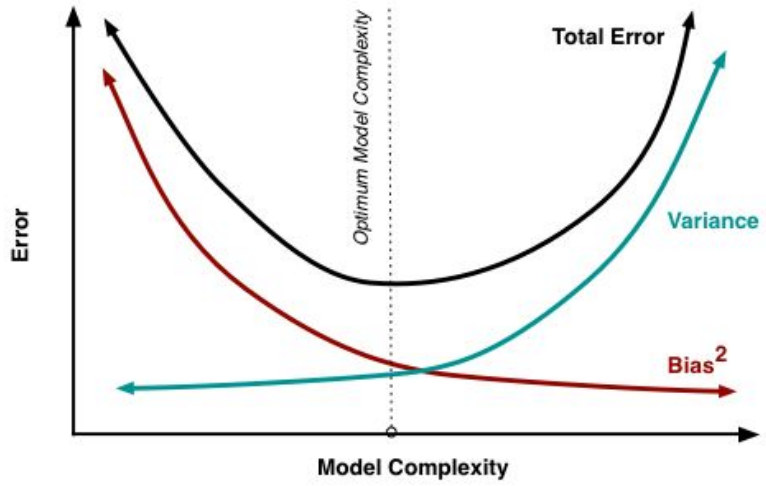
Bias-Variance Tradeoff

El concepto de varianza y sesgo es básico pero de mucha importancia en el campo de ciencia de datos. En la clase vimos que hay una relación inversa entre las dos métricas y también vimos que si el modelo tiene mucho sesgo es por que tenemos un mal ajuste y si es un alto sesgo tenemos sobre ajuste.



Estos dos términos están relacionados a la suma total de errores por medio de la siguiente fórmula,

$$Err(x) = \left(E[\hat{f}\,(x)] - f(x)\right)^2 + E\left[\left(\hat{f}\,(x) - E[\hat{f}\,(x)]\right)^2\right] + \sigma_e^2$$

$$Err(x) = \text{Bias}^2 + \text{Variance} + \text{Irreducible Error}$$

Donde es el modelo poblacional $f(x)$ y el estimado $\hat{f}(x)$.

Bias

El sesgo es la diferencia entre el valor esperado del modelo y el valor correcto que estamos tratando de predecir.

- Al tener solamente una muestra para el entrenamiento es difícil pensar en tener un promedio, es por esto que en clase utilizamos bootstrap para poder encontrar esta media esto está relacionado a $E[\hat{f}(x)]$ de la formula.
- Para $f(x)$ tenemos el mismo problema ya que si conociéramos el modelo que queremos predecir no tendríamos necesidad de hacer la predicción. Para poder poder hacer esto también utilizaremos la técnica de bootstrap.

Varianza

Esta parte del error está relacionada con la variabilidad de nuestra predicción. Una vez más esto está relacionado a tener diferentes muestras para poder determinar la predicción en un punto x_o para luego determinar Var $[\hat{f}(x_0)]$, lo cual hace sentido de la fórmula que tenemos dentro de la suma total de errores.

Bootstrap

Bootstrap es una técnica muy usada para estimar medidas de incertidumbre asociadas a un método estadístico.
Supongamos que tenemos una población de la cual queremos estudiar el parámetro θ si tenemos acceso a la población podríamos tomar n muestras $s_1, s_2, \cdots, s_n$ calcular el parametro $\hat{\theta}$ de cada una de esta muestras y luego promediar estos valores y así obtener una estimación de θ es decir $E[\hat{\theta}] = \theta.$

El problema con esta es que no es factible ir a la población y tomar una nueva muestra y miles de ellas es por esto que se utilizan las muestra que ya se tienen para generar nuevas muestras, utilizando remuestreo con reemplazo. En donde cada punto en nuestra muestras tiene la misma probabilidad de ser seleccionado para nuestra nueva muestra y no se remueve del conjunto de valores del que se puede seleccionar.

Algoritmo

1. Seleccione el número de muestras nuevas que quiere generar B.
2. Seleccione el tamaño de cada muestra n.
3. Utilice el remuestreo con reemplazo para general el conjunto de muestras.
4. Para cada muestra calcule el parámetro a estimar.
5. Calcule el promedio de la estimación, esto nos dará el parámetro poblacional.

El parámetro a estimar dependerá de que es lo que estamos haciendo en el caso de este laboratorio se quiere estimar el sesgo y la varianza de nuestro modelo de regresión.

Estimando la varianza por bootstrap

Para la varianza utilizaremos bootstrap primero para calcular $E[\widehat{f}(x)]$.

1. Para cada muestra calculamos $\hat{f}$ y la llamaremos $\hat{f}_i$
2. Calculamos $\hat{f}_i(x)$ para cada uno de los punto x_j de nuestro train.
3. Calculamos $\frac{1}{B}\sum_{i=1}^{B}\hat{f}_i(x_j)$, $\forall x_j$ esto nos da el valor esperado de la estimacion en cada x_j

Ahora calculamos la varianza para cada $\hat{f}_i$

1. Calculamos Var[$\hat{f}(x_j)$] = $\frac{1}{B-1}\sum_{i=1}^{B}(\hat{f}_i(x_j) - E[\hat{f}(x_j)])^2$, $\forall x_j$. Notar que $E[\hat{f}(x_j)]$ lo calculamos en el paso anterior.
2. Por último calculamos el promedio de cada una de estas varianzas haciendo $\frac{1}{n}\sum_{j=1}^{n} var[\hat{f}(x_j)]$.

Siguiendo estos pasos logramos calcular la varianza del modelo.

Estimando el sesgo

Para estimar el sesgo tenemos que recordar que un estimador puntual tiene que tener la característica que entre mas data se tenga el estimador se aproxima cada vez más al parámetro poblacional que estamos midiendo, entonces

$$\mathbb{E}(\hat{\theta}) - \theta \approx \mathbb{E}(\tilde{\theta}) - \hat{\theta}$$

Entonces podemos usar la media que ya calculamos para la varianza y luego restarle el valor estimado para tener el sesgo, es decir

$$E[\hat{f}(x_j)] = \frac{1}{B}\sum_{i=1}^{B}\hat{f}_i(x_j)$$

Para luego calcular,

$$bias^2[\hat{f}(x_j)] = \left(E[\hat{f}(x_j)] - \hat{f}(x_j) \right)^2$$

Por ultimo calculamos,

$$\frac{1}{n}\sum_{j=1}^{n} bias^2[\hat{f}(x_j)]$$

Ejercicios

Ejercicio 1

Crear una función que calcule la varianza.

Inputs

- Un dataset
- La fórmula de la regresión lineal
- El numero de boots
- El tamaño de cada muestra

Outputs

- La varianza

```
varianza(df,formula,B,n)
```

Ejercicio 2

Crear una función que calcule el sesgo

Inputs

- Un dataset
- La fórmula de la regresión lineal
- El numero de boots
- El tamaño de cada muestra

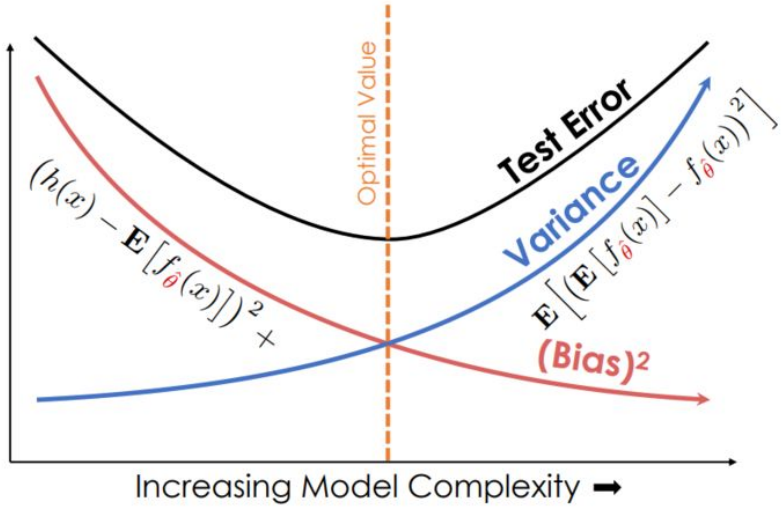
Outputs

- El sesgo

```
sesgo(df,formula,B,n)
```

Ejercicio 3

Definir la complejidad como el grado del polinomio utilizando solamente una variable de input, hacer la siguiente gráfica.



Ejercicio 4

Calcular la varianza y el sesgo de diferentes combinaciones de la variables de input y determinar cuál es el modelo que tiene el mejor balance entre sesgo y varianza.