Stats Final

Study guide

Modeling Basics

$y = f(x) + \varepsilon$

- all models are made up of… response, function, error
- A constant model is just a constant from the data ($\bar{x}$, mean), w/ no variable (not very good)
- we can add a variable called a predictor…
- remember… "all models are wrong, but some are useful"

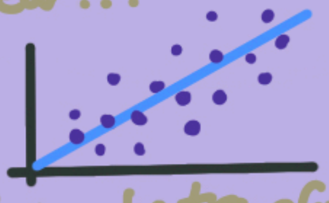
Linear Regression

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

Model

(y → response, β_0 → intercept, β_1 → slope, X → predictor)

- linear…



- but can be transformed…

common one is log…

$\log(y) = \beta_0 + \beta_1 \log(x)$

$y = e^{\beta_0} x^{\beta_1}$

or… $\log(y) = \beta_0 + \beta_1 X$

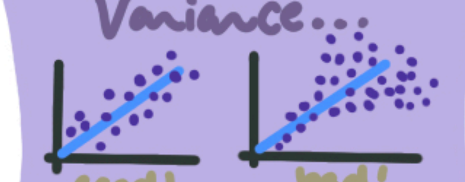
$y = e^{\beta_0} e^{\beta_1 X}$

Conditions

Linearity… resid. v fitted plots

good! even above + below — bad! shaped and fan

Constant Variance…



good! even above + below — bad! not even fan shape

Normality… qq plots…

good! close to line straight — bad! not on line shaped

- Independence
  - knowing one error won't help us find another
- Ø Mean
  - errors centered at mean. guaranteed w/ LSR

Outliers

- not exactly a condition, but still important to check.
- Outliers are points far from the regres. line
- Influential points are points that pull the regression fit.
- studentized Res. test for outliers: >±2 sussy, >±3 bad!
- leverage tests for influential points: $> \frac{4}{n}$ sussy, $> \frac{6}{n}$ bad!
- Cook's Distance combines both. >.5 sussy, >1.0 bad!

Interpretations

- β_0 → intercept. "when the predictor=0 our model predicts a response of β_0"
- β_1 → slope "for every 1 change in our predictor our model predicts a β_1 change in response"
- β_0 (logtrans) → "when the predictor is 1 our expected response is β_0"
- β_1 (logtrans) → "changing our predictor by c our model predicts a change in response of c^{β_1}"

Uses

Hypothesis Test

- is it possible there is no relationship between predictor and response
- $H_0: \beta_1 = 0$ $H_1: \beta_1 \neq 0$
- p-value, if small reject H_0

Confidence Interval

- 95% conf. interval for true slope… "we are 95% sure the true slope is betwe these values"
- Conf. interval for true mean response for predictor x_k "we are 95% con. that the true mean resp. for x_k is …"
- What would I expect the next response for x_k to be… "we are 95% con. that the next response for a predictor x_k will be between these values."

Variability Explained by Model

↳ r^2

$0 \leq r^2 \leq 1$ (0 → bad, 1 → good)

Multiple Regression

$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon$

and…

$y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$

Model

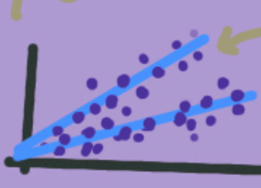
multiple regression…

- 1 variable → line
- 2 variable → plane
- 3 variable → cube

multiple regression w/ cat. var…

- $y = \beta_0 + \beta_1(\text{cat})$, cat = {0 → male, 1 → female}
  - $y = \beta_0 + \beta_1(1)$
  - $y = \beta_0 + 0$
- $y = \beta_0 + \beta_1(\text{cat. }1) + \beta_2(\text{cat. }2)$
  - $y = \beta_0 + 0 + 0$
  - $y = \beta_0 + \beta_1(1) + 0$
  - $y = \beta_0 + 0 + \beta_2(1)$
- $y = \beta_0 + \beta_1(\text{cat.}) + \beta_2 X + \beta_3(\text{cat})X$
  - $y = \beta_0 + \beta_2 X + 0$
  - $y = \beta_0 + \beta_1(1) + \beta_2 X + \beta_3(1)X$

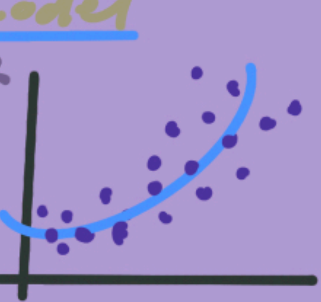
makes 2 lines on same graph



← cat = yes = 1, ← cat = no = 0

polynomial model

$y = \beta_0 + \beta_1 X + \beta_2 X^2$ → include β_1 and β_2 to get flex. curve.



Conditions

- same as simple linear regression. some extra topics…
- multicolinearity
  - We don't want correlation of 2 predictors to be too high since it hinders movement of our model. Check correlation w/ a cor. plot.
- Overfitting

underfitting — just right — overfitting

- remember → we want the model to work on new data.

Interpretation

multiple regression…

- β_0 → "when all explanatory variables = 0 we expect response to be β_0"
- β_K → "with other explanatory variables held constant a one unit increase in predictor K will lead to a β_K increase in the response"

multiple cat. regression…

- β_0 → "the expected response when all cat = Ø and X = 0"
- β_1 → "the change in response when cat = 1 is true and X = 0"
- β_2 → "the change in response when cat are held constant and we see a 1 unit increase in X."
- β_3 → "The change in β_2 when cat = 1 is true"

Uses

Hypothesis Testing

- perform indv. t-test to see if single coef is useful.
- same-ish interp. as single
- perform nested F test on full category…

$H_0: \beta_2 = \beta_3 = 0$

$H_A: \beta_2$ or $\beta_3 \neq 0$

$y = \beta_0 + \beta_1 X + \beta_2 \text{cat} + \beta_3 X \text{cat}$

VS

$y = \beta_0 + \beta_1 X$

- "we have statistically sig. evidence that the true β_2/β_3 is not Ø and the categorical variable has a relationship w/ the response".

Confidence Intervals

- same as last time

Model Selection

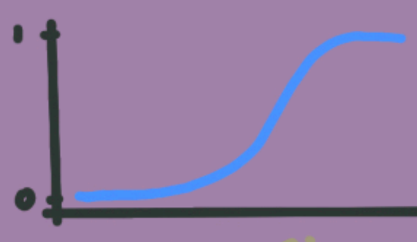
- remember → short and sweet.
- Balance high variability explained (high r^2) w/ a small ish model.
- Can also use AIC
- automated model selection methods are no substitute for thinking

Logistic Regression

$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X$

Model

- if we want to model any response between 0 and 1 we use logistic regression (proportions, percentages)



- We have a sigmoid curve and we can write the model in a couple ways…

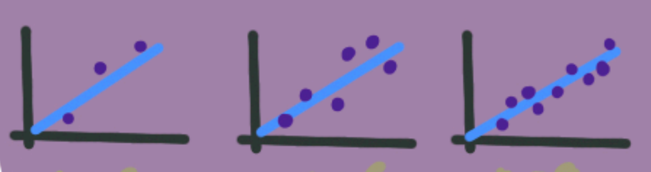
$\log(\text{odds}) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X$

$p = \frac{e^{\beta_0 + \beta_1 X}}{e^{\beta_0 + \beta_1 X} + 1}$

- log odds vs probability model
- odds → ratio of prob. of sucess to prob. of failure
- log odds → the log of the odds

Conditions

- Linearity (logit linear)
  - for binary predictors linearity is automatic
  - for non-binary predictors we create an empirical logit plot…
    1. split x into n buckets
    2. Find the mean x and mean observed response $\hat{p}$ in each bucket
    3. plot logit($\hat{p}$) vs x



n=3 — n=6 — n=9

if the plots are linear enough at multiple n's the it is logit linear.

- Randomness
  - proportions need to come from prob. (not "70% of body is water")
- Independence
  - did outcome of one unit effect outcome of another?

Interpretations

- odds ratio → e^{β_1}, the odds of one event divided by the odds of another
- β_1 → log(odds)… "as our predictor goes up by 1 the log odds go up by β_1"
  - odds… "as our predictor goes up by 1 the odds are multiplied by the odds ratio e^{β_1}"

Uses

Hypothesis Testing

- just like before…

$H_0: \beta_1 = 0$ $H_A: \beta_1 \neq 0$

if our p value is low enough reject H_0.

Confidence Intervals

- Same as before… but now we have to exponentiate (raise to e) the results.
- then conf. interval will be in odds ratio form…
- "We are 95% conf. that as our predictor goes up by 1 our odds are multiplied by ___ to ___."
- Same with prediction and prediction intervals.

Always remember to raise to e (to get odds form) or expit to get prob. form).

Multiple Logistic Regression

$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_K X_K$

Model

- logistic regression w/ categorical…

$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1(\text{cat}_2) + \beta_2(\text{cat}_3)$

$\text{cat}_2 = 1$: log odds $= \beta_0 + \beta_1$ — $\text{cat}_3 = 1$: log odds $= \beta_0 + \beta_2$

- logistic regression w/ multiple variables

$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1(\text{cat}_2) + \beta_2 X + \beta_3(\text{cat}_2)X$

- just like before… makes 2 lines…

Conditions

- same as logistic regression…

Interpretations

Categorical

- β_0 → "log odds of sucess when cat_1 = true"
- β_1 → "change in log odds of sucess from cat_1 = true to cat_2 = true"
- β_2 → "change in log odds of sucess from cat_1 = true to cat_3 = true"

Multiple Log Regression

- β_1 → "change in log odds of sucess from cat_1 = true to cat_2 = true with other variables held constant"
- β_3 → "change in relationship between X and log odds of sucess when cat_2 = true"

Uses

Hypothesis Testing

- for p-values… "prob that there is no change between β_0 and β_1 / β_0 and β_2 and we got coefs by random chance"
- If we want to test all of 1 variable… we test the drop in deviance between the nested model and the full model on X^2 distr.
- if pvalue is low enough there is a stat. sig. relationship between variable and response.

Confidence Intervals

- same as log regression. Just remember to switch out of log odds.

Anova

$y_{ij} = M_i + \varepsilon_{ij}$

(error; test stat for spec. row i; collum j; row i)

Model

	cat_2 A	cat_2 B	cat_2 C	
cat_1 1	data			averages
cat_1 2				
cat_1 3				

- this model cycles through each row and makes a response based on the stats for each spec. row.
- "for cat_1 2 our model predicts their cat_2 will be ___"

Conditions

normality

- qq plots again
- how far is each point from M_i?

Constant variance

- is variance in each category about the same?

zero mean

- same as w/ linear regression
- guaranteed by lsr

Interpretations

- M_i → "the mean score from the data for unit i"

Uses

Hypothesis Testing

- p value tells us if some means are dif.
- But which one? To know that we do a pairwise t test for each pair of cat_1.
- But w/ doing so many t-tests our prob. of a type I error sky rockets! This is called the multiple testing problem
- we use Bonferroni Correction. which picks a higher α to solve this problem.
- And then we can do our tests as normal.

Confidence Intervals

- if we want a confidence interval of the true dif. between two cat_1's means we quickly run into the same problem as above…
- Our conf. level drops as we do 2+ intervals.
- So we use Tukey's Honestly Significant Difference.
- if the int. contains Ø, it tells us same thing as a t-test.

Good Luck