ADV. DATA ANALYSIS

MIDTERM 1

OLD

Lets start w/ reviewing basic R and data analysis steps we learned in STOR 455 ...

DATA ANALYSIS

code ...

summary()
df$newvar=
plot(varnvar,df)
cor(datat,allnumJ)

- Anytime we get a new data set we begin by looking at initial graphical images, 5 # summaries, skews & outliers, and correlations.
- We might want to clean some data or make new variables

LINEAR MODEL

lm(varnvar,df)
summary(lmod)

- form → $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{p-1} X_{p-1} + \epsilon$
  (response; intercept; parameters; predictors; error)
- We find the least squares estimate of the parameters $= \hat{\beta}$, in order to minimize ϵ.
- The difference between $\vec{y}$ (actual response) and predicted response $\hat{\vec{y}}$ are the residuals.
- Variance of the error σ^2
- We can put in qualitative predictors w/ binary dummy vars.
- We can also add interaction terms → $\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_1x_2$ → which allows us to explore how the influence of x_1 on the model changes based on if $x_2 = 0$ or 1.

INTERPRETATION

confint(mod)
cooks.distance()
halfnorm(influence(model)$hat)

- intercept → "given all predictors (X's) are 0, β_0 will be the response"
- coefs → "given all other predictors are held constant we predict a β_n increase in response for every 1 increase in X_n."
- Hypothesis Testing → To determine sign. of predictors in the model. (assume errors are iid.) Run F test and look at p-values.
  Z test → comparing 2 sample means, $\frac{\bar{x}-\mu}{\sigma}$
  t test → used with more unknowns (typically we use for testing sign. of coefs). $(\hat{\beta}-\beta_0)/SE(\hat{\beta})$ w/ df = n-1.
- Confidence Intervals → more useful in judging the effect of a predictor than a pvalue is. 95% space around predictor.
- Judge outliers and influential points w/ Cook statistic. Examine leverage w/ halfnorm plot. leverage ≠ influence

MODEL SELECTION

AIC()
Anova(m1,m2, test=)
step( )

- We can transform variables w/ things like logs if it seems there is not a linear relationship between a var. and response.
- But this can change interpretations!
- We mostly judge the model as a whole based on adjusted R^2 or AIC. We can remove variables that don't contribute. (↑ increase in R^2_a = improve variability explained)
  (↓ lower AIC = better model fit)
- Use Hypothesis testing to test 2 models against each other w/ an Anova test. Mostly we use this w/ nested models.
  (if F test is stat. sign. then larger model is better than nested model.)
- We can also have R run through a larger version of the model and pick out the best variables based on AIC (backward selection)
- Sometimes a normal linear model just isn't right, so we need to start looking at a way to expand on the linear model...

NEW!!

Time to expand on the original linear model w/ logistic regression! We will see it works much better for binary data...

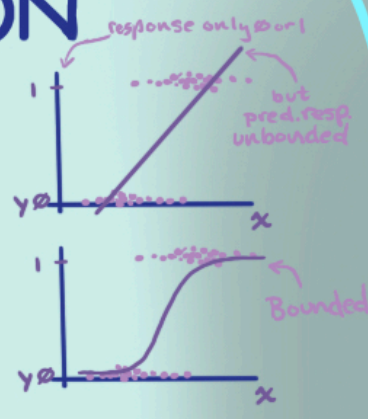
LOGISTIC REGRESSION

code ...

Logit()
ilogit()

glm(, family= binomial)

- We notice w/ a normal lm w/ a binary response (y) that the linear regression will give us pred. outside [0,1], what can we do to get a better model...
  Logistic Regression...



$\eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_q x_{iq}$
(η_i is a link function $\eta_i = g(p_i)$ → $P(y_i=1)=p_i$; this looks like linear regression but ≠ y)

→ we like Logit link function...

$\eta = \log(p/(1-p))$ aka $p = \frac{e^z}{1+e^z}$

- We find the parameters (β_i) w/ the method of maximum likelihood ... $\ell(\beta)$
- But of course now we can't use our linear interpretations for this model

INTERPRETATION

- The most natural way to interpret logistic regression is odds → an unbounded scale of probability...

3 to 1 on odds (for every $3 bet win $1) → 3 to 1, 3/1 = 0 → $p = \frac{3}{3+1} = 75\%$ chance

3 to 1 against odds (for every $1 bet win $3) → 1 to 3, 1/3 = 0 → $p = \frac{1/3}{1+1/3} = 25\%$ Chance

- To translate odds to prob. and vice versa...
  probability = $p = \frac{\theta}{1+\theta}$ odds = $\theta = \frac{p}{1-p}$
- You can see why this translates well to log regression...

$$\log(odds) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots$$
$$odds = e^{\beta_0} e^{\beta_1 x_1} e^{\beta_2 x_2}$$

- coef interp. → "a unit increase in x_i w/ others held constant, changes log odds of sucess by β_i and changes odds of success by a factor of e^{β_i}."
- relative risk → $\left(\frac{\exp(\vec{\beta}\vec{x})}{1+\exp(\vec{\beta}\vec{x})}\right) / \left(\frac{\exp(\vec{\beta}\vec{z})}{1+\exp(\vec{\beta}\vec{z})}\right) = ilogit(\vec{\beta}\vec{x})/ilogit(\vec{\beta}\vec{z})$
  ($\vec{x}$ is the vector of values for the predictors, change 1 value between $\vec{x}$ and $\vec{z}$ holding the rest fixed.)
  "The case where $x_i = \#_1$ is rr% more likely to see sucess than the case where $x_i = \#_2$."
- Hypothesis Tests → use Chi sq. tests instead of F tests. Otherwise works the same.
- Confidence Intervals → find w/ $\beta_i \pm 1.96 \cdot SE(\hat{\beta}_i)$ → note: find confint first then convert from logodds
  (z^* for .95)

MODEL SELECTION

lrtest(generalhoslem)
→ logitgof(y, predict())

- We can't judge the model based on residuals, the raw residuals always seem to form 2 lines (one near 0 and one near 1) because y can only take 2 values. Instead we use Deviance and Deviance Residuals
  (Null Dev. = Dev. of model w/ just intercept; based off likelihood ratio; smaller deviance = better fit; a transformed res. we bin and plot these as a test of fit.)
- Can still use anova (w/ chi sq test) and AIC.
- Again we can use step() / backward elimination to choose variables.
- Other Measures of Goodness of Fit →
  Hosmer-Lemeshow Stat → uses confint. to measure variation, if pvalue ≤ .05 then not a good fit.
  Missclassification → specificity and sensitivity. Test how many correct and incorrect in the model based of outcome in data. Can conceal variation.