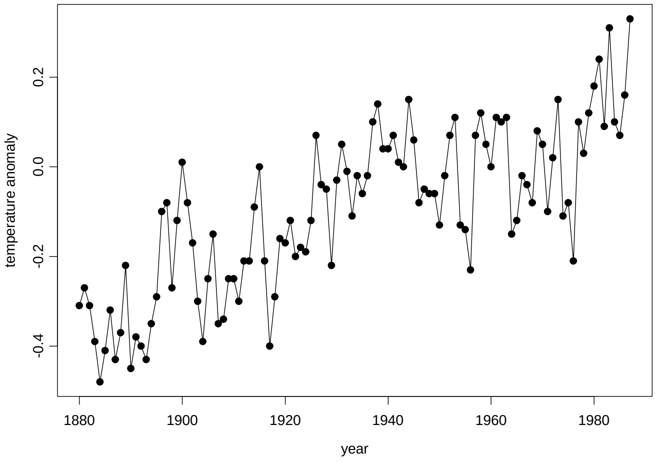
1. Trends in temperature. My SAS code:

```
data temp;
  infile 'c:/philip/stat500/data/temperature.txt' firstobs=2;
  input year temp;

proc reg;
  model temp=year / dwprob;
run;

proc loess;
  model temp = year / smooth=0.6 dfmethod=exact;
  run;

proc mixed;
  model temp = year /solution;
  repeated /subject = intercept type = ar(1);
  run;
```

(a) Estimated slope = 0.00449 , s.e. = 0.00035143

(b) The trend seems to be in three linear pieces. From 1880 to 1945 (approx.) we could fit a straight line, then from 1946 to 1987, the slope is close to zero, then the slope is positive again.



Your conclusion depends on your choice(s) of diagnostic. Some of the likely choices include:

residual plot – does not indicate much problem, but you can see traces of the different slope from the mid 1940's to mid 1980's.

adding quadratic or higher order terms – if you add year^2 to the model, the quadratic coefficient is not significantly different from zero.

chopping year into groups and using an ANOVA lack of fit test. I got p = 0.0021; there is strong evidence that the relationship is more complex than a straight line. The specific result depends on your choice of number of groups and where they start and stop.

using loess with a smoothing parameter of 0.6 gives F = (1.374-1.262) / (106-104.13) / (1.262 / 104.13) = 4.94. This would be compared to an $F_{1.87, 104}$ distribution. Other versions of SAS may give slightly different numbers. The p-value is 0.01.

(c) I got r = 0.45, using proc reg; model … /dwprob;. Other estimators (e.g. using proc mixed or calculating residuals, lagging them, then calculating the correlation) will give slightly different values. If you calculate the correlation between resid and lag(resid), you get 0.458. Fitting an AR(1) model in proc mixed gives r = 0.489. There is no 'right' value.

(d) The Durbin-Watson statistic is 1.069. This is a quite small value indicating positive temporal autocorrelation (because it is < 2). SAS reports p < 0.0001. There is strong evidence of temporally correlated errors.

(e) Using Proc mixed: Estimated slope = 0.00457, s.e. = 0.000595

If you used Cochran-Orcutt: Estimated slope = 0.00460, s.e. = 0.00058043

(f) No, the s.e.'s in (e) are almost twice the s.e. in (a).

Yes. In lecture I said that ignoring positive temporal correlation usually led to a s.e. that is too small. Here (a) is ignoring the correlation; the s.e. in (a) is too small.

2) SAS Code

```
filename website url
"http://www.public.iastate.edu/~pdixon/stat500/data/website.txt";

data website;
  infile website firstobs=2;
  input id deliver backlog team experience process year quarter;
run;

proc print data=website;
     title "Data Set website";
run;

proc reg data=website;
      model deliver = backlog experience process year/partial;
      output out = resids r = resid p = yhat;
run;

/* also possible to construct partial regr plots individually */
/*  code not shown, but following example in notes */
/*  this approach would require 2 additional regressions */
/*  Y= process X = rest,  and Y = experience, X = rest */

/* Breusch-Pagan, by hand */
data resid2;
      set resids;
      esquare =(73/1866.23573)*resid**2;
run;
```

```
proc reg data=resid2;
  model esquare=experience;
  title 'Information for Breusch-Pagan test-experience';
run;
proc reg data=resid2;
  model esquare=process;
  title 'Information for Breusch-Pagan test-process;
run;


/* modified B-P statistic that is less sensitive to normality*/
proc model data=website;
      parms b0 b1 b2 b3 b4;
      deliver=b0 + b1*experience + b2*backlog +b3*process + b4*year;
      fit deliver/breusch=(process);
run;
/* to look at experience, could add a second fit stmt to this proc */
/*  or write another proc model */
```

a) There does seem to be a slight increase in the variance with the means. The residual plot has two clusters of observations. These correspond to the two PROCESS groups.

b) It appears that the process added variable plot has the same spreading pattern seen in the plots of residuals vs. predicted values. This indicates that error variance is associated with process. The same pattern is seen less strongly with experience.

c) The regression of scaled squared residual against experience has SSmodel = 13.98. SSmodel / 2 = 6.99. Comparing this to quantiles of a Chi-square distribution with 1 d.f. gives p = 0.008. The proc model p-value is similar (p=0.01). There is evidence that the variance changes with experience.

d) The SS for the regression against process is even larger, 27.92. SSmodel / 2 = 13.96, which has a p-value of 0.0002. Again, the proc model p-value is similar (p=0.0003). There is a strong association between error variance and process.

3. Estimating the maximum – stainless steel

```
filename steel url
"http://www.public.iastate.edu/~pdixon/stat500/data/steel.txt";

data steel;
     infile steel;
       input temp pratio;
       if _n_ ne 1;
       temp2=temp*temp;
run;

proc print data=steel;
     title "Data Set steel";
run;
```

```
proc reg data=steel;
     title "Fit the Quadratic Regression";
       model pratio=temp temp2/covb;
       Xmax: test temp+800*temp2=0;
run;
```

e) From the output of SAS, the t statistic is -2.76, and the corresponding p-value is 0.0096 < 0.05. There's strong evidence that the quadratic coefficient is not 0.

f) Because $\widehat{\beta_2} < 0$, $\widehat{X_{max}} = -\frac{\widehat{\beta_1}}{2\widehat{\beta_2}} = 486$.

g) Test H_0: Max temp = 400 Celsius is equivalent to test $H_0: \beta_1 + 800\beta_2 = 0$. From the result of SAS, F statistic=3.40 and p-value=0.0743. So there's no evidence (or weak at best evidence) to reject the null hypothesis that the Poisson ratio is maximized at 400 Celsius.

h) Since the p-value is close to 0.05, $486 - 2\sigma$ should be close to 400, i.e. σ is around 40. So the s.e. of $\widehat{X_{max}}$ is moderate. We can also calculate the s.e. of $\widehat{X_{max}}$ directly using delta method which gives the result $s.e.(\widehat{X_{max}}) = 35$. This also suggests that the s.e. of $\widehat{X_{max}}$ is moderate.

(Note: the above discussion about the s.e. of $\widehat{X_{max}}$ is based on asymptotic normality in large sample sizes. The sample size for this problem is not large. The distribution of $\widehat{X_{max}}$ is definitely not normal. If you use simulation to estimate the s.e. of $\widehat{X_{max}}$, you find that it is very large.)

4. Model building problems

Note that other parameterizations are possible. Any reasonable choice was acceptable. Here are mine:

(a) Define:

oak = 1 if red oak, 0 if not

pine = 1 if white pine, 0 if not.

moisture = percent moisture

Model: $Energy = \beta_0 + \beta_1\, oak + \beta_2\, pine + \beta_3\, moisture + \varepsilon$

Desired quantity: $\beta_1 - \beta_2$

(b) X variables and model as in part (a)

Desired test: $\beta_1 = \beta_2 = 0$, full model as in (a); reduced model is $Energy = \beta_0 + \beta_3\, moisture + \varepsilon$

(c) Same X variables as in (a)

Model: $Energy = \beta_0 + \beta_1\ oak + \beta_2\ pine + \beta_3\ moisture + \beta_4\ oak*moisture + \beta_5\ pine*moisture + \varepsilon$

Desired quantity: $\beta_1 - \beta_2 + 10*(\beta_4 - \beta_5)$

(d) *age* is the only X variable

Model: $Y = \beta_0 + \beta_1\ age + \beta_2\ age^2 + \varepsilon$

Desired quantity: $-\beta_1 / 2\ \beta_2$, assuming that $\beta_2 < 0$

(e) If β_2 is the same for both males and females (as given in the problem), then if the age of maximum differs, then β_1 must differ. The desired test is whether males and females have the same β_1. The hint tells you that the intercepts probably differ, so your model should include a term that allows the two sexes to have different intercepts.

Define: *male* = 1 if male, 0 if female.

Model: $Y = \beta_0 + \beta_1 male + \beta_2 age + \beta_3 male * age + \beta_4 age^2 + \varepsilon$.

Test: t-test of $\beta_3 = 0$

(f) *dose* is the dose of the toxicant

Model: $Y = \beta_0 + \beta_1\ dose + \varepsilon$

Desired quantity: This is hardest part here: how is EC_{50} related to the linear regression parameters. If you draw a picture, you will see that $\beta_0 - \beta_0 / 2 =- EC_{50} * \beta_1$, hence $EC_{50} = -\beta_0 / 2\ \beta_1$

(g) There are a couple of ways to write this

Define: *hours* = # hours watching TV per week

$X_1 = 1$ if $hours < 20$, 0 otherwise

Model: $Y = \beta_0 + \beta_1 X_1\ hours + \beta_1 (1-X_1)\ (20) + \varepsilon$

Or, define $X = hours$ if $hours < 20$ and $X = 20$ if $hours >= 20$

Model $Y = \beta_0 + \beta_1 X + \varepsilon$

Desired quantities: β_1 and $(20 – 3)\ \beta_1$