Stat 20 Final Cheat Sheet

Bayesian Estimation

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

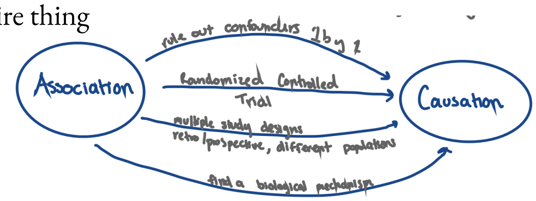
P(data or more extreme | H_0 = True): result of a hypothesis test
Prior P(parameter): probability distribution for a parameter that summarizes the information that you have before seeing the data
Posterior P(parameter | data): distribution of parameter after conditioning on the data

Causality

Descriptive Claims: numerical summaries/graphical summaries, data fundamentals (dataframes and data taxonomy)
Generalization: use data to reason, CIs, boostrapping, hypthesis tests and permutation, account for uncertainty and statistical bias
Causal: A causes B, counterfactual ("if A didn't happen, B will never happen"), experiments challenge this, confounding variables
*Look across time for two states of the same unit, compare two units (ppl) at the same time
Prediction: regression models for multivariate data, uses: prediction description inference

Experimental Design

Observational Study: don't interfere (easier, cheaper, historical data, ethical); **Experiment**: interfere! (establishes causation)
Principles of Exp. Design:
***Replication**: Within a study, replicate by collecting a sufficiently large sample/replicate entire thing
***Control**: compare treatment of interest to control group that isolates the effect of interest
***Blinding**: subjects do not know whether they're in the control or treatment group
***Random Assignment**: randomly assign subjects to treatments
Double-Blinding: subjects and researchers both don't know the treatment assignments
Placebo: fake treatment, often used as the control group for medical studies



Placebo effect: experimental units showing improvement simply because they believe they are receiving a special treatment
Correlation ≠ Causation! (**Confounders**: other factors that could've caused the result, like genetics vs smoking for cancer)

Correlation/Linear Models/Predictions

Correlation Coefficient: measures the strength of the linear relationship (linear, neg/pos, strong/weak); summarize(r = **cor(var1, var2)**)
Linear Model: expresses a predicted value ($\hat{y}$) for y, $y = b_0x + b_1$ $b_1 = \frac{s_y}{s_x}r$ $b_0 = \bar{y} - b_1\bar{x}$

```
summarize(r = cor(Graduates, Poverty),
          sx = sd(Poverty),
          sy = sd(Graduates))
mutate(b1 = sy/sx * r)
```

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Residual: diff between observed and predicted value ($e_i = y_i - \hat{y}_i$)
Estimation: m1 ← **lm**(result ~ cause), attributes(**m1**), coef(m1), fitted(m1), residuals(m1)
Ordinary Least Squares: minimize sum of squared residuals; **Slope:** estimated diff betw 2 vars

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \mathbf{b_0} - \mathbf{b_1}x)^2 = f_{RSS}(b_0, b_1)$$

Numerical Optimization: Nelder-Mead Algo; **Analytical Approach:** solve partials of $f_{RSS} = 0$
Prediction: newx ← tibble ([named x var] = [# of y choices]), predict(m1, newx);
Description (lin relation betw vars) **Prediction** (predicts unknown vals y_i) **Residual Analysis** (deviation of prediction to observation)
R^2 Value: the proportion of total variability in the model's relationship; SSR/TSS, 1=accurate

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = TSS \qquad \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = SSR$$

Interpolation: new datap oint is within original range used to fit the model
Extrapolation: new datapoint outside original range $\mathbf{b} = (X'X)^{-1}X'Y$
Residual Plots: look for heteroskedasticity (inc/dec variance in residuals) and non-linear trends

```
poverty %>%
  mutate(yhat = fitted(m1),
         res  = residuals(m1)) %>%
  ggplot(aes(x = yhat,
             y = res)) +
  geom_point(size = 3) +
  geom_text_repel(
    aes(label = State)) +
  theme_bw(base_size = 18)
```

Inference for Regression

Statistical Inference: use statistics calculated from data to makes inferences about the nature of parameters
Parameters: true slope (B_0) and intercept (B_1), **Statistics**: slope (b_0) and intercept (b_1) we calculate from data
Classical Tools of Inference: confidence intervals, hypothesis tests
H-Test: H_0 = there is no relationship betw var1 and var2, $B_1 = 0$, if there is no relationship, the pairing betw X/Y is artificial → permute!
  1) Generate subsets under H_0 by shuffling X, **2)** Compute new reg line for each dataset and store in b_1, **3)** compare with H_0 dist.
T-Test: compares the means of two samples, used in H-testing, lm() summary always uses t-test
*the test statistic associated w/ *b*'s is distributed like *t* random vars with *n - p* degrees of freedom

```
null <- ump %>%
  specify(change ~ unemp) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 500, type = "permute") %>%
  calculate(stat = "slope")
```

Linearity: linear trend between X and Y, check with residual plot
Independent errors: check with residual plot for serial correlation
Normally-Distributed Errors: look for constant spread in residual plot

$$\frac{b - \beta}{SE} \sim t_{df=n-p}$$

```
get_p_value(obs_stat = obs_slope,
            direction = "both")
```

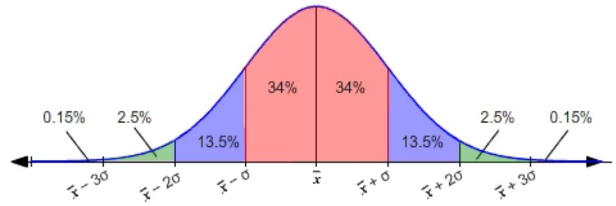
Equal Variance in Errors: look at histogram of residuals
H-tests: small sample sizes require normal errors, large sample sizes (CLT) no normality, permutation+bootstrap needs reasonable size

Multiple Linear Regression

*Allows us to create model to explain one numerical var as a function of many explanatory variables (num or cat), **confint(m1)**
True model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \epsilon; \quad \epsilon \sim N(0, \sigma^2)$ **Estimate fitted model:** $\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_p x_p$
Steps to Model Building: Statistical question, data wrangling, eda, modeling, interpretation
Exploratory: seeks to uncover trends in data, **Confirmatory**: starts with a specific question to confirm

$$R^2_{adj} = 1 - \frac{SSE}{TSS} \cdot \frac{n-1}{n-p-1}$$

Function for $\hat{y}$..
- line / when 1 numerical x
- parallel lines // when 1 numerical x_1 & 1 categorical x_2
- unrelated lines when 1 numerical x_1, 1 categorical x_2, & 1 interaction $x_1 : x_2$
- plane when 2 numerical predictors x_1 & x_2
- parallel planes 2 numerical, 1 categorical
- tilted planes 2 numerical, 1 categorical, 1 interaction



Model Interpretation

Simpson's Paradox: a trend appears in several groups of data but disappears or reverses when the groups are combined
Ecological Fallacy: false assumption that relationships that exist at an aggregated level (ex. betw states) also hold at the individual level
*__avoid__ → restrict interpretations to observational level recorded at, don't use aggregate vs individual for inferences
Response Variable: log(Obs/Exp) observed num, then expectation, then take the log to account for the right skew
Akaike Information Criterion (AIC): used to compare different possible models and determine which one is the best fit for the data
*calculated from **number of indep vars** used and the **MLE** (how well the model can reproduce the data) (lower = better)

PRACTICE TEST

1. Random var, observed 20 draws, left skewed hist w/ outliers, **X ~ Binom(n=8, p=.8)**, higher p = cluster to right, x=#success, y=p(x)
2. **Left skewed** = mean is less than the median, **right skewed** = mean is greater than the median
3. **Bootstrap** = Confidence interval, **Permute** = Testing null hypothesis
4. If the confidence interval goes from **95% to 80%**, the width of the resulting interval **decreases**
5. If the **sample size doubles**, the width of the resulting interval **decreases**
6. A multiple least squares regression model is inappropriate when there is **non-constant variance** in the residual plot
7. Logistic regression = describing a **0-1 relationship** (i.e., season performance ~ Cal team winning against Stanford)
8. Logistic regression has no assumption of **normally distributed errors**, it uses the **log** function to predict 0-1, the intercept determines the **left-right shift** of the s-curve on a scatterplot, can be used when explanatory var are two-level **cat OR num** vars
9. R^2_{adj} and AIC (lst sq, log reg) both **improve** as a model more **closely describes** the data + **worsen** as model **complexity grows**
10. Negative slope = **negative sign** of estimated coeff of the x-axis variable
11. If there is a **dummy variable** (in the legend), compare when its 0 vs 1, if the y-axis is below when it is 1, the estimated coeff is **negative**
12. **Interaction coeff** = look at when the dummy var is on, group like mpg = (b0+b2) + (b1+b3)age to see b1 < b1+b3, so b3 = pos
13. **Sensible population** = similar location, age, and time, like population that also sells/buys diamonds in similar time/city
14. **Confidence interval** CI = x ± z·SD, typically use 2 SDs to calculate 95%, 1 SD to calculate 68%
15. LB/UB = **95% confident** that the parameter associating var1 and var2 is between LB/UB NOT 95% probability cuz P()=100%
16. In the second model, **size is constant** so quality has more impact, but when **not** controlling for size, size is **more impactful** on price
17. Look for **non-linearity** in the graph for concerns about using a linear model
18. Buy with **asking price below** the predicted model and sell them at or above the predicted value
19. For the **first 5 rows**, specify what the range of the variables are, include all variables and an id
20. **Causal claim** is looking at if A causes/affects B, like whether school funding→ academic performance
21. **Negative slope** = negative correlation coefficient

```
diamonds %>%
  filter(carat < 1) %>%
  group_by(quality) %>%
  summarize(avg_p = mean(price))
```

22. **ggplot**(df, aes(x=expenditure, y=sat_score)) + **geom_point**(), make sure to label axis, ticks, and title!
23. With 2 variables vs 1, look at which variable we **controlled** in the first, Simpson's Paradox (trend goes away when group combines)
24. By controlling for a variable, we can gain a better understanding of another, otherwise potentially confounding variable
25. **Building experiment** = mention *random* sampling, experimental group, how to record the data accurately, experimental unit
26. **Test statistic** = **diff in props** of disordered flights between the control and treatment group! **test_stat** = $p_{normal} - p_{disordered}$
27. When running infer code, include **order** = c("treatment", "control"))
28. If the observed test statistic is **center of null**, **p-value = 1(ish)**, so implies it is unrelated and >alpha(.05), **fail to reject**!
29. **Test stat** = sum of the squared difference aka **sum(obs_spot - real_splot)^2** to account for negative nums, close to 0 = accurate!

QUIZZES

1. **Causal claims** = the student did well on the quiz because they attended class → The student did not attend class and did poorly
2. **Random assignment** is useful to balance out all possible **confounding variables** between groups
3. Experiment if the researcher interferes with the test subjects, possible to draw **causal conclusions** from this form of study!
4. **Blinding** can eliminate experimental biases that could arise from a participant or group participating in the experiment
5. **Least squares** is the measured vertical distance to the best fit line
6. **Explanatory var** is stat signif. @ 5% level = h-test regression null: $\mathbf{B_i=0}$, unlikely to observe estimated slope if predictors true slope=0
7. **Learn null distribution**/estimate slope by lm() (caluclate p-vals using t-dist) OR repeatedly permute y var and calculate avg slope
8. Make sure to generalize dummy variables! And pay attention to extrapolation vs interpolation
9. **P-value** = calculate diff in props and see where it falls on graph (like .008 → p=.01 cuz some points fall there)