STA302 METHODS OF DATA ANALYSIS I

FINAL EXAM REVIEW

PROF. KATHERINE DAIGNAULT

STA302 AT A GLANCE

Estimating the trend (M1 & M2)

- Least squares
- Interpretation
- Categorical predictors

Assumptions (M3 & M4)

- What are they?
- How do we check them?
- Interpreting residual plots
- Transform variables

Variable Selection (M5, M6, & M9)

- Hypothesis tests/CIs on coefficients
- Partial F test
- Likelihood criteria
- All possible subsets
- Automated selection

Diagnostics (M7, M8)

- Collinearity (multi)
- Leverage Points
- Outliers
- Influential Points

Model Goodness (M6, M7, M9, M10)

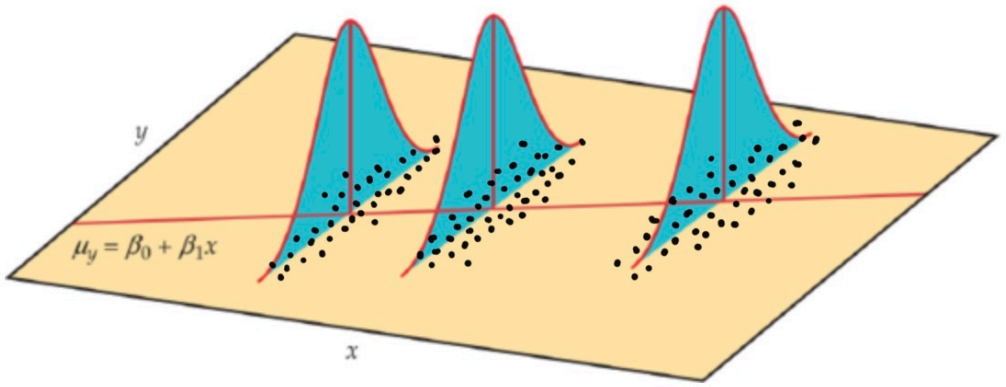
- ANOVA overall test
- R-squared (adjusted and not)
- Likelihood Criteria
- Validation

ESTIMATING THE TREND

ESTIMATION OF A TREND

Inference on a Linear Trend

- Population trend $Y = \beta_0 + \beta_1 X + \varepsilon$, a statistical relationship
  - Y is the *random* response variable
  - X is the *fixed* predictor variable
  - ε is the *random* error, given by $\varepsilon = Y - \beta_0 - \beta_1 X$
- Functional part: $E(Y|X) = \beta_0 + \beta_1 X$
- Sample data: pairs $(x_1, y_1), \ldots, (x_n, y_n)$
- Need to estimate β_0 and β_1 from the sample to estimate these means/the trend



LINEAR REGRESSION IN MATRIX FORM

Simple Linear Regression (SLR)

- Algebraic form: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ for $i = 1, \dots, n$
- Instead create matrices to store these components:

$$\boldsymbol{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \ \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \ \boldsymbol{X} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \ \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

$$n \times 1 \qquad 2 \times 1 \qquad n \times 2 \qquad n \times 1$$

- Each row of $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$ is equal to the algebraic form above through matrix multiplication:

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Multiple Linear Regression (MLR)

- Algebraic form: $y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i$ for $i = 1, \dots, n$
- Similar matrix components, just augmented with extra predictor information:

$$\boldsymbol{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \ \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \ \boldsymbol{X} = \begin{pmatrix} 1 & x_{11} \cdots & x_{1p} \\ \vdots & \vdots \ddots & \vdots \\ 1 & x_{n1} \cdots & x_{np} \end{pmatrix}, \ \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

$$n \times 1 \qquad (p+1) \times 1 \qquad n \times (p+1) \qquad n \times 1$$

- The matrix expression of the multiple linear regression trend is simply $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$

ORDINARY LEAST SQUARES STEPS (SLR)

Least Squares Procedure

1) Set up the estimating equation (RSS) for given model with parameters present
2) Take partial derivatives of your estimating equation with respect to each unknown parameter
3) Set each derivative to 0 to obtain score equation
4) Rearrange equations to solve for each unknown parameter.

$$RSS = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

$$\frac{\partial RSS}{\partial \beta_0} = -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\Longrightarrow \sum_{i=1}^{n} y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i$$

$$\Longrightarrow \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\frac{\partial RSS}{\partial \beta_1} = -2\sum_{i=1}^{n} x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\Longrightarrow \sum_{i=1}^{n} x_i y_i = \hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2$$

$$\Longrightarrow \sum_{i=1}^{n} x_i y_i = (\bar{y} - \hat{\beta}_1 \bar{x})n\bar{x} + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2$$

$$\Longrightarrow \hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

ORDINARY LEAST SQUARES STEPS (MLR)

Least Squares Procedure

1) Set up the estimating equation (RSS) for given model with parameters present
2) Take partial derivatives of your estimating equation with respect to each unknown parameter
3) Set each derivative to 0 to obtain score equation
4) Rearrange equations to solve for each unknown parameter.

$$RSS = \hat{e}^T\hat{e} = \left(Y - X\hat{\beta}\right)^T(Y - X\hat{\beta})$$

$$= (Y^T - \left(X\hat{\beta}\right)^T)(Y - X\hat{\beta}) \text{ by properties of transposes}$$

$$= Y^TY - Y^TX\hat{\beta} - \left(X\hat{\beta}\right)^TY + \left(X\hat{\beta}\right)^T(X\hat{\beta}) \text{ by matrix multiplication}$$

$$= Y^TY - Y^TX\hat{\beta} - Y^TX\hat{\beta} + \hat{\beta}^TX^TX\hat{\beta} \text{ by properties of transposes and scalars}$$

$$= Y^TY - 2Y^TX\hat{\beta} + \hat{\beta}^TX^TX\hat{\beta}$$

$$\frac{\partial RSS}{\partial \hat{\beta}} = -2X^TY + 2(X^TX)\hat{\beta} = \mathbf{0}$$

$$\Rightarrow (X^TX)\hat{\beta} = X^TY$$

$$\Rightarrow \hat{\beta} = (X^TX)^{-1}X^TY$$

*Assuming the inverse exists

Theorem 2.14a. Let $u = \mathbf{a}'\mathbf{x} = \mathbf{x}'\mathbf{a}$, where $\mathbf{a}' = (a_1, a_2, \ldots, a_p)$ is a vector of constants. Then

$$\frac{\partial u}{\partial \mathbf{x}} = \frac{\partial(\mathbf{a}'\mathbf{x})}{\partial \mathbf{x}} = \frac{\partial(\mathbf{x}'\mathbf{a})}{\partial \mathbf{x}} = \mathbf{a}. \qquad (2.112)$$

Theorem 2.14b. Let $u = \mathbf{x}'\mathbf{A}\mathbf{x}$, where $\mathbf{A}$ is a symmetric matrix of constants. Then

$$\frac{\partial u}{\partial \mathbf{x}} = \frac{\partial(\mathbf{x}'\mathbf{A}\mathbf{x})}{\partial \mathbf{x}} = 2\mathbf{A}\mathbf{x}. \qquad (2.113)$$

From Rencher & Schaalje's Linear Models in Statistics, page 56

CONDITIONAL NATURE OF MULTIPLE PREDICTOR MODEL

- We could fit three models to the data:
  - Simple model with X_1 only: $\hat{y}_i = 1.86 + 1.30x_{i1}$
  - Simple model with X_2 only: $\hat{y}_i = 0.86 + 0.78x_{i2}$
  - Two-predictor model we fit before: $\hat{y}_i = 5.375 + 3.012x_1 - 1.285x_2$
- Why is the relationship between Y and X_2 positive in one model and negative in another?
- Estimation and interpretation of coefficients in multiple predictor models conditions on all other predictors
  - i.e. consider only one fixed value of all other predictors when estimating or interpreting the coefficient of interest.

Figure from Rencher & Schaalje's Linear Models in Statistics, page 140

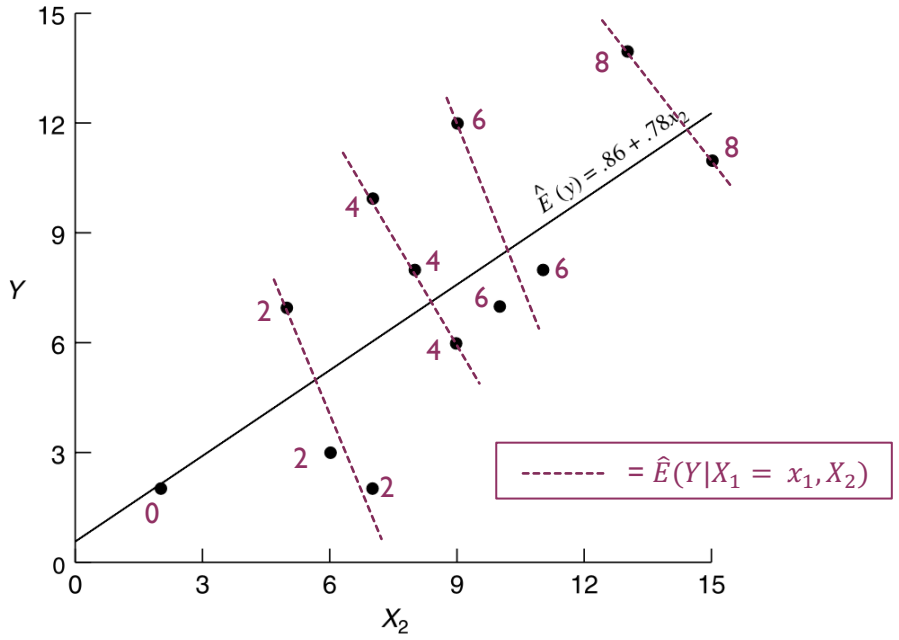
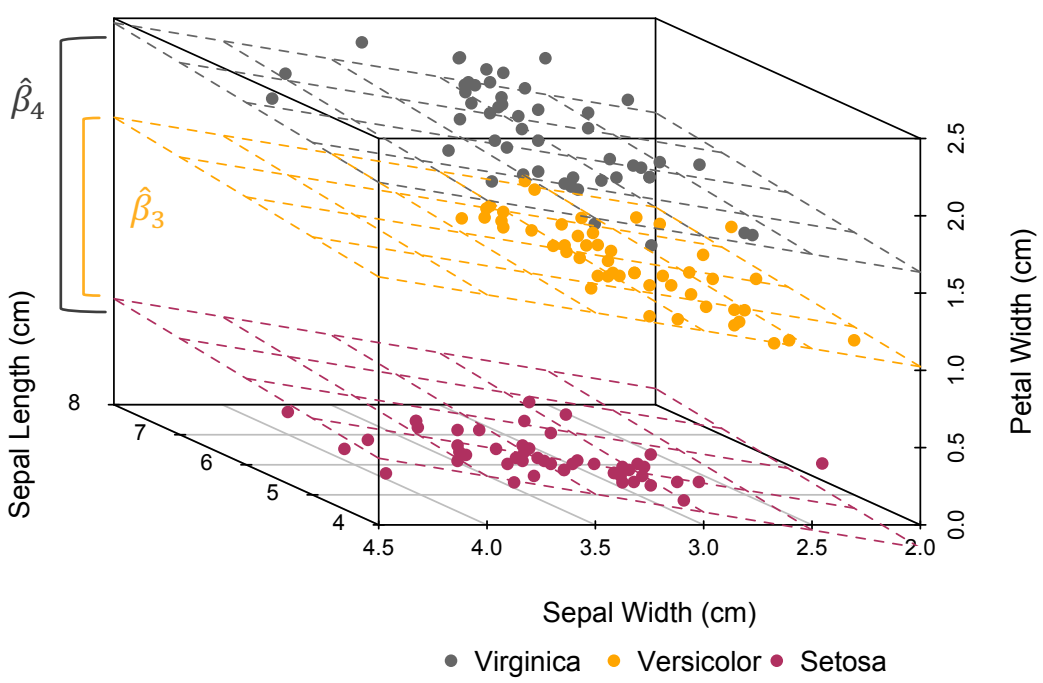


Figure 7.1 Regression of y on x_2 ignoring x_1.

INDICATOR VARIABLE INTERPRETATION

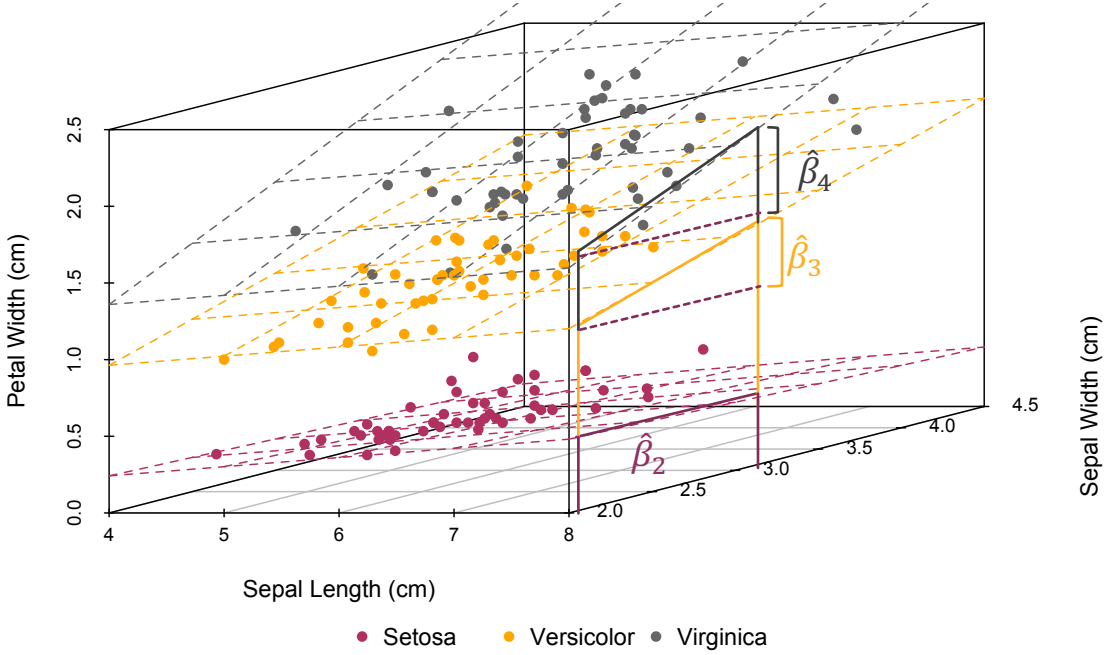
Main Effects only:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1\, Sepal\ Length + \hat{\beta}_2\, Sepal\ Width + \hat{\beta}_3 \mathbb{I}(Versicolor) + \hat{\beta}_4 \mathbb{I}(Virginica)$$



Interaction terms only:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1\, Sepal\ Length + \hat{\beta}_2\, Sepal\ Width + \hat{\beta}_3\, Sepal\ Width * \mathbb{I}(Versicolor) + \hat{\beta}_4\, Sepal\ Width * \mathbb{I}(Virginica)$$



CHECKING AND FIXING ASSUMPTIONS

LINEAR REGRESSION ASSUMPTIONS

1. Linearity of the Relationship (also known as Mean Zero Errors) assumption

$$E(\boldsymbol{\varepsilon}\,|\boldsymbol{X}) = \mathbf{0} \text{ or } E(\boldsymbol{Y}|\boldsymbol{X}) = \boldsymbol{X\beta} \text{ or } \boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$$

2. Uncorrelated Errors (sometimes referred to as Independence) assumption

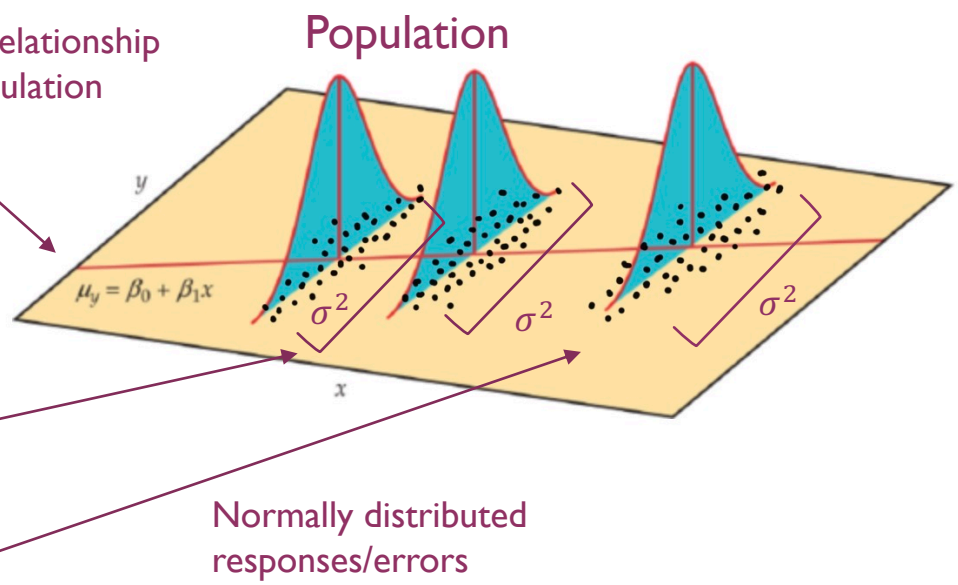
$$Cov(\varepsilon_i, \varepsilon_j) = 0 \text{ or } Cov(y_i, y_j) = 0$$

3. Constant Error Variance (also known as Homoskedasticity) assumption

$$Var(\boldsymbol{\varepsilon}|\boldsymbol{X}) = \sigma^2\boldsymbol{I} \text{ or } Var(\varepsilon_i|X) = Var(y_i|X) = \sigma^2$$

4. Normal Errors assumption

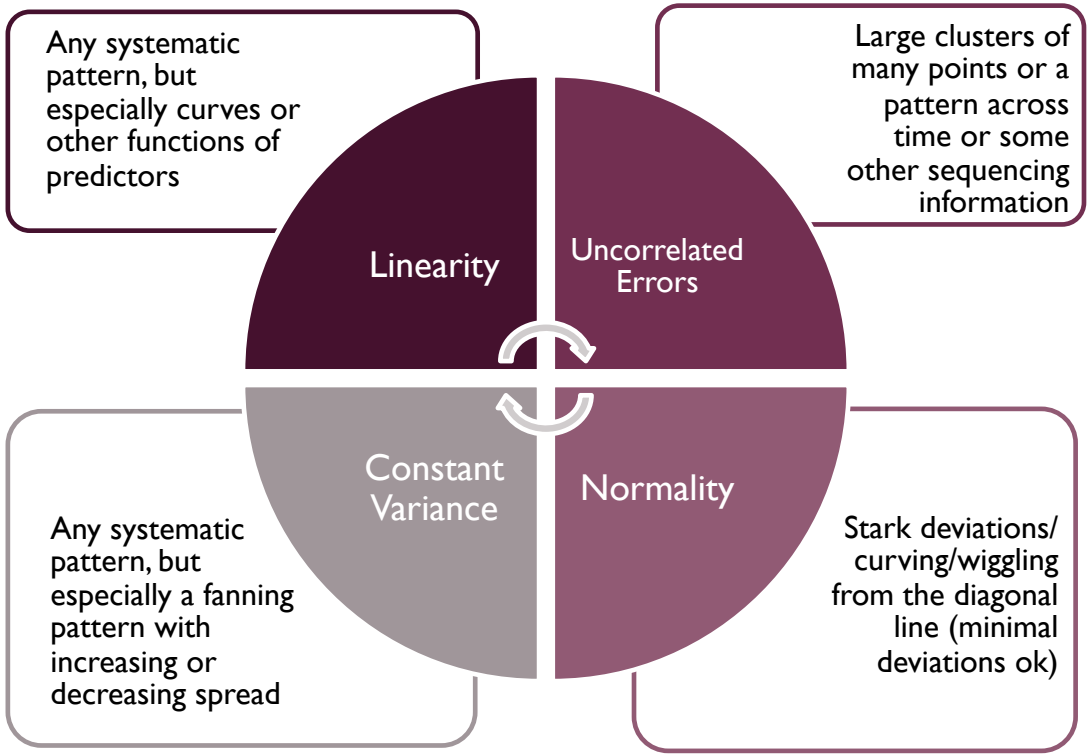
$$\boldsymbol{\varepsilon}|\boldsymbol{X} \sim N_n(\mathbf{0}, \sigma^2\boldsymbol{I}) \text{ or } \boldsymbol{Y}|\boldsymbol{X} \sim N_n(\boldsymbol{X\beta}, \sigma^2\boldsymbol{I}) \text{ or } \varepsilon_i \sim N(0, \sigma^2)$$



LOOK FOR PATTERNS IN RESIDUAL PLOTS

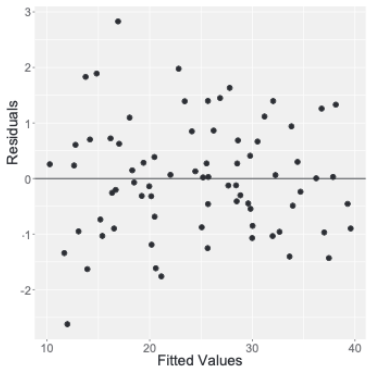
Residuals versus each predictor (scatterplot)	For linearity, uncorrelated errors, and constant variance
Residuals versus fitted values (scatterplot)	For linearity, uncorrelated errors, and constant variance
Normal Quantile-Quantile (QQ) plot	For Normality

Random bands of residuals indicates no violations

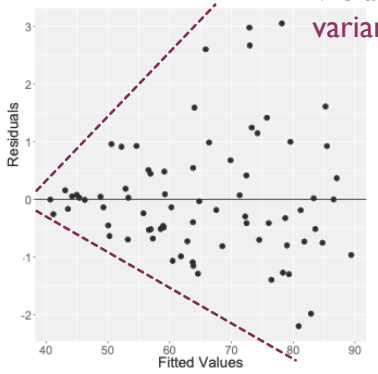


EXAMPLES OF DISTINCT PATTERNS

From Young's Handbook of Regression Methods, pg 55

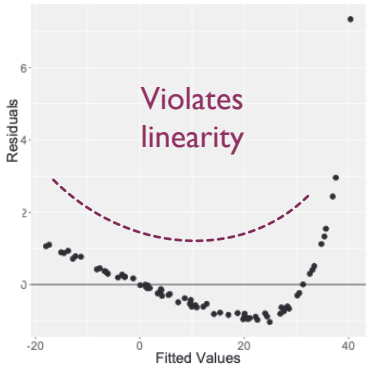


(a)



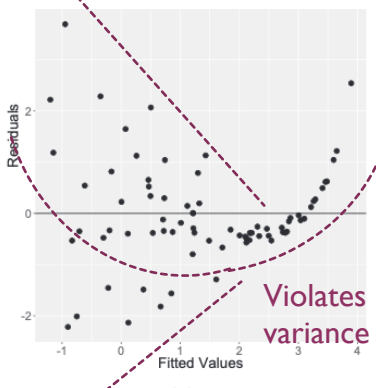
(b)

Violates constant variance



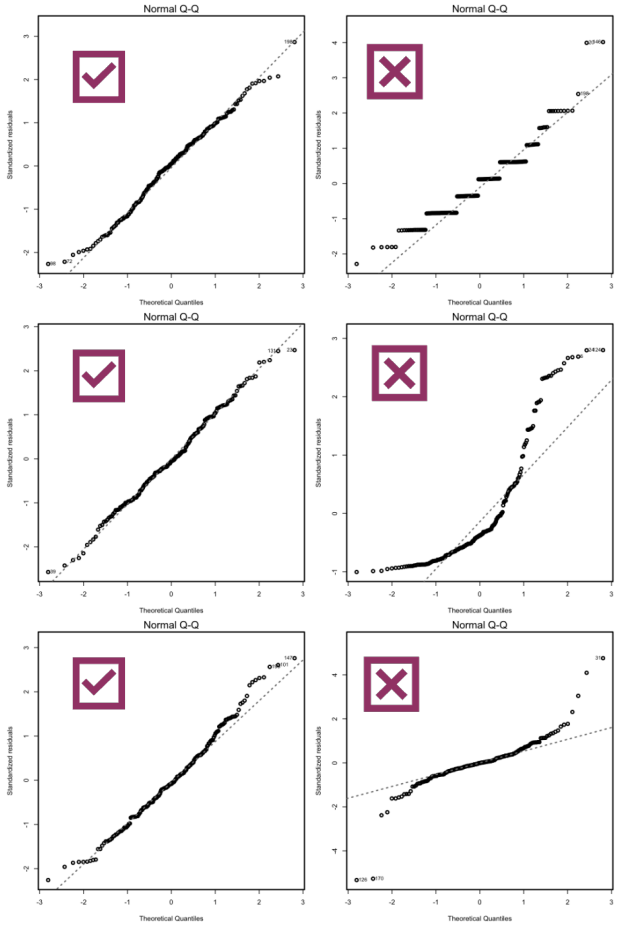
(c)

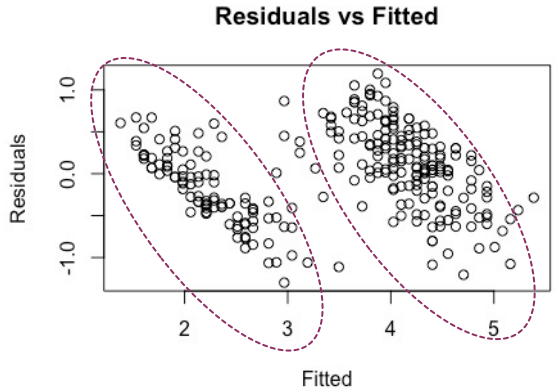
Violates linearity



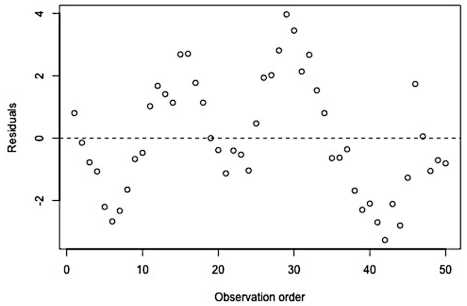
(d)

Violates constant variance and linearity





Violates uncorrelated errors



https://online.stat.psu.edu/stat501/book/export/html/915

García-Portugués, E. (2023). *Notes for Predictive Modeling*. Version 5.9.12.
ISBN 978-84-09-29679-8. Available at https://bookdown.org/egarpor/PM-UC3M/.

ADDITIONAL CONDITIONS FOR MLR

1. Conditional mean response condition: the mean responses are a single function of a linear combination involving $\boldsymbol{\beta}$

$$E(Y_i \,|\boldsymbol{X} = \boldsymbol{x_i}) = g(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})$$

   - $E(Y|X) = log(\beta_0 + \beta_1 x_i)$ satisfies this condition
   - $E(Y|X) = {}^{\beta_1 x_{i1}}/_{\beta_2 x_{i2}} = {}^{g_1(x_1)}/_{g_2(x_2)}$ violates it

2. Conditional mean predictor condition: the mean of each predictor is related to each other predictor in no more complicated way than linearly

$$E(X_i|X_j) = \alpha_0 + \alpha_1 X_j$$

   - Linear or no relationship satisfy condition; anything else violates

Conditions hold if:

1. Conditional mean response	2. Conditional mean predictors
Scatterplot of Response versus Fitted values *Look for random diagonal scatter or an easily identifiable non-linear trend*	All pairwise scatterplots of predictors *Look for lack of curves or other non-linear patterns*

If they fail: Patterns in plots cannot be used to identify a specific violation and can give misleading conclusions

HOW TO FIX VIOLATIONS?

Correlated errors	Non-constant variance	Linearity	Normality
• No fix • Use a different statistical methodology	• Variance stabilizing transformation • ONLY on Y • ln() or sqrt() often work • trial and error	• Box-Cox transformation • Pick something simple/identified from plots • On X and/or Y	• Box-Cox transformation • Pick something simple/identified from plots • Usually only on Y

WAYS TO SELECT VARIABLES/MODELS

PROPERTIES OF SAMPLING DISTRIBUTIONS OF $\widehat{\boldsymbol{\beta}}$

- Our assumptions say $\boldsymbol{Y}|\boldsymbol{X} \sim N(\boldsymbol{X\beta}, \sigma^2\boldsymbol{I})$ and our estimates are $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$
- Using linearity of Normal's, our sampling distribution is $\widehat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1})$
- Let's check that we get the same mean and covariance matrix if we derived them directly.

$$E(\widehat{\boldsymbol{\beta}}|\boldsymbol{X}) = E[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} \mid \boldsymbol{X}]$$
$$= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T E[\boldsymbol{Y}|\boldsymbol{X}]$$
$$= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T E[\boldsymbol{X\beta} + \boldsymbol{\varepsilon} \mid \boldsymbol{X}]$$
$$= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\{\boldsymbol{X\beta} + E[\boldsymbol{\varepsilon} \mid \boldsymbol{X}]\}$$
$$= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X\beta}$$
$$= \boldsymbol{\beta}$$

Linearity: $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$

Linearity: $E(\boldsymbol{\varepsilon}|\boldsymbol{X}) = \boldsymbol{0}$

Like $^c/_c = 1$

- The LS estimators of $\boldsymbol{\beta}$ are unbiased.
- For the covariance matrix:

$$Cov(\widehat{\boldsymbol{\beta}}|\boldsymbol{X}) = Cov((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}|\boldsymbol{X})$$
$$= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T Cov(\boldsymbol{Y}|\boldsymbol{X})\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}$$
$$= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T Cov(\boldsymbol{X\beta} + \boldsymbol{\varepsilon}|\boldsymbol{X})\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}$$
$$= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T Cov(\boldsymbol{\varepsilon}|\boldsymbol{X})\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}$$
$$= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\sigma^2\boldsymbol{I}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}$$
$$= \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}$$
$$= \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$$

Linearity: $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$

Constant variance & uncorrelated errors

Like $^c/_c = 1$

Theorem 3.6d. Let $\mathbf{z} = \mathbf{A}\mathbf{y}$ and $\mathbf{w} = \mathbf{B}\mathbf{y}$, where $\mathbf{A}$ is a $k \times p$ matrix of constants, $\mathbf{B}$ is an $m \times p$ matrix of constants, and $\mathbf{y}$ is a $p \times 1$ random vector with covariance matrix $\boldsymbol{\Sigma}$. Then

(i) $\text{cov}(\mathbf{z}) = \text{cov}(\mathbf{A}\mathbf{y}) = \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}'$, (3.44)

CONCLUDING INFERENCE ON COEFFICIENTS

$(1-\alpha)\%$ CI for β_j: $\hat{\beta}_j \pm t_{\frac{\alpha}{2},n-p-1} s\sqrt{(\boldsymbol{X}^T\boldsymbol{X})^{-1}_{(j+1,j+1)}}$

- Built using data and sampling distribution
- Window that is 2 $t_{\frac{\alpha}{2},n-p-1}$ standard errors wide, centered on estimate $\hat{\beta}_j$
  - Would see the true β_j in this window $(1-\alpha)\%$ of the time
- Interpretation of interval: $(1-\alpha)\%$ of all intervals computed using data repeatedly obtained from the same population would contain the true β_j.
  - Can also say it represents plausible values of β_j with $(1-\alpha)\%$ confidence.

Hypothesis Test of H_0: $\beta_j = \beta_j^0$: $t^* = \dfrac{\hat{\beta}_j - \beta_j^0}{s\sqrt{(\boldsymbol{X}^T\boldsymbol{X})^{-1}_{(j+1,j+1)}}}$

- Testing H_0: $\beta_j = 0$ versus H_A: $\beta_j \neq 0$ is most common
  - Tests the null that no linear relationship exists between X_j and Y (in the presence of other predictors).
- Conclude the test by comparing to sampling distribution:
  - If $|t^*| > t_{\frac{\alpha}{2},n-p-1}$, then reject the null
  - If $P(|T_{n-p-1}| \geq |t^*|) < \alpha$, then reject the null
  - Claim a significant linear relationship exists

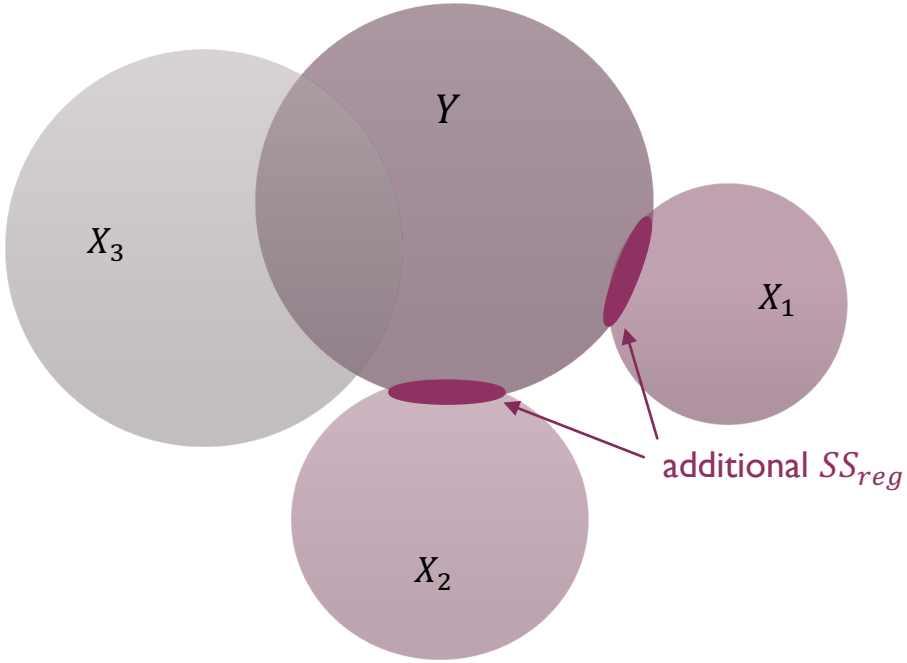
PARTIAL F TEST FOR SUBSETS OF MODELS

- Test is built assuming null hypothesis is true: $H_0: \boldsymbol{\beta}_2 = \mathbf{0}$ versus $H_a: \boldsymbol{\beta}_2 \neq \mathbf{0}$ where $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2)^T$
  - $\boldsymbol{\beta}_2$ is a vector of k coefficients that were removed from the full model to make the reduced model.
    - i.e. reduced model has the $p - k$ predictors matching coefficients in $\boldsymbol{\beta}_1$, so testing if ok to drop k predictors because their β's are 0
- We quantify the difference between the RSS in the two models as $RSS_{drop} = RSS_{reduced} - RSS_{full}$
  - A substantially better full model gives a bigger RSS_{drop}, while similar full and reduced models give a smaller RSS_{drop}
- Like the ANOVA test, we look at a ratio of mean sums of squares, comparing the mean SS difference $\left({}^{RSS_{drop}}/_{k}\right)$ to the model with the smallest mean squares residual (i.e. the full model):

$$F^* = \frac{{}^{RSS_{drop}}/_{k}}{{}^{RSS_{full}}/_{n-p-1}} \sim F(k, n-p-1)$$

CONCLUDING THE PARTIAL F TEST

- The null assumes it is true that the smaller model is better (MSR_{drop} is small relative to MSR_{full})
- If $F^* < F_{(1-\alpha),(k,n-p-1)}$, we fail to reject the null
  - The additional SS_{reg} from k predictors not enough to keep the predictors ← outcome we want if aiming for simpler/smaller models
  - Conclude there does not exist a significant linear relationship between Y and any of the k predictors.
- If $F^* > F_{(1-\alpha),(k,n-p-1)}$, or p-value $< \alpha$, we reject the null
  - Additional SS_{reg} from k predictors explains a lot of variation so we want to keep these predictors
  - Conclude there exists a significant linear relationship between Y and at least one of the k predictors.



COMPARING SETS OF MODELS

- The four main measures to compare models are:
  - Adjusted R^2: $R^2_{adj} = 1 - \frac{RSS/(n-p-1)}{SST/(n-1)}$
  - AIC: $AIC \propto n\, ln\left(\frac{RSS}{n}\right) + 2p$
  - Corrected AIC: $AIC_c \propto n\, ln\left(\frac{RSS}{n}\right) + 2p + \frac{2(p+2)(p+3)}{n-p-1}$
  - BIC: $BIC \propto n\, ln\left(\frac{RSS}{n}\right) + (p+2)\ln(n)$
- The RSS term drives the idea of goodness, while the rest are penalties for number of predictors.
  - we know that models with more predictors automatically have lower RSS even if not useful predictors
- Only a problem for comparing models with different number of predictors
  - if p is fixed between models, then only RSS changes so we'd prefer model with smallest RSS.
  - e.g., for all possible models with 3 predictors, the best one explains the most variation and so has smallest RSS
  - all four criteria would agree on the preferred model
- Criteria are then most helpful in comparing models with different p.
- Idea of comparing all models of the same size to each other is the root of our next tool.

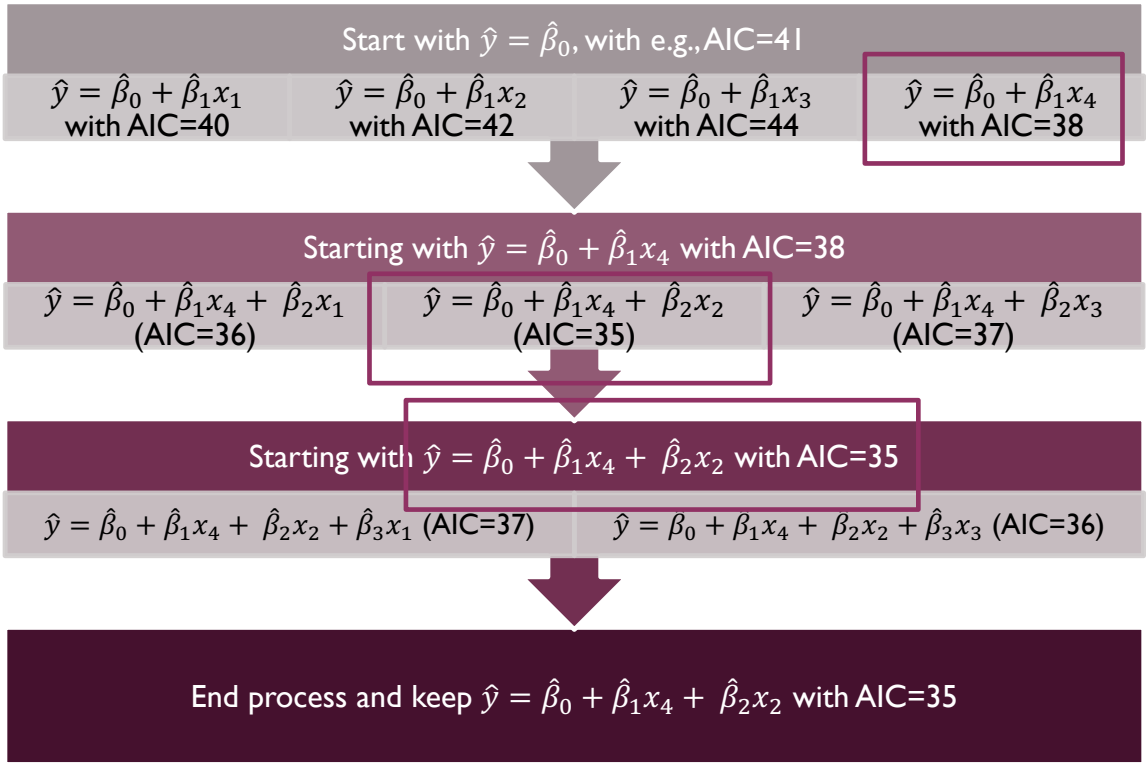
ALL POSSIBLE SUBSETS METHOD OF MODEL SELECTION

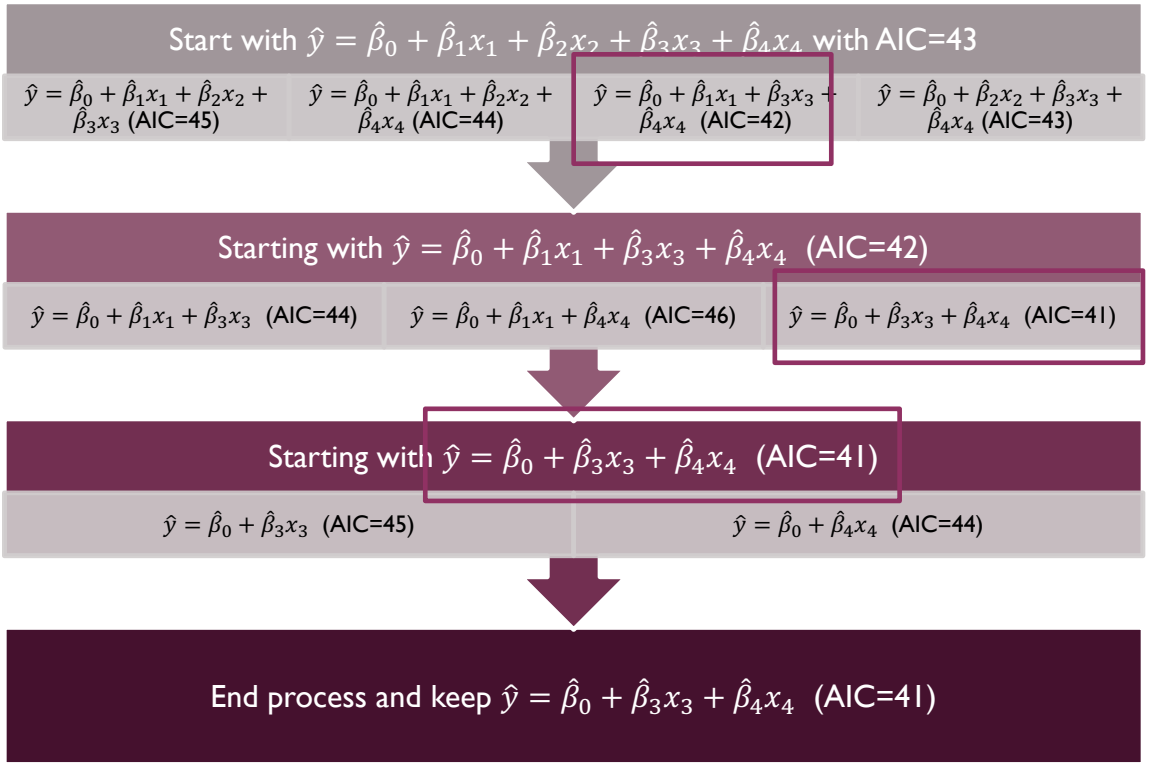
All possible subsets works in two steps:

1. Compare models of each size using adjusted R^2
2. Use all four numerical criteria to pick the best of the best $(R^2_{adj}, AIC, AIC_c, BIC)$

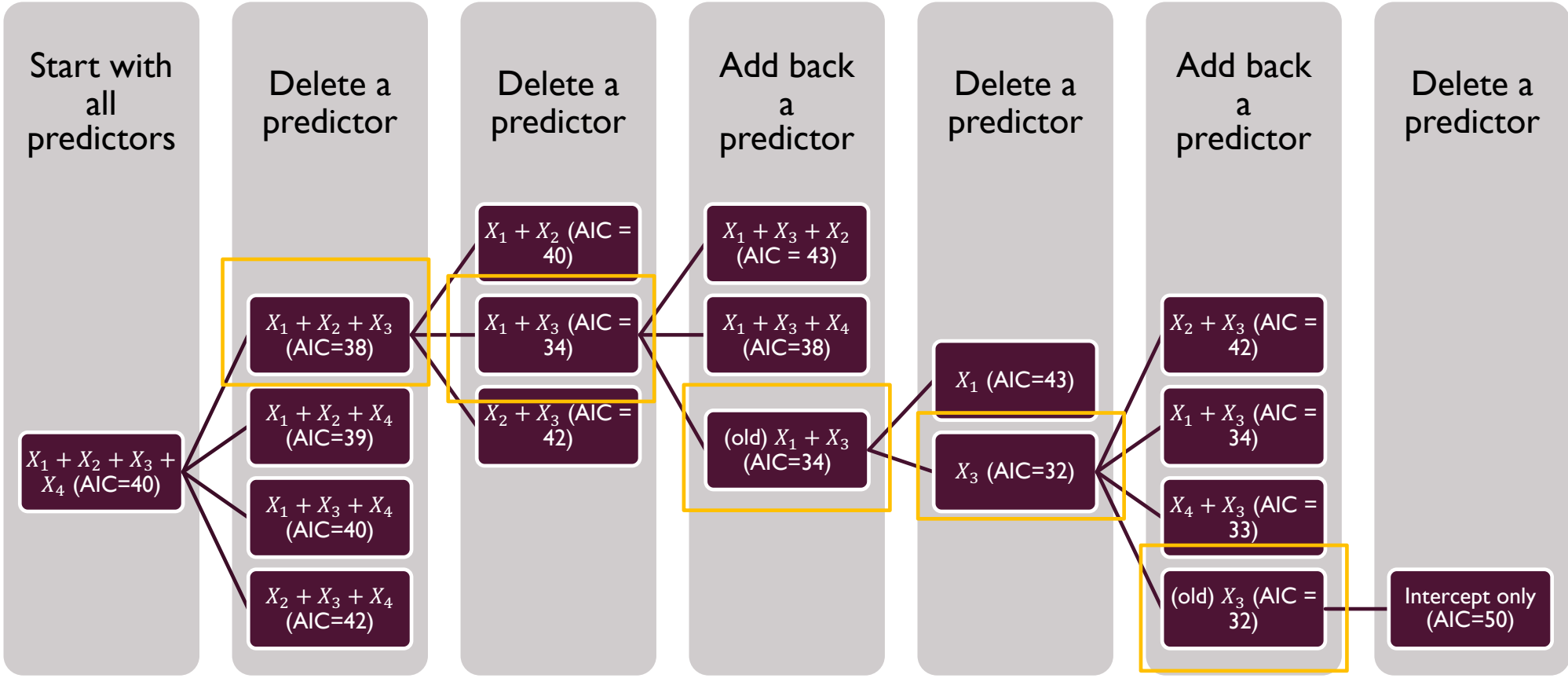
All one-predictor models	All two-predictor models	All three-predictor models	All four-predictor models
• X_1 has $R^2_{adj} = 0.84$	• X_1, X_2 has $R^2_{adj} = 0.87$	• X_1, X_2, X_3 has $R^2_{adj} = 0.87$	• X_1, X_2, X_3, X_4 has $R^2_{adj} = 0.9$
• X_2 has $R^2_{adj} = 0.8$	• X_1, X_3 has $R^2_{adj} = 0.88$	• X_1, X_2, X_4 has $R^2_{adj} = 0.88$	
• X_3 has $R^2_{adj} = 0.83$	• X_1, X_4 has $R^2_{adj} = 0.86$	• X_1, X_3, X_4 has $R^2_{adj} = 0.89$	
• X_4 has $R^2_{adj} = 0.75$	• X_2, X_3 has $R^2_{adj} = 0.89$	• X_2, X_3, X_4 has $R^2_{adj} = 0.85$	
	• X_2, X_4 has $R^2_{adj} = 0.81$		
	• X_3, X_4 has $R^2_{adj} = 0.84$		

FORWARD & BACKWARDS SELECTION PROCEDURES





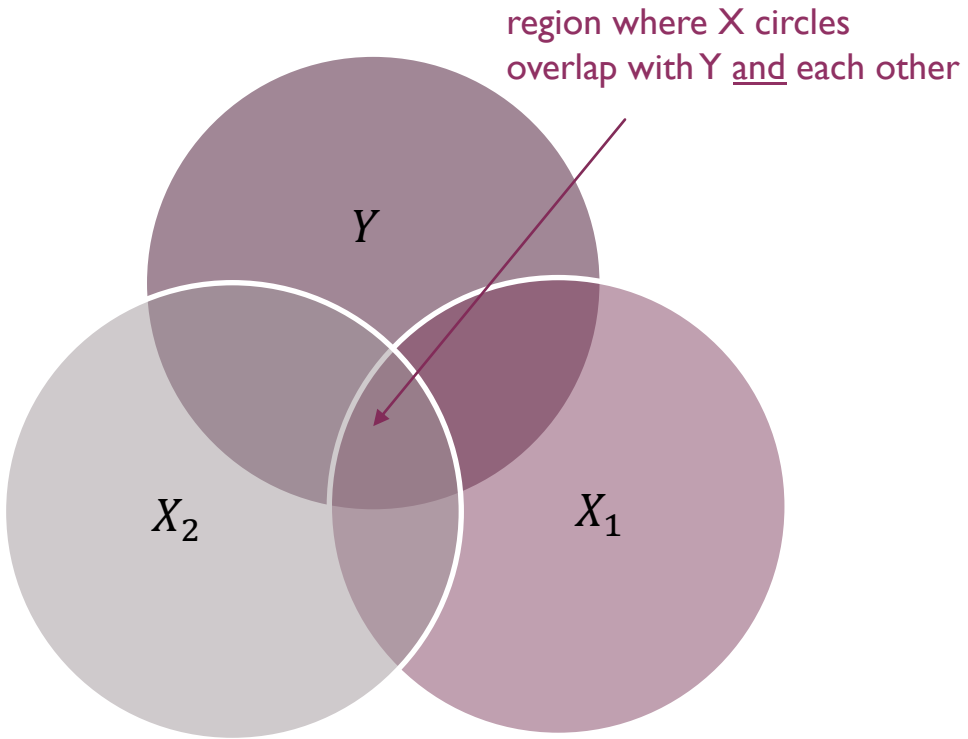
STEPWISE SELECTION PROCEDURE



DIAGNOSTICS ON MODELS

RANK OF DESIGN MATRIX & MULTICOLLINEARITY

- Correlation only tells you if any 2 predictors are related
- We want to know if possibly more than 2 predictors are related
  - this is called multicollinearity
- The strength of the impact of this relationship is due not only to how related the predictors are
  - Also how much those predictors explain Y to begin with (i.e., the overlap with Y)
- Creates an instability in how the predictors are used to estimate the mean responses
  - the model can't distinguish between how much variation is due only to X_1 versus only to X_2



VARIANCE INFLATION FACTOR (2+ PREDICTOR CASE)

- Case of more than two predictors requires a replacement for r_{12}
  - need a measure that also quantifies strength or goodness of linear relationship
  - saw that R^2 measures this as proportion of variance explained by linear relationship
- Here R^2 would be used to say that a model explains variation in a predictor, not a response
  - e.g. suppose we have 3 predictors in our model
  - using, e.g., X_1 as a response, can create a model of the form $X_1 = \beta_0 + \beta_1 x_2 + \beta_2 x_3 + \varepsilon$
  - R^2 of this model measures strength of linear relationship between X_1 and both X_2 and X_3
- R^2 will be used like r_{12} to say how the variance is inflated because a strong linear relationship exists between predictors
- A similar expression for the variance of any $\hat{\beta}_j$ is

$$Var(\hat{\beta}_j) = \frac{1}{1 - R_j^2} \times \frac{\sigma^2}{(n-1)s_{x_j}^2}, \qquad j = 1, \ldots, p$$

- The VIF for $\hat{\beta}_j$ is again the first term, incorporating the R^2 from a model using X_j as response
- Note: unlike the correlation that was squared, R_j^2 is not being squared
  - the square is part of the notation for the quantity

PROBLEMATIC POINTS

delete-one measures

Leverage

- "weird" X value(s)
- $h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2}$ for SLR
- hatvalues() in R for MLR
- $h_{ii} > 2\left({}^{(p+1)}/_{n}\right)$

Outliers

- "weird" Y values/too far off trend
- $r_i = \frac{\hat{e}_i}{s\sqrt{1-h_{ii}}}$
- $r_i \notin [-2,2]$

Influential on all fitted values

- actually changes Y vector
- $D_i = \frac{r_i^2}{(p+1)} \frac{h_{ii}}{(1-h_{ii})}$
- $D_i >$ median of $F(p+1, n-p-1)$

Influential on own fitted value

- changes one element in Y vector
- $DFFITS_i = \left(\frac{h_{ii}}{1-h_{ii}}\right)^{0.5} \frac{\hat{e}_i}{s_{(i)}\sqrt{1-h_{ii}}}$
- $|DFFITS_i| > 2\sqrt{{}^{(p+1)}/_{n}}$

Influential on at least one beta

- changes one estimated slope
- $DFBETAS_{j(i)} = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{\sqrt{s_{(i)}^2 (\boldsymbol{X}^T\boldsymbol{X})^{-1}_{j+1,j+1}}}$
- $\left|DFBETAS_{j(i)}\right| > {}^{2}/_{\sqrt{n}}$

classification may change if transformation applied to model

MEASURING GOODNESS OF A MODEL

ANOVA TEST OF OVERALL SIGNIFICANCE

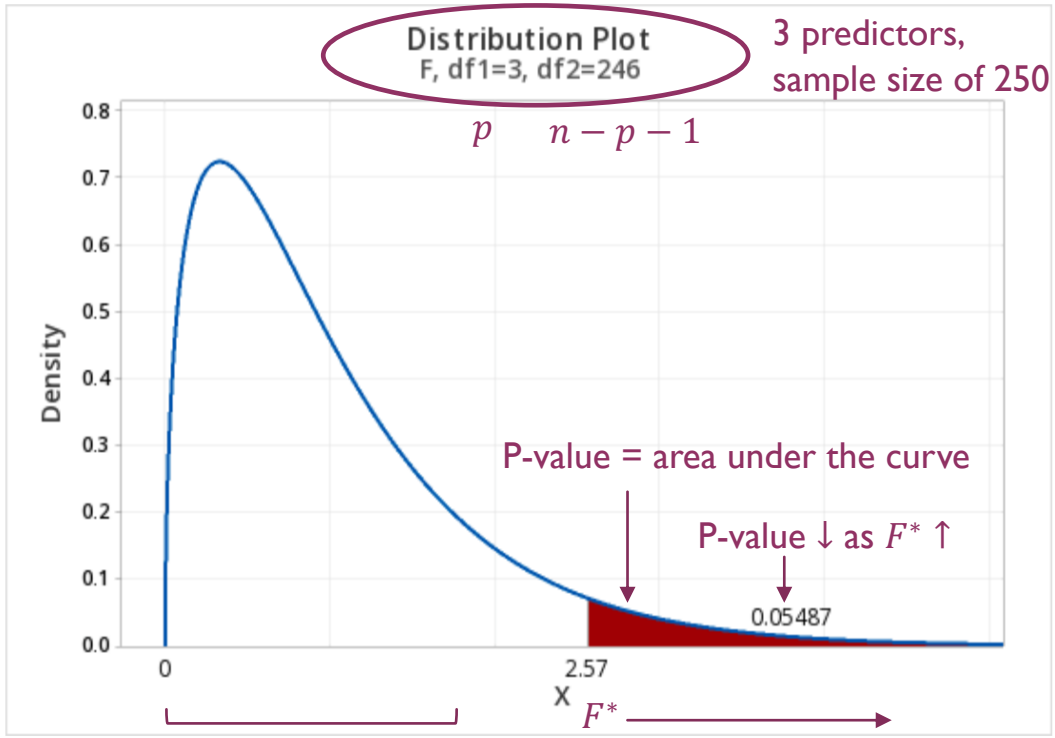
- Analysis of Variance Test of Overall Significance compares RSS and SS_{reg} to identify existence of a linear relationship
- Hypothesis: H_0: $\boldsymbol{\beta}_1 = \mathbf{0}$ versus H_a: $\boldsymbol{\beta}_1 \neq \mathbf{0}$ where we split our vector of coefficients $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_1)^T$
  - Testing the null "all slopes are zero" versus "at least one slope is not zero"
- Assuming the null is true, we test this using the statistic $F^* = \frac{SS_{reg}/p}{RSS/n-p-1} \sim F(p, n-p-1)$
  - Dividing each SS by its df standardizes the quantity, giving us the Mean Squares Regression and Mean Squares Residual
- Test components sometimes summarized in an ANOVA table:

Source	DF	Sum Squares	Mean Squares	F value
Regression	p	SS_{reg}	$MS_{reg} = SS_{reg}/p$	MS_{reg}/MSR
Residual	$n-p-1$	RSS	$MSR = RSS/n-p-1$	–
Total	$n-1$	SST	–	–

CONCLUDING ANOVA TEST

- F distribution assumes the null hypothesis is true.
  - Displays values we expect if no linear relationship exists
  - Expect to see small values of $F^* = \frac{SS_{reg}/p}{RSS/n-p-1}$
- More extreme F^* values will be in the right tail
  - Bigger values are due to $MS_{reg} > MSR$
  - The larger this ratio, the larger the test statistic will be, and therefore the smaller the p-value
- If p-value $< \alpha$, or $F^* > F_{(1-\alpha),(p,n-p-1)}$, then reject H_0 and conclude a statistically significant linear relationship exists for at least one predictor

https://online.stat.psu.edu/stat200/book/export/html/213



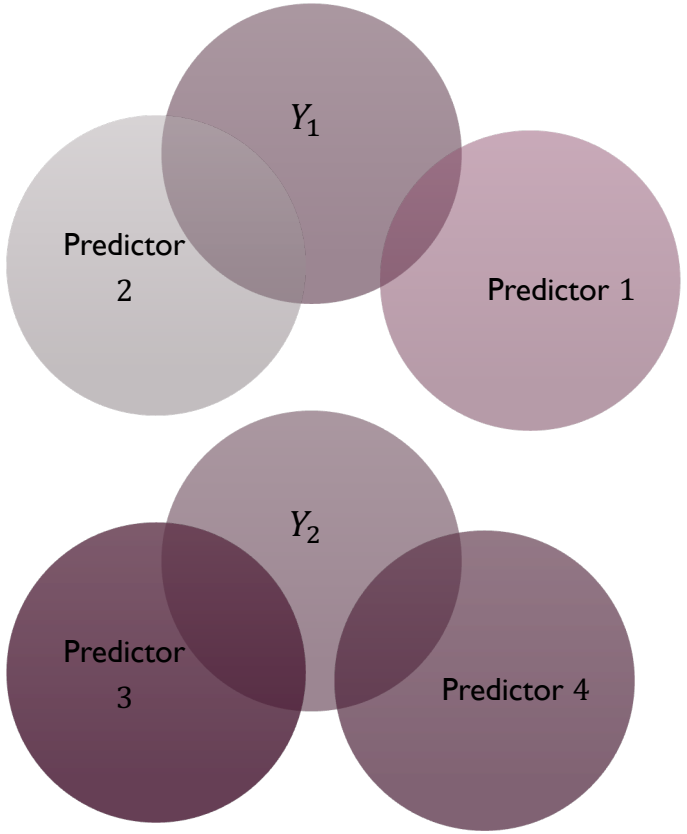
COEFFICIENTS OF DETERMINATION

- The Coefficient of Determination (R^2), given by

$$R^2 = \frac{SS_{reg}}{SST} = 1 - \frac{RSS}{SST}$$

- $0 \leq R^2 \leq 1$ so it represents the proportion of variation in the response that has been explained by the model
- When comparing models with different numbers of predictors, $SS_{reg,big} > SS_{reg,small}$
  - A bigger model will also have a larger R^2 even if extra predictors not significant
- Like the ANOVA test/Partial F test, adjust decomposition with degrees of freedom:

$$R^2_{adj} = 1 - \frac{RSS/(n-p-1)}{SST/(n-1)}$$



LIKELIHOOD GOODNESS CRITERIA

- The four main measures to compare models are:
  - Adjusted R^2: $R^2_{adj} = 1 - \frac{RSS/(n-p-1)}{SST/(n-1)}$
  - AIC: $AIC \propto n\, ln\left(\frac{RSS}{n}\right) + 2p$
  - Corrected AIC: $AIC_c \propto n\, ln\left(\frac{RSS}{n}\right) + 2p + \frac{2(p+2)(p+3)}{n-p-1}$
  - BIC: $BIC \propto n\, ln\left(\frac{RSS}{n}\right) + (p+2)\ln(n)$
- The RSS term drives the idea of goodness, while the rest are penalties for number of predictors.
  - we know that models with more predictors automatically have lower RSS even if not useful predictors
  - Small RSS → large Adjusted R^2 but small AIC, corrected AIC and BIC

VALIDATION

- Good model captures true relationship of population
  - but built using only a sample so could be tailored to data collected (i.e. overfitting)
- Investigate if model performs similarly on another dataset from same population
- No additional sample available so mimic by:
  - randomly splitting dataset into a train and test dataset
  - resampling the data in pieces to see how different subsets of data yield different answers
- Compare properties to determine similarity
- Comparisons:
  - ”heuristic” - compare model fit, assumptions and diagnostics between model in training and test
  - prediction error (MSE) - compare predictions of training set model on test set data to training set data
    - want similar values and small MSE
- Resampling via leave-one-out cross validation
  - omit 1 observation from fitting model and predict that observations value, do for all observations
  - want small MSE averaged across each “test” dataset

EXAM STUDYING AND PROBLEM-SOLVING TIPS

TYPES OF QUESTIONS

Concepts

- Review slides/in-class notes
- Questions from worksheets/polls
- Conclusion of worksheets

Calculations

- Examples from slides/quizzes
- Worksheets for use of R output
- Additional practice problems/tests

Proofs/derivations

- Examples from slides/in-class notes
- Additional practice problems/tests

WAYS TO REVIEW

- Consolidate and synthesize:
  - crib sheets from each module, study cards of terminology/notation
  - draw connections between topics
    - e.g., recognize how all material based on decomposition of SS are related
    - what topics are used to achieve similar purpose
    - what topics depend on other topics (e.g., when do assumptions play a role in ability to make decisions)
  - summarize each topic in your own words, why each is needed/important, and examples of how they occur
- Understand formula components:
  - what does each term represent and what does it tell us about a model
  - where would I find this information in R output or other values that could be provided.
    - related to drawing connections between topics
- Redo exercises, examples, practice problems:
  - do it **without** solutions to identify where you might get stuck and make note of what you needed to move forward → then review this topic
  - think about what the problem is asking and what information told you this

IMPORTANT EXAM DETAILS

- Date and time: December 12 from 7-10pm → BE THERE 30 MINUTES EARLY
- Location: based on last name

STA302H1F	ALL	A - HE,T	BN 210_N	12-Dec	7:00 PM	10:00 PM
STA302H1F	Regular Deferred	~~A - KH~~	~~ES 1050~~	12-Dec	7:00 PM	10:00 PM
STA302H1F	ALL	HE,X - LIU,Y	BN 210_S	12-Dec	7:00 PM	10:00 PM
STA302H1F	Regular Deferred	~~KI - ZZ~~	~~KC KNOX~~	12-Dec	7:00 PM	10:00 PM
STA302H1F	ALL	LIU,Z - WANG,H	BN 322	12-Dec	7:00 PM	10:00 PM
STA302H1F	ALL	WANG,J - YANG,CHENG	KC KNOX	12-Dec	7:00 PM	10:00 PM
STA302H1F	ALL	YANG,CHENY - ZZ	VO AUDB	12-Dec	7:00 PM	10:00 PM

- What to bring:
  - calculator (non-programming/non-graphing)
  - pencil/pen and eraser
  - Tcard or other government issued photo ID
- NO ID CARD MEANS NO ENTRANCE TO EXAM ROOM