



Multiple Regression

Part 3: More Than Two Predictors

STAT 705: Regression and Analysis of Variance

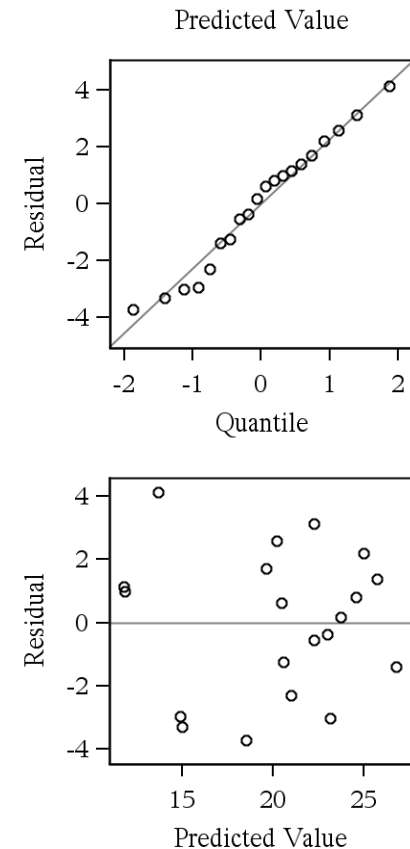
Body Fat Data, Revisited

- Data for another variable was collected on the 20 women in the sample
 - X_3 = thigh measurement
- Can we improve the model by including this variable as a predictor?
- New model: $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i$
- This is an additive model with three predictors

Body Fat Data with 3 Predictors

```
data fat;
  input triceps thigh midarm bodyfat;
  cards;
  19.5  43.1  29.1  11.9
  24.7  49.8  28.2  22.8
  . . . more data lines . . .
;
run;
proc reg data=fat;
model bodyfat = triceps midarm thigh;
plot residual.*predicted.;
run;
```

- Normal probability plot looks great
- Residual plot shows no obvious pattern
- Model assumptions appear to be satisfied



Body Fat Data with 3 Predictors

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	396.98461	132.32820	21.52	<.0001
Error	16	98.40489	6.15031		
Corrected Total	19	495.38950			

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	117.08469	99.78240	1.17	0.2578
triceps	1	4.33409	3.01551	1.44	0.1699
midarm	1	-2.18606	1.59550	-1.37	0.1896
thigh	1	-2.85685	2.58202	-1.11	0.2849

This is testing

$$H_0: \beta_1 = \beta_2 = \beta_3 = 0$$

vs. H_a : At least one of these β 's is not 0

We reject H_0 and conclude that at least one of the predictors is useful for modeling body fat.

1

These are testing whether the individual slopes are 0.

$$H_0: \beta_k = 0 \text{ vs. } H_a: \beta_k \neq 0$$

For all three slopes, we FAIL TO REJECT H_0 , so it seems that NONE of these variables are useful predictors for body fat

2

3

- F test indicates *something* is important.
- t tests indicate *nothing* is important.
- Contradiction?

Correlation Between Predictors

```
proc corr data=fat plots=matrix;  
  var triceps midarm thigh bodyfat;  
run;
```

Pearson Correlation Coefficients			
	midarm	thigh	bodyfat
triceps	0.501	0.909	0.843
midarm		0.098	0.142
thigh			0.878

We want strong correlation between Y and each of the X's

- Body fat and thigh: 0.878, very high
- Body fat and triceps: 0.843, very high
- Body fat and midarm: 0.142, weak

- But we also want the X's to be uncorrelated (i.e., correlation near 0)
 - Midarm and triceps: 0.501, acceptable
 - Midarm and thigh: 0.098, very low (good)
 - Triceps and thigh: 0.909, very high (very bad)

Multicollinearity

- Multicollinearity issues arise when two or more predictors are highly correlated ($> .7$ or so)
- The individual t tests are testing the utility of adding a new predictor *assuming the other predictors are already in the model*
- Correlation between triceps and thigh is 0.909
 - If triceps is in the model, it is already explaining a large proportion (of the variability in Y) that thigh would explain. So... there is not much use in adding thigh to the model
 - If thigh is in the model, it is already explaining a large proportion that triceps would explain. So ... there is not much use in adding thigh to the model.

Variance Inflation Factors (VIF)

- Multicollinearity is measured by variance inflation factors
- One VIF for each predictor in the model: $VIF_k = (1 - R_k^2)^{-1}$
- R_k^2 is the coefficient of determination when X_k is regressed on the remaining $p-1$ predictors
- The variance of the estimator $\hat{\beta}_k$ is a function of VIF_k
- When $R_k^2 = 0$ then $VIF_k = 1$
 - X_k is not linearly related to other predictors in the model
 - Variance is not inflated
- When $R_k^2 > 0$ then $VIF_k > 1$
 - Variances are inflated due to correlation between predictors
- **$VIF_k > 10$ implies serious multicollinearity issues**

VIF in the Body Fat Data

- In SAS, use the option 'VIF' on the model statement

```
proc reg data=fat;  
model bodyfat = triceps midarm thigh / vif;  
run;
```

- Parameter Estimates table has an extra column

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t	Variance Inflation
Intercept	1	117.08469	99.78240	1.17	0.2578	0
triceps	1	4.33409	3.01551	1.44	0.1699	708.84291
midarm	1	-2.18606	1.59550	-1.37	0.1896	104.60601
thigh	1	-2.85685	2.58202	-1.11	0.2849	564.34339

- There are SERIOUS multicollinearity issues in this data set
 - All three VIF's are much greater than 10
- If all three predictors are used, the inference is invalidated by multicollinearity

Remedies for Multicollinearity

- Re-think the research objectives and choose predictors accordingly
- Consider working with composite predictors (e.g. Principal Component Analysis)
 - Drawback: These can be difficult to interpret
- Center the predictors
 - For each predictor, subtract the mean from the observed value
 - Use the difference instead of the original value

VIF vs. Correlation

- Correlation
 - Measures the strength of the linear association between two variables
 - Does not consider any other variables
- Variance Inflation Factor
 - Measures the strength of the linear association between each predictor variable and all the other predictors in the model
 - Is able to detect linear associations that correlation does not find

Which Model?

- There are several models we could create using the three predictors
- We should not include both thigh and triceps in the same model
- Consider models with these predictors
 - Model 1: midarm
 - Model 2: thigh
 - Model 3: triceps
 - Model 4: miarm and triceps
 - Model 5: midarm and thigh

Criteria for Comparing Models

- Principle of parsimony
 - We want to use the simplest model that accurately describes the data
 - We want as few predictors as possible
- MSE
 - Is an estimate for the error variance σ^2
 - Smaller is better
 - Could also use RMSE, the square root of MSE
- R^2
 - Proportion of variability in Y that is explained by the regression model
 - Larger is better
- There will be other criteria later

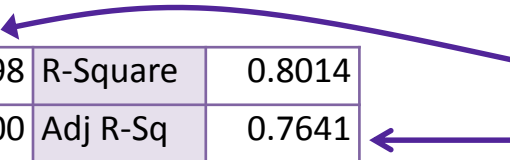
Adjusted R^2

- If we add another predictor to the model, R^2 will get bigger (or stay the same)
 - We want R^2 to be large
 - We also want as few predictors as feasible
 - This is a contradiction
- Solution: Adjusted R^2
 - Adjust the value of R^2 to account for the number of predictors
 - In essence, penalize (reduce) R^2 for models that have a large number of predictors

$$R_{\text{adj}}^2 = 1 - \left(\frac{n-1}{n-p-1} \right) (1 - R^2), \text{ where } p \text{ is the number of predictors}$$

Reading SAS Output

- Look immediately after the ANOVA table



Root MSE	2.47998	R-Square	0.8014
Dependent Mean	20.19500	Adj R-Sq	0.7641
Coeff Var	12.28017		

RMSE (want small)

R^2_{adj} (want large)

- For our five candidate models
 - All of the diagnostic plots are okay, so assumptions appear to be satisfied. (You should run the provided code and verify this.)
 - Criteria are given on the next slide

Five Candidate Models

Predictors in the Model					
Criteria	midarm	thigh	triceps	midarm & triceps	midarm & thigh
MSE	26.96	6.301	7.951	6.231	6.536
R ²	0.0203	0.7710	0.7111	0.7862	0.7757
R ² _{adj}	-0.0341	0.7583	0.6950	0.7610	0.7493

- Worst model: midarm as the only predictor
 - R² is much smaller than others ... and ... R²_{adj} is actually negative!
 - Do not use this model !!!
- The other models are very similar
 - For a one-predictor model, thigh is slightly better than triceps (smaller MSE and higher adj.R²)
 - For a two-predictor model, midarm and triceps is slightly better than midarm and thigh

Common Sense

- MSE and R^2_{adj} are guidelines for selecting a model
- Check the diagnostic plots for each candidate model
 - The assumptions must be verified
- Do not let a minor improvement in these measures dictate which model to use
- For the body fat data, we can
 - Exclude the model with only midarm
 - The other four models have similar MSE and R^2_{adj}
 - Choose the model that will be the easiest to implement and explain

What You Should Know

- Recognize when a data set has multicollinearity issues
- Mitigate the effect of multicollinearity
- Compare models using MSE, RMSE, and/or adjusted R^2
- Interpret F test and t tests