Fall 2025 AMACCS STAC67
Final Review Seminar

Issey Sone

December 9 2025

Table of Contents

Model Building

Q&A

Model Building

Model selection and Validation: Creating the model, how do we know our model is good?
Diagnostics: Anything affecting our model ie. influential points?
Remedial Measures: How can we "fix" or improve our model?

What to know? What to put in your cheat sheet?

- R_p^2 Criterion
- $R_{p,adj}^2$ Criterion
- Mallows' C_p Criterion
- Akaike Information Criterion (AIC)
- PRESS_p Criterion
- DFFITS
- Cook's Distance
- DFBETAS
- VIF
- Forward/Backward Selection
- Stepwise selection
- WLS/OLS

True/False Justify

1. Adding more parameters to our model will either always increase or keep the R_p^2 value the same.
2. In multiple regression model, adding more variables may decrease the adjusted R^2
3. To optimize our model, we want to choose the lowest value of the Mallows' C_p criterion.
4. Stepwise, Backward, and Forward selection based on AIC will always result in the same model in the end

Review

Goal for each Criterion when building the model:

1. R^2_{adj}: Maximize
2. Mallows' C_p criterion: $C_p \approx p'$
3. AIC: minimize
4. BIC: minimize
5. PRESS: minimize

Stepwise Regression output

```
> data <- read.csv("Fish[1].csv")
> null <- lm(Weight ~ 1, data = data)
> full <- lm(Weight ~ Length1 + Length2 + Length3 + Height + Width, data = data)
> nrow(data)
[1] 159
> step(null, scope=list(lower=null, upper=full), direction="forward")
Start:  AIC=1870.99
Weight ~ 1

          Df Sum of Sq      RSS    AIC
+ Length3  1  17251026  2996433 1569.2
+ Length2  1  17085990  3161469 1577.7
+ Length1  1  16978060  3269399 1583.1
+ Width    1  15912356  4335103 1627.9
+ Height   1  10623359  9624100 1754.7
<none>                 20247459 1871.0

... (After a series of steps)

Step:  AIC=1541.73
Weight ~ Length3 + Width

          Df Sum of Sq     RSS    AIC
<none>                 2489410 1541.7
+ Height   1   15996.7 2473414 1542.7
+ Length1  1    2902.9 2486507 1543.5
+ Length2  1    1550.4 2487860 1543.6

Call:
lm(formula = Weight ~ Length3 + Width, data = data)

Coefficients:
(Intercept)      Length3        Width
    -520.94        19.49        70.34
```

Model building Question 1:

Residual standard error: 12.66 on 156 degrees of freedom

1. Based on the output of the stepwise function, from the null model, what will the next model be in the stepwise process and what will the AIC value be?
2. Write out the final regression model from this stepwise process.
3. Calculate the R^2_{adj} for the final model, and comment about this value.
4. Calculate the Mallows' C_p criterion for the final model. Comment about this value.
5. From the final model, we check for influential measures and found that 40th observation had a large studentized deleted residual t_i value of 5.21312. Write the R code to get t critical value

Model question 2

Consider the following R output:

```
> fit <- lm(Y ~ X1 + X2 + X3 + X4, data = data)
> anova(fit)
Analysis of Variance Table

Response: Y
           Df Sum Sq Mean Sq F value    Pr(>F)
X1          1 0.55467 0.55467  62.6109 2.521e-06 ***
X2          1 0.11577 0.11577  13.0683 0.00314 **
X3          1 0.02961 0.02961   3.3421 0.09056 .
X4          1 0.03336 0.03336   3.7654 0.07432 .
Residuals  13 0.11517 0.00886
> with(aprout, round(cbind(which, cp), 3))
  (Intercept) X1 X2 X3 X4     cp
1           1  1  0  0  0 19.176
1           1  0  1  0  0 29.478
1           1  0  0  1  0 32.225
1           1  0  0  0  1 77.073
2           1  1  1  0  0  8.108
2           1  1  0  1  1  8.489
2           1  1  0  1  0  9.645
2           1  1  0  0  1 19.677
3           1  1  1  0  1  3.086
3           1  1  1  1  0  6.765
3           1  1  1  1  1  9.321
3           1  1  0  1  1 10.485
4           1  1  1  1  1  5.000
```

cont.

Consider two models

```
Model 1: Y ~ X1 + X2 + X3 + X4
Model 2: Y ~ X1 + X2 + X4
```

1. Compute AIC for model 1
2. Computer AIC for model 2
3. Which model is better?
4. Calculate R^2_{adj} for $Y \sim X_1$
5. Suppose we choose the model using X_1 and X_4. If correlation between X_1 and X_4 is 0.84 what is the VIF? should we be concerned?

Model building Question 2

Show that the first step in forward selection is the same as selecting the variable most highly correlated with the response

WLS Question 1

Suppose we don't have equal variance assumption within our model, and to remediate this we use WLS. Consider the following simple linear regression model:

$$Y_i = \beta_1 X_i + \epsilon_i, \quad i = 1, \ldots, n,$$

where the errors ϵ_i are independent with mean zero, and the variance of ϵ_i is $\text{Var}(\epsilon_i) = \sigma^2 X_i$.
Find a transformation:

$$Y_i^* = \beta_1 X_i^* + \epsilon_i^*,$$

such that the new error terms ϵ_i^* have constant variance

Ridge Regression Question 1

Consider the linear regression model $Y = X\beta + \epsilon$ with $\epsilon \sim N(0, \sigma^2 I)$. This model (without intercept) is fitted to data using the ridge regression estimator $\hat{\beta}(\lambda) = \arg\min \|Y - X\beta\|_2^2 + \lambda\|\beta\|_2^2$ with $\lambda > 0$. The data is:

$$X' = \begin{pmatrix} -1 & 1 & 1 & -1 \end{pmatrix} \text{ and } Y' = \begin{pmatrix} -1.5 & 2.9 & -3.7 & 0.7 \end{pmatrix}$$

- Evaluate the MLE/OLS estimator of regression parameter for $\lambda = 0$
- Evaluate the ridge regression estimator for $\lambda = 1$
- Verify the ridge regression estimator $\hat{\beta}(\lambda)$ shrinks as λ increases. Hint: use results from part 1 and 2

Q&A

Any Questions?