Statistics 512: Applied Linear Models

Topic 9

Topic Overview

This topic will cover

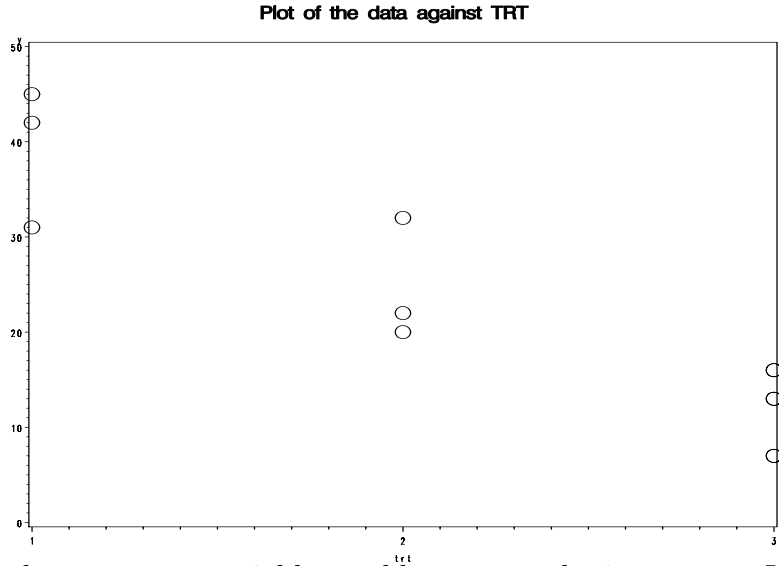
- One-Way Analysis of Covariance (ANCOVA) (§22)
- ANCOVA With More Than One Factor / Covariate (§22)

One-way Analysis of Covariance

ANCOVA is really "ANOVA with covariates" or, more simply, a combination of ANOVA and regression used when you have some categorical factors and some quantitative predictors. The predictors (X variables on which to perform regression) are called "covariates" in this context. The idea is that often these covariates are not necessarily of primary interest, but still their inclusion in the model will help explain more of the response, and hence reduce the error variance.

Example: An Illustration of why ANCOVA can be important

Our response Y is the number of months a patient lives after being placed on one of three different treatments available to treat an aggressive form of cancer. We could analyze these treatments with a one-way ANOVA as follows:

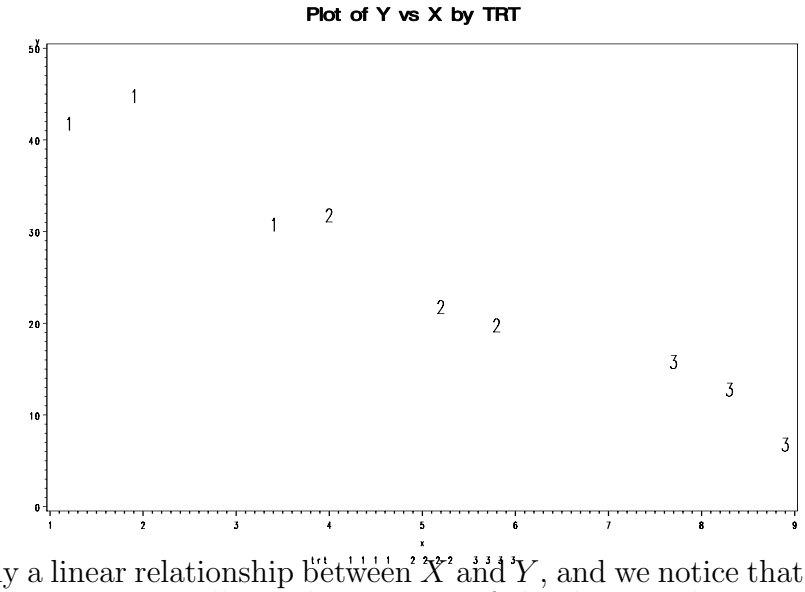


At first glance, the treatment variable would appear to be important. In fact if we run the one-way analysis of variance we get:

```
Dependent Variable: y
                                        Sum of
Source                          DF         Squares     Mean Square    F Value    Pr > F
Model                            2     1122.666667      561.333333      14.43    0.0051
Error                            6      233.333333       38.888889
Corrected Total                  8     1356.000000

          Mean      N    trt
A        39.333      3    1
B        24.667      3    2
C        12.000      3    3
```

The analysis tells us that there is a big difference between the treatments. Treatment 1 is clearly the best as people live longer. Suppose we put a large group of people on Treatment 1 expecting them to live 30+ months only to find that over half of them die prior to 25 months. What did we do wrong????

It turns out that we have neglected an important variable. We need to consider X, the stage to which the cancer has progressed at the time treatment begins. We can see its effect in the following plot:



There is clearly a linear relationship between X and Y, and we notice that the group assigned to the first treatment were all in a lower stage of the disease, those assigned to treatment 2 were all in a mid-stage, and those assigned to treatment 3 were all in a late stage of the disease. We would suspect looking at this plot to find the treatments are not all that different.

The following ANCOVA output leads to the same conclusion:

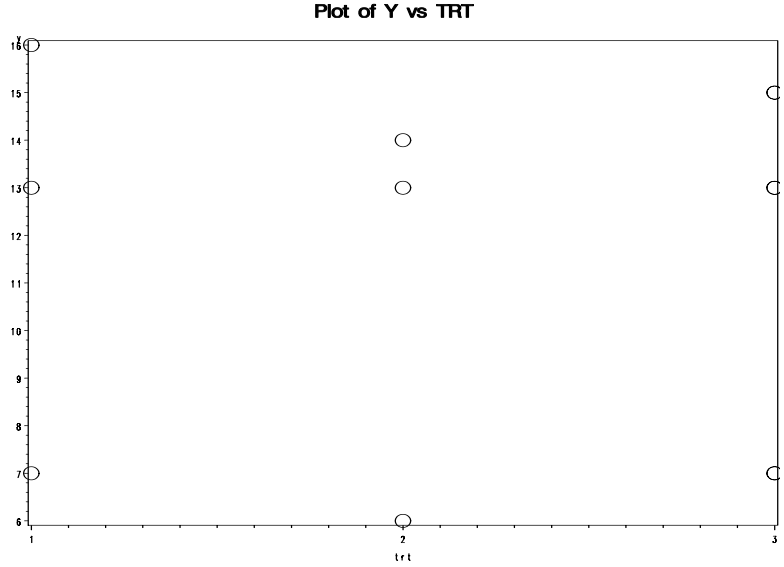
```
                        Sum of
Source          DF         Squares     Mean Square    F Value    Pr > F
x                1     1297.234815     1297.234815     192.86    <.0001
trt              2       25.134378       12.567189       1.87    0.2478
Error            5       33.630807        6.726161
```

```
Total         8     1356.000000

                          LSMEAN
trt        y LSMEAN       Number
1        20.3039393            1
2        23.6762893            2
3        32.0197715            3

      Least Squares Means for Effect trt
   t for H0: LSMean(i)=LSMean(j) / Pr > |t|
i/j              1             2             3
   1                     -0.85813      -1.56781
                          0.4300        0.1777
   2      0.858127                     -1.89665
            0.4300                       0.1164
   3      1.567807       1.89665
            0.1777        0.1164
```

So the stage of the cancer was what actually was affecting the lifetime - it really didn't have anything to do with the choice of treatment. It just happened that everyone on treatment 1 was in an earlier stage of the disease and so that made it look like there was a treatment effect. And notice that if there was to be a difference, treatment 3 actually would have been the best. So to give everyone treatment 1 on the basis of our original analysis could have been a deadly mistake.

A Second Example

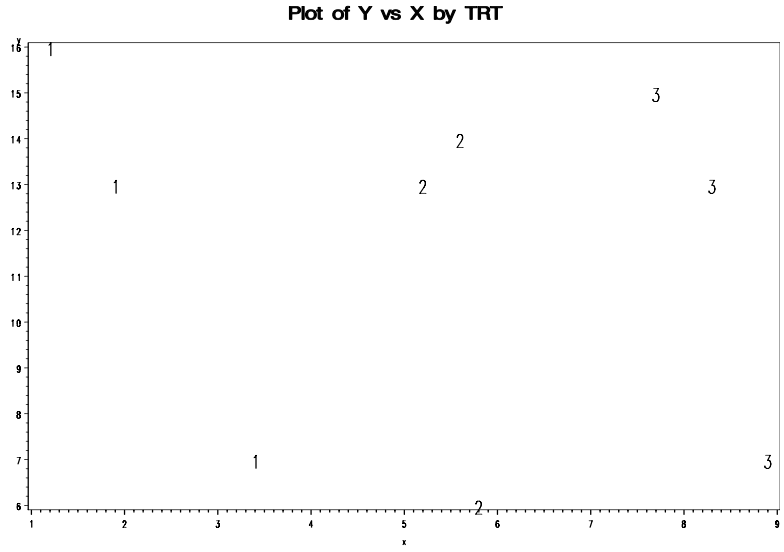
It is also possible to have a difference in means, but not be able to see it unless you first adjust for a covariate. Imagine a similar disease/treatment situation (but different data).




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

Model	2	1.5555556	0.7777778	0.04	0.9604
Error	6	114.6666667	19.1111111		
Total	8	116.2222222			

No significant differences between the treatments, right? WRONG! Consider now what happens when we consider the covariate X = stage of disease:



Now we see that there is probably a difference in means. Again all the treatment 1's were in the early stages of the disease, all the treatment 2's in the middle stages, and all the treatment 3's in the latter stages. But now treatment 3 would appear to be doing a better job since it is keeping those at the advanced stage of cancer alive just as long as those in the initial stages. If we look to the actual analysis:

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
x	1	6.97675624	6.97675624	1.11	0.3407
trt	2	77.76617796	38.88308898	6.18	0.0446
Error	5	31.4792880	6.2958576		
Total	8	116.2222222			

Note that X by itself was not significant. But we had to adjust for it before we could find the differences in the treatments. The output below indicates that treatment 3 is significantly better than the other two treatments. So this time the potentially deadly mistake would be to assume they were equivalent and give out the cheapest (unless you were lucky and that was treatment 3).

trt	y LSMEAN	LSMEAN Number
1	-3.5873786	1
2	11.9844660	2
3	26.2695793	3

Least Squares Means for Effect trt
t for H0: LSMean(i)=LSMean(j) / Pr > |t|

```
            Dependent Variable: y
i/j              1             2             3
   1                     -3.11551      -3.49022
                           0.0581        0.0390
   2      3.115508                     -3.3454
            0.0581                      0.0454
   3      3.490225      3.345401
            0.0390        0.0454
```

Notice that the `lsmean` estimate for the mean of Y with treatment 1 is negative. That's meant to be the mean of Y for an "average" stage of cancer ($\bar{X} = 5.3$) given trt 1. Since all trt 1 patients had $x < 3.5$ this is an unreasonable extrapolation. The interpretation breaks down (it would imply they were dead before treatment began) but the point is made that adjusting for covariates can seriously change your results.

Warning: As these examples illustrate, although ANCOVA is a powerful tool and can be very helpful, it cannot completely compensate for a flawed experimental design. In these two experiments we really haven't a clue how trt 1 behaves in late stage patients, or how trt 3 behaves in early stage patients. It would be foolish not to do another experiment with a proper design.

Data for one-way ANCOVA

- $Y_{i,j}$ is the jth observation on the response variable in the ith group
- $X_{i,j}$ is the jth observation on the covariate in the ith group
- $i = 1, \ldots, r$ levels (groups) of factor
- $j = 1, \ldots, n_i$ observations for level i

KNNL Example (page 926)

(`nknw1020.sas`)
Y is the number of cases of crackers sold during promotion period
Factor is the type of promotion ($r = 3$)

- Customers sample crackers in store
- Additional shelf space
- Special display shelves

$n_i = 5$ different stores per type
The *covariate* X is the number of cases of crackers sold in the preceding period.

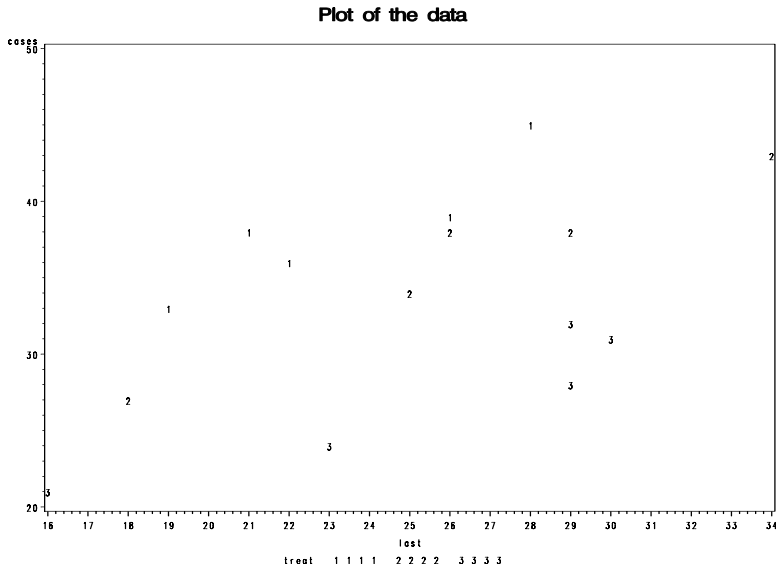
Data

```
data crackers;
infile 'h:\System\Desktop\CH25TA01.DAT';
    input cases last treat store;
proc print data=crackers;
```

Obs	cases	last	treat	store
1	38	21	1	1
2	39	26	1	2
3	36	22	1	3
4	45	28	1	4
5	33	19	1	5
6	43	34	2	1
7	38	26	2	2
8	38	29	2	3
9	27	18	2	4
10	34	25	2	5
11	24	23	3	1
12	32	29	3	2
13	31	30	3	3
14	21	16	3	4
15	28	29	3	5

Plot the data

```
title1 'Plot of the data';
symbol1 v='1' i=none c=black;
symbol2 v='2' i=none c=black;
symbol3 v='3' i=none c=black;
proc gplot data=crackers;
   plot cases*last=treat;
```

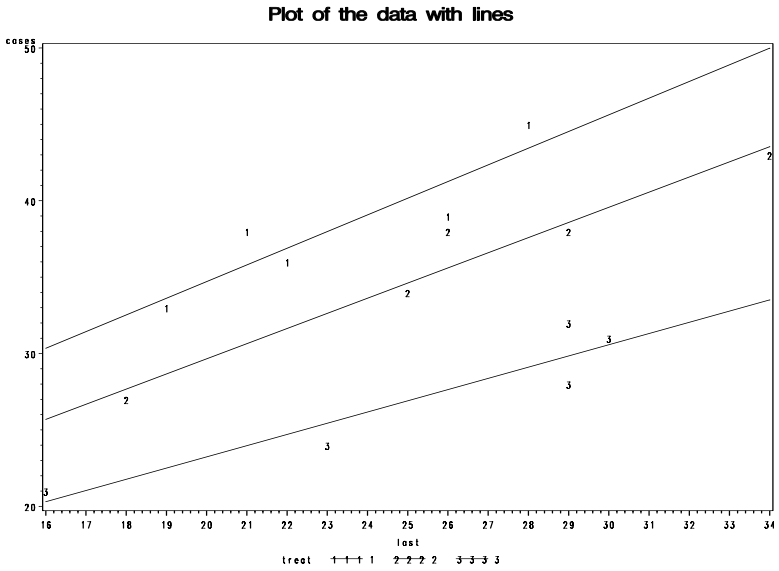


Basic Ideas Behind ANCOVA

- Covariates (sometimes called *concomitant* variables) can reduce the MSE, thereby increasing power for testing. Baseline or pretest values are often used as covariates.
- A covariate can adjust for differences in characteristics of subjects in the treatment groups. It should be related ONLY to the response variable and not to the treatment variables (factors).
- We assume that the covariate will be *linearly related* to the response and that the relationship will be the *same* for all levels of the factor (no interaction between covariate and factor).

Plot of the data with lines

```
title1 'Plot of the data with lines';
symbol1 v='1' i=rl c=black;
symbol2 v='2' i=rl c=black;
symbol3 v='3' i=rl c=black;
proc gplot data=crackers;
   plot cases*last=treat;
```



Cell Means Model

- $Y_{i,j} = \mu_i + \beta_1(X_{i,j} - \bar{X}_{..}) + \epsilon_{i,j}$
- As usual the $\epsilon_{i,j}$ are iid $N(0, \sigma^2)$.
- $Y_{i,j} \sim N(\mu_i + \beta(X_{i,j} - \bar{X}_{..}), \sigma^2)$ independent

- For each i, we have a simple linear regression in which *the slopes are the same*, but the intercepts may differ (i.e. different means once covariate has been "adjusted" out).

Parameters and Estimates

- The parameters of the model are μ_i for $i = 1$ to r; β_1, and σ^2
- We use multiple regression methods to estimate the μ_i and β_1
- We use the residuals from the model to estimate σ^2 (using the MSE)

Factor Effects Model for one-way ANCOVA

- $Y_{i,j} = \mu + \alpha_i + \theta_1(X_{i,j} - \bar{X}_{..}) + \epsilon_{i,j}$
- $\epsilon_{i,j} \sim^{iid} N(0, \sigma^2)$
- The usual constraints are $\sum \alpha_i = 0$ (or in SAS $\alpha_a = 0$)
- Note the deliberate use of θ instead of β for the slope to avoid confusion with a factor B

Interpretation of model

- Expected value of a Y with level i and $X_{i,j} = x$ is $\mu + \alpha_i + \theta_1(x - \bar{X}_{..})$
- Expected value of a Y with level i' and $X_{i',j} = x$ is $\mu + \alpha_{i'} + \theta_1(x - \bar{X}_{..})$
- Of note is that the difference $\alpha_i - \alpha_{i'}$ does NOT depend on the value of x.

```
proc glm data=crackers;
   class treat;
   model cases=last treat/solution clparm;
```

```
                                    Sum of
Source                      DF         Squares     Mean Square    F Value    Pr > F
Model                        3     607.8286915     202.6095638      57.78    <.0001
Error                       11      38.5713085       3.5064826
Corrected Total             14     646.4000000

R-Square     Coeff Var      Root MSE    cases Mean
0.940329      5.540120      1.872560      33.80000

Source                      DF       Type I SS     Mean Square    F Value    Pr > F
last                         1     190.6777778     190.6777778      54.38    <.0001
treat                        2     417.1509137     208.5754568      59.48    <.0001

Source                      DF     Type III SS     Mean Square    F Value    Pr > F
last                         1     269.0286915     269.0286915      76.72    <.0001
```

```
treat                       2     417.1509137     208.5754568      59.48    <.0001

                                      Standard
Parameter          Estimate              Error   t Value   Pr > |t|     95% Confidence Limits
Intercept       4.37659064 B     2.73692149      1.60     0.1381    -1.64733294  10.40051421
last            0.89855942       0.10258488      8.76     <.0001     0.67277163   1.12434722
treat     1    12.97683073 B     1.20562330     10.76     <.0001    10.32327174  15.63038972
treat     2     7.90144058 B     1.18874585      6.65     <.0001     5.28502860  10.51785255
treat     3     0.00000000 B      .               .        .          .            .
```

The estimate for the common slope is $\hat{\theta}_1 = 0.9$, and notice that its confidence interval contains 1 (we'll use that later). The option '`clparm`' can be used to get confidence intervals on the parameters. Note, however, that only CI's for *unbiased* estimates (in this case the slope for `last`) are appropriate.

Interpretation

- The expected value of Y with level i of factor A and $X = x$ is $\mu + \alpha_i + \theta_1(x - \bar{X}_{..})$.
- So $\mu + \alpha_i$ is the expected value of Y when X is equal to the average covariate value
- This is commonly the level of X where the treatment means are calculated (for this to be interpretable, need to make sure this level of X is reasonable for each level of the factor)

`LSMEANS`

- The L(least)S(square) means can be used to obtain these estimates
- All other categorical values are set at an equal mix for all levels (i.e., average over the other factors)
- All continuous values are set at their overall means

Interpretation for KNNL example

- Y is cases of crackers sold under a particular promotion scenario
- X is the cases of crackers sold during the last period
- The `LSMEANS` are the estimated number cases of crackers that would be sold with a given treatment for a store with average cracker sales during the previous period

LSMEANS Statement

```
proc glm data=crackers;
  class treat;
  model cases=last treat;
  lsmeans treat/stderr tdiff pdiff cl;
```

- STDERR gets the standard errors for the means (in the first part of the output)
- TDIFF requests the matrix of statistics (with p-values) that will do pairwise comparisons. You can use this along with ADJUST = TUKEY (or BON, SCHEFFE, DUNNETT) to apply multiple comparison procedures.
- PDIFF requests only the matrix of p-values for the pairwise comparisons (may use ADJUST)
- CL gets confidence limits for the means. When used in conjunction with PDIFF, it also provides confidence limits for the pairwise differences using whatever adjustment you specify.

```
Least Squares Means
                              Standard                    LSMEAN
treat    cases LSMEAN            Error    Pr > |t|       Number
1          39.8174070       0.8575507      <.0001            1
2          34.7420168       0.8496605      <.0001            2
3          26.8405762       0.8384392      <.0001            3

     Least Squares Means for Effect treat
   t for H0: LSMean(i)=LSMean(j) / Pr > |t|
          Dependent Variable: cases
i/j              1             2             3
   1                     4.129808      10.76359
                           0.0017        <.0001
   2      -4.12981                     6.646871
            0.0017                       <.0001
   3      -10.7636      -6.64687
            <.0001        <.0001

treat    cases LSMEAN      95% Confidence Limits
1          39.817407       37.929951    41.704863
2          34.742017       32.871927    36.612107
3          26.840576       24.995184    28.685968

        Least Squares Means for Effect treat
            Difference
               Between    95% Confidence Limits for
i    j           Means       LSMean(i)-LSMean(j)
1    2        5.075390        2.370456     7.780324
1    3       12.976831       10.323272    15.630390
2    3        7.901441        5.285029    10.517853
```

```
NOTE: To ensure overall protection level, only probabilities associated with pre-planned
      comparisons should be used.
```

From this output we see that the means (adjusted for covariate) are significantly different for each treatment and so the first treatment is superior. Allowing food sampling in the store appears to increase sales. Without the covariate we would not see this, as treatments 1 and 2 would test to be equivalent.

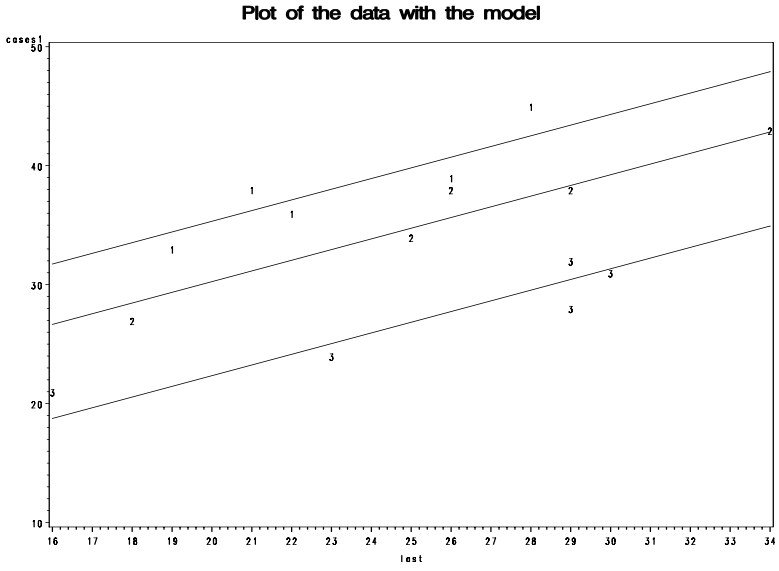
Prepare data for plot

```
title1 'Plot of the data with the model';
proc glm data=crackers;
   class treat;
   model cases=last treat;
   output out=crackerpred p=pred;
data crackerplot; set crackerpred;
   drop cases pred;
   if treat eq 1 then do
     cases1=cases;
     pred1=pred;
     output; end;
   if treat eq 2 then do
     cases2=cases;
     pred2=pred;
     output; end;
  if treat eq 3 then do
     cases3=cases;
     pred3=pred;
     output; end;
proc print data=crackerplot;
```

Obs	last	treat	store	cases1	pred1	cases2	pred2	cases3	pred3
1	21	1	1	38	36.2232	.	.	.	.
2	26	1	2	39	40.7160	.	.	.	.
3	22	1	3	36	37.1217	.	.	.	.
4	28	1	4	45	42.5131	.	.	.	.
5	19	1	5	33	34.4261	.	.	.	.
6	34	2	1	.	.	43	42.8291	.	.
7	26	2	2	.	.	38	35.6406	.	.
8	29	2	3	.	.	38	38.3363	.	.
9	18	2	4	.	.	27	28.4521	.	.
10	25	2	5	.	.	34	34.7420	.	.
11	23	3	1	.	.	.	.	24	25.0435
12	29	3	2	.	.	.	.	32	30.4348
13	30	3	3	.	.	.	.	31	31.3334
14	16	3	4	.	.	.	.	21	18.7535
15	29	3	5	.	.	.	.	28	30.4348

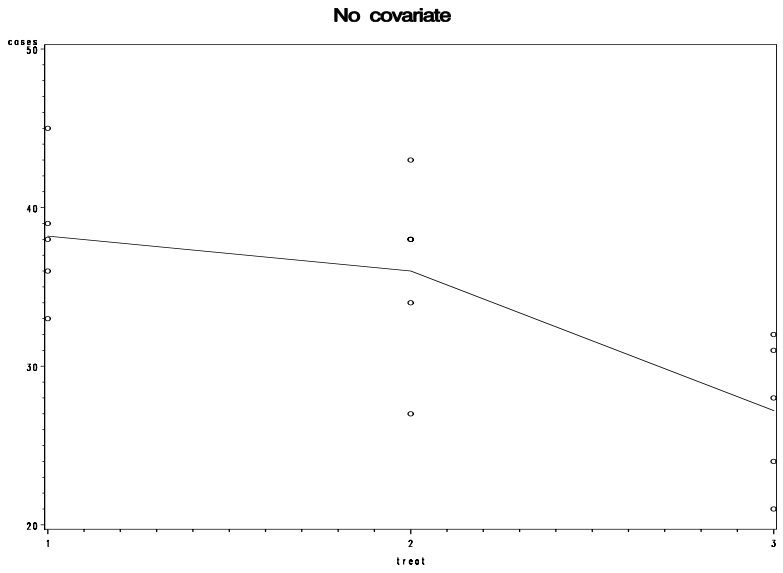
Code for plot

```
symbol1 v='1' i=none c=black;
symbol2 v='2' i=none c=black;
symbol3 v='3' i=none c=black;
symbol4 v=none i=rl c=black;
symbol5 v=none i=rl c=black;
symbol6 v=none i=rl c=black;
proc gplot data=crackerplot;
   plot (cases1 cases2  cases3 pred1 pred2 pred3)*last/overlay;
```



Prepare data for plot without covariate

```
title1 'No covariate';
proc glm data=crackers;
   class treat;
   model cases=treat;
   output out=nocov p=pred;
run;
symbol1 v=circle i=none c=black;
symbol2 v=none i=join c=black;
proc gplot data=nocov;
plot (cases pred)*treat/overlay;
```



Diagnostics and remedies

- Examine the data and residuals (check the three standard assumptions)
- Check the same-slope assumption
- Look for outliers that are influential
- Transform if needed, consider Box-Cox
- Examine variances (standard deviations). Look at MSE for models run separately on each treatment group (use a `BY` statement in `PROC REG` or `GLM`)

Check for equality of slopes

```
title1 'Check for equal slopes';
proc glm data=crackers;
   class treat;
   model cases=last treat last*treat;
```

```
                    Sum of
Source  DF         Squares     Mean Square    F Value    Pr > F
Model    5     614.8791646     122.9758329      35.11    <.0001
Error    9      31.5208354       3.5023150
Total   14     646.4000000

R-Square     Coeff Var      Root MSE    cases Mean
0.951236      5.536826      1.871447      33.80000

Source      DF       Type I SS     Mean Square    F Value    Pr > F
```

```
last         1     190.6777778     190.6777778      54.44    <.0001
treat        2     417.1509137     208.5754568      59.55    <.0001
last*treat   2       7.0504731       3.5252366       1.01    0.4032

Source      DF     Type III SS     Mean Square    F Value    Pr > F
last         1     243.1412380     243.1412380      69.42    <.0001
treat        2       1.2632832       0.6316416       0.18    0.8379
last*treat   2       7.0504731       3.5252366       1.01    0.4032
```

Analysis using differences

Recall that the CI for the slope included 1. So it is not unreasonable to assume a model that looks like

$$Y_{i,j} = \mu + \alpha_i + X_{i,j} + \epsilon_{i,j} \text{ where } \epsilon_{i,j} \sim N(0, \sigma^2)$$

This is the same as considering the one-way ANOVA model

$$Y_{i,j} - X_{i,j} = \mu + \alpha_i + \epsilon_{i,j} \text{ where } \epsilon_{i,j} \sim N(0, \sigma^2)$$

and so we can treat $Y_{i,j} - X_{i,j}$ as our response variable. This corresponds to the *increase* in sales over the previous period.

```
data crackerdiff;
     set crackers;
     casediff = cases - last;
proc glm data=crackerdiff;
     class treat;
     model casediff = treat;
     means treat / tukey;
```

```
                        Sum of
Source       DF         Squares     Mean Square    F Value    Pr > F
Model         2     440.4000000     220.2000000      62.91    <.0001
Error        12      42.0000000       3.5000000
Total        14     482.4000000

R-Square     Coeff Var      Root MSE    casediff Mean
0.912935      21.25942      1.870829         8.800000

           Mean      N    treat
A        15.000      5    1
B         9.600      5    2
C         1.800      5    3
```

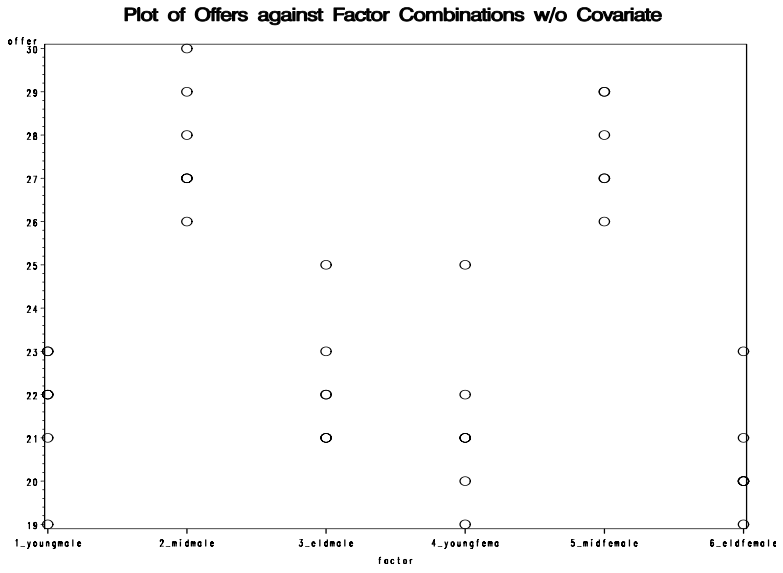
We see that the R^2 is about 0.03 less, but this is because the point estimate for slope was not exactly 1. We do get the same overall results - namely we conclude that treatment 1 is overall best. So this is a perfectly appropriate way to do the analysis in this case.

Two-way ANCOVA Example

- KNNL Problem 22.15 (`nknw1038.sas`)
- Y is offer made by a dealer on a used car (units \$100)
- Factor A is the age of person selling the car (young, middle, elderly)
- Factor B is gender of the person selling the car (male, female)
- Covariate is overall sales volume for the dealer
- This was a planned experiment using the same medium-priced, six-year old car

Plot data without covariate

```
data cash;
infile 'H:\System\Desktop\CH25PR15.DAT';
    input offer age gender rep sales;
*Look at the model without covariate;
data cashplot; set cash;
  if age=1 and gender=1 then factor = '1_youngmale';
  if age=2 and gender=1 then factor = '2_midmale';
  if age=3 and gender=1 then factor = '3_eldmale';
  if age=1 and gender=2 then factor = '4_youngfemale';
  if age=2 and gender=2 then factor = '5_midfemale';
  if age=3 and gender=2 then factor = '6_eldfemale';
symbol1 v=circle h=2;
title1 'Plot of Offers against Factor Combinations w/o Covariate';
proc gplot data=cashplot;
  plot offer*factor;
```



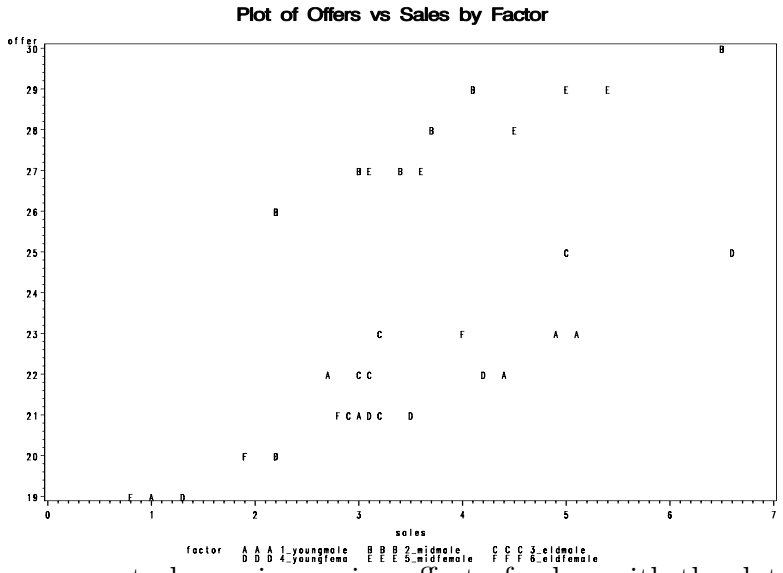
We appear to have differences based on age: namely it appears that dealers may offer less money to the young and elderly. This is backed up by the following two-way ANOVA model output:

```
proc glm data=cash;
  class age gender;
  model offer = age|gender;
  means age gender /tukey;
```

```
                                    Sum of
Source                      DF         Squares     Mean Square    F Value    Pr > F
Model                        5     327.2222222      65.4444444      27.40    <.0001
Error                       30      71.6666667       2.3888889
Corrected Total             35     398.8888889

R-Square     Coeff Var      Root MSE    offer Mean
0.820334      6.561523      1.545603      23.55556

Source                      DF       Type I SS     Mean Square    F Value    Pr > F
age                          2     316.7222222     158.3611111      66.29    <.0001
gender                       1       5.4444444       5.4444444       2.28    0.1416
age*gender                   2       5.0555556       2.5277778       1.06    0.3597

           Mean      N    age
A       27.7500     12    2

B       21.5000     12    1
B
B       21.4167     12    3
```

Now let's consider the covariate:

```
symbol1 v=A h=1 c=black;
symbol2 v=B h=1 c=black;
symbol3 v=C h=1 c=black;
symbol4 v=D h=1 c=black;
symbol5 v=E h=1 c=black;
symbol6 v=F h=1 c=black;
title 'Plot of Offers vs Sales by Factor';
proc gplot data=cashplot;
  plot offer*sales=factor;
```



Notice that there seems to be an increasing effect of sales, with the data dividing into two clusters. The top cluster is the middle-age group. We conduct the two-way ANCOVA:

```
proc glm data=cash;
  class age gender;
  model offer=sales age|gender;
  lsmeans age gender /tdiff pdiff cl adjust=tukey;
```

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	390.5947948	65.0991325	227.62	<.0001
Error	29	8.2940941	0.2860032		
Corrected Total	35	398.8888889			

R-Square	Coeff Var	Root MSE	offer Mean
0.979207	2.270346	0.534793	23.55556

Source	DF	Type I SS	Mean Square	F Value	Pr > F
sales	1	157.3659042	157.3659042	550.22	<.0001
age	2	231.5192596	115.7596298	404.75	<.0001
gender	1	1.5148664	1.5148664	5.30	0.0287
age*gender	2	0.1947646	0.0973823	0.34	0.7142

Source	DF	Type III SS	Mean Square	F Value	Pr > F
sales	1	63.3725725	63.3725725	221.58	<.0001
age	2	232.4894513	116.2447257	406.45	<.0001
gender	1	1.5452006	1.5452006	5.40	0.0273
age*gender	2	0.1947646	0.0973823	0.34	0.7142

Notice that using the covariate allows us to see a significant effect of gender which we could not see before. Age and sales are also both significant. Note also the much-reduced MSE (was 2.4 without covariate (i.e. $s = \$155$), now is 0.29 (i.e. $s = \$53$)). Look at the comparisons for age:

```
                        LSMEAN
age    offer LSMEAN     Number
1        21.4027214          1
2        27.2370766          2
3        22.0268687          3

      Least Squares Means for Effect age
   t for H0: LSMean(i)=LSMean(j) / Pr > |t|
          Dependent Variable: offer
i/j              1             2             3
   1                     -26.507      -2.79334
                          <.0001        0.0241
   2      26.50696                    22.55522
            <.0001                      <.0001
   3      2.793336      -22.5552
            0.0241        <.0001
```

The effect we saw previously regarding age is still there - in addition it appears that the dealer offers young people even less money than the elderly, since groups 1 and 3 are significantly different.

```
                           H0:LSMean1=LSMean2
gender    offer LSMEAN    t Value    Pr > |t|
1           23.7646265       2.32      0.0273
2           23.3464846
```

We also see that the dealers offer slightly less money to women than men. The difference in means is very small ($42) but the standard error is so small that this is significant.