Stor455 Test 2 Study Guide

Multiple Regression

continuing with...

review → linear regression... $y=\beta_0+\beta_1 X+\epsilon$

multiple variable... $y=\beta_0+\beta_1 X_1+\cdots+\beta_k X_k+\epsilon$

Model Selection

→ with add these possible variables how do we decide what to put in the model?

→ 2 Goals
- High R^2 → explains lots of variation
- Parsimony → short and sweet

→ avoid multicolinearity & overfitting

↳ Forward Selection
1. start w/ no predictors and then add best predictor
2. check p-value. if p-value is small enough, check and add another predictor!

↳ Backward Elimination
1. fit full model w/ all terms. Then kick out the worst one (largest p-value)
2. continue until all p-values are small enough

Both of these methods are automated by R, but Automation is no substitute for thinking!

Categorical Variables

→ what if we want to add categorical variables or binary predictors?

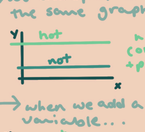
→ we start by coding binary predictors as 1's and 0's...

categories = {0 = hot, 1 = not}

$y=\beta_0+\beta_1(\text{categorical})$ → we end up w/ 2 lines on the same graph...

when hot, $=0$: $=\beta_0+\beta_1(0)=\beta_0$

when not, $=1$: $=\beta_0+\beta_1(1)=\beta_0+\beta_1$



note → (constant + parallel)

$y=\beta_0+\beta_1(\text{cat.})+\beta_2(\text{numeric})$ → when we add a numeric variable...

$=0$: $y=\beta_0+\beta_2 X$

$=1$: $y=\beta_0+\beta_1+\beta_2 X$

→ dif y-intercepts

note → still parallel

$y=\beta_0+\beta_1(\text{cat.})+\beta_2(\text{num.})+\beta_3(\text{cat.})(\text{num.})$ → when we add an Interaction Term, we can get dif. slopes

$=0$: $y=\beta_0+\beta_2 X$

$=1$: $y=\beta_0+\beta_1+\beta_2 X+\beta_3 X$

→ dif y-intercepts, dif slopes

→ But do these two really need to be seperate lines? Is there a statistically significant difference?

→ To answer this we do a Nested F-Test

- we're gonna compare SSE (note: SSE will always go up, we want to see if it goes up a lot or a lot)
- hypotheses $H_0: \beta_2=\beta_3=0$ $H_A: \beta_2\neq 0$ or $\beta_3\neq 0$

$y=\beta_0+\beta_1 X_1+\beta_2 X_2+\beta_3 X_1 X_2$ full model

$y=\beta_0+\beta_1 X_1$ nested model

if H_0 is true SSE$_{full}$ should be just slightly smaller than SSE$_{nested}$

- we compare our statistic to a F-distribution to get a p-value and if p-value below a certain α then we can say that "we are x% confident that the full model accounts for more variability than the nested model"

→ what if we have more than two categories for a variable... Multilevel Categories (markdotorg)

- we make multiple levels! If we have K levels we make K-1 predictors.

categories = {1 = gaslight, 2 = gatekeep, 3 = girlboss} → we leave out cat. 3, it's rep. by the β_0

$y=\beta_0+\beta_1(\text{gaslight})+\beta_2(\text{gatekeep})$

when gaslight=0 gatekeep=0: $y=\beta_0$

when gaslight=1 gatekeep=0: $y=\beta_0+\beta_1$

when gaslight=0 gatekeep=1: $y=\beta_0+\beta_2$

→ and of course we can have an interaction term...

$y=\beta_0+\beta_1(\text{gas})+\beta_2(\text{gate})+\beta_3(\text{gas})(\text{gpa})+\beta_4(\text{gate})(\text{gpa})+\beta_5(\text{gpa})$

→ gas=1: $y=\beta_0+\beta_1+\beta_3(\text{gpa})+\beta_5(\text{gpa})$

→ gate=1: $y=\beta_0+\beta_2+\beta_4(\text{gpa})+\beta_5(\text{gpa})$

→ gas=0 gate=0: $y=\beta_0+\beta_5(\text{gpa})$

Polynomial Regression

→ fits the data curvelinearly (not a straight line) instead of a line!

$y=\beta_0+\beta_1 X+\beta_2 X^2+\epsilon$ → have to include both terms, otherwise really restricts model

3 termed order polynomial model...

$y=\beta_0+\beta_1 X+\beta_2 Z+\beta_3 X^2+\beta_4 Z^2+\beta_5 XZ+\epsilon$

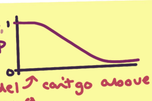
→ Look note... Over Fitting

Brand New! Logistic Regression

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

Logistic Regression

→ if we want to model proportions, rates, or prob. we need logistic regression



→ proportion needed ↗ can't go above 1 (100%) or below 0

→ transform this to live on $(-\infty,\infty)$...

p = probability, odds $=\frac{p}{1-p}$

$\log(\text{odds})=\log\left(\frac{p}{1-p}\right)=\beta_0+\beta_1 X$ ← linear! do this w/ logit in R

$p=\frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$ ← prob. form, do this with expit in R

note: notice how neither have ϵ

Interpret

Odds → ratio of probability of success to probability of failure

logodds → the log of the odds

remember → model is in log odds...

but log odds are hard to understand so lets change to odds...

$\log(\text{odds})=\hat\beta_0+\hat\beta_1 x$

$\text{odds}=e^{\hat\beta_0+\hat\beta_1 x}$ and if we add 1 to our predictor

$\text{odds}=e^{\hat\beta_0+\hat\beta_1(x+1)}$

$\text{odds}=e^{\hat\beta_0+\hat\beta_1 x}(e^{\hat\beta_1})$

↳ so if we add 1 to predictor we multiply odds by $e^{\hat\beta_1}$

this is the odds ratio...

odds ratio → $e^{\hat\beta_1}$, the odds of one event divided by the odds of another.

$\hat\beta_1$ interpretation →

w/ log odds: "as predictor goes up by 1, log odds goes up by $\hat\beta_1$"

w/ odds: "as predictor goes up by 1, odds is multiplied by the odds ratio $e^{\hat\beta_1}$"

p-value interpretations →
- "we have statistically significant evidence that the true slope of the relationship between log odds and predictor is not 0"
- "there is statistically significant relationship between log odds + predictor"

Conditions

Linearity, Randomness, & Independence...

Linearity is the only one we can really look at. the rest we just think about.

Linearity → analogit + linear

→ For binary predictors linearity is automatic

→ For Quantitative predictors we use an empirical logit plot.

How to make empirical logit plot...
1. divide x-values into buckets.
2. compute mean value of x and observed $\hat p$ for each interval
3. compute logit($\hat p$) and plot logit($\hat p$) vs mean x

→ if empirical logit plot is linear enough at multiple intervals then your good to go...

Randomness
- proportions have to come from prob. if we want to use them for tests and intervals.
- if... random assignment, random sampling, randomness in null hypothesis, or plausable randomness then random condition ☑

Independence
- if randomness fails no need to check independence
- did outcome of one unit effect outcome of another?
- if random assignment, random sampling, or reasonable assumption on then independent condition ☑

Tests & Intervals

- We'll look confidence interval, hypothesis test, and prediction. They're basically all the same we just have to convert the units.

Confidence Interval
- find in the normal way, $=\hat\beta_1\pm 2^*SE$
- But... exponentiate the results... raise to e.
- So confidence interval will be in odds ratio form.
- "we are 95% confident that the odds will be multiplied by ___ to ___ as the predictor goes up by 1"

Hypothesis Test
- "are we sure we didn't get our p-value just by chance?"
- just like before, answer w/ p-value from r output
- if p-value is small reject null hypothesis that $\beta_1=0$.

Prediction
- "what outcome will we get if we have a certain predictor?"
- Exact same as before but we have to remember to expit the answer so we get an answer in probability scale rather than log odds scale.

Multivariable Logistic Regression

- let's start w/ some categorical variables... what if we have 2 predictors

Categorical Variables

categories = {1 = Alist, 2 = Blist, 3 = Clist}

$\log(\text{odds})=\beta_0+\beta_1(\text{Blist})+\beta_2(\text{Clist})$

- works same way as before w/ baseline β_0. But remember to interpret either odds ratio or probability and not log odds.
- Also for interpretations... β_2 is change in odds from indicator 1 to 3, not from 2 to 3.

P-value interpretation → "the prob. that there is no change between β_0 and β_1 or β_0 and β_2 we get our coefs by random chance."

testing the whole variable → test $=\beta_0$ vs $=\beta_0+\beta_1 x+\beta_2 x$ We're gonna do something similar to a nested F test but with deviance instead of SSE so it's called Drop-in-Deviance Test. We measure drop in deviance as a χ^2 distributed and the p value tells us if the variable is good.