Excel users: Some of you who used Excel took my instructions about "calculating the residual variation, manually" quite literally -- i.e., using a calculator. But Excel can do these calculations for you if you ask it (i.e. via formulae) So I have put in the Resources for Ch 10 an Excel sheet that does the calculations "manually" for the "Variability of, and trends in, proportions of smokers" from the previous set of exercises. Click on some of the cells to see how the calculations, particularly the squared residuals, were programmed "manually" (Instead of manually, I suppose we should now say "programmed using some of the basic Excel formulae"). If you do not want to, or cannot, use the regression tool in the data analysis toolpak in Excel, you can use this template to get to the same results, but it takes a bit more Excel work on your side. I also put on the web site an analysis of the gas vs. degree-days data using a combination of the Regression tool and "build your own".

-1- Effect of adding insulation on domestic gas consumption

(cf. M&M exercise 2.40; p 150-151; data also under Resources for Ch 10)

M&M asked if 870 ft^3 is evidence that insulation reduced consumption (the predicted consumption is 932 ft^3). But they presented it as a purely arithmetic exercise and missed a perfect opportunity to use the same example to illustrate all of the concepts in the chapter. That is why I am following up this example: to illustrate **PREDICTION INTERVALS**, and to distinguish them from confidence intervals for a mean response (and to wax epidemiological along the way!)

Clearly one should not compare the consumption in a post-insulation month (February, when the average number of degree days for the month was 40) with the corresponding pre-insulation consumption in February of the previous year (when the average number of degree days was 35.5).

The "closest" (in degree-days) other pre-insulation comparison month was December (average number of degree days: 37.8), still not entirely satisfactory.

Instead, we can use the *pattern* of all of the previous year's data to estimate (or predict) what the expected gas consumption *would be* for a month where the degree-days was 40; we can then compare the observed post-insulation consumption with this better (less biased, and more precise) estimate of what the consumption would have been if one had not added insulation. You can think of this as a (mathematically) "temperature- adjusted" comparison. -- i.e., you create a "like-with-like" comparison *synthetically* (i.e., artificially), using a *statistical model.* The model (set of assumptions) is a "poor-person's way of estimating the effect non-experimentally!) The parameters of the model are estimated from *all* of the data . In other words, to come up with estimates of "possible Y's at X=40" we—to use Mosteller's phrase—"borrow strength" from the entire pattern of observed (Y,X) pairs.

Of course this strategy is only as good as the *assumptions* (model) on which it is based: The most important of these *in this example* are

(a) a straight line relationship between average monthly consumption and temperature [i.e., the *linear* relationship of $\mu_{Y|x}$ and x] ;

(b) *equal* and

(c) *Gaussian* variation (vertically) of Y's about this line [i.e., constant $\sigma_{Y|x}$].

The issue of Gaussian variance is not always critical [if all we are interested in is *means*], but it is critical here because (c) all we have is (unfortunately) is a *single* new Y value at X=40, rather than the (more stable and more Gaussian-behaved) average of *several* new Y values at X=40 [i.e., no CLT to help us!] and (b) we do not have enough Y data near X=40 to estimate the "local" Y variance "locally" [we have to "pool"—"*borrow"*—variance estimates from *wherever* we have Y points!]

In part (b) you derived the fitted equation, based only on data points pre-insulation,

$$\hat{\mu}_{Y|x} = 123.2 + 20.22\ x$$

with a RMSE [estimate of SD of all possible Y's at each X] of

$$\hat{\sigma}_{Y|x} = 43 \text{ cubic feet}$$ *[this 43 becomes critical below]*

In order to provide a synthetic comparison, **Question (c)** of the exercise then asked you to predict from the regression equation on average how much gas (Y) would be used at x* = 40-degrees per day without added insulation. So all of you substituted x* = 40, to get

$$\hat{\mu}_{Y|x^*} = 123.2 + 20.22\ x^* = 123.2 + 20.22(40) = 932\ ft^3.$$

M&M then asked you "DID THE INSULATION REDUCE GAS CONSUMPTION?". Most of you compared the actual (observed) 870 ft^3 post-insulation consumption with the "expected" (or predicted") 932 ft^3, obtaining an "observed" reduction of 62 ft^3, and said "YES, IT DID".

Is this conclusion justified? Just because it is stated as a numerical question doesn't mean that you should turn off your "faulty-scientific-reasoning detector" (the McMaster people call it a "C.R.A.P. Detector" [1]).

The "decrease" could be due to any one —or a combination—of several factors. Look again at the "pre" data. Even they do not all lie on the fitted line! Since we don't know the exact reasons for the deviations from this line, we call them "random" or "unexplained" or "we don't know why" variations or "errors" or "residuals", that had nothing to do with the added insulation. They were evident in the data you plotted for **part(a)**, and again as deviations from the fitted line [I know some of you were so impressed with the beautiful coloured line you got from INSIGHT or Excel that you didn't see the variations: we statisticians joke about engineers that they draw the line first and put the data in afterwards!]. Some other very particularistic reasons must be operating again this year -- even without any intervention. Maybe some or all of the people in the home went to Florida for a week, or had to stay longer hours (or more weekends) at work or school, or many other reasons. Or maybe the model is not perfectly linear in temperature anyway (even if kept everything else constant!) . Or for that matter, maybe [like the mother, seeing her army son out of step with the other marching soldiers, says "they are all out of step except my son Johnny"], the line for the previous (pre-added-insulation) year is wrong [unusually high, because ...], and that the new February datapoint is more mainstream and typical!

Of course, we are also Bayesian (and rightly so!) and *expecting* a reduction, *since it makes physical ("biological") sense*, even before seeing the specific new data. But we won't always have such clear or well founded reasons and priors when it comes to less well understood interventions.

So the question is: how likely is a deviation from the line of 63 or more, even with no intervention?

For starters, we could compare the –63 with the s = RMSE = 43 we estimated above. Seen against this 43, the observed "reduction" is only –62/43 = –1.44, which even in a (Z) Gaussian distribution where we *knew* σ = 43, would have some 7.5% of values below it (so P-value = 0.15 if 2-sided). Of course we don't *know* that σ is 43; our *best estimate* of σ is s=43, but it is based on only 9 datapoints leaving only n–2 = 7 "independent estimates of error" or "degrees of freedom with which to estimate σ. With 7 df, and reading from the t_7 table, a ratio of –1.44 stands nearer to the 10% spot, (p=0.20, 2-sided).

Second, who is to say that $\mu_{Y|x^*}$=40 = 932 is correct ? After all, the

---

[1] From a commentary by Ronald J. Kallen, M.D.: When a proposed empiric treatment doesn't "make sense" it sets off my C.R.A.P detector. This is a term I heard used many years ago by Dr. Harriet Dustan (once a prominent cardiology researcher at the Cleveland Clinic Foundation). More recently I have seen it translated as "Circular Reasoning or Anti-Intellectual Pomposity" by GR Norman and DL Streiner in "PDQ Statistics" (1986, published by BC Decker Inc., Toronto,)

line itself, and thus the 932, is only *an estimate*! Note the "hat" in my initial expression for the 932.

So when we subtracted 932 from 870, we forgot that *each* of these numbers has its own separate "error" or "source of variation" or "statistical uncertainty": The 932 has errors of estimation in it. And the 870 contains some amount of randomness from the true "post-insulation" line. It could be an "above the new line" or "below the new line" month.

Thus

$Y|40\text{"}\mathbf{post}\text{"}$ = 870 = $\mu_{Y|40\text{"}\mathbf{post}\text{"}}$ + random variation

$\hat{\mu}_{Y|40\text{"}\mathbf{pre}\text{"}}$ = 932 = $\mu_{Y|40\text{"}\mathbf{pre}\text{"}}$ + estimation error

So...

$$-62 = \mu_{Y|40\text{"}\mathbf{post}\text{"}} - \mu_{Y|40\text{"}\mathbf{pre}\text{"}} + (\text{random variation} - \text{estimation error})$$

I deliberately put the error in estimation of the line in red, to emphasize that the fitted line produced by INSIGHT, is, despite the nice colour, an estimate that contains some error.

How to estimate the amalgam of the two uncertainties in the –62?

We can use "Rule 2" for the variance of the difference of two independent random variables (M&M Chapter 4, p 337), since the two sources of variation/error are entirely separate from each other.

In other words,

Var(estimated difference) =

variance[random variation in new Y] (*)

+

variance[error in estimating $\mu_{Y|40\mathbf{pre}}$] (**)

We do not know whether the variation of a Y about the new line (*) will still have the same SD σ , or variance σ^2, as before, but let's (in the absence of any data for now) assume it will. [We could speculate that if the new line has a flatter slope, it probably will also have a smaller amplitude of random variations from it, but let's keep it simple for now].

This random variation of an individual Y from $\mu_{Y|40\mathbf{post}}$ is an "ε" in the terminology of the Simple Linear Regression Model in the box on M&M p665. So it has variance σ^2, which we estimate by 43^2. Note that the "ε" in this new Y has nothing to do with (is independent of) how well we estimated the $\mu_{Y|40\mathbf{pre}}$ from last year's data.

The error in estimating $\mu_{Y|40\mathbf{pre}}$ (**) is a function of several factors (a) how many data points—9—were used to estimate the line (b) how spread out the temperatures were (c) how far, on average, Y values would be from the line (our best estimate is 43) and (d) how far the X=40 is from the x values used to fit the line (it happens to be at the high end, just beyond the X=37.8. [The further X is from xbar, the bigger the impact of any errors in estimating the slope on the projection]. These four factors are all evident in the formula for the standard error for $\hat{\mu}_{Y|x^*}$ given in the first formula in the box in M&M p690.

SD's or SE's don't add; their squares do! Thus, to get (**), we square the right hand side of the first formula. This tells us how badly or precisely we estimated the "line of means". It tells us how much

faith we should put in the "thin red line", and whether it is realistic to draw it as a hairline with 600 dots per inch laser printer precision or (more realistically here) with a thick paintbrush!

The second formula in page 690 deals with the variation of a single new Y from the true mean Y at X. It may not be it obvious, but the sum of two estimated variances that I have described above, namely

$$s^2 + s^2\left\{\frac{1}{9} + \frac{\text{how far the new X=40 is from the centre of the x's}}{\text{how many and how spread out the x's were}}\right\}$$

is none other than the amalgamated variance

$$s^2\left\{1 + \frac{1}{9} + \frac{\text{how far the new X=40 is from the centre of the x's}}{\text{how many and how spread out the x's were}}\right\}$$

which, when one takes its square-root, gives the $SE[\hat{Y}]$ given on p. 690.

For those who still like to do calculations "manually", M&M pages 686-691 give a fully worked example with n=4.

In our example, xbar is approximately 21.5, so $(40 - 21.5)^2$ is approximately 380. $\Sigma(x - xbar)^2$ is approximately 1440, so that the SE for a new value at X=40 under the "pre" regime is approximately

$$SE[\hat{Y}] = 43\sqrt{1 + 1/9 + 380/1440}$$

$$= 43\sqrt{1 + 0.11 + 0.26}$$

$$= 50.$$

Note that apart from the s of 43, the biggest contribution to the SE comes from the "ε" (the natural variability from month to month even if adjust for temperature), the next largest from the fact that X=40 is a long ways from the data, and lastly from the fact that there are only 9 observations with which to begin at the middle: think of the line as pivoting about the point (xbar,ybar) .

With this $SE[\hat{Y}]$ of 50, the deviation of the new 870 from the estimated 932 is now only t = 62/50 = 1.2 SE's away from the null.

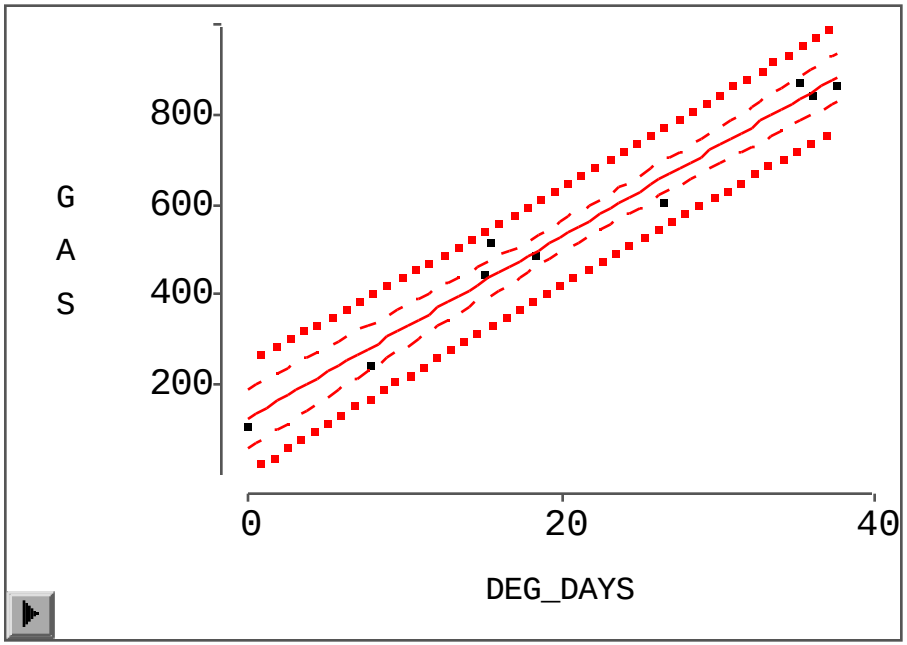
A number of you did consider measuring the 870 not from 932 but say from the lower limit associated with the 932. But from the printouts, it looks like you only considered the uncertainty in the fitted line of means, not the fact that *even if we knew exactly where the line (mean) should be when X=40, not every such month will have the exact same gas consumption*. In fact, as is clear from the calculation, (or from having SAS INSIGHT draw, or SAS PROC REG produce) the "prediction interval" for the individual new Y is a lot wider than the "confidence bands for the mean".

One way to appreciate the difference between confidence and prediction bands is to realize that the inner bands (the CI for the estimated line) can be narrowed by increasing n. But even if we had the large n to narrow—as much as we wished—the inner band towards the true line, the outside bands, which would now simply be $\mu_{Y|40} \pm Z \times SD$, can not be narrowed all the way to a line!

How—via software—to assess the deviation of 870 from value predicted under the null ?

- In Excel.. with Toolpack [see Resources Ch 10]
  (Confidence and Prediction Bands by JH)
- In SAS INSIGHT.. See 1st column, next page
- In SAS PROC REG...
  [see 2nd column next page .. program also in Resources Ch 10]

SAS INSIGHT



Bands:

Inner: – – – 95% CI for mean Consumption (μ_Y) at X

Outer: ▪ ▪ ▪ 95% Prediction interval for a new Y at X

Bands obtained from Curves Menu -- after Fit

Even though INSIGHT calculates the predicted and residual values, adding them to the dataset, it does not produce numerical values for the upper and lower limits of the bands

SAS PROGRAM EDITOR

Below I try to use `Courier font` for program and output, and ***Times font*** for commentary and notes

```
options ls=85 ps=35; run;
data sasuser.mmex2_40;
input deg_days gas;
lines;
15.6 520
26.8 610
37.8 870
36.4 850
35.5 880
18.6 490
15.3 450
 7.9 250
 0.0 110
40.0   .        (see note)
;
run;
```

note: ***Because the Y value is set to missing, this 10th observation won't be used in calculating the parameter estimates, but will have a prediction calculated***

```
Proc means data=sasuser.mmex2_40;
run;
```

Variable	N	Mean	Std Dev	Min	Max	
DEG_DAYS	10	23.4	13.9	0	40	*(37.8: see note)*
GAS	9	558.9	274.4	110	880	

```
Proc reg* data=sasuser.mmex2_40;
 model gas = deg_days / CLI1 CLM2;  **
 plot gas*deg_days = "*"
      predicted.3*deg_days="0" /
          CLEAR COLLECT OVERLAY;
```

*** SEE SYNTAX for PROC REG;**
In SAS 6.12, type "HELP REG" in Command box

1 Prediction Limits for Individual

2 Confidence Limits for Mean
If you specify CLI, CLM or R (residual) P (predicted) is unnecessary. See output.

3 See Syntax. And make sure to put a period at end of keyword [I suspect this is so that SAS can distinguish the various statistics calculated for each observation (predicted, residual, etc.) from 'regular' variables in dataset].

"*" and "0" are symbols you specify for plotting.

**** "Regression Diagnostics" are a set of measures calculated from the residuals, or by omitting 1 observation at a time. Request them with the option INFLUENCE after the "/" in the model statement. These will be covered under multiple regression in the Data Analysis I Course.**

Model: MODEL1
Dependent Variable: GAS

Analysis of Variance

Source[1]	DF	Sum of Squares	Mean Square	F Value[2]	Prob>F
Model	1	589071.3	589071.3	312.0	0.0001
Error	7	13217.5	1888.2		
C Total	8	602288.9			

1 see "Sums of Squares" in Resources for Ch 10
2 see Expected Mean Squares etc.. Interactive Excel spreadsheet

Root MSE	43.5	R-square	0.9781
Dep Mean	558.9	Adj R-sq	0.9749
C.V.	7.8		

<-- ***Coefficient of Variation of Y (%)***

Parameter Estimates

get in the habit of chopping off most of the decimal places given in the computer output, as I have here

Variable	DF[1]	Parameter Estimate[2]	Standard Error	T[3] for H0: Parameter=0	Prob > |T| [4]
INTERCEP	1	123.2	28.6	4.3	0.0035
DEG_DAYS	1	20.2	1.1	17.7	0.0001

[1] Degrees of Freedom. Always one per "term" in the equation; there will be k-1 terms (indicator variables) if "X" variable is categorical with k categories (e.g., k=4 blood groups, so 3 indicator variables or terms in the regression to represent them.

[2] "beta-hat" .
INTERCEP(T) ***has same units as Y. Can think of it as coefficient b_0 of the "variable" (actually a constant!) $X_0 \equiv 1$.***

Other b's have units Y/X. Need to know the X and Y units in order to interpret.

[3] parameter estimate / SE[parameter estimate]

[4] Alternative is 2-sided i.e. $\beta \neq 0$.

Output obtained by requesting the **CLI and CLM** options

Obs	GAS	Dep_Var Value[3]	Predict Std Err Predict*	Lower95% Mean[2]	Upper95% Mean[2]	Lower95% Predict[1]	Upper95% Predict[1]	Residual
1	520	438.7	16.0	400.8	476.5	329.2	548.2	81.3
2	610	665.2	15.6	628.1	702.3	555.9	774.4	-55.1
3	870	887.6	23.5	831.8	943.4	770.7	1004.5	-17.5
4	850	859.3	22.3	806.5	912.1	743.8	974.8	-9.2
5	880	841.1	21.5	790.1	892.1	726.4	955.8	38.9
6	490	499.3	14.8	464.2	534.5	390.7	608.0	-9.3
7	450	432.6	16.1	394.4	470.8	323.0	542.2	17.3
8	250	283.0	21.3	232.6	333.4	168.5	397.4	-32.9
9	110	123.2	28.6	55.6	190.9	0.2	246.3	-13.2
--								
10		**932.1**	**25.6**	**871.5**	**992.7**	**812.8**	**1051.4**	

-- **Space between obs 9 and "obs" 10 added by JH to emphasize the artificial nature of obs 10; it is not used to fit the parameters; rather, it is included so that we can calculate the predicted value and the confidence and prediction limits**

* ***Careful.. this is the SE for the mean, NOT for the individual. SE for individual Y (not shown) is always bigger than RMSE! See a few pages back how the SE of 50 something ft^3 was calculated from the s = 43 ft^3.***

[3] ***Same predicted value used for mean (μ) and individual (Y) ; limits very different!!***

$$SE[\hat{\mu}_{Y|40}] = 43\sqrt{1/9 + 380/1440} = 43\sqrt{0.11 + 0.26} = 25.6$$

Sum of Residuals	0[1]
Sum of Squared Residuals	13216[2]
Predicted Resid SS (Press)	19269[3]

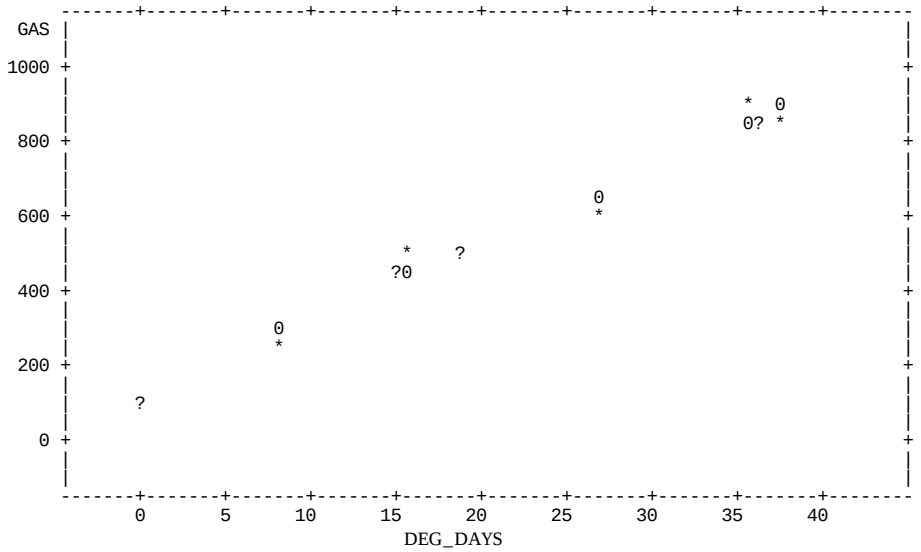
[1] Sum of Residuals = ***0 BY CONSTRUCTION!***

[2] ***Divide this by 9-2 = 7 to give Mean Squared Error. MSE used as estimate of σ^2 . $\sqrt{}$ MSE = RMSE (root of the mean squared error), or s, as an estimate of σ. See ANOVA table.***

[3] ***Leave out 1 observation at a time; see how well other observations predict it: more realistic estimate of performance than testing on same data from which model is estimated.***

Output obtained using the **PLOT** statement

Note the difference and placement between OPTIONS and STATEMENTS in SAS PROCEDURES



I asked SAS to use the "*" symbol for the observed data, and "o" for the fitted line. It uses ? when the two coincide.

Could also ask for (separate) plots of residuals versus predicted, or residuals versus X. CLEAR and COLLECT and OVERLAY specify which plots are to be overlaid and which are new (see Syntax in HELP)

I call the above graph a "typewriter plot" since the vertical resolution is only as good as the available line spacing and the horizontal resolution as good as the type size and the number of characters per line. How many of you have seen the—breakthrough for the time—"IBM Selectric" typewriter, still around and used by secretaries to type in a forms or a label. It had the letters on a removable "type-ball"? Nowadays, inkjet and laser printers, and operating systems that can display graphics, allow 300 and 600 and more dots per inch, rather than the vertical 6 lines per inch, and horizontal 8, 10 or 12 characters per inch available when we oldtimers were starting out.

High Resolution Graphical Output in SAS

SAS INSIGHT automatically uses high resolution graphics rather than typewriter plots.

Programs run as PROCedures in SAS version 8 also generate high resolution graphics. If using PROC REG in the Program Editor in SAS version 6.12, you can request them by specifying the `GRAPHICS` *option in the REG statement itself.*

```
Proc reg

data=sasuser.mmex2_40

          GRAPHICS;

model gas = deg_days /

      CLI clm;

plot

gas       *deg_days ="*"

predicted.*deg_days ="0"
/ CLEAR COLLECT OVERLAY;

run;
```

I copied the graph directly to word 5 on my Mac; one can also save it as a graphics file, and import it in later.

I haven't used this a lot, and am less sure what works well on the Windows side.

This is not be the best that SAS can do with graphics.