Informe Tarea 1 AM: Javier Castro

1. Pregunta 1

1.1. Parte a

Tenemos un conjunto de datos[1] $D = (x_i, y_i)_{i=1,...,N} \subset \mathbb{R}^{M+1} \times \mathbb{R}$ los cuales modelamos con la relación lineal $y_i = \theta^T x_i + \varepsilon_i(\omega)$ donde $(\varepsilon_i(\omega))_{i=1,...,N}$ son realizaciones de variables aleatorias *iid* con $\mathbb{E}(\varepsilon_i) = 0$ y $\mathbb{V}(\varepsilon_i) = \mathbb{E}(\varepsilon_i^2) = \sigma^2$. Es decir, nuestras mediciones del valor de salida y difieren con el modelo lineal $\theta^T x$ por un error aleatorio, proveniente de factores externos e independientes, que lo modelamos como una variable centrada y de varianza finita. El parámetro a estimar en este modelo es $\theta \in \mathbb{R}^{M+1}$, para esto utilizamos un estimador $\hat{\theta}(\varepsilon_1, ..., \varepsilon_N)$. Luego, para un nuevo par de datos (x, y), podemos predecir y mediante la variable aleatoria $\hat{y} = \hat{\theta}^T x$. Observamos que el modelo también aplica para este nuevo dato, entonces $y = \theta^T x + \varepsilon(\omega)$ con ε otra variable *iid* con los ε_i's (proviene de medición independiente al resto). Notar también que lo único aleatorio en esta descripción son las variables $(\varepsilon_i)_{i=1,...,N}$ y por lo tanto podemos calcular el *costo cuadrático*:

$$\begin{aligned}
&\mathbb{E}[(\hat{y} - y)^2] = \mathbb{E}[(\hat{\theta}^T x - \theta^T x - \varepsilon)^2] \\
&= \mathbb{E}[(\hat{\theta}^T x - \theta^T x)^2] - 2\mathbb{E}[(\hat{\theta}^T x - \theta^T x)\varepsilon] + \mathbb{E}[\varepsilon^2] \\
&= \mathbb{E}[(\hat{\theta}^T x - \theta^T x)^2] - 2\mathbb{E}(\hat{\theta}^T x - \theta^T x)\mathbb{E}(\varepsilon) \\
&+ \mathbb{E}(\varepsilon)^2
\end{aligned}$$

Como ε es independiente a todo lo anterior, podemos multiplicar las esperanzas en el doble producto y este término se nos muere porque todos los ε's involucrados tienen esperanza 0. Además, usamos que $\mathbb{E}(\varepsilon^2) = \sigma^2$.

$$\begin{aligned}
&= \mathbb{E}(\hat{\theta}^T x - \theta^T x)^2 + \mathbb{E}(\varepsilon)^2 \\
&= \mathbb{E}[(\hat{\theta}^T x)^2] - 2\mathbb{E}[\hat{\theta}^T x \theta^T x] + \mathbb{E}[(\theta^T x)^2] + \sigma^2 \\
&= \mathbb{E}[(\hat{\theta}^T x)^2] - (\mathbb{E}[\hat{\theta}^T x])^2 + (\mathbb{E}[\hat{\theta}^T x])^2 - 2\mathbb{E}[\hat{\theta}^T x \theta^T x] \\
&+ [\mathbb{E}(\theta^T x)]^2 + \sigma^2 \\
&= \mathbb{V}(\hat{y}) + (\mathbb{E}[\hat{\theta}^T x] - \mathbb{E}[\theta^T x])^2 + \sigma^2
\end{aligned}$$

Donde usamos que $(\mathbb{E}[\theta^T x])^2 = \mathbb{E}[(\theta^T x)^2] = (\theta^T x)^2$ porque $\theta^T x$ es constante para la esperanza. Si definimos $Sesgo(\hat{y}) := \mathbb{E}(\hat{y} - y)$, entonces podemos escribir la descomposición del costo cuadrático como $\mathbb{E}[(\hat{y} - y)^2] = \mathbb{V}(y) - Sesgo(\hat{y})^2 + \sigma^2$.

[1] Asumimos que a los vectores x's ya se les agregó un 1 al final.

1.2. Parte b

Para el modelo lineal anterior tenemos el estimador de mínimos cuadrados regularizados de θ,

$$\hat{\theta}_{MSR}(\varepsilon_1, ..., \varepsilon_N) = (X^T X + \rho I)^{-1} X^T Y \qquad (1)$$

Donde $X \in \mathbb{R}^{N \times (M+1)}$ es la matriz cuya i-ésima fila es x_i^T e $Y = X\theta + \overrightarrow{\varepsilon}$ donde a su vez, $\overrightarrow{\varepsilon}_i = \varepsilon_i$. Dado un par de nuevos datos (x, y), podemos "predecir" y con $x^T \hat{\theta}_{MSR}$. Definamos la matriz $v^T := x^T (X^T X + \rho I)^{-1} X^T \in \mathbb{R}^{1 \times N}$, la cual no es aleatoria, lo que le da aleatoriedad a $x^T \hat{\theta}_{MSR}$ es el vector $\overrightarrow{\varepsilon}$. Así, podemos calcular varianza y sesgo de la predicción $\hat{y} = \hat{\theta}_{MSR}^T x$.

$$\begin{aligned}
&\mathbb{V}(\hat{\theta}_{MSR}^T x) = \mathbb{V}(x^T \hat{\theta}_{MSR}) = \mathbb{V}(v^T X\theta + v^T \overrightarrow{\varepsilon}) \\
&= \mathbb{E}[(v^T X\theta + v^T \overrightarrow{\varepsilon} - \mathbb{E}(v^T X\theta + v^T \overrightarrow{\varepsilon}))^2] \\
&= \mathbb{E}[(v^T \overrightarrow{\varepsilon})^2] = \mathbb{E}[(\sum_{i=1}^{N} v_i \varepsilon_i)^2] = \mathbb{E}(\sum_{i,j=1}^{N} v_i v_j \varepsilon_i \varepsilon_j) \\
&= \sum_{i,j=1}^{N} v_i v_j \mathbb{E}(\varepsilon_i \varepsilon_j) = \sigma^2 \sum_{i=1}^{N} v_i^2 = \sigma^2 \|v\|^2 \\
&= \sigma^2 \left\| x^T (X^T X + \rho I)^{-1} X^T \right\|^2
\end{aligned}$$

Donde usamos la linealidad de la esperanza, que $\mathbb{E}(\varepsilon_i) = 0$ para $i = 1, ..., N$ y que $v^T X\theta$ no es aleatorio. También se usó que, por independencia, $\mathbb{E}(\varepsilon_i \varepsilon_j) = \delta_{ij} \sigma^2$. Usando este cálculo y notando que $\hat{\theta}_{MS}$ es simplemente $\hat{\theta}_{MSR}$ con $\rho = 0$, se deduce que:

$$\mathbb{V}(\hat{\theta}_{MSR}^T x) = \sigma^2 \left\| x^T (X^T X + \rho I)^{-1} X^T \right\|^2 \qquad (2)$$

$$\mathbb{V}(\hat{\theta}_{MS}^T x) = \sigma^2 \left\| x^T (X^T X)^{-1} X^T \right\|^2 \qquad (3)$$

Ahora el sesgo:

$$\begin{aligned} Sesgo(\hat{\theta}^T_{MSR}x) &= \mathbb{E}(x^T\hat{\theta}_{MSR} - x^T\theta) \\ &= \mathbb{E}(v^TX\theta + v^T\overrightarrow{\varepsilon}) - x^T\theta = \mathbb{E}(v^TX\theta) - x^T\theta \\ &= x^T(X^TX + \rho I)^{-1}X^TX\theta - x^T\theta \end{aligned}$$

Donde se usaron los mismos argumento que para calcular la varianza. Ahora, si reemplazamos $\rho = 0$, obtenemos que los sesgos de cada predicción son:

$$Sesgo(\hat{\theta}^T_{MSR}x) = x^T(X^TX + \rho I)^{-1}X^TX\theta - x^T\theta \quad (4)$$
$$Sesgo(\hat{\theta}^T_{MS}x) = x^T(X^TX)^{-1}X^TX\theta - x^T\theta = 0 \quad (5)$$

1.3. Parte c

Notamos que $\mathbb{V}(\hat{\theta}^T_{MSR}x)$, definida en (2), es decreciente como función de ρ y $Sesgo(\hat{\theta}^T_{MSR}x)$, definida en (4), será distinta de 0 $\forall\rho > 0$. Por lo tanto se produce un *trade off* entre la varianza y el sesgo de nuestra estimación dependiendo del valor de ρ. En consecuencia, si usamos **MSR** tendremos poca varianza (más exactitud) de las estimaciones pero éstas tendrán un sesgo distinto de 0. Por otro lado, si usamos **MS** las estimaciones, como variables aleatorias, tendrán el valor esperado exacto pero se desviarán más de este valor porque tienen más varianza.

2. Pregunta 2

Para un set arbitrario de datos $(x_j, y_j)_{j\in J}$ y estimaciones $\hat{y}_\rho(x_j) = \hat{\theta}^T_{MSR}x_j$ calculamos su **error cuadrático medio** (ecm) en función de ρ como $\frac{1}{|J|}\sum_{j\in J}(y_j - \hat{y}_\rho(x_j))^2$. Notar que $\hat{\theta}_{MSR}$ depende implícitamente de ρ como se observa en (1). Por otro lado, si queremos estimar el parámetro asociado a los errores σ^2, se utilizará el estimador insesgado [2] $\hat{\sigma}^2 = \frac{1}{N-(M+1)}\sum_{i=1}^N(\hat{y}_i - y_i)^2$. A continuación se muestran los gráficos obtenidos, pero antes una tabla que muestra los parámetros encontrados con MS para ambos set de datos.

Tabla 1: Info. parámetros

	Entrenamiento	**Validación**
Pendiente	3.622	21.009
Coef. posición	856,479.106	-217,721.932
Norma	856,479.106	217,721.933

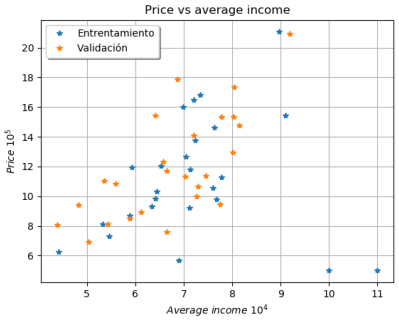

Figura 1: Set de datos completo

Se puede observar una clara tendencia lineal en los datos de *validación*. Por otro lado, los datos de *entrenamiento* presentan dos datos "molestos" que se observan en azul en la parte inferior derecha de la figura (1). Estos datos, al estar en el set de entrenamiento ponen a prueba el método **MSR** y **MS** puesto que ambos dan el mismo peso a los errores, en otras palabras, estos dos datos "tirarán" la recta hacia ellos.
Las siguientes figuras dan cuenta del comportamiento de las estimaciones y parámetros del problema para distintos valores de ρ en el modelo **MSR**.

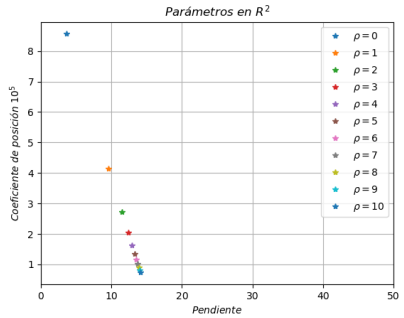

Figura 2: Parámetros en $\mathbb{R}^2$ para distintos valores de ρ

[2] Para demostrar que $\hat{\sigma}^2$ es insesgado, el pazo clave es usar que $tr(ABC) = tr(CBA)$, para A, B y C matrices.

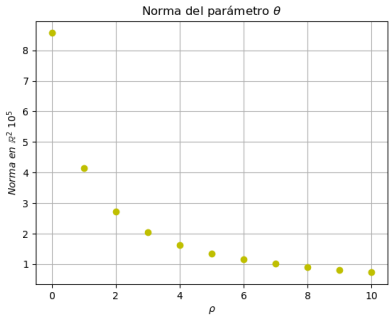

Figura 3: Norma de θ para distintos valores de ρ

En estas imágenes se evidencia la penalización que realiza **MSR** sobre la norma de θ, mientras más grande ρ más penaliza la norma. Es por esto que el parámetro se acerca al origen a medida que ρ aumenta. Notar que la escala de la figura 2 es engañosa, en realidad los puntos van acercándose de forma más rápida al eje.

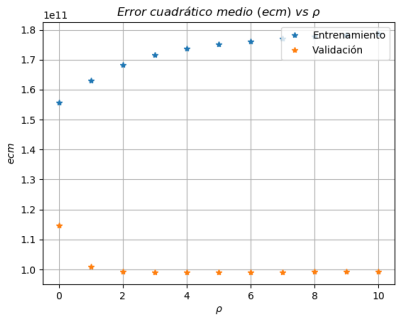

Figura 4: Ecm de la estimación vs ρ con datos de entrenamiento y validación

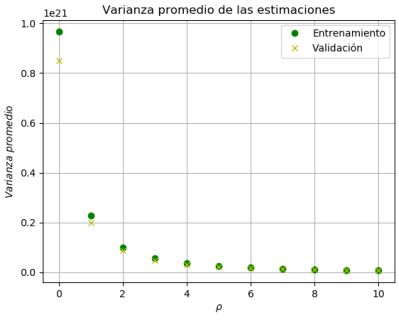

Figura 5: Cada punto del gráfico corresponde al promedio de las varianzas de las estimaciones de los datos de entrenamiento o validación

Para entender la figura (4) recordemos que $\hat{\theta}_{MSR}$ se obtiene de minimizar $\|Y - X\theta\|^2 + \rho\|\theta\|^2$ y por lo tanto cuando ρ crece, $\|\theta\|$ está obligada a disminuir para minimizar, es decir acercarse al origen. Esto hará que θ se aleje del vector que minimiza $\|Y - X\theta\|^2$ y por lo tanto el *ecm* (salvo constantes), aumentará. El comportamiento para los datos de validación es más extraño, como se observa en la tabla 1 el $\hat{\theta}_{MS}$ que se obtiene minimizando con los datos de validación está más cercano al origen que el que se obtiene con datos de entrenamiento. Esto nos dice que al disminuir la norma de θ el *emc* asociado con datos de validación deberá disminuir. No obstante, una vez pasado cierto umbral en la norma, deberá volver a aumentar hasta alcanzar cierta estabilidad al igual que $\|\theta\|$ (Figura 3). Esto último se evidencia graficando hasta un valor más alto de ρ:

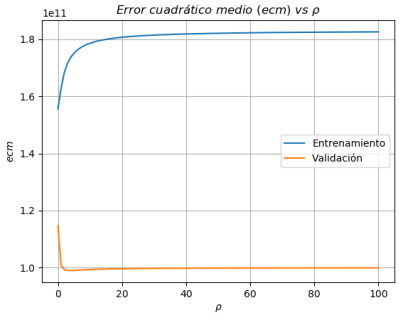

Figura 6: Ecm hasta ρ más alto

Por otro lado, la figura 5 confirma que la varianza calculada en (2) disminuye cuando ρ crece, independiente del dato que se reemplace en (2).

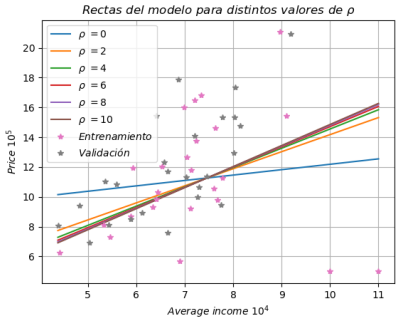

Figura 7: Rectas obtenidas para diferentes valores de ρ junto a datos de entrenamiento y validación

La última figura muestra las distintas rectas encontradas en el modelo.