P1

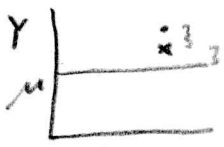
- Categorical ≡ "Quantitative" (of course)

- Regression Analysis used for

  - Prediction (PREDICTIVE)
  - Modeling (DESCRIPTIVE)
  - Testing assoc. of variables to Response. (hyp. testing)

MULTIPLE LIN REG

- Model Params are $\beta_0 \beta_1 \cdots \beta_n, \sigma^2$ (including variance!)



⇒ UNKNOWN ALWAYS! ONLY GOD KNOWS ("GOD EQN")

⇒ We always provide ESTIMATES!

For (p) predictors.

Data: [(x), y_1]

(rows in table) [( ), y_2]

y_n

n ↓ , → p

observations of predictors

(n × p) matrix. This is easier to see for you

RESTROOM
(SORRY)!!

p2

$$Y = \beta X + \varepsilon \quad \text{(these are all matrices!)}$$

$$\underset{n\times 1}{\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}} = \underset{p\times 1}{\begin{bmatrix} \beta_0 \\ \vdots \\ \beta_n \end{bmatrix}} \cdot \underset{n\times p}{\begin{bmatrix} x_1 & \cdots & x_{n,1} \\ & & \\ \cdots & & x_{n,p} \end{bmatrix}} + \underset{n\times 1}{\begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}}$$

"DESIGN MATRIX" "RESIDUAL MATRIX"

→ Order and Interactions of Regression Model.

[Highest Exponent of predictors (2nd order is $\beta_0 X_1 + \beta_1 X_1^2$)
(1st order is $\beta_0 X_1 + \beta_2 X_2 + \cdots$)

when you multiply two predictors ($\beta_1 X_1 X_2$ ≡ 2nd order)

NOT when you multiply same pred!

Ex

1st order :	$\beta_0 + \beta_1 X_1 + \beta_2 X_2 \cdots$	 parallel
2nd order :	$\beta_0 + \beta_1 X_1^2 + \beta_2 X_2 \cdots$	 "spray"
1st order INT :	$\beta_0 + \beta_1 X_1 + \beta_2 X_1 X_2 + \cdots$	 curvy parallel
2nd order INT :	$\beta_0 + \beta_1 X_1^2 + \beta_2 X_1 X_2$	 curvy "spray"

Parameter Estimation

like in simple Regression, try this

NEED $\text{Min}\left[\Sigma(y_i - \hat{y}_i)^2\right] = \text{Min}\left[\sum_{i=1}^{n}\left(y_i - (\hat{\beta}_0 + \cdots \hat{\beta}_p x_{ip})\right)^2\right]$

need some linear algebra voodoo for this one

$\hat{\beta}$

pg

But! $\text{Min}\left[\sum(y_i-\hat{y}_i)^2\right] = (y-X\hat{\beta})^T(y-X\hat{\beta})$

(1) Matrices!

Which works out to: (via the "Orthogonal Decomposition Theorem")

$$X^T(y-\hat{y}) = X^T(y-X\hat{\beta}) = 0$$

→ We want $\hat{\beta}$! (like in simple, we wanted $\hat{\beta}_0, \hat{\beta}_1$)

→ Iff X^TX is invertible (IMP! Multicollinearity!), $\boxed{\hat{\beta} = (X^TX)^{-1}X^T \cdot y}$

$$\Rightarrow \quad \hat{y} = X\hat{\beta} = \underbrace{X\left[(X^TX)^{-1}X^T\right]}_{\text{"H", the Hat Matrix}}y = Hy$$

— So far, only X (design matrix). What about Residuals?

Not so bad. $\boxed{\hat{\varepsilon} = y-\hat{y} = (1-H)y}$

INTERLUDE

- You want to square shit like in simple regression?

Say $B = \begin{bmatrix}\vdots\end{bmatrix}$ $\cdot \Sigma B^2 = B^T \cdot B =$ [illegible] $\begin{bmatrix}\vdots\end{bmatrix}\underset{n\times 1}{[\cdots]}\underset{1\times n}{} =$ Value

---

— Sampling Dist of variance σ^2

$$\hat{\sigma}^2 \approx MSE = \frac{\hat{\varepsilon}^T\hat{\varepsilon}}{n-(p+1)} = \frac{\hat{\varepsilon}^T\hat{\varepsilon}}{n-p-1} \sim \chi^2_{n-p-1}\ (\text{df})$$

???

UNDERSTAND

Why? Because we replaced p βs ... !! ???

Interpretation

"ESTIMATED"
"EXPECTED VALUE"

$\hat{\beta}_0$: Est, EV of response when all others are zero

$\hat{\beta}_i$: Est., EV of response
① holding all others fixed!
② change in response for unit change in i^{th} pred. var
↓
can be x_i or $x_1 x_3^2$, $x_1 x_4$ etc!

+ Marginal & Conditional

- ✓ Simple Linear Reg.
- Assoc. ~~Contribution~~ of just one var without other

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

✓ Mult. Lin. Reg

Assoc. of var to Resp → INTERPRET IN CONTEXT!! based on all other vars

→ This is about the association of vars → Resp!
→ Mult. lin. Reg. cannot be marginal!

+ Statement : Causality and Association

- "This causes that"
- Experimental condition!
  - change X_1, hold others constant, see what happens.
- ISOLATION!

- "Huh ... that looks like its related to that ..."
- When you are observing stuff
- NO ISOLATION!

VARIABLES
- Controlling
- Explanatory
- Prediction

Q1

$Y = \beta_0 + \beta_1 X$ population, theoretical, unknowable!

$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$ sample, estimate of

→ $\hat{\beta}_1$ ESTIMATES β_1 ("TRUE") (have to prove unbiasedness)

→ Use Sample to make INFERENCES about population coeffs

$\hat{\beta}_n$ coeffs β_n

→ Regression ESTIMATES model PARAMETERS (truth)

$\beta_0, \beta_1, \ldots \beta_n, \sigma^2$

→ Eg: $\hat{Y} = 0.89 x_1 + 0.0007 x_2$

"Cannot say this is insignificant!
Eg: x_2 could be in millions! (population of people, elevation)

Why is $Y = \beta_0 + \beta_1 x_1^2 + \beta_2 \log(x_2)$ linear?

→ "Y is still regressed linearly with x_1"

→ You put x_1^2 in Excel, R, they don't know that it is squared!!

Q2

$\hat{Y} = \beta_0 + \beta_1 x_1^2 + \dots$ R^2

$\log(\hat{Y}) = \beta_0 + \beta_1 x_1^2 + \dots$ R^2

} Cannot compare these! ∴ the response is different!

$R^2 = SSR/SST$ → Variance explained by model/Regression / Total variance

$Adj(R^2) = 1 - (1-R^2) \cdot \left(\frac{n-1}{n-k-1}\right)$ $\quad dfF = n-k-1$

→ Adding more variables will only increase R^2

→ Adj R^2 accounts for adding more (k)/Variables

↳ Use to compare adding more k/Variable to model

↳ Can be negative!

→ "fitted values" are $\hat{\beta}_0 \dots \hat{\beta}_n$ in the equation just substitute "1"

$F\text{-Statistic} = \frac{MSE_{Regression}}{MSE_{Error/Residuals}} = \frac{SSR/(k-1)}{SSE/(n-k)} = \frac{SSR/p}{SSE/n-p-1}$

(Bet. Grp / Within Grp) (Explained Var / Unexpl. Var)

Params; observations; ANOVA; MULT. REG.

"fitted values" is basically $\hat{y}_i$

(We assume that $\hat{\varepsilon}_i \sim \hat{\varepsilon}_i \sim N(0, \sigma^2)$

if this is the case, the sampling distrib of $\hat{\varepsilon}_i^2$, variance, is χ^2_{n-2} → ∴ $\hat{\beta}_0$ and $\hat{\beta}_1$ replace the real deal

(for simple regression)

93/ → Model scope is important! It is defined by the data used to train the model. (prediction) Forecasting outside of scope is EXTRAPOLATION and should be avoided !!

~~SIMPLE REG~~

Params : $\beta_0, \beta_1, \sigma^2$

Estimators : $\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2$

(STATISTICS so have Stats properties! like samp. dist!)

↳ mean (Exp. Val) and Var.

$\hat{\sigma}^2$: χ^2_{n-2} ← Samp. dist

dof : $\hat{\beta}_0, \hat{\beta}_1$ replace β_0, β_1

$\hat{\sigma}^2 = MSE = \frac{\sum \hat{\varepsilon}_i^2}{n-2}$ ∴ iff $\varepsilon_i \sim N(0, \sigma^2)$

$\hat{\beta}_1$ mean, $E(\hat{\beta}_1) = \beta_1$

∴ unbiased estimator

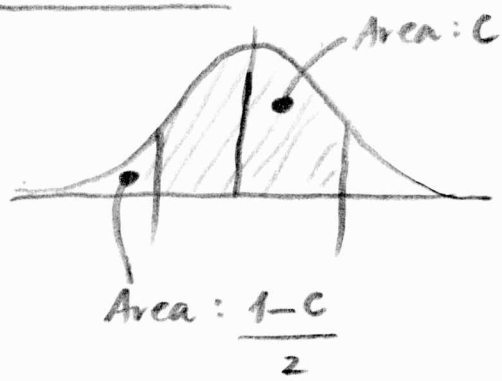
Var, $V(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$

So, Sampling dist of $\hat{\beta}_1$ is $N(\beta_1, \sigma^2/S_{xx})$

~~DIGRESSION~~

Conf. Intervals for α level :

Pop. mean = Sample mean ± Some Error



so, $\beta_1 = \hat{\beta}_0 \pm$ (Some Error)

$= \hat{\beta}_0 \pm \underbrace{t_{(\alpha/2, n-2)} \cdot \sqrt{\frac{MSE}{S_{xx}}}}_{\text{Error term.}}$

$V(\hat{\beta}_1) = \sigma^2/S_{xx}$ but we dont know pop. var σ^2, so replace : $\frac{\hat{\sigma}^2}{S_{xx}} = \frac{MSE}{S_{xx}}$

Q4/

⇒ "Statistically significant" mean that something is different from zero! That it does something, contributes to the model.

Hypothesis testing → $H_0: \beta_1 = 0$ (null hyp) $\quad H_a: \beta_1 \neq 0$

like in Normal Dist, compute the "statistic"

instead of "z-stat", compute "t-stat".

$$t\text{-stat} = \frac{\hat{\beta}_1 - 0}{\sqrt{\hat{\sigma}^2 / S_{xx}}}$$

} if Abs(t-stat) is large, REJECT H_0

(ie) if t-stat $> t_{\frac{\alpha}{2}, n-2}$

→ $H_0: \beta_1 = c, \quad H_A: \beta_1 \neq c$

same shit: $t\text{-stat} = \frac{\hat{\beta}_1 - c}{\sqrt{\hat{\sigma}^2 / S_{xx}}}$

} if Abs(t-stat) is large, REJECT H_0

(ie) if $t_{stat} > t_{\frac{\alpha}{2}, n-2}$

But! This is related to p-value.

p-val = $2 \times \text{Prob}(T_{n-2} > t\text{-stat})$

↓

∴ sum of areas of 2 tails of dist 

larger t_{stat} ⇒ smaller p-val ⇒ more evidence against null hypothesis.

(⇔ smaller error in denominator?)

Can test if β_1 is +ve or -ve.

→ $H_0: \beta_1 = 0 \quad H_A: \beta_1 < 0$

Only interested in one-tail of dist!

REJECT H_0 if $P(T_{n-2} < t_{stat}) < \alpha$

→ $H_0: \beta_1 = 0 \quad H_A: \beta_1 > 0$

one-tail! REJECT if $P(T_{n-2} > t_{stat}) > \alpha$

BACK TO INFERENCE

$\hat{\beta}_0$

mean, $E(\hat{\beta}_0) = \beta_0$ (can prove this...)

variance, $V(\hat{\beta}_0) = \sigma^2 \left[1/n + \bar{x}^2 / S_{xx} \right]$

don't know, so replace with $\hat{\sigma}^2$ or MSE

$\therefore \text{Var}(\hat{\beta}_0) = \text{MSE} \left[1/n + \bar{x}^2 / S_{xx} \right]$ — (Becomes t-dist w/ n-2 dof!)

(pop.) β_0 = (Sample) $\hat{\beta}_0$ + (Some error)

$\beta_0 = \hat{\beta}_0 + t_{\frac{\alpha}{2}, n-2} \times \sqrt{\quad}$

Can make conf int with the above like w/ $\hat{\beta}_1$.

YOU NEED the normality assumption to do any of this shit!

$\varepsilon_i \sim N(0, \sigma^2)$

Residual Analysis for

uncorrelated errors

NOT! independent errors.

G7/ For the line itself, Mean and Var?

$$\text{mean} = E(\hat{Y} \mid x^*) = \beta_0 + \beta_1 x^*$$

$\because \hat{\beta}_0, \hat{\beta}_1$ are unbiased est.!

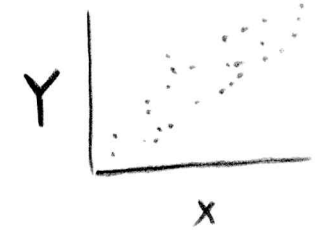
But Variance is diff. if in Estim mode or Pred mode

$\therefore$ uncertainty.

Pred mode has new var. So one more σ^2

LINEARITY $(M=0)(E(\varepsilon)=0)$
CONST VAR $V(\varepsilon_i)=\sigma^2$
IND.
NORM. $\varepsilon_i \sim N(0, \sigma^2)$

Assumptions & Goodness of Fit

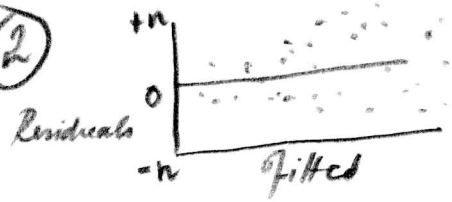
① 

Scatterplot
- LINEARITY
- CONST VAR
- INDEP

$S_{xx} = (x_i - \bar{x})(x_i - \bar{x})$

$S_{xy} = (x_i - \bar{x})(y_i - \bar{y})$

(fit the model)

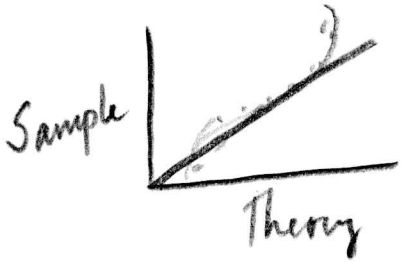
② 

Residual Analysis (Res vs Fitted), ~~QQ Plot~~
- CONST VAR
- INDEP (uncorrelated errors!)

③ 

Histogram
- ⇐⇒ NORMALITY

must be
- symmetric
- unimodal
- no gaps



QQ Plot
- NORMALITY

G8

④ Transform if Problems.

NORM, CONST VAR } Transform Resp Var, $\hat{Y}$

Box-Cox: Change $y^* = y^\lambda$

$\lambda = 1$ NO CHANGE

$\lambda = 0$ $\ln(y)$

(others ...)

LINEARITY } Transform predictors x_i

ANOVA Note: - "Variance" in name but really about means (variance in means).

- About Response Data! (Y_{ij}) (un)known population
- Model: $Y_{ij} = \mu_{ij} + \varepsilon_{ij}$ → Some mean. param to estimate.
- Assume: NORM $\varepsilon_i \sim N(0, \sigma^2)$
  CONS VAR $Var(\varepsilon_i) = \sigma^2$
  INDEP ε_{ij} are iid.

} No linearity ....

Assume Eq. Var! testing means.

- SST = SSE + SSR

(Total var)

SSE: Within group; N-k dof; $MSE = \hat{\sigma}^2 = \frac{SSE}{N-k}$

SSR: Between Group; k-1 dof; $MSR = \frac{SSR}{k-1}$

assume equal mean AND equal var.

$F\text{-Stat} = \frac{\text{Explained}}{\text{Unexpl.}} = \frac{MSR}{MSE} = \frac{\text{Between}}{\text{Within}}$

Q9

You want to jack up F-stat to disprove/reject null hypothesis: Pop. means are equal.

$H_0: \mu_1 = \mu_2 = \cdots = \mu_k$

Simple Regress : One Quantitative

ANOVA : One or more Quantitative

Multiple Reg : One or more Quanti & Qualitat-

Marginal (vs) Conditional

Consider all other params! when saying something.

Causal (vs) Association

Experimental Situation

∴ can change predictors and say something about causation.

Association statement

just observing...

You can always say there's an association (after testing w/ F, t, hypothesis tests of course)

Variables: Controlling → adjust for bias
"Default" vars

Explanatory → Explain / estimate Response

Predictive → Boost predictive power of Resp.
(even if no explanatory power)

ANOVA → Compare means of treatments — k-1

$H_0: \mu_1 = \dots = \mu_k$ means equal! — N-k / N-1

→ Multiple Reg — P model / N-P-1 data / N-1

$H_0: \beta_0 = \beta_1 = \beta_{\dots} = 0$

the predictor coefficients are zero!

Multip Regress: if $\lambda = 1$ no transform

$\lambda = 0 \rightarrow \log(y)$ $(\ln(y))$

$\lambda = 1/2 \rightarrow \sqrt{y}$

$\lambda = -1 \rightarrow 1/y$

	TRUE	ESTIMATE
	$E(\varepsilon_i) = 0$	$E(\hat{\varepsilon}_i) = 0$
	$\text{Var}(\varepsilon_i) = \sigma^2$	$\text{Var}(\hat{\varepsilon}_i) = \sigma^2(1 - h_{ij})$
MATRIX	$E(\boldsymbol{\varepsilon}) = 0$	$E(\hat{\boldsymbol{\varepsilon}}) = 0$
	$\text{Var}(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}$	$\text{Var}(\hat{\boldsymbol{\varepsilon}}) = \sigma^2(\mathbf{I} - \mathbf{H})$

Identity matrix

Slot in Hat matrix

$\Rightarrow$ Est. Residuals do not have constant variance!

we have to standardize!

ie, just divide by standard dev ☺ (of residuals)

$r_i = \hat{\varepsilon}_i / \hat{\sigma}\sqrt{1-h_{ii}}$

Outliers :
- Cannot just discard: analyze!
- "far" from the mean
- LEVERAGE point { far from mean or near obs. space
  - are { MAY OR MAY NOT affect model
- INFLUENTIAL POINTS

DO affect model!

Cooks Dist: Measures how much param value change if i^{th} obs. is removed.

General Rule if CD is $> 4/n$

or CD is > 1

take a look; analyze! do not remove yet!

Many outliers don't mean extreme values/probs

Maybe a heavy-tailed dist!!

correct it by transform ☺

- Coeff. of Determination is $R^2 = 1 - \frac{SSE}{SST} = \frac{SSR}{SST}$

"Var model explains / Total variance"

- GOF only has to do with model assumptions! has nothing to do (does not imply) Pred. Power!

- In Mul Reg, $F\text{-stat} = \frac{MSR}{MSE} = \frac{SSR/p}{SSE/n-p-1}$

- Adj R^2 = Penalize for more predictors $= 1 - (n-1)\left[\frac{1-R^2}{n-p-1}\right]$

- R^2 goes up w/ more p's. Doesn't mean shit.

- R^2 → Explain Var in Response Adj R^2 → Compare models.

- OUTLIERS: Cook's Dist. measures how much the Mul R params change when some data point is removed

Maybe problem if $CD_i > \frac{4}{n}$: investigate.

MULTICOLLINEARITY

Pearson's coeff. can help here w/ diagonals.

linear dependence of one pred var on another.
(can happen if X_a can be estimated from X_b)

Problem? X^TX must be invertible. Prob. if linear dep. for OLS !!

If X^TX is not, SE of coeffs is ∞
SE of resp. is ∞ } theoretically...

↙ Practically (in R), these are "artificially large".

→ R^2 doesn't change! just chill
→ $\hat{\beta}_j$ is unstable (Resp. changes WILDLY!)
→ SE is artificially large
→ F-stat is significant but t-stats are not!

VIF (Variance Inflation Factor) is used to find out.

$$VIF_j = \frac{1}{1 - R_j^2}$$

You are SIMPLY regressing X_j against the rest!

Prop "INCREASE" in $Var(\hat{B}_j)$ compared to if it were uncorrelated to anything else.

Ok if $VIF < Max\left(10, \frac{1}{1 - R^2_{model}}\right)$ or 1 (lowest) (One or more... not clear!)

High VIF ⇒ Wider C.I's,
Less likely that $\hat{\beta}_j$ will be evaled as stat sig.

11

VARIABLE SELECTION

also "Model Selection"

$$Y = \beta_0 + \beta_1 X_1 + \dots \beta_n X_n$$

Art & Science!

More X_n → more BIAS less VAR

Few X_n → more VARIANCE less BIAS in model. → Versus Explaining VARIABILITY!

→ We want a balance bet. BIAS and VAR

→ Reduce the impact of Multicollinearity

→ PREDICTION versus EXPLANATION *

- Want fewer vars that can account for variance
- Easily obtainable!

- Consider incl. Predicting, corr. vars
- DO NOT DO THIS! for prediction!

Other points (Esp. OVER-Interpretation)

- Var. selection is an unsolved problem in stat"
- No "Magic Bullet"
- → Highly influenced by Multicollinearity! – Include one, ignore other" V.S. can just ignore another highly related var! which can explain variability
- Nothing special about the selected variables!
- A kind of "Data Mining"
- For Obs. Studies, CAUSALITY IS RARELY IMPLIED!!!

V2

Notation: p predictors $\rightarrow$ $X_1, X_2, \ldots, X_p$

need subset $S \subset \{1 \ldots p\}$ | so pick X_2, X_3, X_7

re-index $(x_j \text{ for } j \in S)$

"subset of vars with indices in S"

$\hat{\beta}(S)$ → vector of est. coeffs for subset

$X_S = (x_j \text{ for } j \in S)$

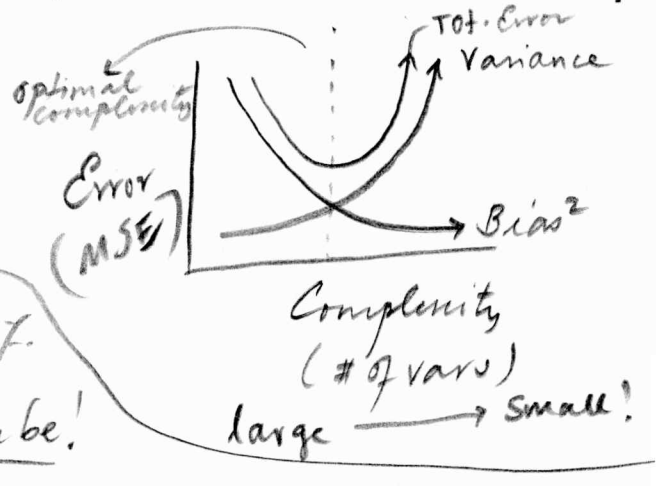
and $\hat{y}(S) = X_S \hat{\beta}(S)$ } The sub-model

Some variable may be significant in full model but not in sub-model (sans another var) and vice-versa.

Intro: Bias ⇒ underfitting → use MSE to measure!

(Bias^2)

Variance ⇒ Overfitting



⇒ None of the coeffs can be signif.

BUT the overall regression can be!

V3/

Prediction Risk Estimation

Few X_i ⇒ ↑ Bias ↓ V
Many X_i ⇒ ↓ Bias ↑ V
very Few X_i ⇒ ↑ Bias ↑ Var

We want low uncertainty for new settings
↳ future obs.

OPTIMIZ We want to min the Risk of making poor prediction

IN FUTUR OBSERV!

How to measure? for sub-model $R(S)$ → Risk.

$$R(S) = \sum_{i}^{n} \frac{E\left[\hat{Y}_i(S) - Y_i^*\right]^2}{n}$$ ← This is future obs!

PRED RISK

↳ fitted model/response.

But don't have future obs !!. so just use values again lol

$$R_{tr}(S) = \frac{\sum^{n} \left[\hat{Y}_i(S) - Y_i\right]^2}{n}$$ → Uses data twice (called "snooping")

TRAINING RISK

??? → Always prefers the more complex model!

(: ~~prefers overfitting so that it is close to zero!~~ ~~Low Var~~!

We obv. need to correct this...

Correction : $R_{tr}(S)$ + Complexity Penalty

(when model becomes too complex, ↓ goes down, ↑ goes up

??? → "Biased upward"

Why? We ♡ simple models!

Higher is worse!! (Lower = Complexity Penalty!)

So: Correction = $R_{tr}(S)$ + Complexity Penalty.

↳ Many of these!

ALL THIS FOR STANDARD LIN-REG!

① Mallow's C_p: $\frac{2|S|\hat{\sigma}^2}{n}$ → No. of predictors / model size; → Est. variance of full model!; → ~~Uses sub-model!~~

② AIC: $\frac{2|S|\sigma^2}{n}$ → This is true var. → Replaced by est. var. in R... → Either FULL or Sub-model .. variance.

Bayesian

③ BIC: $\frac{|S|\sigma^2 \log(n)}{n}$ → true var.; → Corrects for large sample size.; → Same as in AIC

This is used for PREDICTION more than the rest:
it penalizes complex models more!
so it tends to select smaller models
which may have less pred. power than AIC/MCp

④ Leave-one-out Cross-Validation

$$R_{CV}(S) \equiv \frac{1}{n}\sum^{n} E[\hat{f}_{(i)}(S) - Y_i]$$

NOTATION! (i) means "i^{th} value save i^{th} obs.!"

Can show that $R_{CV}(S) \approx R_{tr}(S) + \frac{2|S|\hat{\sigma}^2(S)}{n}$

$\frac{2|S|\hat{\sigma}^2(S)}{n}$: AIC but of sub-model!

Because of this, (use var of sub-model) LOOCV penalizes complexity Less than mallows Cp.

↓ uses FULL model!

VB

For GLM, $L_{Er}(s) = \frac{1}{n} \sum^{n} 2 Y_i \log\left[\frac{Y_i}{\hat{Y}_i(s)}\right]$

(logistic, Poisson.

$+ 2(n_i - Y_i) \log\left[\frac{n_i - Y_i}{n_i - \hat{Y}_i(s)}\right]$

(whatever bro...)

Use AIC and BIC. ∵ they are "a function of the loss-valued function"

In R, Get AIC, BIC, MCp of FULL versus SUB/Reduced Model

lower values are obv. better!

Stepwise Regression \ for p params, 2^p sub-models!
predictors

→ "Greedy" means we take the largest or smallest jump in selection process.

if p is small, fit all
if p is large: use heuristics / greedy algo

→ Usual caveats: ① No guarantee of "optimal" (∵ greedy)
② forward model ≠ back model.
(not always)
③ forward is preferred! Smaller model ♡

FORWARD

→ Start with no preds, Pick Criterion value C_0 ← eg: AIC

Then Add p_k (prefer p_k with smallest Criterion value) ← $C_k < C_0$

(or) Preds that will always be there

V6

BACKWARD :- Cannot use if predictors > sample size
$p > n$

- Computationally expensive
- Large models for larger p.

REGULARIZED REGRESSION! "Penalized" Regression
(mostly large models)

Pred. Risk $R(s) = \frac{1}{n} \sum_{i}^{n} E\left[(\hat{Y}_i(s) - Y_i^*)\right]^2$

$= \underbrace{Bias^2[\hat{Y}_i(s)]}_{\because \text{of underfitting!}} + \underbrace{V(Y_i^*)}_{\text{Irreducible Error!}} + \underbrace{V(\hat{Y}_i(s))}_{\because \text{of overfitting!}}$ → MSE!

This is all about Bias/Var tradeoff!

→ Trying to jointly penalize β's ("multivariate shrinkage")

→ We kinda like small models ∵ they are smaller

In OLS, Estimate coeffs β by

$$\text{Min}\left[\sum (y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}))^2\right]$$

ie, Sum of squared errors is minimized

BUT in Reg. Reg., SSE plus penalty is minimized.

$$\text{Min}\left[\quad + \lambda \cdot \text{Penalty}\right]$$

λ: Constant

Penalty → determine band on βs

V7/ Important: λ balances trade offs!

Lack of fit (OLS part) — Complexity (measure using Penalty)

it penalizes model complexity!!
large $\lambda \Rightarrow$ more penalty for complexity!

So what' the Penalty? (NO INTERCEPT!)

L_0: Look through all sub-models, ~~pick one~~ penalize ~~for~~ ~~with~~ # of non-zero β's LOL → NORM!
not cool for large models...
$\|\beta\|_0 = \#\{j : \beta_j \neq 0\}$

L_1 : Sum abs. val of β's $\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$
LASSO!
Minimizing "forces" β's to be zeros!
* Rewards sparsity!
→ Var Selection
L_0 L_1

Throws out related variables!

L_2 : Sum squared β's. $\|\beta\|_2 = \sum_{j=1}^{p} \beta_j^2$
RIDGE
→ Easy but does not reward sparsity!
→ Do not use for var. selection!

Use for MULTICOLLIN!

eg: $u = \{1, 0, 0 \cdots 0\}$ (p) $\quad v = \{1/\sqrt{p}, 1/\sqrt{p} \cdots 1/\sqrt{p}\}$ (p)

$L_1 = 1$ $\quad L_1 = \sqrt{p}$

$L_2 = 1$ $\quad L_2 = 1$ ←

VI/ **Elastic-Net**: Combine both!

$$Q(B_1 \dots B_p) = \Sigma(OLS) + \lambda_1 \Sigma|\beta| + \lambda_2 \Sigma \beta^2$$

different!

in R_1 use α to turn on/off.

$$Q = OLS + \lambda[\alpha L_1 + (1-\alpha) L_2]$$

when $\alpha = 1$, only LASSO
$\alpha = 0$, only RIDGE
$0 < \alpha < 1$, ELASTIC NET

$$RSE = \sqrt{\frac{SSE}{n-p}}$$

Intercept

→ This is "Bias". It doesn't fit the data properly. ⇒ Underfitting!

"Shrinkage" vs "Selection"
RIDGE LASSO
: Squares!

Cannot go to zero (can go close...) — Can go to zero

predictors

- Ridge is good for $p > n$
- "C_p" → Complexity Penalty.

In R if P-value is 1, model fits perfectly = "Complete Separation"

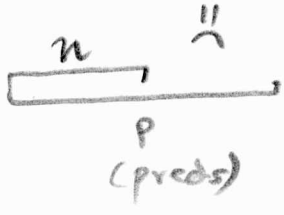
v9

Other fine points:

- Must scale predictors! (maybe unpenalized for Lasso, EN, Ridge)
- Interpret in terms of original
- LASSO: No closed form, numeric algorithm
- RIDGE: closed form
- There are less efficient than OLS
  - ∴ they increase bias! (Statquest video)
  - They inc. Bias but lower Variance.

NO MODEL SELECTION

- LASSO will pick up to 'n' vars
  - usually not a problem ∴ n > p



- RIDGE will dominate LASSO for multicollin...

Elastic

- MUST run OLS after Var. Selection!
- Evaluate this stuff
  - Normal Regression: MSE
  - Logistic Regression: Classification Error
  - Poisson Regression: Sum of Sq. Deviances.
- Mallows unreal when no Control vars!

V10

λ is determined by k-fold C.V.

for λ in (0 .... 1)

① for fold in (1 .... k)

fit model

compute/Evaluate (Eg. MSE)

then...

② Then overall error for all folds.

Now pick whichever λ (from ②) has smallest MSE

L1

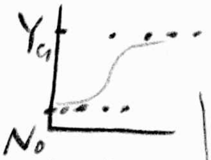
LOGISTIC REGRESSION

For Yes/No questions.

Problem is cannot use Normality Ass. $\rightarrow$ $\epsilon \sim N(0, \sigma^2)$

$\because$ Response Y is Binary..

Turn



We find that S-curve many ways to do this!

The Model

Y is a binary response.

$[(X_{11}, X_{12} \dots X_{1p}), Y_1]$
Y_2
Y_3
$[(X_{n1}, X_{n2} \dots X_{np}), Y_n]$ } Binary Response.

So, we model Probability of success : $P(Y=1 \mid X_1, X_2 \dots X_p)$

say $p = P(Y=1 \mid X_1, X_2 \dots X_p)$

so get a linear model thru a non-linear link function

$$g(p) = g(p(X_1 \dots X_p)) = \beta_0 + \beta_1 X_1 + \dots \beta_p X_p$$

$\longrightarrow$ There is no error term!

ASSUMPTIONS : ① Linearity

This is diff. than ordinary Regression
Saying that whatever g is, it is a linear combo of predicting vars.

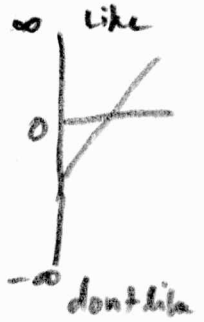
$$g(p(x_1 \dots x_p)) = \beta_0 + \dots \beta_p X_p$$

But g itself, does not have to be linear!

12/

② **Independence**: $Y_1 \dots Y_n$ are ind. rand. vars.

③ **Logit function**: Assume g is $\ln\left(\frac{p}{1-p}\right)$ | See Statquest video!



again, "p" is prob of success!

success!

So, $p = P(Y=1 \mid X_1, \dots X_p)$

so, $\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \dots \beta_p X_p$

④ $$p = \frac{e^{(\beta_0 + \beta_1 X_1 \dots \beta_p X_p)}}{1 + e^{(\beta_0 + \beta_1 X_1 + \dots \beta_p X_p)}}$$

$p/1-p$ is **odds** ... $\ln(p/1-p)$ is **log odds** (duh)

NOT the same as **odds Ratio!**

"So for a 1-unit increase, odds ratio $= \frac{e^{(\beta_0 + \beta_1 (b+1))}}{e^{\beta_0 + \beta_1 b}} = e^{\beta_1}$

ie, $\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$

→ Interpret as the $\left(\frac{\text{increase}}{\text{decrease}}\right)$ in the **log of odds ratio** for a unit increase in X_1

→ THIS IS ODDS OF SUCCESS!

Nothing to do with resp. variable!

L3/

Model Params to Estimate are $\beta_0 \dots \beta_p$

no σ^2 because no error term!

Estimate using MLE (find the best s-curve)

- This is hard. No closed-form :: log(likelihood) is non-linear
- Use Numerical Algo "/
- Get $\hat{\beta}_0\ \hat{\beta}_1 \dots \hat{\beta}_p$
- Assume Bernoulli (???)

Interlude

Bernoulli : "Will X happen or not?"

Binomial : "Prob. of success after n Bernoulli trials"

↓

So in LogReg, if repetitions, use Binomial!

---

when doing e^{β}, $e^{-\beta} = \frac{1}{e^{\beta}}$

if val is 0.732, 27% decrease/lower than ...

if val is 1.274, 27% increase/higher than ...

Odds ratio is the <u>multiplicative effect</u>

Eg: <u>Odds Ratio</u> : $\dfrac{\text{Odds}(\text{Survive} \mid \text{Smoker})}{\text{Odds}(\text{Survive} \mid \text{Non Smoker})}$

OR = 1.46 => 46% higher
OR = 0.73 => 27% lower
} That it.

L5

Inference: MLE is pretty good for large sample sizes!

If not, Type I errors.... ← not reliable otherwise !!

sampling dist of $\hat{\beta} \approx N(\beta, V)$

cannot get this! No error term! Get from software...

∴ No closed form expression...

So! ∵ N.D, Conf Int is Z-interval

$$(1-\alpha)\ Int = \hat{\beta}_j \pm Z_{\alpha/2}\sqrt{V(\hat{\beta}_j)}$$

Hypothesis Testing: Use "Wald Test" (z-test, really)

Same shit as Linear Reg BTW

Thing is... **not** a t-stat ∵ of large sample size!

Overall samples ⇒ P(Type I Error) > α (Sig. level)

Subset Testing: $\text{logit}(p) = \underbrace{\beta_0 + \beta_1 X_1 + \dots \beta_p X_p}_{\text{Controlling vars}} + \underbrace{\alpha_1 Z_1 + \dots \alpha_q Z_q}_{\text{Explanatory vars}}$ ← M_{Full}

$H_0: \alpha_1 = \alpha_2 = \alpha_3 = \dots = \alpha_q = 0$

H_A: At least one α is non-zero

$\text{logit}(p) = \beta_0 + \beta_1 X_1 + \dots \beta_p X_p$ ← $M_{Reduced}$

① Max. Likelihood function for M_1, M_2 to get:

$M_{Full} \rightarrow \mathcal{L}(\hat{\beta}_0, \hat{\beta}_1 \dots \hat{\beta}_p, \hat{\alpha}_1 \dots \hat{\alpha}_q)$

$M_{Reduced} \rightarrow \mathcal{L}(\bar{\beta}_0 + \bar{\beta}_1 \dots \bar{\beta}_p)$ (notice the hats! diff est. ∵ diff models!)

Lb/ Then use $\mathcal{L}_{Full}$ and $\mathcal{L}_{Reduced}$ to get DEVIANCE!

$$D = (\mathcal{L}_{Reduced} - \mathcal{L}_{full})$$

This is χ^2 with q d.o.f (The "p"s just 'cancel out' I guess ...)

so look for p-value = $P(\chi^2_q > D)$

Two Things: ① This is approximate! Large samp. size!
② Not a goodness of fit test!

Overall Regression Test: M_{Full} is the same
$M_{Reduced}$ is $\text{logit}(p) = \beta_0$

$H_0: \beta_1 = \beta_2 = \cdots \beta_p = 0$
H_A: At least one β is non-zero

$D = (\mathcal{L}_{reduced} - \mathcal{L}_{full})$, χ^2 w/ p d.o.f

For both, Reject Null when p is small.
- Subset: "Adding q things has no effect; the q coeffs are not stat. sig"
- Overall: "Adding anything has no effect."

LF

Classification: Log Reg is awesome for this!

for $\hat{p}(x_1, \dots x_p) = \dfrac{e^{(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \dots + \hat{\beta}_p x_p)}}{1 + e^{(\hat{\beta}_0 \dots \hat{\beta}_p x_p)}}$

a classifier $h(x_1, \dots x_p) \equiv \begin{cases} 1 \text{ if } \hat{p} > r \\ 0 \text{ if not} \end{cases}$ → (can be like 0.5)

↓

Its error is $L(h) = 1 - Pr(Y = h(x_1, \dots x_p))$

Proportion of misclassifications is Training Error.

BIASED DOWNWARD! Why? Snooping! Double-dipping! Use to train... then use for error analysis!

(Note! Training RISK is biased upward!)

So what to do? Use *Cross-Validation!* (and no future data X^*)

Split → testing / train!

- LOOCV ($k = n$)
- k-fold
- Random subsampling

Which one to use?

So use
- Random: Expensive, no clear advantage over k-fold
- k-fold: How many folds? 10 is good (LOOCV is extreme...)

Get this higher and you will ↓ BIAS but ↑ Variance !!

L8

Goodness of fit $\not\Leftrightarrow$ Predictive Power!

GOF (Goodness of fit)

Residual Analysis : Can only do this with Binary Data with Replication!

Without Replication,

$$P(Y_i | X_i) \sim \text{Bernoulli}(p(X_i))$$

This is the same thing as $\sim \text{Binomial}(1, p(X_i))$ (1 trial)

With Replication, you can see Y_i more than once for a set of X_{ip}'s

So, $P(Y_i | X_{ip}) \sim \text{Binomial}(n_i, p(X_{ip}))$

n_i trials! $(n_i > 1)$

So now, we can get residuals!

① Pearson Residuals: $r_i = \dfrac{Y_i - n_i \hat{p}_i}{\sqrt{n_i \hat{p}_i (1 - \hat{p}_i)}}$

$\hat{p}_i = \dfrac{e^{(\hat{\beta}_0 + \hat{\beta}_1 X_{\cdots})}}{1 - e^{(\cdots)}}$

L9/ ② Deviance Residuals, d_i

$d_i = \text{Sign}(Y_i - \hat{Y}_i)\sqrt{\text{some shit....}}$

"saturated" model → Response is just obs. response!

fitted model.

kinda makes sense in a way...

BOTH have an approx. SND ist
if the model assumptions hold.

→ How to check? - Histogram of Residuals
- Normality (QQ) plot.

If the deviances/residuals fit, Good fit ☺

→ Also hypothesis test.

H_0: Model fits the data
H_A: Model does not fit data.

Test stat is $D = \sum d_i^2 \sim \chi^2_{n-p-1}$

as usual reject H_0 if p-value $< \alpha$ $(\Pr(\chi^2_{df} > D))$

We actually DONT want to reject H_0!!

We want high p-values!!

L 10

Intro/Intuition

In Logistic Reg, to compare models, you are just comparing the LIKELIHOODS of each model. That's all. (that it will explain stuff)

Pred Power!

Subset Testing: $\mathcal{L}(\text{Reduced}) - \mathcal{L}(\text{Full})$ THIS IS ABT PRED POWER!

GOF Testing: $\mathcal{L}(\text{Saturated}) - \mathcal{L}(\text{~~Full~~ Fitted})$.

There are separate inferences!

* - Pred Power mean its good @ predicting even if one a more assumptions do not hold!

* - GOF $\Rightarrow$ Model Assumptions hold

Intuition

- LogR is not necessarily good for any kind of data!! Its good for prob., but... use your head !!

- Logit func. must be linear combo of X's.

↓GOF Superflat
↑Pred S-curve
↑GOF
↓Pred

Curing GOF probs. ① Add pred vars
② Transform vars.
③ Outliers...?
④ ~~N.B.~~ is not appropriate Binomial dist

:

ⓐ Corr. among responses

ⓑ Heterog in Success that has not been modeled !!

ⓐ and ⓑ lead to OVER DISPERSION

Overdispersion: the variability of the prob. estimator is larger than what would be implied by a Binomial Random Var.

So Maybe use another ~~Logit~~ s-shape function?

Eg: probit, complementary log-log (c-log-log)

probit: tightest s-shape. Inverse of CDF of N.D. ("Cumulative")

c-log-log: fits looong tails really well. Use for skewed dist.

But Logit in "canonical" → $\hat{\beta}_p$ (param estimates) are fully efficient.

also, can speak/interpret param est. in terms of Log-odds ... cannot with others (probit, c-log-log).

Interlude

Simpson's Paradox: "Reversal of association" (change in sign of $\hat{\beta}_p$) when going from marginal → cond.

P1

Poisson Regression

Rates, Counts, Shit that fits a Poisson Dist.

~~or an Exp. Dist~~

~~or a family of Exp Dist~~

But First! Generalized Linear Models

Pred vars: $n \left[\overbrace{\quad\quad}^{p} \; X_{np} \right]$ Response: Y_n

Y_n is a Random Var. from an Exp Dist family { Normal, Poisson } { Binomial, Gamma }

MODEL: "Model some function g as the Expectation of Y as a linear combo of X"

Link function

i.e. $g[E(Y|x_1, \dots x_p)] = \beta_0 + \beta_1 x_1 + \dots \beta_p x_p$

or $E(Y|X) = g^{-1}(\beta_0 + \dots \beta_p x_p)$ Inverse link

Dist	Link	Regression
Normal	$g(m) = m$	$m = x^T\beta$
Poisson	$g(m) = \log(m)$	$m = e^{x^T\beta}$
Binomial/Bernoulli	$g(m) = \log\left(\frac{m}{1-m}\right)$	$m = e^{x^T\beta} / 1+e^{x^T\beta}$
Gamma	$g(m) = 1/m$	$m = \frac{1}{x^T\beta}$

Identity

Log?

Wrong? Gamma ⇒ inverse...?

p2

Intermede

Why is Normal Dist "Exp"?

Anything is in the "Exp-Family" iff its PDF can be written as: (param of dist.)

$$f(y;\theta) = h(y)\, e^{g(\theta)\cdot T(y) - B(\theta)}$$

($g(\theta)$: link function)

To Poisson: $\log[E(Y|X)] = \beta_0 + \beta_1 X_1 \cdots \beta_p X_p$

Why is this diff. than standard Reg w/ $\log(Y)$?

Standard: $E[\log(Y)|X]$

Poisson: $\log[E(Y|X)]$ ↷ flipped!

F := V for poisson...

also, Standard: Var: constant

Poisson: $\mathrm{Var}(Y|X) = e^{(\beta_0 + \beta_1 X_1 \cdots \beta_p X_p)}$

or $\log(\mathrm{Var}(Y|X)) = \beta_0 + \beta_1 X_1 \cdots \beta_p X_p$

NOT constant.

---

* For rates and counts, if you use StReg with log-transform (and instead of Poisson), this will violate const-var. assumption.

ONLY use if you have to (prefer Poisson) if the counts are large with "variance stabilizer" $\sqrt{\mu + \frac{3}{7}}$

P3/ Model, Estimation, Interpretation

Some Rando Var Y has Poisson dist with rate λ

if $P(Y=y) = \frac{e^{-\lambda}\lambda^y}{y!}$ (PDF)

and $E(Y) = V(Y) = \lambda$

So MODEL: $\log(\lambda_i) = \beta_0 + \beta_1 x_{i,1} \cdots \beta_p x_{i,p}$ NO Error term! No Variance!

this is "log-rate"

Interpretation: "Log ratio of Rate" change for unit change in β_p (holding all else fixed)

Interpret as ratio of response rates

$$\frac{e^{\beta_0 + \beta_1(x+1)}}{e^{\beta_0 + \beta_1 x_1}} = e^{\beta_1}$$

Estimation: Use MLE again (max log likelihood)
no closed form, numerically determined...

So just like logistic, $\hat{\beta}_p$'s and their errors are all approx.!

---

ex if $\beta_k = 0.07015$ (Math score) (Award count)

+ unit increase in Math score, the log of the Expected Award count inc by 0.07 (all others fixed)

or, the rate ratio of awards would increase by $e^{0.07015} = 1.072$