Problem 1:

According to above deriving, the $\widehat{\beta_0}$ and $\widehat{\beta_1}$ can express:

$$\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{x}$$

$$\widehat{\beta_1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

From basic properties of the summation operator

$$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n}(x_i - \bar{x})y_i$$

And

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n}(x_i - \bar{x})x_i$$

So $\widehat{\beta_1}$ can rewrite as:

$$\widehat{\beta_1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})x_i}$$

(b) Proof $\mathrm{E}[\widehat{\beta_1}] = \beta_1$:

$$\mathrm{E}[\widehat{\beta_1}] = E[\frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})x_i}]$$

$\because \frac{\sum_{i=1}^{n}(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})x_i}$ is constant

$$\therefore \mathrm{E}[\widehat{\beta_1}] = \frac{\sum_{i=1}^{n}(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})x_i}E[y_i]$$

$$E[y_i] = E[\beta_0 + \beta_1 x_i] = \beta_0 + \beta_1 x_i$$

$$\mathrm{E}[\widehat{\beta_1}] = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\beta_0}{\sum_{i=1}^{n}(x_i - \bar{x})x_i} + \frac{\sum_{i=1}^{n}(x_i - \bar{x})x_i\beta_1}{\sum_{i=1}^{n}(x_i - \bar{x})x_i}$$

$$\mathrm{E}[\widehat{\beta_1}] = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\beta_0}{\sum_{i=1}^{n}(x_i - \bar{x})x_i} + \beta_1$$

$$\because \sum_{i=1}^{n}(x_i - \bar{x}) = 0$$

$$\therefore \mathrm{E}[\widehat{\beta_1}] = \beta_1$$

(a) Proof $\mathrm{E}[\widehat{\beta_0}] = \beta_0$:

$$\because y_i = \beta_0 + \beta_1 x_i + u_i$$

$$\therefore E[y_i] = E[\beta_0 + \beta_1 x_i + u_i] = E[\beta_0] + E[\beta_1 x_i] + E[u_i] = \beta_0 + \beta_1 x_i + 0$$

$$\because \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$\therefore E[\bar{y}] = E[\frac{1}{n}\sum_{i=1}^{n} y_i] = \frac{1}{n}E[\sum_{i=1}^{n} y_i] = \frac{1}{n}\sum_{i=1}^{n} \beta_0 + \beta_1 x_i = \beta_0 + \beta_1 \bar{x}$$

According to (0.9):

$$\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{x}$$

$$\therefore \mathrm{E}[\widehat{\beta_0}] = E[\bar{y} - \widehat{\beta_1}\bar{x}] = E[\bar{y}] - E[\widehat{\beta_1}\bar{x}] = E[\bar{y}] - \bar{x}E[\widehat{\beta_1}] = \beta_0 + \beta_1\bar{x} - \beta_1\bar{x} = \beta_0$$

$$\Longrightarrow \mathrm{E}[\widehat{\beta_0}] = \beta_0$$

Problem 2:

(a) $\widehat{\beta_0} = 32.1427, \widehat{\beta_1} = -0.3189$

(b) According to the formula:

$$math10 = 32.1427 - 0.3189 lnchprg$$

If increase the 10 percent in the number of students for the lunch program,

$$math10 = 32.1427 - 0.3189(lnchprg * 1.1)$$

$$\Rightarrow math10 = 32.1427 - 0.3508 lnchprg$$

the slop become bigger than origin. Now, calculate the percentage of the students passing the math exam.

$$\mathrm{math10}(\%) = \frac{32.1427 - 0.3508 lnchprg - (32.1427 - 0.3189 lnchprg)}{32.1427 - 0.3189 lnchprg} * 100$$

$$\Rightarrow \mathrm{math10}(\%) = \frac{-3.19 lnchprg}{32.1427 - 0.3189 lnchprg}$$

The percentage shows that the different lnchprg will generate different math10(%), and the overhigh lnchprg will cause math10(%) decreasing seriously, if the denominator is positive.

(c) The lnchprg means percentage of students in school lunch program, and the higher lnchprg shows that many students join in this program, this case also point out their maybe have a bad economic state cause poor learning environment. As a result, this situation indirectly effect pass rate.

(d) $math10 = -20.3607 + 6.2297 \log(expend) - 0.3046 lnchprg$

Sample size = 408

$R^2 = 0.1799$

The estimated slopes imply that the more expending increases the pass math rate, and the higher lnchprg decrease the pass math rate.

The expend means students' tuition, if the students' family can afford this expense, the student might not participle in school lunch program, and have a good learning environment. Thus, according to data, the higher expense and lower lnchprg will

have a favorable pass rate.

Problem 3:

(a)Generally, the data are unorganized raw records or objective facts, and lack of classification, they can't definitely represent meaning by itself. As a result, the data need more useful method to process them. With systematically organization and analysis, the data were organized orderly information, they show that more clear things than origin data, and easy to understand what they are mean. The information can assist experts or researchers to study, and deeply investigate them. The knowledge can generate more important items from organized information, it also indicated concrete meaning of this information. The people can learn knowledge to obtain and realize some things that they don't know. The wisdom is ability of creativity, thinking and judgment through learning large amount of knowledge, it can help us solve problem in real world. For example, the people can use fins, which is learned from leg of frog, to swim faster than without it.

Take air quality for examples, the PM2.5, PM10, CO, SO_2 and so on, recorded from every weather observatory. This data seems meaningless for machine, if we send these origin data to machine, but we all know air quality is important for human being. In order to understand relationship between them. First, we need collect to one month or year data of air quality indicators, and the physical condition of certain area people. Second, the raw data will appropriately organize and analyze into useful information, like, table, figure or something organized type. With above assistance, this information can apparently show more significant message, such as, what time the PM2.5 concentration is over high? The concentration distribution of CO in one day, the rate of people infected different disease. Next, the professional can research relationships between these indicators and human healthy, and obtain several results by comparing or analyzing, like PM2.5 not only endanger the respiratory tract of people, but also cause cancer and cardiovascular disease. Finally, when the people see the over-high PM2.5 concentration, they will know today's air quality is bad, and maybe stay at home or wear a mask go outside, etc.

(b)The supervised learning requires labeled data, clearly told to machine what is this, just like teacher teaches students. And then machine automatically learn feature between input and output data, to be more specifically, find a complex function can correspond them. After that, the trained model can predict it, which has been similar to previous items, based on training data. However, the unsupervised learning doesn't need labeled data, this technique automatically generates the results from input data by extracting the features and analyzing the structure. The goal of unsupervised learning is learning by itself, and can achieve learned without a teacher, this case also meets the situation of daily life.

Now for the most part, labeled data need a lot of time to label, and it is hard to obtain the dataset for we want to do. Thus, the supervised learning and unsupervised learning are depending on what task we want to solve. The different methods were used to handle different problem, for example, the supervised learning can deal with classification and regression problem, the unsupervised learning can solve cluster and anomaly detection problem.

In the daily life, the supervised learning and the unsupervised learning are also common, for instance, the supervised learning exists in the young period, the teacher and parent are teaching us, and told us what is this or what is true and false things. After gradually growing, we can naturally react these things, are similar to previous things. However, in the age of information exploration, we can't learn all of the things in teenager, so we may encounter things that we have never seen or met. The unsupervised learning will take an important part in our life, we must learn by myself, whatever surfing online or asking experienced people, this method can assist solve several problems.

Problem 4:

(a) This paper trained a convolutional neural network, including five convolution layers, some max-pooling layers, and three fully connect layers with a final softmax function. The network achieved more better performance than the previous state-of-the-art, to boost the network performance, the network adopts Rectified Linear Units (ReLUs) as activated function instead of tanh function, it also proved the ReLUs is better than tanh neurons, because ReLUs make network fast convergence and obtain high efficiency. To decrease training time, it simultaneously uses multi GPU calculation to reduce huge computations. The Local response normalization will refine and concentrate after ReLUs of certain layers, it can valid decrease error rate. In a large number of parameters, the model is easy overfitting, lead to bad performance, in response to this case, the paper adopts three method to improve them. First, the overlapping pooling, the network allowed pooling kernel overlapping to decrease the opportunity of overfitting. Second, data augmentation, according to exists data to create new data by adjusting color brightness, flipping horizontally and so on, provided data diversity to increase more learnable feature. Finally, the network add dropout in the first and second fully connected layer, the approach makes neurons deal randomly to avoid incorrect feature from over-learning. As a result, the network can achieve the superior performance than the state-of-the-art method, valid deal with overfitting and training time, and get the championship in the ImageNet LSVRC-2010 contest, at the same time, established a good foundation for deep learning.

(b) A well-defined problem makes us easier to solve, and save time, cost and resources, this process usually takes a lot of time to think. As a result, defining problem is important thing, having a right direction is better than work hard. After that, collecting

related data based on above defined problem, the quality of data is also directly determining the performance of the model, so appropriately gathering data is essential part in machine learning. After the data collection is complete, we need to process it before send data to model, this action can help machine more valid learning feature and convergence quickly, like, order randomization avoids machine from learning false feature or split data into training and testing to verify model performance. According to the data we collected, choosing a properly model to train, such as, consider the architecture of network, output label types, variable types and so on. The different task requires different method to handle, an appropriate model can effective boost performance, and get favorable predict results. Training model is extracting feature from input data, compared with predicted results and labeled output, calculate errors between them, and feedback to adjust model parameters, achieved the better performance. Evaluation is significant part in machine learning, it can truly reflect the model performance is good or bad, because training data is smaller part than the all cases, choosing the evaluation metric properly can exam whether the model is successfully trained well or not. If the model obtains poor performance, we will tune the parameters of the model or redesign model architecture to improve the model performance, and repeat the steps until the model can implement the best performance.