IML EX2

Eliyahu Strugo [redacted]

April 29, 2020

1.Solution:

$v \in Ker\left(X^T\right) \iff X^T v = 0_V \iff XX^T v = X0_V = 0_V \iff v \in Ker\left(XX^T\right)$

$\Rightarrow Ker\left(X^T\right) = Ker\left(XX^T\right)$

2.Solution:

$\mathbf{Im}\left(\mathbf{A^T}\right) \subseteq \mathbf{Ker}\left(\mathbf{A}\right)^{\perp}$:

Let $v \in Im\left(A^T\right)$. We have to show that for every $w \in Ker\left(A\right)$ it holds that $\langle v, w\rangle = 0$:

$$v \in Im\left(A^T\right) \Rightarrow \exists u \in V s.t A^T u = v \Rightarrow u^T A = v^T$$

$$\Rightarrow \langle v, w\rangle = v^T w = u^T A w = u^T\left(Aw\right) = u^T 0_V = 0_V$$

$\mathbf{Ker}\left(\mathbf{A}\right)^{\perp} \subseteq \mathbf{Im}\left(\mathbf{A^T}\right)$:

Claim : For any $S, T \leq V$ it holds that $S \subseteq T \Rightarrow T^{\perp} \subseteq S^{\perp}$:

Let $v \in T^{\perp} \Rightarrow \forall w \in T$ it holds that $\langle v, w\rangle = 0$ and since $S \subseteq T$ then $\forall w \in S$ it holds that $\langle v, w\rangle = 0 \Rightarrow v \in S^{\perp}$.

Thus, $Im\left(A^T\right)^{\perp} \subseteq \left(Ker\left(A\right)^{\perp}\right)^{\perp} = Ker\left(A\right) \Rightarrow Ker\left(A\right)^{\perp} \subseteq Im\left(A^T\right)$.

Let $v \in Im\left(A^T\right)^{\perp}$. According to the above claim we have to show that $Av = 0_V$:

$$v \in Im\left(A^T\right)^{\perp} \Rightarrow \forall w \in Im\left(A^T\right)$$

and therefore it holds that :

$$\langle w, v\rangle = 0 \Rightarrow \forall u \in V \ \left\langle A^T u, v\right\rangle = u^T A v = \langle u, Av\rangle = 0$$

.

Hence, for $u = Av$ we get :

$$\langle Av, Av\rangle = 0 \Rightarrow Av = 0 \Rightarrow v \in Ker\left(A\right)$$

as required.

3.Solution:

It's given that X^T is not invertible $\Rightarrow$ the non-homogeneous system $X^T w = y$ has either no solution, or an infite number of solutions.

Therefore, we only have to show that $X^T w = y$ has a solution $\iff y \perp Ker\left(X\right)$:

$y \perp Ker\left(X\right) \iff \forall v \in Ker\left(X\right) \ \langle v, y\rangle = 0 \iff y \in Ker\left(X\right)^{\perp} = Im\left(X^T\right) \iff \exists u \in V,\, u \neq 0 \ s.t \ X^T u = y$

as required.

4.Solution:

$A := XX^T \Rightarrow A$ is a symmetric

$b := Xy$

According to question (1) it holds that $Ker\,(A) = Ker\big(XX^T\big) = Ker\,\big(X^T\big)$ and since A is symmetric then $Ker\,\big(A^T\big) = Ker\,(A) = Ker\,\big(X^T\big)$.

If A is invertible then :

$$Aw = b \Longleftrightarrow A^{-1}Aw = A^{-1}b = \big(X^T\big)^{-1} X^{-1}Xy = \big(X^T\big)^{-1} y \Longleftrightarrow w = \big(X^T\big)^{-1} y$$

$\Rightarrow Aw = b$ has a unique solution.

Otherwise, A is not invertible. Therefore, we can use question (2) which tells us that $Aw = b$ has an infiite number of solutions if and only if $b \perp Ker\,\big(A^T\big) = Ker\big(X^T\big)$.

Indeed, let $u \in Ker\,\big(X^T\big)$ then:

$$\langle b, u\rangle = \langle Xy, b\rangle = y^T X^T u = y^T \big(X^T u\big) = y^T 0_V = 0_V$$

and hence, $b \perp Ker\,\big(A^T\big)$ as required.

5.1.Solution:

For any $A, B \in M_n\,(\mathbb{R})$ it holds that $(A+B)^T = A^T + B^T$.

Therefore,

$$P^T = \left(\sum_{i=1}^{k} v_i v_i^T\right)^T = \sum_{i=1}^{k} \big(v_i v_i^T\big)^T = \sum_{i=1}^{k} \big(v_i^T\big)^T v_i^T = \sum_{i=1}^{k} v_i v_i^T = P.$$

5.2.Solution:

Let f be a linear operator s.t $[f]_E = P$.

It's given that $(v_1 \ldots, v_k)$ is an orthonormal basis of $V \leq \mathbb{R}^p$ so it can be extended by Gram-Schmidt procedure to $\mathcal{B} = (v_1, \ldots, v_p)$ orthogonal basis of $\mathbb{R}^p$.

Let $j \in [p]$.

$$Pv_j = \sum_{i=1}^{k} v_i v_i^T v_j = \sum_{i=1}^{k} v_i \langle v_i, v_j\rangle = \sum_{i=1}^{k} v_i \delta_{ij} = \begin{cases} v_j & j \in [k] \\ 0 & k < j \leq p \end{cases}$$

Hence, $[f]_{\mathcal{B}} = \begin{bmatrix} I_k & 0 \\ 0 & 0 \end{bmatrix}$ and it's eigenvalues are 1 and 0. since P and $[f]_{\mathcal{B}}$ are similar they have the same eigenvalues.

5.3.Solution:

Let f be a linear operator s.t $[f]_E = P$.

According to (5.2) it holds that :

$$\big[f_{|V}\big]_{\mathcal{C}} = I_k = [Id_V]_{\mathcal{C}} \Rightarrow f_{|V} = Id_V.$$

Hence, $\forall v \in V \;\; Pv = v$.

5.4.Solution:

According to (5.2) the eigenvalues of P are $\lambda_0 = 0$ and $\lambda_1 = 1$. Therefore, it holds that $V_{\lambda_0} \oplus V_{\lambda_1} = \mathbb{R}^p$. Let $v \in \mathbb{R}^p$. $V_{\lambda_0} \oplus V_{\lambda_1} = \mathbb{R}^p \Rightarrow \exists u_0 \in V_{\lambda_0}$, $u_1 \in V_{\lambda_1}$ s.t $v = u_0 + u_1$ and then:

$$Pv = P\left(u_0 + u_1\right) = Pu_0 + Pu_1 = 0 + u_1 = u_1$$

and finally:

$$P^2 v = PPv = Pu_1 = u_1 = Pv \Rightarrow P^2 = P.$$

5.5.Solution:

$$P^2 = P \Rightarrow P - P^2 = 0 \Rightarrow I_p P - PP = \left(I_p - P\right) P = 0$$

6.Solution:

$\left(XX^T\right)^{-1} = UD^{-1}U^T$:

$$\begin{aligned} XX^T &= U\Sigma V^T V \Sigma^T U^T = U\Sigma\Sigma^T U^T = UDU^T \\ &\Rightarrow \left(XX^T\right)^{-1} = \left(U^T\right)^{-1} D^{-1} U^{-1} = UD^{-1}U^T \end{aligned}$$

$\left(XX^T\right)^{-1} X = X^{T\dagger}$:

$$\left(XX^T\right)^{-1} X = UD^{-1}U^T U\Sigma V^T = UD^{-1}\Sigma V^T$$

Assume XX^T is invertible then so as D , and since D is also a diagonal matrix then $\forall i \in [d]$ $\sigma_i \neq 0$:

$$\Rightarrow D^{-1}\Sigma = \begin{bmatrix} \sigma_1^{-2} & & \\ & \ddots & \\ & & \sigma_d^{-2} \end{bmatrix}_{d\times d} \begin{bmatrix} diag^{-1}\left(\sigma\right) & 0 \\ 0 & 0 \end{bmatrix}_{d\times m} = \begin{bmatrix} diag^{-1}\left(\sigma\right) & 0 \\ 0 & 0 \end{bmatrix}_{d\times m} = \Sigma^{T\dagger}$$

finally:

$$\left(XX^T\right)^{-1} X = UD^{-1}\Sigma V^T = U\Sigma^{T\dagger}V^T = \left(V\Sigma^\dagger U^T\right)^T = \left(V\Sigma^\dagger U^T\right)^T = \left(X^\dagger\right)^T = X^{T\dagger}$$

as required.

7.Solution:

First we cliam that X is invertible $\iff$ X^T:

X is invertible $\iff d = m$ and $\exists X^{-1}$ s.t $XX^{-1} = X^{-1}X = I_d$, and it holds if and only if :

$$X^T\left(X^{-1}\right)^T = \left(X^{-1}X\right)^T = \left(I_d\right)^T = I_d = \left(I_d\right)^T = \left(X^{-1}\right)^T X^T = \left(XX^{-1}\right)^T$$

Now we can show that XX^T is invertible if and only if $Span\left(x_1, \ldots, x_m\right) = \mathbb{R}^d$:

XX^T is invertible $\iff$ $Ker\left(XX^T\right) = \{0\}$, and according to question(1) it holds if and only if $Ker\left(X^T\right) = \{0\}$.

$Ker\left(X^T\right) = \{0\} \iff X^T$ is invertible $\iff X$ is invertible $\iff Span\left(x_1, \ldots, x_m\right) = Im\left(X\right) = \mathbb{R}^d$.

8.Solution:

Let $\bar{w} \in \mathbb{R}^m$ be a solution , $r = rank\left(X^T\right)$, $z = U^T w$ and $c = V^T y$.

$$\left\|X^T \bar{w} - y\right\|^2 = \left\|V\Sigma^T U^T \bar{w} - y\right\|^2 = \left\|\Sigma^T U^T \bar{w} - V^T y\right\|^2 = \left\|\Sigma^T z - c\right\|^2 = \sum_{i=1}^{r} \left|\sigma_i z_i - c_i\right|^2 + \sum_{i=r+1}^{d} \left|c_i\right|^2$$

Hence, $\bar{w}$ is a solution **if and only if** $\forall i \in [r]$:

$$\sigma z_i - c_i = 0 \Longleftrightarrow z_i = \frac{c_i}{\sigma_i} = \frac{v_i^T y}{\sigma_i}$$

Then for some arbitrary $z_{r+1}, \cdots, z_d \in \mathbb{R}^m$ we get:

$$z = \left(\frac{v_1^T y}{\sigma_1}, \cdots, \frac{v_r^T y}{\sigma_r}, z_{r+1}, \cdots, z_d\right)^T$$

And therefore :

$$z = \Sigma^{T\dagger} V^T y + \sum_{i=r+1}^{d} z_i u_i$$

$$\Rightarrow \bar{w} = Uz = U\Sigma^{T\dagger} V^T y + \sum_{i=r+1}^{d} z_i u_i = X^{T\dagger} y + \sum_{i=r+1}^{d} z_i u_i = \hat{w} + \sum_{i=r+1}^{d} z_i u_i$$

Furthermore,

$$\langle \bar{w}, \hat{w} \rangle = y^T V^T \Sigma^{\dagger} U^T U \begin{bmatrix} 0 \\ \vdots \\ z_{r+1} \\ \vdots \\ z_d \end{bmatrix} = y^T V^T \Sigma^{\dagger} \begin{bmatrix} 0 \\ \vdots \\ z_{r+1} \\ \vdots \\ z_d \end{bmatrix} = 0$$

and hence,

$$\|\hat{w}\|_2^2 \leq \|\hat{w}\|_2^2 + \left\|\sum_{i=r+1}^{d} z_i u_i\right\|_2^2 = \left\|\hat{w} + \sum_{i=r+1}^{d} z_i u_i\right\|_2^2 = \|\bar{w}\|_2^2$$

$\Rightarrow \|\hat{w}\|_2 \leq \|\bar{w}\|_2$ as requried.

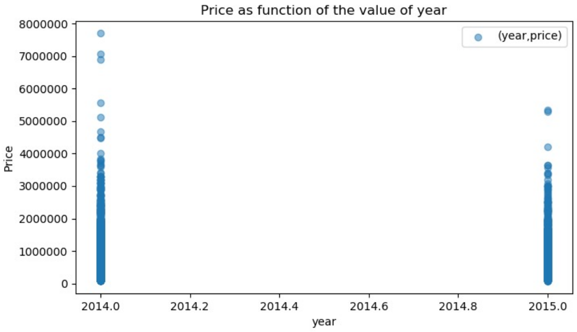
12.Solution:

Checked:

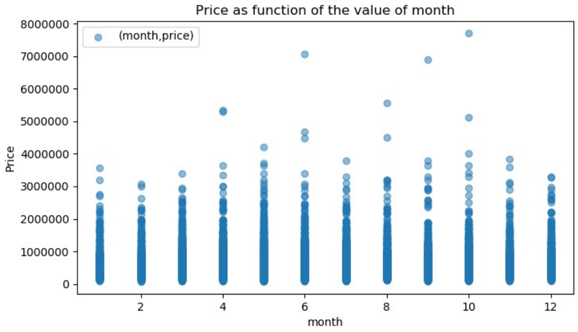
- No duplications
- The price value is integer
- Square values are non-negative
- No 'nan' values
- grad,view and condition in there right range

Droped :

- 'id' : unique values
- zipcode: convert to dummy values (next question)
- date: Since the information on the price for a specific day is either localy or some drastic change (for example new regulations or Tax chagne) that should appear in more consistently in trems of month and years, then i decided to consider only the month and years for examination, and it seems that the year and moth does not have much of effect on the price:



$\rho_{(price, year)} = 0.004872203143780596$

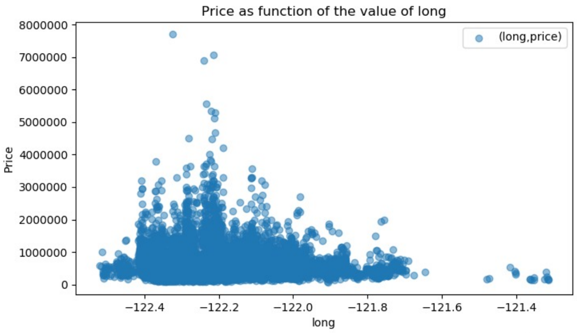


$\rho_{(price, month)} = \text{-}0.010761276721307169$

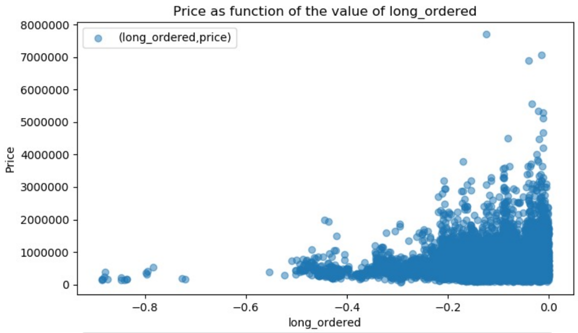
13.Solution:

The categorical feature is actually a single feature - 'zipcode'. Since we lack information on how beging in certain range of zip code affects the price value (if there is any effect), then it is hard to determaine what is the order of the zipcode.

Morever, the 'long' feature seems to have an order, being close to the value -122.2 is 'better' in terms of the price value:

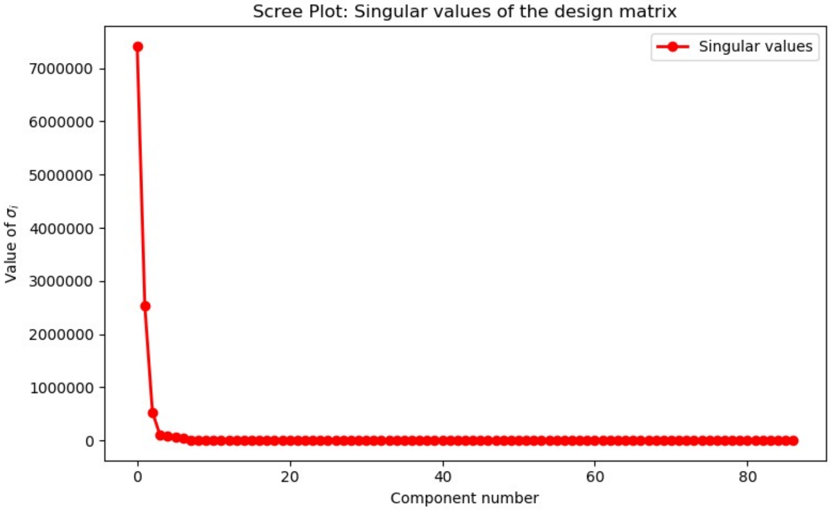


$\rho_{(price, long)} = 0.020607005094107424$



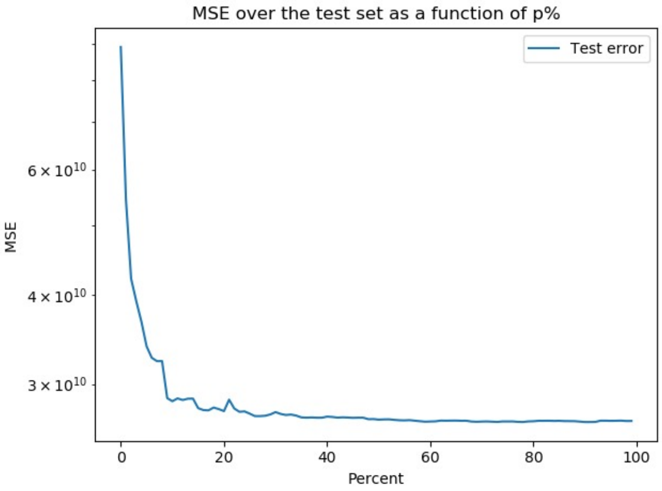
$\rho_{(price, long\_ordered)} = 0.04982239999102782$

15.Solution:



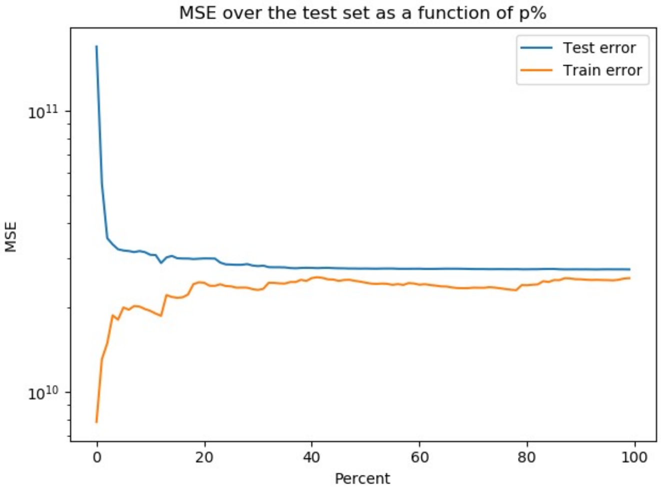
The smallest singular value is $\approx 5 \times 10^{-10}$ which is 'close' to zero and can cause numerical errors when trying to calculate the square matrix $X^T X$ which is a non-singular matrix. The cause for the small values is a high correlation between feartues (and will be described in qusetion 17).

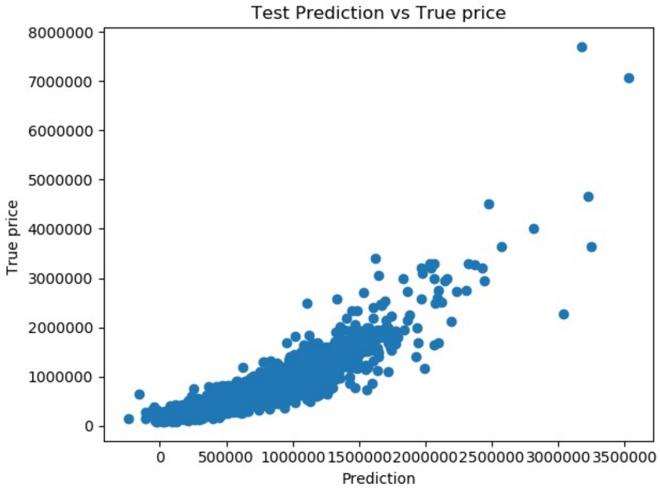
16.Solution:



As expected the error is much higher when there are too few samples. That is, there are too few samples in order to generalize . It seems like the MSE starts to converge when $p \approx 30$.

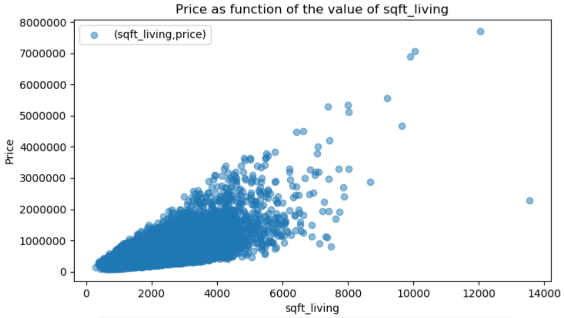
We can consider the train error as well, and compare it to the test error and the predictions compare to the true values :



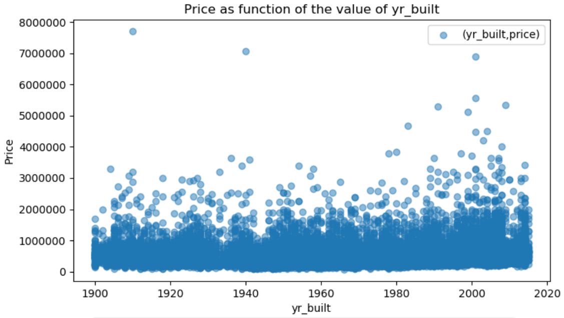


Again as expected, since we use the training set for fitting the model we get better results for the training set, but still it seems to have low variance. Furthermore, by looking at the predictions against the true values it seems that we could decrease the biased if we could get more samples of high prices and deal with outliers.

17.Solution:



$$\boxed{\rho_{(price,\,sqft\_living)} = 0.7020578017894236}$$

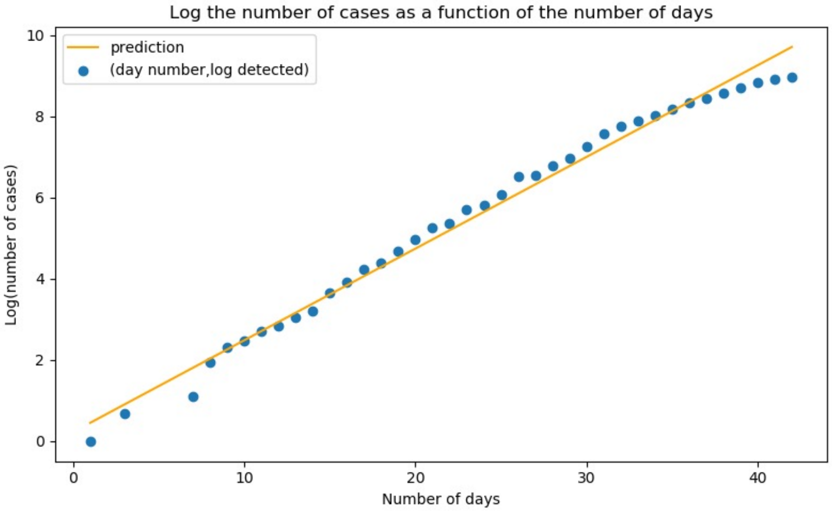


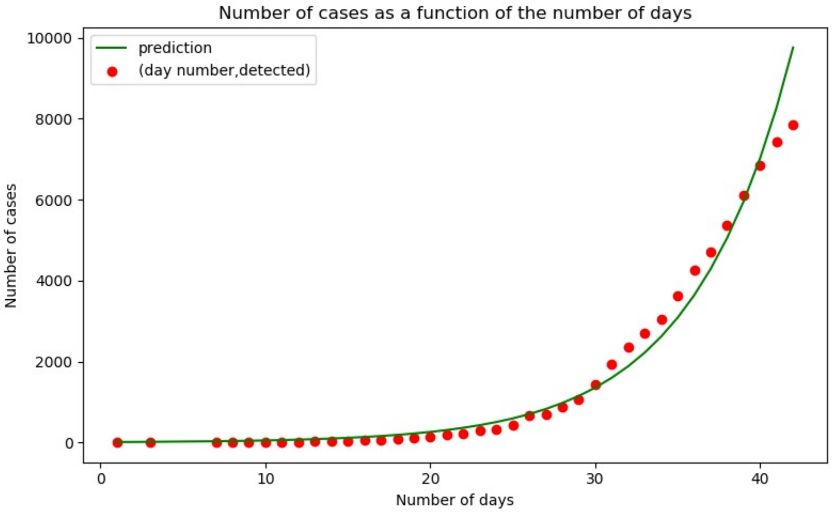
$$\boxed{\rho_{(price,\,yr\_built)} = 0.053083001829779486}$$

Most benficial: I choosed the feature "sqft_living". By looking at the graph it seems that increase in the feature value has the most effect on the value price and it also appears in the value of the Pearson Correlation which has the maximum absolut value.

Least benficial: I choosed the feature "yr_built". By looking at the graph it seems that the values distribution is close unifrom distribution and it's e Pearson Correlation has the maximum absolut value.(except the year featue which I excluded).

21.Solution:





22.Solution:

For the exponential regression we want to minimaize

$$L\left(f_w,(x,y)\right)=\left(e^{\langle w,x\rangle}-y\right)^2$$

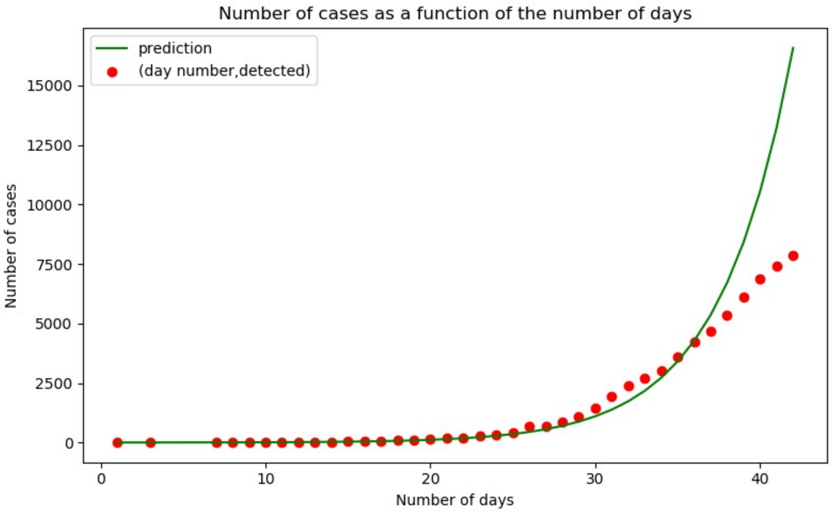
and the ERM becomes:

$$\frac{1}{m}\sum_{i=1}^{m}\left(e^{\langle w,x_i\rangle}-y_i\right)^2$$

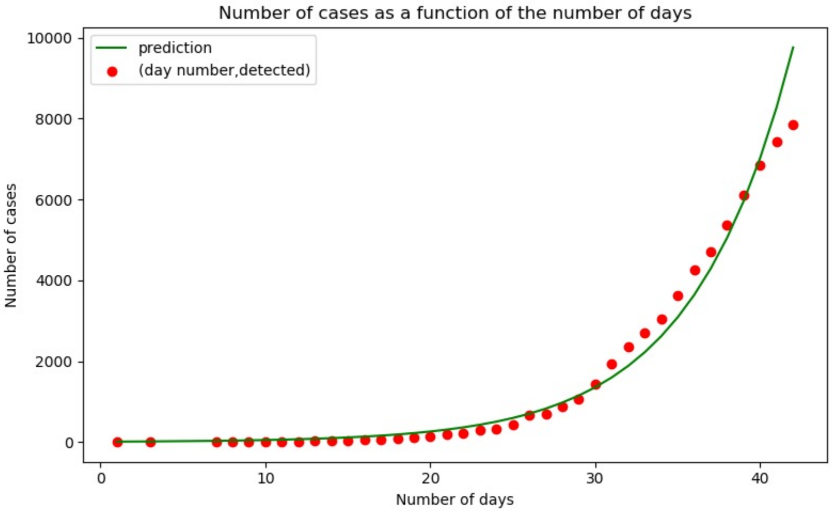
Under the assumtion that $y\approx e^{\langle w,x\rangle}$, If we care of small values of y then in order to find the ERM solution we can minimzie the function $L_{exp}\left(f_w,(x,y)\right)=\left(\langle w,x\rangle-log\left(y\right)\right)^2$. but if this is not the case we might consider to minimize:

$$\sum_i y_i\left(\langle w,x_i\rangle-log\left(y_i\right)\right)^2$$

That way we give less importance for small values of y and get better solution for great values of y, for example in the last question without using this technique we would get :


Number of cases as a function of the number of days
prediction
(day number,detected)
15000
12500
10000
7500
5000
2500
0
Number of cases
0
10
20
30
40
Number of days

Instead of :


Number of cases as a function of the number of days
prediction
(day number,detected)
10000
8000
6000
4000
2000
0
Number of cases
0
10
20
30
40
Number of days