# STOR 455
# STATISTICAL METHODS I

Jan Hannig

# Recall Extra SS (Section 4.9)

- $SSE(X_1, X_2, X_3, X_4, X_5)$ is the SSE for the *full* model
- $SSE(X_1, X_2, X_3)$ is the SSE for the *reduced* model
- $SSR(X_4, X_5 \mid X_1, X_2, X_3)$ is the difference

$$SSR(X_4, X_5 \mid X_1, X_2, X_3) =$$
$$= SSE(X_1, X_2, X_3) - SSE(X_1, X_2, X_3, X_4, X_5)$$
$$= SSR(X_1, X_2, X_3, X_4, X_5) - SSR(X_1, X_2, X_3)$$

# F test

- Numerator is relevant partial MSR
- Denominator is MSE of the full model
- $F \sim F(df_R, df_E)$
- Reject if the P value is small; in which case conclude that the extra variables are important.

# F-test in SAS

```
*F-test;
proc reg data=fat;
   model fat=skinfold thigh midarm;
   test1: test thigh, midarm;
run;
```

Test test1 Results for Dependent Variable fat

| Source | DF | Mean Square | F Value | Pr > F |
|---|---|---|---|---|
| Numerator | 2 | 22.35741 | 3.64 | 0.0500 |
| Denominator | 16 | 6.15031 | | |

# Type I and II SS

- *Type I SS*
  - $SSR(X_1)$
  - $SSR(X_2 \mid X_1)$
  - $SSR(X_3 \mid X_1, X_2)$
  - $SSR(X_4 \mid X_1, X_2, X_3)$
- *Type II SS (sometimes called type III)*
  - $SSR(X_1 \mid X_2, X_3, X_4)$
  - $SSR(X_2 \mid X_1, X_3, X_4)$
  - $SSR(X_3 \mid X_1, X_2, X_4)$
  - $SSR(X_4 \mid X_1, X_2, X_3)$

*F test*

$$F = (SSR/1) / MSE(full) \sim F(1, n-p)$$

# Do it in SAS

```
*Type I and Type II sums of squares;
proc reg data=fat;
  model fat=skinfold thigh midarm /ss1 ss2;
run;
```

# Do it in SAS

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 | | | |

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

Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| | Type I SS | Type II SS |
|---|---|---|---|---|---|---|---|
| Intercept | 1 | 117.08469 | 99.78240 | 1.17 | 0.2578 | 8156.76050 | 8.46816 |
| skinfold | 1 | 4.33409 | 3.01551 | 1.44 | 0.1699 | 352.26980 | 12.70489 |
| thigh | 1 | -2.85685 | 2.58202 | -1.11 | 0.2849 | 33.16891 | 7.52928 |
| midarm | 1 | -2.18606 | 1.59550 | -1.37 | 0.1896 | 11.54590 | 11.54590 |

# Do it in SAS

```
proc glm data=fat;
   model fat=skinfold thigh midarm;
run;
```

(Prints type I and type III SS together with the corresponding F tests.)

# Do it in SAS

The GLM Procedure Dependent Variable: fat

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 3 | 396.9846118 | 132.3282039 | 21.52 | <.0001 |
| Error | 16 | 98.4048882 | 6.1503055 | | |
| Corrected Total | 19 | 495.3895000 | | | |

| R-Square | Coeff Var | Root MSE | fat Mean |
|---|---|---|---|
| 0.801359 | 12.28017 | 2.479981 | 20.19500 |

| Source | DF | Type I SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| skinfold | 1 | 352.2697968 | 352.2697968 | 57.28 | <.0001 |
| thigh | 1 | 33.1689128 | 33.1689128 | 5.39 | 0.0337 |
| midarm | 1 | 11.5459022 | 11.5459022 | 1.88 | 0.1896 |

# Do it in SAS

| Source | DF | Type III SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| skinfold | 1 | 12.70489278 | 12.70489278 | 2.07 | 0.1699 |
| thigh | 1 | 7.52927788 | 7.52927788 | 1.22 | 0.2849 |
| midarm | 1 | 11.54590217 | 11.54590217 | 1.88 | 0.1896 |

| Parameter | Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|
| Intercept | 117.0846948 | 99.78240295 | 1.17 | 0.2578 |
| skinfold | 4.3340920 | 3.01551136 | 1.44 | 0.1699 |
| thigh | -2.8568479 | 2.58201527 | -1.11 | 0.2849 |
| midarm | -2.1860603 | 1.59549900 | -1.37 | 0.1896 |

# Partial correlations

- Measure the strength of a linear relation between two variables taking into account (or conditioning on) other variables.
- *Coefficient of partial determination:* squared partial correlation
- $r^2_{Y1,234}$=*SSR* ($X_1$ |$X_2$, $X_3$, $X_4$)/SSE($X_2$, $X_3$, $X_4$)
  – Different than the F statistics.
  – SAS option /PCORR1 /PCORR2

# Do it in SAS

```
*Partial correlations;
proc reg data=fat;
   model fat=skinfold thigh midarm / pcorr1 pcorr2;
run;
```

Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| | Squared Partial Corr Type I | Squared Partial Corr Type II |
|---|---|---|---|---|---|---|---|
| Intercept | 1 | 117.08469 | 99.78240 | 1.17 | 0.2578 | . | . |
| skinfold | 1 | 4.33409 | 3.01551 | 1.44 | 0.1699 | 0.71110 | 0.11435 |
| thigh | 1 | -2.85685 | 2.58202 | -1.11 | 0.2849 | 0.23176 | 0.07108 |
| midarm | 1 | -2.18606 | 1.59550 | -1.37 | 0.1896 | 0.10501 | 0.10501 |

# Standardized Regression Model

- Numerical problem: roundoff errors
- Interpretation problems: is big coefficient really big?
- $Y = \ldots + \beta X + \ldots$

  $= \ldots + (\beta(s_X/s_Y))\ \ (s_Y)\ \ (X/s_X) + \ldots$
- SAS option /stb
- Choose unit for $X_i$ wisely

# Do it in SAS

```
*Standardized regression;
proc reg data=fat;
   model fat=skinfold thigh midarm / stb;
run;
```

Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| | Standardized Estimate |
|---|---|---|---|---|---|---|
| Intercept | 1 | 117.08469 | 99.78240 | 1.17 | 0.2578 | 0 |
| skinfold | 1 | 4.33409 | 3.01551 | 1.44 | 0.1699 | 4.26370 |
| thigh | 1 | -2.85685 | 2.58202 | -1.11 | 0.2849 | -2.92870 |
| midarm | 1 | -2.18606 | 1.59550 | -1.37 | 0.1896 | -1.56142 |

# Productivity Example

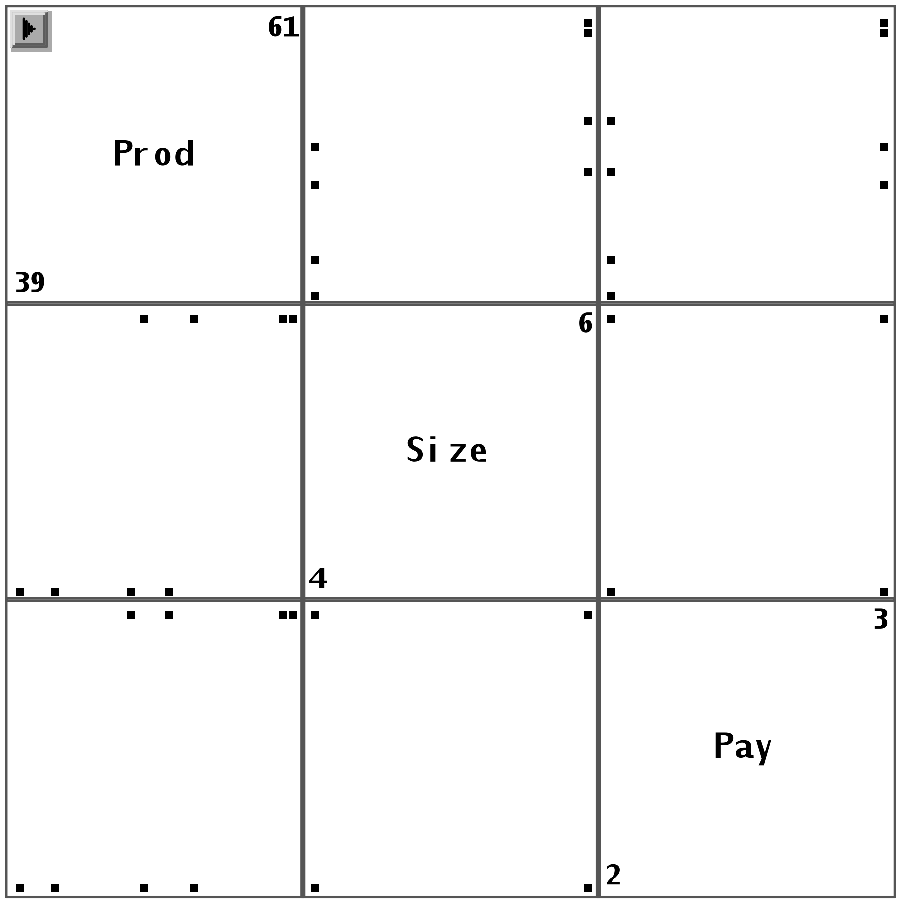
- Response Variable: Productivity
- Explanatory Variable: Crew size and amount of bonus pay
- Designed experiment
- Size and Pay are uncorrelated by design

# Do it in SAS

```
data crewprod;
  infile 'T:\...
  \Ch07ta06.txt';
  input Size Pay Prod;
run;
%include "T:\...
  \scatter.sas";
%scatter(data = crewprod,
  var =Prod Size Pay);
proc print data=crewprod;
run;
```

| Obs | Size | Pay | Prod |
|---|---|---|---|
| 1 | 4 | 2 | 42 |
| 2 | 4 | 2 | 39 |
| 3 | 4 | 3 | 48 |
| 4 | 4 | 3 | 51 |
| 5 | 6 | 2 | 49 |
| 6 | 6 | 2 | 53 |
| 7 | 6 | 3 | 61 |
| 8 | 6 | 3 | 60 |

# Do it in SAS



# Do it in SAS

```
*uncorrelated explanatory variables;
proc reg data = crewprod;
  model Prod = Size Pay /ss1 ss2;
  model Prod = Size;
  model Prod = Pay;
run;
```

Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| | Type I SS | Type II SS |
|---|---|---|---|---|---|---|---|
| Intercept | 1 | 0.37500 | 4.74045 | 0.08 | 0.9400 | 20301 | 0.02206 |
| Size | 1 | 5.37500 | 0.66380 | 8.10 | 0.0005 | 231.12500 | 231.12500 |
| Pay | 1 | 9.25000 | 1.32759 | 6.97 | 0.0009 | 171.12500 | 171.12500 |

# Do it in SAS

Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|---|
| Intercept | 1 | 23.50000 | 10.11136 | 2.32 | 0.0591 |
| Size | 1 | 5.37500 | 1.98300 | 2.71 | 0.0351 |

Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|---|
| Intercept | 1 | 27.25000 | 11.60774 | 2.35 | 0.0572 |
| Pay | 1 | 9.25000 | 4.55293 | 2.03 | 0.0885 |

# Uncorrelated Explanatory Variables

- Simpler interpretation of regression coefficients
- Type I SS same as Type II SS
- Easier to add/remove variables in the model
- Possible in controlled experiments

# Multicollinearity (Section 5.5)

- Strong linear correlation between explanatory variables
- Example: skin fold thickness and thigh circumference; amount of rainfall and hours of sunshine; SAT math, SAT veb and SAT total score
- Numerical and statistical problem

# Body fat example

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 | | | |

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

Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| | Type I SS | Type II SS |
|---|---|---|---|---|---|---|---|
| Intercept | 1 | 117.08469 | 99.78240 | 1.17 | 0.2578 | 8156.76050 | 8.46816 |
| skinfold | 1 | 4.33409 | 3.01551 | 1.44 | 0.1699 | 352.26980 | 12.70489 |
| thigh | 1 | -2.85685 | 2.58202 | -1.11 | 0.2849 | 33.16891 | 7.52928 |
| midarm | 1 | -2.18606 | 1.59550 | -1.37 | 0.1896 | 11.54590 | 11.54590 |

# Body fat example

- The P value for $F(3, 16)$ is $<.0001$
- But the P values for the individual regression coefficients are 0.1699, 0.2849, and 0.1896
- None of these are near our standard of 0.05
- What is the explanation?

# Do it in SAS

```
*compare different model informally;
proc reg data=fat;
   model fat=skinfold;
   model fat=thigh;
   model fat=skinfold thigh;
   model fat=skinfold midarm;
   model fat=thigh midarm;
   model fat=skinfold thigh midarm;
run;
```

# Do it in SAS

Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|---|
| Intercept | 1 | -1.49610 | 3.31923 | -0.45 | 0.6576 |
| skinfold | 1 | 0.85719 | 0.12878 | 6.66 | <.0001 |

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|---|
| Intercept | 1 | -23.63449 | 5.65741 | -4.18 | 0.0006 |
| thigh | 1 | 0.85655 | 0.11002 | 7.79 | <.0001 |

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|---|
| Intercept | 1 | -19.17425 | 8.36064 | -2.29 | 0.0348 |
| skinfold | 1 | 0.22235 | 0.30344 | 0.73 | 0.4737 |
| thigh | 1 | 0.65942 | 0.29119 | 2.26 | 0.0369 |

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|---|
| Intercept | 1 | 6.79163 | 4.48829 | 1.51 | 0.1486 |
| skinfold | 1 | 1.00058 | 0.12823 | 7.80 | <.0001 |
| midarm | 1 | -0.43144 | 0.17662 | -2.44 | 0.0258 |

# Do it in SAS

Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|---|
| Intercept | 1 | -25.99695 | 6.99732 | -3.72 | 0.0017 |
| thigh | 1 | 0.85088 | 0.11245 | 7.57 | <.0001 |
| midarm | 1 | 0.09603 | 0.16139 | 0.60 | 0.5597 |

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

R-Square 0.71, 0.77, 0.78, 0.79, 0.78, 0.80

# Body fat example

- Fit models including different variables
- Regression coefficients for the same variable change dramatically in different models
- Including more variables in the model increases $R^2$

# Effects of multicollinearity

- Numerical problem: X'X is close to singular, difficult to invert accurately
- Type I SS and Type II SS will differ
- Type II SS and t-test may be misleading
- $R^2$ and predicted values are usually ok

# Multicollinearity (2)

- Regression coefficients and its standard error can not be well estimated
- Difficult to interpret the regression coefficients
- Redundancy in the model

# Multicollinearity (3)

- Extreme cases can help us to understand the problem
- Scatter plot and pair-wise correlation can detect some collinearity but not all
- More diagnostic and remedy of collinearity in Section 5.5

# Variance Inflation Factor

- VIF = $1/(1 - R^2_k)$
- $R^2_k$ is the squared multiple correlation obtained in a regression where all other explanatory variables are used to predict $X_k$
- One suggested rule: a value of 10 or more indicates excessive multicollinearity
- Tolerance: TOL=1/VIF=$(1 - R^2_k)$

# Body fat example

```
* check collinearity using VIF/TOL;
proc reg data = fat;
  model fat = skinfold thigh midarm /
  VIF TOL;
run;
```

Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| | Tolerance | Variance Inflation |
|---|---|---|---|---|---|---|---|
| Intercept | 1 | 117.08469 | 99.78240 | 1.17 | 0.2578 | . | 0 |
| skinfold | 1 | 4.33409 | 3.01551 | 1.44 | 0.1699 | 0.00141 | 708.84291 |
| thigh | 1 | -2.85685 | 2.58202 | -1.11 | 0.2849 | 0.00177 | 564.34339 |
| midarm | 1 | -2.18606 | 1.59550 | -1.37 | 0.1896 | 0.00956 | 104.60601 |

# Regression Diagnostics Recommendations

- Examine the tolerance/VIF for each X
- If there are variables with low tolerance, you need to do some model building
  - Recode variables
  - Variable selection (coming later)