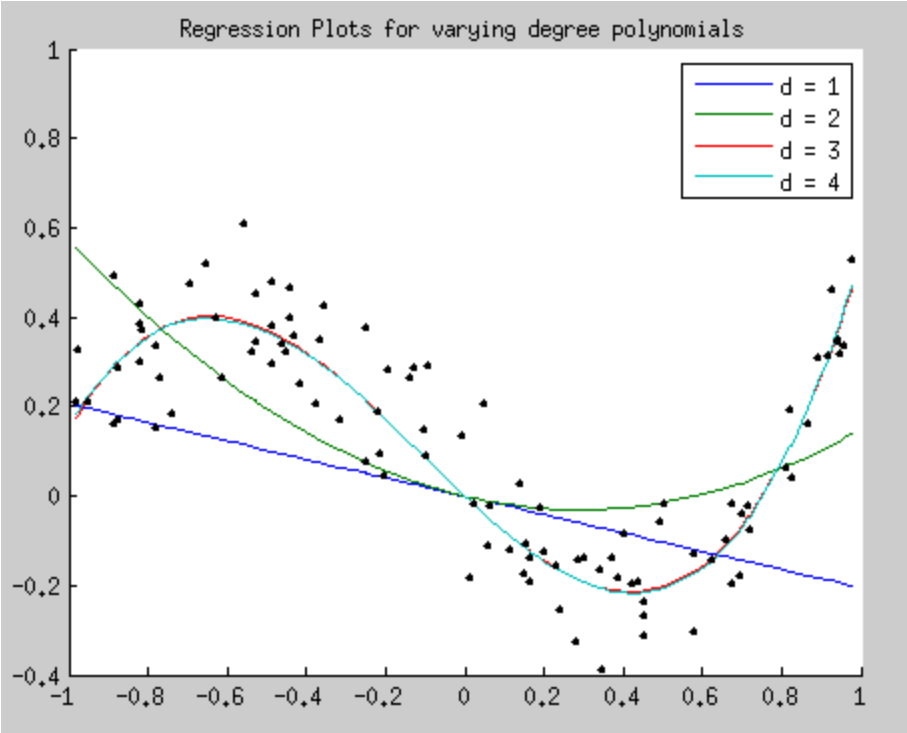
Mukul Murthy - 21803720
Thomas Chow - 22215698

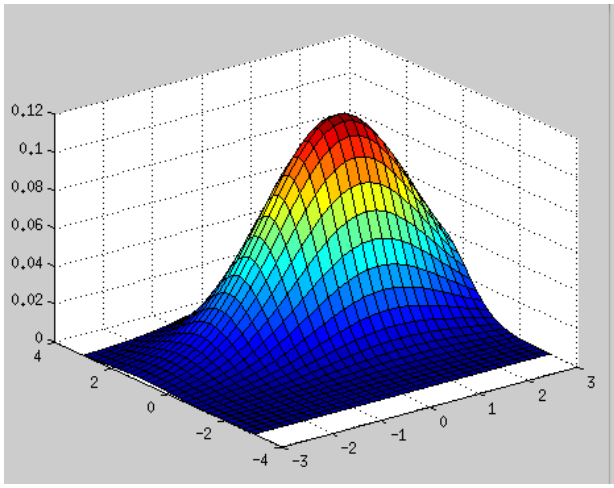
Homework 3

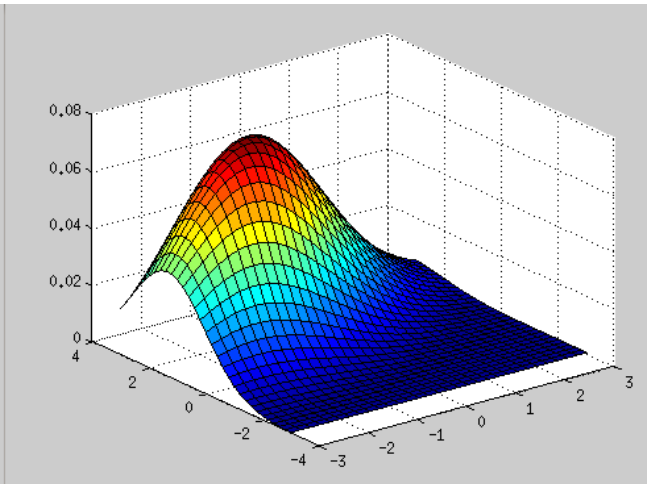
1a) Of the values from 0 to 10, d=10 minimizes the cost function. This makes sense, because a higher degree polynomial will always fit data equal or better to lower degree polynomials. As d approaches infinity, the model fits the data better and better. However, choosing a value of d that is too high could lead to overfitting.



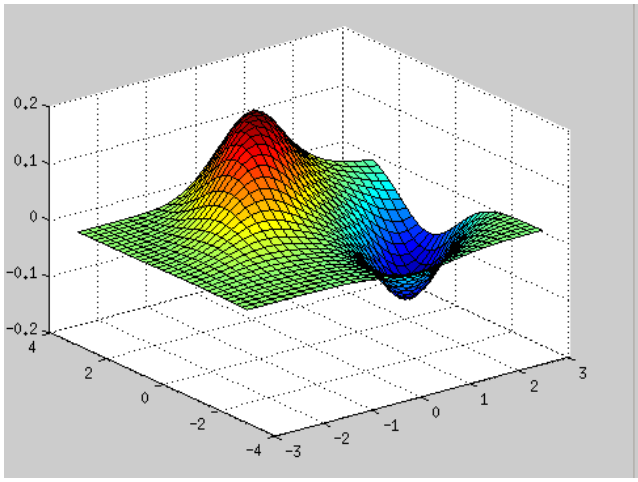
1b) The error is 4.8148 for d=3 and 4.7738 for d=10. This says that there is not much difference between d=3 and d=10. Additionally, d=10 is a much more complicated model, and the slight decrease in error for this data set may not be worth the extra complexity.

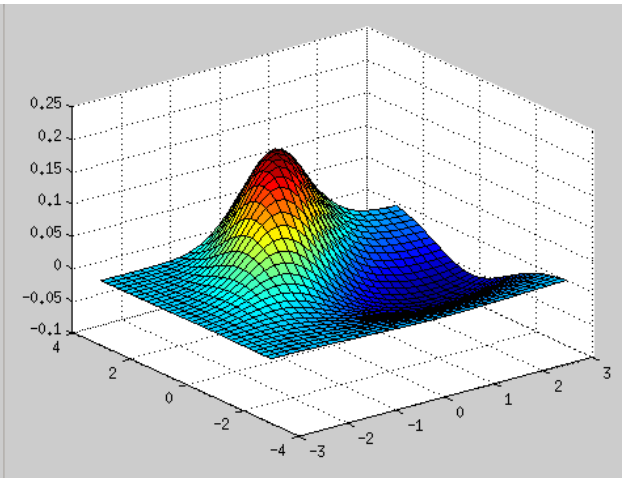
2. i) and ii)



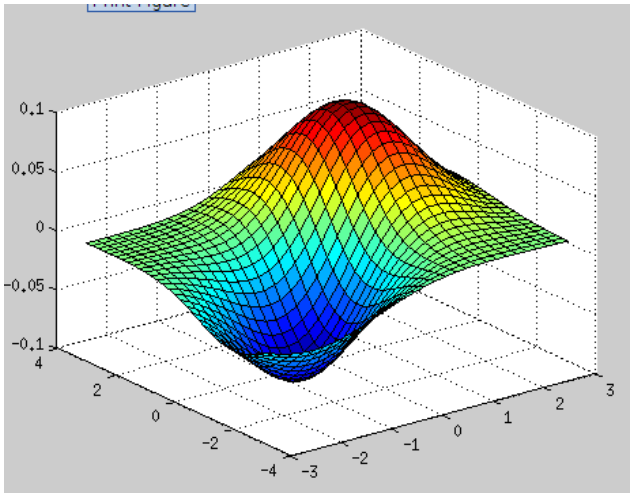


iii) and iv)





v)



3 i)

The maximum likelihood estimate for mean is the sample mean. The maximum likelihood estimator for the covariance is described by the following second equation:

$$\hat{\mu} = \frac{1}{N}\sum_i x_i = \overline{x}$$

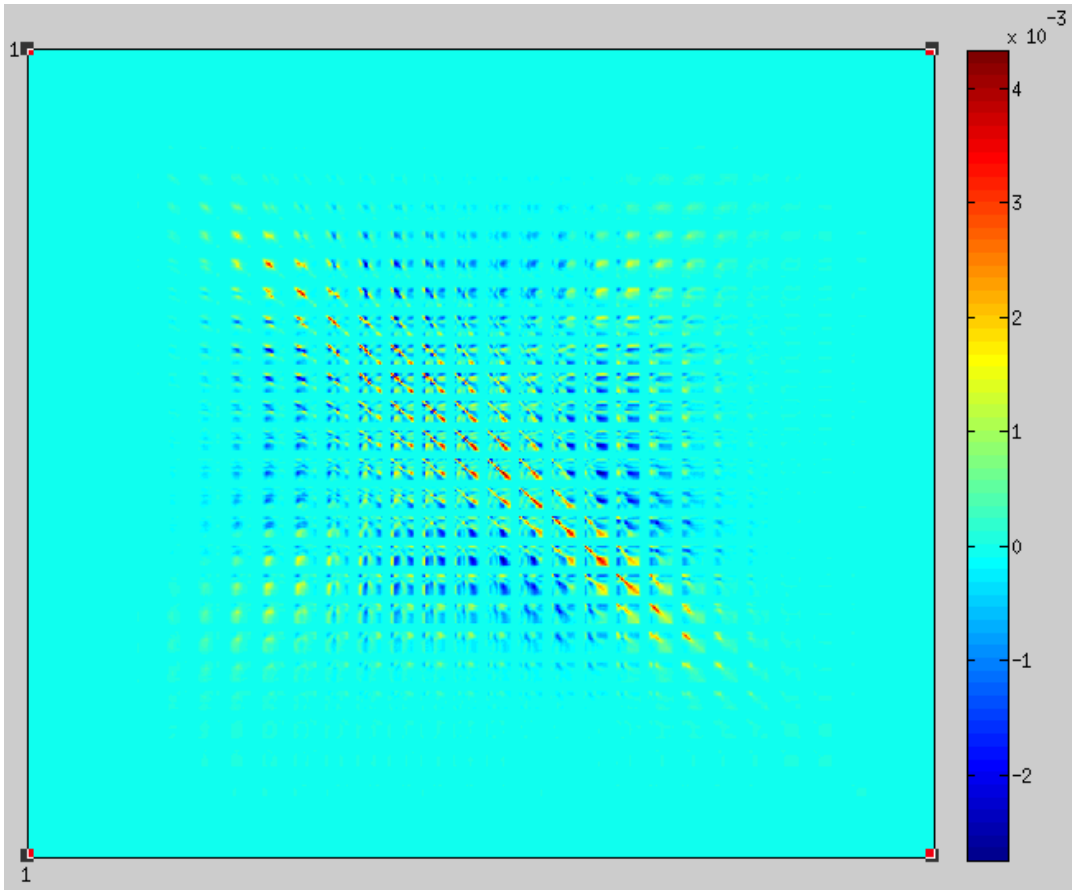
$$\hat{\sigma}^2 = \frac{1}{N}\sum_i (x_i - \overline{x})^2 = \left(\frac{1}{N}\sum_i x_i^2\right) - (\overline{x})^2$$

The maximum likelihood estimators are unbiased because the expected values for the mean and covariance matrix are the same as the true values.

ii) Before observing any values, a given sample has the same probability of belonging to a class as the proportion of samples that are part of that class. To find the prior probabilities, we counted how many samples there were for each class and divided by the number of total samples (Bernoulli distribution). The prior probabilities (in order from 0 to 9) are as follows:

0.0989 0.1124 0.0993 0.1022 0.0972 0.0905 0.0985 0.1045 0.0976 0.0989

iii) The covariance matrix shows the covariance of each pixel with each other pixel (entry i,j shows the covariance between pixels i and j). The diagonal is bright; this makes sense because the entries on the diagonal simply show the variance in that pixel. Entries close to the diagonal are also bright. It makes that pixels close together have high covariances; they are likely to change together. However, going a little further away from each pixel results in negative covariance. This says there is likely to be a light (no writing) pixel somewhat close to a dark (marked) pixel, which makes sense because the font used has a limited thickness.

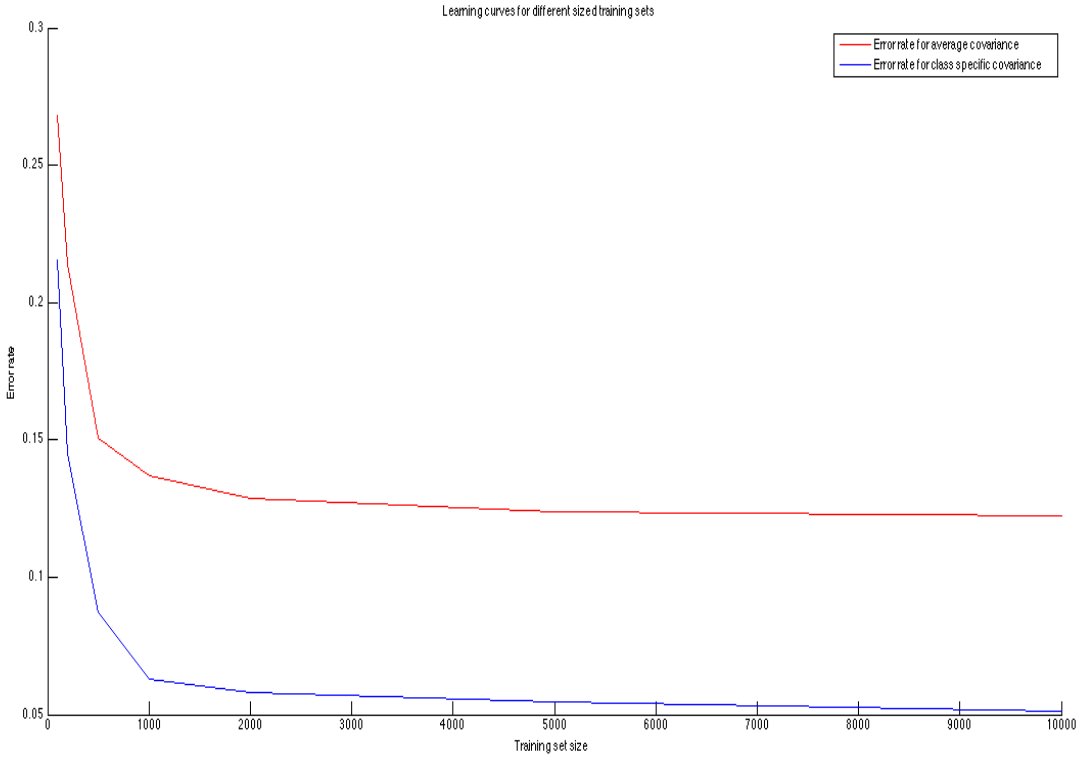


iv)

a) The form of the decision boundary is linear. This is because each class draws from a distribution with the same covariance matrices.

b) The form of the decision boundary is quadratic. This is because each class draws from a distribution with different covariance matrices, so the shape of the gaussian is different for different classes. This creates multiple places of intersection between the gaussian contours, which requires a higher order decision boundary to characterize.

c) The difference in performace arises because the covariance specific to a class label is more representative of the correlation between pixels for a particular class. The decision boundary for the average covariance is linear, which means it is more rigid and may not be able to conform to a curved decision boundary. The quadratic decision boundary fits around curved decision boundaries, which makes it better for distinguishing samples of different classes that have very similar features and may fall on the same side of a linear decision boundary.



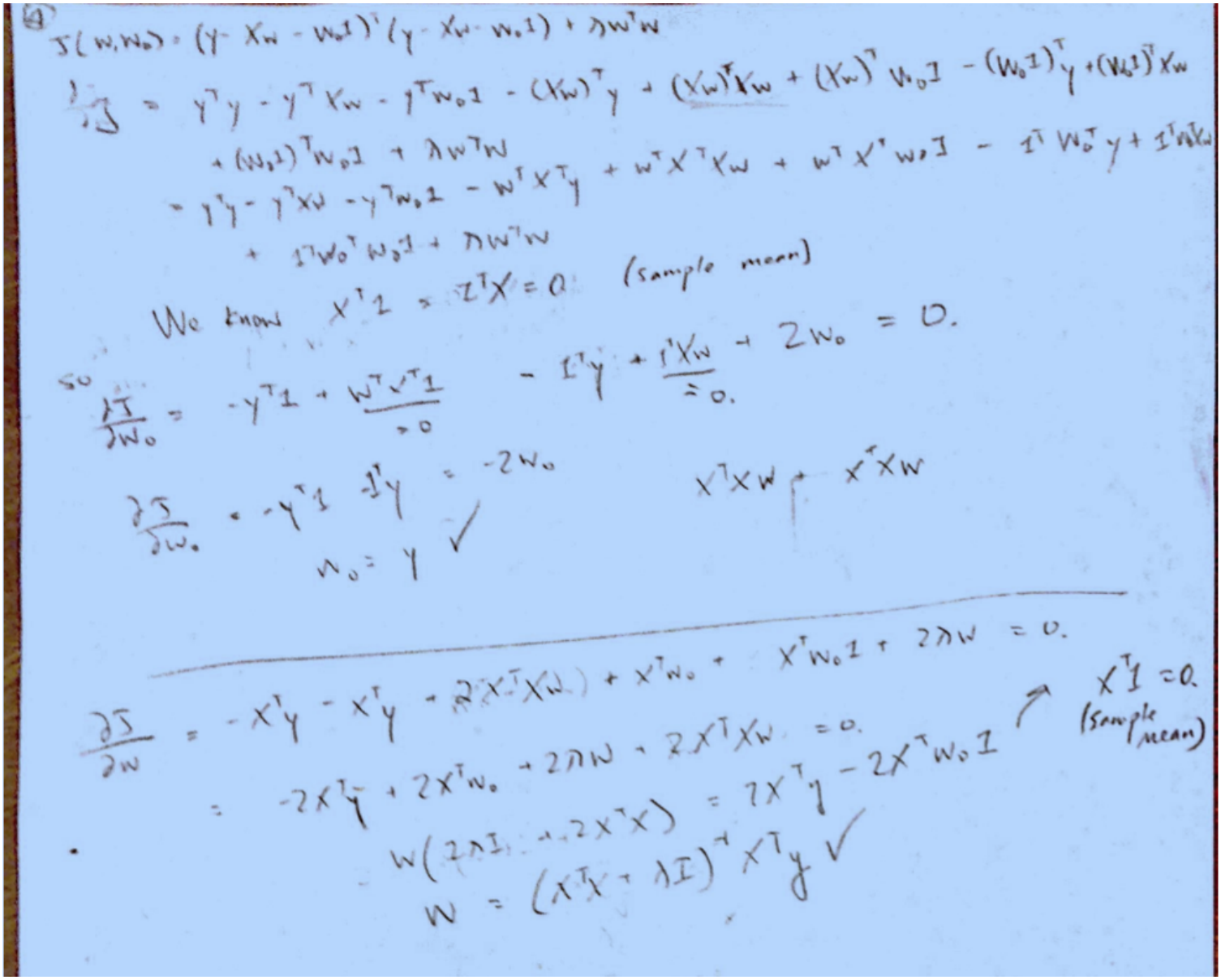
error for average covariances =

0.2681 0.2142 0.1506 0.1369 0.1287 0.1237 0.1225

error for class specific covariances =

0.2156 0.1448 0.0872 0.0630 0.0581 0.0548 0.0512

4.


④
$J(w,w_0) = (y - Xw - w_0 1)^T (y - Xw - w_0 1) + \lambda w^T w$
$= y^T y - y^T Xw - y^T w_0 1 - (Xw)^T y + (Xw)^T Xw + (Xw)^T w_0 1 - (w_0 1)^T y + (w_0 1)^T Xw$
$+ (w_0 1)^T w_0 1 + \lambda w^T w$
$= y^T y - y^T Xw - y^T w_0 1 - w^T X^T y + w^T X^T Xw + w^T X^T w_0 1 - 1^T w_0^T y + 1^T w_0^T Xw$
$+ 1^T w_0^T w_0 1 + \lambda w^T w$
We know $X^T 1 = 1^T X = 0$ (sample mean)
so $\frac{\partial J}{\partial w_0} = -y^T 1 + \underbrace{w^T X^T 1}_{=0} - 1^T y + \underbrace{1^T Xw}_{=0} + 2w_0 = 0.$
$\frac{\partial J}{\partial w_0} = -y^T 1 - 1^T y = -2w_0$
$w_0 = \bar{y}$ ✓
$X^T Xw \rightarrow x^T Xw$
$\frac{\partial J}{\partial w} = -X^T y - X^T y + 2X^T Xw + X^T w_0 + X^T w_0 1 + 2\lambda w = 0.$
$= -2X^T y + 2X^T w_0 + 2\lambda w + 2X^T Xw = 0.$
$w(2\lambda I + 2X^T X) = 2X^T y - 2X^T w_0 1$
$X^T 1 = 0.$ (sample mean)
$w = (X^T X + \lambda I)^{-1} X^T y$ ✓

5.

5. $P(y_i|x_i) \sim N(\mu_i, \sigma^2)$ where $\mu_i = w_0 + w_1 x_i$.

$$L(D|\theta) = \prod_i \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y_i-\mu_i)^2}{2\sigma^2}}$$

$$\log(L) = \sum_i \log\left(\frac{1}{\sigma\sqrt{2\pi}}\right) + -(y_i-\mu_i)^2 \cdot \frac{1}{2\sigma^2}$$

$$\frac{\partial \log(L)}{\partial w_0} = -2(y_i-\mu_i)\cdot\frac{1}{2\sigma^2} = 2(y_i - w_0 - w_1 x_i)$$

set derivative to zero

$$\sum_i y_i - w_0 - w_1 x_i = 0.$$

$$\sum_i w_0 = \sum_i (y_i - w_1 x_i)$$

$$w_0 = \frac{1}{n}\sum_i (y_i - w_1 x_i)$$

$$\boxed{w_0 = \bar{y} - w_1 \bar{x}} \checkmark$$

$$\frac{\partial \log(L)}{\partial w_1} = -\sum_i 2(y_i - w_0 - w_1 x_i) x_i$$

Set derivative to zero.

$$\sum_i y_i x_i - w_0 x_i - w_1 x_i^2 = 0.$$

$$\sum_i y_i x_i - (\bar{y} - w_1 \bar{x}) x_i - w_1 x_i^2 = 0.$$

$$\sum_i y_i x_i - \bar{y} x_i + w_1(\bar{x} x_i - x_i^2) = 0.$$

$$w_1 = \frac{\sum_i (y_i x_i - \bar{y} x_i)}{\sum_i (x_i^2 - x_i \bar{x})}$$

$$\boxed{w_1 = \frac{\sum_i x_i y_i - n\bar{y}\bar{x}}{\sum_i x_i^2 - n\bar{x}^2}} \checkmark$$