

# Temperatura Crítica de Superconductores

*¿Es suficiente una regresión múltiple?*

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

Elección de un modelo

Ridge Regression

Extra

Conclusiones

Referencias bibliográficas



# Contenido

## Avance 1

Nuevos modelos

Elección de un modelo

Ridge Regression

Extra

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

## Resumen del Avance 1

- ▶ El objetivo era predecir la Temperatura Crítica de los Superconductores, con un modelo de regresión lineal simple.
- ▶ Se limpió la base de datos: de **169** variables se pasaron a **34**.
- ▶ Se hizo un modelo de regresión simple con la variable `std_ThermalConductivity`, ya que es modelo con mejor  $R^2$  respecto `critical_temp` ( $R^2 = \mathbf{0.43}$ ).

## Resumen del Avance 1

- ▶ Buscamos alternativas para mejorar el  $R^2$ .
- ▶ Se crearon **7** bases de datos según `range_Valence`, ya que es la variable categórica que mejores correlaciones nos da.
- ▶ Finalmente obtuvimos **7** modelos para predecir la variable respuesta, con un  $R^2$  conjunto igual a **0.56**.

## Objetivo del Avance 2

- ▶ Predecir la temperatura crítica de los superconductores en base a nuestra variable respuesta, aplicando nuevas herramientas para mejorar los resultados obtenidos en el Avance 1.



# Contenido

Avance 1

Nuevos modelos

Elección de un modelo

Ridge Regression

Extra

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. Extras.

## Nuevos modelos

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



## Nuevos modelos

- ▶ Se utilizó la base de datos limpiada en el Avance 1 para trabajar solo con **34** variables.
- ▶ Se solucionó el problema de multicolinealidad en cada modelo viendo el *VIF* (Excepto en *Ridge* y *Extras*).

## Nuevos modelos

- ▶ Se utilizó la base de datos limpiada en el Avance 1 para trabajar solo con **34** variables.
- ▶ Se solucionó el problema de multicolinealidad en cada modelo viendo el *VIF* (Excepto en *Ridge* y *Extras*).
- ▶ En todos los modelos se usó criterio AIC (excepto en Modelo con *VIF*).

## Modelo con *Backward*

- ▶ Multicolinealidad → **2** variables eliminadas.



## Modelo con *Backward*

- ▶ Multicolinealidad → **2** variables eliminadas.
- ▶ Modelo conformado finalmente por **27**  $\beta$ 's.

## Modelo con *Forward*

- ▶ Multicolinealidad → **4** variables eliminadas.



## Modelo con *Forward*

- ▶ Multicolinealidad → **4** variables eliminadas.
- ▶ Modelo conformado finalmente por **28**  $\beta$ 's.

## Modelo con *Backward-Forward*

- ▶ Multicolinealidad → **2** variables eliminadas.



## Modelo con *Backward-Forward*

- ▶ Multicolinealidad → **2** variables eliminadas.
- ▶ Modelo conformado finalmente por **27**  $\beta$ 's.



## Modelo con add1

- ▶ Multicolinealidad → **3** variables eliminadas.

## Modelo con add1

- ▶ Multicolinealidad → **3** variables eliminadas.
- ▶ Modelo conformado finalmente por **28**  $\beta$ 's.

## Modelo con drop1

- ▶ Multicolinealidad → **1** variable eliminada.

## Modelo con drop1

- ▶ Multicolinealidad → **1** variable eliminada.
- ▶ Modelo conformado finalmente por **25**  $\beta$ 's.

## Modelo con *VIF*

- ▶ Se consideró el modelo conformado por todas las variables de la base de datos.

## 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 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.



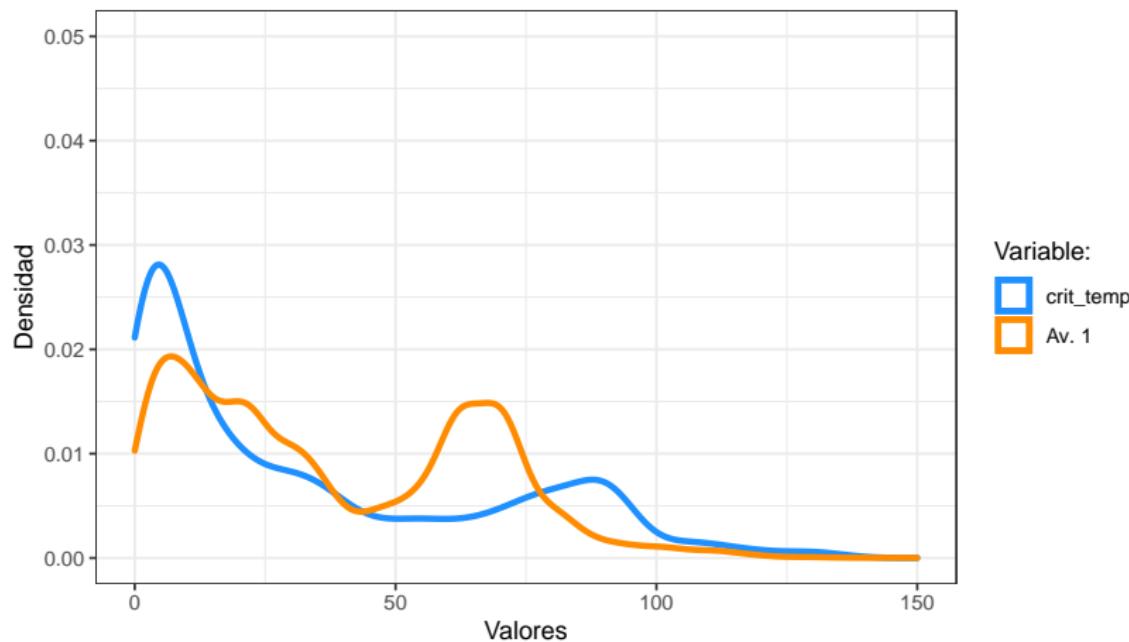
## Modelo con la idea del Avance 1

- ▶ Se crearon **7** bases de datos según `range_Valence`.
- ▶ Se creó un modelo para cada base de datos mediante selección *Backward*.

## 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*.
- ▶ 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.

## Modelo Avance 1: Densidad de Ajustados y Reales



# Contenido

Avance 1

Nuevos modelos

Elección de un modelo

Ridge Regression

Extra

Conclusiones

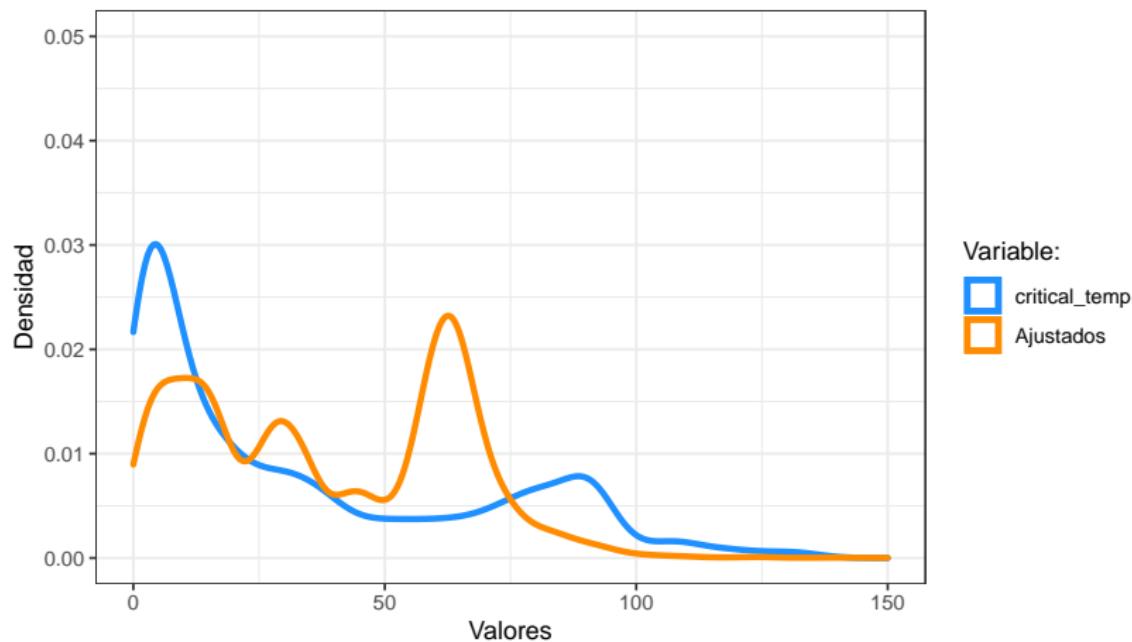
Referencias bibliográficas



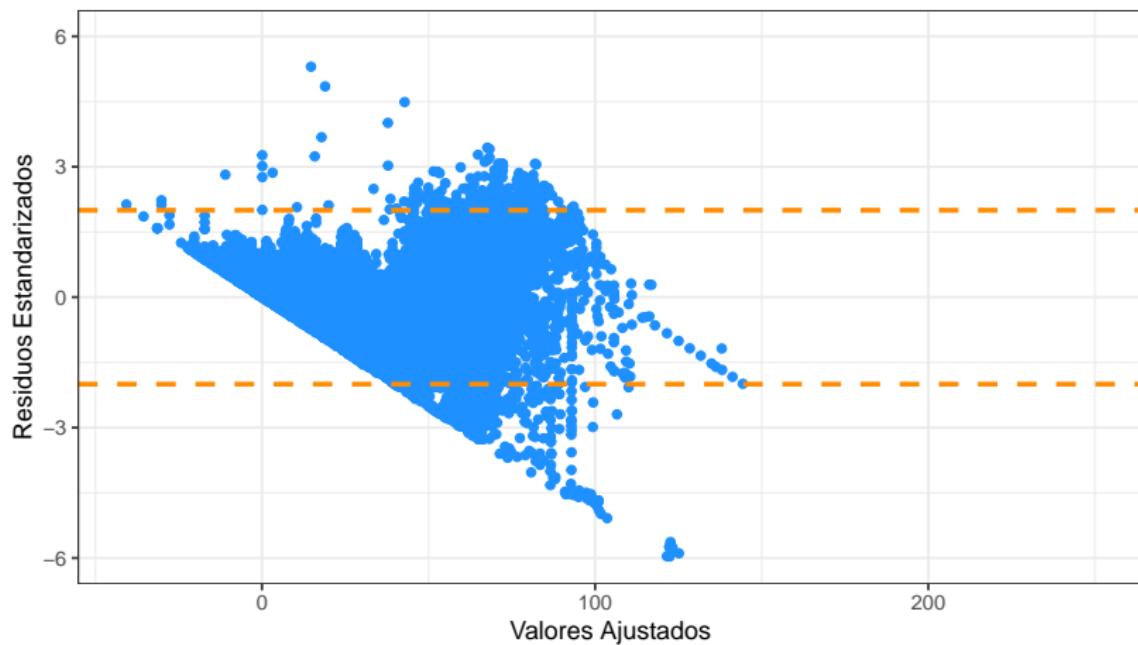
## AIC, BIC y $R^2$ de los modelos.

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	0.66
Idea Avance 1	121021.6	121840.3	0.74

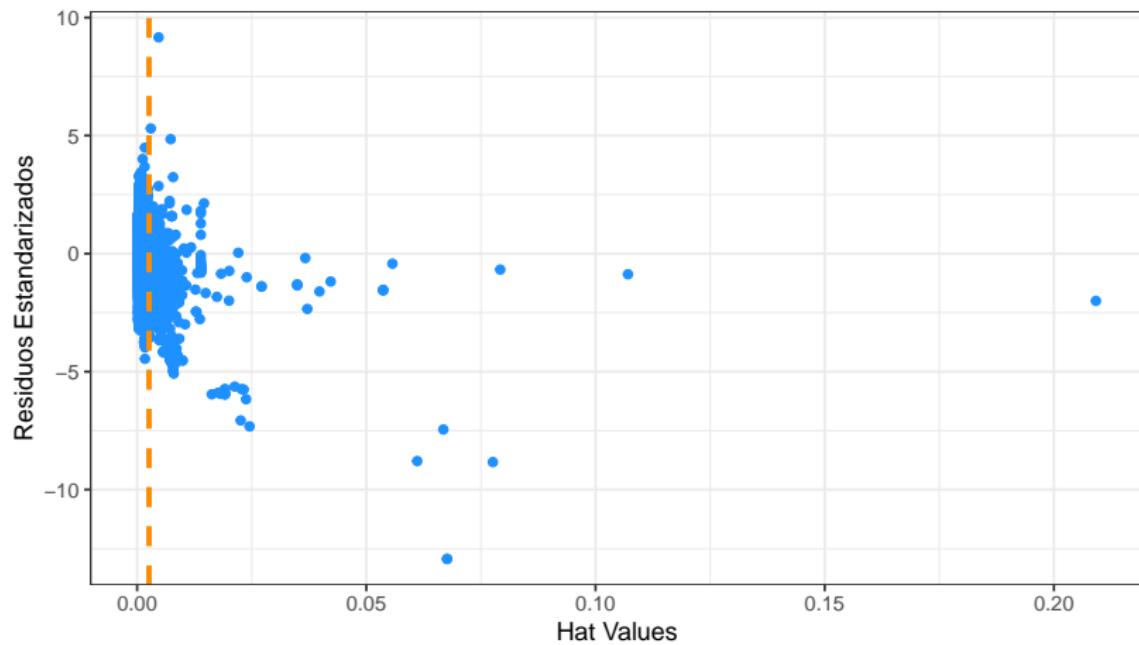
## Modelo Elegido: *Backward*



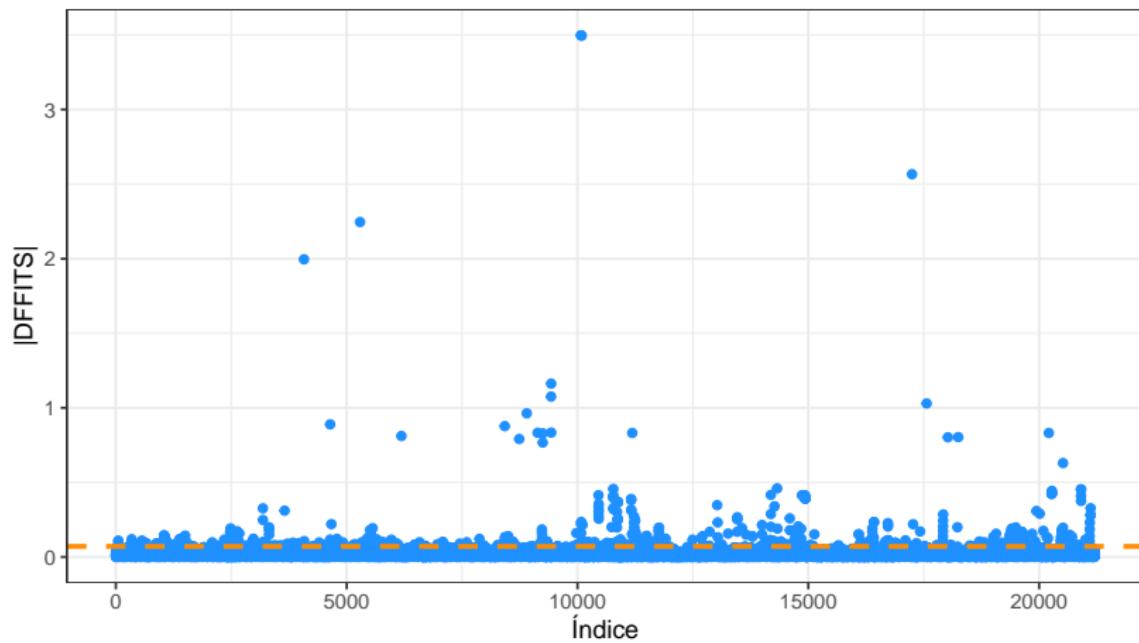
## Análisis de Puntos: *Outliers*



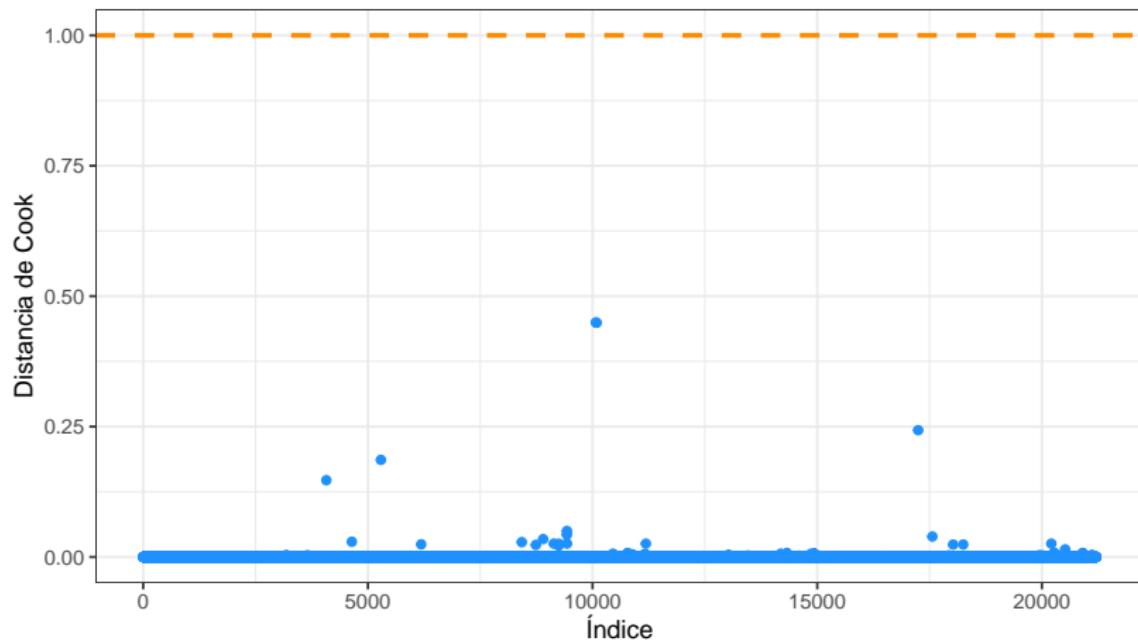
## Análisis de Puntos: *Leverage*



## Análisis de Puntos: *DFFITS*



## Análisis de Puntos: *Distancia de Cook*



## Supuesto de *Independencia*

- ▶ Se utilizó el Test de *Durbin-Watson*.



## Supuesto de *Independencia*

- ▶ Se utilizó el Test de *Durbin-Watson*.
- ▶ Independencia de residuos  $\Leftrightarrow$  valor-p > **0.05**.

## Supuesto de *Independencia*

- ▶ Se utilizó el Test de *Durbin-Watson*.
- ▶ Independencia de residuos  $\Leftrightarrow$  valor-p > **0.05**.
- ▶ No se cumple el supuesto.

## Supuesto de *Normalidad*

- ▶ Se utilizó el Test de *Kolmogorov-Smirnov*.



## Supuesto de *Normalidad*

- ▶ Se utilizó el Test de *Kolmogorov-Smirnov*.
- ▶ Criterio: valor-p > **0.05**.



## Supuesto de *Normalidad*

- ▶ Se utilizó el Test de *Kolmogorov-Smirnov*.
- ▶ Criterio: valor-p > **0.05**.
- ▶ El modelo no cumple con este supuesto.

## Supuesto de *Normalidad*

- ▶ Se utilizó el Test de *Kolmogorov-Smirnov*.
- ▶ Criterio: valor-p > **0.05**.
- ▶ El modelo no cumple con este supuesto.
- ▶ *Primera solución aplicada:* Transformación de *Box-Cox*.

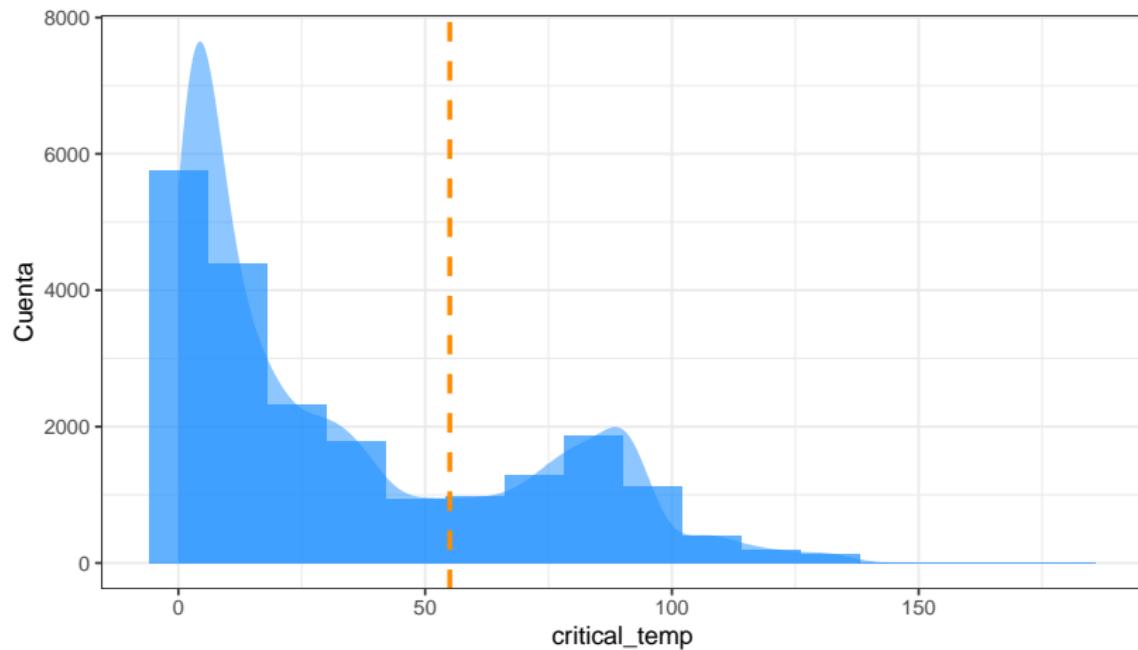
## Supuesto de *Normalidad*

- ▶ Se utilizó el Test de *Kolmogorov-Smirnov*.
- ▶ Criterio: valor-p > **0.05**.
- ▶ El modelo no cumple con este supuesto.
- ▶ *Primera solución aplicada:* Transformación de *Box-Cox*.
- ▶ *Segunda solución aplicada:* Transformación de *Johnson*.

## Supuesto de *Normalidad*

- ▶ Se utilizó el Test de *Kolmogorov-Smirnov*.
- ▶ Criterio: valor-p > **0.05**.
- ▶ El modelo no cumple con este supuesto.
- ▶ *Primera solución aplicada:* Transformación de *Box-Cox*.
- ▶ *Segunda solución aplicada:* Transformación de *Johnson*.
- ▶ Tercera solución aplicada:

## Supuesto de *Normalidad*



## Supuesto de *Homocedasticidad*

- ▶ Se utilizó el Test de *Breusch-Pagan*.



## Supuesto de *Homocedasticidad*

- ▶ Se utilizó el Test de *Breusch-Pagan*.
- ▶ Criterio: valor-p > **0.05**.



## Supuesto de *Homocedasticidad*

- ▶ Se utilizó el Test de *Breusch-Pagan*.
- ▶ Criterio: valor-p > **0.05**.
- ▶ El modelo no cumple con este supuesto.

## Supuesto de *Homocedasticidad*

- ▶ Se utilizó el Test de *Breusch-Pagan*.
- ▶ Criterio: valor-p > **0.05**.
- ▶ El modelo no cumple con este supuesto.
- ▶ Solución propuesta para Heterocedasticidad:

*Weighted Least Squares Regression*.

# Contenido

Avance 1

Nuevos modelos

Elección de un modelo

Ridge Regression

Extra

Conclusiones

Referencias bibliográficas



## Ridge Regression

- ▶ **Objetivo:** Minimizar RSS.



## 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  $\beta$ .
  - ▶  $\lambda \rightarrow \infty$  :  $\beta \rightarrow \vec{\mathbf{0}}$ .



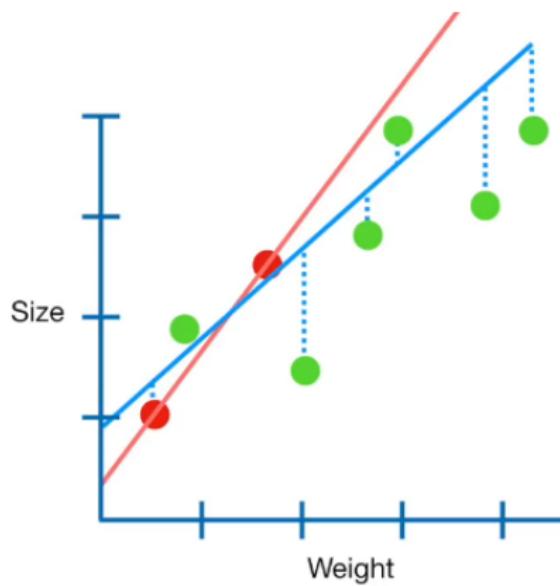
## Ridge Regression: $\lambda$ óptimo

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

## 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.



## Ridge Regression: Ventajas

- ▶ Reduce la varianza.
- ▶ Datos de Entrenamiento vs. Datos de Prueba.



## 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*.



## Lasso Regression

- ▶ Misma idea que en *Ridge Regression*.
- ▶ Realiza selección de predictores.

## Lasso Regression

- ▶ Misma idea que en *Ridge Regression*.
- ▶ Realiza 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.



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

- ▶ Usamos uno u otro dependiendo del escenario.
- ▶ *Ridge Regression*: cuando los  $\beta'$ s  $\neq \mathbf{0}$  y tienen la misma magnitud aproximadamente.

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

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

## Resultados de la implementación en R

- ▶ Se usó el package `glmnet`.



## Resultados de la implementación en R

- ▶ Se usó el package `glmnet`.
- ▶ Se usó la misma fórmula que el modelo resultante con *Backward* en *Ridge Regression*.



## Resultados de la implementación en R

- ▶ Se usó el package `glmnet`.
- ▶ Se usó la misma fórmula que el modelo resultante con *Backward* en *Ridge Regression*.
- ▶ En *Lasso Regression* se consideró el modelo completo.

## Resultados de la implementación en R

- ▶ El  $\lambda$  óptimo en los modelos nos dió:



## Resultados de la implementación en R

- ▶ El  $\lambda$  óptimo en los modelos nos dió:
  - ▶ *Ridge Regression:* **0.05.**
  - ▶ *Lasso Regression:* **0.01.**

## Resultados de la implementación en R

- ▶ El  $\lambda$  óptimo en los modelos nos dió:
  - ▶ *Ridge Regression:* **0.05.**
  - ▶ *Lasso Regression:* **0.01.**
- ▶ Veamos algunos coeficientes importantes de los modelos.

## Coeficientes de los modelos en orden creciente

	Coeficiente	<i>Backward</i>	<i>Ridge</i>	<i>Lasso</i>
Intercepto		-48.2633	-46.8465	-44.4442
gmean_ThermalConductivity		-0.3338	-0.3301	-0.3171
:				
wtd_range_atomic_radius				0
:				
wtd_entropy_TConductivity		6.9492	6.7131	7.684503
Ba		9.3430	9.3538	10.6381

## Comparación entre modelos

- ▶ La función que nos permite hacer *Ridge* y *Lasso Regression* no nos aporta información suficiente para calcular la *Log-Verosimilitud*.

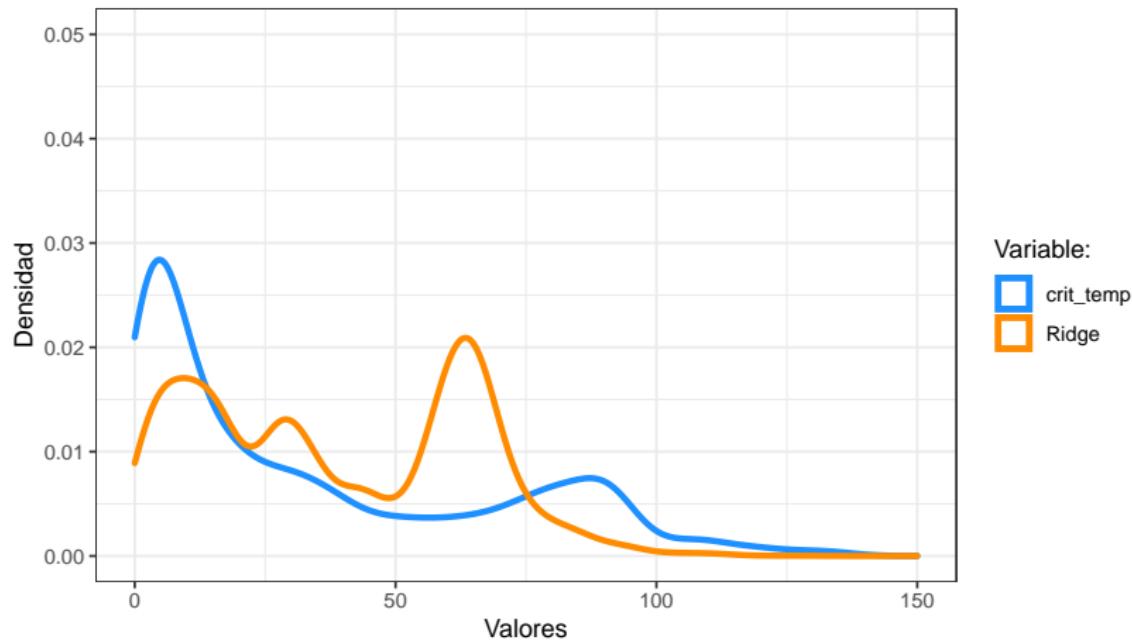
## Comparación entre modelos

- ▶ La función que nos permite hacer *Ridge* y *Lasso Regression* no nos aporta información suficiente para calcular la *Log-Verosimilitud*.
- ▶ Un criterio de comparación es el  $R^2$  ajustado:

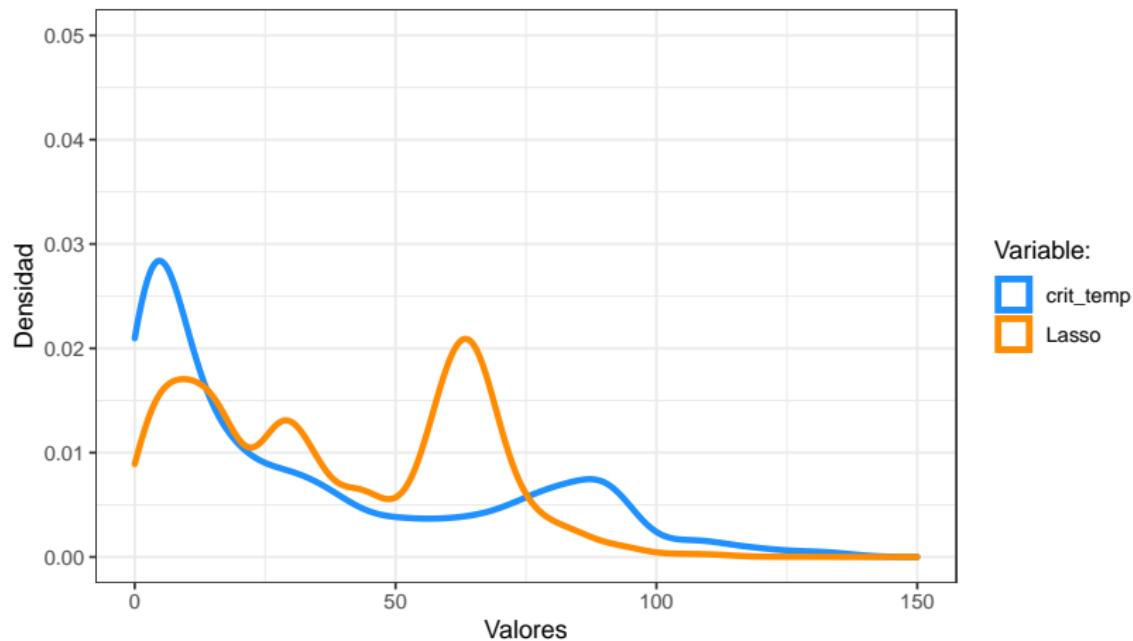
## Comparación entre modelos

- ▶ La función que nos permite hacer *Ridge* y *Lasso Regression* no nos aporta información suficiente para calcular la *Log-Verosimilitud*.
- ▶ Un criterio de comparación es el  $R^2$  ajustado:
  - ▶ *Backward*: **0.66**.
  - ▶ *Ridge Regression*: **0.66**.
  - ▶ *Lasso Regression*: **0.65**.

## Ridge Regression: Densidad de Ajustados y Reales



## Lasso Regression: Densidad de Ajustados y Reales



# Contenido

Avance 1

Nuevos modelos

Elección de un modelo

Ridge Regression

Extra

Conclusiones

Referencias bibliográficas



## Random Forest

- Árbol de decisión: *Definición.*



## Random Forest

- ▶ Árbol de decisión: *Definición.*
- ▶ Conjunto de estos árboles.

## Random Forest

- ▶ Árbol de decisión: *Definición.*
- ▶ Conjunto de estos árboles.
- ▶ Bootstrap.

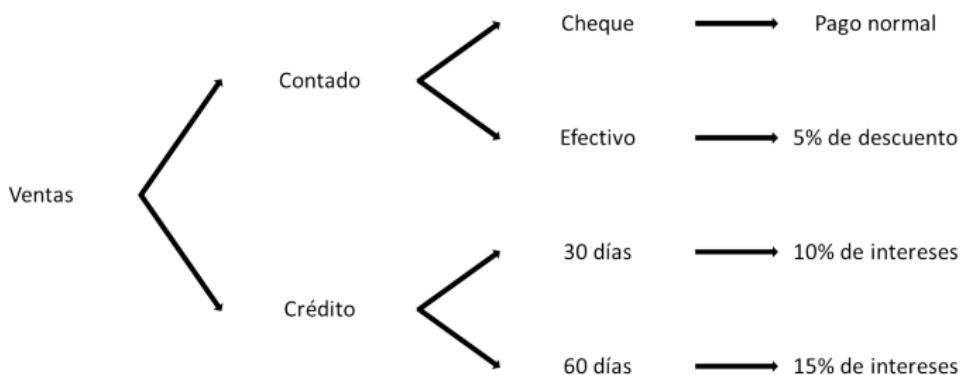
## Random Forest

- ▶ Árbol de decisión: *Definición.*
- ▶ Conjunto de estos árboles.
- ▶ Bootstrap.
- ▶ Robustez.

## Random Forest

- ▶ Árbol de decisión: *Definición.*
- ▶ Conjunto de estos árboles.
- ▶ Bootstrap.
- ▶ Robustez.
- ▶ Poco control sobre el modelo.

## Random Forest



## *Random Forest: Implementación*

- ▶ Package randomForest.



## *Random Forest: Implementación*

- ▶ Package `randomForest`.
- ▶ **500** árboles de decisión.

## *Random Forest: Implementación*

- ▶ Package `randomForest`.
- ▶ **500** árboles de decisión.
- ▶ El resultado es considerablemente mejor.

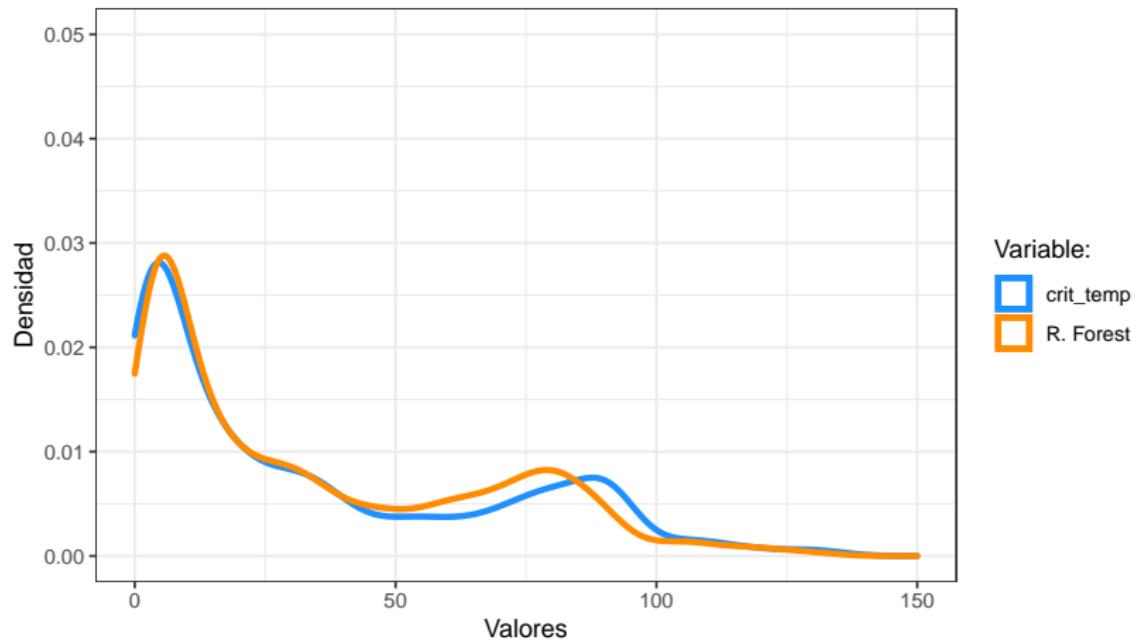
## *Random Forest: Implementación*

- ▶ Package `randomForest`.
- ▶ **500** árboles de decisión.
- ▶ El resultado es considerablemente mejor.
- ▶  $R^2 = 0.92$ .

## *Random Forest: Implementación*

- ▶ Package `randomForest`.
- ▶ **500** árboles de decisión.
- ▶ El resultado es considerablemente mejor.
- ▶  $R^2 = 0.92$ .
- ▶ Veamos la gráfica de Valores Ajustados y Valores Reales:

## Random Forest: Densidad de Ajustados y Reales



# Contenido

Avance 1

Nuevos modelos

Elección de un modelo

Ridge Regression

Extra

Conclusiones

Referencias bibliográficas



# Conclusiones

- ▶ Sobre los nuevos modelos.

## Conclusiones

- ▶ Sobre los nuevos modelos.
- ▶ Sobre el cumplimiento de los supuestos.

## Conclusiones

- ▶ Sobre los nuevos modelos.
- ▶ Sobre el cumplimiento de los supuestos.
- ▶ *Ridge y Lasso Regression.*



## Conclusiones

- ▶ Sobre los nuevos modelos.
- ▶ Sobre el cumplimiento de los supuestos.
- ▶ *Ridge y Lasso Regression.*
- ▶ *Random Forest.*

## Conclusiones

- ▶ Sobre los nuevos modelos.
- ▶ Sobre el cumplimiento de los supuestos.
- ▶ *Ridge y Lasso Regression.*
- ▶ *Random Forest.*
- ▶ Contraste con Avance 1.

## Conclusiones

- ▶ Sobre los nuevos modelos.
- ▶ Sobre el cumplimiento de los supuestos.
- ▶ *Ridge y Lasso Regression.*
- ▶ *Random Forest.*
- ▶ Contraste con Avance 1.
- ▶ **¿Es suficiente una regresión múltiple?**

# Contenido

Avance 1

Nuevos modelos

Elección de un modelo

Ridge Regression

Extra

Conclusiones

Referencias bibliográficas



## Referencias bibliográficas



[https://rpubs.com/Joaquin\\_AR/255596](https://rpubs.com/Joaquin_AR/255596)

*A.D., Random Forest ....*

2017



[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

