PROJECT I

part A) The assumption we have made is that there exists a continous function $f(x)$ and a normal distributed error $\varepsilon \sim N(0, \sigma^2)$ which describes our data

$$y = f(x) + \varepsilon$$

Function $f(x)$ is approximated with our model $\tilde{y}$ where we minimized $(y - \tilde{y})^2$, with

$$\tilde{y} = X\beta$$

Show that the expectation value of y for a given element i is

$$\mathbb{E}(y_i) = \sum_j x_{ij}\beta_j = X_{i,*}\beta$$

And that its variance is

$$\mathrm{Var}(y_i) = \sigma^2$$

Let us start the proof for the first element by introducing the expectation value for y.

$$\mathbb{E}[y] = \frac{1}{n}\sum_{i=0}^{p-1} y_i \Rightarrow \frac{1}{n}\sum_{i=0}^{p-1}(f(x_i) + \varepsilon_i)$$

$$f(x) \simeq X\beta = \tilde{y}$$

Our function f has been approximated through minimization of cost function $(y-\hat{y})^2$, also known as MSE.

$$c(\beta) = \frac{1}{n}\sum_{i=0}^{n-1}(y_i - \hat{y}_i)^2$$

As we see, the cost function is an extension of the mathematical expression for expectation of y. Now in order to find the optimal parameters β giving the best approximation of f, we have to compute the first derivative of $c(\beta)$ with respect to β.

$$\frac{\partial c(\beta)}{\partial \beta} = (X^T(y - X\beta))$$

Now in order to find the optimal parameters, we set the first derivative equal to zero.

$$\Rightarrow \quad X^T(y - X\beta) = 0$$

$$X^T y - X^T X\beta = 0$$

Now if we assume that X^TX is an invertible matrix, we have the solution : $\beta = (X^TX)^{-1}X^Ty$

$$\epsilon = y - \tilde{y} = y - X\beta$$

$$\Rightarrow X^T(y - X\beta) = 0 \quad \text{we have} \quad X^T\epsilon = X^T(y - X\beta) = 0$$

The usual assumption is to write the expected value of residuals error as equal to zero as its density distribution has a true mean of 0 according to earlier assumption.

I wish to elaborate further on the value of residuals in the proof. If we take a look at one of our earlier:

$$X^T e = X^T(y - X\beta) = 0$$

$$\Rightarrow X^T e = X^T y - X^T X\beta$$

Now inserting the OSL expected value for β, we get

$$\Rightarrow X^T e = X^T y - X^T X (X^T X)^{-1} X^T y = 0$$

$$= X^T y - X^T y = 0$$

Thus as we see, it's not only a simple expectation, but the value of the RSS, $\mathbb{E}(\varepsilon) = 0$. We can now continue on with our proof.

(In the final version of the report, remember to include the respective set that each of the variables included belongs to $\Rightarrow y \in \mathbb{R}^m$, $X \in \mathbb{R}^{n \times p}$, $\beta \in \mathbb{R}^p$, and the value of $\hat{\beta} = (X^T X)^{-1} X^T y$))

From earlier we have written that:

$$\mathbb{E}[y] = \frac{1}{n} \sum_{i=0}^{n-1} y_i$$

Now given X, each of its rows is symbolized by writing $X_{i,*}$. So the expectation value of y for a given element can be written:

$$\mathbb{E}(y_i) = \mathbb{E}[X_{i,*}\beta] + \mathbb{E}[\varepsilon_i]$$

$$= X_{i,*}\beta + 0 = X_{i,*}\beta \Leftrightarrow \sum_{j=0}^{p-1} X_{i,j}\beta_j$$

Continuing on, we now wish to prove that the variance of y for any given element i is $Var(y_i) = \sigma^2$. From the lecture notes and "Pattern Recognition and Machine Learning" by Christopher M. Bishop, variance can be written in the following way.

$$Var(y_i) = \mathbb{E}\left[[y_i - \mathbb{E}(y_i)]^2\right] = \mathbb{E}(y_i^2) - (\mathbb{E}(y_i))^2$$

$$= \mathbb{E}[(X_{i,*}\beta + \varepsilon_i)^2] - (X_{i,*}\beta)^2$$

$$= \mathbb{E}[(X_{i,*}\beta)^2 + 2\varepsilon_i X_{i,*}\beta + \varepsilon_i^2] - (X_{i,*}\beta)^2$$

$$= (X_{i,*}\beta)^2 + 2\mathbb{E}(\varepsilon_i)X_{i,*}\beta + \mathbb{E}(\varepsilon_i^2) - (X_{i,*}\beta)^2$$

If we now look back at the information we have available about the distribution of residual error, we see that $\mathbb{E}(\varepsilon_i) = 0$ and $\mathbb{E}(\varepsilon_i^2) = \sigma^2$

$$= \mathbb{E}(\varepsilon_i^2) = Var(\varepsilon_i) = \underline{\underline{\sigma^2}}$$

- Moving on, we now have to show that $\mathbb{E}(\hat{\beta}) = \beta$ using Ordinary Least Squares expression for the optimal parameters $\hat{\beta}$. From before, we have that

$$\frac{\partial C(\beta)}{\partial \beta} = X^T(y - X\beta)$$

$$\Rightarrow X^T y = X^T X \beta$$

- Assuming that the matrix $X^T X$ is invertible, we can write the expression for optimal OSL parameters:

$$\hat{\beta}_{OSL} = (X^T X)^{-1} X^T y$$

If the matrix X^TX causes any problems, by being near- or singular, we can rewrite it using Singular Value Decomposition. Back to the proof, we can now write:

$$\mathbb{E}(\hat{\beta}) = \mathbb{E}((X^TX)^{-1}X^Ty) = (X^TX^{-1})X^T\mathbb{E}(y)$$

$$= (X^TX)^{-1}X^TX\beta = \underline{\underline{\beta}}$$

Thus our model using ordinary least squares regression is unbiased. In other words, this simply means that the expected value is equal to the true value of the parameter. Having obtained a proof for $\mathbb{E}(\hat{\beta})$, we can now finish this part by showing that the variance of parameters β is

$$Var(\hat{\beta}) = \sigma^2(X^TX)^{-1}$$

$$\Rightarrow Var(\hat{\beta}) = \mathbb{E}\{[\beta - \mathbb{E}(\beta)]^2\} = \mathbb{E}\{[\beta - \mathbb{E}(\beta)][\beta - \mathbb{E}(\beta)]^T\}$$

$$= \mathbb{E}\{[(X^TX)^{-1}X^Ty - \beta][(X^TX)^{-1}X^Ty - \beta]^T\}$$

$$= (X^TX)^{-1}X^Tyy^TX(X^TX)^{-1,T} - (X^TX)^{-1}X^Ty\beta^T - \beta y^TX(X^TX)^{-1,T} + \beta\beta^T$$

Before continuing, we can pay attention to the fact that we can use $\mathbb{E}(yy^T) = X\beta\beta^TX^T + \sigma^2 I$.

$$= (X^TX)^{-1}X^T\mathbb{E}(yy^T)X(X^TX)^{-1,T} - (X^TX)^{-1}X^Ty\beta^T - \beta y^TX(X^TX)^{-1,T} + \beta\beta^T$$

$$= (X^TX)^{-1}X^T\mathbb{E}(yy^T)X(X^TX)^{-1} - (X^TX)^{-1}X^T\mathbb{E}(y)\beta^T - \beta\mathbb{E}(y^T)X(X^TX)^{-1} + \beta\beta^T$$

$$= (X^TX)^{-1}X^T\mathbb{E}(yy^T)X(X^TX)^{-1} - (X^TX)^{-1}X^TX\beta\beta^T - \beta\beta^TX^TX(X^TX)^{-1} + \beta\beta^T$$

$$= (X^TX)^{-1}X^T\mathbb{E}(yy^T)X(X^TX)^{-1} - \beta\beta^T$$

- The lecture notes significantly shorten further calculations. We will however write full calculations.

$$\rightarrow (X^TX)^{-1}X^T(X\beta\beta^TX^T + \sigma^2 I)X(X^TX)^{-1} - \beta\beta^2$$

$$= \left((X^TX)^{-1}X^TX\beta\beta^TX^T + (X^TX)^{-1}X^T\sigma^2\right)X(X^TX)^{-1} - \beta\beta^2$$

$$= (X^TX)^{-1}X^TX\beta\beta^TX^TX(X^TX)^{-1} + (X^TX)^{-1}X^T\sigma^2X(X^TX)^{-1} - \beta\beta^2$$

- $$= \beta\beta^T + \sigma^2(X^TX)^{-1} - \beta\beta^T = \underline{\sigma^2(X^TX)^{-1}}$$