67577 | Introduction to Machine Learning | Exercise 2

Guy Lutsker 207029448

Let X be the design matrix of a linear regression problem with p rows (variables) and n columns (samples). Let $y \in \mathbb{R}^n$ be the response vector corresponding the samples in X. Recall that for some vector space $V \subseteq \mathbb{R}^p$ the orthogonal complement of V is: $V^\perp = \{x \in \mathbb{R}^p | \langle x, v \rangle = 0 \ \forall v \in V\}$

$\mathcal{Q}uestion$ 1

Prove that: $Ker\left(X^\top\right) = Ker\left(XX^\top\right)$

Solution :

Let $X \in \mathbb{R}^{pXn}$ be a matrix we will prove $Ker(X^T) \subset Ker(XX^T) \wedge Ker(XX^T) \subset Ker(X^T)$

$$Ker(X^T) \subset Ker(XX^T) :$$

Let $v \in Ker(X^T)$ then $XX^Tv = X \cdot 0 = 0 \rightarrow v \in Ker(XX^T)$

$$Ker(XX^T) \subset Ker(X^T) :$$

let $v \in Ker(XX^\intercal) \rightarrow XX^\intercal v = 0$ lets look at $0 = v^\intercal XX^\intercal v = (X^\intercal v)^\intercal (X^\intercal v) \rightarrow X^\intercal v \overset{vector\ that\ orthogonal\ to\ itself\ is\ the\ 0\ vector}{=} 0 \rightarrow v \in Ker(X^\intercal)$

$\mathcal{Q}uestion$ 2

Prove that for a square matrix A: $Im\left(A^\top\right) = Ker\left(A\right)^\perp$

lets prove the statement:

$$Im(A^T) \subset Ker(A)^\perp :$$

let $v \in Im(A^T)$ and let $w \in Ker(A)$ we need to prove that $\langle v, w \rangle = 0$.
we know that $Aw = 0$ and that $\exists u, s.t\ v = A^Tu = v$
$\langle v, w \rangle = \langle A^Tu, w \rangle \overset{parseval}{=} (A^Tu)^Tw = u^TAw \overset{w \in Ker(A)}{=} u^T \cdot 0 = 0$

$$Ker()^\perp \subset Im(A^T) :$$

let $v \in Ker(A)^{\perp}$, we need to prove that $\exists u, s.t\ v = A^T u$
suppose there is no u s.t $\forall w \in Ker(A)$ it holds that $\langle A^T u, w\rangle \neq 0$
from parseval $\langle A^T u, w\rangle = u^T A w = 0$

Question 3

Let $y = X^{\top} w$ be a non-homogeneous system of linear equations. Assume that $X^{\top}$ is not invertible. Show that the system has ∞ solutions $\Leftrightarrow$ $y \perp Ker\,(X)$.

we know that the system has ∞ solution meaning there exists at least one solution w.
so we know that $y \in Im(A^T)$→from Q2 we know that $Im(A^T) = Ker(A)^{\perp}$→$y \in Ker(A)^{\perp}$
→$y \perp Ker(A)$

Question 4

Consider the (normal) linear system $XX^{\top}w = Xy$. Using what you have proved above prove that the normal equations can only have a unique solution (if $XX^{\top}$ is invertible) or infinitely many solutions (otherwise).

we know that $XX^T w = Xy$, now if $XX^T$is invertible then we can multiply from the left both sides by $(XX^T)^{-1}$ and get $w = (XX^T)^{-1}Xy$ as a unique solution to the equation

if however $XX^T$is invertible, we know that it has ∞ solutions $\Leftrightarrow X^T y \perp Ker(X^T X) \overset{Q1}{=} Ker(X)$
lemma: $X^T y \perp Ker(X)$. proof: let $v \in Ker(X)$ lets look at $\langle X^T y, v\rangle = y^T X v = 0$
so we proved that $X^T y \perp Ker(X)$ which means that the system has ∞ solutions, as required.

Question 5

In this question you will prove some properties of orthogonal projection matrices seen in recitation 1. Let $V \subseteq \mathbb{R}^d$, $dim\,(V) = k$ and let $v_1, \dots, v_k$ be an orthonormal basis of V. Define the orthogonal projection matrix $P = \sum_{i=1}^{k} v_i v_i^\top$ (Notice this is an outer product). Show that:

(a) P is symmetric

(b) The eigenvalues of P are 0 or 1 and that $v_1, \dots, v_k$ are the eigenvectors corresponding the eigenvalue 1

(c) $\forall v \in V \;\; Pv = v$

(d) $P^2 = P$

(e) $(I - P)\,P = 0$

(a) notice that $\forall v \in V,\ v \cdot v^\intercal$ is a symmetric matrix s.t if $v = \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix}$, $[v \cdot v^\intercal]_{i,j} = [a_i \cdot a_j],\ \ v \cdot v^\intercal =$

$$\begin{bmatrix} a_1^2 & \dots & a_1a_i & a_1a_n \\ \vdots & \ddots & & \cdots \\ a_1a_i & \cdots & a_ia_j & a_ia_n \\ a_1a_n & \cdots & \cdots & a_na_n \end{bmatrix}$$

such a matrix is symmetric since $\begin{bmatrix} a_1^2 & \dots & a_1a_i & a_1a_n \\ \vdots & \ddots & & \cdots \\ a_1a_i & \cdots & a_ia_j & a_ia_n \\ a_1a_n & \cdots & \cdots & a_na_n \end{bmatrix}^\intercal = \begin{bmatrix} a_1^2 & \dots & a_1a_i & a_1a_n \\ \vdots & \ddots & & \cdots \\ a_1a_i & \cdots & a_ia_j & a_ia_n \\ a_1a_n & \cdots & \cdots & a_na_n \end{bmatrix}$.

and so P is a sum of symmetric matrices, and its not hard to see that a sum of symmetric matrices is a symmetric matrix as well.

(b) we know that $\mathbb{R}^d = Im(P) \oplus Ker(P)\ s.t, \forall v \in Im(P)^{-1}\ Pv = v$
(didn't see we need to prove it is c, but suppose the proof is here :))
meaning each vector in $\mathbb{R}^d$ is either in the kernel, meaning it is an eigenvector of value 0 or it is in the Image meaning it stays where it was- eigenvector of value 1.
if we take the basis of V and expand to the $\mathbb{R}^d \rightarrow v_1 ... v_k, v_{k+1} ... v_d$ we will notice that $v_1 ... v_k$ span V and so they are in Image(P) meaning they are the eigenvectors with value 1.
and since that a subspace of dim k, that means for a fact that the rest of the basis vectors are in the kernel.

(c) let $v \in V$ we can represent v as a linear combination of the base vectors $v_1 ... v_n$ which are also eigenvectors s.t $v = \sum_{i=1}^{n} a_i v_i$ lets look at :

$$Pv = P(\sum_{i=1}^{n} a_i v_i) = \sum_{i=1}^{n} a_i P v_i \overset{v_i\ is\ an\ eigenvector}{=} \sum_{i=1}^{n} a_i v_i = v$$

(d) lets look at:

$$P^2 = (\sum_{i=1}^{k} v_i v_i^\intercal)^2 = \sum_{j=1}^{k}(v_j v_j^\intercal \cdot \sum_{i=1}^{k} v_i v_i^\intercal)$$

and now for each j we can see that:

$$v_j v_j^\intercal \cdot \sum_{i=1}^{k} v_i v_i^\intercal = \sum_{i=1}^{n} v_j \cdot \langle v_j, v_i \rangle v_i^\intercal \overset{\langle v_j,v_i\rangle=\delta(v_i,v_j)}{=} v_j v_j^\intercal$$

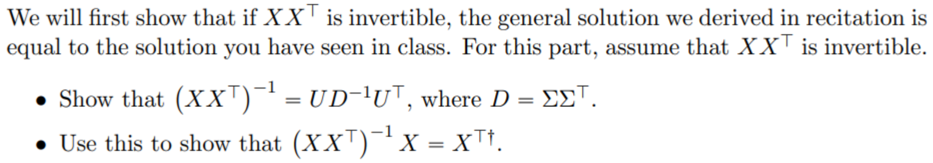
meaning that:

$$P^2 = \sum_{j=1}^{k}(v_j v_j^\intercal \cdot \sum_{i=1}^{k} v_i v_i^\intercal) = \sum_{i=1}^{k} v_i v_i^\intercal = P$$

(e)

$$(I-P)P = (I-P)\sum_{i=1}^{k} v_i v_i^\intercal = \sum_{i=1}^{k}(I-P)v_i \cdot v_i^\intercal \overset{v_i\ eigenvector\ of\ value\ 1}{=} \sum_{i=1}^{k}(v_i - v_i)v_i^\intercal = 0$$

Question 6

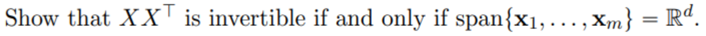


We will first show that if $XX^\top$ is invertible, the general solution we derived in recitation is equal to the solution you have seen in class. For this part, assume that $XX^\top$ is invertible.

- Show that $\left(XX^\top\right)^{-1} = UD^{-1}U^\top$, where $D = \Sigma\Sigma^\top$.
- Use this to show that $\left(XX^\top\right)^{-1}X = X^{\top\dagger}$.

lets look at $(XX^\intercal)^{-1} = (U\Sigma V^\intercal(U\Sigma V^\intercal)^\intercal)^{-1} = (U\Sigma V^\intercal(V\Sigma^\intercal U^\intercal))^{-1} \overset{V\ is\ Orthogonal}{=}$
$(U\Sigma\Sigma^\intercal U^\intercal)^{-1} \overset{denote\ D=\Sigma\Sigma^\intercal}{=} (UD^{-1}U^\intercal)^{-1}$
now lets look at $(XX^\intercal)^{-1}X = (UD^{-1}U^\intercal)^{-1}X = UD^{-1\intercal}U^\intercal U\Sigma V^\intercal = U(\Sigma\Sigma^\intercal)^{-1}\Sigma V^\intercal$
$U\Sigma^{\intercal^{-1}}\Sigma^{-1}\Sigma V^\intercal = U\Sigma^{\intercal^{-1}}V^\intercal \overset{XX^\intercal\ is\ invertible}{=} U\Sigma^{\intercal^\dagger}V^\intercal = (V^\intercal\Sigma^\dagger U)^\intercal = X^{\dagger\intercal} = X^{\intercal^\dagger}$

Question 7



Show that $XX^\top$ is invertible if and only if $\text{span}\{\mathbf{x}_1, \ldots, \mathbf{x}_m\} = \mathbb{R}^d$.

if XX^T is invertible, that means its kernel is 0 and from Q1
we know that means $X^\intercal$ kernel is zero too,
that means its column space is all of $\mathbb{R}^d$ meaning that $span(x_1...x_m) = \mathbb{R}^d$

if $span(x_1...x_m) = \mathbb{R}^d$ then X is full rank meaning its column space and row space
are of rank d meaning that $deg(X) = deg(X^\intercal) = d \rightarrow X^\intercal$has kernel 0
and from Q1 that means XX^T is invertible.

Question 8

Recall that if $XX^\top$ is not invertible then there are many solutions. Show that $\hat{\mathbf{w}} = X^{\top\dagger}\mathbf{y}$ is the solution whose L_2 norm is minimal. That is, show that for any other solution $\overline{\mathbf{w}}$, $\|\hat{\mathbf{w}}\|_2 \leqslant \|\overline{\mathbf{w}}\|_2$ (Why is it ok to do so?)

lets assume that there exists a partition of X's singular values which we will denote as $x_1..x_n$ s.t \$\forall 1 \leq i \leq k\ x_i > 0$ and $\forall k+1 \leq j \leq n\ x_j = 0$
that means that for each solution w their first k components must be identical for them to be a solution to the equation system.

denote $\bar{w} = \begin{bmatrix} \bar{w}_1 \\ \vdots \\ \bar{w}_n \end{bmatrix}$, $\hat{w} = \begin{bmatrix} \hat{w}_1 \\ \vdots \\ \hat{w}_n \end{bmatrix}$ notice that $\forall k+1 \leq j \leq n,\ \bar{w}_j = 0$ and lets look at:

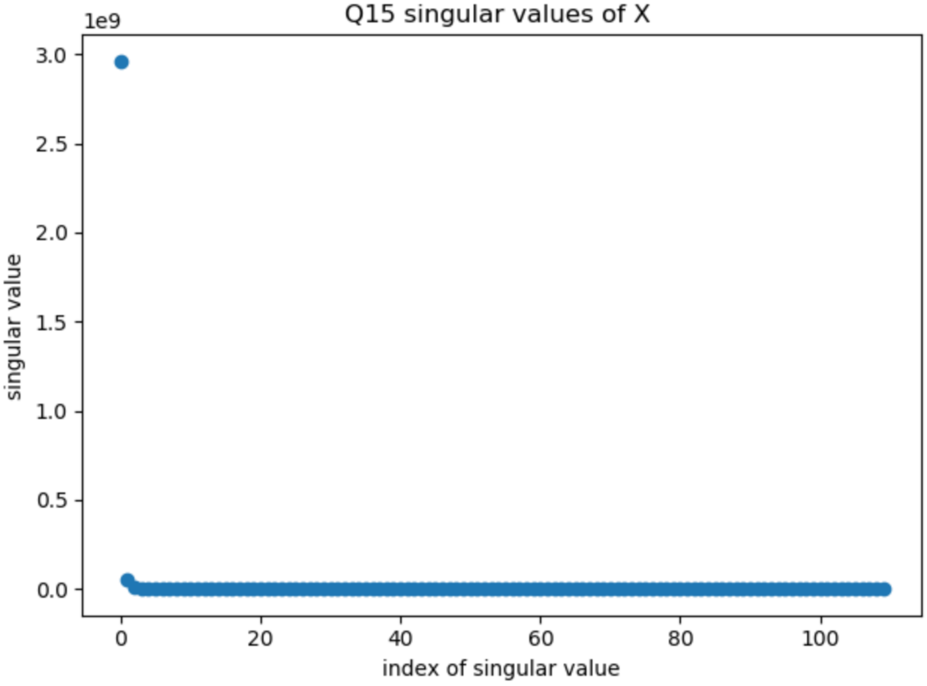
$$\|\hat{w}\| = \sqrt{\sum_{i=1}^{n} (\hat{w_i})^2} = \sqrt{\sum_{i=1}^{k} (\hat{w}_i)^2 + \sum_{i=k+1}^{n} (\hat{w}_i)^2} = \sqrt{\sum_{i=1}^{k} (\hat{w}_i)^2 + \sum_{i=k+1}^{n} 0}$$
$$\leq \sqrt{\sum_{i=1}^{k} (\bar{w_i})^2} = \|\bar{w}\|$$

Question 13

I chose as categorical features zipcode, lan, long.
I chose these since i thought those features couldn't impact the price in a linear way.
meaning maybe these features can affect the price, but its hard to see how it could affect it linearly.
and so i chose to make them dummy features in my data set.

Question 15

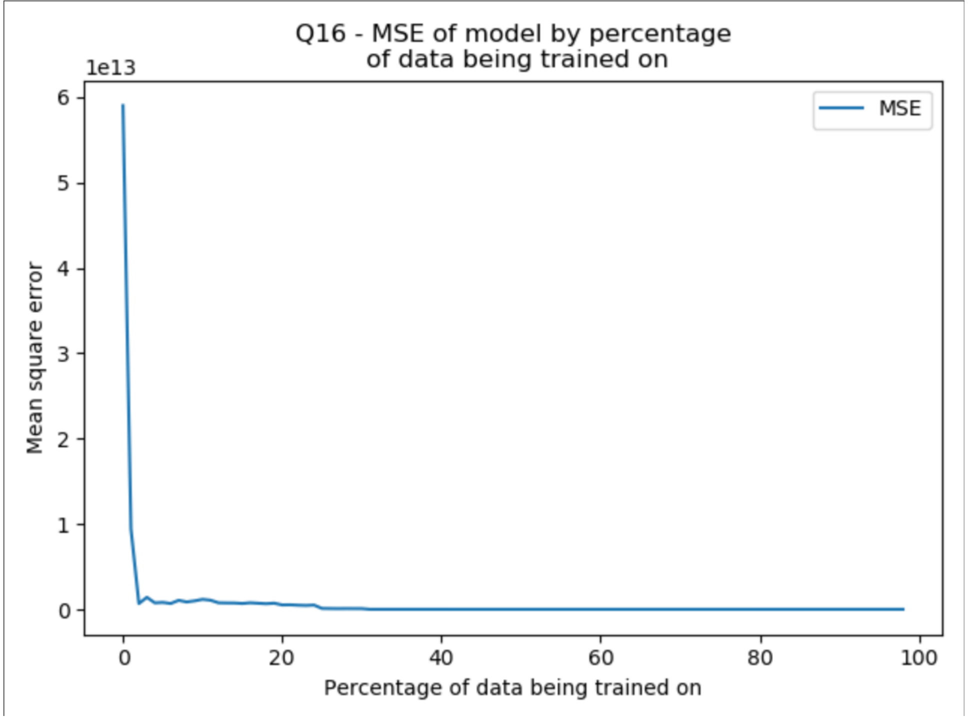
scree plot:



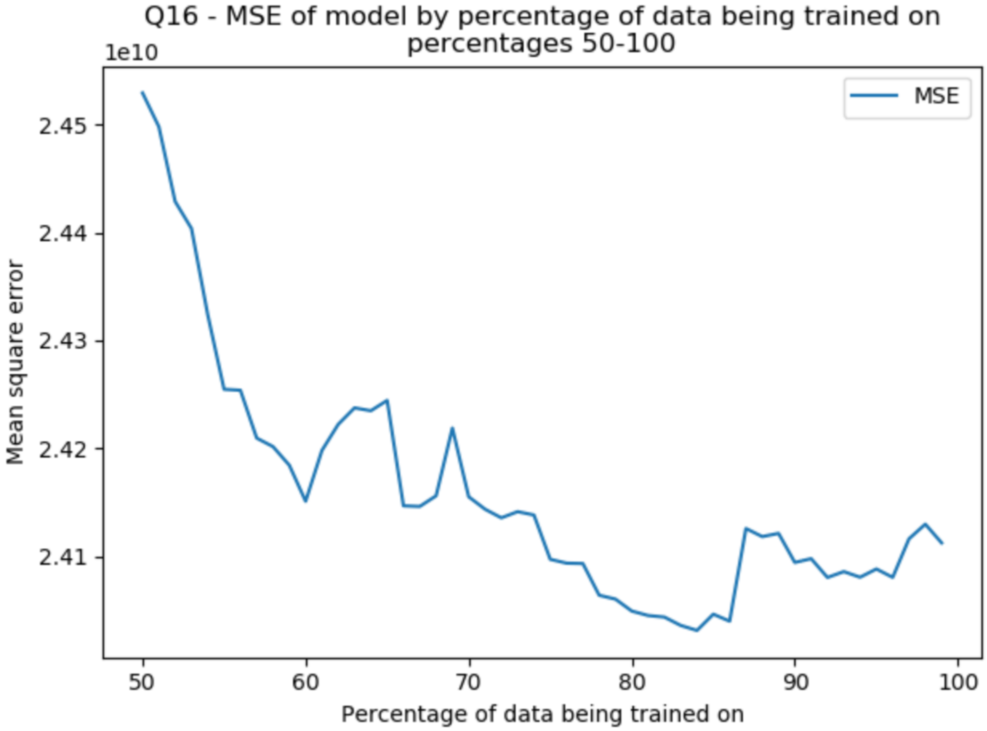
There are a couple of singular values which are zero, and so the matrix is invertible.

Question 16

MSE plot:



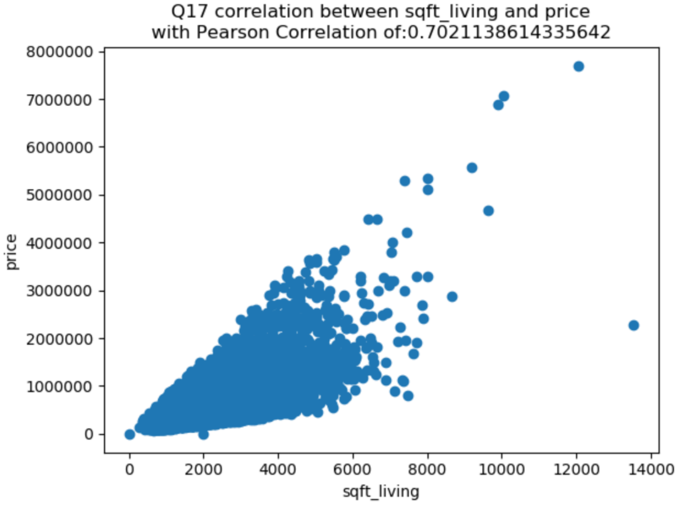
and the last 50 results:

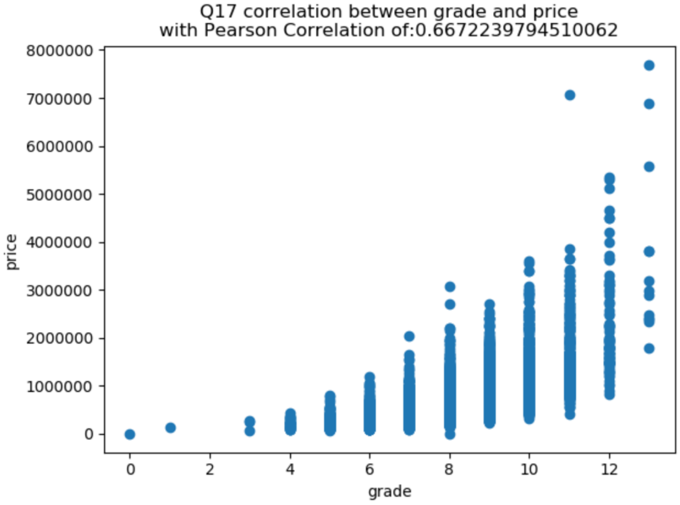


explaining: we can clearly wee that as the percentage of data we train our model on the accuracy increases.
we can see this by the fact that the error rate drops as the percentage increases.

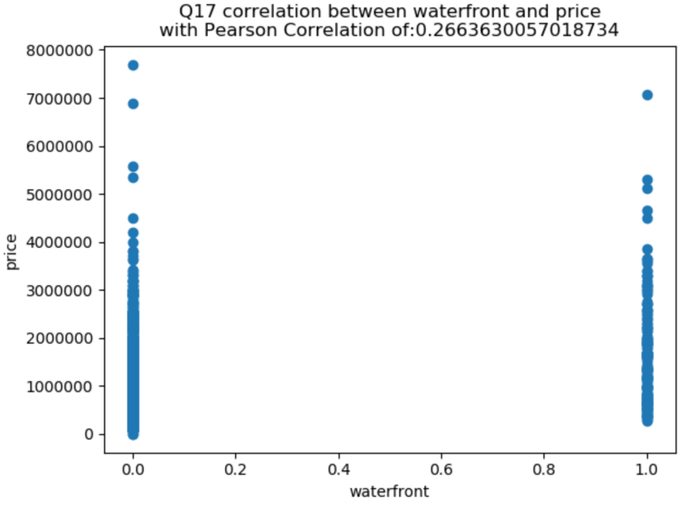
Question 17

we can see that the highest correlation features with price are grade and sqrt_living:


Q17 correlation between sqft_living and price
with Pearson Correlation of:0.7021138614335642

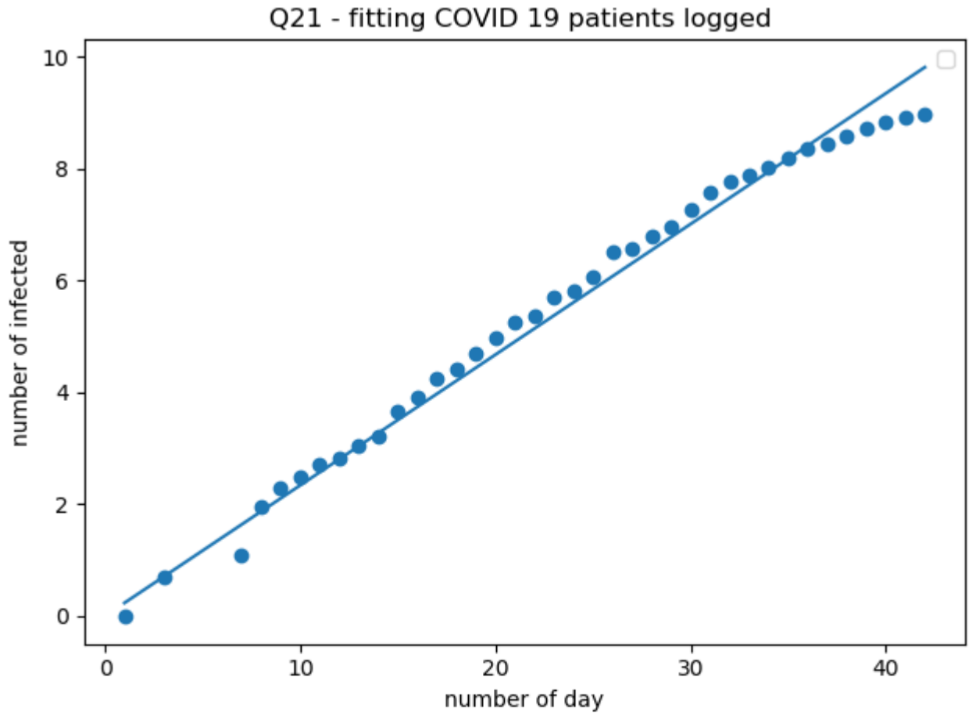

Q17 correlation between grade and price
with Pearson Correlation of:0.6672239794510062

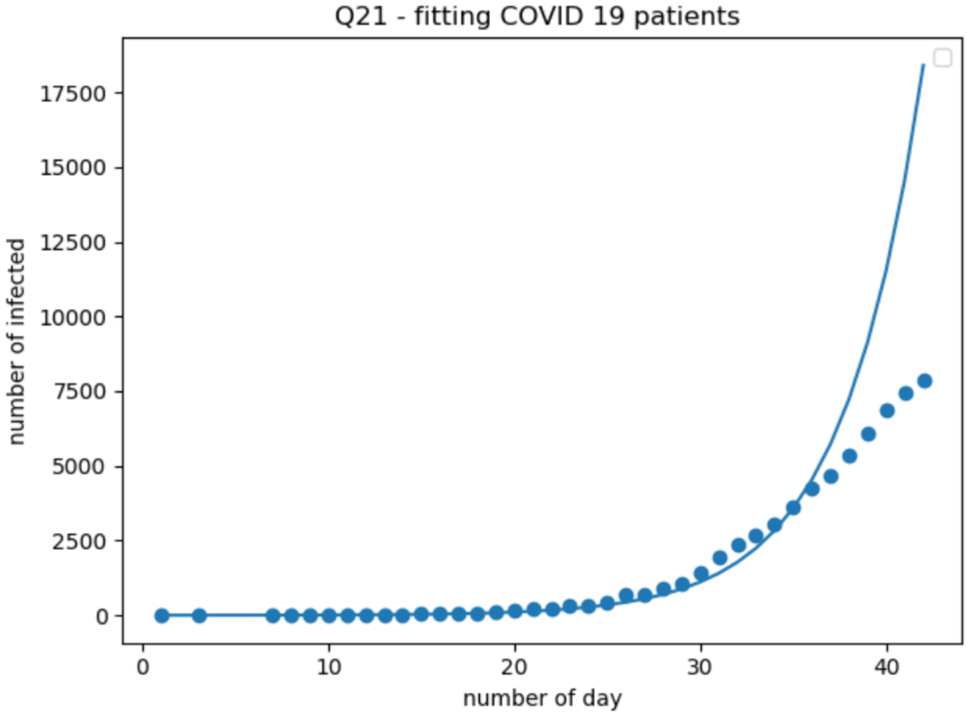
and we can see that waterfront, doesn't seem to be very beneficial(low correlation) for the model:


Q17 correlation between waterfront and price
with Pearson Correlation of:0.2663630057018734

Question 21

Graphs of our model for COVID19:





Question 22

Given a sample (x, y) and a weight $w \in \mathbb{R}$, in this question we used the following loss function:

$$L_{exp}(f_w, (x, y)) = (\langle w, x\rangle - \log(y))^2.$$

If we want to use the least squares loss we saw in class for the exponential regression scenario (i.e., using y instead of $\log(y)$), how should the loss look like? How will you find the ERM solution in that case? (explain shortly the steps you'll preform, you do not need to provide a closed-form solution for the estimator). Answer in the submitted PDF.

we could try to use :

$$L_{exp}(f_w, (x, y)) = (e^{\langle w,x\rangle} - y)^2 = e^{2\langle w,x\rangle} - 2ye^{\langle w,x\rangle} + y^2 = e^{2\sum_{i=1}^n w_i x_i} - 2ye^{\sum_{i=1}^n w_i x_i} + y^2$$

$$= \prod_{i=1}^n e^{2w_i x_i} - 2y\prod_{i=1}^n e^{w_i x_i} + y^2$$

so we can look at:

$$\frac{\partial}{\partial w_i} = 2x_i\prod_{i=1}^n e^{2w_i x_i} - 2yx_i\prod_{i=1}^n e^{w_i x_i} = 2x_i e^{\langle w,x\rangle}(e^{\langle w,x\rangle} - y)$$

and try to find a vector that zeros out the equation, s,t it is a min point.
ans so, for m samples, we need to find $\hat{w} = min\{\sum_{i=1}^m (e^{\langle w,x\rangle} - y_i)^2\}$