STA302 METHODS OF DATA ANALYSIS 1

MODULE 10: MODEL VALIDATION

PROF. KATHERINE DAIGNAULT

MODULE 10 OUTLINE

1. Why do we want to validate a model?
2. How is validation done?
3. What happens if a model is not validated?

DESCRIPTIVE VS PREDICTIVE MODELS

- Purpose of model is important in how a preferred model is selected
  - helps to know if a model with more or fewer predictors is preferable
- Consider whether you want model to:
  - simply describe the relationship or historical versus current situation, plus estimate the effects directly
  - accurately predict a future event or observation
- Changes decisions made during analysis and how goodness is measured
  - e.g., choice of transformations, predictors, interpretability, etc.

Descriptive Models

- focus on explanation and understanding of population through
  - variables associated with response
  - estimated effect of variable on response
  - how past behaviours/trends might affect future

Predictive Models

- extend current trends into the future, rather than comprehensive understanding
- not used to
  - estimate effects on response
  - identify variables associated to response
- only provide snapshot of future values with error

DIFFERENT ANALYTICAL APPROACHES

Descriptive Models

- Interpretability of the model
  - easy to understand and describe the relationship
  - variables are reasonable, make sense, and easy to interpret
- Explain the variance
  - want a high amount of variation explained by model
  - tells us we have captured the predictors that affect the response well without too many extraneous ones
- Of course, always ensure model assumptions hold

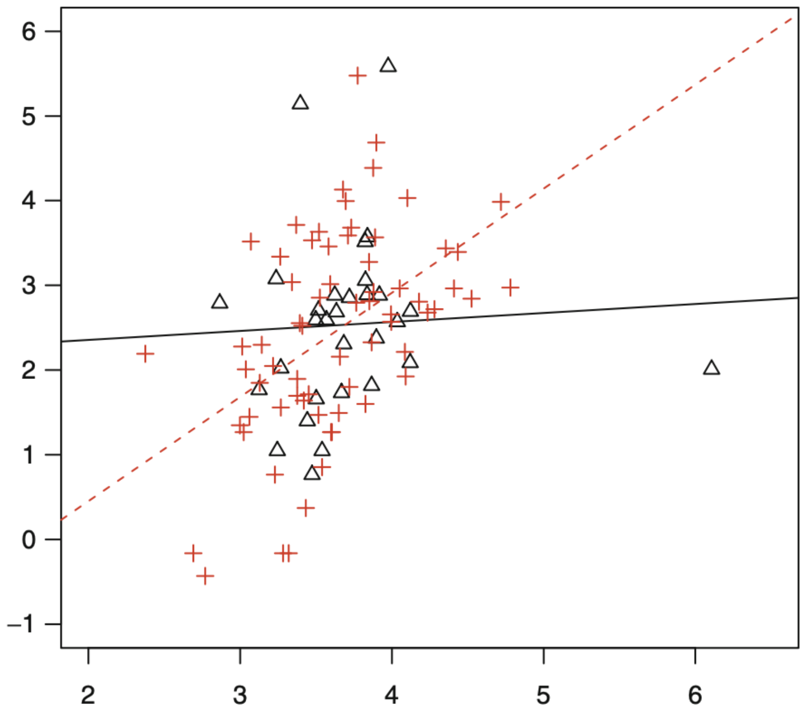
Predictive Models

- Explain the variance
  - want a high amount of variation explained by model
  - ensure as many predictors as reasonable included without overfitting the data
- Correct structure of variables
  - predictors/response are included in a format that increases accuracy of predictions
  - interpretation not prioritized so complicated transformations can be used
    - in so far as to satisfy model assumptions

OVERFITTING & VALIDATION

- Overfitting the data means the model only provides good predictions on the data used to build it
  - e.g., in figure using black line to predict on red dataset would yield inaccurate predictions
- Occurs when there are a large number of predictors in a model
  - many of which have no clear or significant relationship to the response
- While good prediction generally needs more predictors
  - want to avoid having too many that make model to specific to your dataset

From Sheather's "A Modern Approach To Regression with R", pg 249



USING AN INDEPENDENT DATASET

- Model Validation is then the process of checking how your model performs in an independent dataset
  - we can see if the model does overfit the data
- Need a dataset that is independent from information used to build model
  - has no common observations or overlap
- Sampling more data from population for this purpose would be ideal
  - but impractical or infeasible in most situations
- Mimic this by splitting current data into two entirely separate parts

- One part is the training dataset, used to build, assess, and select the final model
- Other is the test/validation dataset, used only to understand model's generalizability to new data
- Key to independent splitting of dataset:
  - decide on split proportion, ensuring both parts are sufficiently large (e.g., 50/50 is usually good choice)
    - any proportion can be used
  - divide/sample the data randomly to avoid sampling bias.
- Works as long as original data was a random and representative sample from population

MODULE 10 OUTLINE

1. Why do we want to validate a model?
2. How is validation done?
3. What happens if a model is not validated?

VALIDATION IN THE ANALYSIS PROCESS

Split data

- Training Set
  - Explore data and note interesting features
- Testing Set
  - Explore data and note interesting features

In training data

- Perform all model building
- Conduct all model assessments and diagnostics
- Compare models to arrive at a few preferred/best models

where everything we have learned occurs

In testing data

- Fit chosen model(s) on test data
  - no new model selection/building occurs
- Compare model components between test dataset and training dataset (i.e., validate)

WHAT TO LOOK FOR TO VALIDATE MODEL HEURISTICALLY

To say a model is validated means that all characteristics of the model look similar in both datasets

- Minimal differences (< 2 s.e.'s) in estimated coefficients:
  - implies a similar trend/relationship occurs in both datasets
- Same predictors are significantly linearly related:
  - more likely to be true significance rather than due to overfitting
    - seeing a reduction in number of significant predictors can indicate overfitting
- Have a similar R^2_{adj}, so similarly good explanation of variance
- No additional or worsening model violations
  - means transformations only helped in training data
- Similar numbers and types of problematic observations
  - model fit is influenced in a similar way in each dataset
  - helps explain why a model may not have been validated
- Similar amount of multicollinearity
  - indicates if test data exhibits higher correlation among predictors.
  - helps explain why significance may differ

EXAMPLE

```
> data <- read.table("prostateAlldata.txt", header=T)
> nrow(data)
[1] 97
>
> # split data using sample()
> s <- sample(1:nrow(data), 67, replace=F)
> train <- data[s ,]
> test <- data[-s,]
```

indices of all observations

to sample

Build model in each dataset:

```
> model_train <- lm(lpsa ~ lcavol + lweight + svi + lbph
+                   data = train)
> # now build same model in test set
> model_test <- lm(lpsa ~ lcavol + lweight + svi + lbph,
+                  data = test)
```

```
> vif(model_train)
  lcavol  lweight      svi     lbph
1.649669 1.379835 1.642061 1.324427
> vif(model_test)
  lcavol  lweight      svi     lbph
1.201774 1.232421 1.199374 1.242755
```

```
Call:
lm(formula = lpsa ~ lcavol + lweight + svi + lbph, data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-1.8709 -0.3903 -0.0172  0.5676  1.4227

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.32592    0.77998  -0.418   0.6775
lcavol       0.50552    0.09256   5.461 8.85e-07 ***
lweight      0.53883    0.22071   2.441   0.0175 *
svi          0.67185    0.27323   2.459   0.0167 *
lbph         0.14001    0.07041   1.988   0.0512 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7275 on 62 degrees of freedom
Multiple R-squared:  0.6592,	Adjusted R-squared:  0.6372
F-statistic: 29.98 on 4 and 62 DF,  p-value: 6.911e-14
```

# of significant predictors ✗

different estimates, but within 2 SE ✓

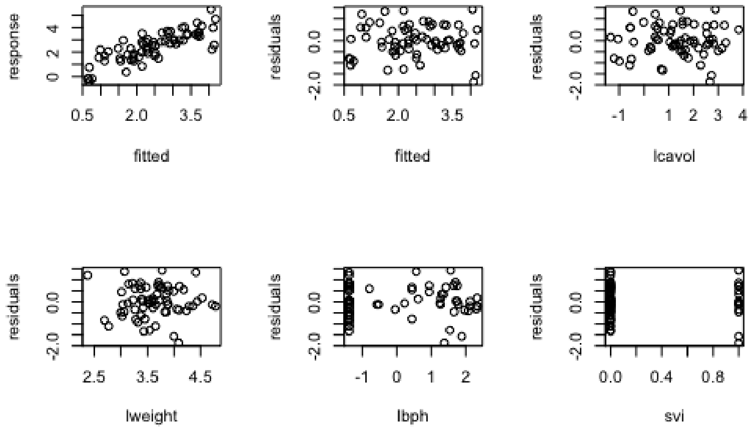
reasonably similar R^2_{adj} ✓

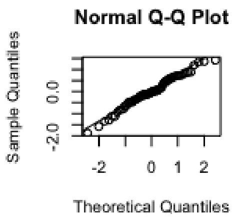
similar VIFs ✓

```
Call:
lm(formula = lpsa ~ lcavol + lweight + svi + lbph, data = test)

Residuals:
     Min       1Q   Median       3Q      Max
-1.08087 -0.40229 -0.05645  0.49322  1.01647

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.52957    0.93066   0.569   0.5744
lcavol       0.59555    0.12655   4.706 7.98e-05 ***
lweight      0.26215    0.24492   1.070   0.2947
svi          0.95051    0.32214   2.951   0.0068 **
lbph        -0.05337    0.09237  -0.578   0.5686
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6445 on 25 degrees of freedom
Multiple R-squared:  0.6703,	Adjusted R-squared:  0.6175
F-statistic:  12.7 on 4 and 25 DF,  p-value: 8.894e-06
```

EXAMPLE (CONTINUED)

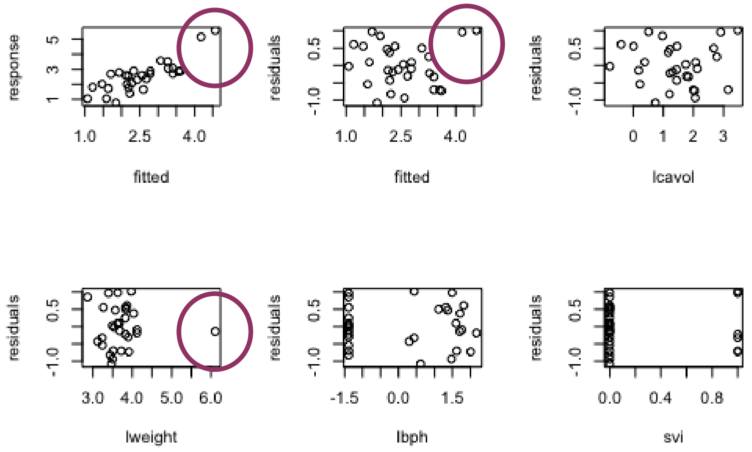
Model in Training Data

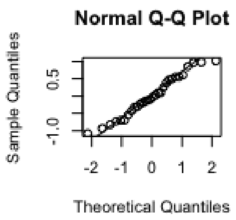




No new or worsening violations

Model in Test Data





	Training	Test
Leverage	3	1
Outlier	4	0
Influence (Cooks)	0	0
Influence (DFFITS)	5	2
Influence (DFBETA)	Between 3 and 8	Between 1 and 4

In training set: observation 45 problematic in many ways;
In test set: observation 9 problematic in many ways

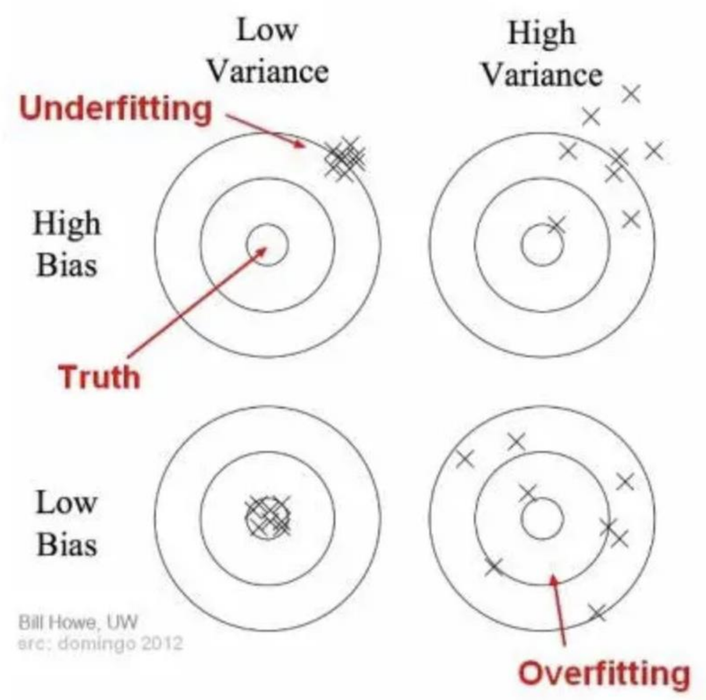
Differences in # of problematic observations

NUMERICAL CRITERION FOR VALIDATION

- Overfitting occurs when model only accurately predicts responses in the training dataset.
- Result of the bias-variance tradeoff
  - Mathematically written as $E(MSE) = Bias^2 + Var + \sigma^2$
  - models that have been trained too closely on one dataset will have low bias but high variance in the predictions on new data - overfitting
  - aiming for a model that captures the truth (low bias) but is not affected by the data used to make the prediction (low variance)
- Mean squared error used to measure how well model fits dataset

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

- Model validated if MSE is small when training model is used to make predictions on test data



https://towardsdatascience.com/understanding-the-bias-variance-tradeoff-165e6942b229

EXAMPLE (CONTINUED)

Recall: residual standard error is

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

Mean square error is

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

So $MSE = \frac{s^2(n-p-1)}{n}$

```
Call:
lm(formula = lpsa ~ lcavol + lweight + svi + lbph, data = train)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.8709 -0.3903 -0.0172  0.5676  1.4227 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.32592    0.77998  -0.418   0.6775    
lcavol       0.50552    0.09256   5.461 8.85e-07 ***
lweight      0.53883    0.22071   2.441   0.0175 *  
svi          0.67185    0.27323   2.459   0.0167 *  
lbph         0.14001    0.07041   1.988   0.0512 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7275 on 62 degrees of freedom
Multiple R-squared:  0.6592,	Adjusted R-squared:  0.6372 
F-statistic: 29.98 on 4 and 62 DF,  p-value: 6.911e-14
```

Training $MSE = \frac{0.7275^2(62)}{67} = 0.4898$

```
> # predictions in test set
> fitted_test <- predict(model_train, newdata = test)

> # compute MSE
> mean((test$lpsa - fitted_test)^2)
[1] 0.5115283
```

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

MSEs are relatively small and relatively similar so no evidence of overfitting

RESAMPLING METHODS

- So far, we've only considered one test dataset to validate a model
- Resampling methods split the original dataset multiple times, allowing you to see how the model performs in multiple test sets
  - allows you to average your MSE so less likely to fail to validate a model on a single "weird" test dataset
- Leave-one-out cross validation (LOOCV) is one such method
  - creates a training set of $n - 1$ observations and a test set of 1 observation
  - repeats for each observation resulting in n test sets
- Same idea as how we identify problematic points
  - remove a single observation to see how fitted value changes without this observation
  - now interested in the fitted value of this omitted observation
- For each of the n test sets:
  - build model using the other $n - 1$ observations
  - use this model to make a prediction in the test set
  - compute the squared error for this prediction (i.e. $(y_i - \hat{y}_i)^2$)
- Average these n prediction errors to get the MSE
- Always looking for small MSE to indicate a validated model

EXAMPLE (CONTINUED)

```
> # create storage for prediction errors
> se <- NULL
>
> # look through all n=nrow(data) observations
> for(i in 1:nrow(data)){
+   #create training set by removing observation i
+   LOOtrain <- data[-i, ]
+   #create testing set containing only observation i
+   LOOtest <- data[i, ]
+
+   #fit chosen model
+   model <- lm(lpsa ~ lcavol + lweight + svi + lbph,
+               data = LOOtrain)
+   #make prediction for observation i
+   fitted <- predict(model, newdata=LOOtest)
+
+   #compute prediction error
+   se_i <- (LOOtest$lpsa - fitted)^2
+   #store this with others
+   se <- c(se, se_i)
+ }
>
```

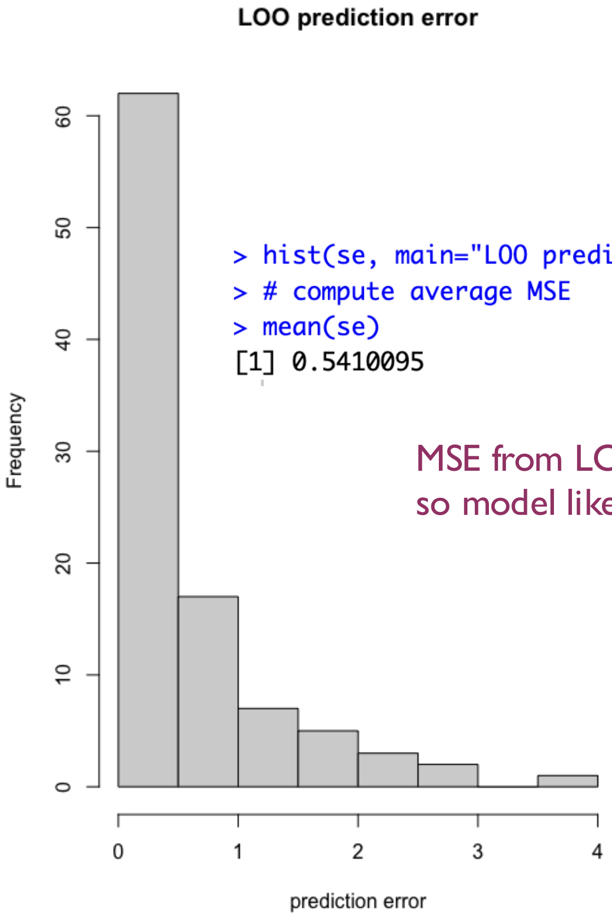
Do this process n times

Create training and test set

Fit model on training data

Get predicted value in test set

Compute $(y_i - \hat{y}_i)^2$




```
> hist(se, main="LOO prediction error", xlab="prediction error")
> # compute average MSE
> mean(se)
[1] 0.5410095
```

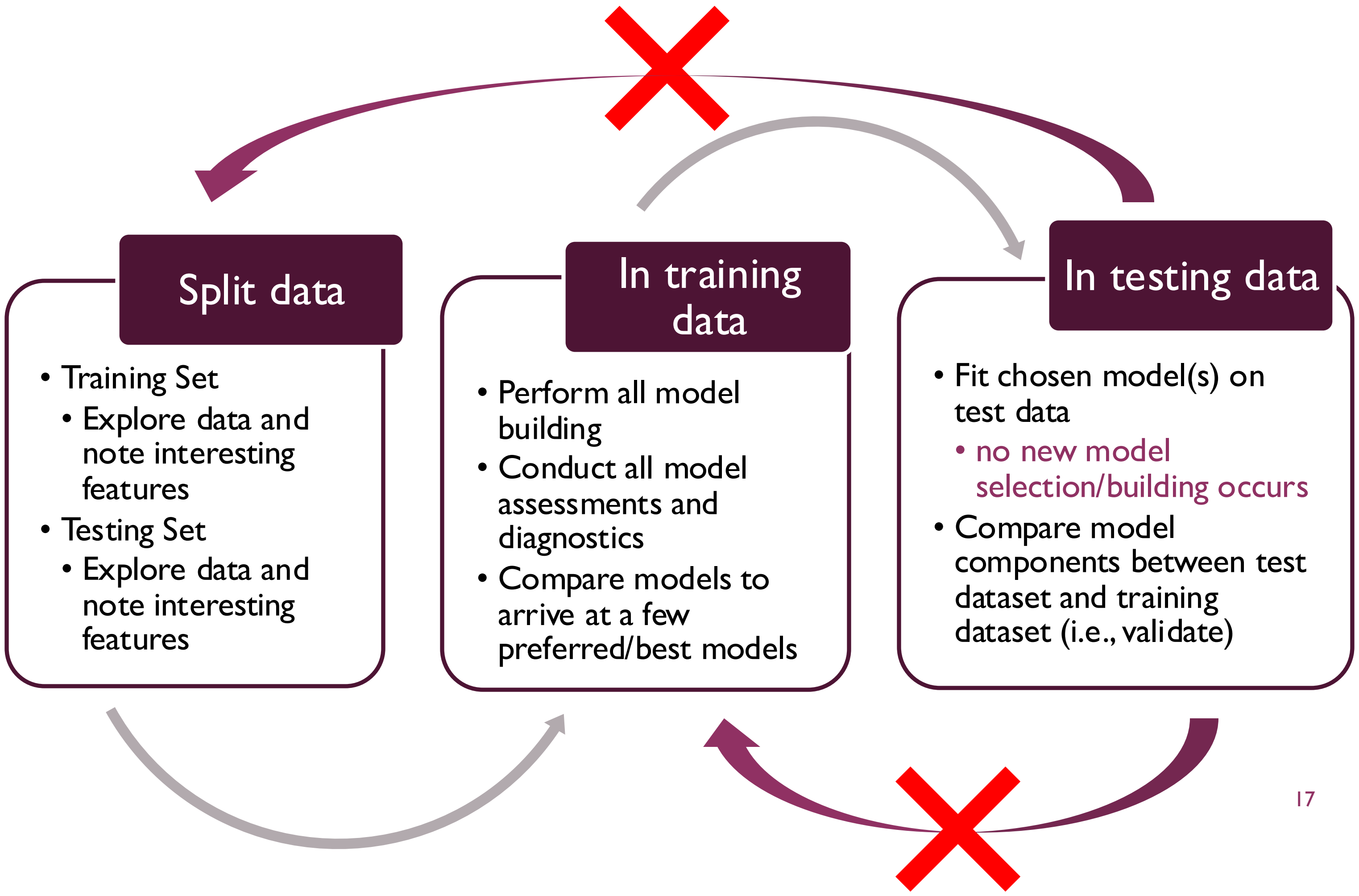
MSE from LOOCV still relatively small so model likely validated

MODULE 10 OUTLINE

1. Why do we want to validate a model?
2. How is validation done?
3. What happens if a model is not validated?

WHAT HAPPENS IF MY MODEL ISN'T VALIDATED?

- Don't panic - there are several reasons why this could happen
- It DOES NOT mean you did anything wrong or that you should change anything
  - DO NOT go back and select another test set
  - DO NOT go back and change your model or any other step in your analysis
  - The validation step is always the LAST STEP in your analysis
- Can help you pick a preferred model if many were validated



WHY WAS THE MODEL NOT VALIDATED?

- Only option is to understand why model could not be validated
  - need to investigate model or dataset characteristics that may have led to this
- Cannot go back and change anything about model based on what was seen in validation set
  - means validation set is used to build a model
  - but able to use information gained to better understand chosen model's performance
- Understand how lack of validation is a limitation of the model's performance.

1. Differences Between Training and Test Datasets

- Simplest reason why model was not validated
- Random sampling of original dataset may still provide datasets with very different characteristics
  - may have higher spread/variation in one dataset
  - may have different centers (e.g., means or medians)
  - may have different shapes or distributions
- Big discrepancies in any of these characteristics can result in a failure to validate a model.
- These can be detected early during the EDA stage.

WHY WAS THE MODEL NOT VALIDATED?

Presence of Highly Influential Observations

- Influential points yield drastic differences in fitted regression trend
- Depending on which dataset they are present, can tell you which model is "off"
  - if identified in model from training set, then likely training set model estimation is off
  - if identified in test set, then likely test set model is off
  - if in both datasets, likely both models are off and unclear which is closer to truth
- Noting the presence and impact of these points helps understand why validation failed

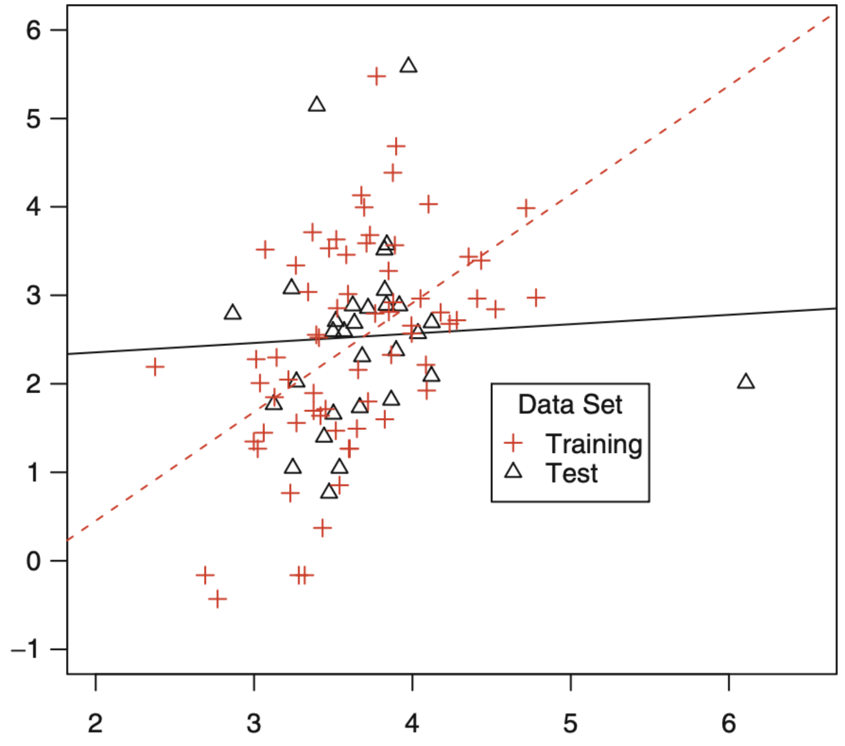
Complicated Transformations

- Use of complicated transformations may be warranted
  - e.g., severe violations of assumptions, predictive models, etc.
- Often contribute to lack of validation of a model
  - complicated transformation is more tailored to the training dataset and exact issue present there
- Simpler transformations more generalizable to different dataset
  - better luck validating model in external data

LACK OF VALIDATION IS A LIMITATION

- We cannot change anything about our model or validation procedure
  - only option is to address the problem as a limitation of our model
- To discuss lack of validation as a limitation of the model:
  - was it validated, and if not, in what way? (i.e., in what part of the comparison did it fail?)
  - why might this have happened? (i.e., what attributes of datasets or model might have led to this?)
  - what is the impact of a lack of model validation? (i.e., what does this mean for how your model might be used by others?)

From Sheather's "A Modern Approach To Regression with R", pg 249

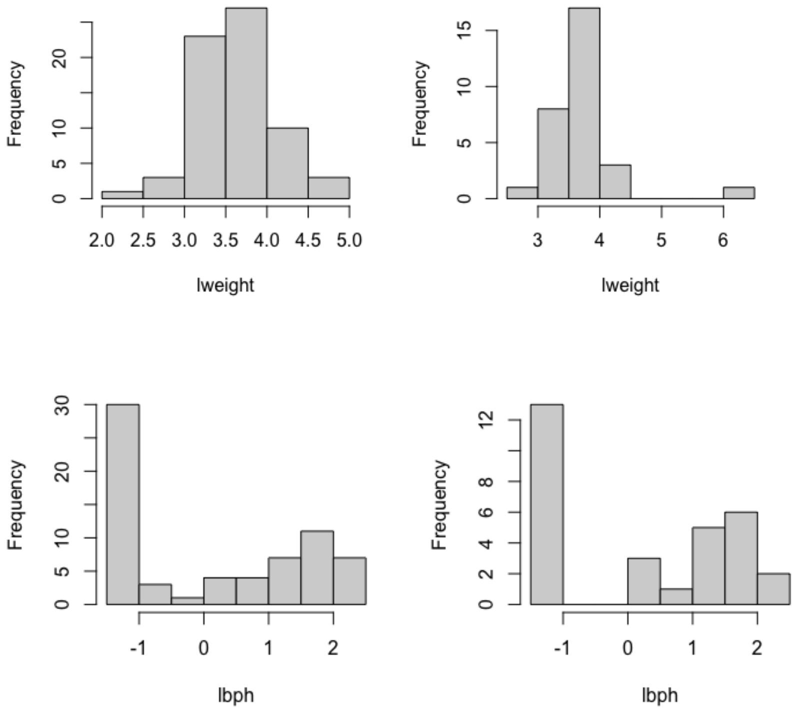


EXAMPLE (CONTINUED)

Differences in # of problematic observations ⟹ Different # of significant predictors

```
> describe(train[, c(2,3,5,6)])
        vars  n mean   sd median
lcavol     1 67 1.31 1.24   1.47
lweight    2 67 3.63 0.48   3.60
lbph       3 67 0.07 1.46  -0.05
svi        4 67 0.22 0.42   0.00
```

```
> describe(test[, c(2,3,5,6)])
        vars  n mean   sd median
lcavol     1 30 1.43 1.04   1.44
lweight    2 30 3.71 0.54   3.65
lbph       3 30 0.16 1.44   0.44
svi        4 30 0.20 0.41   0.00
```



Limitations:

- Differences arose in some estimated coefficients and in the number of problematic observations between original training and test sets.
- Differences occur in the distributions of some predictors, and there is one observation in each dataset that is problematic in many ways. These could explain the small differences in MSE and heuristic approach.
- LOOCV indicated model still predicts reasonably well in external datasets so likely validated

MODULE TAKE-AWAYS

1. For what reason do we try to validate our model?
2. Where does model validation occur in the flow of our analysis?
3. How do we know if our model is validated?
4. What do we do if we are not able to validate a model?