Resumen cap 3 HASTIE - Regresión lineal - Conceptos

josebenjumedarubio

November 2020

1 Introduction

La regresión lineal es una herramienta muy útil a la hora de predecir valores cuantitativos. Veremos el enfoque del mínimo error cuadrático medio. Estudiamos varias variables aleatorias que, esperemos, están relacionadas: inversión en propaganda de un producto, de tres tipos, que son televisiva (TV), por radio o en prensa, y ventas de ese producto (sales).

Preguntas importantes: 1 - ¿Existe una relación entre estas dos variables? Contraste de hipótesis. 2- ¿Cómo de fuerte es esta relación entre las variables? Equivalente a: Dado un valor para la inversión en TV, ¿podemos predecir las ventas con un alto nivel de exactitud? 3 - ¿Cuál de las tres publicidades contribuye más a las ventas? 4 - ¿Con cuanta precisión podemos estimar el impacto de cada publicidad en las ventas? 5 - ¿Es la relación lineal? 6 - ¿Tiene algún efecto la interacción entre publicidades? Es decir, ¿es mejor invertir 50.000 dólares en radio y otros 50.000 en televisión que invertir 100.000 nada más en uno de los dos?

La regresión lineal predice una variable a partir de otra asumiendo una relación lineal entre ambas, con un error que sigue una distribución normal de media 0:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

Donde:
β_0 es el corte de la recta con el eje OY
β_1 es la pendiente de la recta.

En la práctica no se encuentran variables aleatorias que cumplan una relación completamente lineal, con ϵ idénticamente 0, sino que siempre hay cierto error, cierto comportamiento de Y que es independiente de X. Entonces decimos que la variable aleatória X consigue explicar cierto porcentaje de la variable Y.

Para estimar los valores de β_0 y β_1, necesitamos una muestra de tamaño n, y querremos encontrar unas estimaciones tales que, al aplicar la función de nuestra recta de regresión a cada valor de la muestra x_i, el valor $\hat{y}_i$ que obtengamos sea lo más cercano posible a y_i.

Para medir lo cercanos que son nuestras estimaciones $\hat{y}_i$ de los valores de la muestra y_i, utilizamos el error cuadrático medio.

Sea $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ la predicción para Y basada en el i-ésimo valor de X, $e_i = y_i - \hat{y}_i$ es el i-ésimo error.

El error cuadrático se define como:

$$RSS = e_1^2 + e_2^2 + ... + e_n^2 = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + ... + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2$$

Como hemos dicho que vamos a intentar que todas nuestras predicciones estén cerca de sus respectivos valores muestrales, nuestras estimaciones de β_0 y β_1 serán los valores que minimicen esa cantidad.

La manera de obtenerlos ya la he explicado en el otro documento.

$$\hat{\beta}_1 = \frac{\overline{xy} - \overline{x}.\overline{y}}{\overline{x^2} - \overline{x}^2} = \frac{COV(x,y)}{V(x)}$$

$$\beta_0 = \overline{y} - \overline{x}\frac{COV(x,y)}{V(x)}$$

Otra manera de escribir que asumimos que la relación real entre X e Y es

$$Y \approx \beta_0 + \beta_1 X$$

es decir

$$Y = \beta_0 + \beta_1 X + \epsilon$$

donde ϵ es el error, que es una normal de media cero. Normalmente ϵ es independiente de X. Si fuese linealmente dependiente, simplemente modificamos la función $Y = f(X)$ que también es lineal, para que incluya el error, y si tuviese otra dependencia, supongo que diríamos que X e Y no son linealmente dependientes. Podemos decir que el modelo dado por $Y = f(X) + \epsilon$ define la linea de regresión poblacional, con $f(X) = \beta_0 + \beta_1 X$.

Nuestra estimación de la recta de regresión es una estimación insesgada, es decir, de media, acierta. El significado de esto se ve bien en el siguiente ejemplo: Tenemos un modelo cuya recta de regresión es

$$Y = \beta_0 + \beta_1 X + \epsilon$$

. Generamos una muestra de tamaño n, con la que estimamos una recta de regresión

$$Y = \hat{\beta}_{0_0} + \hat{\beta}_{1_0} X$$

. Cada parámetro puede ser algo menor o algo mayor que el valor poblacional, es decir, podemos haber sobrestimado o subestimado el valor real. Sin embargo, si hacemos k muestras más, siendo k un número muy grande, y obtenemos k rectas de regresión, con sus respectivas k estimaciones de los parámetros $\hat{\beta}_{0_1}, \hat{\beta}_{0_1}, ..., \hat{\beta}_{0_k}$ y $\hat{\beta}_{1_1}, \hat{\beta}_{1_2}, ..., \hat{\beta}_{1_k}$, la media de cada uno de ellos coincide con el valor poblacional:

$$\beta_0 = \frac{1}{n}(\hat{\beta}_{0_1} + \hat{\beta}_{0_1} + ... + \hat{\beta}_{0_k}) \qquad \beta_1 = \frac{1}{n}(\hat{\beta}_{1_1} + \hat{\beta}_{1_1} + ... + \hat{\beta}_{1_k})$$

¿Cómo hemos conseguido que esto ocurra? En el desarrollo de las derivadas de la función del error cuadrático medio, hemos hecho estas dos sustituciones

$$\frac{1}{n}\sum_{i=0}^{n} x_i = \overline{x} \qquad \frac{1}{n}\sum_{i=0}^{n} y_i = \overline{y}$$

Es decir, hemos estimado la media poblacional con la media muestral. Veamos cual es la media, si tomásemos k muestras, de este estimador:

$$E[\,\overline{x}\,] = E[\,\frac{1}{n}\sum_{i=0}^{n} x_i\,] = \frac{1}{n}\sum_{i=0}^{n} E[\,x_i\,] = \frac{1}{n}\sum_{i=0}^{n} \mu = \mu$$

Conclusión: hemos construido un estimador a partir de otros, y como estos otros eran insesgados, el nuevo también lo es.

Ahora sabemos que nuestra estimación, de media, acierta, pero ¿cómo de cerca del valor real estará una sola estimación? Para explicarlo, utilizaremos el ejemplo de la media poblacional estimada mediante la media muestral:

$$\sigma^2(\hat{\mu}) = \sigma(\hat{\mu})^2 = \frac{\sigma^2(X)}{n}$$

La cantidad $\sigma(\hat{\mu})$ es la distancia a la que $\hat{\mu}$ va a estar de su media, que coincide con μ, así que en esta fórmula tenemos la cantidad que buscábamos.

De la misma manera, podemos estudiar cómo de cerca están las estimaciones $\hat{\beta}_0$ y $\hat{\beta}_1$ de β_0 y β_1 respectivamente, con las siguientes fórmulas:

$$\sigma(\hat{\beta}_1) = \frac{\sigma^2(\epsilon)}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \qquad \sigma(\hat{\beta}_0) = \sigma^2(\epsilon)\left(\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right)$$

Para que sean estrictamente válidas es necesario que el error ϵ y la covarianza sean incorrelados, aunque hay muchos casos en que esto no ocurre y aún así las fórmulas dan muy buenas aproximaciones. De ellas podemos destacar que al estar la varianza de x en el denominador, esto nos beneficia. Cuanto más extendidas estén las x, mejor estimaremos los parámetros. Podemos imaginar que queremos trazar una recta uniendo dos puntos: cuando estos están muy cerca, un pequeño error lleva a una recta muy distinta de la que queríamos en principio, y en cambio, si están lejos, para que la recta varíe tanto necesitaríamos haber cometido un error mucho más grande.

Normalmente, ϵ es desconocido,y se estima mediante el error residual estándar (RSE en inglés):

$$RSE = \sqrt{\frac{RSS}{n-2}}$$

donde RSS es, como dijimos antes, el error cuadrático (Residual Sum of Squares):

$$RSS = e_1^2 + e_2^2 + ... + e_n^2$$

Cuando estimemos el error a partir de los datos, ya no estaremos haciendo un estudio exacto de la variable aleatoria $\hat{\beta_1}$, sino una estimación, así que deberemos escribir $\widehat{SE}(\hat{\beta_1})$.

Estas fórmulas son muy útiles porque nos dan un intervalo de confianza del 95% muy sencillo:
Para β_0 : $\hat{\beta}_0 \pm 2SE(\hat{\beta}_0)$ Para β_1 : $\hat{\beta}_1 \pm 2SE(\hat{\beta}_1)$

Los errores cuadráticos también son útiles a la hora de hacer contraste de hipótesis, que consiste en formular una hipótesis o afirmación sobre los datos, y comprobar si estos aportan, una vez fijado cierto nivel de confianza, suficiente evidencia como para probar que la afirmación es falsa, es decir, rechazar H_0 y aceptar la contraria, H_1.

Un contraste de hipótesis muy frecuente es afirmar que no existe ninguna relación entre X e Y. Matemáticamente:

$$\begin{aligned} H_0 &\equiv \beta_1 = 0 \\ H_1 &\equiv \beta_1 \neq 0 \end{aligned}$$

Para comprobar la hipótesis alternativa, tenemos que asegurarnos de que β_1 está lo suficientemente lejos de 0 como para estar seguros de que no ha sido fruto del azar. Para esto, suponemos que β_1 sí que es 0, y damos un intervalo de confianza α en el que obtendremos nuestra estimación si éste es el caso. Si la estimación que obtenemos a partir de la muestra no pertenece al intervalo, entonces rechazamos.

El tamaño del intervalo de confianza está directamente relacionado con la varianza de nuestro estimador. Si la varianza es alta, es probable que teniendo, por ejemplo, $\beta_1 = 0$, obtengamos una estimación $\hat{\beta}_1 >> 0$, con lo cual, será más difícil rechazar la hipótesis, y el intervalo de confianza α será mayor. En cambio, si tenemos una varianza muy cercana a cero, muestras que nos proporcionen estimaciones también cercanas a cero pueden ser suficiente para descartar $\hat{\beta}_1 = 0$.

Utilizaremos el estadístico $t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$, para ver cuántas $SE(\hat{\beta}_1)$ dista $\hat{\beta}_1$ de 0. Si verdaderamente no hay ninguna relación entre X e Y, es decir, si $\hat{\beta}_1$ es 0, esperamos tener una distribución t-student, con n-2 grados de libertad, que

significa que lo más probable es obtener valores muy cercanos a 0, incluido el propio 0, sin importar si son algo mayores o menores (simétrica), y los que se vayan alejando más serán mucho menos probables. A partir de ±30, la t-student y la normal(0,1) son muy parecidas. Así, teniendo una distribución fija conocida, podemos calcular la probabilidad de cada valor obtenido. Digamos que el intervalo de confianza es del 95%. Si obtenemos entonces un valor de $\hat{\beta}_1$ que pertenezca a $t_{\alpha/2}$ o a $t_{1-\alpha/2} = -t_{\alpha/2}$, descartamos la hipótesis.

2 P-valor

Continuando con el ejemplo, suponemos que nuestra muestra nos da un valor de t que pertenece justo al borde del intervalo $t_{0.17}$. Esto significa que los valores que hay desde $\hat{\beta}_1$ hacia la derecha, tienen en conjunto una probabilidad 0.17.

Para α tales que $\alpha/2 < 0.17$, no podríamos rechazar, pues este valor de $\hat{\beta}_1$ aún estaría en nuestro intervalo de confianza α. Es decir, hay un rango de valores de α para los que podemos rechazar, y otro rango para los que no. Se define entonces el p-valor como el ínfimo de los valores de α que nos permite rechazar la hipótesis, es decir, el valor de α para el que la estimación de $\hat{\beta}_1$ pertenece a t_α y tal que no existe $\epsilon > 0$ para el que $\hat{\beta}_1 - \epsilon \in t_\alpha$

3 Regresión lineal múltiple - segunda parte del resumen

La regresión lineal simple es útil cuando predecimos una respuesta en base a una única variable, pero cuando hay varias hay que proceder de manera distinta.

En un caso en el que tengamos que predecir respecto de n variables, podríamos hacer n regresiones lineales independientes, pero habría que decidir qué valor de Y le corresponde a cada muestra, pues cada modelo nos daría un valor y acabaríamos teniendo n valores. Además, cada predicción se basaría en una sola variable, ignorando las otras, y esto puede reducir la calidad de las predicciones si las variables tienen correlación unas con otras.

Lo que vamos a hacer es modificar el modelo de regresión simple para que tenga en cuenta tantas variables como queramos, dándole a cada variable una pendiente, quedándonos el siguiente modelo:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \epsilon$$

Donde X_i representa la iésima variable, y β_i cuantifica la asociación entre esa variable y la variable respuesta. La interpretación de β_i es el efecto medio sobre Y de un incremento de una unidad sobre la variable X_i, manteniendo fijos los demás parámetros.

3.1 Estimar los coeficientes de regresión

Igual que ocurría con la regresión lineal simple, los coeficientes de regresión son desconocidos, y tenemos que estimarlos. Procedemos de la misma manera, minimizando el error cuadrático (Residual Sum of Squares):

$$RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 X_1 - \hat{\beta}_2 X_2 - \ldots - \hat{\beta}_n X_n)^2$$

Al contrario que con la estimación simple, las fórmulas que nos permiten minimizar esta ecuación y obtener las estimaciones de los $n + 1$ parámetros son bastante complicadas, y, dado que las podemos obtener mediante cualquier software de estadística, no las incluiremos aquí.

3.2 Regresión lineal múltiple vs múltiples regresiones lineales

Hemos mencionado que la correlación entre las variables hay que tratarla con cuidado, y ahora vamos a ver porqué.

Suponemos que tenemos dos variables, X_0 y X_0, para estimar otra variable Y. Puede darse el caso de que la correlación entre X_0 e Y sea muy alta, es decir, que Y se explique muy bien mediante X_0, y que, además, el aumento de la primera variable suela ir acompañado del aumento de la segunda, con lo cual, también tengamos una correlación alta entre X_0 y X_1. En este caso, aunque de manera indirecta, Y también se explicará bien mediante X_1.

Si hacemos dos modelos de regresión simple, pensaremos que ambas variables son igual de determinantes para el valor de Y, cuando en realidad, solo uno de los dos es necesario, y podemos prescindir del aumento del otro.

Para darnos cuenta de esto es imprescindible hacer un modelo de regresión múltiple, en el que el coeficiente de cada variable es efecto medio sobre Y de un incremento de una unidad sobre la variable X_i, **manteniendo fijos los demás parámetros**. Aquí será donde nos demos cuenta de que un aumento de la variable X_1 dejando fija X_0 no nos proporcione ningún aumento de Y.

Un ejemplo sencillo y quizá algo absurdo sería estudiar la cantidad de ataques de tiburones en las playas de cierto lugar, en relación a la temperatura y a la cantidad de helados vendidos en los quioscos de la playa.

Cualquiera puede pensar que al aumentar la temperatura, la gente irá más a la playa, y por tanto se consumirán más helados, y además habrá más ocasiones en las que puedan darse ataques de tiburones que si no se baña nadie. Es decir, todas las variables aumentan a la vez, pero no tiene sentido pensar que

vender más helados aumenta la cantidad de ataques de tiburones. Sin embargo, si hiciésemos dos modelos de regresión lineal independientes, uno estudiando los ataques frente a la temperatura, y otro estudiándolos frente a la cantidad de helados vendidos, concluiríamos que las ambas variables son igual de determinantes, pues su aumento hace aumentar también la cantidad de ataques de tiburones.

3.3 Preguntas importantes, y sus respuestas

3.3.1 ¿Es al menos una de las variables útil a la hora de predecir la respuesta?

Para responder a esto, podemos hacer un contraste de hipótesis con

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_p$$

$$H_1 : \exists i : \beta_i \neq 0$$

Para testear esta hipótesis utilizamos el estadístico F, que se define:

$$F = \frac{(TSS - RSS)/p}{RSS/(n-p-1)}$$

donde

$$TSS = \sum_{i=0}^{n} (y_i - \hat{y})^2$$

$$RSS = \sum_{i=0}^{n} (y_i - \hat{y}_i)^2$$

Si la suposición de linealidad es correcta, se cumple

$$E[RSS/(n-p-1)] = \sigma^2$$

Y si H_0 es cierta, se cumple

$$E[(TSS - RSS)/p] = \sigma^2$$

Por lo tanto, si el estadístico F tiene un valor lejos de 1, podemos rechazar H_0, y si no, no. ¿Cómo de lejos? Depende de n y p. Cuanto más grande sea n, menos lejos tendrá que estar. Respecto a p, cuando H_0 es cierta y los errores ϵ_i se distribuyen como una normal, el estadístico F sigue una distribución de Fisher-Snedecor, y podemos calcular el p-valor y utilizarlo para decidir si rechazamos H_0 o no. Cuanto más cerca de 0 esté el p-valor, mayor evidencia tenemos para rechazar H_0.

Hay veces que queremos comprobar si un subconjunto de q coeficientes son 0. Para ello, tomamos:

$$H_0 : \beta_p - q + 1 = \beta_p - q + 2 = \ldots = \beta_p = 0$$

donde las q variables que queremos estudiar están al final del conjunto (y por tanto los coeficientes también).

Ahora, para un segundo modelo que utiliza todas las variables excepto las q últimas, tenemos:

$$F = \frac{(RSS_0 - RSS)/q}{RSS/(n-p-1}$$

Este estadístico F nos dice el efecto de añadir estas q variables a un modelo en el que ya estaban todas las demás.

Sin embargo, parece que la información que nos da el estadístico F, ya la hemos obtenido en otro momento, al calcular cada coeficiente β_i como el incremento medio de la respuesta al incrementar en una unidad la variable X_i, estando fijas las demás, así que, ¿que nos aporta el estadístico F que no tengamos ya?

La respuesta la tenemos al observar un modelo con muchas variables: digamos $P = 100$. Vamos a suponer además, que ninguna de estas variables tiene relación alguna con la variable respuesta, es decir, $\beta_1 = \beta_2 = \ldots = \beta_p = 0$. En este caso, y **teniendo en cuenta que el nivel de confianza es 0.05**, hay una probabilidad del 5% de que un p-valor esté por debajo de 0.05, así que, de 100, podemos esperar encontrar 5 p-valores por debajo del 0.05, que es bastante cercano a 1, y que nos indicaría que los datos aportan una fuerte evidencia para rechazar H_0. Si es probable encontrar 5, podemos decir casi con total seguridad, que al menos habrá uno, y por lo tanto siempre fallaremos.

Esto no ocurre con el estadístico F, que tiene en cuenta la cantidad de variables, y haya la cantidad que haya, la probabilidad de obtener un p-valor inferior a 0.05 es del 5%. Tenemos un 95% de probabilidad de acertar.

Por último, y muy importante, el estadístico F funciona cuando p es "relativamente" pequeño (con 100 aún funciona bien, parece), y mucho más pequeño si lo comparamos con n. Cuando p se acerca a n o incluso lo supera, tenemos que utilizar otras técnicas, y no podemos aplicar casi nada de lo que se ha visto hasta ahora.

3.3.2 Selección de variables

El primer paso a la hora de hacer una regresión es hallar el estadístico F y calcular su p-valor. Si a partir de éste concluimos que existe al menos una variable linealmente relacionada con la variable respuesta, hay que encontrar cuál/cuáles

es/son. Normalmente ocurre que solo un subconjunto de las variables nos serán útiles.

Si tenemos p variables, para probar todas las posibles combinaciones, habría que hacer 2^p modelos, cosa que rara vez es factible, así que hay que probar otros métodos: regresión progresiva, **backward selection**, y **mixed selection**.

La regresión progresiva consiste en partir de un modelo sin ninguna variable, de la forma $Y = constante$. Entonces, se hacen p modelos para cada una de las variables, relacionándolas con la variable respuesta. Añadimos al modelo principal la que tenga menor error cuadrático. A continuación vemos que error cuadrático obtenemos al añadir al modelo principal, que entonces tendría dos variables, cada una de las restantes, y nos quedamos con la que de un error cuadrático menor. Así sucesivamente hasta cumplir una regla de parada que definamos. Es posible que acabemos incluyendo en nuestro modelo variables redundantes.

La **backward selection** (necesario $p < n$) consiste en partir de un modelo con todas las variables e ir quitándolas una a una empezando por la que tenga mayor p-valor, y en cada paso recalculándolo, ya que cambia según las variables que haya. Así sucesivamente hasta cumplir una regla de parada que definamos. El p-valor que calculamos para la variable X_i es el resultante de contrastar la hipótesis $H_0 : \beta_i = 0$.

Mixed selection consiste en una mezcla de las dos técnicas anteriores. Básicamente, procedemos como en regresión progresiva, pero cada vez que añadimos una variable, comprobamos el p-valor de las que llevamos elegidas hasta ese momento, pues ya hemos visto que dependiendo de las variables del modelo, el p-valor puede cambiar, y cuando supere algún límite establecido previamente, eliminamos esa variable. Así sucesivamente hasta cumplir una regla de parada que definamos.

3.3.3 Bondad del ajuste

Dos de las medidas numéricas de la bondad del ajuste de un modelo son el error cuadrático medio y el estadístico R^2. El error cuadrático medio se calcula de manera análoga a como se hacía en regresión simple.

El estadístico R^2, que en regresión simple era el cuadrado de la correlación entre la variable y la respuesta, ahora es el cuadrado de la correlación entre la respuesta y la respuesta estimada. Un valor cercano a 1 indica que la varianza del modelo explica gran parte de la varianza de la variable respuesta. Hay que tener cuidado con R^2, pues siempre aumenta al añadir variables al modelo, tengan estas algo que ver con la variable respuesta o no, así que lo que hay que mirar es cuánto aumenta. Podemos también apoyarnos, para cada variable que

añadamos, en el p-valor de su coeficiente. Si es muy cercano a cero hemos hecho bien en añadirla.

Además de estas dos medidas, es bueno pintar los datos y el modelo en un gráfico, pues hay información que no se recoge ni en R^2 ni en el error cuadrático medio, como un efecto **synergy** entre dos o más variables, que en conjunto aportan más que la suma de lo que aportan por separado, y otras relaciones no lineales.

3.3.4 Predecir

Una vez tenemos el modelo, hacer una predicción es tan simple como evaluar nuestra recta en el punto en que queramos, pero hay tres tipos de incertidumbre a considerar.

Los coeficientes $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p$ son estimaciones de los coeficientes poblacionales $\beta_1, \beta_2, \ldots, \beta_p$, es decir, el plano de regresión

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \ldots + \hat{\beta}_p X_p$$

es una estimación del plano de regresión poblacional

$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

. Al error que obtengamos aqui lo llamaremos error reducible.

Para gestionar esto tenemos los intervalos de confianza, que nos dan una idea de cómo de cerca estará $\hat{Y}$ de $f(X)$.

En la práctica, no encontramos un modelo poblacional como acabamos de ver, sino que hay que añadir un término de error, representando la cantidad de varianza de Y que las variables no son capaces de explicar, así que la ecuación es $f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon$, donde a ϵ lo llamaremos el error irreducible. Esto quiere decir que aunque conociésemos los verdaderos coeficientes, aún tendríamos un error de predicción derivado de desconocer el error antes mencionado. Y además, no vamos a conocer tampoco los verdaderos coeficientes. Aquí utilizamos un intervalo de predicción, que considera tanto el error reducible como el irreducible.

Por ejemplo, al hacer una predicción, obtenemos un intervalo de confianza, por ejemplo, con una confianza del 95%. Esto quiere decir que en el 95% de las predicciones, el valor de f(X) (esta es la ecuación sin el término de error irreducible) estará contenido en ese intervalo. Además, tenemos un intervalo de predicción, que será más grande, con la confianza correspondiente, digamos también 95%, que indicará que el 95% de las veces, el valor de $Y(X)$ estará en ese intervalo (aquí $Y(X)$ es el modelo con el término de error).

4 Otras consideraciones sobre el modelo de regresión

4.1 Variables cualitativas

Son las variables cuyos valores no son números.

4.2 Variables cualitativas con solo dos posibles valores

Supongamos que tenemos p variables $X_1, X_2, \ldots, X_p$, y una variable de respuesta Y, tal que la variable X_1 es cualitativa: se refiere al género de una población, con dos posibles valores, hombre y mujer. Podemos crear una nueva variable, basada en ésta, de la siguiente forma:

Si hacemos una regresión lineal simple de Y respecto a **ERROR** obtenemos una ecuación $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, en la que el parámetro β_0 se puede interpretar como el valor medio de Y para la población de mujeres, y el parámetro β_1 como cuánto más grande será el valor de Y si la población son hombres.

Si por el contrario hacemos

El parámetro $\beta_0 sería el punto medio entre la media de los hombres y la media de las mujeres, y \beta_1$ es cuánto está por encima de β_0 la población de hombres y cuánto está por debajo la de mujeres. Si la media sin tener en cuenta género es 10, la media de los hombres es 4 y la de mujeres 12, β_0 sería 8, y $beta_1$ sería -4.

Recuerdo: el nivel de significación es la probabilidad de cometer el error más grave: rechazar H_0 cuando es cierta. Si nos piden hacer un contraste de hipótesis con nivel de significación α, no podemos rechazar H_0 a no ser que la probabilidad de rechazarla siendo cierta sea menor o igual que α. Si tenemos $\alpha = 1$ podemos rechazar siempre, porque la probabilidad de equivocarnos siempre es, como la probabilidad de cualquier suceso, menor o igual que 1. Si tenemos $\alpha = 0$ no podemos rechazar nunca, porque siempre habrá una mínima probabilidad de haber rechazado mal, y estaremos incumpliendo el criterio del nivel de significación.

Sigo recordando: dada una hipótesis H_0 y una muestra, el p-valor es el máximo nivel de significación que necesitaríamos para no rechazar H_0. Digo el máximo porque si fuese el mínimo, con un valor de significación $\alpha = 0$ ya no podemos rechazar y estaría solucionado, pero sin obtener información alguna. Cuanto menor es el nivel de significación, menos probabilidad puede haber de cometer el error más grave, es decir, tenemos que curarnos en salud y no arriesgarnos a rechazar aunque esté claro que H_0 es falsa, a no ser que la probabilidad de cometer el error sea menor o igual que el nivel de significación α.

5 cosas que tengo mal

Cosas que digo mal Un estimador insesgado cumple que E[mugorro]=mu; corregir en la captura para ponerlo todo muestral, con gorritos. Definir la regresión como Y = a + bx + error y decir que el error sigue una distribución normal de media cero. Dejar claro que el modelo es la recta con el error, y yo intento aproximarlo con una recta. epsilon es una variable aleatória, que no depende linealmente de nada. En concreto es independiente de x. El error de cada muestra x es aleatório, y se comporta como una normal de media 0 . borrar el parrafo de la segunda captura sobre epsilon epsilon es la parte de Y que no se explica con X. Para ver la bondad del ajuste, comparamos la cantidad de Y que se explica con X y la que no. La siguiente captura mirar los sumatorios que empiezan en 0, y la n y la k. Dentro de esa captura revisar todos los subíndices y tal. Quitar toda esa parte por ahora.

Puntos suspensivos con el más es cldots y ldots.

Distribución fija no, distribución conocida, para cuando hable de los intervalos de confianza.

Cada semana ser capaz de explicar a modo de presentación lo que he avanzado. Dedicar tiempo a ordenarlo y hacer un esquema Apuntar dudas

Para la semana que viene probar ejemplos del boston housing de regresión con una variable y luego con varias.

Repasar dónde he puesto RSE, RSS y error cuadrático medio, que casi con total seguridad habrá fallos.