Linear Statistical Models — Quiz 05 — Full Name ______________________________

1. In class, we fit a factor-numeric model to the housing data. Below is output from fitting that model, but only to data from zip codes 63105, 63122, and 63130.

```
Call: lm( log(sale_price) ~ zip_code + log(total_living_area) )

Coefficients:          Estimate  Std. Error   t value   Pr(>|t|)
(Intercept)              5.14        0.11     45.41     <2e-16 ***
zip_code63122           -0.22        0.01    -11.56     <2e-16 ***
zip_code63130           -0.42        0.01    -21.44     <2e-16 ***
log(total_living_area)  1.08        0.01     77.16     <2e-16 ***

 Residual standard error: 0.2307 on 1565 degrees of freedom
 Multiple R-squared:  0.8355,    Adjusted R-squared:  0.8352
 F-statistic:  2650 on 3 and 1565 DF,  p-value: < 2.2e-16
```

(i) (4) Write down the statistical model for y_i (logarithm of sale price) that corresponds to the output above. Define all of your notation.

$$Y_i = b_0 + b_{j(i)} + c_1 x_i + \varepsilon_i$$

$j(i)$ = levels for zip code

x_i = log total living area

$\varepsilon_i \overset{iid}{\sim} N(0, \theta^2)$

(ii) (2) How many observations (home sales) were used in this analysis?

$1565 + 4 = 1569$

(iii) (3) How do you interpret the (Intercept) coefficient?

Expected log sale price when log total living area is zero for zip code 63105

intercept for 63105.

(iv) (3) What is the intercept for 63122?

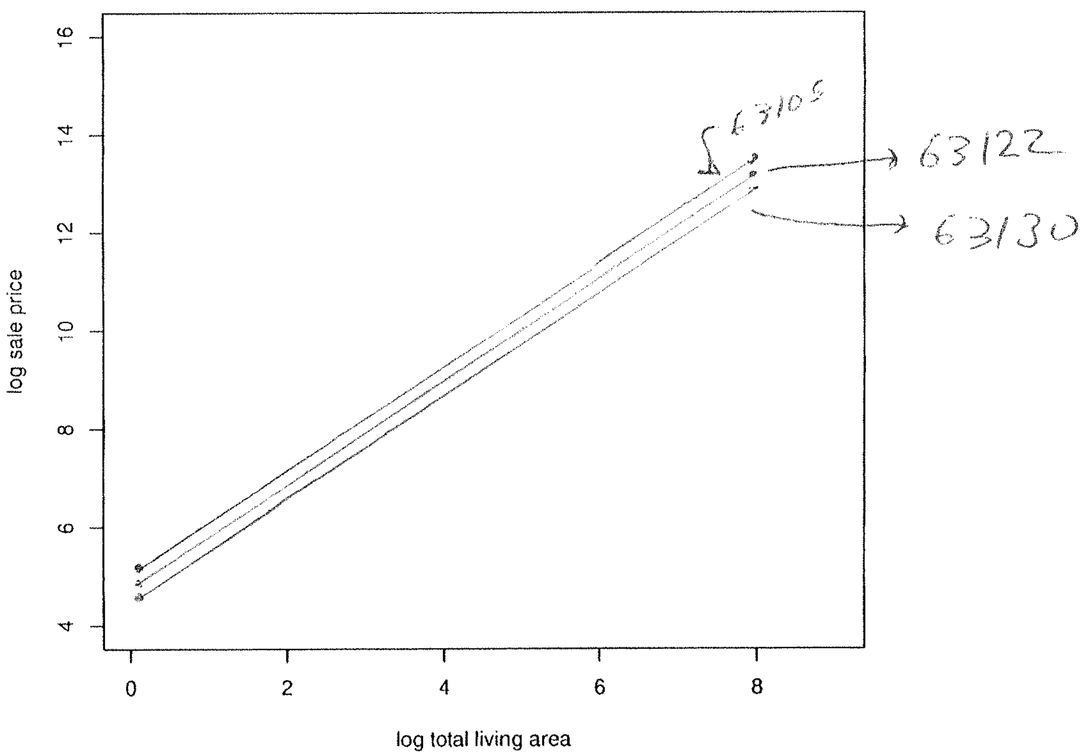
$$5.14 - 0.22 = 4.92$$

(v) (3) What is the estimated expected log sale price for a 3000 sq ft house in zip code 63122? log(3000) ≈ 8. You may leave your answer as a mathematical expression.

$$5.14 - 0.22 + 1.08(8)$$

$$4.92 + 8 + 0.64 = 13.56$$

(vi) (3) As best you can, draw and label the 3 regression lines for the 3 zip codes on the plot below.



(vii) (3) Write the definition of an additive model, and say how that definition applies to this model.

The effect of one variable does not depend on the value of the other. Here, the effect of log(total-living-area) (its slope) does not depend on zip code.

2. Below is a table of estimated expected values for a two-factor additive model

k

		1	2	3	4
	1	2	-4	-1	4
j	2	8	2	5	10
	3	6	0	3	8

(i) (3) Fill in the missing estimated expected values.

(ii) (3) Write down the two-factor additive model using mathematical notation. Define all terms.

$Y_i = b_0 + b_{j(i)} + c_{k(i)} + \varepsilon_i$

$j(i)$ = level of Factor 1

$k(i)$ = level of Factor 2

$\varepsilon_i \overset{ind}{\sim} N(0, \sigma^2)$

(iii) (3) Write out the parameter estimates for all of the regression coefficients, using R's convention for setting certain parameters to zero. (Hint: Check that your answers are consistent with the table.)

$\hat{b}_0 = 2$ $\hat{c}_1 = 0$

$\hat{b}_1 = 0$ $\hat{c}_2 = -6$

$\hat{b}_2 = 6$ $\hat{c}_3 = -3$

$\hat{b}_3 = 4$ $\hat{c}_4 = 2$

0

(2) Bonus, Problem 1. According to the model, if you have two houses in the same zip code, but one is twice the size of the other, do you expect the larger one to cost more than double, or less than double, and why?

$E(Y_i) = b_0 + b_{j(i)} + c_1 x_i$ where Y_i = log sale price

House 2 has double size of house 1

$$\frac{e^{x_2}}{e^{x_1}} = 2 \Rightarrow e^{x_2 - x_1} = 2 \Rightarrow x_2 - x_1 = \log 2$$

Interested in ratio of prices

$$\frac{e^{E(Y_2)}}{e^{E(Y_1)}} = \exp\left(E(Y_2) - E(Y_1)\right)$$

$$= \exp\left(b_0 + b_1 + c_1 x_2 - b_0 - b_1 - c_1 x_1\right)$$

$$= \exp\left(c_1 (x_2 - x_1)\right)$$

$$= \exp\left(c_1 \log 2\right)$$

$$= \left(e^{\log 2}\right)^{c_1} = 2^{c_1} = 2^{1.08} > 2 \checkmark$$

More than double.