Multiple Regression (Review)

1. We have a dataset measuring the price ($), size ($ft^2$), number of bedrooms, and age (years) of 518 houses in Easton, Pennsylvania. We fit a regression model to explain price in terms of the other variables.

```
The regression equation is
PRICE = 25875 + 39.2 SIZE - 1145 BEDROOM - 354 AGE

Predictor    Coef  SE Coef      T      P
Constant    25875     3555   7.28  0.000
SIZE       39.196    2.138  18.34  0.000
BEDROOM     -1145     1153  -0.99  0.321
AGE        -353.8    266.9  -1.33  0.186

S = 12612.2   R-Sq = 51.0%   R-Sq(adj) = 50.7%


Analysis of Variance

Source           DF           SS           MS       F      P
Regression        3  85029785549  28343261850  178.18  0.000
Residual Error  514  81760176401    159066491
Total           517  1.66790E+11
```

   (a) Interpret the estimated coefficient of Bedroom in the context of the fitted regression model.

   > **Solution:** In a regression model with Size, Bedroom and Age, holding hold Size and Age constant, if we increase Bedroom by 1, then mean Price *decreases* by $1145.

   (b) What does the result of the t test on the coefficient of Size indicate?

   > **Solution:** The coefficient is significant ($p < 0.001$). Size has the ability to explain Price beyond what is explained by Bedroom and Age.

   (c) What does the result of the t test on the coefficient of Bedroom indicate?

   > **Solution:** The coefficient is not significant ($p = 0.321$). Bedroom does not convey additional information in explaining Price Price beyond what is explained by Size and Age.

(d) What does the result of the F test indicate?

Solution: The test statistic is significant ($p < 0.001$). Thus, there is statistically significant evidence that the model is useful in explaining Price.

Multiple Regression with Qualitative Predictors

2. We asked 46 NYU students how much time they spend on social media, and what their primary computer is (Mac or PC). We are going to use regression to find out if one type of computer associated is with more social media usage. We have the response variable

$$\text{Social} = \text{amount of time (in minutes per week) using social media}$$

We would like to use "OS" as a predictor variable, which is a categorical (qualitative) variable taking values in the set $\{\text{Mac}, \text{PC}\}$.

(a) Why does the model $\text{Social} = \beta_0 + \beta_1 \text{OS} + \varepsilon$ not make sense?

Solution: The variable "OS" is categorical, not quantitative. It doesn't make sense to multiple the value of OS by a number.

(b) Give two different models to explain Social in terms of OS.

Solution: Define two dummy variables for OS:

$$\text{PC} = \begin{cases} 1 & \text{if OS} = \text{PC} \\ 0 & \text{otherwise;} \end{cases}$$

$$\text{Mac} = \begin{cases} 1 & \text{if OS} = \text{Mac} \\ 0 & \text{otherwise.} \end{cases}$$

There are two possible models:

$$\text{Social} = \beta_0 + \beta_1 \text{PC} + \varepsilon$$

or

$$\text{Social} = \beta_0 + \beta_1 \text{Mac} + \varepsilon$$

Both models are equivalent, though the interpretations of the coefficients β_0 and β_1 are different.

(c) Consider the model from part (b) involving the dummy variable "PC". What is the interpretation of β_0?

Solution: For the model $\text{Social} = \beta_0 + \beta_1 \text{PC} + \varepsilon$ The coefficient β_0 is equal to the mean social usage for Mac users.

(d) Again, consider the model from part (b) involving the dummy variable "PC". What is the interpretation of β_1?

Solution: For the model Social $= \beta_0 + \beta_1 \text{PC} + \varepsilon$ The mean social usage for Mac is β_0, and the mean social usage for PC is $\beta_0 + \beta_1$. Thus, β_1 represents the difference in the mean social usage between PC and Mac users.

3. Using the data from problem 2, we fit the regression model in Minitab, and got the following output.

```
The regression equation is
Social = 295 - 132 PC

Predictor     Coef  SE Coef      T      P
Constant    295.20    57.09   5.17  0.000
PC         -132.34    84.49  -1.57  0.124

S = 285.436   R-Sq = 5.3%   R-Sq(adj) = 3.1%
```

(a) What is the estimated mean social usage for Mac users?

Solution: $\hat{\beta}_0 = 294.20$ minutes per week.

(b) What is the estimated mean social usage for PC users?

Solution: $\hat{\beta}_0 + \hat{\beta}_1 = 294.20 - 132.34 = 161.86$ minutes per week.

(c) What is the interpretation of the p-value for the test on the coefficient of PC?

Solution: The p-value is for a hypothesis test of the following null and alternative:

$H_0 : \beta_1 = 0$ (the mean social usage is the same for Mac and PC users)
$H_a : \beta_1 \neq 0$ (the mean social usage is different for Mac and PC users)

Since the p-value is 0.124, which is greater than .05, we do not reject the null. There is not statistically significant evidence that the mean social usage is different for Mac and PC users.

4. We use the same data as in the previous problem, but now we are interested in whether or not texting behavior differs by cell phone type (Blackberry, iPhone, other smart phone, or standard cell phone).

   (a) Introduce dummy variables to encode cell phone type.

   Solution: We can encode cell phone type using four dummy variables

$$\text{Blackberry} = \begin{cases} 1 & \text{if Cell = Blackberry} \\ 0 & \text{otherwise;} \end{cases}$$

$$\text{iPhone} = \begin{cases} 1 & \text{if Cell = iPhone} \\ 0 & \text{otherwise;} \end{cases}$$

$$\text{Other} = \begin{cases} 1 & \text{if Cell = Other smart phone} \\ 0 & \text{otherwise;} \end{cases}$$

$$\text{Standard} = \begin{cases} 1 & \text{if Cell = Standard cell phone} \\ 0 & \text{otherwise.} \end{cases}$$

   (b) Using the variables you defined in part (a), devise a regression model which explains text usage in terms of cell phone type.

   Solution: We can choose to use any of the categories as the baseline. For example, if we choose "Standard" as the baseline, then the model is

$$\text{Text} = \beta_0 + \beta_1 \text{Blackberry} + \beta_2 \text{iPhone} + \beta_3 \text{Other} + \varepsilon.$$

   Different choices of the baseline category give different models (all are valid).

   (c) What is the interpretation of β_0, the intercept?

   Solution: The coefficient β_0 is the mean value of Text for the baseline category (Standard cell phone, in our case).

   (d) What are the interpretations of the other coefficients in your model?

   Solution: We first note that the mean value of Text for Blackberry owners is $\beta_0 + \beta_1$. Thus, β_1 is the difference in the mean value of Text between Blackberry owners and Standard cell phone owners. The meanings of β_2 and β_3 can be similarly derived.

5. We fit a model that explains Text in terms of cell phone type using dummy variables for cell phone type.

```
The regression equation is
Text = 132 + 91 Blackberry + 349 iPhone + 68 Smartphone

Predictor    Coef  SE Coef     T      P
Constant    131.7    316.9  0.42  0.680
Blackberry   90.8    501.0  0.18  0.857
iPhone      349.0    354.2  0.99  0.330
Smartphone   68.3    388.1  0.18  0.861

S = 776.121   R-Sq = 3.9%   R-Sq(adj) = 0.0%

Analysis of Variance

Source          DF        SS      MS     F      P
Regression       3   1025437  341812  0.57  0.640
Residual Error  42  25299274  602364
Total           45  26324711
```

(a) What is the estimated mean Text usage for people without smart phones?

Solution: $\hat{\beta}_0 = 131.7$.

(b) What is the estimated mean Text usage for people with iPhones?

Solution: $\hat{\beta}_0 + \hat{\beta}_2 = 131.7 + 349.0 = 480.7$.

(c) Is there statistically significant evidence that people with iPhones exhibit different texting behavior (volume) than people without smart phones?

Solution: We note that β_2 is equal to the difference in the mean value of Text between people with iPhones and people without smart phones. This question asks us to test the hypotheses

$$H_0 : \beta_2 = 0 \quad \text{(no difference in means)}$$
$$H_a : \beta_2 \neq 0$$

We use a t test on the coefficient of iPhone; the p-values is 0.330. Since this is above .05, there is not significant evidence of a difference (we do not reject H_0).

(d) Is cell phone type useful for predicting Text?

Solution: For this question, we are asked to test the hypotheses

$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ (cell phone type is useless for predicting Text)
$H_a : \beta_j \neq 0$ for some $j = 1, 2, 3$

We use an F test for this. The p-value is 0.640, which is above .05, so we do not reject the null. There is not significant evidence that cell phone type is useful for predicting Text.