1. Economic growth and Democratic presidential vote

(a) $\hat{\beta}_1 = \frac{4.149}{385.256} = 0.0108$, $\hat{\beta}_0 = \frac{8.288}{17} - 0.0108\frac{32.18}{17} = 0.4671$

(b) $t = \frac{0.0108 - 0}{\sqrt{\frac{0.03653/15}{385.26}}} = 4.29$ $p - value = 2P\left(t_{17} > 4.29\right) < 0.001$

There is very strong evidence that the linear slope relating growth rate to the Democratic vote is not 0.

(c) $\hat{Y} = 0.4671 + 0.0108(5) = 0.521$

$$\hat{Y} \pm t_{15,0.995}\sqrt{\frac{0.03653}{15}\left(\frac{1}{17} + \frac{\left(5 - 32.18/17\right)^2}{385.26}\right)} = 0.521 \pm 2.946(0.0143) = (0.479, 0.563)$$

(d) $$\hat{Y} \pm t_{15,0.995}\sqrt{\frac{0.03653}{15}\left(1 + \frac{1}{17} + \frac{\left(5 - 32.18/17\right)^2}{385.26}\right)} = 0.521 \pm 2.946(0.0514) = (0.37, 0.67)$$

The prediction interval is much wider because it concerns a prediction for one year.

(e)

Source	d.f.	SS	MS	F
Model	1	0.04468	0.04468	18.34
Error	15	0.03653	0.00244	
Total	16	0.0812		

Yes, the F statistic is, at least within round off error, 4.29^2

(f) R^2 = 0.04468 / 0.0812 = 0.55 or 55% Yes, this is a model with R^2 larger than 50%.

(g) $(0.4671 + 0.0108 * 1.9) \pm t_{15,0.975}\sqrt{\frac{0.03653}{15}\left(1 + \frac{1}{17}\right)} = 0.488 \pm 2.13(0.0508) = (0.38, 0.60)$

I don't think this is a very precise prediction. Comparing to my suggested criteria:

The 95% prediction is almost as wide as the range of the data

It is much wider than what is needed to predict close races. The prediction interval is over twice as wide as the desired 0.10 width.

2. Snow gauge calibration

a) X is density, Y is gain. The values of X are fixed by the experimenter's choice of block densities. The values of Y are measured with error. The error variation is associated with gain, so gain is the Y variable.

b) The proposed model was: $Gain = \beta_0 + \beta_1 density + \varepsilon$. My SAS code is:

```
data snow;
  infile 'snow1.txt';
  input density gain;

/* I added a line with density = 0.2 and gain = . to the data file */
proc glm;
  model gain = density;
  estimate 'gain at 0.2' intercept 1 density 0.2;
  output out = resids r = resid p = yhat stdp = stderr stdi = stdobs
         lclm = lline uclm = uline lcl = lpred ucl = upred;

proc print;
  where gain = .;
run;
```

I got:

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	348.4059986	13.40911904	25.98	<.0001
density	-579.9308681	33.49515224	-17.31	<.0001

So, the estimated intercept is 348 and the estimated slope is -579.

- Notice the wording of the question. A regression of gain on density is talking about a regression using Y=gain and X=density. Sometimes folks have misinterpreted the wording and reversed X and Y.

c) Yes. t-value = -17.31 with p-value < 0.0001. There is very strong evidence that the linear slope is not zero.

d) The predicted gain is 232 with a 95% ci of (215, 250) = 232 +/- ($t_{88.0.975}$)*8.72.

SAS gives you the hard part (the s.e.) either from the estimate statement output:

Parameter	Estimate	std err.	t Value	Pr > |t|
gain at 0.2	232.419825	8.72288481	26.64	<.0001

Or from the stdp, lclm, and uclm parts of the output statement:

```
  density       yhat    stderr  stdobs  lline   uline   lpred   upred
2 0.200 . . . . 232.420 8.72288 72.0301 215.085 249.755 89.2752 375.564
```

* When predicting the average, i.e. the position on the line, the appropriate variance is

$$se_{\hat{y}} = \sqrt{MSE\left(1/n + \frac{(x-\bar{x})}{\sum(x-\bar{x})^2}\right)}$$

e) The 95% prediction interval is (89, 375), computed as 232 +/- ($t_{88,\ 0.975}$) * 72.03.

* When predicting individual values at a specific x, the appropriate variance is

$$\sqrt{MSE\left(1+1/n + \frac{(x-\bar{x})}{\sum(x-\bar{x})^2}\right)} = \sqrt{MSE + \left(1/n + \frac{(x-\bar{x})}{\sum(x-\bar{x})^2}\right)} = \sqrt{MSE + (se_{\hat{y}})^2}$$

SAS gives you se(Ynew) and the prediction interval from the stdi, lcl and ucl parts of the output statement.

2. Anscombe data sets.

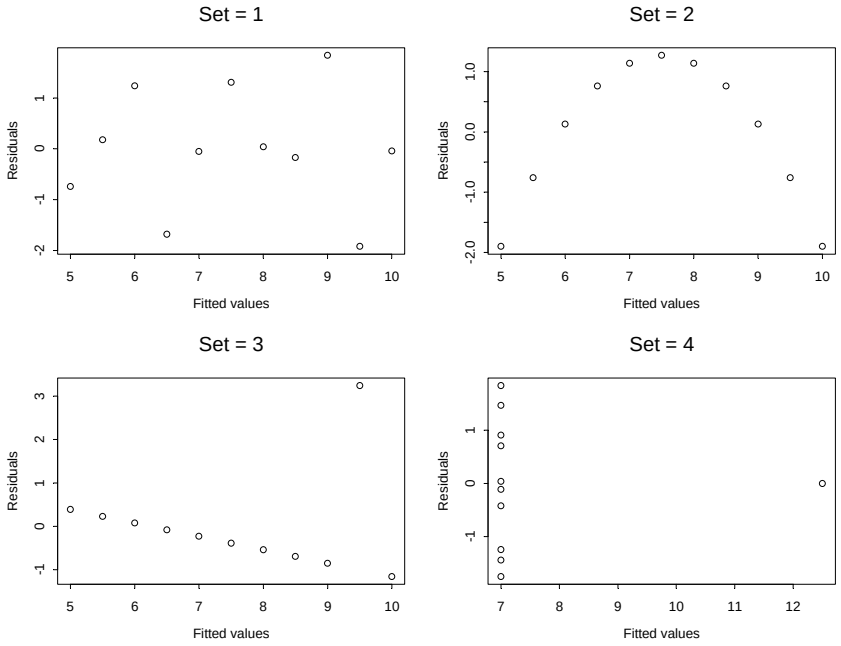
My SAS code:

```
data anscombe;
  infile 'anscombe.txt';
  input set x y;
proc sort; by set;   /* just in case not sorted properly */
proc glm;   /* reg will do exactly same things */
  by set;
  model y = x;
run;
```

(a) All four sets give practically the same numerical values so, if we decide the linear regression is a good description of the relationship between y and x for any particular set, we would have to make the same decision for every other set.

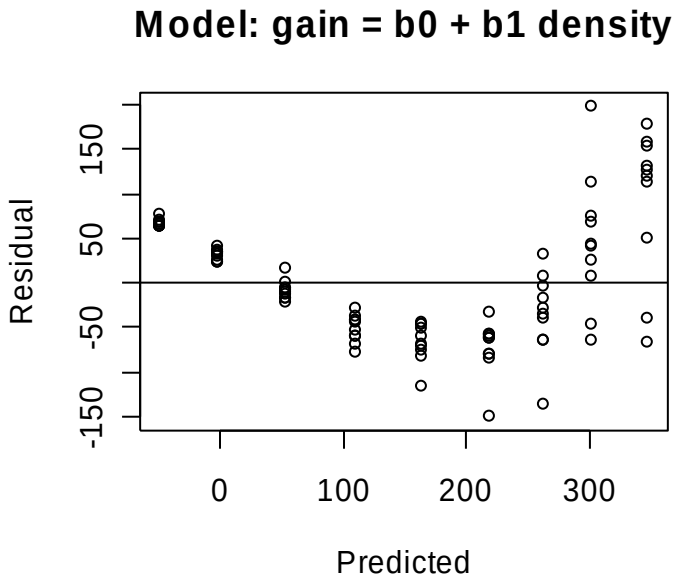
```
SET=1          R-Square     Coeff Var      Root MSE        y Mean
               0.666542      16.48605      1.236603      7.500909

       Parameter         Estimate      Std  Error    t Value    Pr > |t|
       Intercept      3.000090909      1.12474679       2.67      0.0257
       x              0.500090909      0.11790550       4.24      0.0022

SET=2          R-Square     Coeff Var      Root MSE        y Mean
               0.666242      16.49419      1.237214      7.500909

       Parameter         Estimate      Std  Error    t Value    Pr > |t|
       Intercept      3.000909091      1.12530242       2.67      0.0258
       x              0.500000000      0.11796375       4.24      0.0022

SET=3          R-Square     Coeff Var      Root MSE        y Mean
               0.666324      16.48415      1.236311      7.500000

       Parameter         Estimate      Std  Error    t Value    Pr > |t|
       Intercept      3.002454545      1.12448123       2.67      0.0256
       x              0.499727273      0.11787766       4.24      0.0022

SET=4          R-Square     Coeff Var      Root MSE        y Mean
               0.666707      16.47394      1.235695      7.500909

       Parameter         Estimate      Std  Error    t Value    Pr > |t|
       Intercept      3.001727273      1.12392107       2.67      0.0256
       x              0.499909091      0.11781894       4.24      0.0022
```

b) Plots shown below clearly suggest that the linear regression is a good description of the relationship between y and x in `Set=1` and not in the other three sets, since the plot for `Set=1` shows a random scatter of residuals around zero, and the other three plots show specific patterns.



3. Diagnostics for snow gauge problem

a) This residual plot should have you all screaming:
lack of fit: residuals not centered at 0 for all predicted values
unequal variances: vertical spread not the same at all predicted values

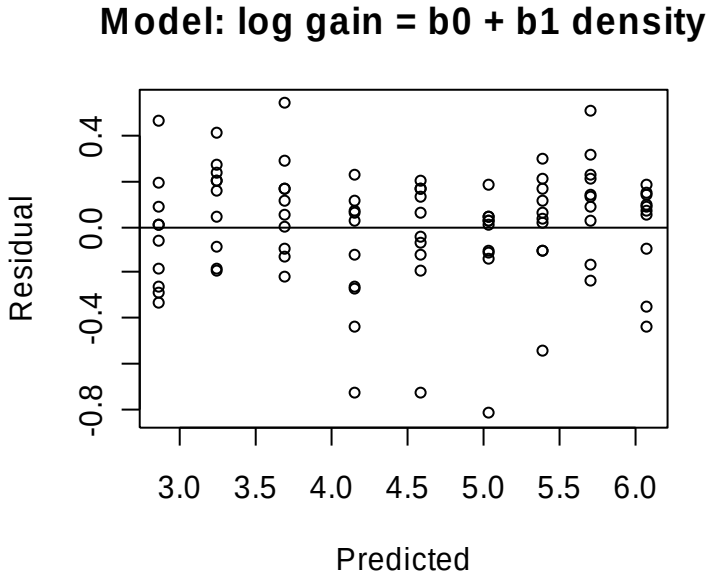


(b) I get a slope of 0.93, which suggests a log transformation or something close to that. I usually round to the nearest 'common' transformation, i.e. log for these data.

- Notice that this transformation is chosen on the basis of the variance-mean relationship. It may (or may not) result in a straightline relationship between f(Y) and X.
- My SAS code (for this part)

```
proc sort;
  by density;
proc means noprint;
  by density;
  var gain;
  output out = means mean = mean stddev = sd;

data means2;
  set means;
  logmean = log(mean);
  logsd = log(sd);
proc plot;
  plot logsd*logmean;

proc glm;
  model logsd = logmean;
```

c) This looks a lot better. Certainly no sign of unequal variances. You might be concerned about a few possible outliers (esp at Yhat = 4.6 and 5.0), but I don't know anything unusual about those points, so I would leave them in the analysis. You might be concerned about the 'wiggle': the residuals seem to go up then down then back up. We'll see about that potential lack of fit in the next part



d) Unless you know a SAS trick, this requires two runs of proc glm: one to estimate error SS for the regression model and the second to estimate error SS for the ANOVA model. Those give:

For the regression:

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	100.0145843	100.0145843	1605.56	<.0001
Error	88	5.4817599	0.0622927		
Corrected Total	89	105.4963442			

For the ANOVA:

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	100.6427554	12.5803444	209.95	<.0001
Error	81	4.8535888	0.0599208		
Corrected Total	89	105.4963442			

The lack of fit SS and d.f. are the difference (either in error quantities or the difference in model quantities) between the two models. Those are SS= 0.6282 with df=7. You can either compute the F statistic directly or put all the values into one ANOVA table to compute the F statistic. Either way, the denominator is the pure error MS = MSE from ANOVA = 0.0599. You get F = (0.6282/7) / 0.0599 = 1.50. You should compare this to quantiles of the F 7,81 distribution. The book tabulates F 7,60. The observed value, 1.5, is between the 0.5 quantile (0.917) and 0.9 quantile (1.82), so the p-value is between 0.1 and 0.5. The actual p-value is 0.17.

There is no evidence of lack of fit.

* Be careful in wording the conclusion. Here, H0 is that the line fits. Rejecting H0 implies that the line does not fit. Remember t-tests from early in the semester; accepting H0 does not imply equal means. Same idea here; accepting H0 does not mean that the line fits or that the response is a straight line.

* There is a trick you can use to get SAS to compute the lack of fit. Basically, create a second copy of the X variable and define it as a class variable in proc glm;. Fit the anova after fitting the regression and the type I SS for the anova will give you lack of fit F test.

```
data snow2;
  set snow;
  lackoffit = density;
proc glm;
  class lackoffit;
  model loggain = density lackoffit;
run;
```

e) This is a calibration problem. The appropriate regression is still Y = log(gain), X = density, so we have to work backwards. The fitted regression is log gain = 6.084 – 4.685 density. So, for measured gain = 152, i.e. log gain = 5.0239, the predicted density is density = (5.0239 – 6.084)/(-4.685) = 0.226.

To get the s.e. of the predicted density, we need the se for predicting an individual observation at x=0.226. This will be slightly larger than sqrt(MSE). You can compute this by hand or rerun the model, adding a new data point with X=0.226 and loggain = . I get se = 0.251, which gives se(density) = 0.251 / abs(-4.685) = 0.054.