Multiple Regression
Part 2: Body Fat Example

STAT 705: Regression and Analysis of Variance

Example

- Continuing with the example from last lesson
  - Y = Body fat (in points)
  - X_1 = triceps skinfold thickness (in mm)
  - X_2 = midarm circumference (in cm)
- Random sample of n = 20 females
- Model: $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i \quad i = 1, \ldots n$
- Use SAS to fit the model

Fit the Model with SAS

```
- - - - - - - - - - - - - - - - - - - -

... data step goes here ...

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

title 'Regression analysis of bodyfat
       on skinfold thickness and
       midarm circumference';

proc reg data=fat;
  model bodyfat = triceps midarm /
    clb covb p clm cli ;
   plot residual.*predicted.;
run;

- - - - - - - - - - - - - - - - - - - -
```

Full SAS program is in the file MultReg.BodyFatExample.sas

PROC CORR: Examine the strength of the relationship between the response (body fat) and the two predictors.

Title statements are optional, but recommended.

PROC REG:
Options on model statement are after the slash.

- **'clb'** for confidence limits on the betas.
- **'covb'** for the variance-covariance matrix of the beta hats
- **'p'** for predicted (fitted) values
- **'clm'** for confidence limits for the mean response
- **'cli'** for confidence limits for an individual response

Plot statement generates diagnostic plots.

PROC CORR Output

Correlation matrix is always symmetric, i.e., the lower left entries are the mirror image of the upper right entries. Some software will print only half of this matrix.

Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	triceps	midarm	bodyfat
triceps	1.00000 21	0.50102 0.0207 21	0.84327 <.0001 20
midarm	0.50102 0.0207 21	1.00000 21	0.14244 0.5491 20
bodyfat	0.84327 <.0001 20	0.14244 0.5491 20	1.00000 20

Each cell contains

(1) sample correlation coefficient (Pearson's)

(2) p-value for testing H_0: $\rho = 0$ vs. H_a: $\rho \neq 0$

(3) number of observations used in calculation

Diagonal entries are always 1
(This is the correlation of a variable with itself.)

PROC CORR Output

Pearson Correlation Coefficients Prob > |r| under H0: Rho=0 Number of Observations			
	triceps	**midarm**	**bodyfat**
triceps	1.00000 21	0.50102 0.0207 21	0.84327 <.0001 20
midarm	0.50102 0.0207 21	1.00000 21	0.14244 0.5491 20
bodyfat	0.84327 <.0001 20	0.14244 0.5491 20	1.00000 20

We are concerned with three correlations.

(1) predictor 'triceps' and response 'bodyfat' highest correlation at 0.84

(2) predictor 'midarm' and response 'bodyfat' correlation 0.14 (pretty low)

(3) between two predictors: correlation 0.50

If the correlation between the predictors is greater than 0.7 (or so), we can have problems with the fitted model. This should not be an issue with the current data.

Interpret Correlation

- Predictor triceps is more strongly correlated with body fat than is midarm (0.84 vs. 0.14)
  - We should definitely include triceps in the model
- Even though midarm is weakly correlated with body fat (0.14), it may still be useful in the model
- This is consistent with information in the scatterplot matrix (from the last lesson)

ANOVA Table for Regression

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	p	SSReg	SSReg / dfReg	MSReg / MSE	
Error	n – p – 1	SSE	SSE / dfE		
Corrected Total	n – 1	SSTotal			

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	2	389.46	194.73	31.25	<.0001
Error	17	105.93	6.23		
Corrected Total	19	495.39			

- Similar to the ANOVA tables we have already seen.
- These relationships are all still true:
  - dfReg + dfE = dfTotal
  - SSReg + SSE = SSTotal
  - MS = SS / df
  - F = MSReg / MSE
  - SSTotal = (n – 1) Var(Y)
  - SSE = sum of squared residuals
- dfReg is "DF Model" in the table
- p is number of predictors

How are they different? In simple linear regression, p = 1. Now p > 1.

Information in the ANOVA Table

Model: $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i \quad i = 1, \ldots n$

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	2	389.46	194.73	31.25	<.0001
Error	17	105.93	6.23		
Corrected Total	19	495.39			

SSE = $\sum_{i=1}^{n} r_i^2$

Point estimate of the error variance

MSE = $\hat{\sigma}^2$

These are the F statistic and p-value for testing

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

vs. $H_a: \beta_1 \neq 0$ or $\beta_2 \neq 0$

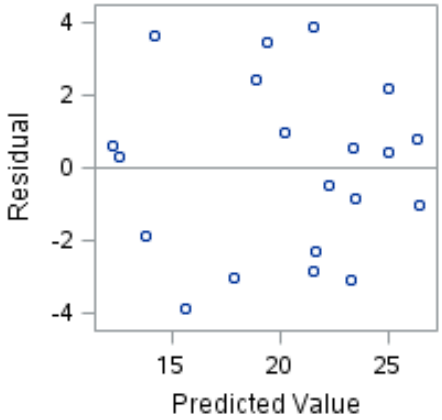
If the assumptions are violated, this test NOT VALID and should be ignored.

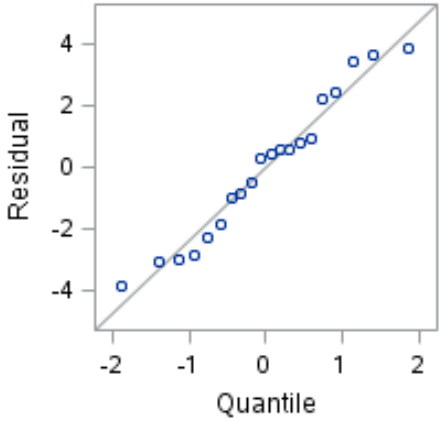
If H_0 is true, then the model becomes

$Y_i = \beta_0 + \varepsilon_i$

This is called the "reduced" model.

Check the Assumptions





- Do this **before** interpreting any of the other output
- Residual plot (on the top) is okay
  - Points show no obvious pattern
- Normal probability plot (on the bottom) is okay
  - Points follow the line
  - No evidence that residuals are not normal
- Assumptions appear to be satisfied
- We can proceed with the analysis

Parameter Estimates Table

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	6.79163	4.48829	1.51	0.1486	-2.67783	16.26109
triceps	1	1.00058	0.12823	7.80	<.0001	0.73004	1.27113
midarm	1	-0.43144	0.17662	-2.44	0.0258	-0.80407	-0.05882

Each row in the table is estimating and testing one parameter in the model:

$$(\text{BodyFat})_i = \beta_0 + \beta_1(\text{Triceps})_i + \beta_2(\text{Midarm})_i + \varepsilon_i$$

For Triceps: From the above table $\hat{\beta}_1 = 1.00$ and $se\left(\hat{\beta}_1\right) = 0.128$

- $\hat{\beta}_1$ follows a t distribution with df = dfE $= 17$
- Critical value is 2.110 (from t table, α / 2 = 0.025)
- 95% Conf. Int. for β is $1.00 \pm (2.110)(0.128) \Rightarrow 1.00 \pm 0.27 \Rightarrow (0.73, 1.27)$
- The columns 't Value' and 'Pr > | t | ' are testing the single parameter

  $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$

t Tests and F test

Parameter Estimates			
Variable	**. . .**	**t Value**	**Pr > |t|**
Intercept	. . .	1.51	0.1486
triceps	. . .	7.80	<.0001
midarm	. . .	-2.44	0.0258

Analysis of Variance			
Source	**. . .**	**F Value**	**Pr > F**
Model	. . .	31.25	<.0001
Error	. . .		
Corrected Total	. . .		

- Parameter Estimates table reports individual t tests
  - Intercept: $H_0: \beta_0 = 0$ vs. $H_a: \beta_0 \neq 0$; t = 1.51, p-value = 0.1486
  - Triceps: $H_0: \beta_1 = 0$ vs. $H_a: \beta_1 \neq 0$; t = 7.80, p-value < 0.0001
  - Midarm: $H_0: \beta_2 = 0$ vs. $H_a: \beta_2 \neq 0$; t = -2.44, p-value = 0.0258
- Analysis of Variance table reports overall F test
  - $H_0: \beta_1 = 0$ and $\beta_2 = 0$ vs. $H_a: \beta_1 \neq 0$ or $\beta_2 \neq 0$
  - F = 31.25, p-value < 0.0001
- Where β_0, β_1, and β_2 are as specified in the model

Interpret Hypothesis Tests

- After verifying the assumptions, look at F test
  - p-value < 0.0001
  - Reject null hypothesis that both β_0 and β_1 are 0
  - We conclude that there is a significant relationship between body fat and at least one of the predictors
- F test is significant, so use the individual t tests to decide which predictors are significant

Inference on Individual Coefficients

- F test is significant, so look at the t tests
  - Triceps: p-value < 0.0001 ⇨ Conclude $H_a: \beta_1 \neq 0$
  - Midarm: p-value = 0.0258 ⇨ Conclude $H_a: \beta_2 \neq 0$
- Conclude both triceps and midarm are significant
- Both of these t tests are marginal (or partial) tests
  - Given that the other predictor(s) are already in the model, the t test evaluates the additional contribution of the variable being tested

Interpret the Intercept

- Least Squares solution:

  E(body fat) = 6.79 + 1.00(triceps) – 0.43(midarm)

- β_0 is the intercept and represents E(Y) when $X_1 = X_2 = 0$.
  - $\hat{\beta}_0 = 6.79$
  - For a woman who has triceps = 0 mm and midarm = 0 cm, we expect her body fat to be 6.79 points
  - This does make any sense – it is not possible to have triceps = 0 or midarm = 0

(It is often the case that the intercept is nonsensical, but we usually keep it in the model unless there is a good reason to take it out.).

Interpret the Slopes

- The slope for the k^{th} variable is the change in E(Y) for each unit increase in the variable, **while keeping all the other *X's* constant**
- $\hat{\beta}_1$ = 1.00
  - ⇨ If the triceps measurement increases by 1 mm **and the midarm stays the same**, the expected body fat increases by 1.00 points
- $\hat{\beta}_2$ = –0.43
  - ⇨ If the midarm increases by 1 cm **and the triceps stays the same**, the expected body fat decreases by 0.43 points

Interval Estimation

- We have concluded
  (1) model assumptions are satisfied , and
  (2) both predictors are significant
- Now we can use the model for prediction and estimation
- Two quantities of interest
  - Both quantities require that we specify a value for every predictor
  - The mean response is the average value of the response for all items in the population that have the specified values of the predictors
  - The individual predicted value is the value for **one** of the items in the population that has the specified values of the predictors

Point Estimate

- A single value that estimates the quantity of interest
  - for mean response: point estimate is $\hat{Y}$
  - for individual response: point estimate is $\hat{Y}$
  - (this is not a typo – the point estimates are the same)

Interval Estimates

- Are a range of values that estimate the quantity of interest
- Are of the form: (point estimate) $\pm$ (critical value)*(SE)
- Critical value is from the t distribution with df = dfE
- SE is the standard error of the point estimate
  - SE is different for the two different quantities
  - SE for an individual response is larger because an individual value fluctuates more than the mean value

Confidence vs. Prediction Intervals

- Mean response
  - interval estimate is a confidence interval
- Individual response
  - interval estimate is a prediction interval
- In SAS
  - option ‘clm’ to get confidence limits for mean
  - option ‘cli’ to get confidence limits for individual value
  - add a line to the data set (see next slide)

Partial SAS Code

```
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 here ...


  25.2  51.0  27.5  21.1
  20.0  50.0  20.0     .
;
proc reg data=fat;
  model bodyfat = triceps midarm / clm cli ;
run;
```

This line is for estimation and prediction based on triceps measurement 20 mm and midarm measurement 20 cm. We are currently not using the thigh measurement of 50.

The single period at the end of the line indicates a missing value for the response. SAS does not use observations with missing values, so this data line is not used to fit the model.

Confidence limits for the mean response (clm) and for individual values (cli).

Partial SAS Output

Output Statistics								
Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict		Residual
1	11.9000	13.7481	1.0546	11.5231	15.9731	8.0307	19.4655	-1.8481
2	22.8000	19.3394	0.5791	18.1176	20.5612	13.9329	24.7460	3.4606
. . .	. . . . .	. . . . .	. . . . .	. . . . .	. . . . .	. . . . .	. . . . .	. . . . .
20	21.1000	20.1417	0.5585	18.9633	21.3201	14.7448	25.5386	0.9583
21	.	18.1745	1.3219	15.3856	20.9634	12.2150	24.1340	.

- Among all women with midarm 20 cm and triceps 20 mm, the mean body fat is in the interval (15.3856, 20.9634), with 95% confidence.
- For a single woman with midarm 20 cm and triceps 20 mm, we predict her body fat is in the interval (12.2150, 24.1340), with probability 95%.
- Note that the individual prediction interval is wider than the mean confidence interval. This will always be true.

What You Should Know

- Write the SAS code to fit a regression model
- Interpret the estimated model coefficients
- For <u>each</u> test that is automatically generated by SAS
  - Write the hypotheses (H_0 and H_a)
  - Interpret the results
- For confidence intervals and prediction intervals
  - Know the difference between them
  - Be able to generate them in SAS and interpret the results