STAT 5000

Statistical Methods I

Week 13

Fall 2024

Dr. Danica Ommen

Unit 3

Multiple Linear Regression

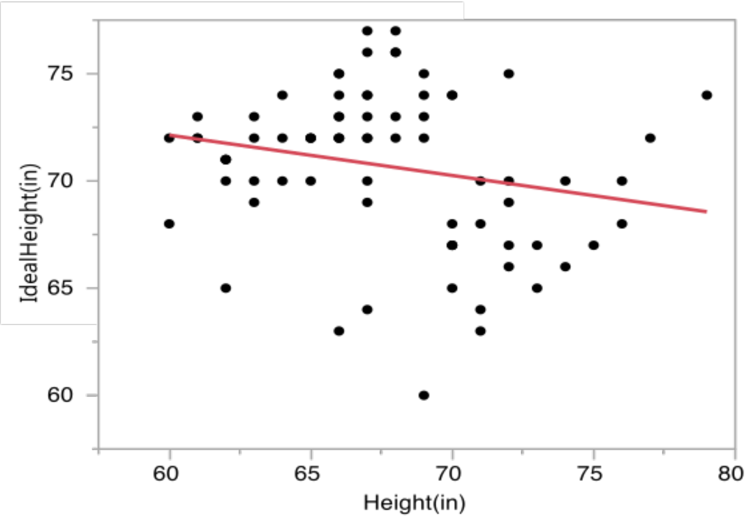
Categorical Predictors

MLR with Categorical Predictors

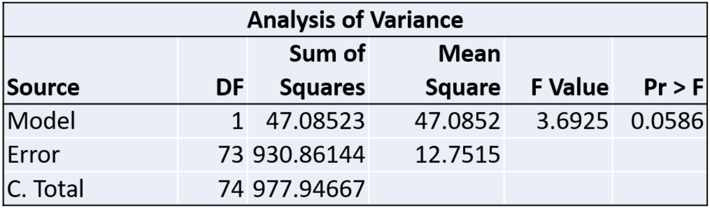
Motivating Example

- Students in STAT 101 at Iowa State University were asked in a recent semester to provide demographic data for use during the semester
- A random sample of 75 students were selected and a few of the variables collected were:
  - Height (inches)
  - Height of the student's ideal romantic partner (inches)
- SLR Model: $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$
  - $Y =$ Height of ideal romantic partner (inches)
  - $x =$ Height (inches)

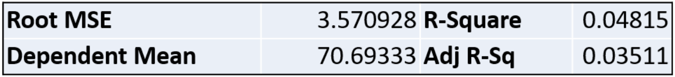
Motivating Example



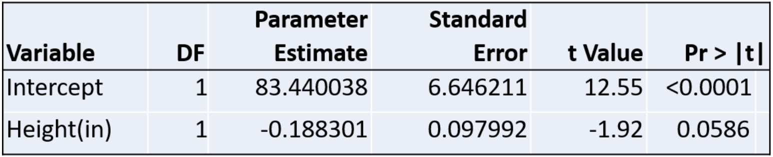
Motivating Example



Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	47.08523	47.0852	3.6925	0.0586
Error	73	930.86144	12.7515		
C. Total	74	977.94667			



Root MSE	3.570928	**R-Square**	0.04815
Dependent Mean	70.69333	**Adj R-Sq**	0.03511



Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	83.440038	6.646211	12.55	<0.0001
Height(in)	1	-0.188301	0.097992	-1.92	0.0586

Motivating Example

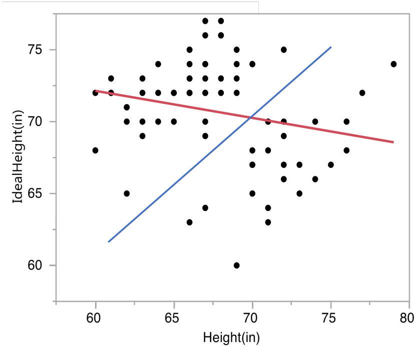
- Linear relationship between student's height and the height of their ideal romantic partner is very weak and negative
- $R^2 = 4.815\%$

  > *Only 4.815% of the variation in the height of a student's ideal romantic partner can be explained by the simple linear regression with the student's height*

- Student's height is statistically significant at the 10% level, but not the 5% level

Motivating Example: What's going on?

- Two clusters of points in scatterplot:
  - To left of blue line
  - To right of blue line
- What are those clusters related to?



Motivating Example: What's going on?

- Variable for student's gender should be added:
  - $x_{1i} =$ student's height
  - $x_{2i} = \begin{cases} 1, & \text{student is female} \\ 0, & \text{student is male} \end{cases}$
- Model:

$$Y_i = \beta_0 + \beta_1\, x_{1i} + \beta_2\, x_{2i} + \epsilon_i$$

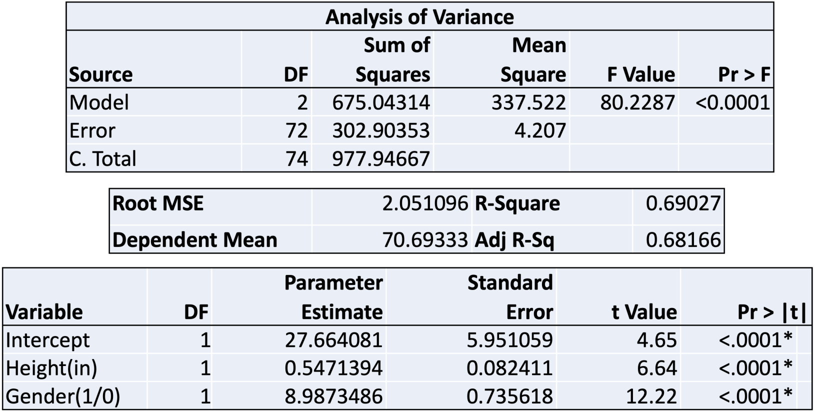
  - Females ($x_{2i} = 1$): $Y_i = \beta_0 + \beta_1\, x_{1i} + \beta_2 \ + \epsilon_i$
  - Males ($x_{2i} = 0$): $Y_i = \beta_0 + \beta_1\, x_{1i} + \epsilon_i$

MLR with Categorical Predictors

Motivating Example: Comparing the Two Models

- Same Slope (β_1):
  - Assumes relationship between height of ideal romantic partner (response variable Y) and student's height (quantitative explanatory variable x_1) is the same for both groups in the categorical explanatory variable x_2 (Gender = female and Gender = male).
- Different Intercepts:
  - For females: $\beta_0 + \beta_2$
  - For males: β_0
- Result: parallel regression lines

Motivating Example: ANOVA Table



Analysis of Variance					
		Sum of	Mean		
Source	DF	Squares	Square	F Value	Pr > F
Model	2	675.04314	337.522	80.2287	<0.0001
Error	72	302.90353	4.207		
C. Total	74	977.94667			

Root MSE	2.051096	**R-Square**	0.69027
Dependent Mean	70.69333	**Adj R-Sq**	0.68166

		Parameter	Standard		
Variable	DF	Estimate	Error	t Value	Pr > |t|
Intercept	1	27.664081	5.951059	4.65	<.0001*
Height(in)	1	0.5471394	0.082411	6.64	<.0001*
Gender(1/0)	1	8.9873486	0.735618	12.22	<.0001*

Motivating Example: ANOVA Summary

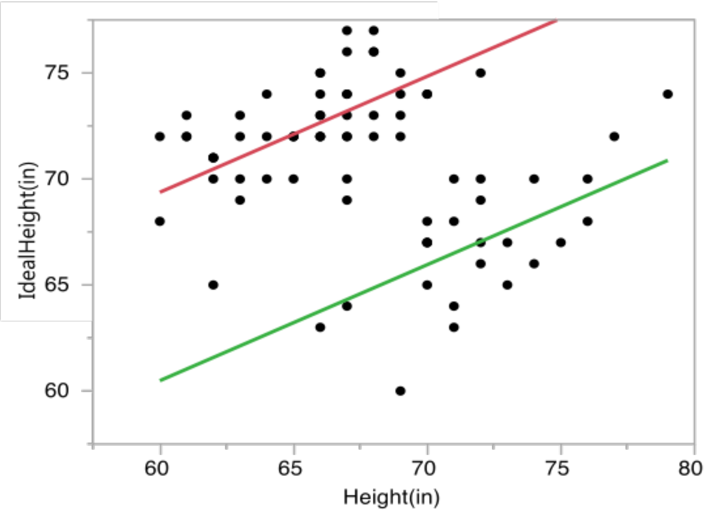
- Model is highly significant in explaining the height of the students' ideal romantic partner (p-value < 0.0001)
- $R^2 = 0.6903$:
  69.03% of the variation in the height of the students' ideal romantic partner can be explained by the multiple linear regression model with height and Gender of the student
- Given height in the model, Gender is highly significant (p-value < 0.0001)
- Given Gender in the model, height is highly significant (p-value < 0.0001)

Motivating Example: Model Summary

- Using the parameter estimates from the MLR, we can determine the estimated intercept and slope of the two models - one for females and one for males:

$$\hat{Y}_i = \begin{cases} 36.651 + 0.547\, x_{1i}, & \text{females} \\ 27.664 + 0.574\, x_{1i}, & \text{males} \end{cases}$$

Motivating Example: Model Summary



Motivating Example: Interaction Model

- We can consider a model with an interaction term between x_1 and x_2 - height and gender:

$$Y_i = \beta_0 + \beta_1\, x_{1i} + \beta_2\, x_{2i} + \beta_3\, x_{1i}x_{2i} + \epsilon_i$$

- This model will allow for a different relationship between students' height and the height of their ideal romantic partner

MLR with Categorical Predictors

Motivating Example: Interaction Model

- Two Models:
  - Females ($x_{2i} = 1$):

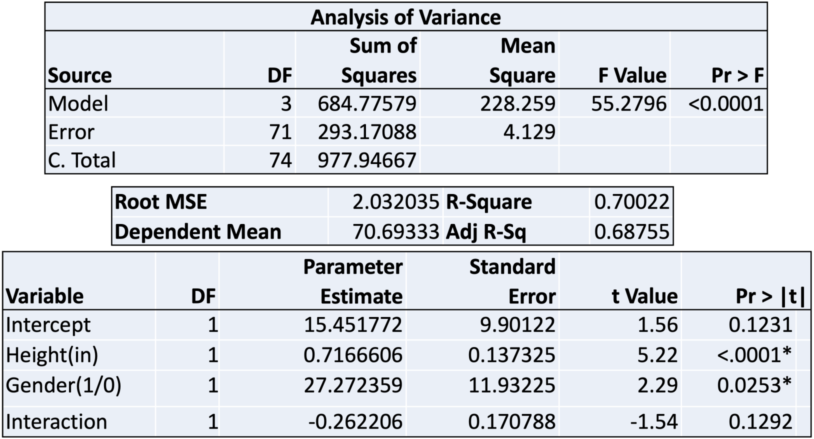
$$Y_i = (\beta_0 + \beta_2) + (\beta_1 + \beta_3)\, x_{1i} + \epsilon_i$$

  - Males ($x_{2i} = 0$):

$$Y_i = \beta_0 + \beta_1\, x_{1i} + \epsilon_i$$

- Comparison
  - Different Slopes: $(\beta_1 + \beta_3)$ vs. β_1
  - Different Intercepts: $\beta_0 + \beta_2$ vs. β_0

Motivating Example: Interaction ANOVA



Analysis of Variance					
		Sum of	Mean		
Source	DF	Squares	Square	F Value	Pr > F
Model	3	684.77579	228.259	55.2796	<0.0001
Error	71	293.17088	4.129		
C. Total	74	977.94667			

Root MSE	2.032035	**R-Square**	0.70022
Dependent Mean	70.69333	**Adj R-Sq**	0.68755

		Parameter	Standard		
Variable	DF	Estimate	Error	t Value	Pr > |t|
Intercept	1	15.451772	9.90122	1.56	0.1231
Height(in)	1	0.7166606	0.137325	5.22	<.0001*
Gender(1/0)	1	27.272359	11.93225	2.29	0.0253*
Interaction	1	-0.262206	0.170788	-1.54	0.1292

Motivating Example: Interaction Summary

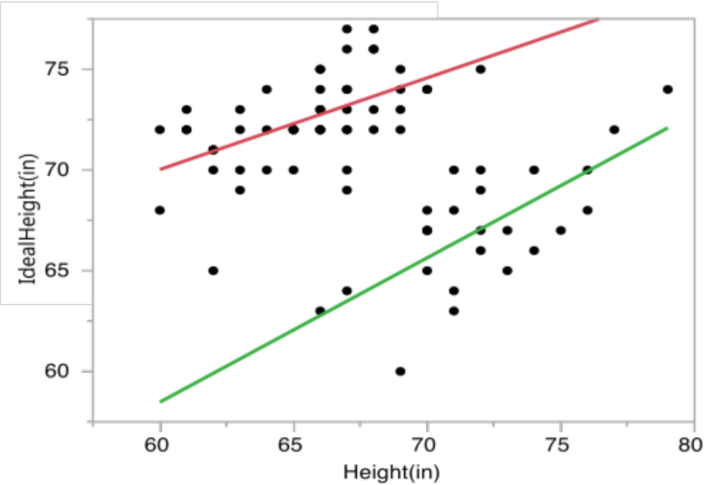
- Model is highly significant in explaining the height of the students' ideal romantic partner (p-value < 0.0001)
- $R^2 = 0.7002$
  70.02% of the variation in the height of the students' ideal romantic partner can be explained by the multiple linear regression model with height and gender of the student
- The interaction term is not statistically significant (p-value $= 0.1292$)

Motivating Example: Interaction Summary

- Using the parameter estimates from the MLR model, we can determine the estimated intercept and slope of the models for females and for males:

$$\hat{Y}_1 = \begin{cases} 42.724 + 0.455\, x_{1i}, & \text{females} \\ 15.452 + 0.717\, x_{1i}, & \text{males} \end{cases}$$

Motivating Example: Interaction Summary



MLR with Categorical Predictors

Motivating Example: 2 MLR Model Comparison

- MLR Model
  - $SS_{\text{error}} = 302.90353$
  - $MS_{\text{error}} = 4.207$
  - $R^2 = 69.03\%$
  - adj-$R^2 = 68.17\%$
- MLR Model with Interaction
  - $SS_{\text{error}} = 293.17088$
  - $MS_{\text{error}} = 4.129$
  - $R^2 = 70.02\%$
  - adj-$R^2 = 68.76\%$

MLR with Categorical Predictors

Motivating Example: Model Selection?

- The two models are very similar
- MLR model with interaction term has slightly better values for MS_{error} and adj-R^2
- Interaction term is not statistically significant (p-value $= 0.1292$)
- Estimated MLR model with interaction term appears to fit the points in the scatterplot slightly better than the MLR model

MLR with Categorical Predictors

Motivating Example: Alternative Parameterization

- Baseline coding of categories:

$$x_{2i} = \begin{cases} 1, & \text{student is female} \\ 0, & \text{student is male} \end{cases}$$

- Sum-to-Zero coding of categories:

$$x_{2i} = \begin{cases} 1, & \text{student is female} \\ -1, & \text{student is male} \end{cases}$$

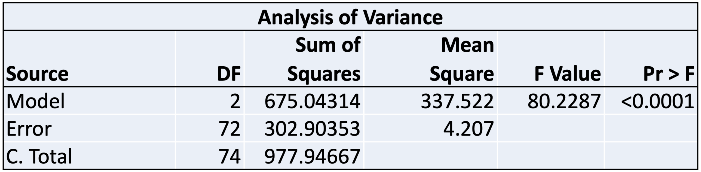
Motivating Example: Sum-to-Zero Constraint

- Explanatory Variables:
  - $x_{1i} =$ student's height
  - $x_{2i} = \begin{cases} 1, & \text{student is female} \\ -1, & \text{student is male} \end{cases}$
- Model:

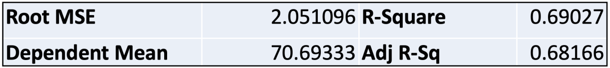
$$Y_i = \beta_0 + \beta_1 \, x_{1i} + \beta_2 \, x_{2i} + \epsilon_i$$

  - Females ($x_{2i} = 1$): $Y_i = \beta_0 + \beta_1 \, x_{1i} + \beta_2 \; + \epsilon_i$
  - Males ($x_{2i} = -1$): $Y_i = \beta_0 + \beta_1 \, x_{1i} - \beta_2 \; + \epsilon_i$
- Comparison
  - Same Slope: β_1
  - Different Intercepts: $\beta_0 + \beta_2$ vs. $\beta_0 - \beta_2$

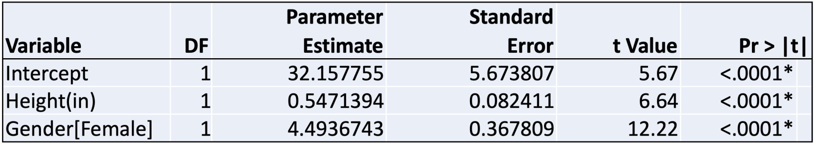
Motivating Example: Sum-to-Zero ANOVA



Analysis of Variance					
		Sum of	Mean		
Source	DF	Squares	Square	F Value	Pr > F
Model	2	675.04314	337.522	80.2287	<0.0001
Error	72	302.90353	4.207		
C. Total	74	977.94667			



Root MSE	2.051096	**R-Square**	0.69027
Dependent Mean	70.69333	**Adj R-Sq**	0.68166



		Parameter	Standard		
Variable	**DF**	**Estimate**	**Error**	**t Value**	**Pr > |t|**
Intercept	1	32.157755	5.673807	5.67	<.0001*
Height(in)	1	0.5471394	0.082411	6.64	<.0001*
Gender[Female]	1	4.4936743	0.367809	12.22	<.0001*

Motivating Example: Sum-to-Zero Model Summary

- Using the parameter estimates from the MLR, we can determine the estimated intercept and slope of the two models - one for females and one for males:

$$\hat{Y}_i = \begin{cases} (32.157 + 4.494) + 0.547\, x_{1i}, & \text{females} \\ (32.157 - 4.494) + 0.547\, x_{1i}, & \text{males} \end{cases}$$

$$= \begin{cases} 36.651 + 0.547\, x_{1i}, & \text{females} \\ 27.664 + 0.547\, x_{1i}, & \text{males} \end{cases}$$

MLR with Categorical Predictors

Motivating Example: Sum-to-Zero with Interaction

- We can consider a model with an interaction term between x_1 and x_2 - height and gender
  - For Females:

$$Y_i = (\beta_0 + \beta_2) + (\beta_1 + \beta_3)\, x_{1i} + \epsilon_i$$

  - For Males:

$$Y_i = (\beta_0 - \beta_2) + (\beta_1 - \beta_3)\, x_{1i} + \epsilon_i$$

- This model will allow for a different relationship between students' height and the height of their ideal romantic partner
  - Different Slopes: $(\beta_1 + \beta_3)$ vs. $(\beta_1 - \beta_3)$
  - Different Intercepts: $(\beta_0 + \beta_2)$ vs. $(\beta_0 - \beta_2)$

Motivating Example: Sum-to-Zero with Interaction ANOVA

Analysis of Variance					
		Sum of	Mean		
Source	DF	Squares	Square	F Value	Pr > F
Model	3	684.77579	228.259	55.2796	<0.0001
Error	71	293.17088	4.129		
C. Total	74	977.94667			

Root MSE	2.032035	**R-Square**	0.70022
Dependent Mean	70.69333	**Adj R-Sq**	0.68755

		Parameter	Standard		
Variable	DF	Estimate	Error	t Value	Pr > |t|
Intercept	1	29.087952	5.966125	4.88	<.0001*
Height(in)	1	0.5855576	0.085394	6.86	<.0001*
Gender(1/0)	1	13.63618	5.966125	2.29	0.0253*
Interaction	1	-0.131103	0.085394	-1.54	0.1292

Motivating Example: Sum-to-Zero with Interaction

- Using the parameter estimates from the MLR model, we can determine the estimated intercept and slope of the models for females and for males:

$$\hat{Y}_1 = \begin{cases} (29.088 + 13.636) + (0.586 - 0.131)\, x_{1i}, & \text{females} \\ (29.088 - 13.636) + (0.586 + 0.131)\, x_{1i}, & \text{males} \end{cases}$$

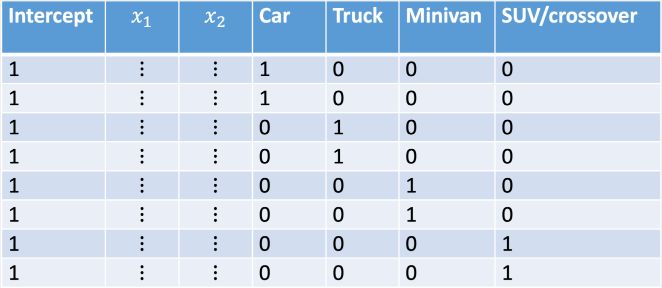
$$= \begin{cases} 42.724 + 0.455\, x_{1i}, & \text{females} \\ 15.452 + 0.717\, x_{1i}, & \text{males} \end{cases}$$

MLR with Categorical Predictors

Categorical Predictors with 3+ Levels

- We can also add categorical variables with more than two categories to our multiple linear regression model by adding columns to the design matrix:
- For example, the type of vehicle has 4 categories: car, truck, minivan, SUV/crossover

Will the following design matrix work?

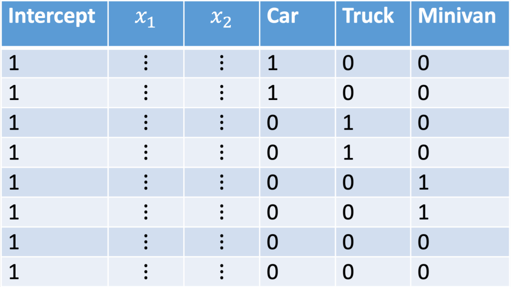


Intercept	x_1	x_2	Car	Truck	Minivan	SUV/crossover
1	⋮	⋮	1	0	0	0
1	⋮	⋮	1	0	0	0
1	⋮	⋮	0	1	0	0
1	⋮	⋮	0	1	0	0
1	⋮	⋮	0	0	1	0
1	⋮	⋮	0	0	1	0
1	⋮	⋮	0	0	0	1
1	⋮	⋮	0	0	0	1

MLR with Categorical Predictors

Categorical Predictors with 3+ Levels

- The problem is that design matrix is over-parameterized
- To fix this problem, we will constraint the value of one of the groups (called the baseline group)
- For example, let's use the SUV/crossover group as our baseline group



Intercept	x_1	x_2	Car	Truck	Minivan
1	⋮	⋮	1	0	0
1	⋮	⋮	1	0	0
1	⋮	⋮	0	1	0
1	⋮	⋮	0	1	0
1	⋮	⋮	0	0	1
1	⋮	⋮	0	0	1
1	⋮	⋮	0	0	0
1	⋮	⋮	0	0	0

Categorical Predictors with 3+ Levels

- The model is:

$$Y_i = \beta_0 + \beta_1\, x_{1i} + \beta_2\, x_{2i} + \beta_3\, x_{3i} + \beta_4\, x_{4i} + \beta_5\, x_{5i} + \epsilon_i$$

  where

  - $x_{3i} = \begin{cases} 1, & \text{car} \\ 0, & \text{otherwise} \end{cases}$
  - $x_{4i} = \begin{cases} 1, & \text{truck} \\ 0, & \text{otherwise} \end{cases}$
  - $x_{5i} = \begin{cases} 1, & \text{minivan} \\ 0, & \text{otherwise} \end{cases}$

Categorical Predictors with 3+ Levels

- The parameters β_3, β_4, and β_5 have special interpretations in this model. They are:
  - $\beta_3 =$ the difference in the expected value of the response variable between cars and SUV/crossovers
  - $\beta_4 =$ the difference in the expected value of the response variable between trucks and SUV/crossovers
  - $\beta_5 =$ the difference in the expected value of the response variable between minivans and SUV/crossovers

MLR with Categorical Predictors

Categorical Predictors with 3+ Levels

- Testing for the statistical significance of the type of vehicle requires testing:

$$H_0 : \beta_3 = \beta_4 = \beta_5 = 0 \quad \text{vs.} \quad H_a : \text{ at least one } \beta_j \neq 0, j = 3, 4, 5$$

- Use the partial F-test to do it
  - Reduced model: the model without the categorical variable
  - Full model: the model with the categorical variable
  - The F-statistic is

$$F = \frac{(SSE_{\text{reduced}} - SSE_{\text{full}})/(m-1)}{MSE_{\text{full}}}$$

    where $m =$ the number of levels/categories in the categorical variable ($m - 1 = 3$ in our example)
  - Reject H_0 if $F > F_{m-1, n-(k+1), 1-\alpha}$ where $n - (k+1)$ is the error d.f. for the full model

Unit 3

Multiple Linear Regression

Model Selection

MLR: Model Selection

Importance of Model Selection

- Including too few variables in the model leads to inaccurate estimates of coefficients and response means
- Including too many variables leads to unnecessary excess variability in estimates of the coefficients and mean response

MLR: Model Selection

Theory

- Consider two models
  - Model A ("fit"): $\mathbf{Y} = X\beta + \epsilon$
  - Model B ("true"): $\mathbf{Y} = X\beta + Z\gamma + \epsilon$
- Fitting model A (omitting the variables in Z) leads to a biased estimate of the regression coefficients:

$$\mathbf{b} = (X^TX)^{-1}X^T\mathbf{Y}$$

$$\begin{aligned}
E(\mathbf{b}) &= E((X^TX)^{-1}X^T\mathbf{Y}) \\
&= (X^TX)^{-1}X^TE(\mathbf{Y}) \\
&= (X^TX)^{-1}X^T(X\beta + Z\gamma) \\
&= (X^TX)^{-1}X^TX\beta + (X^TX)^{-1}X^TZ\gamma \\
&= \beta + (X^TX)^{-1}X^TZ\gamma
\end{aligned}$$

Theory

- Estimates of the mean responses based on Model A may be biased:

$$\hat{\mathbf{Y}} = X\mathbf{b} = X(X^TX)^{-1}X^T\mathbf{Y} = P_X\mathbf{Y}$$

  where $P_X = X(X^TX)^{-1}X^T$ projects vectors into the space spanned by the columns of X.

- Then,

$$\begin{aligned} E(\hat{\mathbf{Y}}) &= E(P_X\mathbf{Y}) \\ &= P_XE(\mathbf{Y}) \\ &= P_X(X\beta + Z\gamma) \\ &= X\beta + P_XZ\gamma \end{aligned}$$

  because $P_XX = X(X^TX)^{-1}X^TX = XI_{n\times n} = X$.

MLR: Model Selection

Theory

- Bias in $\hat{\mathbf{Y}}$ based on model A:

$$\begin{aligned} \delta &= E(\hat{\mathbf{Y}}) - E(\mathbf{Y}) \\ &= (X\beta + P_X Z\gamma) - (X\beta + Z\gamma) \\ &= (P_X - I)Z\gamma \end{aligned}$$

- No bias if $\gamma = \mathbf{0}$ or each column in Z is in the column space of X, e.g. the correct model has $E(\mathbf{Y}) = X\beta$

MLR: Model Selection

Theory

- Estimate of the error variance based on Model A may be biased:

$$E(MS_{error}) = \sigma^2 + \frac{\gamma^T Z^T (I - P_X) Z\gamma}{n-k-1} = \sigma^2 + \frac{1}{n-k-1}\Sigma_i \mathrm{Bias}(\hat{Y}_i)^2$$

- $p = k + 1$ is the number of parameters in the MLR model
- Omitting useless terms ($\gamma = 0$): $E(MS_{error}) = \sigma^2$
- Omitting needed terms ($\gamma \neq 0$): $E(MS_{error}) > \sigma^2$

Theory

- The total variance of $\hat{\mathbf{Y}}$:

$$\begin{aligned}\sum_i \mathrm{Var}(\hat{Y}_i) &= \mathrm{trace}(\mathrm{Var}(\hat{\mathbf{Y}})) \\ &= \mathrm{trace}(\mathrm{Var}(P_X\mathbf{Y})) \\ &= \mathrm{trace}(P_X(\sigma^2 I)P_X^T) \\ &= \sigma^2\mathrm{trace}(P_X) \\ &= \sigma^2(k+1)\end{aligned}$$

- Because P_X is symmetric and idempodent, i.e.,

$$P_X P_X^T = P_X P_X = P_X,$$

it has $k+1$ eigenvalues equal to one, and the rest are zero

MLR: Model Selection

Theory

- Adding a predictor to Model A (adding a column to X that is not a linear combination of the columns already in X)
  - decreases bias (or may leave it the same) of $\hat{\mathbf{Y}}$
  - increases the total variance of the estimates of the response means $\hat{\mathbf{Y}}$, because the column rank of X, which is also the rank of the new P_X, increases by 1
- If we fit the "true" model, Model B, then
  - bias $= 0$
  - variance $= \sum_i \text{Var } \hat{Y}_i = \sigma^2(k + 1 + \dim(Z))$

Model Selection Criteria

- How many explanatory variables? Which ones?
- Criteria for identifying the “best” model
  - R^2
  - adj R^2
  - C_p
  - AIC
  - BIC

MLR: Model Selection

Criterion: R²

$$R^2 = \frac{SS_{\text{model}}}{SS_{\text{Total}}}$$

- Larger values indicate better model
- Maximizing R^2 is equivalent to minimizing SS_{error}
- R^2 never decreases when adding an explanatory variable to model
- Most useful for comparing two models with the same number of explanatory variables

MLR: Model Selection

Criterion: adj R²

$$\text{adj } R^2 = 1 - \frac{MS_{\text{error}}}{SS_{\text{Total}}/(n-1)}$$

- Larger values indicate better model
- Maximizing adj R^2 equivalent to minimizing $MS_{\text{error}} = \hat{\sigma}^2$
- Does not necessarily increase when adding an explanatory variable to model
- Most useful in comparing models with different numbers of explanatory variables

MLR: Model Selection

Criterion: $\mathbf{C_p}$

$$C_p = \frac{SS_{\text{error}}}{\hat{\sigma}^2} - [n - 2(k+1)]$$

- SS_{error} from fitted model
- $\hat{\sigma}^2$ is MS_{error} for model containing all explanatory variables
- $p = k + 1$ is the number of coefficients in the fitted model

Criterion: $\mathbf{C_p}$

- The rationale behind C_p statistic is to minimize $E[\Sigma_i(\hat{Y}_i - E(Y_i))^2]$, the mean squared error of the predictions, *MSEP*= bias² + variance

$$\begin{aligned} MSEP &= E[\Sigma_i(\hat{Y}_i - E(\hat{Y}_i) + E(\hat{Y}_i) - E(Y_i))^2] \\ &= E[\Sigma_i(\hat{Y}_i - E(\hat{Y}_i))^2] + E[\Sigma_i(E(\hat{Y}_i) - E(Y_i))^2] \\ &= \Sigma_i \mathrm{Var}(\hat{Y}_i) + \Sigma_i \mathrm{Bias}(\hat{Y}_i)^2 \\ &= \sigma^2(k+1) + E(SS_{error}) - \sigma^2(n-k-1) \\ &= E(SS_{error}) - \sigma^2(n-2(k+1)) \end{aligned}$$

- The second to last line uses the previously obtained relationship of bias and $E(MS_{error})$

MLR: Model Selection

Criterion: $\mathbf{C_p}$

$$C_p = \frac{SS_{\text{error}}}{\hat{\sigma}^2} - (n - 2(k+1))$$

- Full name: Mallow's C_p
- Good models have C_p around $p = k + 1$
    - Why?
- $C_p < p$ is no problem (sampling error)
- Large C_p indicates poor model
- Let m denote the size of biggest possible model with $m - 1$ explanatory variables and m regression coefficients.
- For the model containing all explanatory variables, $C_p = m$
- Limited to MLR models

MLR: Model Selection

Criterion: $\mathbf{C_p}$

- C_p is related to the F-test that the sub-model with only p explanatory variables is acceptable

  $$C_p = (m - p)(F - 1) + p$$

  If $F < 2$ then $C_p < m$ and the data do not provide enough evidence on bias to reject the sub-model
- C_p focuses on prediction
- You can think of using C_p to minimize

  $$SS_{error} + \text{[penalty for p]}$$

  (so do AIC and BIC)

MLR: Model Selection

Criterion: AIC

$$\text{AIC} = n \log(SS_{\text{error}}/n) + 2(k+1)$$

- Full name: Akaike Information Criterion
- Smaller values indicate better models
- Favors models with a slightly larger number of explanatory variables, i.e., may include a few non-significant explanatory variables
- Not limited to MLR models

MLR: Model Selection

Criterion: BIC

$$\text{BIC} = n \log(SS_{\text{error}}/n) + (k+1)\log(n)$$

- Full name: Bayesian Information Criterion
- Smaller values indicate better models
- Leads to smaller models than AIC (larger penalty for explanatory variables)
- Not limited to MLR models

MLR: Model Selection

Summary

- Many different approaches
- Measures that focus on fit
  - R^2 (fit using SS_{error}): bad
  - adjusted R^2 (fit using MS_{error})
- Measures that combine fit and complexity
  - general idea: fit + penalty for model complexity
  - Mallows C_p: least penalty
  - AIC: larger penalty
  - BIC: largest penalty (usually)
  - Often C_p, AIC and BIC lead to same model
    - When they differ, smaller penalty $\Rightarrow$ more variables
    - C_p selects most variables
    - BIC selects fewest

MLR: Model Selection

Example: Grandfather Clocks

Model	R^2	adj R^2	C_p	AIC	BIC
Numbid	15.62%	12.81%	484.299	379.953	382.884
Age	53.24%	51.68%	255.914	361.065	363.997
Age & Numbid	89.23%	88.49%	39.361	316.065	320.462
Age * Numbid	95.39%	94.89%	4	290.938	296.801

- Significance of explanatory variables depends on presence of other explanatory variables in model.
- Cannot make independent decisions about significance of explanatory variables.
- How do we decide which explanatory variables to be included in the final model?

MLR: Model Selection

Selection Techniques

- Different methods for searching among models
  - All possible subsets of a given group of explanatory variables
  - Stepwise model selection
    - Backward elimination
    - Forward selection
    - Stepwise (Mixed) selection

MLR: Model Selection

Selection Techniques: All Possible Subsets

- Set of k explanatory variables
- Fit all $2^k - 1$ possible models
- Compare models using some criterion (adj-R^2, C_p, AIC, BIC)
- Works up to about $k = 20$ (i.e., takes a reasonable amount of time to process $2^k - 1$ possible models)
- Review the best models of each size: $1, 2, \ldots, k$

MLR: Model Selection

Selection Techniques: Stepwise Methods

- Enter or delete one variable at a time from model according to algorithm
- Less time to compute than all possible subsets
- Possible algorithms
  - Forward selection
  - Backward elimination (selection)
  - Stepwise selection

MLR: Model Selection

Stepwise Methods: Forward Selection

1. Start with only intercept in model
2. Fit all one variable models, select the explanatory variable with the largest correlation with the response as long as effect test for variable is statistically significant (p-value $< \alpha_{entry}$)
3. Add to the model the next explanatory variable that reduces the SS_{error} the most as long as effect test for variable is statistically significant (p-value $< \alpha_{\text{entry}}$)
4. Repeat step 3 until no significant variables can be added to the model

MLR: Model Selection

Stepwise Methods: Backward Elimination

1. Begin with the largest possible model (all *k* explanatory variables)
2. Do an effects test for each explanatory variable and compute the *p*-value
3. Delete the variable with the least significant effect test (largest *p*-value) as long as p-value $\geq \alpha_{stay}$
4. Fit the model again & repeat step 3
5. Stop when there is no explanatory variable with an effect test with p-value $\geq \alpha_{stay}$

MLR: Model Selection

Stepwise Methods: Stepwise Selection

1. Start with only intercept in model
2. Fit all one variable models, select the explanatory variable with the largest correlation with the response as long as effect test for variable is statistically significant (p-value $< \alpha_{entry}$)
3. Add to the model the next explanatory variable that reduces the SS_{error} the most as long as effect test for variable is statistically significant (p-value $< \alpha_{\text{entry}}$)
4. Examine each variable in the current model to make sure effect test for variable is still significant (p-value $< \alpha_{stay}$). If not, delete variable from model.
5. Repeat steps 3-4 until there are no changes
6. Note: need $\alpha_{entry} \leq \alpha_{stay}$ to avoid never ending loops

MLR: Model Selection

Difficulties with Model Selection

- Multicollinearity: high correlation between some explanatory variables
- Example: Suppose x_j and x_l have a high correlation (near -1 or 1) and are both in model
  - Significance test for either β_j or β_l: does x_j or x_l significantly add to the model that includes all other explanatory variables?
  - Once one of the variables is in the model, the other is not likely to significantly add to model due to their close association

MLR: Model Selection

Assessing Impact of High Correlation

- Pairwise correlation matrix for explanatory variables
  - $r > |0.7|$
- Fit models with and without highly correlated explanatory variables
  - Large change in estimated coefficients, standard errors, and p-values
- Models with a significant *F*-test statistic and many or all non-significant *t*-test statistics

MLR: Model Selection

Variance Inflation Factor (VIF)

- Measures the degree to which the standard error of an estimated coefficient $\hat{\beta}_j$ is inflated by the correlations with the other explanatory variables:

$$\text{VIF}_j = \frac{1}{1 - R_j^2}$$

  where R_j^2 is the R^2 values from the MLR with response variable x_j on the remaining explanatory variables
- Explanatory variables with $\text{VIF}_j > 4$ should be investigated further
- Explanatory variables with $\text{VIF}_j > 10$ indicate severe multicollinearity

MLR: Model Selection

Solutions to Multicollinearity

- Fit model as is; don’t assess significance of individual explanatory variables
- Select only certain explanatory variables for model to remove highly correlated variables
- Rely more on theoretical or contextual basis (rather than statistical) for inclusion of variables in model

Misuses of Model Selection

Observational studies:

- Including an explanatory variable in the model does not imply a causal relationship. Wrong to claim that:
  - Included ⇒ variable *causes* change in Y
  - Omitted ⇒ variable has *no effect* on Y
  - Omitted ⇒ variable is *unimportant*
- DO NOT focus only on estimated coefficients for the selected model
  - Depends on which other variables are included in the model
  - Could be many other reasonable models

MLR: Model Selection

Misuses of Model Selection

- Overemphasis on choice of variables in model
  - e.g. repeat a study on a different population
  - Find predictors A, B, D, H in pop. 1 and predictors A, F, L, M in pop. 2
  - Is A more important?
  - Do the two populations respond differently?
- Extrapolation
  - Model provides good predictions across region of X values included in the study
  - May not be valid outside that region

MLR: Model Selection

After Selection?

- Still need to examine model assumptions (diagnostics)
- Need to examine case diagnostics
- Have we overfit to this particular data set?
  - Rule of thumb: sample size $n > 6\text{-}10 \times m$
  - If sample size is a lot fewer, e.g. 15 candidate variables, $n = 40$ obs, fitted model predicts current data well, new data poorly
  - Model validation

Unit 3

Multiple Linear Regression

Diagnostics

MLR: Diagnostics

MLR Model and Assumptions

$$\begin{aligned} Y_i &= \mu_{Y|\mathbf{x}} + \epsilon_i \text{ where } \epsilon_i \text{ i.i.d. } N(0, \sigma^2) \\ &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i \end{aligned}$$

- Observations Y_i are independent
- Values of **x** are fixed
- $\mu_{Y|\mathbf{x}}$ is a linear function of **x**
- Homogeneous error variance: $Var(\epsilon_i) = \sigma^2$
- Normally distributed errors: ϵ_i i.i.d. $N(0, \sigma^2)$

MLR: Diagnostics

Independence

- Check independence of observations through details of data collection
- Beware of
  - Observations over time
  - Clustering of observations
  - Spatial elements to observations
- Crucial assumption - must use other methods if violated

MLR: Diagnostics

Fixed Values of x

- Assume **x** is measured without error
- Check through definition of variables and through details of data collection
- If violated for some x_j, model the error in those x_j using random effects

MLR: Diagnostics

Linearity

- Scatterplot: Plot of Y versus each x_j
  - Linear patterns
- Residual Plot: Plot of residuals $\mathbf{e}$ versus each x_j
  - No patterns

Violations of Linearity

- Transform Y values so that relationship with each x_j is linear
- Transform each of the x_j variables to have linear relationship with Y
- Common transformations:
  - Power: Y^2, Y^3, $\sqrt{Y}$, etc.
  - Exponential: $\exp(Y)$, $\ln(Y)$
- Conduct analysis with transformed Y and/or $\mathbf{x}$ values
- Undo transformation in drawing conclusions

MLR: Diagnostics

Homogeneous Variance

- Residual Plots: scatterplots of residuals with predicted values $\hat{Y}$ and with each x_j
  - Look for changes in variability around the horizontal line at 0
  - Megaphone shaped pattern: variability of e increases or decreases as either $\hat{Y}$ or specific x_j increases
- Impact: confidence intervals for conditional mean and prediction intervals

MLR: Diagnostics

Violations of Homogeneous Variance

- Transform Y or x_j
- Use Weighted Least Squares

MLR: Diagnostics

Weighted Least Squares

- Assume $Var(\epsilon_i) = \sigma_i^2$ for $i = 1, \ldots, n$
- Define diagonal matrix W to have elements $w_{ii} = 1/\sigma_i^2$
- Weighted least squares estimate of β is

$$(X^T WX)^{-1} X^T WY$$

Weighted Least Squares

- Observations with smaller σ_i^2 get a larger weight in the weighted least squares estimate than observations with larger σ_i^2
- Must know or be able to estimate values of w_{ii}
  - If the *i*th observation is an average of n_i equally variable observations, then $Var(Y_i) = \sigma^2/n_i$ and $w_{ii} = n_i$.
  - If the *i*th observation is a total of n_i observations, then $Var(Y_i) = n_i\sigma^2$ and $w_{ii} = 1/n_i$.
  - If variance is proportional to some predictor x_j, then $Var(Y_i) = x_{ij}\sigma^2$ and $w_{ii} = 1/x_{ij}$.
  - In some cases, the values of the weights may be based on theory or prior research

MLR: Diagnostics

Weighted Least Squares

- The difficulty in applying weighted least squares in practice is determining the weights (estimate of error variances)
- Estimation schemes exist for estimating weights based on other characteristics (megaphone shape or upward trend in residual plots)
- Least squares and weighted least squares estimates are usually similar in value
- Differences occur with inference and prediction

MLR: Diagnostics

Normality

- Distribution of Residuals
  - Histogram of residuals
  - Normal probability plot of residuals
  - Tests for normality of residuals
- Affects inference, especially for smaller sample sizes

MLR: Diagnostics

Violations of Normality

- Remedies
  - Check for outliers
  - Transform Y
  - Conduct robust regression

MLR: Diagnostics

Model Selection Assessment

- Multiple Testing Problem
  - Adjust significance for all models considered?
  - Ignore the issue (most common practice) i.e. assume selected model is correct
  - Explore conclusions from several of the best models
  - Model averaging (using AIC or BIC)

Model Selection Assessment

- Stepwise procedures tend to overfit the sample data. Would the model perform as well in making predictions for new cases randomly selected from the population?
- Model validation: Split data into two parts
  - Training sample (perhaps 2/3 of the data)
  - Validation sample (the remainder of the data)
  - Use training sample to select model
  - Use validation sample to assess model performance and fit

Model Selection Assessment

- Compute

$$MSE_{\text{Validation}} = \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} (Y_i - \hat{Y}_i)^2$$

- Should be approximately equal to MSE_{Training} from selected model
- $MSE_{\text{Validation}}$ will be substantially larger if model is over fit to the training sample
- Use as model selection technique - fit many models to the training sample and compute $MSE_{\text{Validation}}$ for the validation sample for each selected model
- PRESS is doing this over-and-over with the size of the test sample equal to one case

Model Selection: PRESS Criterion

- PRESS (predicted residual error sum of squares)
  - Predict each response using the other $n-1$ cases to estimate the model parameters
  - $PRESS = \sum_i (Y_i - \hat{Y}_{i(-i)})^2$
  - nice idea, not used very often

MLR: Diagnostics

Case Diagnostics

- Leverage
- Outliers
- Influential Points

MLR: Diagnostics

Case Diagnostics: Leverage

- Extreme values in **x**'s are called high leverage cases because they exert a large "pull" on the fitted regression model
- Measured using the projection matrix P_X (also called the hat matrix = H)

$$\hat{\mathbf{Y}} = H\mathbf{Y} = P_X\mathbf{Y} = X(X^TX)^{-1}X^TY$$

- For an observation i, can write

$$\hat{Y}_i = \sum_{j=1}^{n} h_{ij}Y_j = h_{ii}Y_i + \sum_{j\neq i} h_{ij}Y_j$$

- h_{ii} is the (i, i) element of P_X and it is called the *leverage* of the i^{th} case

MLR: Diagnostics

Case Diagnostics: Leverage

- Properties of h_{ii}
  - Measures the extent to which the i^{th} observation dictates its own fitted value
  - $0 \leq h_{ii} \leq 1$
  - $\sum_{i=1}^{n} h_{ii} = k + 1$
  - Measures the "distance" between the vector of values for the explanatory variables for the *i*th observations and the average vector of values of explanatory variables
- Often use $2(k+1)/n$ or $3(k+1)/n$ as a guide for determining large h_{ii}
- In addition to an absolute cutoff, look for large h_{ii} by examining the distribution of h_{ii} values across cases

Case Diagnostics: Outliers

- Extreme Y_i value for a given $\mathbf{x}$
- Three assessment methods
  - Residuals
  - Internally studentized residuals
  - Externally studentized residuals

Case Diagnostics: Residuals

- Residuals

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

- $\text{Var}(e_i) = \sigma^2(1 - h_{ii})$
- Observations with higher leverage will have residuals with smaller variance
- Residual values with absolute value
  - Less than 2 are fine
  - Between 2 and 3 indicate potential outliers
  - Greater than 3 indicate outliers

MLR: Diagnostics

Case Diagnostics: Residuals

- Internally studentized residuals

$$r_i = \frac{e_i}{\sqrt{MSE(1 - h_{ii})}}$$

  - r_i will have mean zero and approximately equal variance
  - Outliers will inflate MSE

- Externally studentized residuals

$$t_i = \frac{e_i}{\sqrt{MSE_{(-i)}(1 - h_{ii})}}$$

where $\text{MSE}_{(-i)}$ is MSE without the *i*th observation

  - t_i will have mean zero and approximately equal variance

MLR: Diagnostics

Case Diagnostics: Outliers

- Outliers inflate value of $\hat{\sigma}^2$
- Will lower values of t and F test statistics
- Will inflate widths of confidence intervals for parameters and prediction intervals

MLR: Diagnostics

Case Diagnostics: Influence

- Concerned about unusual cases that have a big influence on both:
  - $\hat{Y}_i$ for some $\mathbf{x}_i$
  - regression coefficient $\hat{\beta}_j$
- Could delete the case, refit model and examine the change

MLR: Diagnostics

Case Diagnostics: Influence

- COOK'S D: effect deleting the i^{th} case on the entire set of fitted values

$$D_i = \frac{\sum_j (\hat{Y}_j - \hat{Y}_{j(-i)})^2}{(k+1)MSE} = \left(\frac{r_i^2}{k+1}\right)\left(\frac{h_{ii}}{1-h_{ii}}\right)$$

- D_i is large when r_i is large and h_{ii} is large
- There is no gold-standard for the cutoff of Cook's D
  - SAS uses $4/n$.
  - $D_i > 2 * \sqrt{2/n}$ indicates substantial influence
  - $D_i > F_{k+1,n-k-1,0.5}$ indicates substantial influence
  - Can also judge D_i relative to other D_j's

Case Diagnostics: Influence

- DFFITS$_i$ - effect of i^{th} case on fitted value for Y_i

$$\text{DFFITS}_i = \frac{\hat{Y}_i - \hat{Y}_{i(-i)}}{\sqrt{MSE_{(-i)} h_{ii}}} = t_i \sqrt{\frac{h_{ii}}{1 - h_{ii}}}$$

- $|\text{DFFITS}_i| > 2$ is considered large in small or medium sized samples
- $|\text{DFFITS}_i| > 2\sqrt{\frac{k+1}{n}}$ is considered large in big samples

MLR: Diagnostics

Case Diagnostics: **Influence**

- DFBETAS: effect of deleting the i^{th} case on the estimate of a single coefficient

$$\text{DFBETA}_{k,i} = \frac{b_k - b_{k(-i)}}{\sqrt{MSE_{(-i)} c_{kk}}}$$

- $k = 0$ for population intercept β_0
- $k = j$ for population slope β_j
- c_{kk} is (k, k) element of $(X^T X)^{-1}$
- DFBETA larger than 2 (small or medium size samples) or larger than $2n^{-1/2}$ (large samples) may be worthy of attention

Questions?

Contact me:

Email: dmommen@iastate.edu
Student Office Hours: Thursdays @ 10-11 am