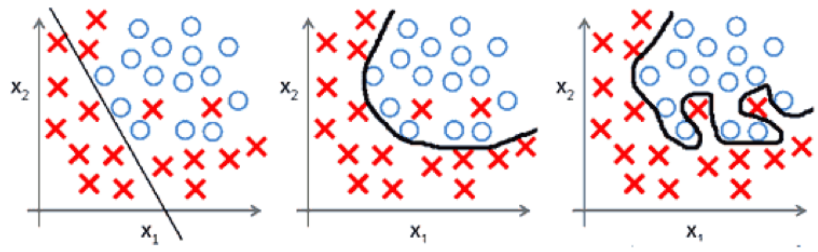
1 Bias, Variance and Regularization

(a) **(5 points)** The figures below show the decision boundary of three different classifiers. In which of the figures below does the classifier have a larger bias and in which figure does the classifier has a larger variance? Draw out an approximate graph where you demonstrate how the training and test error for each classifier will change over time during the course of training the model, and explain how do you expect each classifier to perform (accuracy) on the test set.
Note: You don't need to calculate the errors, but you should show how the training and test errors compare for different classifiers.

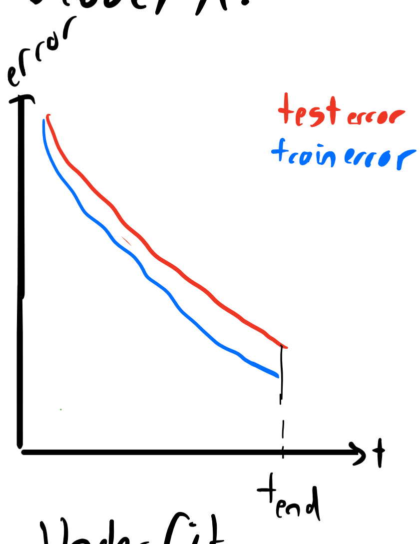


Model A   Model B   Model C

Bias: A > B > C , underfitting

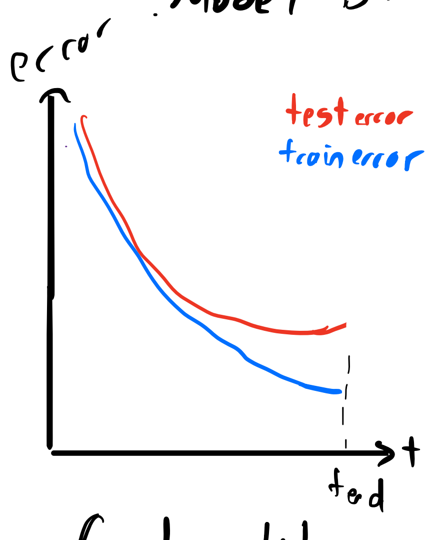
Variance: C > B > A , overfitting

Model A:



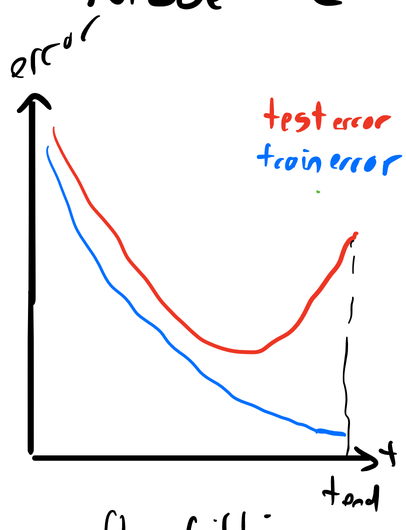
Underfit

Model B:



Good model

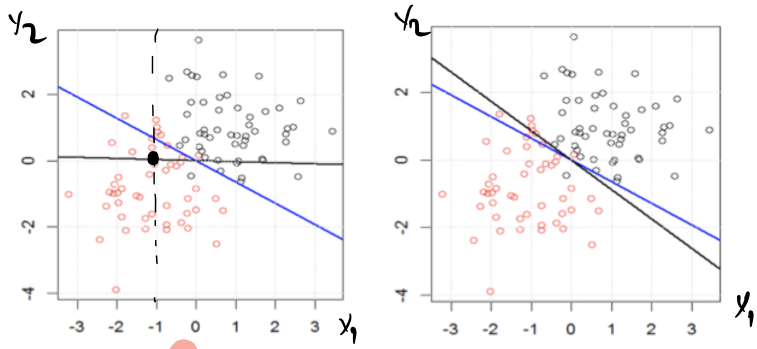
Model C.



Overfitting

Test set performance: Model B will perform best due to moderate variance and moderate bias. Model A and B can not be compared since we do not have data, but overfitting model generally beats underfitting, since it is trained more. However, early stopping technique is used for some models, therefore some underfitting models can perform better.

(b) **(5 points)** One strategy to reduce variance and improve generalization is regularization. In figure below, the blue lines are the logistic regression without regularization and the black lines are logistic regression with L1 or L2 regularization. In which figure L1 regularization is used and why?



(a) (b)

Model 1:

$\beta_1^{(1)} x_1 + \beta_2^{(1)} x_2 + \beta_0^{(1)}$

Model 2:

$\beta_1^{(2)} x_1 + \beta_2^{(2)} x_2 + \beta_0^{(2)}$

L1 regularization is used for Model 1, graph a. Because changing x_2 has no effect on decision boundary. Therefore, coefficient of x_2 should be zero. L1 regularization made x_2's coefficient to be zero.

Also, suppose make x_1 constant, changing x_2 has no effect on score. Therefore coefficient $\beta_2^{(1)}$ should be zero.

2 Classification Metrics

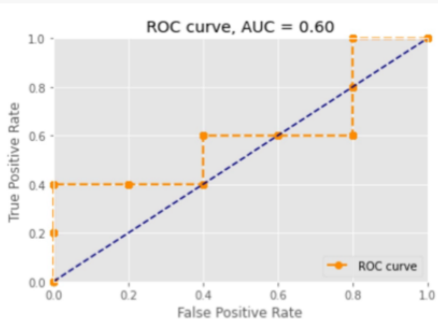
You have trained a Logistic Regression classifier which gives you the following predictions on a test set:

Correct Label	Predicted Probability
1	0.97
1	0.93
1	0.65
1	0.52
1	0.37
0	0.89
0	0.71
0	0.56
0	0.54
0	0.16

(a) **(5 points)** Draw the Receiver Operating Characteristic (ROC) curve. Show clearly the points where the curve changes direction by e.g. including ticks on the x- and y-axis in the corresponding locations.

```python
import numpy as np
from sklearn import metrics
import matplotlib.pyplot as plt
score = np.array([0.97, 0.93, 0.65, 0.52, 0.37, 0.89, 0.71, 0.56, 0.54, 0.16])
y = np.array([1,1,1,1,1,0,0,0,0,0])
FPR = []
TPR = []
thresholds = np.arange(0.0, 1.01, 0.0001)
P = sum(y)
N = len(y) - P
for thresh in thresholds:
    FP=0
    TP=0
    thresh = round(thresh,2)
    for i in range(len(score)):
        if (score[i] >= thresh):
            if y[i] == 1:
                TP = TP + 1
            if y[i] == 0:
                FP = FP + 1
    FPR.append(FP/N)
    TPR.append(TP/P)

auc = metrics.auc(FPR, TPR)
plt.plot(FPR, TPR, linestyle='--', marker='o', color='darkorange', lw = 2, label='ROC curve', clip_
plt.plot([0, 1], [0, 1], color='navy', linestyle='--')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.0])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('ROC curve, AUC = %.2f'%auc)
plt.legend(loc="lower right")
plt.savefig('AUC_example.png')
plt.show()
```



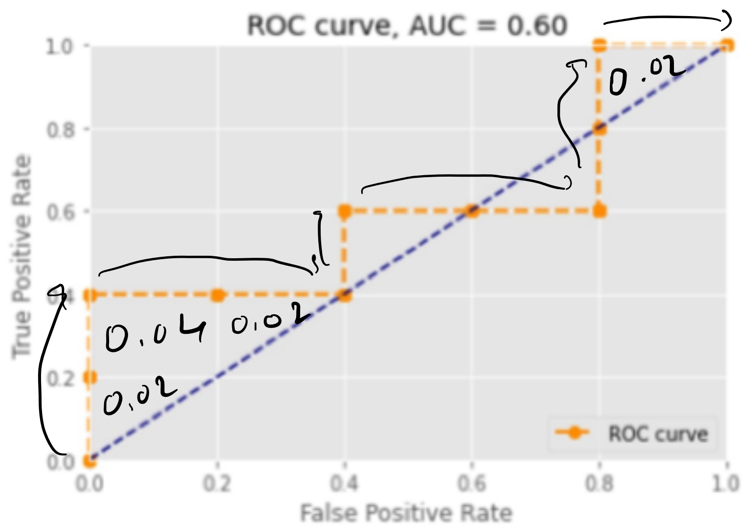


Figure 1: ROC Curve

(b) **(2 points)** Compute the AUC score.

$$AUC = 0.04 + 3 * 0.02 + 0.5$$
$$= \boxed{0.60}$$

(c) **(2 points)** Draw the confusion matrix when the decision threshold is 0.5.

Actual \ Prediction	−	+
−	1 TN	4 FP
+	1 FN	4 TP

Correct Label	Predicted Probability	
1	0.97	→ 1 → TP
1	0.93	→ 1 → TP
1	0.65	→ 1 → TP
1	0.52	→ 1 → TP
1	0.37	→ 0 X → FN
0	0.89	→ 1 X → FP
0	0.71	→ 1 X → FP
0	0.56	→ 1 X → FP
0	0.54	→ 1 X → FP
0	0.16	→ 0 → TN

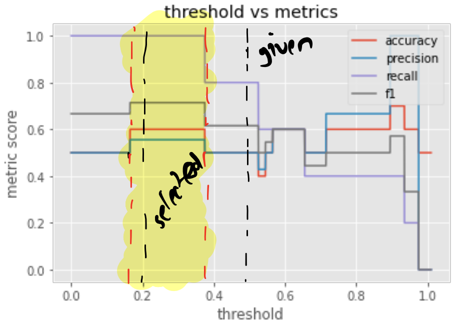
Recall $\frac{TP}{TP+FN}$, Precision $\frac{TP}{TP+FP}$

(d) **(2 points)** Using the previous confusion matrix, compute Accuracy, Precision, Recall, and F1 score.

Accuracy: $\frac{5}{10} = 0.5$, Precision: $\frac{4}{8} = 0.5$, Recall: $\frac{4}{5} = 0.8$

$$F1 = \frac{tp}{tp + 0.5(fp+fn)} = \frac{4}{4+\frac{1}{2}(5)} = \frac{4}{6.5} = 0.6153846$$

(e) **(5 points)** Can we improve any of the previous scores (without a negative effect on any of the other scores) by changing the threshold? If yes, which threshold value would you choose and why? If not, explain why not.



Threshold = 0.2

Correct Label	Predicted Probability	
1	0.97	1 → TP
1	0.93	1 → TP
1	0.65	1 → TP
1	0.52	1 → TP
1	0.37	1 → TP
0	0.89	1 → FP
0	0.71	1 → FP
0	0.56	1 → FP
0	0.54	1 → FP
0	0.16	0 → TN

Figure 2: threshold vs metrics

Actual \ Prediction	−	+
−	1 TN	4 FP
+	0 FN	5 TP

→ threshold = 0.2

$\boxed{\text{Threshold} \in (0.16, 0.37)}$

Accuracy: $\frac{6}{10} = 0.6$, Precision: $\frac{5}{9} = 0.55$, Recall: $\frac{5}{5} = 1$

F1 $= \frac{tp}{tp + 0.5(fp + fn)} = \frac{5}{5 + \frac{1}{2}(4)} = \frac{5}{7} = 0.71$

With threshold $\in (0.16, 0.37)$ we were able to classify one more positive label correctly, therefore all metrics have increased.

there are no other interval.

3 Logistic Regression

Suppose we fit a multiple logistic regression: $\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$.

(a) **(2 points)** Suppose we have $p = 2$, and $\beta_0 = 3, \beta_1 = 5, \beta_2 = -8$. When $X_1 = X_2 = 0$, what are the odds and probability of the event that $Y = 1$?

odds: $\frac{P(Y=1)}{1-P(Y=1)}$, $P(Y=1) = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2)}}$

$P(Y=1 \mid X_1 = X_2 = 0) = \frac{1}{1+e^{-\beta_0}} = 0.9525$

$\frac{P(Y=1)}{1-P(Y=1)} = 20.0855691 \rightarrow \ln(20.08) = 3 = \beta_0$

(b) **(2 points)** Suppose we increase the X_1 value by 2, how does it change the log odds and odds of the event that Y = 1? What if instead, we decrease the X_2 value by 2?

$P(Y=1 \mid X_1 = 2, X_2 = 0) = \frac{e^{\beta_0 + 2\beta_1}}{1 + e^{\beta_0 + 2\beta_1}} = \frac{e^{3+2 \cdot 5}}{1+e^{13}}$

$= 0.999997$

odds: $= 442418 \Rightarrow$ log odds $= 13$ ↑ increases

increases

log odds and odds.

$P(Y=1 \mid X_1 = 0, X_2 = -2) = \frac{e^{\beta_0 - 2\beta_2}}{1+e^{\beta_0 - 2\beta_2}} = \frac{e^{3+16}}{1+e^{19}} = 0.9999999944$

odds: $= 178571422 7.6 \Rightarrow$ log ods $= 19$

increases

increases

(c) **(2 points)** Explain how increasing or decreasing β_0, β_1 or β_2 affect our predictions.

$\beta_0 + \beta_1 x_1 + \beta_2 x_2 = score$

$$P(Y=1) = \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2}}{1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2}} > 0.5 \text{ or threshold}$$

$$\beta_0: \; P(Y=1) = \frac{e^{\beta_0 + c}}{1 + e^{\beta_0 + c}} \quad \begin{aligned} &\Rightarrow \beta_0 \uparrow \; \frac{e^{\infty}}{1+e^{\infty}} = 1, \\ &\Rightarrow \beta_0 \downarrow \; \frac{e^{-\infty}}{1+e^{-\infty}} = 0 \end{aligned}$$

When β_0 increases we are more likely to classify our data as 1, when β_0 decreases we are more likely to classify our data as 0.

$$\beta_i : \beta_1 \text{ or } \beta_2 \qquad P(Y=1) = \frac{e^{c + \beta_i x_i}}{1 + e^{c + \beta_i x_i}} \quad \begin{cases} x_i > 0 & \begin{cases} \beta_i \uparrow \; \frac{e^{\infty}}{1+e^{\infty}} = 1 \\ \beta_i \downarrow \; \frac{e^{-\infty}}{1+e^{-\infty}} = 0 \end{cases} \\ x_i < 0 & \begin{cases} \beta_i \uparrow \; \frac{e^{-\infty}}{1+e^{-\infty}} = 0 \\ \beta_i \downarrow \; \frac{e^{\infty}}{1+e^{\infty}} = 1 \end{cases} \end{cases}$$

When β_i, $i \neq 0$ increases and the feature x_i is positive. We are more likely to classify our data as 1. However if our feature is negative. We are more likely to classify our data as 0. If our feature is close to zero or $\beta_i x_i$ is smallor than other terms, changing β_i has no effect on classification.

When β_i, $i \neq 0$ decreases and the feature x_i is positive. We are more likely to classify our data as 0. However, if our feature is negative. We are more likely to classify our data as 0. If our feature is close to zero or $\beta_i x_i$ is smaller than other terms, changing β_i has no effect on classification.

(d) **(2 points)** What is the formulation of the decision boundary? Which points are on the decision boundary?

Decision boundary of logistic regression is the set x that satisfy

$\Rightarrow P(Y=1 \mid X) = P(Y=0 \mid X) = \frac{1}{2}$ if threshold is 0.5

$$\Rightarrow \frac{e^{\beta_0+\beta_1 x_1+\dots+\beta_n x_n}}{\cancel{1+e^{\beta_0+\beta_1 x_1+\dots+\beta_n x_n}}} = \frac{1}{\cancel{1+e^{\beta_0+\beta_1 x_1+\dots+\beta_n x_n}}}$$

>0 >0

$\Rightarrow e^{\beta_0+\beta_1 x_1+\dots+\beta_n x_n} = 1 \Rightarrow \beta_0+\beta_1 x_1+\dots+\beta_n x_n = \ln 1 = 0$

D.B $\beta_0+\beta_1 x_1+\dots+\beta_n x_n = 0$

$\Rightarrow$ Points that satisfy $\beta_0+\beta_1 x_1+\dots+\beta_n x_n = 0$ are on the decision boundary

when $P(Y=1 \mid X) \neq P(Y=0 \mid X)$:

$P(Y=0 \mid X) = \frac{w}{w+1}$ $w=1$ previous case

$P(Y=1 \mid X) = \frac{1}{w+1}$ $\boxed{0 \le w \le \infty}$ with w we can adjust threshold.

$$\frac{1}{w} = e^{\beta_0+\dots+\beta_n x_n} \Rightarrow \boxed{\beta_0+\dots+\beta_n x_n = -\ln(w)}$$

$w=2$: $P(Y=0 \mid X) < \frac{2}{3}$ $\Rightarrow$ probabilities less than 0.33 will be predicted as 0 and probabilities greater than 0.33 will be predicted as 1.

$P(Y=1 \mid X) \geq \frac{1}{3}$

Lastly, points that satisfy $\beta_0 + \cdots + \beta_n x_n = -\ln(u)$ are on the decision boundary.

(e) **(2 points)** Suppose we fit another two logistic regression models: one with only X_1 and the other one with only X_2, and we observe that the coefficients of X_1 and X_2 in the two models are different than those specified in part (a). Explain what is the potential reason and why it could be problematic that the coefficients are different than those specified in part (a).

Main reason is the colinearity between predictor variables. Variables are not independent from each other.
Multicolinearity increases the standard errors of their coefficients and make those coefficients unreliable and unstable estimates.
Multicolinearity can be easily found by the variance inflation factor.

4 Logistic Regression with Interaction Term

You are analyzing how the birth weight of a baby (normal weight=0, low weight=1) depends on the age of the mother (number of years over 23, e.g. a 25-year-old will have value 2) and the frequency of physician visits during the first trimester of pregnancy (0=not frequent, 1=frequent). You have also decided to include an interaction term for age and frequency. Your logistic regression coefficients are as follows:

Feature	Coefficient
Intercept	-0.48
Age	0.06
Frequency	-0.45
Age $\times$ Frequency	-0.19

(a) **(4 points)** Discuss the meaning of each coefficient, and explain what does the coefficient of the interaction term show.

Intercept is the bias term. It is the expected value when all predictor variables are zero. When predictor variables are not zero, it is the mean value of prediction. In the case of logistic regression, these coefficients have multiplicative effect on the odds. When all numerical features are zero and categorical variable is zero, then the estimated odds are exp(beta0).

Age is a predictor variable, and if you are older your baby will likely have low weight. Therefore, prediction and age is positively correlated. The effect of coefficient of age is different from linear regression case. A change in age by one unit changes the odds ratio multiplicatively by a factor of exp(beta1).

Note that, we have also a binary variable, frequency. Binary coefficient splits our model into two models. Frequency's coefficient adjusts the intercept. Therefore, odds are different in the case of frequency = 1 and frequency = 0. The multiplicative difference is exp(coef_frequency).

We have also an interaction term. Interaction terms effect is similar to effect to intercept. Age_Frequency's coefficient adjust the coefficient of age. Therefore, odds are different in the case of frequency = 1 and frequency = 0. The multiplicative difference is exp(coef_frequency)

When f = 1, intercept becomes -0.43-0.48= -0.93. When f = 0, intercept becomes -0.19+0.06=-0.13. Note that, the sign the of age's coefficient changes, therefore the logistic curve flips. Age becomes negatively correlated with weight.

In brief, beta0/beta1 shifts logistic regression curve and beta1 decides on the slope of logistic regression. Frequency variable changes these default coefficients.

In the context of the question, for frequent visitors, odds is exp(-0.13) = 0.87 times for a unit increase in age.
For non-frequent visitors, odds is exp(0.06) = 1.06 times for a unit-increase in age. The ratio of these two odds is the exponentiated coefficient for the interaction term of age_frequency: 0.82 = exp(-0.19)

(b) **(3 points)** Specify the logistic regression models when the mother visited the physician frequently and when they didn't. Explain how the mother's age affects the odds in each scenario.

$f=1$

$$P(Y=1 \mid X) = \frac{e^{-0.93-0.13\,Age}}{1+e^{-0.93-0.13\,Age}}$$

$f=0$

$$P(Y=1 \mid X) = \frac{e^{-0.48+0.06\,Age}}{1+e^{-0.48+0.06\,Age}}$$

two different models

When f = 1, if age increases odds decreases, therefore we are more likely to label as 0. However when f = 0, if age increases odds increases, therefore we are more likely to label as 1. In result, sign of the age coefficient changes and the correlation between age and weight of baby changes with frequency variable.

$f=1$, Age $\uparrow$, odds increases and approaches to ∞

odds: $\frac{P(Y=1)}{1-P(Y=1)} \Rightarrow \infty$

Age and odds positively correlated,

$f=0$, Age $\uparrow$, odd decreases and approaches to 0

odds: $\frac{P(Y=1)\downarrow}{1-P(Y=1)\downarrow} \Rightarrow 0$

age and odds negatively correlated.

(c) **(4 points)** Compare how physician visits affect odds of low weight at ages 18, 23, 25, 28, 30, by calculating the odds ratio of low birth weight for mothers with frequent physician visits over those with non-frequent physician visits, in the following table (fill the "Odds Ratio" column in the table below). *Note: for age, you should use number of years over 23.*

Odds Ratio: $\dfrac{\frac{P(Y=1 \mid X=1)}{1-P(Y=1 \mid X=1)}}{\frac{P(Y=1 \mid X=0)}{1-P(Y=1 \mid X=0)}}$

Age:	Age	Odds Ratio	95% Confidence Interval
-5	18	1.65	(0.705, 4.949)
0	23	0.64	(0.325, 1.201)
2	25	0.44	(0.262, 1.036)
5	28	0.25	(0.206, 0.916)
7	30	0.17	(0.050, 0.607)

$f=1$

$$P(Y=1 \mid X) = \frac{e^{-0.93-0.13\,Age}}{1+e^{-0.93-0.13\,Age}}$$

$f=0$

$$P(Y=1 \mid X) = \frac{e^{-0.48+0.06\,Age}}{1+e^{-0.48+0.06\,Age}}$$

$\Rightarrow \dfrac{e^{-0.93-0.13*(-5)}}{1+} = 0.4304 \; \dfrac{P(Y=1)}{1-P(Y=1)} \Rightarrow 0.755783$

$\Rightarrow \dfrac{e^{-0.48+0.06(-5)}}{1+e^{-0.48+0.06(-5)}} = 0.3143 \; \dfrac{P(Y=1)}{1-P(Y=1)} \Rightarrow 0.4584$

$\Rightarrow \dfrac{e^{-0.93-0.13*(0)}}{1+} = 0.2829 \; \dfrac{P(Y=1)}{1-P(Y=1)} \Rightarrow 0.394557$

$\dfrac{e^{-0.48+0.06(0)}}{1+e^{-0.48+0.06(0)}} = 0.3822 \; \dfrac{P(Y=1)}{1-P(Y=1)} \Rightarrow 0.61878$

$\Rightarrow \dfrac{e^{-0.93-0.13*(2)}}{1+} = 0.23325 \; \dfrac{P(Y=1)}{1-P(Y=1)} \Rightarrow 0.3042$

$\dfrac{e^{-0.48+0.06(2)}}{1+e^{-0.48+0.06(2)}} = 0.4109 \; \dfrac{P(Y=1)}{1-P(Y=1)} \Rightarrow 0.6976$

$\Rightarrow \dfrac{e^{-0.93-0.13*(5)}}{1+} = 0.1707 \; \dfrac{P(Y=1)}{1-P(Y=1)} \Rightarrow 0.2059$

$\dfrac{e^{-0.48+0.06(5)}}{1+e^{-0.48+0.06(5)}} = 0.4551 \; \dfrac{P(Y=1)}{1-P(Y=1)} \Rightarrow 0.8352$

$\Rightarrow \dfrac{e^{-0.93-0.13*(7)}}{1+} = 0.130519 \Rightarrow 0.1588$

$\dfrac{e^{-0.48+0.06(7)}}{1+e^{-0.48+0.06(7)}} = 0.4850 \; \dfrac{P(Y=1)}{1-P(Y=1)} \; 0.94176$

Interpretation: Odds ratio decreases as age increases. Between 1.65 and 1, the difference between frequent and non-frequent visitors decreases, however between 1 and 0.17, the difference between frequent and non-frequent visitors increases. The main reason behind this switch is the sign of the age coefficient is different for each frequency value. The odds are higher at 18 when comparing frequent over non-frequent. However, the odds are lower at other ages when comparing frequent over non-frequent.

(d) **(4 points)** Interpret the numbers in the "Odds Ratio" column, considering the listed confidence intervals. *Hint: what does an odds ratio of 1 mean (holding other predictors fixed)?.*

Interpretation: Odds ratio decreases as age increases. Between 1.65 and 1, the difference between frequent and non-frequent visitors decreases, however between 1 and 0.17, the difference between frequent and non-frequent visitors increases. The main reason behind this switch is the sign of the age coefficient is different for each frequency value. The odds are higher at 18 when comparing frequent over non-frequent. However, the odds are lower at other ages when comparing frequent over non-frequent.

Age:
-5
0
2
5
7

Age	Odds Ratio	95% Confidence Interval
18	1.65	(0.705, 4.949)
23	0.64	(0.325, 1.201)
25	0.44	(0.262, 1.036)
28	0.25	(0.206, 0.916)
30	0.17	(0.050, 0.607)

Odds ratio of 1: If odds ratio is higher than 1, frequent visitors odds is higher than the non-frequent visitors odds. However, odds ratio switches in our case as age changes. If odds ratio is lower than 1, frequent visitor's odds is lower than the non-frequent visitors odds. If odds ratio is 1, binary value has no effect on odds, since two odds are equal to each other which also means probabilities are equal to each other.

For age 18, confidence interval is (0.705, 4.949), therefore the odds when f = 1 is much likely higher than the odds when f = 0. However, we can not give definiteness since (0.705,1) is included in the confidence interval.
For age 23, confidence interval is (0.325, 1.201), therefore the odds when f = 1 is much likely lower than the odds when f = 0. However, we can not give definiteness since(1,1.201) is included.
For age 25, confidence interval is (0.262, 1.036), therefore the odds when f = 1 is much likely lower than the odds when f = 0. We can slightly give definiteness to our answer since(1,1.036) is a very narrow interval.
For age 28, confidence interval is(0.206,0.916), therefore the odds when f = 1 is very likely lower than the odds when f = 0. Since (1,a) interval is not included in this age, we can give definiteness to our answer.
For age 30, confidence interval is(0.050, 0.607), this is most likely case that odds when f = 1 is lower than the odds when f = 0. The confidence interval is very far away from the odds ratio value 1.

(e) **(5 points)** compare the "difference in probability" of low birth weight for mothers at ages 18, 23, 25, 28, 30, in the table below. Interpret your results and compare your interpretation to part (d).

Age	Difference in probability	95% Confidence Interval
18	0.1111	(-0.788,0.393)
23	-0.09932	(-0.197,0.088)
25	-0.1777	(-0.232,0.046)
28	-0.2843	(-0.315,-0.016)
30	-0.3479	(-0.540,-0.092)

Difference in probability: $P(Y=1 \mid f=1) - P(Y=1 \mid f=0)$

The pattern we see here is very similar to the pattern at d. Rather than thinking about the odds ratio 1 boundary, here we need to focus on the boundary where difference in probabilities is 0.

For age 18, difference of probability indicates that the probability of giving low weight birth is higher when frequent visits are made. Confidence interval also validates and creates the indefiniteness like in the case of odds ratio. Confidence interval is (-0.788,0.393) which means that we can not be sure whether probability is higher in the case of frequency = 1 or frequency = 0. The confidence interval includes both ends (-0.788,0) and (0, 0.393)

For age 23, the difference of probability indicates that the probability of giving low weight birth is lower when frequent visits are made. Confidence interval also validates like in the case of odds ratio. Confidence interval is (-0.197,0.088) which means we can not be sure whether probability is higher in the frequency = 1 case or frequency = 0 case. The confidence interval includes both ends (-0.197,0) and (0, 0.088). However, note that the negative side is larger than the positive side which supports our deduction at the first sentence.

For age 25, the difference of probability indicates that the probability of giving low weight birth is lower when frequent visits are made. Confidence interval also validates like in the case of odds ratio. Confidence interval is (-0.232,0.046) which means we can not be sure whether probability is higher in the frequency = 1 case or frequency = 0 case. The confidence interval includes both ends(-0.232,0) and (0,0.045). However,note that the negative side is larger than the positive side which supports our deduction at the first sentence.

For age 28, it is similar to interpretation made for age 25. In this case there is no interval where p>0. Therefore, it is highly likely that probability of giving low weight birth is lower when frequent visits are made. The probability of giving low weight birth is higher when non-frequent visits are made.

For age 30, it is extreme case for age 28. We are very far away from probability 0 which indicates it is highly likely that probability of giving low weight birth is lower when frequent visits are made.