

# Temperatura Crítica de Superconductores

Presentamos nuevos modelos

Grupo A - Estadística

Pontificia Universidad Católica de Chile  
Facultad de Matemáticas  
EYP2307 - Análisis de Regresión

1 de Diciembre de 2020



# Contenido

Avance 1

Nuevos modelos

Elegimos modelo

Ridge y Lasso Regression

Conclusiones

Referencias bibliográficas



# Contenido

## Avance 1

## Nuevos modelos

## Elegimos modelo

## Ridge y Lasso Regression

## Conclusiones

## Referencias bibliográficas



# Recursos Utilizados

1. Usamos RStudio.
2. R Markdown y R Sweave.
3. GitHub.
4. Bases de datos.
  - ▶ `train.csv`
  - ▶ `unique_m.csv`



# Objetivo Avance 1

- Predecir la temperatura crítica de los superconductores, en base a nuestra variable respuesta `critical_temp`.



# Limpieza de la base de datos

- ▶ Como se tenían **169** variables en total, se decidió limpiar la base de datos.



# Limpieza de la base de datos

- ▶ Como se tenían **169** variables en total, se decidió limpiar la base de datos.
- ▶ Al hacer la limpieza nos quedamos solo con **34** variables.



## Elección del modelo

- ▶ Se hizo un análisis de correlación.





## Elección del modelo

- ▶ Se hizo un análisis de correlación.
- ▶ La variable `std_ThermalConductivity` tuvo la correlación más alta de **0.65**, por lo tanto se utilizó para nuestro modelo de regresión lineal simple.



## Elección del modelo

- ▶ Se hizo un análisis de correlación.
- ▶ La variable `std_ThermalConductivity` tuvo la correlación más alta de **0.65**, por lo tanto se utilizó para nuestro modelo de regresión lineal simple.
- ▶ Al hacer el análisis de la varianza explicada:  $R^2 = \mathbf{0.43}$ .



## Elección del modelo

- ▶ Se hizo un análisis de correlación.
- ▶ La variable `std_ThermalConductivity` tuvo la correlación más alta de **0.65**, por lo tanto se utilizó para nuestro modelo de regresión lineal simple.
- ▶ Al hacer el análisis de la varianza explicada:  $R^2 = \mathbf{0.43}$ .
- ▶ Se decidió buscar alternativas para intentar aumentar este último valor.



# Buscando Alternativas

- ▶ Nos decidimos por un nuevo modelo.



## Buscando Alternativas

- ▶ Nos decidimos por un nuevo modelo.
- ▶ Utilizamos la variable `range_Valence` por ser una variable discreta y así nos quedaron **7** modelos.



## Buscando Alternativas

- ▶ Nos decidimos por un nuevo modelo.
- ▶ Utilizamos la variable `range_Valence` por ser una variable discreta y así nos quedaron **7** modelos.
- ▶ El modelo final nos quedó:



## Buscando Alternativas

- ▶ Nos decidimos por un nuevo modelo.
- ▶ Utilizamos la variable `range_Valence` por ser una variable discreta y así nos quedaron **7** modelos.
- ▶ El modelo final nos quedó:
  - ▶  $\rho = \mathbf{0.75}$ .



# Buscando Alternativas

- ▶ Nos decidimos por un nuevo modelo.
- ▶ Utilizamos la variable `range_Valence` por ser una variable discreta y así nos quedaron **7** modelos.
- ▶ El modelo final nos quedó:
  - ▶  $\rho = \mathbf{0.75}$ .
  - ▶  $R^2 = \mathbf{0.56}$ .





## Objetivo del Avance 2

- Predecir la temperatura crítica de los superconductores en base a nuestra variable respuesta, utilizando modelos de regresión lineal múltiple para mejorar los resultados obtenidos en el Avance 1.



# Contenido

Avance 1

Nuevos modelos

Elegimos modelo

Ridge y Lasso Regression

Conclusiones

Referencias bibliográficas



# Nuevos modelos

- ▶ Creamos una serie de nuevos modelos de regresión lineal múltiple:
  1. *Backward*.



# Nuevos modelos

- ▶ Creamos una serie de nuevos modelos de regresión lineal múltiple:
  1. *Backward*.
  2. *Forward*.



# Nuevos modelos

- ▶ Creamos una serie de nuevos modelos de regresión lineal múltiple:
  1. *Backward*.
  2. *Forward*.
  3. *Backward-Forward*.



# Nuevos modelos

- ▶ Creamos una serie de nuevos modelos de regresión lineal múltiple:
  1. *Backward*.
  2. *Forward*.
  3. *Backward-Forward*.
  4. *add1*.



# Nuevos modelos

- ▶ Creamos una serie de nuevos modelos de regresión lineal múltiple:
  1. *Backward*.
  2. *Forward*.
  3. *Backward-Forward*.
  4. *add1*.
  5. *drop1*.



# Nuevos modelos

- ▶ Creamos una serie de nuevos modelos de regresión lineal múltiple:
  1. *Backward*.
  2. *Forward*.
  3. *Backward-Forward*.
  4. *add1*.
  5. *drop1*.
  6. *VIF*.





# Nuevos modelos

- ▶ Creamos una serie de nuevos modelos de regresión lineal múltiple:
  1. *Backward*.
  2. *Forward*.
  3. *Backward-Forward*.
  4. *add1*.
  5. *drop1*.
  6. *VIF*.
  7. Modelo con la idea del Avance **1**.



# Nuevos modelos

- ▶ Creamos una serie de nuevos modelos de regresión lineal múltiple:
  1. *Backward*.
  2. *Forward*.
  3. *Backward-Forward*.
  4. *add1*.
  5. *drop1*.
  6. *VIF*.
  7. Modelo con la idea del Avance 1.
  8. *Ridge Regression*.



# Nuevos modelos

- ▶ Creamos una serie de nuevos modelos de regresión lineal múltiple:
  1. *Backward*.
  2. *Forward*.
  3. *Backward-Forward*.
  4. *add1*.
  5. *drop1*.
  6. *VIF*.
  7. Modelo con la idea del Avance 1.
  8. *Ridge Regression*.
  9. *Lasso Regression*.



# Nuevos modelos

- ▶ Se utilizó la base de datos limpiada en el Avance **1** para trabajar solo con **34** variables.



## Modelo con *Backward*

- ▶ Criterio AIC.
- ▶ Multicolinealidad → **0** variables eliminadas.
- ▶ Modelo conformado finalmente por **27** variables.



## Modelo con *Forward*

- ▶ Criterio AIC.
- ▶ Multicolinealidad → **0** variables eliminadas.
- ▶ Modelo conformado finalmente por **28** variables.



## Modelo con *Backward-Forward*

- ▶ Criterio AIC.
- ▶ Multicolinealidad → **0** variables eliminadas.
- ▶ Modelo conformado finalmente por **27** variables.



## Modelo con add1

- ▶ Criterio AIC.
- ▶ Multicolinealidad → **0** variables eliminadas.
- ▶ Modelo conformado finalmente por **28** variables.





# Modelo con drop1

- ▶ Criterio AIC.
- ▶ Multicolinealidad → **0** variables eliminadas.
- ▶ Modelo conformado finalmente por **25** variables.



## Modelo con $VIF$

- ▶ Se consideró el modelo conformado por todas las variables de la base de datos.
- ▶ Se fue eliminando el problema de multicolinealidad progresivamente.
- ▶ Modelo conformado finalmente por **28** variables.



# Modelo con la idea del Avance 1

- ▶ Se crearon **7** bases de datos según `range_Valence`.
- ▶ Se creó un modelo para cada base de datos mediante selección *Backward* con criterio AIC.
- ▶ Cantidad de variables:
  1. Modelo para `range_Valence` = 0: **25** variables.
  2. Modelo para `range_Valence` = 1: **23** variables.
  3. Modelo para `range_Valence` = 2: **24** variables.
  4. Modelo para `range_Valence` = 3: **25** variables.
  5. Modelo para `range_Valence` = 4: **22** variables.
  6. Modelo para `range_Valence` = 5: **19** variables.
  7. Modelo para `range_Valence` = 6: **27** variables.



# Contenido

Avance 1

Nuevos modelos

**Elegimos modelo**

Ridge y Lasso Regression

Conclusiones

Referencias bibliográficas



## AIC, BIC y $R^2$ de los modelos.

Modelo	AIC	BIC	$R^2$
<i>Backward</i>	126489.9	127064.9	0.66
<i>Forward</i>	126880.5	127103.5	0.66
<i>Backward-Forward</i>	126849.9	127064.9	0.66
add1	126858.4	127081.4	0.66
drop1	126849.6	127048.7	0.66
VIF	126880.5	127103.5	0.6
Idea Avance 1	121021.6	121840.3	0.74



## Supuesto de *Independencia*

- ▶ Se utilizó el Test de *Durbin-Watson*.
- ▶ Independencia de residuos  $\Leftrightarrow$  Valor D entre **1.5** y **2.5**.
- ▶ Ningún modelo mencionado cumple este supuesto.



## Supuesto de *Normalidad*

- ▶ Se utilizó el Test de *Kolmogorov-Smirnov*.
- ▶ Criterio: Valor-p  $> 0.05$ .
- ▶ Ningún modelo mencionado cumple este supuesto.
- ▶ Primera solución aplicada: Transformación de *Box-Cox*
- ▶ Segunda solución aplicada:



# Diapositiva

► abc





# Diapositiva

► abc



# Diapositiva

► abc



# Diapositiva

► abc



# Diapositiva

► abc



# Diapositiva

► abc



# Diapositiva

► abc



# Contenido

Avance 1

Nuevos modelos

Elegimos modelo

Ridge y Lasso Regression

Conclusiones

Referencias bibliográficas



# Ridge Regression

- ▶ Objetivo: Minimizar **RSS**.
- ▶ *Shrinkage Penalty* :  $RSS_{\text{Ridge}} = RSS_{\text{AMC}} + \lambda \sum_{j=1}^p \beta_j^2$ .
  - ▶  $\lambda = \mathbf{0}$  :  $RSS_{\text{Ridge}} = RSS_{\text{AMC}}$ .
  - ▶  $\lambda \geq \mathbf{0}$  : Impacto en valores de  $\beta$ .
  - ▶  $\lambda \rightarrow \infty$  :  $\beta \rightarrow \vec{\mathbf{0}}$ .



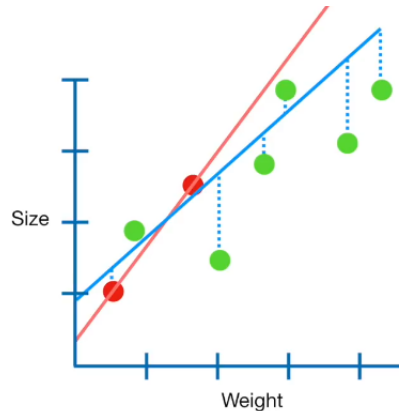


## Ridge Regression: $\lambda$ óptimo

- ▶ Es aquel que reduce la mayor varianza del modelo sin apenas perder ajuste.
- ▶ *Validación cruzada.*



## Ridge Regression: Visualización



## *Ridge Regression: Ventajas*

- ▶ Reduce la varianza.
- ▶ Datos de Entrenamiento vs. Datos de Prueba.
- ▶ Minimiza la influencia sobre el modelo de los predictores menos relacionados con la variable respuesta.



## *Ridge Regression: Limitación*

- ▶ Modelo final incluye todos los predictores.



# Lasso Regression

- ▶ Misma idea que en *Ridge Regression*.
- ▶ Selección de predictores.
- ▶ *Shrinkage Penalty* :  $RSS_{\text{Lasso}} = RSS_{\text{AMC}} + \lambda \sum_{j=1}^p |\beta_j|$ .



## Comparación entre *Ridge* y *Lasso Regression*

- ▶ Usamos uno u otro dependiendo del escenario.
- ▶ *Ridge Regression*: cuando los  $\beta \neq \vec{0}$  y tienen la misma magnitud aproximadamente.
- ▶ *Lasso Regression*: cuando un gran grupo de parámetros  $\approx \mathbf{0}$ .



# Resultados de la implementación en R

► *Ridge Regression:*



# Contenido

Avance 1

Nuevos modelos

Elegimos modelo

Ridge y Lasso Regression

Conclusiones

Referencias bibliográficas





# Conclusiones

► abc



# Contenido

Avance 1

Nuevos modelos

Elegimos modelo

Ridge y Lasso Regression

Conclusiones

Referencias bibliográficas



## Referencias bibliográficas



<https://online.stat.psu.edu/stat501/lesson/13/13.1>

*Weighted Least Squares.*

2018



[https://rpubs.com/Joaquin\\_AR/242707](https://rpubs.com/Joaquin_AR/242707)

*Selección de predictores: Ridge y Lasso.*

2016



<https://rstatisticsblog.com/data-science-in-action/machine-learning/ridge-regression-in-r/>

*Simple Guide To Ridge Regression In R.*

2020

