3.1 Alternate Variable Types and Interactions

Dr. Bean - Stat 5100

1 Why Interactions?

Example (HO 3.1.1): Y = `cycles`, X_1 = `charge_rate`, X_2 = `temperature`

All models we have discussed in this class assume that the effects of the explanatory variables are **additive**.

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$$

In other words, the effect of each explanatory variable can be considered **separate** from all other explanatory variables.

What if the **real** effect of X_1 on Y actually depends on X_2 as well?

What would it mean for the effect of `charge_rate` on `cycles` to depend on `temperature`?

- We "know": higher `charge_rate` → lower `cycles`, and
  higher `temperature` → higher `cycles`
- But maybe: higher `charge_rate` **and** higher `temperature` → **much** higher `cycles`
- "**much**" higher here: significantly more than could be attributed to the sum of the effects of `charge_rate` and `temperature` only (often called **synergy**)

Whenever the effect of an explanatory variable (X_k) on the response (Y) *depends on* the values of other explantory variables, you have an **interaction effect**.

Metaphor: The bachelorette - the relationship of each potential suitor (X_k) with the bachelorette (Y) is partially depends upon the other potential suitors.

How is an interaction effect different from multicollinearity?

Muticollinearity only has to do with relationships among the X_k and has nothing to do with Y. Interactions have everything to do with the relationship between the X_k's and Y.

Define an interaction term as a new predictor variable:

$$\begin{aligned} X_3 &= X_1 \cdot X_2 \\ Y_i &= \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i \\ &= \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \varepsilon_i \end{aligned}$$

Note: sometimes β_{12} instead of β_3

1.1 How to interpret interaction terms?

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

- if X_1 increases by 1 unit, then we expect an average change of $\beta_1 + \beta_3 X_2$ in Y
  - the effect of X_1 on Y depends on X_2
  - if the interaction term is non-zero, we *cannot* separate the effect of X_1 from the effect of X_2. We must consider them jointly (unless X_1 or $X_2 = 0$).

1.2 Best Practices

- Don't check all possible interactions. Only include an interaction term in a linear model if its output is interpretable.

- Include all lower-ordered terms that compose an interaction term, regardless of the significance of the lower interaction term.
  - Prevents forcing lower ordered coefficients to zero.
  - Maintains a flexible response surface and facilitates interpretation.

1.3 Things to remember about interactions:

- Unless the X_k are standardized, the interaction term $X_3 = X_1 * X_2$ is likely to be collinear with either X_1 or X_2.
  - This will ruin inference for the "lower order" terms, but not the interaction term.

- Two-way interactions are often interpretable, but higher order interactions (ex: $X_4 = X_1 * X_2 * X_3$) become difficult to interpret.
  - A plot of residuals from a non-interaction model against the potential interaction term may help to determine inclusion (if a trend is apparent).

- If your problem is best solved by including multiple, high-ordered, interaction terms, then regression trees/random forests is likely a better approach (more in Module 4).

1.4 Polynomial Predictors

- Up to this point, we have limited ourselves to modeling variables that share a linear relationship.

- If a variable X_k shares a quadratic, or higher-order (often called "curvilinear") relationship with Y, then that means that the effect of X_k on Y *depends upon itself* (i.e. interacts with itself).

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \beta_4 X_1^2 + \beta_5 X_2^2 + \varepsilon$$

- Handle higher-ordered terms the same way we handle other interaction terms:
  - include lower-order terms
  - standardize to reduce multicollinearity

  - coefficient interpretations important: – if X_1 increases by 1 unit (and X_2 held constant), then we expect an average change in Y of $\beta_1 + \beta_3 X_2 + 2\beta_4 X_1$

For those who have taken calculus, you may see a relationship between one unit increase in X_k with the $\frac{\partial Y}{\partial X_k}$.

2 Alternate Variable Types

Up to this point we have only focused on **quantitative variables**:

- Values are represented as numbers where number *order* and *magnitude* matters.
- Quantiative variables can be either:
  - Continuous: can take on any value (theoretically infinite number of decimal places) within a range.
  - Discrete: can only take on a discrete (countable) set of values.

We now wish to also consider **qualitative variables**

- Cannot be measured/ordered on a numerical scale.
- SAS can't recognize words/letters in a regression model, and it will treat a set of numbered factored levels as quantitative (and thus order the levels).
- Because of this, we use **dummy/indicator variables** to include qualitative predictors in a model.

2.1 Dummy Variables

Consider the following student demographic variables (qualitative in bold): (age, height, **Utah residency status**, weight, **major college**)

Use an indicator variable to include residency status in model

$$X = I_{\text{resident}} = \begin{cases} 1 & \text{if student is resident of Utah} \\ 0 & \text{otherwise} \end{cases}$$

Things get a little more complicated for major college as we have to create multiple dummy variables to represent a single categorical variable:

$$\begin{aligned} X_1 &= I_{\text{College of Science}} = \begin{cases} 1 & \text{if student's major is within the college of science} \\ 0 & \text{otherwise} \end{cases} \\ X_2 &= I_{\text{College of Engineering}} \\ &\vdots \\ X_7 &= I_{\text{School of Business}} \end{aligned}$$

If there are eight colleges in the University, why would I only have seven dummy variables?

Values of 0 for all seven indicator variables means the person is a member of the eighth college. This college would be referred to as the base class on which all things are compared.

3 Example (See HO 3.1.1)

Y = months, X_1 = size, X_2 = type of firm

Note that $X_2 = I_{[\text{firm = stock}]} = \begin{cases} 1 & \text{if firm = stock} \\ 0 & \text{otherwise} \end{cases}$

Model with only qualitative predictor:

$$Y = \beta_0 + \beta_2 X_2 + \varepsilon$$

- equivalent to a two-sample t-test
- special case of one-way ANOVA model (`proc glm`, STAT 5200)

$$\begin{aligned} Y_{i,j} &= \mu_i + \epsilon_{i,j}, \qquad i = 1,2; j = 1, \ldots, n_i \\ &= \mu + \alpha_i + \epsilon_{i,j}, \qquad \sum_{i=1}^{2} \alpha_i = 0 \\ \epsilon_{i,j} &\ iid\ N(0, \sigma^2) \end{aligned}$$

Model with both qualitative and quantitative predictor:

- Additive

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

- Interaction

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

Note how the additive and interaction models differ:
(in the size (X_1) vs. months (Y) relationship for each firm type)

- Additive:
  - stock ($X_2 = 1$): $Y = (\beta_0 + \beta_2) + \beta_1 X_1 + \varepsilon$
  - mutual ($X_2 = 0$): $Y = \beta_0 + \beta_1 X_1 + \varepsilon$
- Interaction

  - stock ($X_2 = 1$): $Y = (\beta_0 + \beta_2) + (\beta_1 + \beta_3)X_1 + \varepsilon$
  - mutual ($X_2 = 0$): $Y = \beta_0 + \beta_1 X_1 + \varepsilon$

Note that the additive model results in *two parallel lines*, where the difference between stock and mutual firms are separated by a constant distance β_2. Whereas in the interaction model, both the slope *and* the intercept are different.

3.1 Note on interactions between qualitative predictors.

- possibly very interesting
- numerically much easier in [two-way] ANOVA setting (`proc glm`, STAT 5200), as ANOVA doesn't require the use of dummy variables.

3.2: Variable Selection

Dr. Bean - Stat 5100

1 Why Variable Selection

- Up until now, we have focused on trying to make predictions/inference using all the potential explanatory variables we have available to us.
- We now wish to consider several candidate models, ultimately making a judgment as to which model is "best."
    - Selection is more than an art than it is a science: no "right" decisions, several *wrong* decisions, several "reasonables."
    - This is an iterative process, that makes it difficult to know when we are "done" (see Figure 1 on last page).
- One element of the model building process involves **selecting a subset** of potential explanatory variables for use in the final model.
    - Follows the Ockham's razor principle: *entia non sunt multiplicanda praeter necessitatem*

"Entities should not be multiplied without necessity" (i.e. all else equal: simpler answers are better).

2 Methods of Variable Selection

How to pick the "best" subset of variables?

- Whenever possible, remove variables based on **context**, which comes with **expertise**.
- Automatic Methods:
    - **All possible regressions:** Consider all possible combinations of predictor variables, select the "best" model according to some measurement criteria.
    - **Stepwise methods:** Take a structured approach that takes a (semi) intelligent search through a subset of all possible models.
    - **Penalized regression:** more in Module 4.

2.1 All Possible Regressions

Consider all subsets of predictor variables $X_1, \ldots, X_{p-1}$.

- Number of subsets of size $p-1 = \binom{P-1}{p-1} = \frac{(P-1)!}{(p-1)!(P-p)!}$.
- Number of subsets of all possible sizes: $\sum_{p=1}^{P} \binom{P-1}{p-1} = 2^{P-1}$.

2.1.1 Measures of "goodness"

- R-square - but which model will always have the highest R^2?

$$R_p^2 = 1 - \frac{SS_{Error,p}}{SS_{Total}}$$

- Adjusted R-square - balances against # of predictors

$$R_{a,p}^2 = 1 - \frac{n-1}{n-p} \cdot \frac{SS_{Error,p}}{SS_{Total}}$$

  As p increases, $R_{a,p}^2$ first increases, then decreases

- Mallow's C_p - for a certain subset of $p-1$ predictors:

$$C_p = \frac{SS_{Error} \text{ from model with } p-1 \text{ predictors}}{MSE \text{ from model with } P-1 \text{ predictors}} + 2p - n$$

  When a subset of $p-1$ predictors gives unbiased $\hat{Y}$'s, $E[C_p] \approx p$.
  – so look for model with smallest p such that $C_p \approx p$,
  i.e., want $C_p \approx$ # predictors + 1.

- Akaike's information criteria & Schwarz's Bayesian criterion
  – both penalize larger numbers of predictors (want small):

$$\begin{aligned} AIC_p &= n \log SS_{Error,p} - n \log n + 2p \\ SBC_p &= n \log SS_{Error,p} - n \log n + p \log n \end{aligned}$$

- Prediction sum of squares – based on leave-one-out philosophy ($\hat{Y}_{i(i)}$)

$$PRESS_p = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_{i(i)} \right)^2$$

  – look for models with small $PRESS_p$

2.2 Stepwise Selection

Stepwise methods:

- automatically select a model based on some criterion (convenient)
- less satisfactory, do not "guarantee" the "right" model
- best used as "confirmatory" approaches
- three main: backward (okay), forward (worst), stepwise (hybrid)

Backward Elimination – basic algorithm

1. Fit model with all $P-1$ predictors
   (a) Compare each predictor's individual P-value to some threshold (`slstay`; default in SAS is 0.10)

   (b) If any predictor's P-value > `slstay`, drop predictor with largest P-value

2. Repeat with $P-2$ predictors

3. Continue until all predictors remaining have P-values below `slstay`

Forward Selection – basic algorithm

1. Find predictor with highest correlation with response
   (a) Regress response on this predictor
   (b) Leave predictor in model if P-value is below some threshold (`slentry`; default in SAS is 0.50)

2. Given the previously entered predictor, find the predictor with the highest partial correlation with response
   (a) Add this predictor to the model
   (b) Leave in model if P-value is below `slentry`

3. Continue until no more predictors warrant inclusion
   (P-value of "next" predictor above threshold)

Big problem here: best 2-variable model does not necessarily contain best 1-variable model (first step(s) can throw everything off)
Stepwise Selection – basic algorithm:

1. Take a "forward" step: add "best" predictor with P-value below `slentry` (default 0.15)

2. Take a "backward" step: evaluate all predictors in model and drop the variable with the highest P-value above `slstay` (default 0.15)

3. Iterate "forward" and "backward" steps until model stays the same

Note: in all these automatic stepwise procedures (backward, forward, stepwise), the `slentry` and `slstay` thresholds are deceptive. After the first step (really a hypothesis test), they are not significance levels (α), but "conditional" significance levels, which are harder to interpret.

2.3 Remember this...

- In order to have reliable results, we need $n >> P$ (often 6*10 times larger).
- Each described technique measures how well your models fit the data you already have, which might not translate to new data (in production).

We get a sense of how our models perform on new data by:

- Splitting our data into "training" and "test" sets.
- Fit each model using only the training data, then use the model to predict on the test data.
- Calculate the mean square prediction error:

$$MSPR = \frac{\sum_{i=1}^{n^*}\left(Y_i - \hat{Y}_i\right)^2}{n^*}$$

FIGURE 9.1
Strategy for Building a Regression Model.

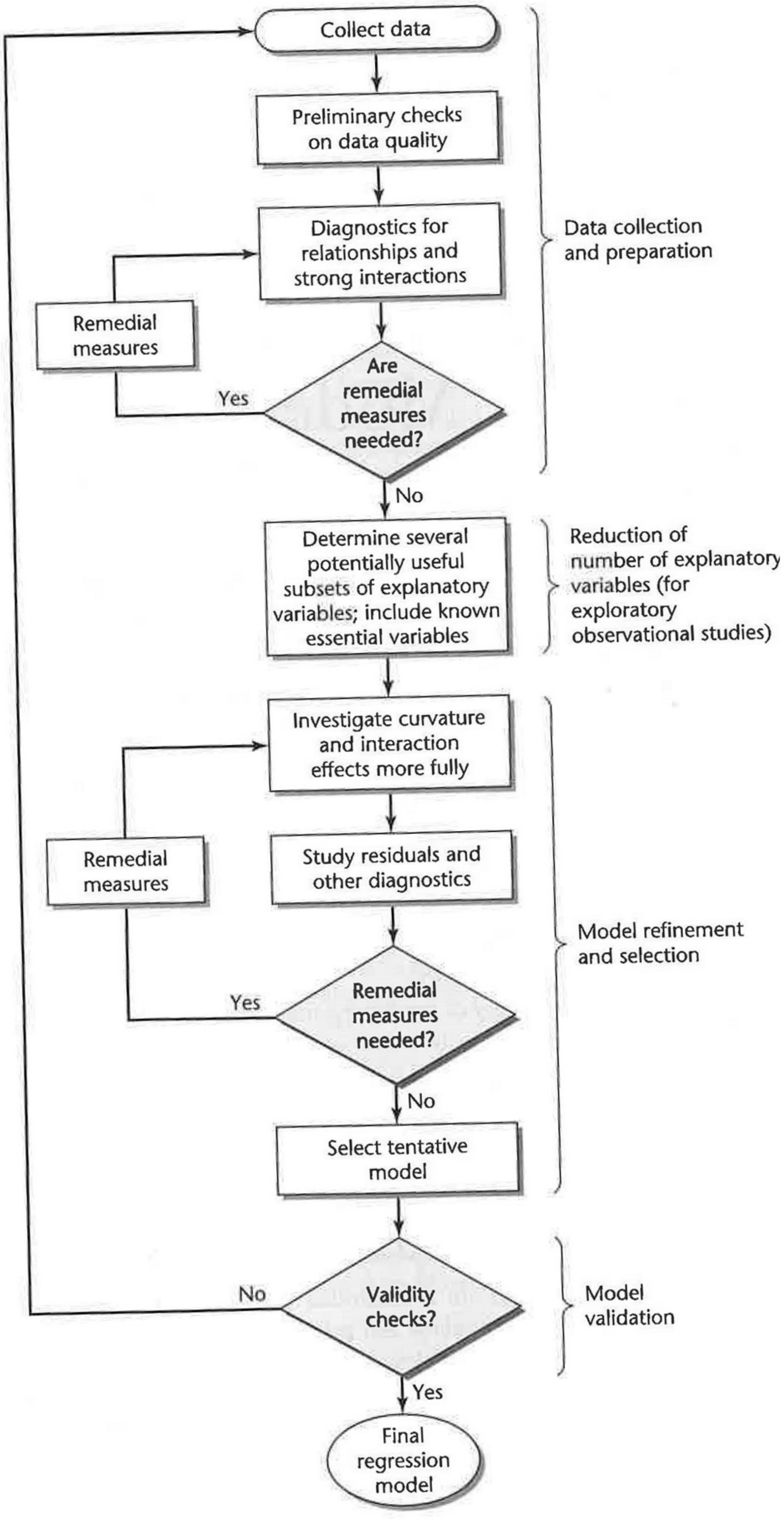
Collect data
Preliminary checks on data quality
Diagnostics for relationships and strong interactions
Remedial measures
Yes
Are remedial measures needed?
No
Data collection and preparation
Determine several potentially useful subsets of explanatory variables; include known essential variables
Reduction of number of explanatory variables (for exploratory observational studies)
Investigate curvature and interaction effects more fully
Remedial measures
Study residuals and other diagnostics
Model refinement and selection
Yes
Remedial measures needed?
No
Select tentative model
No
Validity checks?
Model validation
Yes
Final regression model

Figure 1: General model for multiple regression model selection (taken from Kutner et. al. (2004)).

3.3: Influential Observations and Outliers

Dr. Bean - Stat 5100

1 Why Care About Influential Observations/Outliers?

When we specify a model form of

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_{p-1} X_{p-1} + \varepsilon$$

we assume that all observations in the data are generated from the same source (i.e. the theoretical line).

If we have observations that are **not** from the same source as the rest, OLS regression will try to **force** the model to fit the data, perhaps compromising the estimated coefficients and or inference.

Two things to watch for (not mutually exclusive):

- **Outliers** - observations with values of Y that are not well-explained by the model.
- **Influential Points** - observations that unduly influence the estimated coefficients b_k or predicted values $\hat{Y}$.

2 Ways to detect outliers or influential points

- (Primary) Scatterplots of X_k vs Y
- Other Diagnostics for Influential Observations
  - Hat matrix diagonals
  - DFBETAS
  - DFFITS
  - Cooks Distance
- Other Diagnostics for Outliers
  - Residuals
  - Studentized Residuals
  - Studentized Deleted Residuals

2.1 Hat Matrix Diagonals

Recall the linear algebra representation of the OLS regression model:

$$Y = X\beta + \varepsilon \qquad b = (X'X)^{-1}X'Y$$

$$\hat{Y} = Xb = X(X'X)^{-1}X'Y = HY$$

In other words, the predicted values of Y are simply linear combinations of the actual values of Y where each observation "weight" is determined by the X matrix.
Let $h_{i,l}$ be the element in row i and column l of H
- sometimes called "leverage" (influence of obs. i on its fitted value)

Since $\hat{Y} = HY$, then $\hat{Y}_i = \sum_{l=1}^{n} h_{i,l}Y_l$

(Individual) What would a "larger" diagonal element $h_{i,i}$ mean?

It means that the value of Y_i is more influential in its own prediction $(\hat{Y}_i)$. We care about this because if the influence of a particular point is large enough, then the model is likely fitting that particular point at the sacrifice of the rest of the data.

We usually consider a point to be influential if:

- rule of thumb: $h_{i,i} > \frac{2p}{n}$ or $h_{i,i} > \frac{3p}{n}$
- can plot $h_{i,i}$ against observation number, with reference lines at $2p/n$ (SAS default) and/or $3p/n$

Another graphical diagnostic with $h_{i,i}$:

- leverage plots/partial regression/added variable plots); for X_1:
  1. Regress X_1 on $X_2, \ldots, X_{p-1}$ and obtain residuals $e_{X_1|X_2,\ldots,X_{p-1}}$
  2. Regress Y on $X_2, \ldots, X_{p-1}$ and obtain residuals $e_{Y|X_2,\ldots,X_{p-1}}$
  3. Plot $e_{Y|X_2,\ldots,X_{p-1}}$ vs. $e_{X_1|X_2,\ldots,X_{p-1}}$, and add regression line
     - slope will be b_1 from multiple regression model
     - Helps to visualize the marginal effect of adding X_1 in the model after already including all other X variables.
     - Influential points fall significantly farther away from the line than other points.
- (possible) modification here: point-size in leverage plot proportional to corresponding $h_{i,i}$ NOT shown in the SAS output provided in HO 3.3.1.
  - then this is called a proportional leverage plot
  - influential observations will be the points with big "bubbles" that appear to "pull" the regression line in their direction

2.2 DFBETAS

Provide a measure of how **different** ("DF") an estimate of β_k would be if we removed one observation from the data.

$$\begin{aligned}
b_k &= \text{estimate of } \beta_k \text{ using full data} \\
b_{k(i)} &= \text{estimate of } \beta_k \text{ when observation } i \text{ is ignored} \\
MSE_{(i)} &= \text{Mean SS for error when observation } i \text{ is ignored} \\
C_{kk} &= k^{th} \text{ diagonal element of } (X'X)^{-1} \\
DFBETAS_{k(i)} &= \frac{b_k - b_{k(i)}}{\sqrt{MSE_{(i)} C_{kk}}}
\end{aligned}$$

Interpreting DFBETAS:

- DFBETAS$_{k(i)}$ positive: obs. i "pulls" b_k up
- DFBETAS$_{k(i)}$ negative: obs. i "pulls" b_k down

How "large" to declare observation i "influential" on b_k?

- *Rough* rule of thumb:

$$|DFBETAS_{k(i)}| > 1 \qquad \text{for } n \leq 30$$

$$|DFBETAS_{k(i)}| > 2/\sqrt{n} \quad \text{for } n > 30 \text{ (SAS)}$$

- Graphical diagnostics probably better for DFBETAS:
  - Histograms or boxplots for each k
  - Proportional leverage plot with "bubble" size prop. to DFBETAS$_{k(i)}$
  - Plot DFBETAS$_{k(i)}$ against obs. number for each k (Provided by SAS, unlike the others)

2.3 DFFITS

Similar to DFBETAS: how different would $\hat{Y}_i$ be
if observation i were not used to fit the model

$$DFFITS_i = \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\sqrt{MSE_{(i)} h_{i,i}}}$$

How large DFFITS to declare obs. i as influential on $\hat{Y}_i$?

- *Rough* rule of thumb:

$$|DFFITS_i| > 1 \qquad \text{for } n \leq 30$$

$$|DFFITS_i| > 2\sqrt{\tfrac{p}{n}} \quad \text{for } n > 30 \text{ (SAS)}$$

- Good graphical diagnostics for DFFITS:
  - Plot DFFITS vs. Observation Number

  - Plot Residuals vs. Predicted Values, with point sizes proportional to corresponding DFFITS_i

(DFBETAS_{ij} vs. DFFITS_i) vs. $h_{i,i}$

- somewhat related, so "conclusions" will quite often agree
- BUT: if two or more points exert "influence" together then the drop-one diagnostics (DFBETAS and DFFITS) may not detect them
  - these are leverage points - need to look at $h_{i,i}$

2.4 Cooks Distance

Kind of an overall measure of effect of obs. i on all of the $\hat{Y}_l$ values:

$$D_i = \frac{\sum_{j=1}^{n} \left(\hat{Y}_j - \hat{Y}_{j(i)}\right)^2}{p \cdot \text{MSE}}$$

Diagnostics:

- Numerical:
  - simple: compare D_i with $4/n$ (SAS)
  - more useful: compare D_i with the $F_{p,n-p}$ distribution (See 3.3.1 pg 8 for example of how to do this "by hand")
    - ∗ percentile 10-20: little influence
    - ∗ percentile 50+: major influence
- Graphical: plot D_i (or percentile from $F_{p,n-p}$) vs. observation number i

2.5 Residuals

$e_i = Y_i - \hat{Y}_i$

Sometimes a large $|e_i|$ indicates an outlier

- not well-explained by fitted model
- but how "large" it needs to be depends on the residuals:
  - Recall $\varepsilon \sim N(0, \sigma^2)$, so $e_i \sim N(0, \sigma^2(1 - h_{ii}))$
    – because $\hat{Y} = HY$ results in $e = Y - HY = (I - H)Y$
  - Could compare e_i with the normal critical values, but need to estimate variance (including σ^2) $\Rightarrow$ normal approx. not appropriate; need Student's t

2.6 Studentized Residuals

$$r_i \quad = \quad \frac{e_i}{\sqrt{MSE \cdot (1 - h_{ii})}} \qquad\qquad (MSE = \hat{\sigma}^2)$$

If ε_i iid $N(0, \sigma^2)$, then the r_i follow the t_{n-p} distribution; diagnostics:

- Numerical: compare $|r_i|$ with upper $\alpha/2$ critical value of t_{n-p}
- Graphical: plot $\hat{Y}_i$ vs. r_i, with ref. lines at upper $\alpha/2$ critical value of t_{n-p}

2.7 Studentized Deleted Residuals

If obs. i really is an outlier, then including it in the data will inflate MSE
- So consider dropping it and re-calculating the studentized residual:

$$e_i^* \quad = \quad \frac{e_i}{\sqrt{MSE_{(i)}(1 - h_{ii})}} \qquad\qquad \text{(Text uses } t_i \text{ instead of } e_i^*\text{)}$$

2.8 Other Diagnostics (similar to studentized residuals)

- plot $\hat{Y}_i$ vs. e_i^*
- compare to $|e_i^*|$ to some critical value of t_{n-p} (for each of $i = 1, \ldots, n$)

  BUT: α = probability of type I error (calling obs. i outlier when it's not)

  - actually want α to be probability of *at least one* type I error in all n tests
    – a family-wise error rate
  - many ways to adjust the critical value; here, we'll use Bonferroni correction:

    compare $|e_i^*|$ to upper $\alpha/(2n)$ critical value of t_{n-p}

3 Remedial Measures for Influential Observations or Outliers

1. Look for:
   - typos in data (more common than would like to think)
   - fundamental differences in observations
     - drop obs. if from a different "population"
   - very skewed distributions of predictors
     - remember that in general, there is no assumption regarding the distribution of X's
     - sometimes transforming X will reduce influence of obs. with extreme values
   - **Caution:** Important to not remove outlier points simply because they are outliers.
     - `https://rss.onlinelibrary.wiley.com/doi/full/10.1111/j.1740-9713.2017.01016.x`
2. Look at potential changes to model:

- will a transformation “bring in” the observations?
- should a curvilinear or other predictor be added?
  - look at leverage plot for the possible predictor
  - any trend suggests adding it to model

3. Could obtain estimates differently (instead of OLS, robust regression - more in Module 4):

   - LAD (least absolute deviation) regression
   - IRLS (iteratively reweighted least squares) regression

3.4: Model Validation

Dr. Bean - Stat 5100

1 Why Model Validation?

Recall that there are two, distinct, goals of linear modeling and we don't always care about both at the same time:

- Inference: Is there a significant, linear relationship between X_k and Y, after accounting for the effect of a set of other X variables?
  - Example: Do students who use the tutor center see a significant positive affect to their GPA after accounting for study time and demographics?
- Prediction: Given a set of variables that are *easy* to measure, can I predict a variable that is hard to measure?
  - Use car weight (easy to measure) to predict car safety (hard to measure).

For **prediction**, there are a lot of alternatives to linear regression for which measures such as AIC, SBC, $C(p)$, and even R^2 are not relevant.

We need an *objective* way to compare the effectiveness of models with incomparable forms.

Why is the data we are using to fit our models not a fair measure of model effectiveness?

We ultimately want a model that can predict well on new data. Complex models have incentive to overfit the current data at the sacrifice of good predictions on new data.

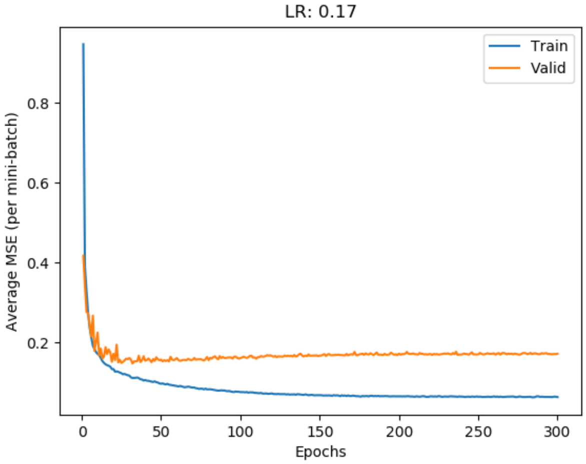

Figure 1: Comparison of accuracy on training and test sets for a neural network.

2 Validation Details

Terminology

- **Training set:** the data that is used to fit each model.
- **Test set:** data not used in model fitting that is used to compare model accuracy.
- **Validation set (optional):** If you perform too many comparisons with the test set, you run the risk of overfitting the test data. A validation set is a third set of data that is also withheld and only used to validate the best one or two models based on the test set.

Example in SAS: proc surveyselect can randomly assign observations to training and test sets.

3 Cross Validation

Whenever you have enough data, withholding a subset of the data prior to model building is ideal.

However, collecting new data can be very expensive such that creating a “test set” is not feasible. **Cross Validation:** is a method that tries to estimate test set error using training data.

The process:

- Randomly separate our data into k-groups (usually five or ten).
- Treat all but one of the groups as a training set, the remaining group as a test set.
- Fit a model using the training data, predict for the test data.
- Repeat the process, each time treating a different group as the test data until all observations have a prediction.

SAS does not have an easy method for performing custom cross validation. For this purpose, we will stick to validation accuracy from a test set in our projects.
In R, custom cross validation is fairly easy. However, working with validation accuracy from a test set is also perfectly okay so we will also opt for this method in R as well.

However, certain procedures use cross validation as a means of performing variable selection such as proc glmselect.

Cautions and Considerations:

- *Any* variable selection techniques or other forms of training must be included as part of the cross validation process. In other words, you can’t use all of the data to select variables, then act “blind” to that same data in the model validation step.
  - The consequence of such a move is that you will likely overestimate your model’s predictive capability.
  - Trying to embed variable selection into cross validation is extraordinarily difficult and not necessarily stable.
  - Check out this optional video for a more detailed explanation: `https://www.youtube.com/watch?v=r64tRyHFAJ8&list=PLOg0ngHtcqbPTlZzRHA2ocQZqB1D_qZ5V&index=23`
- The more groups you create, the more models you must fit, which can get computationally expensive.

- Too many groups makes it hard to estimate the true "test set" error.
  - Less groups, more bias, less variance in the test set error estimation. Try to select a number of cross validation groups that balance the bias and variance (usually five or ten groups).
- Check out chapter 5 in this book for more details on cross validation and other forms of model validation: `http://faculty.marshall.usc.edu/gareth-james/ISL/ISLR%20Seventh%20Printing.pdf`