CSE 544

PROBABILITY AND STATISTICS FOR DATA SCIENCE

ASSIGNMENT 6

Team members:

Name	SBU ID
Venkata Satwik Chilukuri	113263262
Akshay Somayaji	113322316
Shardul Churi	114403283

1(a) $D = \{x_1, x_2 \dots x_n\} \sim$ Normal (θ, σ^2), with known σ
Also, prior of $\theta \sim$ Normal (a, b^2). To find: Posterior of θ, i.e, $f(\theta|D)$.

→ we know, $f(\theta|D) \propto \mathcal{L}(D) \cdot f(\theta)$ (i)

Now, $\mathcal{L}(D) = f(D|\theta) = f(\{x_1 \dots x_n\}|\theta)$

$= \prod_{i=1}^{n} f(x_i|\theta)$ ($\because x_i$ are conditionally $\perp$ on θ)

$\Rightarrow \mathcal{L}(D) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-1}{2}\left(\frac{x_i-\theta}{\sigma}\right)^2} = \prod_{i=1}^{n} \frac{1}{(2\pi\sigma^2)^{1/2}} \cdot e^{\frac{-(x_i-\theta)^2}{2\sigma^2}}$

$\therefore \mathcal{L}(D) = \frac{1}{(2\pi\sigma^2)^{n/2}} \cdot e^{\frac{-\sum_i (x_i-\theta)^2}{2\sigma^2}}$ (ii)

→ prior of θ $f(\theta) \sim$ Normal (a, b^2)

$\Rightarrow f(\theta) = \frac{1}{(2\pi b^2)^{1/2}} \cdot e^{\frac{-(\theta-a)^2}{2b^2}}$ (iii)

Using (ii) & (iii) in (i), we get

$f(\theta|D) \propto \frac{1}{(2\pi\sigma^2)^{n/2} \cdot (2\pi b^2)^{1/2}} \cdot e^{\frac{-1}{2}\left(\frac{(\theta-a)^2}{b^2} + \sum_{i=1}^{n} \frac{(x_i-\theta)^2}{\sigma^2}\right)}$

↓

This constant term can be ignored, since we are checking for proportionality, so we need only consider terms involving θ.

$\Rightarrow f(\theta|D) \propto e^{\frac{-1}{2}\left(\frac{(\theta-a)^2}{b^2} + \sum_{i=1}^{n} \frac{(x_i-\theta)^2}{\sigma^2}\right)}$

$\Rightarrow f(\theta|D) \propto e^{-1/2\left(\frac{\theta^2+a^2-2a\theta}{b^2} + \sum_{i=1}^{n} \frac{x_i^2+\theta^2-2x_i\theta}{\sigma^2}\right)}$

$\Rightarrow f(\theta|D) \propto e^{\frac{-1}{2}\left(\frac{\theta^2+a^2-2a\theta}{b^2} + \frac{\sum x_i^2 + n\theta^2 - 2\theta\sum x_i}{\sigma^2}\right)}$

$\Rightarrow f(\theta|D) \propto e^{\frac{-1}{2}\left(\frac{\sigma^2\theta^2 + a^2\sigma^2 - 2a\sigma^2\theta + b^2\sum x_i^2 + nb^2\theta^2 - 2\theta b^2\sum x_i}{\sigma^2 b^2}\right)}$

$$\Rightarrow f(\theta|D) \propto e^{\frac{-1}{2}\left(\frac{a^2\sigma^2 + b^2\sum x_i^2}{b^2\sigma^2}\right)} \cdot e^{\frac{-1}{2}\left(\frac{\theta^2(\sigma^2+nb^2) - 2\theta(a\sigma^2 + b^2\sum x_i)}{b^2\sigma^2}\right)}$$

↳ Constant term, can be removed like before

$$\Rightarrow f(\theta|D) \propto e^{\frac{-1}{2}\left(\frac{\theta^2(\sigma^2+nb^2) - 2\theta(a\sigma^2 + nb^2\bar{x})}{b^2\sigma^2}\right)} \quad \left(\because \bar{x} = \sum_{i=1}^{n} x_i / n\right)$$

dividing by $\sigma^2 + nb^2$

$$\Rightarrow f(\theta|D) \propto e^{\frac{-1}{2}\left(\frac{\theta^2 - 2\theta\left(\frac{a\sigma^2 + b^2 n\bar{x}}{\sigma^2+nb^2}\right)}{\frac{b^2\sigma^2}{\sigma^2+nb^2}}\right)} \quad \text{(iv)}$$

→ Let us add and subtract $\left(\frac{a\sigma^2 + nb^2\bar{x}}{\sigma^2+nb^2}\right)^2$ term in the numerator (subtracted)

The negative term can be ignored as it is a constant term with respect to θ, as before.

Hence, the exponent numerator becomes -

$$\theta^2 - 2\theta\left(\frac{a\sigma^2 + nb^2\bar{x}}{\sigma^2+nb^2}\right) + \left(\frac{a\sigma^2 + nb^2\bar{x}}{\sigma^2+nb^2}\right)^2$$

$$= \left(\theta - \left(\frac{a\sigma^2 + nb^2\bar{x}}{\sigma^2+nb^2}\right)\right)^2$$

$$= \left(\theta - \left(\frac{n a se^2 + n b^2\bar{x}}{n se^2 + n b^2}\right)\right)^2 \quad \left(\because se^2 = \frac{\sigma^2}{n}\right)$$

$$= \left(\theta - \left(\frac{a.se^2 + b^2\bar{x}}{se^2 + b^2}\right)\right)^2 \quad //$$

→ Replacing above result in (iv), we get

$$f(\theta|D) \propto e^{\frac{-1}{2}\frac{\left(\theta - \left(\frac{a.se^2 + b^2\bar{x}}{se^2+b^2}\right)\right)^2}{\frac{b^2\sigma^2}{n}}} \quad \text{(v)}$$

$\frac{b^2\sigma^2}{n} = \frac{b^2 n.se^2}{n se^2 + n b^2}$

For some distribution $T \sim \text{Normal}(x, y^2)$, its pdf can be represented proportionally as

$$f_T(\theta) \propto e^{-\frac{1}{2}\frac{(\theta - x)^2}{y^2}} \quad \text{(vi)}$$

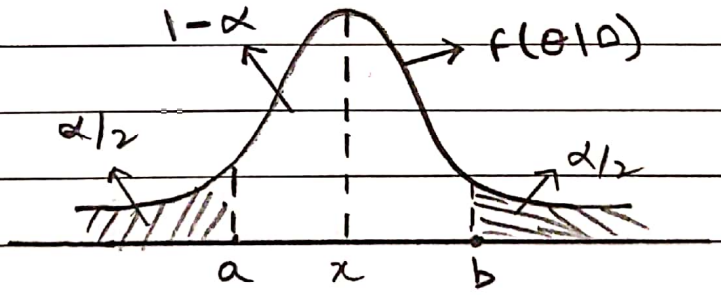
→ Comparing (v) and (vi), it is clear that

$$x = \frac{a \cdot se^2 + b^2 \bar{X}}{se^2 + b^2},$$

$$y = \frac{b^2 se^2}{se^2 + b^2}$$

b) To find - $1-\alpha$ posterior interval for θ.

→ we take 2 points (a, b) on the posterior dist^n curve, such that the area between them is $(1-\alpha)$. By symmetry, we have area to the left of a = area to the right of b = $\alpha/2$, as shown



→ From above graph, it is clear that

$$Pr(\theta < a \mid D) = Pr(\theta > b) = \alpha/2$$

$$\Rightarrow Pr(\theta \in [a, b] \mid D) = 1 - \left(\frac{\alpha}{2} + \frac{\alpha}{2}\right)$$

$$\therefore Pr(a \le \theta \le b \mid D) = 1 - \alpha \quad \text{(i)}$$

→ Now, $\theta|D \sim \text{Normal}(x, y^2)$. To convert it to standard normal, we use trans^n property

$$Z = \frac{\theta - x}{y}, \text{ where } x \text{ is mean \& } y \text{ is std-dev}$$

(ii)

→ Subtracting x & dividing by y in (i) inequality, we get $Pr\left(\frac{a-x}{y} \leq \frac{\theta-x}{y} \leq \frac{b-x}{y} \mid \theta\right) = 1-\alpha$

$\Rightarrow Pr\left(\frac{a-x}{y} \leq Z \leq \frac{b-x}{y}\right) = 1-\alpha$ (from (ii))

→ From standard normal properties, we know

$$Pr\left(-Z_{\alpha/2} \leq Z \leq Z_{\alpha/2}\right) = 1-\alpha$$

On comparing above 2 equations, we get

$$\frac{a-x}{y} = -Z_{\alpha/2}, \quad \frac{b-x}{y} = Z_{\alpha/2}$$

$$\Rightarrow a = x - Z_{\alpha/2} \cdot y, \quad b = x + Z_{\alpha/2} \cdot y$$

∴ the $(1-\alpha)$ posterior interval for θ is the interval $(x - Z_{\alpha/2} \cdot y, \; x + Z_{\alpha/2} \cdot y)$ =

2. Bayesian Inference in action

a) sigma = 3

- **Sequence of steps to calculate Posterior over 5 iterations**

Step 1: Prior_1 ~ Normal(0,1), Posterior_1 ~ Normal(4.590762, 0.082569)

Step 2: Prior_2 ~ Normal(4.590762, 0.082569), Posterior_2 ~ Normal(4.813524, 0.043062)

Step 3: Prior_3 ~ Normal(4.813524, 0.043062), Posterior_3 ~ Normal(4.921257, 0.029126)

Step 4: Prior_4 ~ Normal(4.921257, 0.029126), Posterior_4 ~ Normal(4.972837, 0.022005)

Step 5: Prior_5 ~ Normal(4.972837, 0.022005), Posterior_5 ~ Normal(4.983966, 0.017682)

Calculated Posterior Means and Variances

Step	Mean Estimate	Variance Estimate
1	4.590762	0.082569
2	4.813524	0.043062
3	4.921257	0.029126
4	4.972837	0.022005
5	4.983966	0.017682

- **Observations of graph (Refer ./graphs/Q2_a_sigma3.png)**

- With each iteration, that is, as we encounter more data, the width of the posterior curves are getting narrower, that is, its variance is reducing.
- The final posterior (Step 5) has been shifted away from the initial (Step 1) prior.

b) sigma = 100

- **Sequence of steps to calculate Posterior over 5 iterations**

Step 1: Prior_1 ~ Normal(0,1), Posterior_1 ~ Normal(0.058716, 0.990099)

Step 2: Prior_2 ~ Normal(0.058716, 0.990099), Posterior_2 ~ Normal(0.095009, 0.980392)

Step 3: Prior_3 ~ Normal(0.095009, 0.980392), Posterior_3 ~ Normal(0.138226, 0.970874)

Step 4: Prior_4 ~ Normal(0.138226, 0.970874), Posterior_4 ~ Normal(0.171219, 0.961538)

Step 5: Prior_5 ~ Normal(0.171219, 0.961538), Posterior_5 ~ Normal(0.218918, 0.952381)

Calculated Posterior Means and Variances

Step	Mean Estimate	Variance Estimate
1	0.058716	0.990099
2	0.095009	0.980392
3	0.138226	0.970874
4	0.171219	0.961538
5	0.218918	0.952381

- **<u>Observations of graph (Refer ./graphs/Q2_a_sigma100.png)</u>**

❖ The width of the posterior curves are almost constant across each iteration.

❖ The final posterior (Step 5) is still close to the initial (Step 1) prior.

c) Conclusion

- We can probably conclude that for small sigma (known standard deviation of X1…Xn data), the posterior moves away from the prior and the distribution gets fit tighter around the mean with each iteration (variance reduces).
- For large sigma, the posterior does not move far from the prior, and there is minimal reduction in the variance across each iteration.

Question 3 : Regression Analysis

(a) In order to find a good estimator, we try to minimise the error.

We make use of SSE or sum of square error and attempt to minimize it

$$S = \sum_{i=1}^{n} (\hat{\varepsilon}_i)^2$$

$$= \sum_{i=1}^{n} (y_i - \hat{y})^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

partially differentiating w.r.t $\hat{\beta}_0$

$$\frac{\partial S}{\partial \hat{\beta}_0} = \sum_{i=1}^{n} \frac{\partial}{\partial \hat{\beta}_0} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = 0$$

$$0 = \sum_{i=1}^{n} 2 \cdot (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) \cdot \frac{\partial}{\partial \hat{\beta}_0} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)$$

$$0 = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)$$

$$\therefore \sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{\beta}_0 + \sum_{i=1}^{n} \hat{\beta}_1 x_i$$

$$\therefore n\bar{y} = n \cdot \hat{\beta}_0 + n \cdot \hat{\beta}_1 \bar{x}$$

$$\therefore \bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$$

$$\therefore \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Partially differentiating w.r.t $\hat{\beta}_1$

$$\frac{\partial S}{\partial \hat{\beta}_1} = \sum_{i=1}^{n} \frac{\partial}{\partial \hat{\beta}_1} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = 0$$

$$0 = \sum_{i=1}^{n} 2 \cdot (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) \cdot \frac{\partial}{\partial \hat{\beta}_1} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)$$

(y_i → 0, $\hat{\beta}_0$ → 0, $\hat{\beta}_1$ → 1)

$$0 = \sum_{i=1}^{n} 2 x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)$$

$$0 = \sum_{i=1}^{n} x_i y_i - \hat{\beta}_0 \sum_{i=1}^{n} x_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2$$

Substituting value of $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

$$\therefore 0 = \sum_{i=1}^{n} x_i y_i - \bar{y} \sum_{i=1}^{n} x_i + \hat{\beta}_1 \bar{x} \sum_{i=1}^{n} x_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2$$

$$\hat{\beta}_1 n \cdot \bar{x}^2 - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = n \cdot \bar{x} \bar{y} - \sum_{i=1}^{n} x_i y_i$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n \bar{x} \bar{y}}{\sum_{i=1}^{n} x_i^2 - n \bar{x}^2}$$

$$= \frac{\sum_{i=1}^{n} x_i y_i - n \bar{x} \bar{y} - n \bar{x} \bar{y} + n \bar{x} \bar{y}}{\sum_{i=1}^{n} x_i^2 - 2n \bar{x}^2 + n \bar{x}^2}$$

$$= \frac{\sum_{i=1}^{n} x_i y_i - \bar{y} \sum_{i=1}^{n} x_i - \bar{x} \sum_{i=1}^{n} y_i + \sum_{i=1}^{n} \bar{x} \bar{y}}{\sum_{i=1}^{n} x_i^2 - 2 \bar{x} \sum x_i + n \cdot \bar{x}^2}$$

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

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

(b) $$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

To show, $\hat{\beta}_1$ and $\hat{\beta}_0$ are unbiased.

for an estimator to be unbiased, its expectation must be equal to its TRUE VALUE.

$$\therefore E(\hat{\beta}_1) = \beta_1 \; ; \quad E(\hat{\beta}_0) = \beta_0$$

Proving $\hat{\beta}_1$,

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

considering numerator only,

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

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

$$= \sum_{i=1}^{n} (x_i - \bar{x}) y_i - \bar{y} \sum_{i=1}^{n} (x_i - \bar{x}) \quad [\bar{y} \rightarrow \text{constant}]$$

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

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

Similarly, considering denominator only,

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

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

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

$$= \sum_{i=1}^{n} (x_i - \bar{x}) x_i - \bar{x} \sum_{i=1}^{n} (x_i - \bar{x}) \quad \left[\bar{x} \to \text{const}\right]$$

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

Now, computing $E(\hat{\beta}_1 | X)$,

$$E(\hat{\beta}_1 | X) = E\left(\frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \,\middle|\, X \right)$$

$$= \frac{E\left(\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \,|\, X \right)}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Rewriting numerator,

$$= \frac{E\left(\sum_{i=1}^{n} (x_i - \bar{x}) y_i \,|\, X \right)}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$= \frac{\sum_{i=1}^{n} (x_i - \bar{x}) E(y_i | X)}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$= \frac{\sum_{i=1}^{n} (x_i - \bar{x})(\beta_0 + \beta_1 x_i)}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$= \frac{\sum_{i=1}^{n} (x_i - \bar{x})\, \beta_0}{\sum_{i=1}^{n} (x_i - \bar{x})^2} + \frac{\sum_{i=1}^{n} (x_i - \bar{x})\, \beta_1 x_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$= \frac{\beta_0 \sum_{i=1}^{n} (x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} + \beta_1 \frac{\sum_{i=1}^{n} (x_i - \bar{x})\, x_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

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

* rewriting denominator,

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

$$= \beta_1$$

$\therefore E(\hat{\beta}_1 | x) = \beta_1$

$\therefore \hat{\beta}_1$ is unbiased.

Now, proving for $\hat{\beta}_0$,

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

$$E(\hat{\beta}_0 | x) = E(\bar{y} - \hat{\beta}_1 \bar{x} | x)$$
$$= E(\bar{y}|x) - E(\hat{\beta}_1 \bar{x} | x)$$
$$= \bar{y} - \bar{x}\, E(\hat{\beta}_1 | x)$$
$$= \bar{y} - \beta_1 \bar{x}$$
$$= \beta_0$$

$\therefore E(\hat{\beta}_0 | x) = \beta_0$.

$\therefore \hat{\beta}_0$ is unbiased.

4. More on Regression and Time series analysis

a)

- For Total Population, intercept **β0_cap = -5234.821** and X coefficient **β1_cap = 2.7575**. **SSE** for Total Population dataset = **123.8494**

- For 65+ Population, intercept **β0_cap = -1178.5724** and X coefficient **β1_cap = 0.6076**. **SSE** for 65+ Population dataset = **176.0354**

- The Total Population data is suitable for linear regression, as from the plotted graph *(./graphs/Q4_a_population_total.png*), we can see the rate of change in total US population seems to have a near linear relationship with time (in years).

- The 65+ Population may not be suitable for linear regression, as from the plotted graph *(./graphs/Q4_a_population_65.png*), there seems to be a non-linear (skewed) relationship between rate of change of 65+ year old population over the years (It appears to be increasing gradually from 1980-1996, and then slows down between 1996-2013, after which there is a rapid increase in population.

b)

i) Train with 1980-2018 dataset -

- For 65+ Population from 1980-2018, intercept **β0_cap = -1131.1471** and X coefficient **β1_cap = 0.5838.** Regression Equation: $Y_{population} = -1131.1471 + 0.5838X_{year}$

- **Predicted 65+ Population in 2060** using 1980-2018 data = **71.4809 million**

- **SSE** for 1980-2018 65+ Population data = **137.6926**

ii) Train with 2008-2018 dataset

- For 65+ Population from 2008-2018, intercept **β0_cap = -2778.3124** and X coefficient **β1_cap = 1.4025**. Regression Equation: $Y_{population} = -2778.3124 + 1.4025X_{year}$

- Predicted 65+ Population in 2060 using 2008-2018 data = **110.8376 million**

- **SSE** for 2008-2018 65+ Population data = **2.0088**

- We should trust the second prediction more, as the from the graph that plots the regression fit of 65+ Year Old Population vs Year *(./graphs/Q4_a_population_65.png*), we can see a steep slope from 2013 denoting a sharp rate of increase of the 65+ population, which is better accounted for when we train the 2008-2018 data. Also, the SSE is lesser in case of the second prediction. Hence, the media may be right with the prediction that the 65+ population will double, as supported by our prediction of 110.84 million, which is more than double the population in 2018 of 52 million.

c) From our python program,
- Actual 65+ Population ratio in 2019 = **0.1646**. Actual Total Population = 328.329953, Actual 65+ Population = 54.058263
- Predicted 65+ Population ratio in 2019 using method 1 = **0.1623**
- Predicted 65+ Population ratio in 2019 using method 2 = **0.1618**. Predicted Total population = 329.6887 million, Predicted 65+ population = 53.3351 million

- Inference:
The ratio predicted by the first method (0.1623) seems to be more accurate as it is closer to the actual ratio (0.1646), than the second method which predicted a ratio of 0.1618. However as we can see, the difference in both methods is marginal (0.0005), so we can say both the methods work well.

5. Multiple Linear Regression (MLR)

a) All 7 features

	Beta Values
β1	-0.002911
β2	0.003232
β3	0.019910
β4	0.000576
β5	0.023193
β6	0.130898
β7	0.056820

- Linear Equation (Not including β0 intercept):

$$Y_{Admit\_Chance} = -0.002911X_{GRE} + 0.003232X_{TOEFL} + 0.019910X_{Uni\_rating} + 0.000576X_{SOP} + 0.023193X_{LOR} + 0.130898X_{GPA} + 0.056820X_{Research}$$

- SSE using all **7 features** is **0.3162**

b) TOEFL, SOP and LOR

	Beta Values
β1	0.003887
β2	0.041874
β3	0.048257

- Linear Equation (Not including β0 intercept):

$$Y_{Admit\_Chance} = 0.003887X_{TOEFL} + 0.041874X_{SOP} + 0.048257X_{LOR}$$

- **SSE** using **TOEFL, SOP and LOR** features is **0.6403**

c) GRE SCORE AND GPA

	Beta Values
β1	-0.004106
β2	0.235712

- Linear Equation (Not including β0 intercept):

$$Y_{Admit\_Chance} = -0.004106X_{GRE} + 0.235712X_{GPA}$$

- SSE using **GRE and GPA** features is **0.4648**

d) Observations:

- We obtain least SSE in the first case when we're using all 7 features, which may mean that the model using these features has a higher accuracy than the other 2 models.

- The 2nd model using 3 training features has the most SSE, which may mean that one or more of the features is/are not sufficient to make the right prediction, and not as important a factor as the others in determining the admit chance %.

6a) The soil in the farm is represented by RV $H = \{0, 1\}$, where 0 - good soil, 1 - bad soil.
Different samples of water content in the soil is represented by $w = \{w_1, w_2, \dots w_n\}$

→ The hypothesis test is given as below:

$H_0 : H = 0$, $H_1 : H = 1$

→ Also given, $P(H=0) = p$, $P(H=1) = 1-p$,
$f_w(w | H=0) = \text{Normal}(w; -\mu, \sigma^2)$ and
$f_w(w | H=1) = \text{Normal}(w; \mu, \sigma^2)$

→ also, $C = \begin{cases} 0 & \text{if } P(H=0|w) \geq P(H=1|w) \\ 1 & \text{otherwise,} \end{cases}$

which implies H_0 is chosen when $C = 0$, i.e,
when $P(H=0|w) \geq P(H=1|w)$ (i)

⇝ Using Baye's theorem, we have

$$P(H=0|w) = \frac{P(w|H=0) \cdot P(H=0)}{P(w)}$$

$$P(H=1|w) = \frac{P(w|H=1) \cdot P(H=1)}{P(w)}$$

⇒ using above results in (i), we get

$$P(w|H=0) \cdot p \geq (1-p) \cdot P(w|H=1)$$

Now, we are assuming that the samples in $w = \{w_1, \dots w_n\}$ are conditionally ⊥ given hypothesis

$$\therefore P(w|H=0) = P(\{w_1, \dots w_n\} | H=0)$$
$$= \prod_{i=1}^{n} P(w_i | H=0)$$
$$= \prod_{i=1}^{n} f_w(w_i | H=0) \quad \text{(ii)}$$

→ Using (ii) in (i), we get

$$\prod_{i=1}^{n} f_w(w_i | H=0) \cdot p = \prod_{i=1}^{n} f_w(w_i | H=1) \cdot (1-p)$$

$$\Rightarrow \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{w_i+\mu}{\sigma}\right)^2} \cdot p \geq \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{w_i-\mu}{\sigma}\right)^2} \cdot (1-p)$$

($\because$ given $f_w(w|H)$ is distributed normally)

$$\Rightarrow e^{-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{w_i+\mu}{\sigma}\right)^2} \cdot p \geq e^{-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{w_i-\mu}{\sigma}\right)^2} \cdot (1-p)$$

$$\Rightarrow e^{\frac{-1}{2\sigma^2}\sum(w_i^2+\mu^2+2w_i\mu)} \cdot p \geq e^{\frac{-1}{2\sigma^2}\sum(w_i^2+\mu^2-2w_i\mu)} \cdot (1-p)$$

$$\Rightarrow e^{\frac{1}{2\sigma^2}\left(-\sum w_i^2 - \sum\mu^2 - \sum 2w_i\mu + \sum\mu^2 + \sum w_i^2 - \sum 2w_i\mu\right)} \geq \frac{1-p}{p}$$

$$\Rightarrow e^{\frac{-4\mu\sum w_i}{2\sigma^2}} \geq \frac{1-p}{p} \Rightarrow e^{\frac{+2\mu\sum w_i}{\sigma^2}} \leq \frac{p}{1-p}$$

taking ln on both sides, we get

$$\frac{2\mu\sum w_i}{\sigma^2} \leq \ln\left(\frac{p}{1-p}\right)$$

$$\therefore \boxed{\sum_{i=1}^{n} w_i \leq \ln\left(\frac{p}{1-p}\right) \cdot \frac{\sigma^2}{2\mu}}$$ is the condition for choosing H_0

b) For $P(H_0) = 0.1$, the hypothesis selected are :: 0100101101

For $P(H_0) = 0.3$, the hypothesis selected are :: 0100101101

For $P(H_0) = 0.5$, the hypothesis selected are :: 0100101101

For $P(H_0) = 0.8$, the hypothesis selected are :: 0100101101

c) Average Error probability is given by

$$AEP = P(C=0 \mid H=1) \cdot P(H=1) + P(C=1 \mid H=0) \cdot P(H=0)$$

→ Now, $P(C=0 \mid H=1)$ can be read as choosing $H=0$ given that H is actually 1. From part a), it is same as the probability of

$$\sum_{i=1}^{n} w_i \leq \ln\left(\frac{p}{1-p}\right) \cdot \frac{\sigma^2}{2\mu}$$

Similarly, we can read $P(C=1 \mid H=0)$ as choosing $H=1$ given H is actually 0, which is probability of

$$\sum_{i=1}^{n} w_i > \ln\left(\frac{p}{1-p}\right) \cdot \frac{\sigma^2}{2\mu}$$

Let $c = \ln\left(\frac{p}{1-p}\right) \cdot \frac{\sigma^2}{2\mu}$

∴ using above results, we get

$$P(C=0 \mid H=1) = P\left(\sum w_i \leq c \mid H=1\right)$$
$$P(C=1 \mid H=0) = P\left(\sum w_i > c \mid H=0\right) \quad (i)$$

→ Now, we have to find distribution of $\sum_{i=1}^{n} w_i$

We know that $f_w(w \mid H=0) = N(-\mu, \sigma^2)$.

By using weighted sum of $\perp$ Normals (since $w_1, \ldots, w_n$ are conditionally $\perp$), we have

$$\Sigma w_i \mid H=0 \sim N(-n\mu, n\sigma^2) \quad \text{(ii)}$$

Similarly, $\Sigma w_i \mid H=1 \sim N(n\mu, n\sigma^2)$

→ Now, $P(\Sigma w_i \leq c \mid H=1) = F_{\Sigma w_i | H=1}(c)$

To convert above to std normal z, we have to apply transformation $z = \frac{x - \mu_x}{\sigma_x}$ at point x on $X \sim N(\mu_x, \sigma_x)$

→ ∴ For $\Sigma w_i \mid H=1 \sim N(n\mu, n\sigma^2)$, we have

$$z = \frac{c - n\mu}{\sigma\sqrt{n}} \quad \text{(iii)}$$

$$\therefore P(\Sigma w_i \leq c \mid H=1) = F_z\left(\frac{c - n\mu}{\sigma\sqrt{n}}\right)$$

$$= \Phi\left(\frac{c - n\mu}{\sigma\sqrt{n}}\right)$$

→ Similarly, $P(\Sigma w_i > c \mid H=0) = 1 - P(\Sigma w_i \leq c \mid H=0)$

$$= 1 - F_{\Sigma w_i | H=0}(c)$$

$$= 1 - \Phi\left(\frac{c + n\mu}{\sigma\sqrt{n}}\right) \quad (\text{From (ii)})$$

∴ $AEP = P(C=0 \mid H=1) \cdot P(H=1) + P(C=1 \mid H=0) \cdot P(H=0)$

$= P(\Sigma w_i \leq c \mid H=1) \cdot (1-p) + P(\Sigma w_i > c \mid H=0) \cdot p$

$$\therefore AEP = (1-p) \cdot \Phi\left(\frac{c - n\mu}{\sigma\sqrt{n}}\right) + p \cdot \left(-\Phi\left(\frac{c+n\mu}{\sigma\sqrt{n}}\right) + 1\right)$$

$$= \boxed{(1-p)\,\Phi\left(\frac{\ln\left(\frac{p}{1-p}\right) \cdot \frac{\sigma^2}{2\mu} - n\mu}{\sigma\sqrt{n}}\right) + p \cdot \left(1 - \Phi\left(\frac{\ln\left(\frac{p}{1-p}\right) \cdot \frac{\sigma^2}{2\mu} + n\mu}{\sigma\sqrt{n}}\right)\right)}$$

(using value of constant c)