Examen

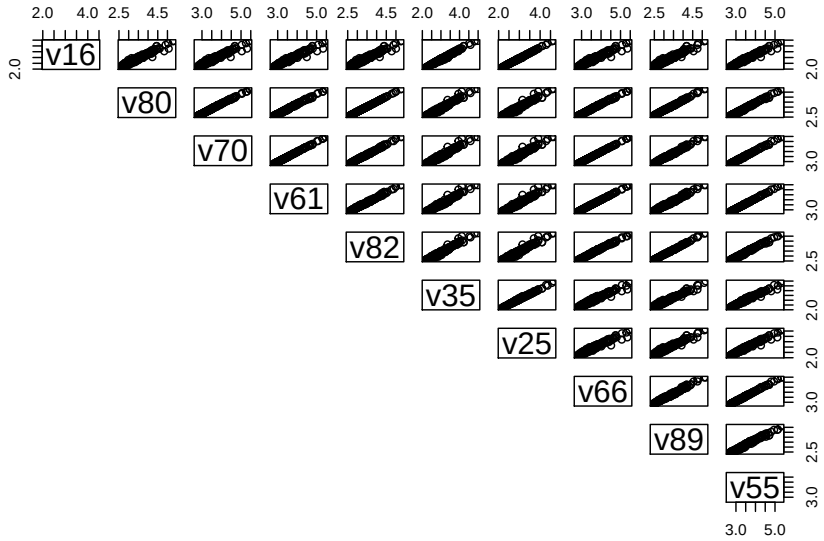
Imanol Núñez Morales

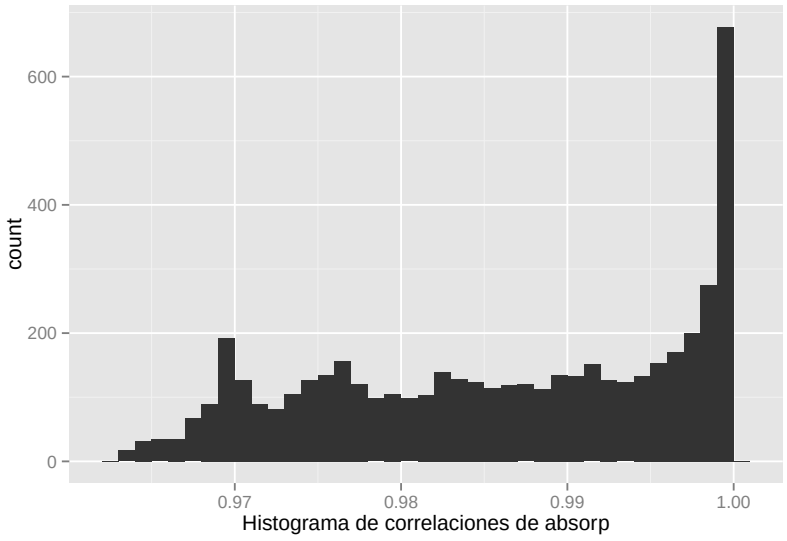
29/9/2015

Ejercicio 11

La tasa base de error para el modelo lineal es 3.12967.

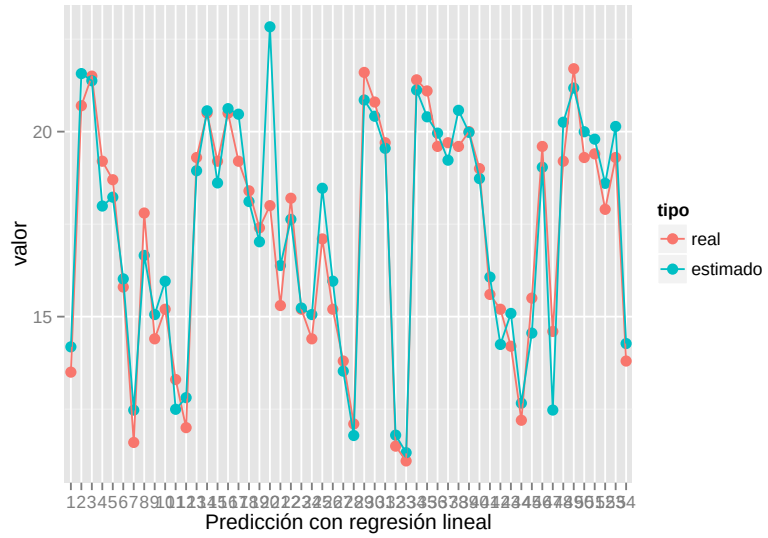
Para las gráficas para las variables de absorción primero se tomaron diez variables al azar y luego se hizo un histograma de las correlaciones entre las variables.

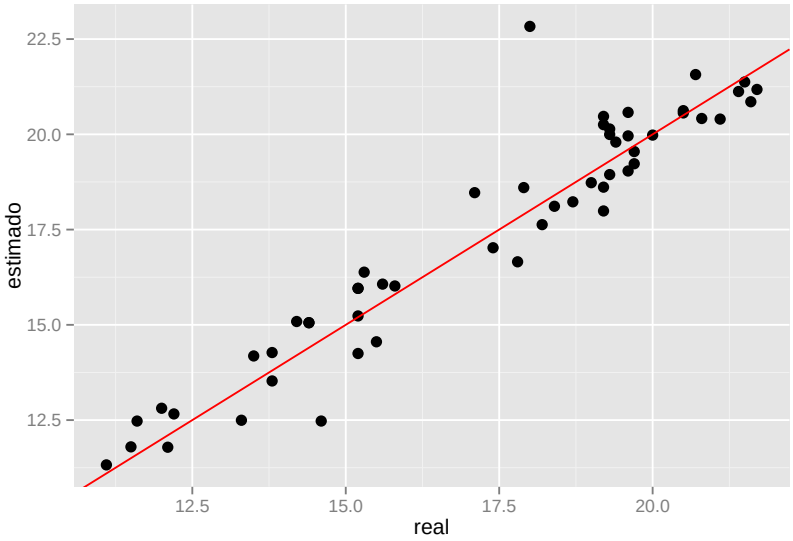




Aquí notamos que mi creencia de que las variables de absorp están correlaciondas es cierta y esto se debe a que las ondas de los espectros son continuas.

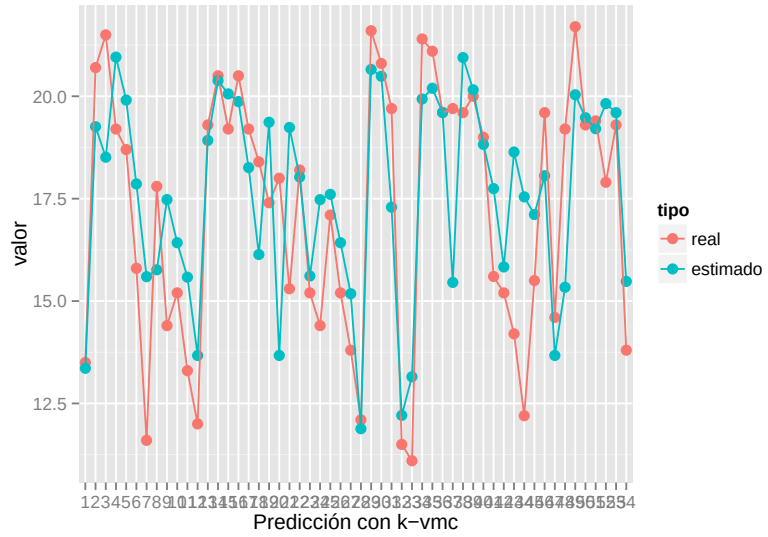
Ajustando el modelo lineal por mínimos cuadrados, obtenemos que a raíz cuadrada del error cuadrático medio es 0.97848 y, graficando las predicciones de éste contra la muestra de prueba, obtenemos lo siguiente.

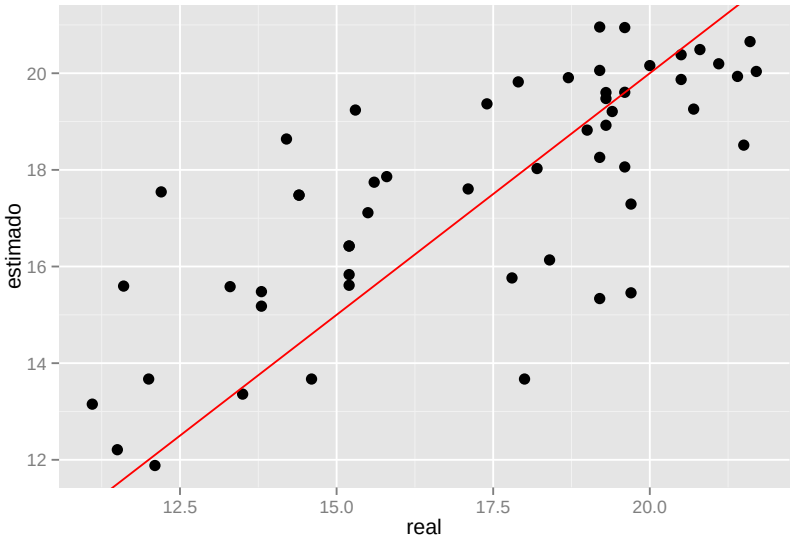




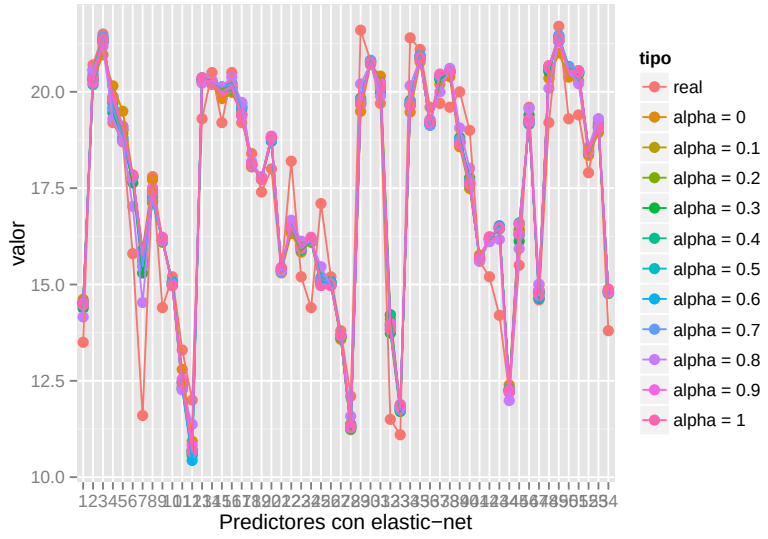
Ahora ajustamos un modelo de k-vecinos más cercanos con validación cruzada y resulta que la mejor k es 5.

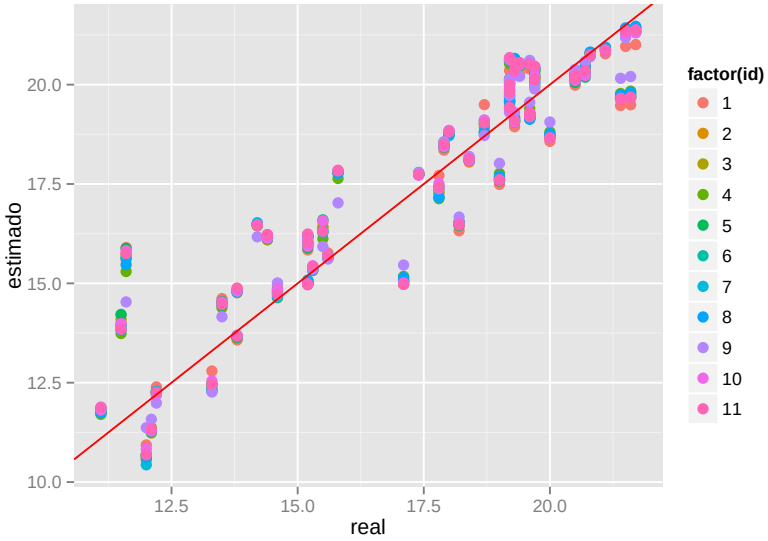
La raíz del error cuarático medio que resulta del ajuste por k-vecinos más cercanos es 2.08831 y si comparamos los valores de la muestra de prueba contra los valores de la predicción obtenemos:



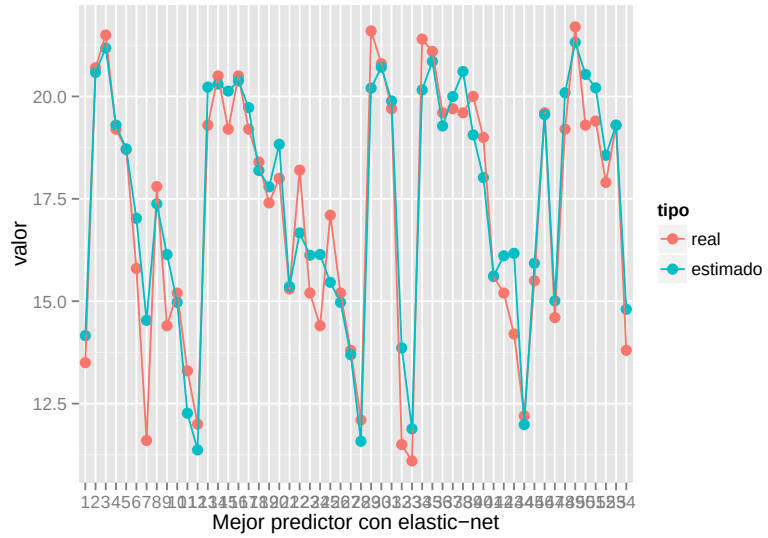


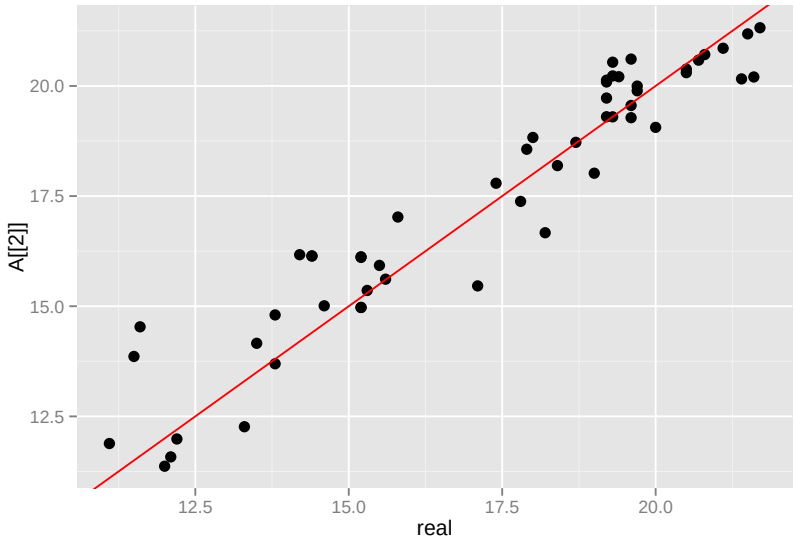
Ahora, para $\alpha \in \{0, 0.1, ..., 1\}$ se ajustan modelos con penalización elastic-net y se hace validación cruzada para escoger la mejor λ y se grafican contra los valores reales para cada α.





Resulta que en términos del error cuadrático medio, que vale 0.96306, la mejor α es $\alpha = 0.8$ y para este valor se predicen los valores reales y se grafica.





Viendo el vector de las raíces de los errores cuadráticos medios de las regresiones con penalización elastic-net vemos que los valores de α que mejoraron la estimación respecto a regresión por mínimos cuadrados son:

```
## [1] 0.8
```

La mejora que hubo con elastic-net sobre mínimos cuadrados es que al no permitirse valores de β grandes no hay predicciones que se desvíen mucho del valor real. En cambio, k-vecinos más cercanos tuvo un desempeño más pobre para estimar los valores reales que las otras dos soluciones propuestas.

Pregunta 12

Separando en muestra de entrenamiento y muestra de prueba, tenemos que la tasa base de error es 0.504. Se utiliza lasso con validación cruzada para escoger la λ, cuyo valor es 0.02114 al usar la mayor λ que está a una desviación estándar de la λ_{min}, que más nos ayude para hacer la predicción. Evaluando el modelo que obtenemos tenemos la matriz de proporciones de confusión es:

```
##
##        0    1
##   0 0.73 0.29
##   1 0.27 0.71
```

De la tabla anterior tenemos que la sensibilidad es 0.73413 y la especificidad es 0.70565. Además podemos encontrar las quince palabras que indican con mayor fuerza que una reseña sea positiva y también para que sea negativa.

Las palabras que indican con mayor fuerza que una reseña es positiva son:

```
##       [,1]
##  [1,] "job"
##  [2,] "seen"
##  [3,] "great"
##  [4,] "life"
##  [5,] "making"
##  [6,] "person"
##  [7,] "political"
##  [8,] "pulp"
##  [9,] "insurance"
## [10,] "second"
## [11,] "quite"
## [12,] "very"
## [13,] "cameron"
## [14,] "most"
## [15,] "osment"
```

Las palabras que indican con mayor fuerza que una reseña es negativa son:

```
##       palabras.negativas
##  [1,] "bad"
##  [2,] "try"
##  [3,] "villain"
##  [4,] "supposed"
##  [5,] "women"
##  [6,] "production"
##  [7,] "paris"
##  [8,] "any"
##  [9,] "minutes"
## [10,] "griffin"
## [11,] "vince"
## [12,] "team"
## [13,] "could"
## [14,] "there's"
## [15,] "eve"
```

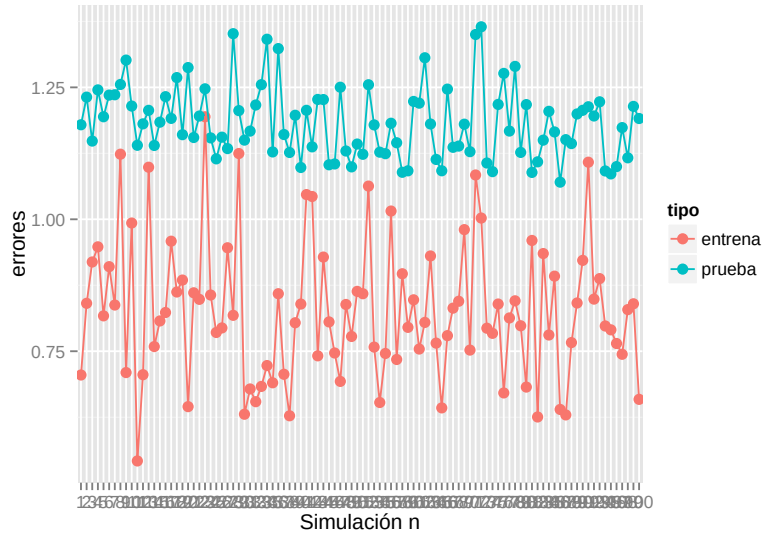
Al parecer las reseñas malas las escriben misóginos a los cuales no les gustan los villanos; lo cual es extraño porque si una película es violenta entonces parece que tendrá reseñas positivas, así como si es de política.

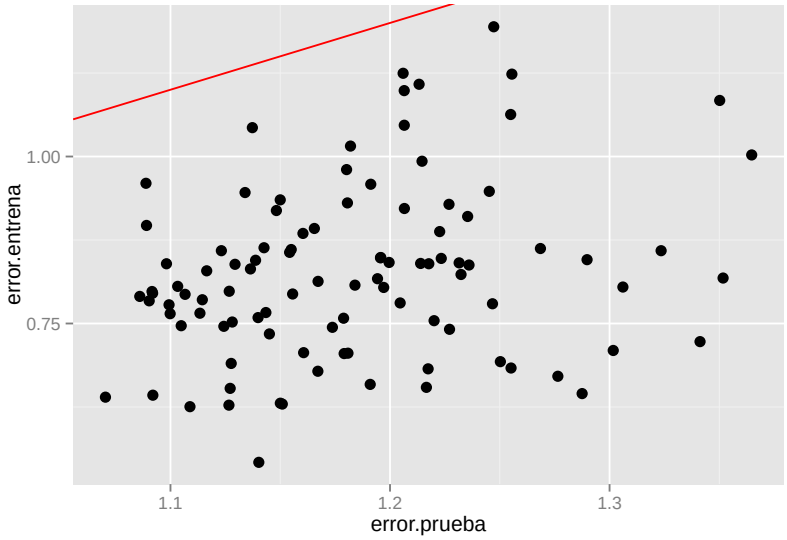
Pregunta 13

Aquí se simula 100 muestras de tamaño $n = 30$ y se tiene una muestra de prueba grande de lo siguiente:

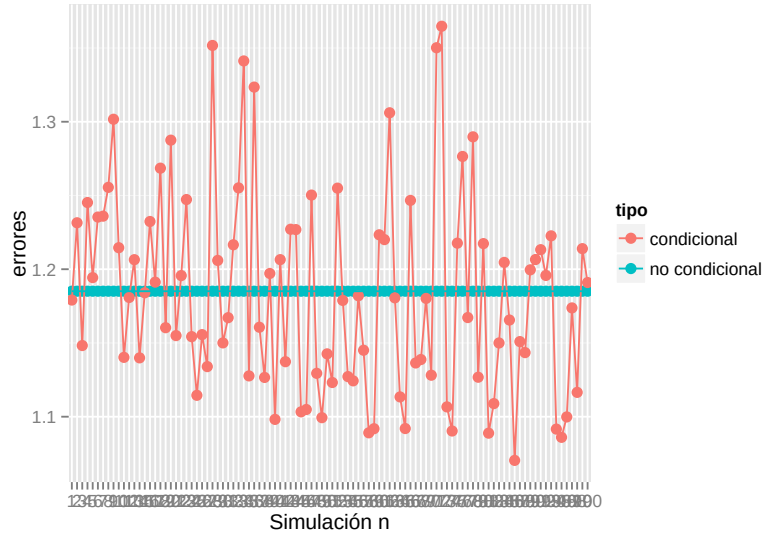
$$x \sim U(0,1) \qquad y = \left| x - \frac{1}{2} \right| + \varepsilon \qquad \text{con} \qquad \varepsilon \sim N(0,1).$$

Usando la raíz del error cuádratico medio obtenemos dos gráficas en las cuales se compara el error de entrenamiento con el error de predicción y en ambas se ve que en general el error de entrenamiento es menor al error de predicción y no parece que exista correlación entre ambos.





El error de predicción no condicional se calcula como la raíz cuadrada de la media de los errores de predicción, bajo el error cuadrático medio. El valor de este error es 1.18509. La siguiente gráfica muestra que la variabilidad del error de predicción no condicional alrededor del error de predicción condicional es demasiada.[1]



[1] Se usó la raíz del error cuadrático medio para que los números no fueran tan grandes.