1 variable EDA

- categorical
  - bar graph
  - percent summaries
- quantitative
  - histogram
  - center: $\overline{x}$, median
  - spread: stddev, IQR, range
  - five number summary/box plot

1 variable transformations

- need normal distributions?
- $x^{\frac{1}{n}}$, $\log(x+c)$ so everything is > 1.
- the above's inverses
- quantile plots (qqplot) can help us diagnose if normal enough (look for straight line)

2 variable EDA

- explanatory x axis $\rightarrow$ response y axis

1 variable inference

- statistics ($\overline{x}, S_x, ...$) predicts parameters ($\mu, \sigma, ...$)
- components:
  - point estimation: estimate via single number calculated
  - interval estimation: give plausible interview and how plausible
  - significance testing about hypotheses: assess evidence for/against claim about
- 95% confidence interval for μ is $\overline{X} \pm 2 \cdot \text{SE}_{\overline{X}}$
  - (works for arbitrary parameter/statistic estimate)
  - any sample Standard Error SE is $\frac{S}{\sqrt{n}}$ with sample stddev S (but remember, we just use software)
  - technically, 2 should be t_{crit} which varies with n, but it approximates to 2 for 95% confidence when large n
- hypotheses testing
  - H_0 vs H_A
  - "p value is compared to significance level. we do (not) reject the null hypothesis. we do (not) have sufficient evidence that ..."
  - remember: p finds boolean evidence of difference from norm, not magitude of difference

Statistical Model Primer

- statistical models are often of form: quantity = expectation + error
- in 1 variable, eg: $Y_i = \mu + \varepsilon_i$ where μ is the prediction and ε_i is the error at i.
  - we also specify the distribution and mean + stddev of the errors
- in 2 variables, eg: for some X axis value, $Y_i = \mu_{Y|X} + \varepsilon_i$
  - we also specify the shape, center, spread of the distribution of errors

Simple Linear Regression

- our model idea is $Y_i = \beta_0 + \beta_1 X + \varepsilon_i$ where we assume the errors are
  - independent, mean 0, constant stddev/spread (for required for least squares)
  - are normal (required for inference)
  - (can be denoted iid, $N(\mu = 0, \text{variance} = \sigma^2)$)
- our **sample** regression equation is $\hat{y} = b_0 + b_1 X$
- notice that we have three parameters: β_0, β_1, σ
  - they are estimated by b_0, b_1 (when using least squares), and $\hat{\sigma}$: what R calls "Residual standard error"
- to apply the model:
  1. **state** the model
     - eg: "we use the SLR model. vision $\text{distance} = \beta_0 + \beta_1 \cdot \text{age} + \varepsilon_i$ where errors are independent, mean 0, constant stddev, normal.
  2. **validate** the data works for the model
     - linearity: visual inspection
     - errors are:
       - independent: residual plot. residuals "patternlessly" above and below 0 line.
       - mean 0: residual plot. reasonably centered around 0.
       - constant stddev: residual plot. reasonably constant spread, scanning left to right
       - normal: normal qqplot, follows line
     - if there are problems, consider diff model/transformations
  3. **estimate** the parameters
     - use software to find $b_0, b_1, \hat{\sigma}$
  4. **inference**: is data probably showing a relationship between X and Y?
     - t test for $\beta_1 =$ or $\neq 0$
  5. **measure strength** of model with R^2 (if not chance)
     - R^2 is the percent of variability in Y that can be attibuted to the linear relationship with X
     - "Multiple R-squared" in R. NOT "Adjusted"
  6. **predict** of Y from X (for individual with X or all people with X)
     - the equation predicts the point estimate of Y given X
     - get prediction vs confidence interval via R for probable values of Y for individual or all at X

Nonlinear Relationships?

- can use a nonlinear model (same four error assumptions)
- can transform it
- transformations often preferred: fewer parameters make a simpler model
- make sure to not overfit!

Multiple Regression

- we're often interested in predicting a Y from multiple explanatory X_i
- when contribution from each X_i is linear, we have *multiple linear regression*:

$$Y_i = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p + \varepsilon_i$$

where errors are
  - independent
  - mean 0
  - contant stddev
  - normal
- $p + 2$ parameters: $\beta_{\{0-p\}}$ and σ
  - like SLR, σ is stddev of errors, ie typical deviation of Y from regression hyperplane
  - $\hat{\sigma}$ in R is still "residual standard error"
- each β_i is the avg change in Y when X_i increases by 1 unit and the other Xs remain fixed
- eg `school.mod = lm(GPA ~ IQ + SelfConcept, data=school)`
- to apply the model:
  1. **state** the model
  2. **validate** the data works for the model with EDA
     - scatterplots of Y against *each* explanatory (w/ `pairs` plot). linearity: visual inspection
     - error conditions (also just use a residuals/qqplot):
       - independent: residual plot. residuals "patternlessly" above and below 0 line.
       - mean 0: residual plot. reasonably centered around 0.
       - constant stddev: residual plot. reasonably constant spread, scanning left to right
     - if there are problems, consider diff model/transformations
     - low multicollinearity (each X_i weakly correlated with each other) (might otherwise get mathematically impossible/conceptually inappropriate, misleading results. see `media/high_multicollinearity`)
       - can *informally* investigate via: correlation matrix, odd parameter estimates, oddly large estimate stderrs
       - mathematically diagonse via variance inflation factor (vif)
         - let a model be $Y \sim X_1 + X_2 + X_3$
         - vif of X_i is $\frac{1}{1-R^2}$, with R^2 from $X_i \sim$ the other Xes.
         - i.e., vif of X_1 depends on $X_1 \sim X_2 + X_3$
         - BUT: just use software.
         - when high multicol., drop variables: check diff subsets of Xes, recheck diagnostics for each. find best model with R's *adjusted R-squared* (adjusts for different number of explanatory variables. otherwise, R-squared would be higher with more variables, rmbr?)
         - BUT: also just use software (best subsets routine)
         - vif ≥ 2.5 is concerning
  3. **estimate** parameters w/ software
  4. **inference**: is data probably showing a relationship between X_i and Y?
     - F-statistic: tests if *any* of X_i are important for predicting Y
     - individual T-tests: tests if *each* X_i is a significant predictor *in the presence of all other explanatories*
  5. **predict**: use model, with R^2 for its effectiveness
     - multiple R-squared: proportion of variation in Y that can be explained by all of X_i. has a few properties:
       - closer to 1 = better "fit"

  - can only increase with more predictors
  - diminshing returns

including categorical explanatories?

- check for no interaction between predictors parallel lines iff X_2 doesn't depend on X_1 iff no interaction between X_1 and X_2 iff X_1 effect doesn't depend on X_2.

  why is this so verbose from the slides :/

- assuming no interaction between predictors:
  - include a binary indicator/dummy variable (1 if smoker, 0 else)
  - call the category defined as 0 a "baseline" category
- if a categorical variable has, say, 3 options, we get *2* dummy variables, both binary with 0 representing baseline group.
  - "controlling for years of seniority, dept A makes X less than dept C on average"
  - "holding dept constant, we estimate for every extra year of seniority, salary increases by X on average"

what if categorical explanatories *have* interaction?

- let us investigate a situation where calories $\sim$ carbs, but with slopes that differ depending on whether the item is meat.
- new model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 \cdot \text{DummyMEAT} + \beta_3 (X_1 \cdot \text{DummyMEAT}) + \varepsilon$$

  - capture the difference in slopes with an "interaction term" (the β_3 term above)
  - `lm(Calories ~ Carbs + Meat + Carbs:Meat, data=fastfood)`
  - also try actual code: `summary(lm(Sepal.Length ~ Petal.Width * Species, data=iris))`
- assumptions:
  - population relationship linear within each level of Dummy
  - within each level of Dummy, the errors are indep, mean 0, const stddev, normal (i.i.d., $N(0, \sigma^2)$)
- IMPT: if interaction term stat. significant, then those explanatories must be kept (regardless of their individual variable p-values)
  - this just means that their slopes are indeed different, i think
  - so if they are not significant go back to normal multiple regression ig
- coefficient β_3 interpretation:
  1. difference in slopes; i.e., how the quantitative X_1 effect depends on the group Dummy value
     - "For every unit increase in X_1, the change in Y is β_3 greater/less on avg in Dummy_1 than in Dummy_0
  2. equivalently: how the vertical difference between the lines changes; i.e., how the group Dummy effect depend son the quantitative X_1 value
     - "For a particular value x_1 of the quantitative variable, "

good luck on exam 1 <3

One way ANOVA

- recall: we first learn the regression models (Q → Q)
- now we investigate the ANOVA (anal. of variance) models (C → Q)
- big picture: evidence of difference in means of our cat. groups?
  - $H_0 : \mu_1 = \mu_2 = ...$

    H_1 : means not all same
- model:

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

  where j indexes the *independent* cat. populations, i the individuals

  same errors as regression: $i.i.d., N(0, \sigma^2)$

  - $k + 1$ parameters: each population mean and stddev of pop. errors
- NOTE: One way ANOVA is stat. identical to mult lin reg with categorical X, $k - 1$ dummy variables
- apply model steps are similar to regression:
  1. **state** the model
  2. **validate** the data works for the model with EDA
     - side by side boxplots (Y vs groups), stats for each gruop
       - if largest stddev ÷ smallest stddev ≥ 2, use $\alpha = 0.025$ in F test, instead of $\alpha = 0.05$ (called Keppel Correction/spread rule of thumb)
     - groups independent: nature of study. eg no time dependence, 1 person 1 group
     - error conditions:
       - independent: residual plot. residuals "patternlessly" above and below 0 line.
       - mean 0: residual plot. reasonably centered around 0.
       - constant stddev: residual plot. reasonably constant spread, scanning left to right
       - normal: qqplot
     - if there are problems, consider diff model/transformations (OF THE RESPONSE)
  3. **estimate** parameters w/ software
  4. **inference**: is data probably showing a relationship?
     - significant with ANOVA F-test (see above)
     - if yes, suppl. with 'multiple comparisons'
  5. **predict**: use model, with R^2 for its effectiveness
     - multiple R-squared: proportion of variation in Y that can be explained by all of X_i. has a few properties:
- we may derive the ANOVA F-statistic (dist shaped soooorta like chisq):

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Major	3	939.9	313.28	46.6	<2e-16
Residuals	136	914.3	6.72		

$$F = 46.6 = \frac{313.28}{6.72} = \frac{939.9/3}{914.3/136}$$

$$R^2 = \frac{939.9}{939.9+914.3} = 50.69\%$$

- and ANOVA R^2 value:

- ok, now we believe means not equal. but *which* means?
  - pairwise multiple comparison of means.
  - we can do this via manual CI inspection for means, but more groups mean much higher false positives.
  - so we use Tukey test (check tukey intervals don't overlap):

Logistic Regression (Simple Binary case)

- simple (one X) binary (Y = 0 or 1)
  - we can get proportion from a dataset eg with `prop.table(table(dataframe$succeeded))`
- logistic regression: Q $\to$ C wow!
- probability $p = P(Y = 1 \mid X = x)$
- suppose Y binary, p defn above.
  - odds favoring Y = 1 are $\frac{p}{1-p}$ (IMPT: this just means if odds=3, odds of Y = 1 are 3:1.)
  - equivalently, $p = \frac{\text{odds}}{1+\text{ odds}}$ via math
- simple binary logistic reg model is

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X$$
$$\text{i.e. odds}_1 = e^{\beta_0+\beta_1 X}$$
$$\text{i.e. } p = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$$

  no error term! $|\beta_1|$ measures "steepness" of log
- linreg: eyeball linearity. hard with logistic. use "goodness of fit" test in R (rg though labs for `ResourceSelection` library, `hoslem.test`)
  - output p value: H_0 : good fit.
- once GOF test passes for sample, test model for appropriateness for population. check β_1 hypotheses test ($H_A : \beta_1 \neq 0$, sig. relationship)
- interpretation:
  - e^{β_0}: odds, on avg, favoring Y = 1 when X = 0.
  - e^{β_1}: for every unit incr in X, odds are *multiplied* by this. (odds ratio!)
- inference: sign. reln? just look at p value for β_1 as usual
  - let's say a CI for β_1 is (lower, upper). to get CI for odds ratio, just e^{lower} etc
- in linreg, R^2 measures association. here, there is no mathy nice measure.
- we use a rough measure: percentage of "concordant pairs"
  - consider: 11 success 14 failures. $11 \cdot 14 = 154$ pairs.
  - a pair is concordant iff the success has a higher prob than the failure in the pair (discordant otherwise)
  - where prob is taken from predicting with the model
- putting it all together:

```r
titaniclogit <- glm(
  factor(survived) ~ sex + age + factor(pclass),
  family = binomial(link="logit"),
  data = titanic
)
summary(titaniclogit)
```

- Multiple case: adding more betas just like in linear reg (... is ... times more ... on avg or whatever)

(Multi)nomial Logistic Regression

- Y more than 2 groups
- odds favoring finishing category n over reference c, each has their own logistic reg model
- c categories: $c-1$ bin log. reg.s w/ $k+1$ params (has unique $\beta_{\{0-k\}}$ per cat.)
- "estimate that for every year older, the odds of pref candidate A over mayor are multplied by 0.82 on avg, controlling for ..."
- "estimate that the odds of favoring B over the mayer for female voters are 3.29 times the odds that male voters favor B over the mayor, on avg, controlling for ..."

Ordinal Logistic Regression

- Y more than 2 groups
- $P(Y \leq 1)$
  $P(Y \leq 2)$
  ...
  $P(Y \leq c-1)$ (at c prob is always 1. not interesting.)
- cum odds $\frac{P(Y \leq j)}{1-P(Y \leq j)} = Y \leq j$ on top
- otherwise same as multinomial ... EXCEPT only the constant betas differ.
- "odds of no higher than..."

Proportional-Odds Cumulative Logit Model

(end material for exam 2)

Introduction to Statistical Learning

Classifiers

- commonly, we search for cat. response (e.g. pedestrian or not?)
- find boundary in variable space to separate
- EDA: pairs plot

Logistic Regression as Classification

- take log reg, then separate by which side of 0.5 p is on

Discriminant Analysis

- only defined for quantitative X
- assume within each class, variables are Normal
- Linear DA: boundaries are lines/hyperplanes
- Quadratic DA: boundaries are quadratic hypersurfaces

- assumed Normals can be “stretched” or “twisted”

Classification Trees

- split data into those hyperrectangular things: make decision tree!
- higher nodes are more important
- can use both cat. and quant. X
- CAREFUL overfitting!
  - prune to min decisions manually or via algo
  - split b/t train and test data
  - ensembles
    - bagging trains models and takes avg (RANDOM FOREST!)
    - boosting sequentially combines and error corrects (takes weighted avg)

More about classification tress: The Gini Index

- tree algo decides splits via on purity, often via Gini index
- Gini $\in [0, 1]$ of decreasing purity

Intro to Clustering

- UNSUPERVISED!
- create clusters in variable space

hierarchical aka agglomerative

- start w/ each obs. in a cluster, aggregate based on defn of “closeness”
- “closeness”
  - single linkage: close as nearest nodes (form snakes)
  - complete linkage: close as furthest nodes (form balls)
  - avg linkage: balanced
- assess via adjusted rand index (ARI) $\in [-1, 1]$ of increasing quality

k-means

- hyperspherical clusters in variable space
- very fast algo, simple, better for data w/ Normal around averages
- minimize within-cluster sum of squares (WSS)
- choose k via elbow plot: incr k untill WSS doesn’t steeply drop (care overfitting!)