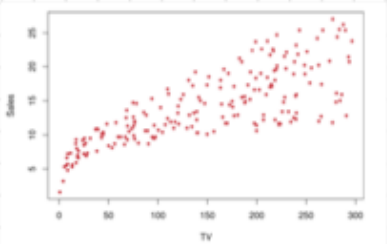
Lecture 2 - 29/08/24



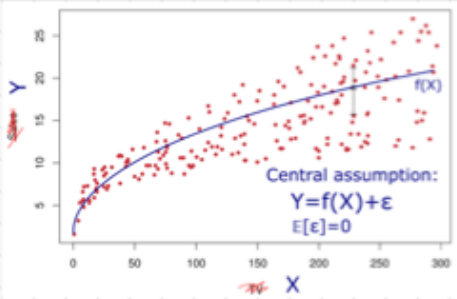
- This is a classic case of regression problem!
- The ∴ are the past observations for a company
- A question might be, How much will I sell if I spend 50€ on TV advertising? 200€ ⋮

✱ We are talking now about a linear regression problem, we see the classical input which in this case are TV ads & we wanna predict how much sells will we be able to make.

Approach:
- Basically what happens is that we wanna try to fit a line
- This line should be as precise as possible, which means that in all the parts of the graph it should be as close as possible to all the samples (∴)

We know from high school that a line is: $y = mx + c$

Result:



The hard part of regression resides in deciding what function to choose to compute our regression!

Note: Regression Not linear R.

Regression Setting

- We assume a functional relationship between X & y:

$$y = f(X) + \epsilon$$

- The noise has mean zero in this case & most of the times uncorrelated between observations, it also has costant variance.
- We can also describe this as → $E(y|X) = f(X)$

This means given X as input, then output y is given by f(x)

Linear Regression

- In linear regression to get back to what we said before, we choose f(x) as a linear function:

$$y = \beta_0 + \beta_1 X_1 + \epsilon$$

β_0 is the intercept which sometimes is called bias

This is the slope

Both are coefficients

The noise terms are assumed to be
- Independent
- Gaussian with mean 0
- Costant Variance σ^2

- In matrix formulation our $f(x)$ becomes:

$$y = X\beta + \varepsilon \quad \text{Expands} \rightarrow \quad \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

↳ The matrix X is called **design matrix**

- Now what happens if $n = p = 1$?

$$[y_1] = [1 \;\; x_{11}] \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + [\varepsilon_1]$$

$$y_1 = \beta_0 + \beta_1 x_{11} + \varepsilon_1$$

- Sometimes though variables are **not continuous**:

  Has kids? yes/no (no continuity)

  Is a student? yes/no (no continuity)

↳ This is when we **then introduce dummy variables** $\forall$ A of feature X

$$\mathbb{1}\{x = A\} = \begin{cases} 1, & x = A \\ 0, & x \neq A \end{cases}$$

This is called **one-hot encoding**

- There are **many ways of parameterising a model** with a group-specific mean:

i.e.

$$y = \beta_0 \mathbb{1}\{x = A\} + \beta_1 \mathbb{1}\{x = B\} + \varepsilon$$ where the **coefficients are the group means**

or

$$y = \beta_0 + \beta_1 \mathbb{1}\{x = B\} + \varepsilon$$ where β_0 is the mean in group A, β_1 the diff in group means between groups B & A

- **Sometimes something more complex is needed**:

* not always inputs are linear in input variables x_1, x_2
* we might need a **non-linear term** → $x_1 x_2$

This leads to → $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \varepsilon$

? now how can we **rename** this to make it look linear....

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

This is called **transformation of features**

* we tend to call this complex terms → **interaction terms**
* the models with this term have a nice interpretation

Now consider what happens between y & x_1 for $\forall$ values of $x_2 = a$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 a + \beta_3 x_1 a$$

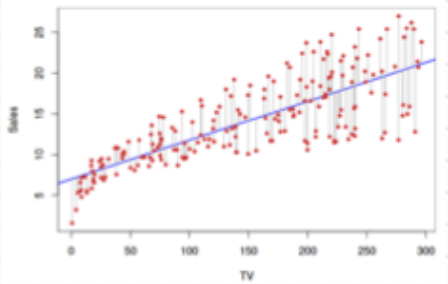
IT THEN REDUCES TO:

$$y = (\beta_0 + \beta_2 w) + x_1(\beta_1 + \beta_3 w)$$

FIRST APPROACH TO MODEL BUILDING!

① MAKE SOME SCATTER PLOTS AGAINST SINGLE FEATURES! (ALWAYS!)

② ESTIMATE THE COEFFICIENTS!



- CLEARLY A BAD MODEL IN THE PICTURE DUE TO THE BIG DISTANCE BETWEEN DOTS & THE REGRESSION LINE!

* IN THIS CASE WE TALK ABOUT ORDINARY LEAST SQUARES REGRESSION, IT AIMS TO FIND β THAT MINIMISES THE SUM OF SQUARED ERRORS, WHICH IS ALSO NAMED - THE RESIDUAL SUM OF SQUARES!

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y})^2 = \sum_{i=1}^{n} (y_i - x_i^T \beta)^2$$

ESTIMATED PARAM

$\hat{y}$ = ESTIMATED VALUE

y_i = THE REAL FEATURE

^2 IS JUST TO REMOVE THE NEGATIVE SIGN (-)

BASICALLY IF THE DIFF BETWEEN THE TWO IS VERY HIGH, OUR MODEL IS BAD!
WE ARE TRYIN TO MINIMIZE THIS DIFF BY FINDING THE BEST β'S

$$\hat{y} = \beta_0 + \beta_1 x_1 + \ldots + \epsilon$$

* IN GENERAL WE PREFER TO FIT OUR MODEL TO MAXIMIZE THE MODEL LIKELIHOOD

↘ THIS BASICALLY MEANS:

- WHAT ARE THE X's (FEATURES) THAT WILL MORE LIKELY REPRODUCE REALITY?
- GIVEN (X_i, y_i), THE $\hat{y}$ IS AS CLOSE AS POSSIBLE TO y_i
- LIKELIHOOD IS HOW LIKELY PARAMS WILL GIVE US THE DATA WE WANT

↘ REMEMBER THAT WE ASSUME GAUSSIAN NOISE, THIS MEANS MINIMIZING RSS & USING MAXIMUM LIKELIHOOD WILL YIELD THE SAME RESULT.

↘ MAXIMUM LIKELIHOOD STARTS WITH DEFINING A LIKELIHOOD FUNCTION:

$$p_Y(y_i \mid X) = \frac{1}{(2\pi\sigma^2)^{1/2}} e^{-\frac{1}{2}\frac{(y_i - x_i^T\beta)^2}{\sigma^2}}.$$

For a choice of β & σ the likelihood of the model given all data points would be:

$$\prod_{i=1}^{n} p_Y(y_i \mid X) = \prod_{i=1}^{n} \frac{1}{(2\pi\sigma^2)^{1/2}} e^{-\frac{1}{2}\frac{(y_i - x_i^T\beta)^2}{\sigma^2}}.$$

The Goal is to find the β & σ that maximize this.

Or, minimize the following negative logarithm

$$\ell(\beta, \sigma^2) = -\frac{1}{2}\left(n \log \sigma^2 + \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - x_i^T\beta)^2\right)$$

Strategy:

1. Solve for β
2. Plug in $\hat{\beta}$ & maximise the profile likelihood for σ^2

Minimizing RSS to get β

$$\text{RSS} = \sum_{i=1}^{n}(y_i - x_i^T\beta)^2 = (Y - X\beta)^T(Y - X\beta)$$

What we wanna do is differentiate & set equal to zero for each $\beta_{r}, r = 0 \ldots p$

$$\frac{\partial \text{RSS}}{\partial \beta_r} = 2\sum_{i=1}^{n} x_{ir}(y_i - x_i^T\beta)$$

Why do we differentiate?

From calculus we know that when we want to minimize or maximize a function, one way to do that is by finding that point by taking the derivative of the function with respect to the variables we are tryin' to optimize

Set it equal to zero to find the min which is where the function is neither inc or dec

Remember!!

differentiable $\forall \beta$

3 Estimate Residual Variance

- We've seen that by fixing β at $\hat{\beta}$ & minimizing log-likelihood in σ^2:

  we get $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i^T\hat{\beta})^2$ → also known as MSE

What we normally do is normalization for unbiasedness:

$$\text{RSS}/(n-p)$$

Given a large n & small p, we'll not see the difference

④ Make Predictions from the linear model

- Given a set of features $x_i = (x_{i1}, \dots, x_{ip})$ now we can predict y_i
- We say that a natural prediction is the mean of y_i
  - ↳ this is simply the val of the regression line at x_i
- We replace $\hat{\beta}$ with the expression we found for it using max likelihood:

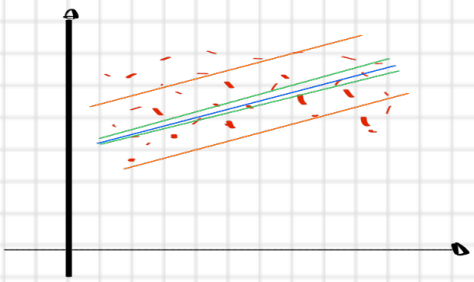
$$\hat{y} = X\hat{\beta} = \underbrace{X(X^tX)^{-1}X^t}_{H}y$$

This matrix is called **hat matrix** because "puts a hat on y" on the transformation $y \to \hat{y}$.
It's diagonal elems h_{ii} are called **leverage** & are important in model checking

$$\begin{bmatrix} h_{11} & & \\ & h_{ii} & \\ & & h_{nn} \end{bmatrix}$$

⑤ Confidence Intervals & Prediction Intervals

- Prediction Interval is **always wider** than confidence interval
  - ↳ Confidence Interval → How much the regression line will change given new data
  - ↳ Prediction Interval → Where will the predicted value $\hat{y}_i$ fall



- Regression line
- Confidence Interval
- Prediction Interval
- Real values y

From the plot:

- Normally the p-value is $p = 0.05$ or lower, this means:
  1. we are **95% confident** that for new data points, if the will change, it will stay in range.
  2. we are **95% sure** that the predictions $\hat{y}_i$ of our model will fall in the range

Building & Scrutinizing models! (recap)

- **Choose** features to be in the model **by** doing **Scatter plots**
- Features should **enter the model by**:
  - Transformations
  - Interactions between features
- **Fit the model.**
- Do **model check** to see if the model fits well! (scrutinize)

Scrutinize

- Scrutinize means validate / test.
- I.e. a test for $\beta_j = 0$ can be tested by comparing $\frac{\beta_j - 0}{\hat{SE}(\beta_j)}$ to a t-distribution with n-2 degrees of freedom.
- Model output always gives the estimated SE for coefficients, $\hat{SE}(\beta_j)$.

  Alternatively... use F-static.

This is normally used to check whether specific β's are significantly $\neq 0$ in our linear model

In simple terms it sees if the selected β is actually influencing a lot

- We can also test whether several coefficients are zero

  Ex.

  $M_1: y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2 + \varepsilon$

  $M_0: y = \beta_0 + \beta_1 x_1 + \varepsilon$

  We can see that this model is a special M_1 model where $\beta_2 = 0$ & $\beta_3 = 0$

* Note!

Now we have created two different models, M_0 & M_1, they will have a different regression line which means different result, so now we will simply compare the sum of residuals (RSS) of both & we will be able to see how influencial our β_2 & β_3 are

The F-test statistic measures how much the RSS changes when we use the simpler model instead of the more complex one.

$$F = \frac{(\text{RSS}_{M_0} - \text{RSS}_{M_1})/q}{\text{RSS}_{M_1}/(n-p-1)} = \left(\frac{\text{RSS}_{M_0}}{\text{RSS}_{M_1}} - 1\right) / \frac{q}{(n-p-1)}$$

Where q is the number of parameters dropped in reducing M_1 to M_0.
Higher F if more reduction in RSS for M_1 or/and q is small compared to $n-p-1$.

- We can also have a special test where all coefficients are 0.

  $M_1: y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2 + \varepsilon$

  $M_0: y = \beta_0 + \varepsilon$

Hypothesis Testing

- As said two nested models can be compared by F-tests.
- We though, can also compare by information criterion. → AIC or BIC

Strategies:

- Forward Selection (start by including no/few vars)
- Backward Selection (start by including all/many vars)
- Alternating Forward & Backward

Remember A

> Don't test too much!
> Use model inspection to see well-fitting model

Residuals

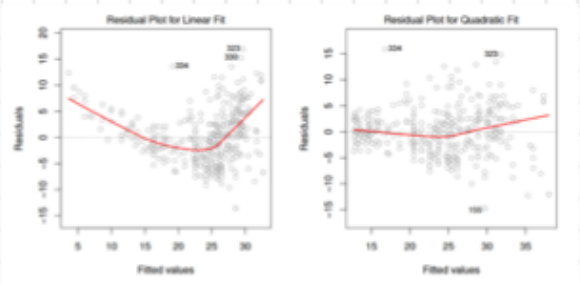
- We check residuals to see whether our model is wrong & in which ways.
- The (raw) residuals are the estimated errors:

$$e_i = y_i - \hat{y}_i = y_i - x_{i\cdot}^{\top}\beta$$

- What we can do is also have a look to standardized residuals that have been scaled by their standard error:

$$\frac{e_i}{\sqrt{\hat{\sigma}^2(1-h_{ii})}}$$

Ex.



& we see the fitted values vs the predicted which means real vs prediction

& we see the ----- at 0 because the residuals are the diff (real, estimated) so the closer this diff to 0, the more precise the estimate was.