Lecture Outline (week 10)

Polynomial Regression Models
(quadratic models, Partial F tests, Type 1 SS, Residuals, Partial Correlation)
Multiple Regression Models
Models with One Qualitative and One Quantitative Variables

Polynomial Regression Models
(quadratic models, Partial F tests, Type 1 SS)

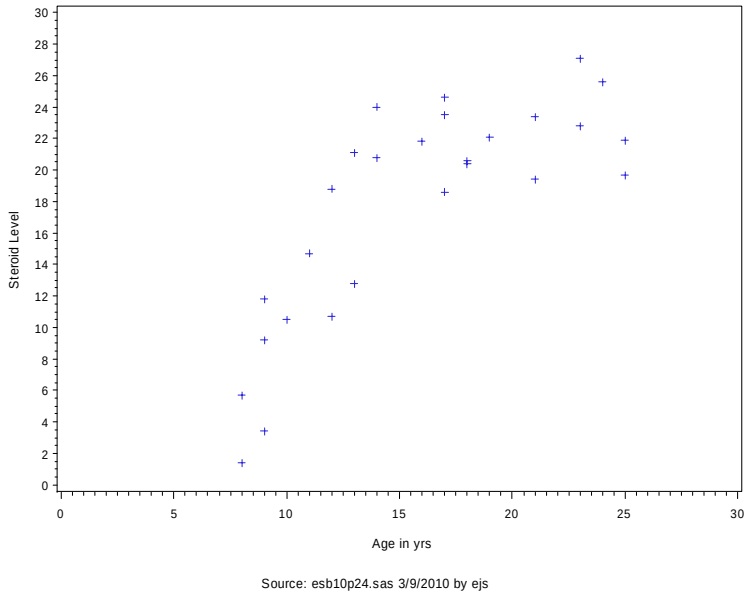
Example 1: Steriod Levels among Women age 8-25

(Problem 8 (p336) Chapter 8 (Kutner, Nachtsheim, neter, and Li (5th edition: Applied Linear Statistical Models) (program esb10p24.sas, esb10p25.sas))

Construct Scatter Plots with Independent Variables

```
FOOTNOTE "&prg";
AXIS1  LABEL=(ANGLE=90  " Steroid Level " ROTATE=0) ORDER=0 to 30 by 2;
* Defines label for y axis;
AXIS2  LABEL=( " Age in yrs" ROTATE=0) ORDER=0 to 30 by 5  ;
PROC GPLOT DATA=d2;
  PLOT steroid*age/ vaxis=axis1 haxis=axis2;
  TITLE1 "Figure 1.  Scatter plot of Steroid by age ";
  RUN;
```

Figure 1. Scatter plot of Steroid by age



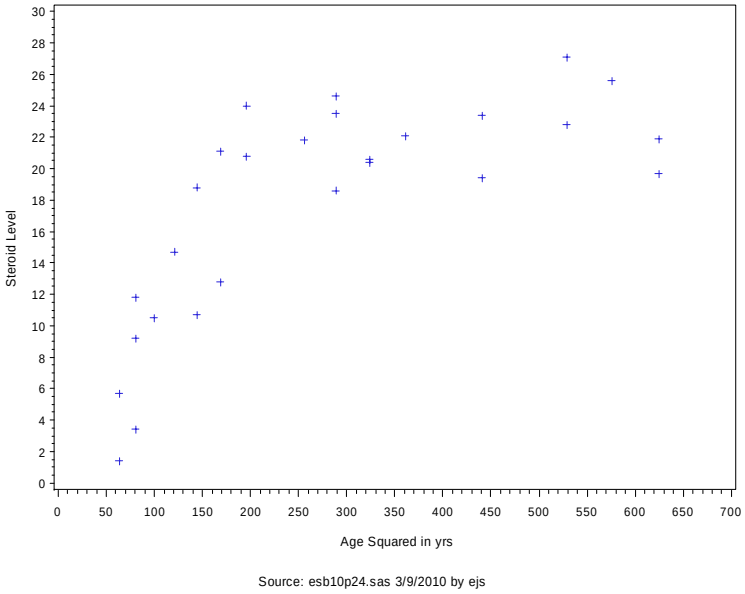
```
DATA d2;
  SET d1;
  age2=age*age;
  RUN;
PROC SORT DATA=d2;
  BY age;
RUN;
PROC PRINT DATA=d2 (OBS=10) NOOBS;
  VAR steroid age age2;
  TITLE2 "Table 1.  Example of Data for Steroids";
  RUN;
```

Table 1. Example of Data for Steroids

steroid	age	age2
1.4	8	64
5.7	8	64
9.2	9	81
11.8	9	81
3.4	9	81
10.5	10	100
14.7	11	121
10.7	12	144
18.8	12	144
21.1	13	169

Source: esb10p24.sas 3/9/2010 by ejs

Figure 2. Scatter plot of Steroid by age squared



Source: esb10p24.sas 3/9/2010 by ejs

Notes:

This scatter plot doesn't look like a straight line would provide a good fit. Still, it appears that a line would be better than a horizontal line.

This scatter plot does not account for what might be explained by a linear regression with age.

Evaluate Correlation of Steroids with age, age squared.

```
PROC CORR DATA=d2;
  VAR steroid age age2;
  TITLE2 "Table 3.  Correlation of Steroids with other variables";
RUN;
```

```
Table 3.  Correlation of Steroids with other variables

The CORR Procedure

   3  Variables:    steroid  age      age2
```

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
steroid	27	17.64444	7.02963	476.40000	1.40000	27.10000
age	27	15.77778	5.50058	426.00000	8.00000	25.00000
age2	27	278.07407	181.68060	7508	64.00000	625.00000

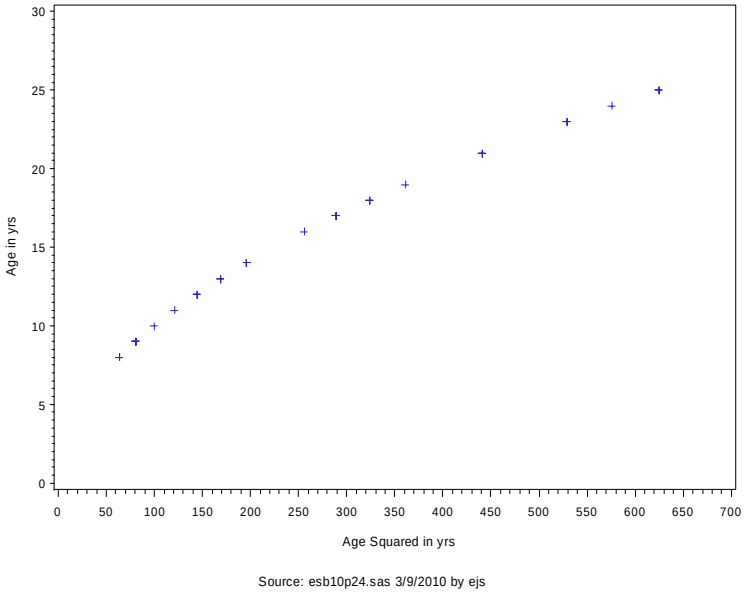
Pearson Correlation Coefficients, N = 27
Prob > |r| under H0: Rho=0

	steroid	age	age2
steroid	1.00000	0.78577 <.0001	0.71312 <.0001
age	0.78577 <.0001	1.00000	0.98943 <.0001
age2	0.71312 <.0001	0.98943 <.0001	1.00000

```
Source: esb10p24.sas 3/9/2010 by ejs
```

Notes:
There is a large correlation of age with age squared, but not a perfect correlation. When two variable are highly correlated, they are called 'colinear'. A scatter plot illustrates this.

Figure 3. Scatter plot of Age by Age Squared



Fit Regression Models

Model 1: $Y_i = \beta_0 + X_{1i}\beta_1 + E_i$ X_{1i}=Age

Model 2: with polynomial regression, only consider hierarchical models (models where the lower order polynomial terms are included). This means that we would not consider fitting a model with only a quadratic term.

Model 3: $Y_i = \beta_0 + X_{1i}\beta_1 + X_{2i}\beta_2 + E_i$ X_{1i}=Age X_{2i}=Age squared

Table 4. Regression of Steroids with age (Model 1)

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	793.28051	793.28051	40.35	<.0001
Error	25	491.52616	19.66105		
Corrected Total	26	1284.80667			

Root MSE	4.43408	R-Square	0.6174
Dependent Mean	17.64444	Adj R-Sq	0.6021

Table 6. Regression of Steroids with age and age squared (Model 3)

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	1046.26586	523.13293	52.63	<.0001
Error	24	238.54081	9.93920		
Corrected Total	26	1284.80667			

Root MSE	3.15265	R-Square	0.8143
Dependent Mean	17.64444	Adj R-Sq	0.7989

Notes:
Corrected Total Sums of squares is the same in all models.

The Adjusted R-square is largest with Model 3.

Comparing Models

Comparison of Model 3 with Model 1:

Model 1: $Y_i = \beta_0 + X_{1i}\beta_1 + E_i$ X_{1i}=Age

Model 3: $Y_i = \beta_0 + X_{1i}\beta_1 + X_{2i}\beta_2 + E_i$ X_{1i}=Age X_{2i}=Age squared

Null Hypothesis: There is no difference between Model 3 and Model 1

or $H_0 : \beta_2 = 0$

Alternative Hypothesis: $H_a : \beta_2 \neq 0$

To test this Hypothesis, we use a Partial F-test

Extra Sum of Squares

Define: Regression sum of Squares:

$SSR(X_1)$ = sums of squares explained by including X_1

Examples:

For Model 1: $SSR(X_1)$=793.28 df=1

For Model 3: $SSR(X_1, X_2)$=1046.26 df=2

Extra sum of squares:

$$SSR(X_2 \mid X_1) = SSR(X_1, X_2) - SSR(X_1)$$
$$= 1046.26 - 793.28 \quad \text{df=2-1=1}$$
$$= 252.98$$

Extra Mean Square: $MSR(X_2 \mid X_1)$ = 252.98/1=252.98

Partial F-test: $F_{cal} = \frac{MSR(X_2 \mid X_1)}{MSE} = \frac{252.98}{9.939}$ =25.45 (denominator is MSE for Model 3)

Compare with F with 1 and 24 DF

```
PROC REG DATA=d2;
  MODEL steroid=age age2 /SS1;
    PLOT p.*age steroid*age/OVERLAY;
  TITLE2 "Table 6.  Regression of Steroids with age and age squared (Model
3)";
  RUN;
```

```
Table 6.  Regression of Steroids with age and age squared

Dependent Variable: steroid

                           Analysis of Variance
                                  Sum of           Mean
Source                     DF        Squares         Square    F Value    Pr > F
Model                       2     1046.26586      523.13293      52.63    <.0001
Error                      24      238.54081        9.93920
Corrected Total            26     1284.80667

Root MSE              3.15265    R-Square     0.8143
Dependent Mean       17.64444    Adj R-Sq     0.7989
Coeff Var            17.86766


                              Parameter Estimates
                    Parameter       Standard
Variable     DF      Estimate          Error    t Value    Pr > |t|      Type I SS
Intercept     1     -26.32541        5.88154      -4.48      0.0002     8405.81333
age           1       4.87357        0.77515       6.29      <.0001      793.28051
age2          1       -0.11840       0.02347      -5.05      <.0001      252.98535

Source: esb10p24.sas 3/9/2010 by ejs
```

The partial F-test is $F_{cal} = (-5.05)^2$.

Residuals

Fit Model of Steriods on Age, and Get Residuals

```
PROC REG DATA=d2;
  MODEL steroid=age ;
  OUTPUT OUT=e1  p=yhat r=yresid;
  TITLE2 "Simple regression model on age";
  RUN;
PROC PRINT DATA=e1 (OBS=10) NOOBS;
  TITLE2 "Table 7. List of Residuals from Reg of Steroids on Age";
  RUN;
```

Table 7. List of Residuals from Reg of Steroids on Age

steroid	age	age2	yhat	yresid
27.1	23	529	24.8970	2.20304
22.1	19	361	20.8802	1.21982
21.9	25	625	26.9054	-5.00535
10.7	12	144	13.8508	-3.15082
1.4	8	64	9.8340	-8.43404
18.8	12	144	13.8508	4.94918
14.7	11	121	12.8466	1.85338
5.7	8	64	9.8340	-4.13404
18.6	17	289	18.8718	-0.27179
20.4	18	324	19.8760	0.52401

Source: esb10p25.sas 3/23/2010 by ejs

Construct Residual Plots and Studentized Residuals

Studentized Residual: Fit model without the observation, calculate the residual using the observation, divide by the standard error based on fitted model. (If error is normally distributed, residuals should be between -2 and 2.)

```
FOOTNOTE "&prg";
PROC REG DATA=d2;
  MODEL steroid=age ;
  OUTPUT OUT=e1  p=yhat r=yresid;
  PLOT r.*age /NOMODEl NOSTAT ;
  PLOT rstudent.*age /NOMODEl NOSTAT ;
  TITLE1 "Figure 3.  Residuals from Regression on Age (Model 1)";
  RUN;
FOOTNOTE ;
TITLE1 "&prg" ;
```

Figure 3. Residuals from Regression on Age (Model 1)

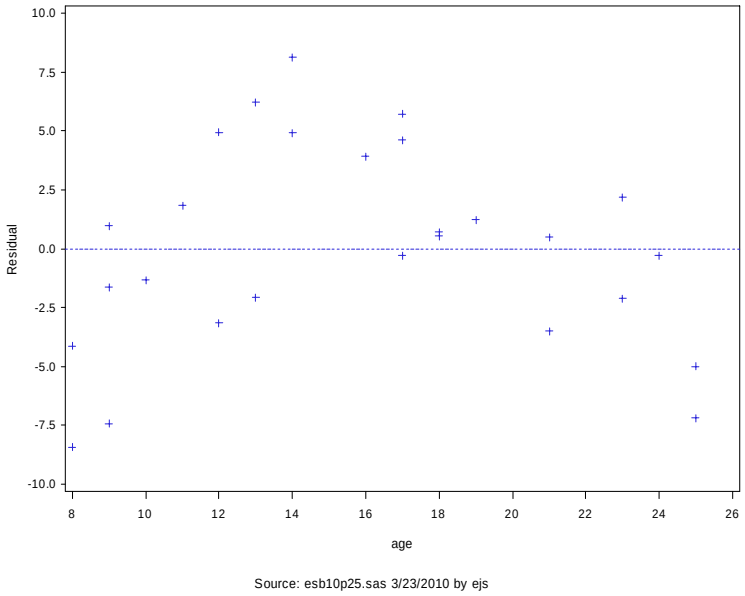
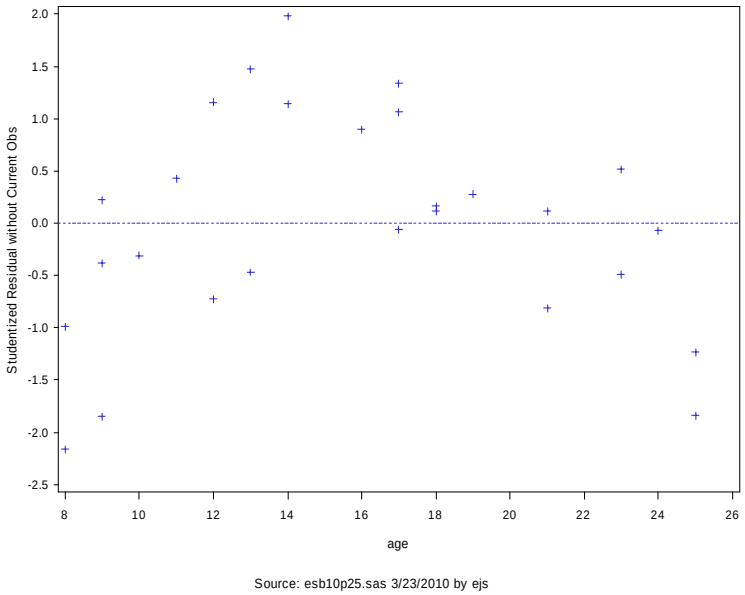
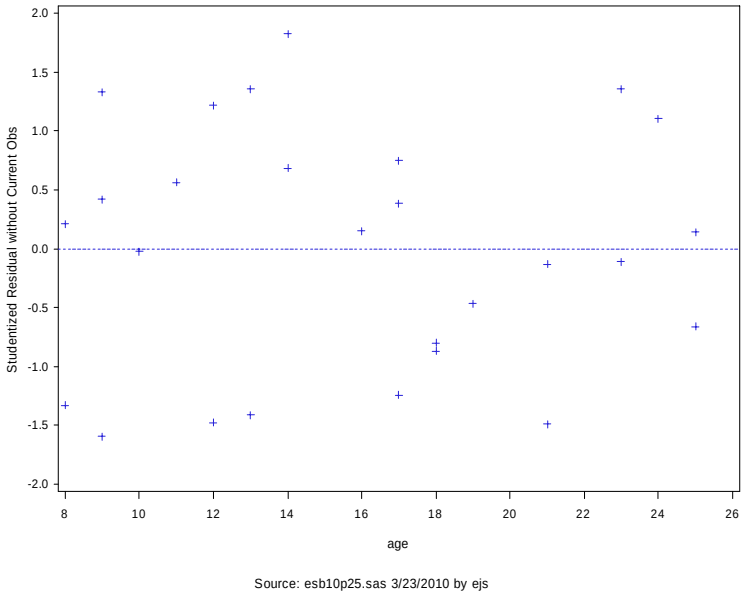


Figure 3. Residuals from Regression on Age (Model 1)




```
FOOTNOTE "&prg";
PROC REG DATA=d2;
  MODEL steroid=age age2;
  OUTPUT OUT=e1  p=yhat r=yresid;
  PLOT rstudent.*age /NOMODEl NOSTAT ;
  TITLE1 "Figure 4.  Residuals from Regression on Age and Age2 (Model 3)";
  RUN;
FOOTNOTE ;
TITLE1 "&prg" ;
```

Figure 4. Residuals from Regression on Age and Age2 (Model 3)



Source: esb10p25.sas 3/23/2010 by ejs

Figure 4. Residuals from Regression on Age and Age2 (Model 3)

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	1046.26586	523.13293	52.63	<.0001
Error	24	238.54081	9.93920		
Corrected Total	26	1284.80667			

Building Regression Models by new models for Residuals:

Fit Model of Steriods on Age, and Get Y-Residuals
Fit Model of Age2 on Age, and Get X2-Residuals

Fit Model of Y-Residuals on X2-Residuals

```
PROC REG DATA=d2;
  MODEL steroid=age ;
  OUTPUT OUT=e1  p=yhat r=yresid;
  TITLE2 "Table 7.  Simple  Regression on Age (Model 1)";
  RUN;
PROC REG DATA=d2;
  MODEL age2=age ;
  OUTPUT OUT=e2   r=age2_resid;
  TITLE2 "Table 8.  Simple  Regression on Age (Model 1)";
  RUN;
DATA d3;
  MERGE e1 (KEEP=yresid age)
        e2 (KEEP=age2_resid age2);
RUN;
PROC PRINT DATA=d3 (OBS=10) NOOBS;
  VAR yresid age2_resid;
  TITLE2 "Table 8.  List of residuals ";
RUN;

PROC REG DATA=d3;
  MODEL yresid=age2_resid;
  TITLE2 "Table 9.  Regression of Residuals on Age Squared (like Model 3)";
  RUN;
```

```
Table 8.  List of residuals
                age2_
 yresid         resid
 2.20304      14.9021
 1.21982     -22.3770
-5.00535      45.5416
-3.15082     -10.6154
-8.43404      40.1055
 4.94918     -10.6154
 1.85338      -0.9352
-4.13404      40.1055
-0.27179     -29.0166
 0.52401     -26.6968
```

Table 7. Simple Regression on Age (Model 1)
Dependent Variable: steroid

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	793.28051	793.28051	40.35	<.0001
Error	25	491.52616	19.66105		
Corrected Total	26	1284.80667			

Root MSE	4.43408	R-Square	0.6174
Dependent Mean	17.64444	Adj R-Sq	0.6021
Coeff Var	25.13016		

Table 9. Regression of Residuals on Age Squared (like Model 3)
Dependent Variable: yresid Residual
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	252.98535	252.98535	26.51	<.0001
Error	25	238.54081	9.54163		
Corrected Total	26	491.52616			

Root MSE	3.08895	R-Square	0.5147
Dependent Mean	-3.8323E-15	Adj R-Sq	0.4953
Coeff Var	-8.06026E16		

Source: esb10p25.sas 3/23/2010 by ejs

Compare to Model 3:

Table 6. Residuals from Regression on Age and Age2 (Model 3)
Dependent Variable: steroid
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	1046.26586	523.13293	52.63	<.0001
Error	24	238.54081	9.93920		
Corrected Total	26	1284.80667			

Root MSE	3.15265	R-Square	0.8143
Dependent Mean	17.64444	Adj R-Sq	0.7989
Coeff Var	17.86766		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS
Intercept	1	-26.32541	5.88154	-4.48	0.0002	8405.81333
age	1	4.87357	0.77515	6.29	<.0001	793.28051
age2	1	-0.11840	0.02347	-5.05	<.0001	252.98535

Source: esb10p25.sas 3/23/2010 by ejs

Notes:
The accounting of the SS is the same using Tables 7 and 9, as in Table 6. The DF for the Error for Table 6 is correct- The DF in Table 9 does not account for having fit age in the model.

Correlation and Partial Correlation

```
PROC CORR DATA=d2;
  VAR steroid age age2;
  TITLE2 "Table 1.  Correlation of Steroids with other variables";
RUN;
```

```
Table 1.  Correlation of Steroids with other variables
    Pearson Correlation Coefficients, N = 27
           Prob > |r| under H0: Rho=0

              steroid          age          age2
steroid       1.00000      0.78577      0.71312
                            <.0001       <.0001

age           0.78577      1.00000      0.98943
               <.0001                    <.0001

age2          0.71312      0.98943      1.00000
               <.0001       <.0001
Source: esb10p25.sas 3/23/2010 by ejs
```

Partial Correlation: Correlation of Residuals

```
PROC CORR DATA=d3;
  VAR yresid age2_resid;
  TITLE2 "Table 10.  Correlation of Residuls of Y on age with age2 with age";
RUN;
```

```
Table 10.  Correlation of Residuls of Y on age with age2 with
age

   2  Variables:    yresid     age2_resid

Pearson Correlation Coefficients, N = 27
       Prob > |r| under H0: Rho=0
                                  age2_
                    yresid        resid
yresid            1.00000     -0.71742
Residual                       <.0001

age2_resid       -0.71742      1.00000
Residual          <.0001
```

```
PROC CORR DATA=d2;
  PARTIAL age;
  VAR steroid age2;
  TITLE2 "Table 11.  Partial correlation of age squared wih steroids";
  RUN;
```

```
Table 11.  Partial correlation of age squared wih steroids

   1 Partial Variables:    age
   2         Variables:    steroid  age2

Pearson Partial Correlation Coefficients, N = 27
       Prob > |r| under H0: Partial Rho=0

              steroid          age2
steroid       1.00000      -0.71742
                            <.0001

age2         -0.71742       1.00000
               <.0001

Source: esb10p25.sas 3/23/2010 by ejs
```

Notes:
Partial correlations allow you to see what variable is most highly correlated after accounting for the previous variable.

Example 2: Body Fat (Y) and its relationship to
X1 tricepts skinfold thickness
X2 thigh circumference
X3 midarm circumference

What is the best model for estimating Body Fat (CH07TA01) based on these other measures? (Chapter 7 (p257) (Kutner, Nachtsheim, neter, and Li (5th edition: Applied Linear Statistical Models) (program esb10p26.sas))