STOR 455
STATISTICAL METHODS I

Jan Hannig

Interactions

- Interaction: the effect of one variable depends on the value of another variable
  - Reinforcement or interference
    - The effect of going from color D to E is bigger for larger diamonds (carat*E interaction if D is baseline)
    - The effect of going from color D to E is larger for flawless diamonds (color*clarity interaction)
  - Make sure you can interpret the interactions you include
    - Carat*certification interaction is hard to interpret (everyone can measure the weight accurately – better not to include)
    - Color*certification interaction is still tricky but one can explain it (different standards in grading color?)

Peeps, Smoking
&
Alcohol

http://www.peepresearch.org/smoking.html

A Three-Part Study

Introduction: Physicians are continually warning the public of the risks involved in smoking and drinking. Here, we investigate whether these same health risks apply to Peeps.

Methods: To fully understand the effects of these substances, we decided to expose Peeps to each condition individually, and then in combination.

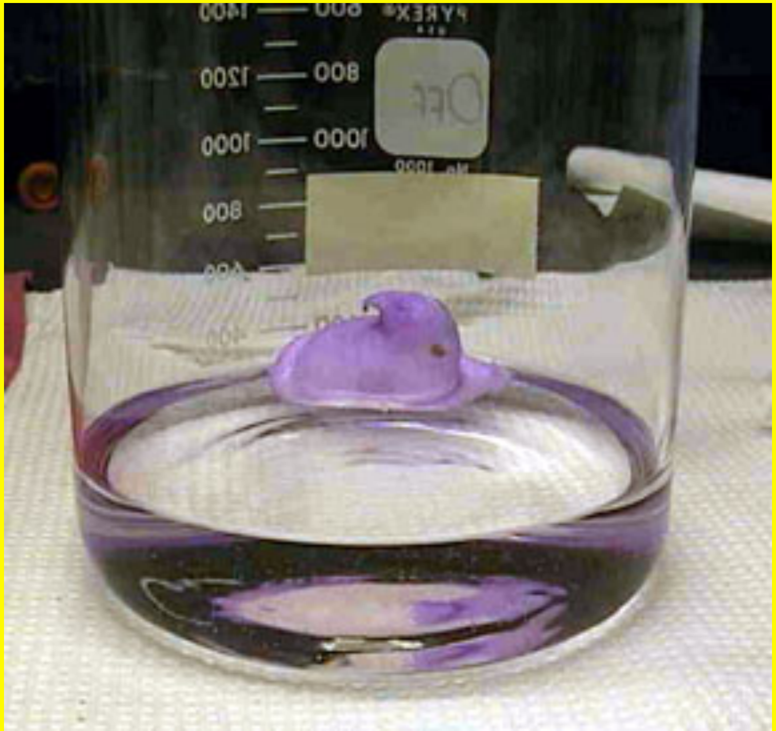
Step 1: Peeps and Alcohol
Step 2: Peeps and Cigarettes
Step 3: Peeps, Alcohol and Cigarettes

Step 1: Peeps and Alcohol

Materials:
This study began (as many do) with one gallon of 95% [190 proof] ethyl alcohol. The precise text on the label may be viewed by clicking on the image to the right.

McCormick
DISTILLING CO, INC.
INDUSTRIAL ALCOHOL DIVISION
USP
ETHYL ALCOHOL
K
95% – 190 PROOF
WARNING: FLAMMABLE, KEEP AWAY FROM SPARKS, HEAT AND OPEN FLAME
FLASH POINT 17 C° (60 F°)
MFG DATE 01/22/97
1 GALLON (3.785 liter)

Methods:
The peep imbibed liberally. Note that this Peep is having trouble swimming upright. This same peep bumped into the sides of the experimental chamber several times over the duration of this study.



Results: The test subject did report a moderate headache and some nausea the day following this study. Otherwise, however, this Peep was not substantially or permanently impaired.

Step 2: Peeps and Cigarettes

Materials: Our Peep subject was permitted to select the cigarette brand of its choice. Please remember that representatives of the tobacco industry insist that they have never marketed their products to young chicks.

The peep grabbed a smoke...

And lit up.

The subject began smoking...

and continued until its nicotine craving was fully satiated.

Results: The Peep showed no adverse signs to smoking this cigarette, although it expressed concern that it might someday be forced to huddle outside public buildings in the snow. Of course, it can quit anytime it wants to...

Step 3: Peeps, Cigarettes and Alcohol

Lesser scientists might have decided that alcohol and tobacco are benign substances after seeing the first two studies. We decided, however, to plunge deeper into this investigation...

Materials: Here, we bring together the elements from the first two parts of the study.

Unfortunately, the subject began to show some adverse signs to the combination of treatments. Note in particular the blackening along the lower extremities, the faint flame surrounding the subject, and the mild scent of caramelizing sugar.

The initial symptoms began to be expressed more definitively...

...soon the flames became quite substantial, initiating a metamorphosis. The subject entered the stage which scientists call "ball of charred goo".

Flames were extinguished shortly thereafter.

Conclusion: The synergistic effect of smoking and alcohol in Peeps produces a rapidly exothermic oxidation reaction, leading to a chemical and morphological divergence from the wild-type Peep phenotypes.

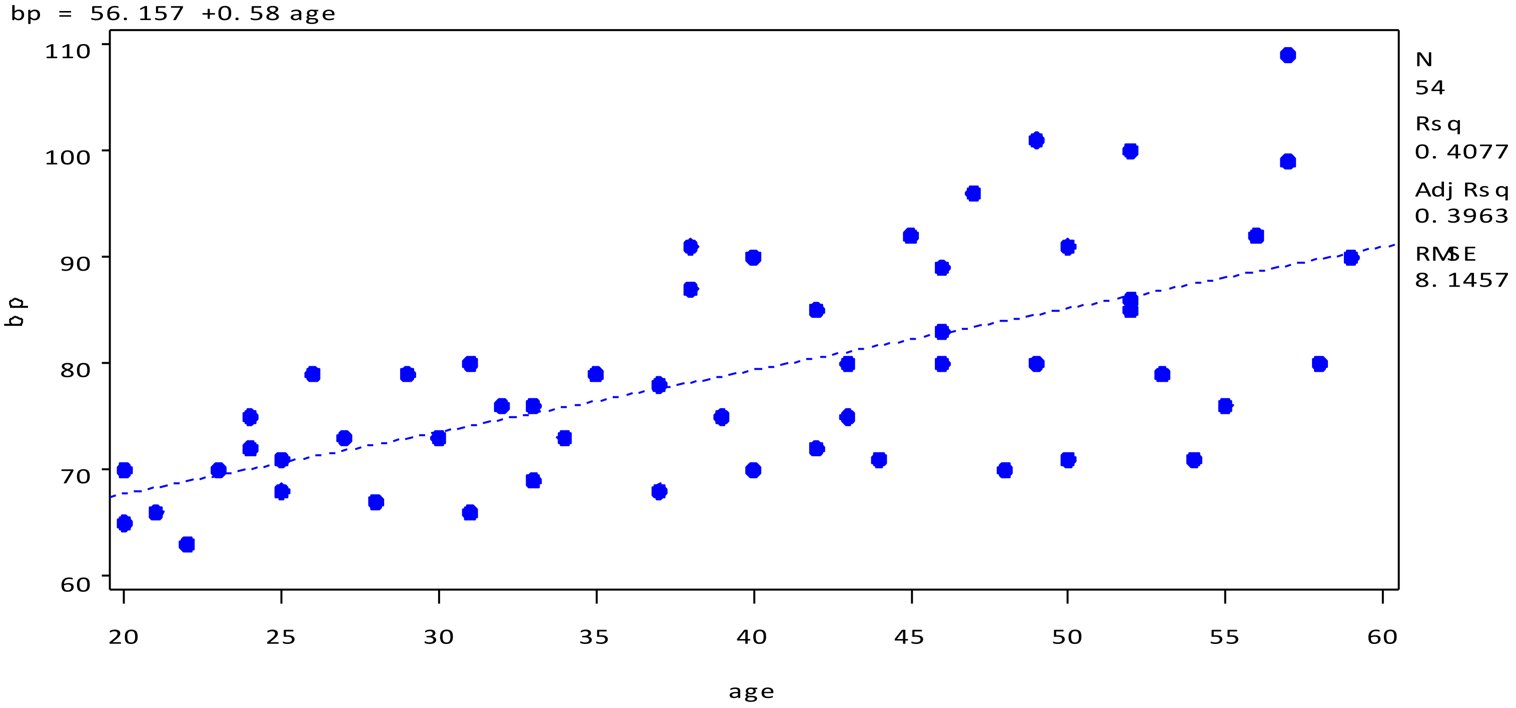
Remedial Measures

- If transformations fail we have other options
  - Unequal variance: Weighted least squares (Section 8.2)
  - Multicollinearity: Ridge regression
  - Influential cases: Robust regression (Section 8.3)
  - Nonlinear response: Nonparametric regression Evaluating uncertainty: bootstrapping

Blood Pressure Example

- Y is diastolic blood pressure
- X is age
- n = 54 healthy adult women aged 20 to 60 years old
- Scatter plot show non-constant variance

Do it in SAS


bp = 56. 157 +0. 58 age
110
100
90
80
70
60
bp
20
25
30
35
40
45
50
55
60
age
N
54
Rsq
0. 4077
Adj Rsq
0. 3963
RMSE
8. 1457

Weighted Least Squares (Section 8.2)

- Transformation may create other problems
- Generalize regression model: relax the assumption to allow different variances
- LS estimators still unbiased and consistent, but no longer have minimum variance
- WLS: minimize the the sum of weighted squared residuals

Weighted Least Squares

- OLS minimize

$$SSE = \sum_{i=1}^{n}(y_i - b_0 - b_1 x_{1,i} + \cdots + b_{p-1}x_{p-1,i})^2 = (Y - \mathbf{X}b)'(Y - \mathbf{X}b)$$

- WLS minimize

$$WSSE = \sum_{i=1}^{n} w_i(y_i - b_0 - b_1 x_{1,i} + \cdots + b_{p-1}x_{p-1,i})^2 = (Y - \mathbf{X}b)'W(Y - \mathbf{X}b)$$

$$W = \begin{pmatrix} w_1 & 0 & \dots & 0 \\ 0 & w_2 & \dots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \dots & w_n \end{pmatrix}$$

  - This gives $b_w = (X'WX)^{-1}(X'WY)$
  - Confidence intervals and tests are similar to before (see the book for formulas)
  - W and cW give the same results

Weighted Least Squares

- If the model is $Yi=\beta_0+b_1\ x_{1i} + \ldots +b_{p-1}\ x_{p-1i}+g_i\ \xi_i \quad \xi_i$ iid $N(0,\sigma^2)$
- Optimal weights: proportional to inverse variance $w_i=1/g_i^2$
- Often g_i themselves are related to the $\mathbf{x}_i$ and can be estimated from the residuals.

Determine the weights

- Method I: find a relationship between the absolute/ squared residuals and another variable and use this as a model for the standard deviation/variance
- Method II: use grouped data or approximately grouped data to estimate the variance
- Method III: use nonparametric method to estimate variance function
- Weights are proportional to the inverse of the estimated variance

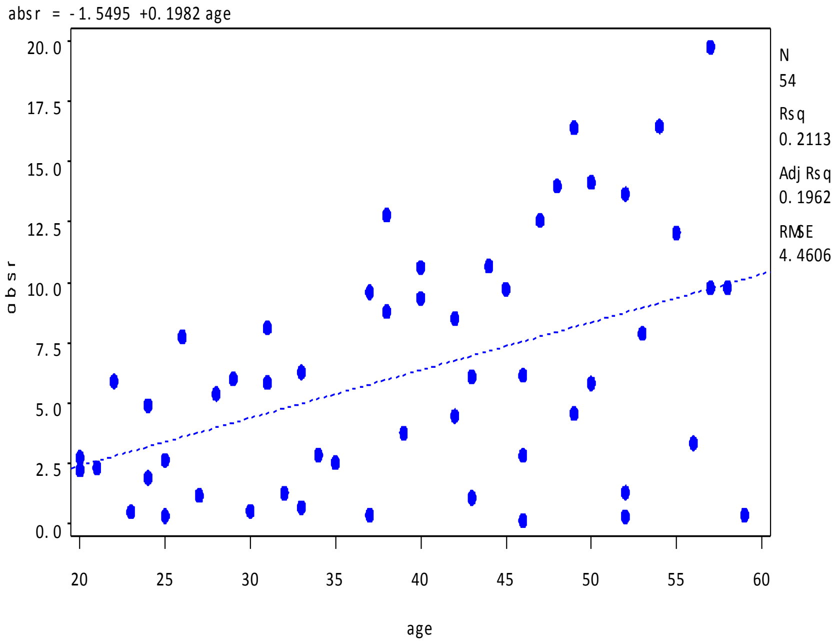
Do it in SAS

```
* Output residuals from
  proc reg;
proc reg data=dias;
  model bp = age /clb;
  output out=d1
   r=residual;
run;
* transform residuals;
data d1; set d1;
  absr = abs(residual);
run;
```

```
* estimate the s.d. using
  LS;
proc reg data = d1;
  model absr = age;
  output out = d2 p = s ;
run;

* Weights correspond to
  inverse variance;
data d2;
  set d2;
  w = 1/(s**2);
run;
```

Do it in SAS



```
*regression with
  weights;
proc reg data =
  d2;
  weight w;
     model bp =
  age / clb;
run;
```

Do it in SAS: OLS

```
                         Analysis of Variance

                              Sum of          Mean
Source              DF       Squares        Square    F Value    Pr > F

Model                1    2374.96833    2374.96833      35.79    <.0001
Error               52    3450.36501      66.35317
Corrected Total     53    5825.33333


        Root MSE              8.14575    R-Square     0.4077
        Dependent Mean       79.11111    Adj R-Sq     0.3963
        Coeff Var            10.29659


                          Parameter Estimates

                 Parameter     Standard
Variable    DF    Estimate        Error   t Value   Pr > |t|     95% Confidence Limits

Intercept    1    56.15693      3.99367     14.06     <.0001     48.14304     64.17082
age          1     0.58003      0.09695      5.98     <.0001      0.38548      0.77458
```

Do it in SAS: WLS

```
                         Analysis of Variance

                             Sum of          Mean
Source               DF      Squares         Square    F Value    Pr > F

Model                 1     83.34082      83.34082      56.64    <.0001
Error                52     76.51351       1.47141
Corrected Total      53    159.85432


         Root MSE             1.21302    R-Square     0.5214
         Dependent Mean      73.55134    Adj R-Sq     0.5122
         Coeff Var            1.64921


                     Parameter Estimates

               Parameter     Standard
Variable    DF      Estimate       Error   t Value   Pr > |t|     95% Confidence Limits

Intercept    1     55.56577     2.52092     22.04    <.0001     50.50718     60.62436
age          1      0.59634     0.07924      7.53    <.0001      0.43734      0.75534
```

Ridge regression

- Remedy for multicollinearity
- Allow bias while reduce variance
- (X' X) is difficult to invert (near singular) approximate it by inverting (X' X+cI), c bias constant (a small positive number)
- Interesting but has not turned out to be a useful method in practice

Robust regression

- Remedy for influential cases /outliers
- Basic idea: Procedures insensitive to outliers
- LAR regression: minimize sum of absolute values of residuals
- LMS regression: minimize Median of the squares of residuals
- IRLS regression: down weight cases with high residuals
- The book gives a particular robust regression method for straight line (not covered).

Nonparametric regression

- Remedy for nonlinear response surface
- Several versions: locally weighted regression, smoothing splines, etc.
- Compromise between over fit and over smoothing
- Choose smooth parameter

Bootstrap

- Very important theoretical development that have a major impact on applied statistics
- Based on simulation
- Sample *with* replacement from the data or residuals and get the distribution of the quantity of interest
- CI based on quantiles of the sampling distribution