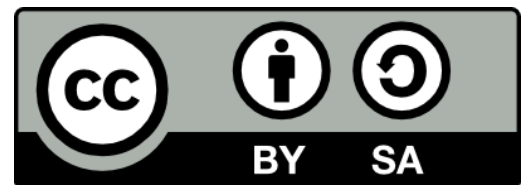
Big Data I:

Ingeniería de datos

Felipe Ortega
Dpto. de Estadística e Investigación Operativa
Universidad Rey Juan Carlos

April 8, 2015

Universidad
Rey Juan Carlos



Objetivos

- Presentar algunas estrategias (planificación) y tácticas (ejecución) que conviene tener en cuenta a la hora de realizar proyectos de análisis de datos, especialmente con big data.
- Modelado de datos, interpretación de resultados, advertencias contra malas prácticas, etc.

Teorías en busca de datos...

- A veces, los investigadores o analistas pueden comenzar con una teoría que creen fundamentada, y buscan algún conjunto de datos que la valide [1].
- Problema: La teoría debe demostrarse para *varios* (muchos) conjuntos de datos diferentes, no para uno solo.
- Ejemplo: tomamos muchas muestras de nuestro conjunto de datos, repitiendo el experimento hasta que para alguna de ellas obtengamos un resultado significativo.
  - Según la teoría de inferencia estadística (frequentista) esto está garantizado.

Y datos en busca de teorías

- *Data fishing* [1].
- R. H. Coase: “If you torture the data enough, nature will always confess”.
- Ejemplo: Si aumentamos mucho el tamaño de nuestra muestra, siempre podremos llegar a obtener resultados significativos. El problema es que solo estaremos informando acerca de “ruido significativo”.
- Importancia de incluir intervalos de confianza y tamaño del efecto (*effect size*) [4].
- P-valores dependen del tamaño del efecto *y* del tamaño muestral.

Sobre los p-valores

- *The earth is round* ($p < 0.05$).
- `http://mark.reid.name/blog/the-earth-is-round.html`
- Famoso artículo de J. Cohen que despeja cualquier duda sobre la correcta interpretación de los p-valores.
- **"el p-valor no es la probabilidad de que la hipótesis nula sea cierta dados los datos observados"**.

Validación cruzada

- A veces el analista opta por intentar utilizar todos los datos disponibles para crear un modelo. Esto no suele ser una buena idea por varios motivos:
  1. Podemos tener varios subgrupos / subpoblaciones en nuestros datos, que quedan enmascaradas por el conjunto.
     - Cada subgrupo puede verse afectado por un conjunto de factores distinto.
  2. Globalmente, los datos pueden contener mucho ruido o errores. Usando muestras de menor tamaño podemos controlar mejor estos aspectos.
- Validación cruzada: Dividimos los datos en múltiples particiones, y ejecutamos múltiples simulaciones del modelo usando en cada caso un conjunto diferente de particiones para la fase de entrenamiento (o modelado) y prueba (o validación).

Overfitting

- Consiste en dedicar grandes esfuerzos para crear un modelo que se ajusta casi perfectamente al conjunto de datos analizado...
- Pero que, por desgracia es inservible para una nueva muestra de datos generada por el mismo proceso.
- Ejemplo: Podemos aumentar arbitrariamente el grado de un polinomio en un modelo lineal para crear curvas que pasen exactamente por todos los puntos. Sin embargo, el modelo no tendría validez *descriptiva* ni *predictiva*.

Paradoja de Simpson

- Ilustrada mediante datos de matriculación en UC Berkeley.
- Ocurre cuando la tendencia que aparece en diferentes grupos individuales desaparece al combinarlos.
- Ej: parecía que las mujeres tenían menor tasa de admisiones que los hombres, pero se debía a que las mujeres solicitaban mayoritariamente plazas en departamentos con elevadas tasas de rechazo de admisión.
- `http://en.wikipedia.org/wiki/Simpson%27s_paradox`.

Modelos generadores de datos

- Principales ideas contenidas en [2].
- Los procesos (naturales o de otra índole) producen datos que podemos capturar para intentar estudiarlos y comprenderlos.
- Para ello creamos modelos que también producen datos (artificiales). Buscamos entonces que los datos producidor por el modelo propuesto se parezcan lo más posible a los datos generados por el proceso real.
  1. El modelo produce datos.
  2. El modelo tiene parámetros desconocidos, que debemos estimar.
  3. Los datos del proceso real se pueden usar para reducir la incertidumbre sobre los parámetros desconocidos.

Estimación mediante función de verosimilitud

- Likelihood function.
- Nos da una idea de la probabilidad de obtener los datos observados en función de los valores que adquieran los parámetros desconocidos de nuestro modelo.
- MLE: Escogemos los valores de los parámetros que maximizan la probabilidad de obtener los datos generados por el proceso real.

Correlaciones y causalidad

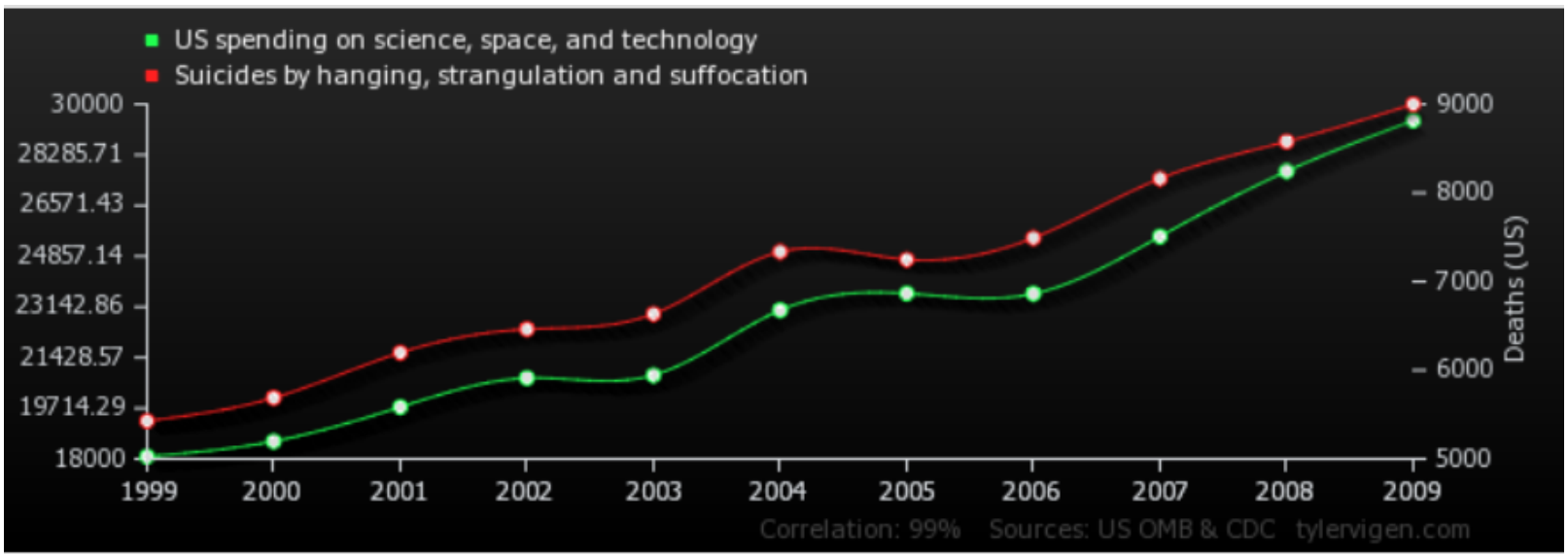
- “Correlation does not even imply correlation”. A. Gelman.
- El hecho de que encontremos correlaciones en los datos que estamos analizando no implica que esas correlaciones existan entre las dos poblaciones de interés que comparamos.
- Sin embargo, es cierto que la correlación es una de las pocas (si no la única) formas en las que podemos detectar causalidad.
- Importancia de los métodos experimentales para poder determinar con certeza relaciones causa-efecto.

Idem con series temporales

- `http://svds.com/post/avoiding-common-mistake-time-series`.
- Basta con añadir una tendencia parecida a dos series temporales totalmente aleatorias para que muestren correlación entre ambas.

Correlaciones espúreas

- http://www.tylervigen.com/



Sesgo de grandes datos

- Existe una tendencia generalizada a creer que los grandes conjuntos de datos pueden aportar resultados mucho más validos que conjuntos de datos más pequeños.
- Ejemplo: dificultades para identificar marcadores biológicos en estudios de medicina sobre enfermedades.

Silos de datos

- Ejemplo mencionado en [3].
- En muchas organizaciones (especialmente las de gran tamaño) existen numerosos almacenes de datos que no están interconectados entre sí.
- A veces no se intercambian datos por puro desconocimiento, pero muchas otras ocasiones no se intercambian por otros motivos (privacidad, recelo, etc.).
- En ocasiones el analista puede esperar obtener más información con diferentes conjuntos de datos que describan el mismo proceso, pero pueden surgir muchos problemas:
  - Inconsistencias.
  - Diferencias en nomenclatura, identificadores etc.
- Ejemplo: Unificación de bases de datos de compañías de telefonía móvil en España.

Referencias

1. Jules J. Berman. *Principles of big data*. Morgan Kaufmann. 2013.
2. Westfall, P., Kenning, K.S.S. *Understanding Advanced Statistical Methods*.
3. Manoochehri, M. Data Just Right. Addison-Wesley Professional. 2014.
4. Coe, R. (2002). It's the effect size, stupid: What effect size is and why it is important.