An Example Data Analysis

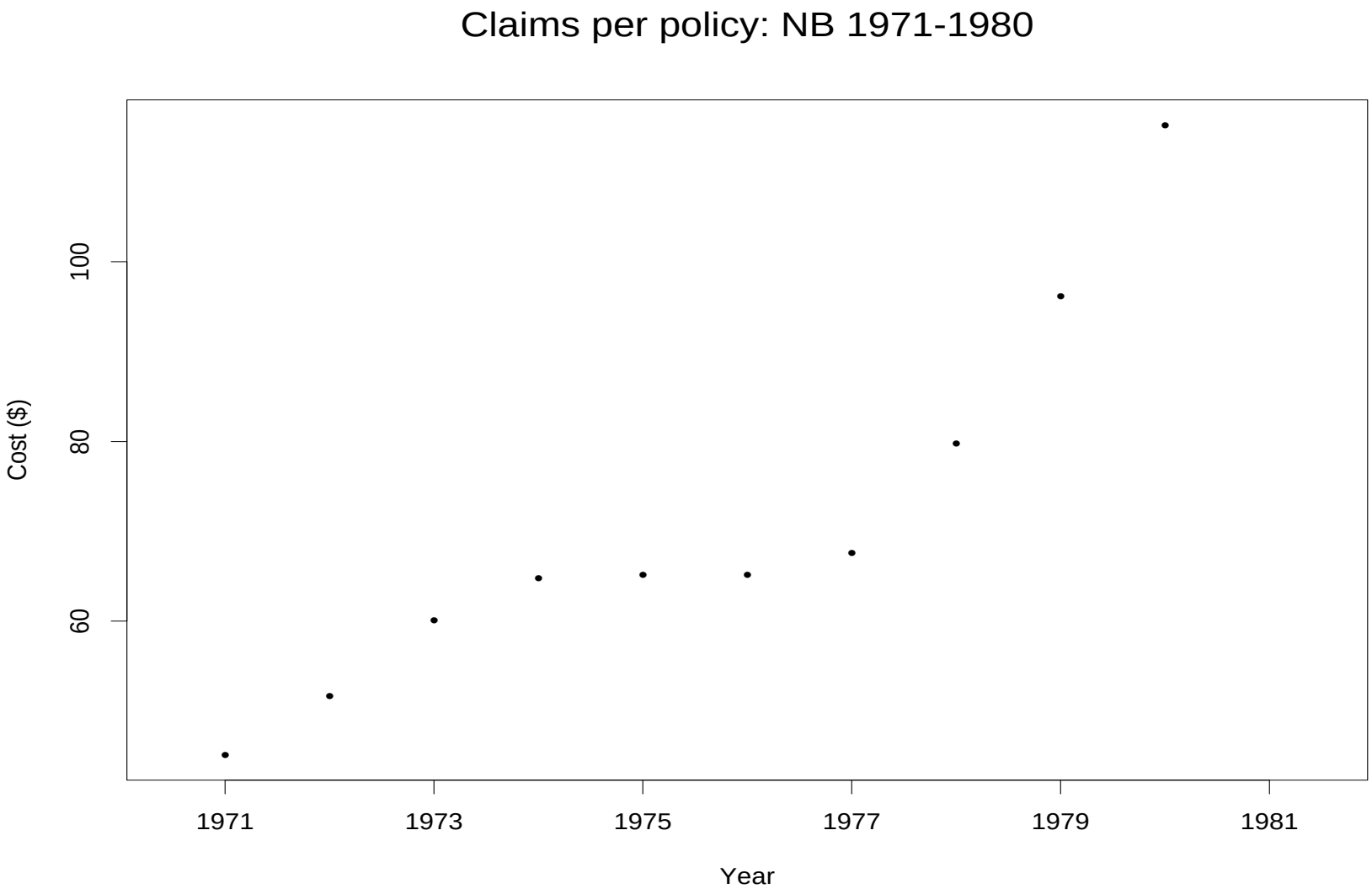
- Fit a polynomial model to a small data set.
- Use software to complete ANOVA table.
- Discuss use of estimates and standard errors: tests and confidence intervals.
- Discuss use of F-tests.
- Examine residual plots.
- Then go back to theory to justify tests and so on.

Polynomial Regression

Data: average claims paid per policy for automobile insurance in New Brunswick in the years 1971-1980:

Year	1971	1972	1973	1974	1975
Cost	45.13	51.71	60.17	64.83	65.24
Year	1976	1977	1978	1979	1980
Cost	65.17	67.65	79.80	96.13	115.19

Data Plot


Claims per policy: NB 1971-1980
Cost ($)
60
80
100
1971
1973
1975
1977
1979
1981
Year

- One goal of analysis is to extrapolate the Costs for 2.25 years beyond the end of the data; this should help the insurance company set premiums.
- We fit polynomials of degrees from 1 to 5, plot the fits, compute error sums of squares and examine the 5 resulting extrapolations to the year 1982.25.
- The model equation for a pth degree polynomial is

$$Y_i = \beta_0 + \beta_1 t_i + \cdots + \beta_p t_i^p + \epsilon_i$$

where the t_i are the covariate values (the dates in the example).

Notice:

- $p+1$ parameters (sometimes there will be p parameters in total and sometimes a total of $p+1$ – the intercept plus p others)
- β_0 is the intercept

The design matrix is given by

$$X = \begin{bmatrix} 1 & t_1 & \cdots & t_1^p \\ 1 & t_2 & \cdots & t_2^p \\ \vdots & \vdots & \cdots & \vdots \\ 1 & t_n & \cdots & t_n^p \end{bmatrix}$$

Other Goals of Analysis

- estimate βs
- select good value of p. This presents a trade-off:
  - large p fits data better BUT
  - small p is easier to interpret.

Edited SAS code and output

```
options pagesize=60 linesize=80;
data insure;
  infile 'insure.dat';
  input year cost;
  code = year - 1975.5 ;
  c2=code**2 ;
  c3=code**3 ;
  c4=code**4 ;
  c5=code**5 ;
proc glm  data=insure;
   model cost = code c2 c3 c4 c5 ;
run ;
```

NOTE: the computation of *code* is important. The software has great difficulty with the calculation without the subtraction. It should seem reasonable that there is no harm in counting years with 1975.5 taken to be the 0 point of the variable time.

Here is some edited output:

Dependent Variable: COST

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	3935.2507732	787.0501546	2147.50	0.0001
Error	4	1.4659868	0.3664967		
Corr Total	9	3936.7167600			

Source	DF	Type I SS	Mean Square	F Value	Pr > F
CODE	1	3328.3209709	3328.3209709	9081.45	0.0001
C2	1	298.6522917	298.6522917	814.88	0.0001
C3	1	278.9323940	278.9323940	761.08	0.0001
C4	1	0.0006756	0.0006756	0.00	0.9678
C5	1	29.3444412	29.3444412	80.07	0.0009

From these sums of squares I can compute error sums of squares for each of the five models.

Degree	Error Sum of Squares
1	608.395789
2	309.743498
3	30.811104
4	30.810428
5	1.465987

- Last line is produced directly by SAS.
- Each higher line consists of the sum of the line below together with the Type I SS figure from SAS.
- So, for instance, the ESS for a degree 4 fit is just the ESS for a degree 5 fit plus 29.3444412, the ESS for a degree 3 fit is the ESS for a degree 2 fit plus 0.006756, and so on.

Same Numbers Different Arithmetic

R Code	Error Sum of Squares.
lm(costc̃ode)	608.4
lm(costc̃ode+c2)	309.7
lm(costc̃ode+c2+c3)	30.8
lm(costc̃ode+c2+c3+c4)	Not computed
lm(costc̃ode+c2 +c3+ c4+ c5)	1.466

The actual estimates of the coefficients must be obtained by running SAS proc glm 5 times, once for each model. The fitted models are

$$y = 71.102 + 6.3516t$$

$$y = 64.897 + 6.3516t + 0.7521t^2$$

$$y = 64.897 + 1.9492t + 0.7521t^2 + 0.3005t^3$$

$$y = 64.888 + 1.9492t + 0.7562t^2 + 0.3005t^3 - 0.0002t^4$$

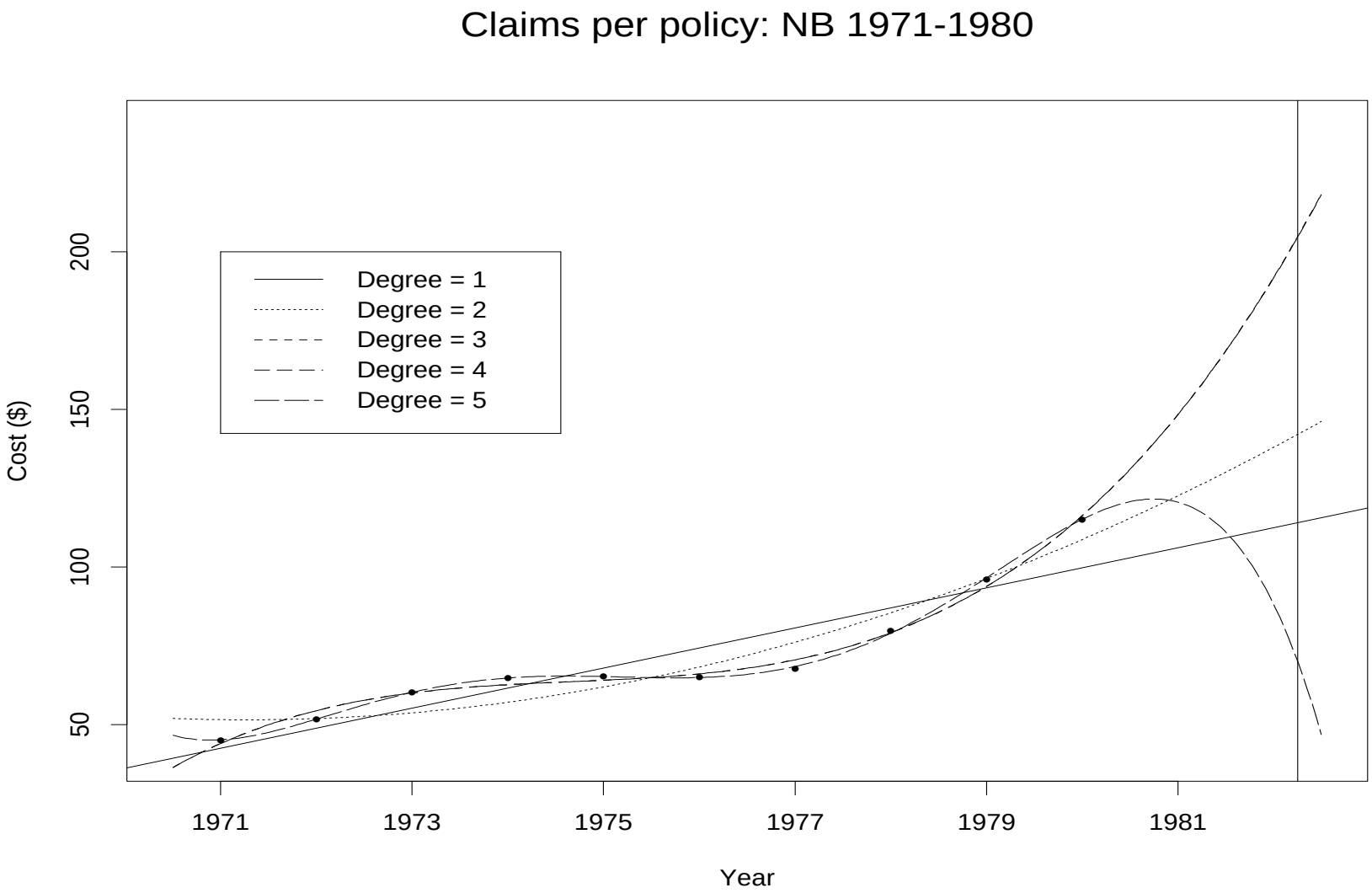
$$y = 64.888 - 0.5024t + 0.7562t^2 + 0.8016t^3 - 0.0002t^4 - 0.0194t^5$$

You should observe that sometimes, but not always, adding a term to the model changes coefficients of terms already in the model.

These lead to the following predictions for 1982.25:

Degree	$\hat{\mu}_{1982.25}$
1	113.98
2	142.04
3	204.74
4	204.50
5	70.26

Here is a plot of the five resulting fitted polynomials, superimposed on the data and extended to 1983.



- Vertical line at 1982.25 to show that the different fits give wildly different extrapolated values.
- No visible difference between the degree 3 and degree 4 fits.
- Overall the degree 3 fit is probably best but does have a lot of parameters for the number of data points.
- The degree 5 fit is a statistically significant improvement over the degree 3 and 4 fits.
- But it is hard to believe in the polynomial model outside the range of the data!
- Extrapolation is very dangerous and unreliable.

We have fitted a sequence of models to the data:

Model	Model equation	Fitted value
0	$Y_i = \beta_0 + \epsilon_i$	$\hat{\mu}_0 = \begin{bmatrix} \bar{Y} \\ \vdots \\ \bar{Y} \end{bmatrix}$
1	$Y_i = \beta_0 + \beta_1 t_i + \epsilon_i$	$\hat{\mu}_1 = \begin{bmatrix} \hat{\beta}_0 + \hat{\beta}_1 t_1 \\ \vdots \\ \hat{\beta}_0 + \hat{\beta}_1 t_n \end{bmatrix}$
$\vdots$	$\vdots$	$\vdots$
5	$Y_i = \beta_0 + \beta_1 t_i + + \cdots + \beta_5 t_i^5 + \epsilon_i$	$\hat{\mu}_5$

This leads to the decomposition

$$Y = \underbrace{\hat{\mu}_0 + (\hat{\mu}_1 - \hat{\mu}_0) + \cdots + (\hat{\mu}_5 - \hat{\mu}_4) + \hat{\epsilon}}_{\text{7 pairwise } \perp \text{ vectors}}$$

We convert this decomposition to an ANOVA table via Pythagoras:

$$||Y - \hat{\mu}_o||^2 = \underbrace{||\hat{\mu}_1 - \hat{\mu}_0||^2 + \cdots + ||\hat{\mu}_5 - \hat{\mu}_4||^2}_{\text{Model SS}} + ||\hat{\epsilon}||^2$$

or

$$\text{Total SS (Corrected)} = \text{Model SS} + \text{Error SS}$$

Notice that the Model SS has been decomposed into a sum of 5 individual sums of squares.

Summary of points to take from example

1. When I used SAS I fitted the model equation

$$Y_i = \beta_0 + \beta_1(t_i - \bar{t}) + \beta - 2(t_i - \bar{t})^2 + \cdots + \beta_p(t_i - \bar{t})^p + \epsilon_i$$

What would have happened if I had not subtracted $\bar{t}$? Then the entry in row $i+1$ and column $j+1$ of $X^T X$ is

$$\sum_{k=1^n} t_k^{i+j}$$

For instance, for $i = 5$ and $j = 5$ for our data we get

$$(1971)^{10} + (1972)^{10} + \cdots = \text{HUGE}$$

Many packages pronounce $X^T X$ singular. However, after **recoding** by subtracting $\bar{t} = 1975.5$ this entry becomes

$$(-4.5)^{10} + (-3.5)^{10} + \cdots$$

which can be calculated "fairly" accurately.

2. Compare (in case $p = 2$) for simplicity:

$$\mu_i = \alpha_0 + \alpha_1 t_i + \alpha_2 t_i^2 + \cdots + \alpha_p t_i^p$$

and

$$\mu_i = \beta_0 + \beta_1(t_i - \bar{t}) + \beta_2(t_i - \bar{t})^2 + \cdots + \beta_p(t_i - \bar{t})^p$$

$$\begin{aligned}\alpha_0 + \alpha_1 t_i + \alpha_2 t_i^2 &= \beta_0 + \beta_1(t_i - \bar{t}) + \beta_2(t_i - \bar{t})^2 \\ &= \underbrace{\beta_0 - \beta_1\bar{t} + \beta_2\bar{t}^2}_{\alpha_0} + \underbrace{(\beta_1 - 2\bar{t}\beta_2)}_{\alpha_1} t_i + \underbrace{\beta_2}_{\alpha_2} t_i^2\end{aligned}$$

So the parameter vector α is a linear transformation of β:

$$\begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & -\bar{t} & \bar{t}^2 \\ 0 & 1 & -2\bar{t} \\ 0 & 0 & 1 \end{bmatrix}}_{=A \text{ say}} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}$$

It is also an algebraic fact that

$$\hat{\alpha} = A\hat{\beta}$$

but $\hat{\beta}$ suffers from much less round off error.

3. Extrapolation is very dangerous — good extrapolation requires models with a good physical / scientific basis.
4. How do we decide on a good value for p? A convenient informal procedure is based on the **Multiple** R^2 or **Multiple Correlation** $(=R)$ where

$$R^2 = \text{fraction of variation of } Y \text{ "explained" by regression}$$
$$= 1 - \frac{ESS}{TSS(\text{adjusted})}$$
$$= 1 - \frac{\sum(Y_i - \hat{\mu}_i)^2}{\sum(Y_i - \bar{Y})^2}$$

For our example we have the following results:

Degree	R^2
1	0.8455
2	0.9213
3	0.9922
4	0.9922
5	0.9996

Remarks:

- Adding columns to X always drives R^2 up because the ESS goes down.
- 0.92 is a high R^2 but the model is **very bad** — look at residuals.
- Taking $p = 9$ will give $R^2 = 1$ because there is a degree 9 polynomial which goes exactly through all 10 points.

Effect of adding variables in different orders

Decomposition of Model SS depends on order in which variables are entered into the model in SAS. Examples and ANOVA tables:

```
options pagesize=60 linesize=80;
data insure;
  infile 'insure.dat';
  input year cost;
  code = year - 1975.5 ;
  c2=code**2 ;
  c3=code**3 ;
  c4=code**4 ;
  c5=code**5 ;
proc glm  data=insure;
   model cost = code c2 c3 c4 c5 ;
run ;
```

Edited output:

Dependent Variable: COST

Source	DF	Type I SS	Mean Sq	F Value	Pr > F
CODE	1	3328.3210	3328.3210	9081.45	0.0001
C2	1	298.6523	298.6523	814.88	0.0001
C3	1	278.9324	278.9324	761.08	0.0001
C4	1	0.0007	0.0007	0.00	0.9678
C5	1	29.3444	29.3444	80.07	0.0009
Model	5	3935.2508	787.0502	2147.50	0.0001
Error	4	1.4660	0.3665		
C Totl	9	3936.7167			

Edited output: Changing the model statement in *proc glm* to

```
   model cost =  code c4 c5 c2 c3 ;
```

gives

```
Dependent Variable: COST
           Sum of     Mean
Source DF Squares   Square    F Value Pr > F

Model  5 3935.2508 787.0502 2147.50 0.0001
Error  4    1.4660   0.3665
C Totl 9 3936.7168

Source DF Type I SS Mean Sq  F Value Pr > F

CODE   1 3328.3210 3328.3210 9081.45 0.0001
C4     1  277.7844  277.7844  757.95 0.0001
C5     1  235.9181  235.9181  643.71 0.0001
C2     1   20.8685   20.8685   56.94 0.0017
C3     1   72.3588   72.3588  197.43 0.0001
```

Source	DF	Type III SS	Mean Square	F Value	Pr > F
CODE	1	0.88117350	0.88117350	2.40	0.1959
C4	1	0.00067556	0.00067556	0.00	0.9678
C5	1	29.34444115	29.34444115	80.07	0.0009
C2	1	20.86853994	20.86853994	56.94	0.0017
C3	1	72.35876312	72.35876312	197.43	0.0001

Parameter	Estimate	T for H0: Parameter=0	Pr > |T|	Std Error of Estimate
INTERCEPT	64.88753906	176.14	0.0001	0.36839358
CODE	-0.50238411	-1.55	0.1959	0.32399642
C4	-0.00020251	-0.04	0.9678	0.00471673
C5	-0.01939615	-8.95	0.0009	0.00216764
C2	0.75623470	7.55	0.0017	0.10021797
C3	0.80157430	14.05	0.0001	0.05704706

Discussion

- For CODE the SS is unchanged.
- But after that, the SS are all changed.
- The MODEL, ERROR and TOTAL SS are unchanged, though.
- Each Type 1 SS is the sum of squared entries in the difference in two vectors of fitted values.
- So, e.g., line C5 is computed by fitting the two models

$$\mu_i = \beta_0 + \beta_1 t_i + \beta_4 t_i^4$$

and

$$\mu_i = \beta_0 + \beta_1 t_i + \beta_4 t_i^4 + \beta_5 t_i^5 \, .$$

- To compute a line in the Type III sum of squares table you also compare two models,
- But, in this case, the two models are the full fifth degree polynomial and the model containing every power *except* the one matching the line you are looking at.
- So, for example, the C4 line compares the models

$$\mu_i = \beta_0 + \beta_1 t_i + \beta_2 t_i^2 + \beta_3 t_i^3 + \beta_5 t_i^5$$

and

$$\mu_i = \beta_0 + \beta_1 t_i + \beta_2 t_i^2 + \beta_3 t_i^3 + \beta_4 t_i^4 + \beta_5 t_i^5 \, .$$

- For polynomial regression this comparison is silly;
- No one would expect a model like the fifth degree polynomial in which the coefficient of t^4 is exactly 0 to be realistic.
- In many multiple regression problems, however, the type III SS are more useful.

It is worth remarking that the estimated coefficients are the same regardless of the order in which the columns are listed. This is also true of type III SS. You will also see that all the F P-values with 1 df in the type III SS table are matched by the corresponding P-values for the t tests.

Selection of Model Order

An informal method of selecting p, the model order, is based on

$$\begin{aligned} R^2 &= \text{squared multiple correlation} \\ &= \text{coefficient of multiple determination} \\ &= 1 - \frac{\text{ESS}}{\text{TSS(Adjusted)}} \end{aligned}$$

Note: adding more terms always increases R^2.

Formal methods can be based on hypothesis tests. We can test $H_o : \beta_5 = 0$ and then, if we accept this test $H_o : \beta_4 = 0$ and then, if we accept that test $H_o : \beta_3 = 0$ and so on stopping when we first reject a hypothesis. This is "backwards elimination".
Justification: Unless $\beta_5 = 0$ there is no good reason to suppose that $\beta_4 = 0$ and so on.
Apparent conclusion in our example: $p = 5$ is best; look at the P values in the SAS outputs.

Problems arising with that conclusion:

- $p = 5$ gives lousy extrapolation
- there is no good physical meaning to a fifth degree polynomial model.
- there are too many parameters for $n = 10$
- the correct relation is probably not best described by a polynomial.
- If, for instance, $\mu(t) = \alpha_0 e^{\alpha_1 t}$ then the best polynomial approximation might well have high degree.