Regresiones

Aprendizaje supervisado.

Cuando una categoría no me sirve, no ataca el problema que quiero solucionar. Tenemos una instancia a predecir que es un **número** (discreto o continuo). El valor de entrada es un feature y la salida es un target (puede haber más de un target a predecir). Puede entrar más de un feature.

¿Cómo evaluamos un modelo de regresión?
Métricas que vimos: accuracy, recall, etc.
No sirve el accuracy porque es muy preciso. Como la relación es de igual a igual (y = algo) por ahí le pifiamos por muy poco a la predicción (No es lo mismo errar por 1 peso que por 1 millon de pesos al precio de una casa)

Por eso usamos el **Error cuadrático medio**.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}\underbrace{\left(y^{(i)} - \hat{y}^{(i)}\right)}_{\text{error}}^2$$

Cuanto más me alejo, más aumenta el error y es al cuadrado porque tiene que ser positivo.
Lo más importante de las métricas de regresión es que sean diferenciables.

Los modelos que vimos previamente podían con un ligero cambio ser una regresión:

<u>Árboles de regresión</u>
Lo que cambia:

- Para cada nodo de decisión se elige el feature que menos desvío estándar tenga.
- Cuando se llega a una hoja, se elige promedio de los Y que estaban en esa hoja.

<u>KNN para regresión</u>
En vez de devolver las probabilidades o la clase mayoritaria, se devuelve la media de los K más vecinos.

<u>Regresión lineal simple</u>
Consiste en predecir una respuesta numérica Y en base a una única variable de entrada X1, suponiendo una relación lineal. Predigo el Y con una recta.

$$Y \approx \beta_0 + \beta_1 \cdot X_1$$

Resultado a predecir — Variables a encontrar — Variable de entrada

Estimación de coeficientes

Residuo o error de predicción: el valor que debería haber dado menos el que aproxime. El índice (i) es cada caso puntual.

$$e^{(i)} = y^{(i)} - \hat{y}^{(i)}$$

Residual sum of squares.

$$\mathrm{RSS} = \sum_{i=1}^{n} \left(y^{(i)} - \hat{\beta}_0 - \hat{\beta}_1 \cdot x_1^{(i)}\right)^2$$

$$\frac{\delta \mathrm{RSS}}{\delta \hat{\beta}_0} = 0 \qquad \frac{\delta \mathrm{RSS}}{\delta \hat{\beta}_1} = 0$$

El cuadrado es para que sea diferenciable y para sacarle el signo.
Lo malo de que sea al cuadrado es que hace más grande los outliers (valores de error grande). Con robustez zafamos de eso.

como obtenemos β0 y β1 en RSS? derivada parcial de RSS en funcion de ambas variables. obtenemos los nuevos β

Regresión lineal múltiple

Todos los features se van a relacionar linealmente (asumimos). Usamos múltiples features.

$$Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p$$

Las medidas de error sirven de igual forma que para regresión lineal simple. Los coeficientes se estiman de forma análoga.

Evaluación del modelo (enfoque de aprendizaje automático)
Usar RMSE No sirve para esta técnica porque no se pueden hacer las derivadas parciales para hallar los b

Evaluación del modelo (enfoque estadístico)
De forma estadística podemos medir la importancia de cada beta de mi regresión lineal, ya sea múltiple o simple. La hipótesis es que el feature influye en la predicción de Y, y la hipótesis nula es que no influye. Usa el p valor. Calcula un p valor para cada coeficiente.

Podemos medir estadísticamente la importancia de cada feature. Como un enfoque estadístico para la ingeniería de features.

RSS: variabilidad no explicada por el modelo.
TSS: variabilidad total de los datos.

$$\mathrm{RSS} = \sum_{i=1}^{n} \left(y^{(i)} - \hat{y}^{(i)}\right)^2$$

$$\mathrm{TSS} = \sum_{i=1}^{n} \left(y^{(i)} - \overline{Y}\right)^2$$

R^2 : Proporción de la variabilidad explicada por el modelo.

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}}$$

Cuando RSS = 0 hay overfitting.

Esto es una forma de mirar cómo se comporta el modelo con los datos:
$R^2 \rightarrow 0$ cuando el modelo se ajusta poco de la variabilidad de los datos.
$R^2 \rightarrow 1$ cuando el modelo se ajusta mucho de la variabilidad de los datos.

Me conviene que $R2 \rightarrow 1$ porque implica que el RSS es cercano a cero. Y porque se ajusta bien a la variabilidad de los datos. Pero que no sea 1 posta porque sino tenemos overfitting

Supuestos/ sesgo que tiene este modelo

El sesgo de este modelo es que asume que las variables independientes se relacionan linealmente y hay un error random(no puedo obtener información de él, no hay un patrón de este error):

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1} + \varepsilon_i$$

Usamos el dataset que no vio el entrenamiento para evaluar en RSS y cómo esos datos no los memorizo es más útil.

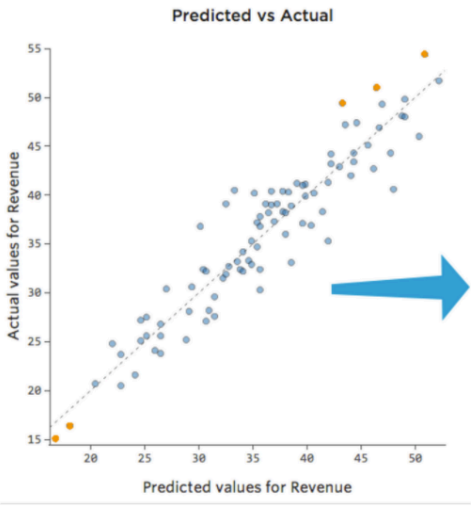
Diagnóstico de regresión

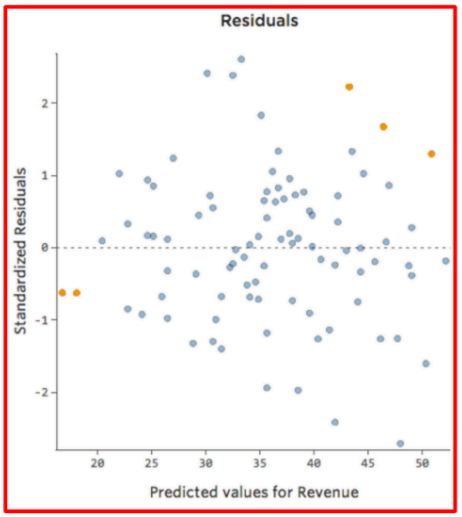
¿Cómo validamos los supuestos de que hay una regresión lineal? Vamos a estudiar los residuos para ver si los errores son random, de forma gráfica. Vamos a ver que si hay un patrón gráfico en esos residuos, entonces hay información que todavía no extraje, y estos sesgos están como refutados porque ese error debería ser random.

Si el ajuste no proporciona residuos que parezcan razonables, entonces dudamos de que el modelo lineal sea apropiado para nuestros datos.

Residual plot
El modelo lineal $Y = \beta_0 + \beta_1 X + \varepsilon$ supone que los errores ε tienen media poblacional cero y varianza constante, y que son independientes para distintas observaciones.

Para chequear que los supuestos del modelo lineal son apropiados para un conjunto de datos, suelen hacerse una serie de gráficos.

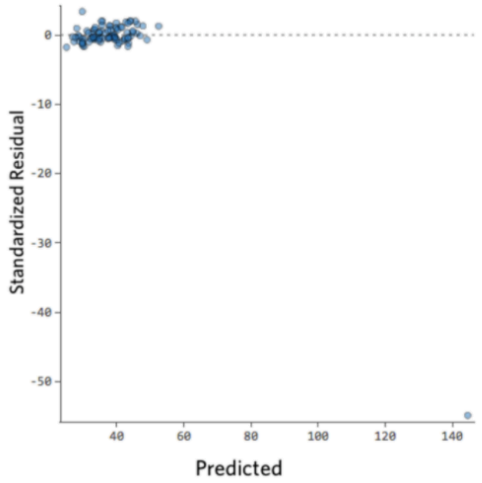




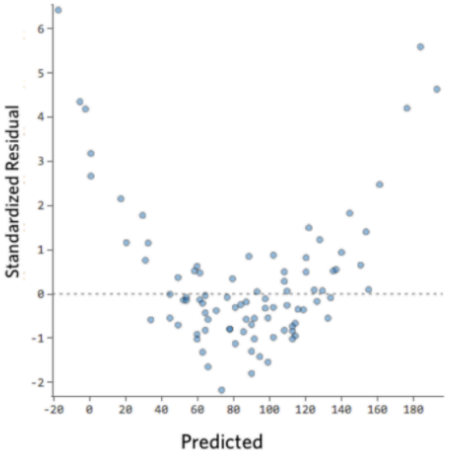
No hay patrón aparente -> Anda bien :)

Podemos observar que no hay un patrón, su media está en cero y es una nube. El preconcepto en el que se basa el modelo de que el error es aleatorio se cumple.

En los siguientes graficos vemos que si tenemos patrones:



Caso: outlier
(mirar puntito abajo a la derecha)



Caso: no lineal
(observar la curvatura de los residuos)

En el primer caso tenemos un outlier, un punto muy alejado de la nube. Y en el segundo nos da información de que entre lo predicho y lo verdadero tenemos un residuo, tenemos un patrón. Esto quiere decir que tenemos información que podemos seguir trabajando.

¿Cómo tratar estos problemas?
Si se detecta una curva en el gráfico podemos intentar transformar algunas de las variables y luego aplicar un modelo lineal.
Se puede aplicar el logaritmo a la variable Y, no estaría prediciendo bien pero para entrenar lo puedo aplicar. Es un truco para darle importancia tanto a chicos como muy grandes.
Otra posibilidad es proponer modelos que contemplen un vínculo más general entre X e Y (un polinomio por ejemplo o combinamos features entre sí)