## Contents

13 Model Choice in Multiple Regression ............................... 5

13.1 Overview .......................................................... 5

13.2 ANOVA versus Linear Continuous Models ....................... 5

13.3 Choice of Transformation ......................................... 9

13.4 Example: Beta Carotene Supplements ............................ 12

13.4.1 Overview ......................................................... 12

13.4.2 Methods for modeling dose response .......................... 14

13.4.3 ANOVA analysis ................................................ 16

13.4.4 Binary dose: Placebo versus Active ........................... 21

13.4.5 Linear, continuous dose ....................................... 23

13.4.6 Polynomial models of dose .................................... 25

13.4.7 Highest order polynomial models versus ANOVA ........... 27

13.4.8 Threshold at 0 and Linear Term ............................... 28

13.5 Data driven model selection ....................................... 32

13.5.1 Simulation ....................................................... 32
13.5.2 Post hoc adjustments for multiple comparisons . . . . . . . 34
Chapter 13

Model Choice in Multiple Regression

13.1 Overview

· There are many different ways we can model predictors

· What alternative models will make scientific sense

· What is the impact of letting the data drive the selection of a model

· I am going to discuss in terms of a clinical trial where we have replicates at dose levels
  – We can find dose-specific means and compare modeling approaches

13.2 ANOVA versus Linear Continuous Models

· Compare power of linear continuous models versus ANOVA as a function
  – of trend in means AND
– standard errors within groups

• ANOVA (dummy variables)
  – Uses indicator variables for every dose (group) level
    ∗ Again, I am thinking about “dose” in a general sense that could including covariates like age, cholesterol, blood pressure, etc.
    
    ∗ Traditionally, dose would just be dose of some treatment
  
  – Fits group means exactly
    ∗ One way ANOVA: One categorical predictor
    
    ∗ Two way ANOVA: Two categorical predictors
      · Fit with the interactions to get group means exactly
  
  – Does not mix random error with systematic error
    ∗ Systematic error: Error due to differences from sample means from predicted means
    
    ∗ Random error: Error that cannot be explained after controlling for dose
  
  – ANOVA ignores the ordering of the groups, so it gains no power from trends
    ∗ e.g. does not assume that the difference between dose=15 and dose=30 group is similar to the difference between the dose=30 and dose=45 groups
    
    ∗ In fact, the same level of significance is gained no matter what permutation of dose groups is used

• Linear continuous models
  – Borrows information across groups
 Accurate and efficient if the model is correct

- If model is incorrect, mixes random and systematic error
  - Will have some systematic error because the means are not predicted exactly

- Can gain power from ordering of groups in order to detect a trend
  - But, no matter how low the standard error is, if there is no trend in the mean, there is no statistical significance
CHAPTER 13. MODEL CHOICE IN MULTIPLE REGRESSION

Linear: High power; ANOVA: High Power

Linear: Mod power; ANOVA: Low Power

Linear: No power; ANOVA: High Power

Linear: No power; ANOVA: Low Power
Other options for modeling continuous predictors
- Combinations of linear trends and indicator variables
- Splines
- Fractional polynomials
- etc.

13.3 Choice of Transformation

- The exact form used to model predictors should be based on scientific (first) and statistical (second) criteria

- Scientific issues
  - The form used to model predictors must address the specific scientific question
    - Should be the next logical step in the process of investigating the overall goal
    - Remember binary search from the Scientist game
    - First, establish some sort of an association
    - Second, detect a first order trend
    - Third, detecting specific forms of non-linearities
      - Threshold effects?
      - U- or S-shaped trends?
• Finally, more complex models

– When the scientific question relates to prediction, it is imperative that the regression model accurately reflects the true relationship between predictors and the summary measure of response
  • Failure to have the correct model will guarantee that some groups may not have the correct predicted response

– When the scientific question relates to detection of associations, the importance of having the true model depends on the statistical role of the predictor
  • With the predictor of interest, the most important issues is to protect the validity of the statistical inference
    ∙ Data driven decision will inflate the type I error rate
  
  • With precision variables, it is not as crucial that the true relationship be modeled
    ∙ An approximate model will provide most of the precision gains

  • With confounders, failure to accurately model the relationship between the confounder and the response may lead to residual confounding
    ∙ Sometimes will use very flexible models for continuous confounders (e.g. fractional polynomials)

– As the goal of any analysis is to communicate findings to the greater scientific community, it is also important that modeling of predictors is easy to understand
  • This is an issue that matters most for your predictor of interest

  • We are generally not worried about making inference about precision variables or confounders

• Statistical issues
The greatest statistical precision will be gained when the model reflects the true relationship between the predictor and the response:

- Accurate modeling of the relationship will avoid introducing systematic error in the estimates of the standard errors.

- Parsimony: Using the fewest parameters to model the relationship will allow greater precision.

- Precision is a trade-off between parsimony and increased accuracy from including more parameters.

We should select the form of modeling the predictor before looking at the data:

- Data drive selection of transformations will tend to lead to inaccurate (anti-conservative) statistical inference.

- Overfitting of the data leads to spuriously low estimates of the within group variability.
  - Thus standard errors estimates are too low.
  - Type-I errors are inflated.
  - Confidence interval are too narrow (inaccurate coverage probabilities).

- Data-driven model selection will also lead to coefficient estimates that are biased away from the null (leading you to overstate your scientific effects).
13.4 Example: Beta Carotene Supplements

13.4.1 Overview

- Before doing large scale clinical trials, it is important to understand the pharmacokinetics of a drug

- Phase II prevention trials often administer a drug in various doses to volunteers, and pertinent plasma levels are then measured at regular intervals

- Of particular interest is how dose level affects the build up of drug in the plasma over time, as well as how the dose level might affect other blood chemistries

- Forty-six (46) volunteers were randomly assigned to receive one of five doses of beta-carotene (0, 15, 30, 45, or 60 mg/day) for 9 months in a double blind fashion

- The specific aim was to determine how different dose levels affected the serum beta-carotene levels after 9 months

- Other measured variables available in this data set include subject age, sex, weight, body mass index, percent body fat, and serum cholesterol level at baseline
. tabstat carot3, by(dose) stat(n mean sd min q max)

Summary for variables: carot3
by categories of: dose (dose)

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>mean</th>
<th>sd</th>
<th>min</th>
<th>p25</th>
<th>p50</th>
<th>p75</th>
<th>max</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>7</td>
<td>186.3214</td>
<td>87.79767</td>
<td>84.5</td>
<td>126</td>
<td>149</td>
<td>286</td>
<td>323</td>
</tr>
<tr>
<td>15</td>
<td>8</td>
<td>1253.583</td>
<td>570.4673</td>
<td>576.75</td>
<td>695.375</td>
<td>1250</td>
<td>1771.208</td>
<td>2018.75</td>
</tr>
<tr>
<td>30</td>
<td>9</td>
<td>1504.611</td>
<td>479.0258</td>
<td>849.3333</td>
<td>1157.333</td>
<td>1498.5</td>
<td>1840</td>
<td>2248.5</td>
</tr>
<tr>
<td>45</td>
<td>7</td>
<td>1749.081</td>
<td>579.049</td>
<td>950.25</td>
<td>993</td>
<td>1848.25</td>
<td>2247.667</td>
<td>2310.4</td>
</tr>
<tr>
<td>60</td>
<td>9</td>
<td>1877.63</td>
<td>429.8801</td>
<td>1233.333</td>
<td>1724.667</td>
<td>1865</td>
<td>1917.667</td>
<td>2855</td>
</tr>
<tr>
<td>Total</td>
<td>40</td>
<td>1350.416</td>
<td>734.4823</td>
<td>84.5</td>
<td>799.6667</td>
<td>1528.917</td>
<td>1914.667</td>
<td>2855</td>
</tr>
</tbody>
</table>

In this randomized trial, we can consider several potential response variables

- Plasma level at the end of treatment

- Change in plasma level over the treatment period

- Either of the above adjusted for baseline plasma (ANCOVA model)

Accounting for baseline

- Dose group $i$, subject $j$, time $t$

- $Y_{ijt} \sim (\mu_{it}, \sigma^2)$; corr($Y_{ij0}, Y_{ij9}$) = $\rho$

\[
\begin{align*}
\bar{Y}_{i.9} & \sim (\mu_{i9}, \sigma^2/n) \\
\bar{Y}_{i.9} - \bar{Y}_{i.0} & \sim (\mu_{i9} - \mu_{i0}, 2\sigma^2(1 - \rho)/n) \\
Y_{i.9} - \rho\bar{Y}_{i.0} & \sim (\mu_{i9} - \rho\mu_{i0}, \sigma^2(1 - \rho^2)/n)
\end{align*}
\]

- Compared variances of the above three equations
  * When are the variances equal, smaller, larger

  * Which is always smallest
CHAPTER 13. MODEL CHOICE IN MULTIPLE REGRESSION

- By randomization, there will be equal means at baseline
  - $\mu_{T,0} = \mu_{P,0}$ where $T$ is any of the treatment doses and $P$ is placebo

- Contrast across dose groups
  \[
  \begin{align*}
  \overline{Y}_{T,9} - \overline{Y}_{P,9} & \sim (\mu_{T,9} - \mu_{P,9}, 2\sigma^2 / n) \\
  (\overline{Y}_{T,9} - \overline{Y}_{T,0}) - (\overline{Y}_{P,9} - \overline{Y}_{P,0}) & \sim (\mu_{T,9} - \mu_{P,9}, 4\sigma^2(1 - \rho)/n) \\
  (\overline{Y}_{T,9} - \rho\overline{Y}_{T,0}) - (\overline{Y}_{P,9} - \rho\overline{Y}_{P,0}) & \sim (\mu_{T,9} - \mu_{P,9}, 2\sigma^2(1 - \rho^2)/n)
  \end{align*}
  \]

- Simple linear regression
  - Regress $Y$ on $X$
    \[
    \begin{align*}
    * Y_i & \sim (\mu_Y, \sigma_Y^2); X_i \sim (\mu_X, \sigma_X^2) \\
    * \text{corr}(Y_i, X_i) &= \rho \\
    * \beta_0 &= \mu_Y - \beta_1\mu_X \\
    * \beta_1 &= \rho \frac{\sigma_Y}{\sigma_X}
    \end{align*}
    \]

- Analysis of Covariance
  - Dose group $i$, subject $j$, time $t$
    \[
    \begin{align*}
    * Y_{ijt} & \sim (\mu_{it}, \sigma^2); \text{corr}(Y_{ij0}, Y_{ij9}) = \rho \\
    \end{align*}
    \]

  - Regression model: $E[Y_{ij9}|Y_{ij0}] = \beta_0 + \beta_1Y_{ij0}$
    \[
    \begin{align*}
    * \beta_1 &= \rho \\
    \end{align*}
    \]

13.4.2 Methods for modeling dose response

- In a randomized clinical trial, we will tend to have the greatest precision if we adjust for baseline as a predictor in a linear regression model
A wide variety of models may be considered for examining the relationship between dose and plasma levels

- Dummy variables where we model each dose level independently, without borrowing information across groups (ANOVA)

- Linear continuous predictors (transformed or untransformed)

- Dichotomization (at any of several thresholds)

- Polynomials, splines, other flexible methods

- Combinations of the above

- Even more complex models

I will compare possible models

- Graphically: Show data and fitted values without adjustment for baseline

- Numerically: Show regression estimates and tests after adjustment for baseline

- Note that this is an academic exercise and not something you would do in practice to come up with the “best” model

Predicted values

- After computing a regression command, Stata will provide predicted values for each case

  * Mathematically, this is just the intercept plus the regression parameters multiplied by the covariates for each case

  * Stata command: predict varname
13.4.3 ANOVA analysis

- Fits each group independently

- Does not use the ordering of the dose groups when looking for an effect
  - Completely ignores the magnitude and ordering of the $x$-axis

- A priori, we might expect this is not the most efficient method if the alternative hypothesis is true
  - We expect larger plasma levels with increasing dose
  - We will thus have less power to detect a first-order trend

```
.xi: regress carot3 i.dose carot0, robust
i.dose   _Idose_0-60 (naturally coded; _Idose_0 omitted)

Linear regression
        Number of obs =        40
       F(    5,    34) =     47.68
       Prob > F      =    0.0000
       R-squared    =     0.7184
       Root MSE     =     417.46

------------------------------------------------------------------------------
       | Robust
     carot3 | Coef.  Std. Err.    t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   _Idose_15  | 1224.19   213.56    5.73  0.000     790.19   1658.20
   _Idose_30  | 1439.84   155.80    9.24  0.000    1123.22   1756.47
   _Idose_45  | 1679.00   167.15   10.04  0.000    1339.30   2018.71
   _Idose_60  | 1791.01   152.95   11.71  0.000    1480.19   2101.83
    carot0   |  1.9028   0.5370    3.54  0.001     0.8115   2.9942
    _cons   | -361.45   167.54   -2.16  0.038    -701.94   -20.96
------------------------------------------------------------------------------
```
Model: Dummy Variables (ANOVA)

Fitted
Group Means
CHAPTER 13. MODEL CHOICE IN MULTIPLE REGRESSION

· Testing for the dose effect
  
  – We must use the _testparm command (or test) because the model includes the baseline measurement
  
  – _testparm is similar to test, but allows testing multiple parameters using wildcards

```
. testparm _I*

  ( 1) _Idose_15 = 0
  ( 2) _Idose_30 = 0
  ( 3) _Idose_45 = 0
  ( 4) _Idose_60 = 0

      F(  4,  34) =  59.47
      Prob > F =    0.0000
```

· We would have had the same fitted values (and thus inference) if we had decided to drop a different dose group
  
  – Example: Making my own dummy variables for dose, with dose at 60 being the reference group

```
. regress carot3 dose0 dose15 dose30 dose45 carot0, robust

Linear regression                     Number of obs =      40
                                  F(  5,  34) =    47.68
                                  Prob > F    =  0.0000
                                  R-squared   =    0.7184
                                  Root MSE    =  417.46

------------------------------------------------------------------------------
            | Robust             Coef.  Std. Err.     t    P>|t|     [95% Conf. Interval]
-------------+------------------------------------------------------------------
      carot3  |                  -1791.009   152.946   -11.71  0.000    -2101.833   -1480.185
     dose0   |                  -566.8194  240.2233   -2.36   0.024    -1055.012   -78.62694
     dose15  |                  -351.1717  188.5790   -1.86   0.071    -734.4104   32.067
     dose30  |                  -112.0251  204.5116   -0.55   0.587    -527.6427   303.5925
     dose45  |                   1.902792   0.537001    3.54  0.001       .8114738   2.99411
      carot0  |                  1429.557   176.8189    8.08   0.000       1070.218   1788.896
     _cons   |                  1429.557   176.8189    8.08   0.000       1070.218   1788.896
------------------------------------------------------------------------------
```

· Note that the parameter estimates all will lead to the same fitted values
– e.g. Intercept in above model (1430) equals the intercept + dose60 coefficient (-361 + 1791) in previous model

• Overall F statistics, R-squared, Root MSE all the same

• Partial t-tests tend to differ as we are making comparisons to different reference groups

• Could also fit the same model with no intercept
  – Would then have to include all five dose groups
  – We can get Stata to include fit all five dose groups and no intercept using the \texttt{noconstant} option
  – In R, fit a model without an intercept by adding a \(-1\) in the model equation (e.g. \(y \sim -1 + x\))
  – Not including the intercept changes the overall F statistic and the R-squared measures

```stata
. regress carot3 dose0 dose15 dose30 dose45 dose60 carot0, robust noconstant
```

| carot3 | Coef. | Std. Err. | t    | P>|t| | [95% Conf. Interval] |}
|--------|-------|-----------|------|-----|---------------------|}
| dose0  | -361.4516 | 167.5432 | -2.16 | 0.038 | -701.9404 -20.96284 |}
| dose15 | 862.7379 | 241.9456 | 3.57 | 0.001 | 371.0453 1354.431 |}
| dose30 | 1078.386 | 178.9827 | 6.03 | 0.000 | 714.6491 1442.122 |}
| dose45 | 1317.532 | 223.2838 | 5.90 | 0.000 | 863.7649 1771.3 |}
| dose60 | 1429.557 | 176.8189 | 8.08 | 0.000 | 1070.218 1788.896 |}
| carot0 | 1.902792 | .5370015 | 3.54 | 0.001 | .8114738 2.99411 |}
• Correspondence of the no-intercept model compared to previous models
  – Some textbooks refer to this as a “cell means” coding system
    * If we didn’t have baseline beta carotene in the model, the dose parameters would correspond directly to the means in each dose group
    * With baseline beta carotene in the model, the dose parameters are the means when carot0 is 0
  – In terms of model fit, the model is the same as before
    * No intercept means each dose group is compared to a mean of 0
  – Fitted values will be the same
  – Test of dose effect will need to test equality of all five dose covariates
    * This is not a test that these 5 parameters are 0
    * \( H_0 : \text{dose}0 = \text{dose}15 = \text{dose}30 = \text{dose}45 = \text{dose}60 \)
    * \( H_1 : \) at least one of the above is not equal

  . test dose0=dose15=dose30=dose45=dose60

    ( 1) dose0 - dose15 = 0
    ( 2) dose0 - dose30 = 0
    ( 3) dose0 - dose45 = 0
    ( 4) dose0 - dose60 = 0

    \( F( 4, 34) = 59.47 \)
    Prob > F = 0.0000

• Difference in interpretation of the no-intercept model to previous models
  – Overall F statistic
    * Test removing all covariates, leaving a mean 0
  – Multiple R squared
∗ “Percent of explained variation”

∗ With intercept, compares to variance around overall mean

∗ Without intercept, compares to variance about 0

13.4.4 Binary dose: Placebo versus Active

· Dichotomize into dose 0 versus dose > 0
  – Will be an accurate model if all (or virtually all) of the effect is attained at the lowest dose level
  – Often used when little is know about a treatment, or when dose is difficult to quantify
    ∗ e.g. Smoking

∗ We are relatively certain of a smoking effect, so our major scientific interest is likely related to the dose-response relationship above the lowest dose

. regress carot3 trt carot0, robust

Linear regression

Number of obs = 40
F( 2, 37) = 84.00
Prob > F = 0.0000
R-squared = 0.6434
Root MSE = 450.3

------------------------------------------------------------------------------
| Robust
| carot3 | Coef. | Std. Err. | t | P>|t| | [95% Conf. Interval] |
------------------------------------------------------------------------------
| trt | 1544.22 | 120.0337 | 12.86 | 0.000 | 1301.009 | 1787.432 |
| carot0 | 2.059907 | .7091411 | 2.90 | 0.006 | .6230503 | 3.496763 |
| _cons | -406.6816 | 215.3205 | -1.89 | 0.067 | -842.9623 | 29.59913 |
------------------------------------------------------------------------------
Model: Dichotomous Dose

- X Fitted
- ○ Group Means

Plasma Beta Carotene

Dose of Supplementation

0 10 20 30 40 50 60

0 500 1000 1500 2000 2500
13.4.5 Linear, continuous dose

- Estimates the best fitting straight line to response
  - Accurate if the response is linear

- Often used when little is known about the treatment and a general trend is expected
  - In this particular application, we are relatively certain of an effect, so our major interest is in modeling the dose response relationship above 0.

```
. regress carot3 dose carot0, robust
Linear regression
Number of obs = 40
F( 2, 37) = 25.47
Prob > F = 0.0000
R-squared = 0.5622
Root MSE = 498.94

------------------------------------------------------------------------------
| Robust
|     | Coef.  | Std. Err. | t    | P>|t|   | 95% Conf. Interval |
-------------+------------------------------------------------------------------
  carot3 |  dose | 25.46276 | 3.632547 | 7.01 | 0.000 | 18.10252  32.823 
         |  carot0 | 1.334088  | 0.657003 | 2.03 | 0.050 | 0.0028737 2.665303
         |    _cons | 245.0003  | 222.814  | 1.10 | 0.279 | -206.4638 696.4644
------------------------------------------------------------------------------
```

- To test the treatment effect, could either use the `test` command for dose or use the output directly as we are only testing one parameter.
Model: Linear Dose

![Graph showing data points and fitted line for plasma beta carotene vs. dose of supplementation. The graph includes points marked as X for fitted values and O for group means.]
13.4.6 Polynomial models of dose

- Fit terms involving dose, dose squared
  - Often used to fit U-shaped trends
  - In general, a quadratic is a pretty strong assumption in that it assumes constant curvature over dose

```
. regress carot3 dose dosesqr carot0, robust
```

```
Linear regression
Number of obs = 40
F( 3, 36) = 59.30
Prob > F = 0.0000
R-squared = 0.6824
Root MSE = 430.81

------------------------------------------------------------------------------
| Robust
| carot3 | Coef. Std. Err. t P>|t| [95% Conf. Interval]
-------------+----------------------------------------------------------------
dose | 67.14673 8.207062 8.18 0.000 50.50203 83.79142
dosesqr | -.6723087 .1451496 -4.63 0.000 -.9666857 -.3779317
carot0 | 1.728068 .5638738 3.06 0.004 .5844789 2.871657
_cons | -195.5986 179.34 -1.09 0.283 -559.3169 168.1197
------------------------------------------------------------------------------
```
CHAPTER 13. MODEL CHOICE IN MULTIPLE REGRESSION

Model: Quadratic Dose

\[ \text{Model: Quadratic Dose} \]

\[ \text{X} \quad \text{Fitted} \]

\[ \text{Group Means} \]

\[ \text{Plasma Beta Carotene} \]

\[ \text{Dose of Supplementation} \]

0 10 20 30 40 50 60

0 500 1000 1500 2000 2500
The partial t-test for dosesqr can be interpreted as a test for linear dose response
- It is highly significant, suggesting departure from linearity

To test the treatment effect, we need to test the two dose covariates

```
. test dose dosesqr
   ( 1) dose = 0
   ( 2) dosesqr = 0

F( 2, 36) = 84.56
Prob > F = 0.0000
```

13.4.7 Highest order polynomial models versus ANOVA

- With 5 discrete dose levels, a 4th degree polynomial will fit the means exactly

- Thus, the model will have the same fit as the ANOVA model using dummy variables for each levels of dose
  - Higher order polynomials are borrowing less information across dose groups
  - Highest order polynomial borrows no information across dose groups
. gen dosecub = dose^3
. gen dosequad = dose^4
. regress carot3 dose dosesqr dosecub dosequad carot0, robust

Linear regression
Number of obs = 40
F( 5,  34) = 47.68
Prob > F = 0.0000
R-squared = 0.7184
Root MSE = 417.46

------------------------------------------------------------------------------
 | Robust
| carot3 | Coef. Std. Err. t P>|t| [95% Conf. Interval]
-------------+----------------------------------------------------------------
dose | 157.876 61.94333 2.55 0.015 31.99197 283.7599
dosesqr | -6.943752 5.066692 -1.37 0.180 -17.24051 3.353004
dosecub | .1385695 .1313523 1.05 0.299 -.1283706 .4055096
dosequad | -.0009734 .0010718 -0.91 0.370 -.0031515 .0012047
carot0 | 1.902792 .5370015 3.54 0.001 .8114738 2.99411
_cons | -361.4516 167.5432 -2.16 0.038 -701.9404 -20.96284
------------------------------------------------------------------------------

. xi: regress carot3 i.dose carot0, robust
i.dose _Idose_0-60 (naturally coded; _Idose_0 omitted)

Linear regression
Number of obs = 40
F( 5,  34) = 47.68
Prob > F = 0.0000
R-squared = 0.7184
Root MSE = 417.46

------------------------------------------------------------------------------
 | Robust
| carot3 | Coef. Std. Err. t P>|t| [95% Conf. Interval]
-------------+----------------------------------------------------------------
_Idose_15 | 1224.19 213.5586 5.73 0.000 790.1863 1658.193
_Idose_30 | 1439.837 155.7948 9.24 0.000 1123.224 1756.45
_Idose_45 | 1678.984 167.1502 10.04 0.000 1339.294 2018.674
_Idose_60 | 1791.009 152.946 11.71 0.000 1480.185 2101.833
carot0 | 1.902792 .5370015 3.54 0.001 .8114738 2.99411
_cons | -361.4516 167.5432 -2.16 0.038 -701.9404 -20.96284
------------------------------------------------------------------------------

13.4.8 Threshold at 0 and Linear Term

· Threshold at 0 and linear dose

· To fit, use a dummy variable for dose0 plus dose (continuous)
– Fits dose 0 by its group mean

– Fits dose > 0 by a line (an intercept and slope)

– Allows us to address two scientific questions
  * Is there any effect of dose? (test both slopes)

  * Is there any additional benefit beyond the lowest dose? (test linear term’s slope)

```
.regress carot3 trt dose carot0, robust
```

```
Linear regression
Number of obs = 40
F( 3, 36) = 81.26
Prob > F = 0.0000
R-squared = 0.7170
Root MSE = 406.69

------------------------------------------------------------------------------
carot3 | Coef. Std. Err. t P>|t| [95% Conf. Interval]
-------------+----------------------------------------------------------------
  trt | 1050.836 223.418 4.70 0.000 597.7237 1503.949
  dose | 12.81144 4.794314 2.67 0.011 3.088122 22.53476
  carot0 | 1.904584 .5164903 3.69 0.001 .8570932 2.952075
   _cons | -361.9675 161.4254 -2.24 0.031 -689.3534 -34.58168
------------------------------------------------------------------------------
```
Model: Threshold and Linear Dose

- X Fitted
- O Group Means

Dose of Supplementation vs. Plasma Beta Carotene
· Testing the effect of treatment
  – Two variables model dose, so we need to test both

  – If response increases from dose 0 to lowest dose OR

  – ... response increases as dose increase, THEN

  – ... we will declare an effect of treatment

· The partial t-test for the \text{trt} term can be used to test for linear dose response
  – Here, it is highly significantly different from 0, indicating that just a linear model is not adequate

· The partial t-test for the \text{dose} term can be interpreted as a test for any added effect above the lowest dose
  – It is significantly different from 0 \((p = 0.011)\)

  – There is a multiple comparison issue here, but many people are comfortable doing this 'step down' test after they have already tested from any treatment effect

\texttt{. test trt dose}

\begin{verbatim}
( 1) \text{trt} = 0
( 2) \text{dose} = 0
\end{verbatim}

\begin{verbatim}
\text{F}( 2, \text{ 36}) = 121.88
\text{Prob > F} =  0.0000
\end{verbatim}
13.5 Data driven model selection

· Suppose we look at a scatterplot before deciding which model we fit and choose a model that can fit the data well
  – If the data looks like a straight line, choose the model linear in dose
  – If the data looks like a U, choose a quadratic
  – If the data is a complicated pattern of differences among groups, we might choose dummy variables or splines
  – etc.

· This approach would tend to mimic the behavior of fitting several different models and choosing the model with the lowest $p$-value
  – When our eye sees some trend in the data, we would be most likely to pick the model giving the lowest $p$-value

13.5.1 Simulation

· Using the 46 subjects in this dataset, I can randomly permute the dose they received
  – Effectively, randomize subjects to a different dose
  – But, keep their 9-month and baseline beta carotene levels the same (not permuted)
    * Should remove any association between dose and beta carotene

· Next, fit each of the five models (linear, quadratic, ANOVA, dichotomized, and dichotomized plus linear)

· Repeat the process 1000 times (representing 1000 studies)
– Calculate how often each model rejects the null hypothesis of a dose effect

• Individual Model Results

  – Empirical type I error for each method of analysis individually

<table>
<thead>
<tr>
<th>Model</th>
<th>Emp. Type-I error</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANOVA</td>
<td>0.049</td>
</tr>
<tr>
<td>Linear</td>
<td>0.046</td>
</tr>
<tr>
<td>Quadratic</td>
<td>0.046</td>
</tr>
<tr>
<td>Dichotomized</td>
<td>0.050</td>
</tr>
<tr>
<td>Dichot + linear</td>
<td>0.041</td>
</tr>
</tbody>
</table>

• Multiple comparison issues

  – With 5 hypothesis tests at a nominal 0.05 level, experiment-wise error rate is at most 0.25 (0.05 × 5)

  – Worst-case assumes that all tests are mutually exclusive
    * e.g. If the linear dose-response model is significant, no other model is more likely to be significant

    * In fact, the tests will be correlated

  – How many of the 1000 simulated trials had at least on model with a $p$-value < 0.05?
    * From the simulation, I found this to be 122 or 12.2%

    * Note that there is error in this estimate (due to the simulation randomness)
      * 95% CI: [10.2%, 14.2%]

• General statistical issues
– The true type 1 error rate for such data driven analyses will depend on several factors
  * The number of tests performed
  * The models considered
    · Similar models will tend to reject the null hypotheses on the same dataset
  * The distribution of the data
    · In particular, heavy tailed distributions decreases the concordance between the tests

• When you have multiple models you are considering, the conclusions are less strong
  – The p-values (or other metrics) can still be useful in ordering the associations

  – Among all of the models considered, it appears as if SNP X is the most strongly associated with CVD
    * Would be useful to put a CI around this ranking as well

13.5.2 Post hoc adjustments for multiple comparisons

• In frequentist reasoning, we try to ensure that our error rate is held at some level $\alpha$
  – When only considering one decision, this is relatively easy

  – When making multiple decisions, we must consider the experiment-wise error rate

• In the worst case scenario, an error rate of $\alpha$ one each decision could lead to an experiment-wise error rate that is as high as $k \times \alpha$
– Such would be the case if all of our errors were mutually exclusive

• If all error were independent of each other, then the experiment-wise error rate is
  \[ 1 - (1 - \alpha)^k \]

• Experiment-wise error rates ($\alpha = 0.05$ at each decision)

<table>
<thead>
<tr>
<th>Number of Comparisons</th>
<th>Worst Case Scenario Errors</th>
<th>Independent Errors</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.0500</td>
<td>0.0500</td>
</tr>
<tr>
<td>2</td>
<td>0.1000</td>
<td>0.0975</td>
</tr>
<tr>
<td>3</td>
<td>0.1500</td>
<td>0.1426</td>
</tr>
<tr>
<td>5</td>
<td>0.2500</td>
<td>0.2262</td>
</tr>
<tr>
<td>10</td>
<td>0.5000</td>
<td>0.4013</td>
</tr>
<tr>
<td>20</td>
<td>1.0000</td>
<td>0.6415</td>
</tr>
<tr>
<td>50</td>
<td>1.0000</td>
<td>0.9231</td>
</tr>
</tbody>
</table>

• When making multiple comparison which all tend address the same scientific question, we may adjust our level of significance to protect the experiment-wise error rate
  – The problem with this approach is does not adjust for any bias in parameter estimates

• Bonferroni Correction
  – Assumes the worst case scenario
  – When making $k$ comparisons, either
    * Tests individual $p$-values against $\frac{\alpha}{k}$
    * Multiply $p$-values by $k$ and compare to $\alpha$ (keeping the $p$-values < 1)

• Bonferroni is easy and it can be applied in all settings
– Extremely conservative when the statistics from various tests are positively correlated

• Many other varieties of adjusting after performing multiple comparisons
  – Tukey, Scheffe, etc.

– None are great

– Did they really adjust for all of the comparisons they made? Probably not.

– My strong preference is to avoid multiple comparisons in the first place
  * If there was some model fitting involved to get to the final model, acknowledge that fact in the paper

  * Understand the science

  * Avoid data-driven approaches when you care about correct statistical inference (CIs and p-values)