$

8. a) $b_0 = 12.48$
$b_1 = 2.24$

b) $H_0: \beta_1 = 0$
$H_A: \beta_1 \neq 0$

$$t = \frac{2.24}{0.36} = 6.22$$

Reject the null at 1% significance.
There is evidence of a difference in wages between men and women.

9. a) The estimated wage-gender gap is -2.33 (the coefficient on "genderfemale"). It is estimated that women earn $2.33 less than men per hour, on average.

b) This test has already been performed for us in R. The relevant t-stat is -6.007. We reject H_0.

c) R has also tested this already. The p-value is 0.201. We fail to reject the hypothesis that there is no difference in earnings between ~~men and women~~ married and single individuals.

d) $\bar{R}^2 = 0.2497$, so 24.97% of variation in "wage" can be explained using the regressors in the model.

e) $\widehat{wage} = -4.96 + 0.83(14) + 0.11(36) - 2.33(0) + 0.54(0)$

$= 10.62$

f) $R^2 = 1 - \frac{RSS}{TSS} = 1 - (1-\bar{R}^2)\frac{(n-k-1)}{n-1}$

$= 1 - (1-0.2497)\frac{(534-4-1)}{(534-1)}$

$= 0.2553$.

10.) In the first regression, the variable "Living.Area" is omitted. This is causing the estimated effect of "Fireplaces" on "Price" to be biased (much too large).

The problem is that Fireplaces and Living.Area are correlated, and that Living.Area is important in determining the Price. When Living.Area is omitted, its effect channels through Fireplaces.

When there are more Fireplaces, the house is larger! This is the main reason Price increases.

This is a sitcation of omitted variable bias. The omitted variable (Living.Area) is correlated with the included variable (Fireplaces). Furthermore, the omitted variable is an important determinant of Price.