Cuadrados Mínimos Lineales

Carlos Iguaran

Métodos numéricos

Referencias

Que pueden citar en el TP

- An Introduction to Statistical Learning - Gareth. Capítulo 3.
  - Se puede descargar legalmente aquí
- https://online.stat.psu.edu/stat462/
- Demos

Agenda

Seguramente no cubramos todos los temas, pero la presentación quedará como referencia.

- Cuadrados mínimos lineales
- Regresión estadística
- Interpretando el modelo
- Estimadores de ajuste
    - R^2, R^2 ajustado, Residual standard error
- Gráficos de diagnóstico
    - Residuals, standarized residuals, studentized residuals, leverage
- Selección de modelos
- Variables categóricas
- Interacciones
- Colinearidad

Cuadrados mínimos lineales

- Teniendo un dataset
    - $X^1, \ldots, X^n \in R_m$
    - $Y^1, \ldots Y^n \in R$
- Sabemos que podemos ajustar una función
- $\hat{y}^j = \beta_0 + \sum \beta_i T_i$, donde
- T_i es una transformación sobre X. Esto es clave. La función es lineal en los parámetros β. No tiene por qué serlo en X.
- El criterio de cuadrados mínimos busca encontrar los Beta tales que $\sum (y^j - \hat{y}^j)^2$ sea mínimo

TP: Regresión lineal estadística

- Es decir, qué sucede cuando existe una Población, tenemos una muestra y queremos aplicar regresión sobre esa muestra.

- Esta es una técnica central del análisis de datos. Buscamos entender la contribución de variables a un resultado final. En general llamamos variables independientes a las X y variable dependiente a la Y.

- Regresores = variables independientes

- Ejemplo: Y es la capacidad pulmonar.

- X = (edad, cantidad de cigarrillos diarios, cantidad de pulmones, cantidad de veces que corre por semana, altura, peso).

Ejemplo fumadores

- Nos gustaría entender:
- Tienen relación los regresores con la variable independiente?
- Todos son útiles, o algunos están de más?
- Los features alcanzan para explicar los datos? Qué tan bien los explican?
- Si tuviera que predecir la capacidad pulmonar de alguien, sabiendo sus datos, que tan certera será esa predicción?

- Hasta ahora, nos preocupaba: tenemos un dataset, encontremos la función que mejor ajusta. En otras palabras, buscamos los mejores parámetros. Podemos considerar este como un problema de optimización resoluble via métodos numéricos.

- Ahora nos va a interesar extraer conclusiones a partir de esos parámetros y que tengan una semántica. Como las conclusiones las extraemos con información parcial |dataset| < |fumadores| debemos incorporar cuestiones estadísticas.

- Repensemos el modelo como:

- $y_i = \beta_0 + \sum \beta_j T_j + \epsilon_i$

- Donde el ruido

- $\epsilon_i \sim N(0,\sigma^2)$

- $y_i = \beta_0 + \sum \beta_j T_j + \epsilon_i$

- $\epsilon_i \sim N(0,\sigma^2)$

- Eso quiere decir que y es una variable aleatoria con distribución normal. (por propiedades de una v.a con distribución normal).

- Recordando de proba y estadística, una forma de estimar los parámetros de una distribución a partir de datos, es via el método de maxima verosimilitud.

- **Teorema**: la solución de cuadrados mínimos es equivalente al estimador de máxima verosimilitud.

- Demostración: Sección 13.2 All of Statistics - Wasserman)

La linea de regresión es en el fondo un estimador

- Tenemos una población desconocida (fumadores)
- Tenemos una muestra.
- De la misma forma que para μ usamos el estimador $\hat{\mu} = \frac{\sum y_i}{n}$
- Podemos pensar que existen β y que queremos estimarlos con $\hat{\beta}$

Demo

Interpretando el modelo

- β_i como el efecto promedio sobre Y del incremento en una unidad de T_i
- Si Ti es “peso en kilos” Bi representa el aumento (o disminución, dependiendo del signo) promedio en capacidad pulmonar al aumentar un kilo de peso.
- Cada β tiene su propia escala, lo cual no permite comparar los coeficientes entre sí. Para poder comparar, es práctica común normalizar los datos, restando la media, y dividiendo por el desvío standard.
- $X_i^j \leftarrow \dfrac{X_i^j - \mu_i}{\sigma_i}$

Demo

```
VIF for BMI 10.654528641585493
                            OLS Regression Results
==============================================================================
Dep. Variable:        Life expectancy   R-squared:                       0.709
Model:                            OLS   Adj. R-squared:                  0.701
Method:                 Least Squares   F-statistic:                     86.40
Date:                Thu, 14 Oct 2021   Prob (F-statistic):           1.24e-45
Time:                        10:11:55   Log-Likelihood:                -551.94
No. Observations:                 183   AIC:                             1116.
Df Residuals:                     177   BIC:                             1135.
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         49.6489      2.384     20.828      0.000      44.945      54.353
x1            -0.7502      0.090     -8.357      0.000      -0.927      -0.573
x2            -0.0916      0.145     -0.630      0.530      -0.379       0.195
x3             0.0039      0.004      1.009      0.314      -0.004       0.012
x4             0.1402      0.023      6.127      0.000       0.095       0.185
x5             0.2734      0.037      7.467      0.000       0.201       0.346
==============================================================================
Omnibus:                        2.604   Durbin-Watson:                   2.127
Prob(Omnibus):                  0.272   Jarque-Bera (JB):                2.652
Skew:                           0.015   Prob(JB):                        0.265
Kurtosis:                       3.589   Cond. No.                         781.
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Aclaración sobre p-valores e intervalos predictivos.

Estimadores de ajuste

- Nuestro modelo, explica los datos?
- En que casos sobre o sub estima el target?
- Hay puntos que modifican fuertemente el ajuste final?

- Varias técnicas, algunas numéricas, otras gráficas.
- Todas tienen ventajas/desventajas.
- En general, se usan en paralelo

R^2

- $RSS = \sum (y - \hat{y})^2$
- Los datos tienen variación que le es propia.
- Total sum of squares (TSS) = $\sum (y^i - \bar{y})^2$
- TSS sólo depende de los datos. Es una métrica de la variabilidad intrínseca de los datos, sin aplicar regresión.
- RSS es una métrica de la variabilidad que el modelo no puede explicar. TSS-RSS es la variabilidad que sí puede explicar.
- $R^2 = \dfrac{\text{variabilidad explicada}}{\text{variabilidad total}} = \dfrac{TSS - RSS}{TSS}$
- Idea: “Cuanta mas variabilidad explica el modelo, mejor”.
- Realidad: R2 se puede inflar/desinflar artificialmente.
- Variante R2 ajustado: tiene en cuenta la cantidad de variables p, y penaliza por agregar variables.
- $R^2_{ajustado} = 1 - \dfrac{(1 - R^2)(N - 1)}{N - p - 1}$

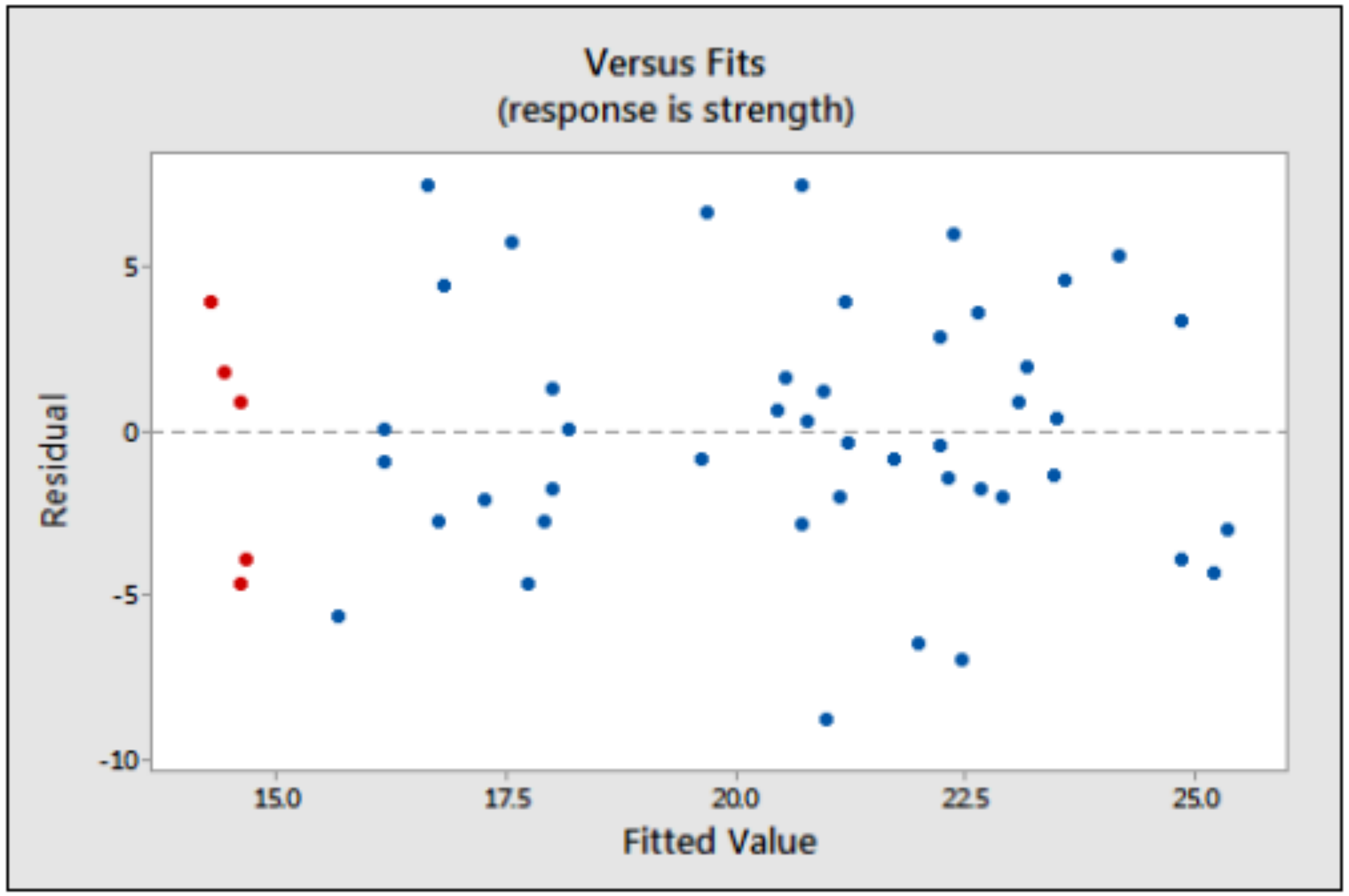
Demo

Gráficos de diagnóstico

Gráficos de diagnóstico

- residuals vs fitted values
- Coef change when excluding/including data. (Dfbeta)
- Influence, leverage, que tanto cambian los residuos cuando se incluye/excluye datos. (hatvalues)
- QQ plot https://online.stat.psu.edu/stat462/node/122/
- Studentized residuals
- Standarized residuals
- Cooks distance

Residuals vs fitted values



Es una forma de ver si la relación es lineal o no.

Deberíamos no ver ningún patrón, ser IID, tener media 0.

Permite ver si tenemos algún ‘outlier’

Referencia: https://online.stat.psu.edu/stat462/

Outliers

- Coloquialmente: un valor extremo, distante de la mayoría de las otras observaciones.

- En regresión: un y^i lejano de su valor estimado. Se divide por el desvio standard, para tener en cuenta la variación de los valores de la muestra. Se los llama ***standarized residuals***

- Dicho de otra forma: Número de desvíos standards lejos de la linea de regresión.

- En un contexto “big data” puede que no sea tan importante, pero cuando la muestra no es enorme (como en el TP) podría cobrar sentido.

- Podríamos detectar errores en el input de datos o casos interesantes de anomalías reales.

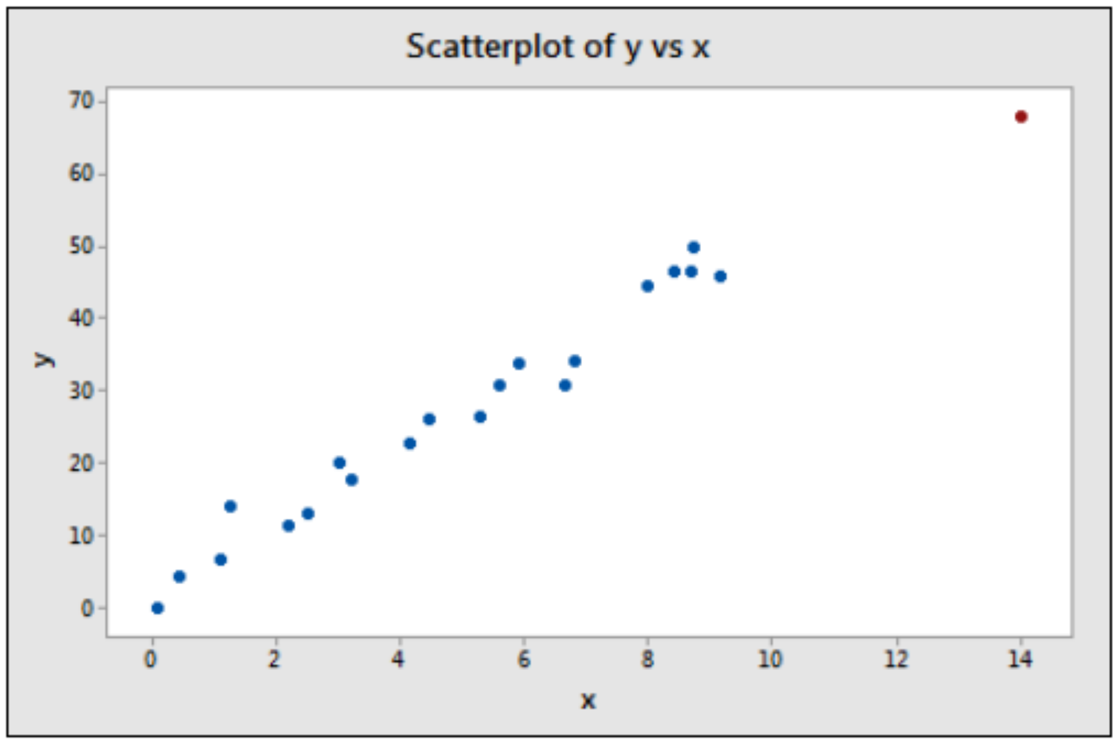
Demo

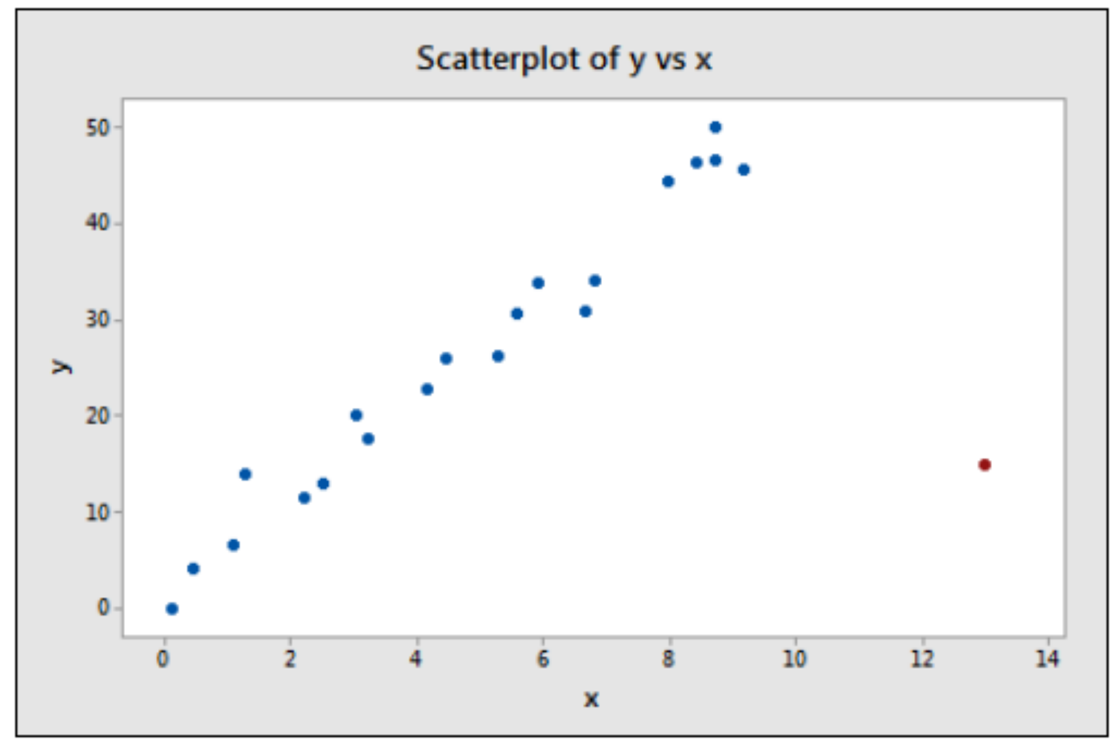
DEMO sección no lineal + outliers

https://colab.research.google.com/drive/1jOkhjpfFkcfQYuyJuKFX-cfM1ykQpHdI#scrollTo=IkTAhPDqjttw

Leverage

- Leverage: que tan anormales son los features X de una observación.
- Outliers -> Y vs Leverage -> X
- Son candidatos a modificar la curva de manera antinatural, pero no necesariamente.



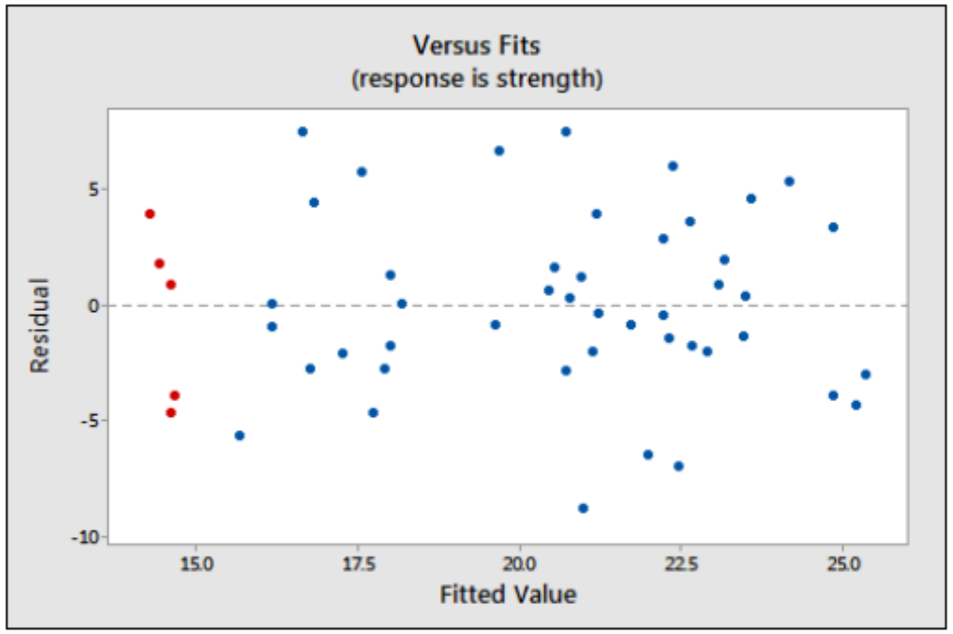


- Si un par de puntos afectan mucho la linea final, podrían empeorar el fit respecto al resto. En multiple regression, es difícil de graficar por la dimensionalidad.
- Para una explicación con ejemplos ver esta referencia.

Derivación leverages

- $y_i = \beta_0 + \sum \beta_j T_j + \epsilon_i$
- En notación matricial usamos y, β y T
- Ecuación normal: $T^t T \beta = T^t y$ -> $\beta = (T^t T)^{-1} T^T y$
- Por otro lado, $\hat{y} = T\beta$, con lo cual
- $\hat{y} = T(T^t T)^{-1} T^t y$
- Es decir, $H = T(T^t T)^{-1} T^t$ vincula $\hat{y}$ con y, por eso se la llama \`H hat\`
- $\hat{y}$ se puede escribir como una combinación lineal de las filas de H, $H_{i,i}$ cuantifica el peso que tiene y_i sobre el valor de $\hat{y}_i$
- H tiene propiedades útiles
    - $H_{i,i}$ Mide que tanto se aleja el x_i de la media
    - 0<=1 $H_{i,i}$ <= 1
    - $\sum H_{i,i} = k + 1$ (coeficientes + intercept)
- $H_{i,i}$ se los llama leverages, midiendo el **potencial** de que un punto afecte el resultado final del fit.
- Solo depende de los predictores, no del target
- Para ver si un punto tiene influencia, necesitamos ver también el target

Influencia de outliers



- Si un outlier en Y 'tracciona' el fit, va a tener un residuo pequeño, pero va a penalizar a todo el resto.
- No se ve en versus fit, para eso se usan **studentized residuals**
- $\hat{y}_{(i)}$: valor predicho para i, filtrando el modelo sin i.
- $d_i = y_i - \hat{y}_{(i)}$
- Buscamos detectar si sacar el i-ésimo valor genera una diferencia muy grande
- Studentized residual -> normalizar los d. Se pueden demostrar varias equivalencias (ver referencia)

$$t_i = \frac{d_i}{s(d_i)} = \frac{e_i}{\sqrt{MSE_{(i)}(1-h_{ii})}} \qquad t_i = r_i\left(\frac{n-k-2}{n-k-1-r_i^2}\right)^{1/2},$$

Distancia de Cook

Cook's distance D_i of observation i (for $i = 1, \dots, n$) is defined as the sum of all the changes in the regression model when observation i is removed from it[5]

$$D_i = \frac{\sum_{j=1}^{n} \left(\widehat{y}_j - \widehat{y}_{j(i)}\right)^2}{ps^2}$$

where $\widehat{y}_{j(i)}$ is the fitted response value obtained when excluding i, and $s^2 = \frac{\mathbf{e}^\top \mathbf{e}}{n-p}$ is the mean squared error of the regression model.[6]

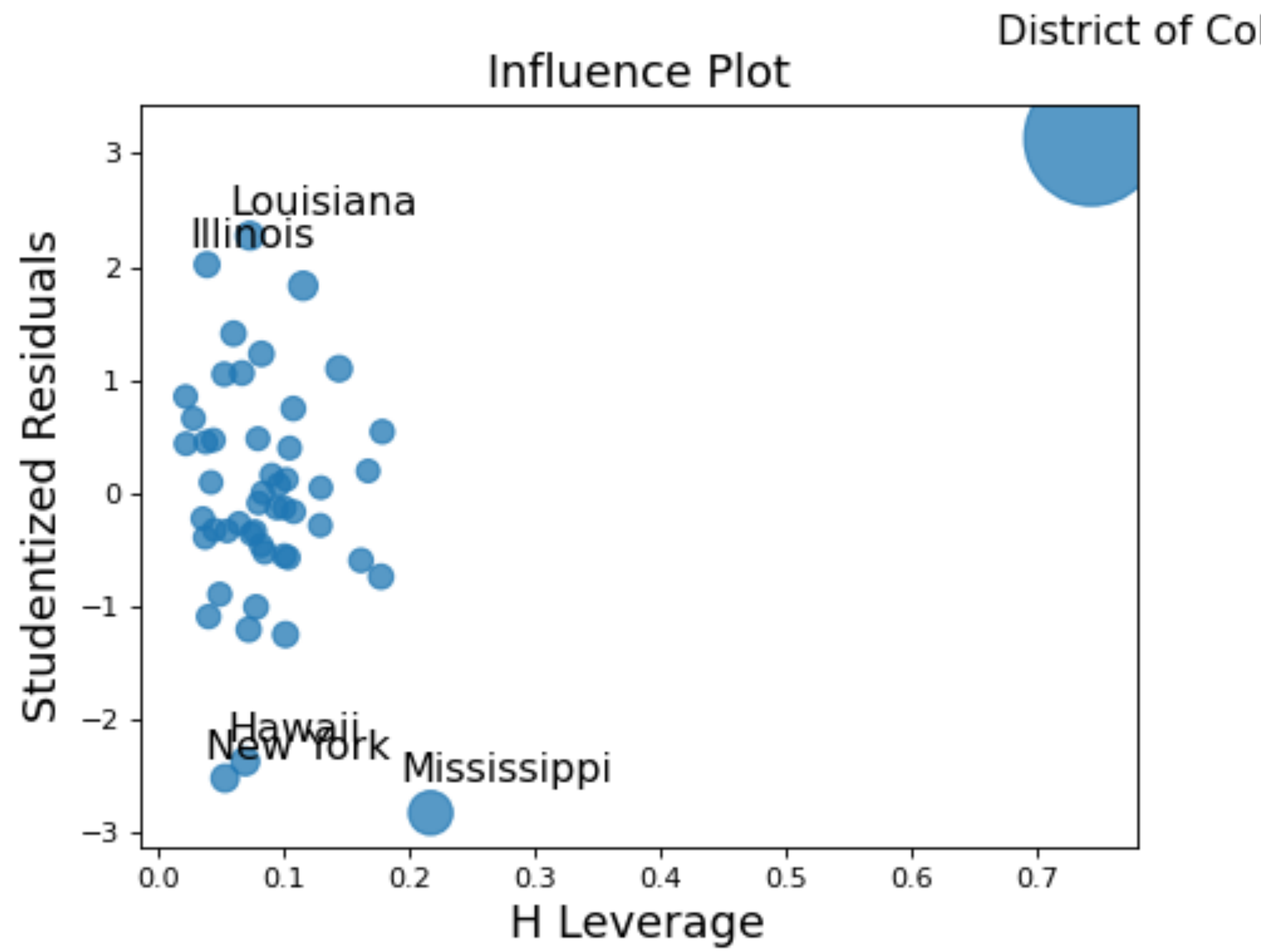
Como usarla:

- $d_i > 0.5$: sospechoso y $d_i > 1$ seguro (por motivos estadísticos)

- d_i muy separado del resto

Influence plot

- Combina:
  - Leverage
  - Residuos studentized
  - Cook



Fuente del plot

Selección de modelos

Selección de modelos

- Podría ser que no todos los features son útiles para construir un modelo. Dados P features tenemos 2^P posibles regresiones para ajustar.
- NO queremos incluir todos los features para mejorar el ajuste. Hay un tradeoff entre agregar features vs tener menos pero que expliquen bien el problema. Navaja de Ockham.
  - Agregar o quitar va a generar nuevas instancias de todos los plots/ estimadores que vimos hasta ahora.
- Podrían ir usando distintos criterios basados en entender el problema.
- Algunas estrategias cuando tienen muchos features
  - Forward selection: empezamos sin ningún predictor (solo un intercept) y vamos agregando uno a uno el predictor que más contribuya al R^2. Podemos cortar si el agregado no contribuye suficiente.
  - Backward selection: empezamos con todos los predictores, eliminando aquellos que no sean estadísticamente significativos.

Interacciones

Extendiendo el modelo

Lineal + Aditivo

- El modelo lineal asume que la relación entre predicadores y respuesta es aditiva y lineal.
- Aditiva: el efecto del predictor X_j sobre Y es independiente del resto.
- Por ejemplo: tengo un experimento donde una habitación tengo una máquina que genera una cantidad de chispazos (X1) y pólvora en gramos (X2). Sea Y la cantidad de fuego generada.
- Lineal: un cambio de una unidad en X_j tiene un efecto fijo sobre Y_j, no depende del valor de X_j.

Interaction term

- Agregamos un *interaction term*:
- $Y = \beta_0 + \beta_1 x_1 + \beta 2 x_2 + \beta_3 x_1 x_2$
- Hierachical principle: si agregamos la interacción x1x2, debemos incluir x1 y x2 como predictores.
- Ojo, x1x2 agrega colinearidad respecto a x1 y x2

Variables categóricas

Categóricas

Factor variables, dummy variables, cualitativas

- Por ejemplo: género, continente, etc.
- Encodeando género ($x_d = 1$ o $x_d = 0$)
- $y_i = \beta_0 + \sum \beta_i x_i + \beta_d x_d$
- Esto generaliza a variables no binarias, llamadas multinivel.
- Podemos reinterpretar el modelo como dos lineas distintas, pero esas nuevas lineas tienen coeficientes distintos a la original.

Demo

Colinearidad

Colinearidades

- Complica la regresión, porque al modelo le cuesta separar el efecto de cada predictor colineal (no los puede distinguir).
- Como problema de optimización, agregar colinearidad hace que tengamos múltiples soluciones con mismo RSS.
- Ejemplo Y=Capacidad pulmonar, X1=Cigarrillos X2 = Pastillas de menta.

Correlación implica colinearidad

Correlation plot

- Primer punto de entrada para reducir colinearidades en el modelo.

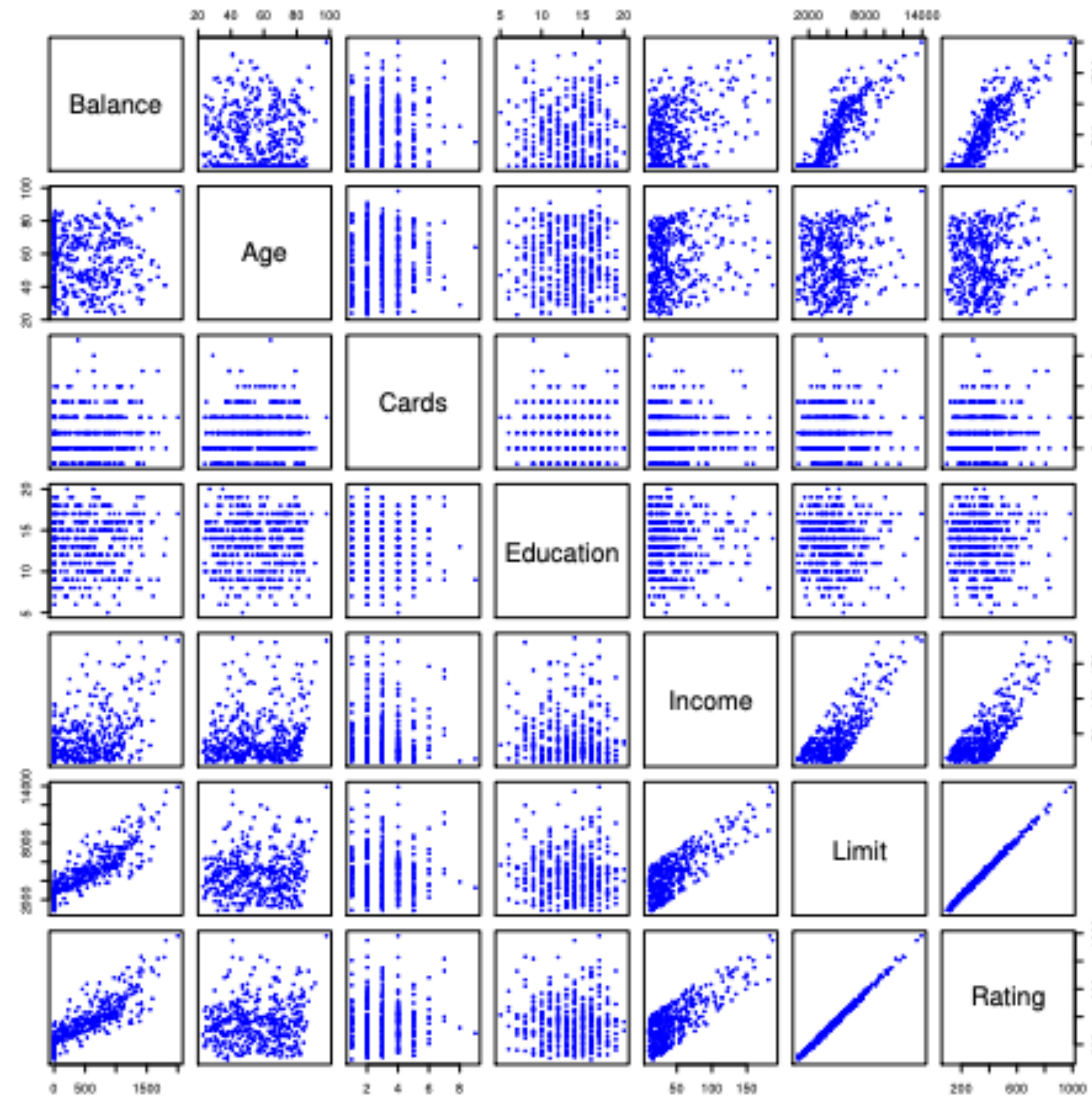


FIGURE 3.6. *The* `Credit` *data set contains information about* `balance`, `age`, `cards`, `education`, `income`, `limit`, *and* `rating` *for a number of potential customers.*

Elements of Statistical Learning, pag 83

Multicolinearidad

- Podría haber multicolinearidad entre más de un par de variables. Una forma de medirla es via el Variance Inflation Factor
- Sea X un conjunto de variables, notamos $X_j | X_{-j}$ a aplicar la regresión con X_j cómo target y todos los otras variables como predictores.
- $VIF(X_j) = \dfrac{1}{1 - R^2(X | X_{-j})}$
- En general:
    - \>5 es problemático
    - y > 10 si o si hay que modificarlo. Es un **factor**
- Soluciones ante colinearidad: eliminar variables, o combinarlas.
- Eliminar: si no baja el R^2 significativamente
- Combinar en caso contrario

Extras que suman en el TP

- Akaike's information criteria
- Bayesian information criteria
- Mallows CP
- Difference in fits (variante para Cook's distance).

En resumen

- CML es una forma de resolver el problema de regresión lineal
- Regresión lineal puede servir como herramienta de análisis de datos. El altamente interpretable.
- Dos formas de medir ajuste R2 y RSE
- Residual plots
- Selección de modelos
- Colinearidad (correlación y VIF).
- Interaction terms y variables categóricas

Bonus track: EDA

EDA

- Mostrar comparaciones. Por ejemplo, grupo control vs grupo treatment. Variables con respecto a categorías

- Estructura sistemática de los datos. Idea/visión de la estructura causal (si bien no es propiamente un análisis de causalidad).

- Integrar evidencia, mezclar plots con texto, números, point estimates para complementar.

- Describir la evidencia, etiquetas, fuentes, escalas

- Contar una historia (contenido por encima de metodología).

EDA

Gráficos

- Exploratorios:
  - Entender propiedades de los datos
  - Encontrar patrones
  - Sugerir estrategias de modelados
  - Debug del análisis
- Comunicar resultados