ST505/697R: Fall 2012. Homework 4 Solution.

1. 2.21. No, R^2 measures the percent of variation explained by X, but even if R^2 is close to 1 the remaining variation may still be such that prediction Y is difficult. Said, another way it may be that R^2 is close to 1 yet the variance σ^2 (or it's estimate MSE) is big, leading to uncertain predictions.

2. 2.41. No, a yes answer would not lead to a meaningful analysis. When the X's are under the choice of the experimenter, as they are here, the true correlation between X and Y is not a well defined quantity. We also saw that if we change X values what is being estimated on average by r, or R^2 changes and it can be made arbitrarily large or small by maninpulating the X's

3. Show that
$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_i (Y_i - \bar{Y})^2 - b_1^2 \sum_i (X_i - \bar{X})^2.$$
Since $e_i = Y_i - \hat{Y}_i = Y_i - \bar{Y} - b_1(X_i - \bar{X})$,
$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_i (Y_i - \bar{Y})^2 + b_1^2 \sum_i (X_i - \bar{X})^2 - 2b_1 \sum_i (Y_i - \bar{Y})(X_i - \bar{X}).$$
Since $b_1 = \sum_i (Y_i - \bar{Y})(X_i - \bar{X}) / \sum_i (X_i - \bar{X})^2$, then $\sum_i (Y_i - \bar{Y})(X_i - \bar{X}) = b_1 \sum_i (X_i - \bar{X})^2$. So, the last term is $-2b_1 b_1 \sum_i (X_i - \bar{X})^2 = -2b_1^2 \sum_i (X_i - \bar{X})^2$. Hence $SSE = \sum_i (Y_i - \bar{Y})^2 - b_1^2 \sum_i (X_i - \bar{X})^2$.
Note that $SSR = SSTO - SSE = SSTO - [SSTO - b_1^2 \sum_i (X_i - \bar{X})^2] = b_1^2 \sum_i (X_i - \bar{X})^2$
$$= \frac{[\sum_i (X_i - \overline{X})(Y_i - \overline{Y})]^2 \sum_i (X_i - \bar{X})^2}{[\sum_i (X_i - \overline{X})^2]^2} = \frac{[\sum_i (X_i - \overline{X})(Y_i - \overline{Y})]^2}{\sum_i (X_i - \overline{X})^2}.$$
So
$$R^2 = SSR/SSTO = \frac{[\sum_i (X_i - \overline{X})(Y_i - \overline{Y})]^2}{\sum_i (X_i - \overline{X})^2 \sum_i (Y_i - \overline{Y})^2}$$
and the square root of this is the absolute value of r.

4. As done out in class, substituting $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ and $\hat{Y}_i = b_0 + b_1 X_i$ leads to $Y_i - \hat{Y}_i = \epsilon_i + G_i$ where $G_i = \beta_0 + \beta_1 X_i - b_0 - b_1 X_i$. In G_i all but b_0 and b_1 are constants so $E(G_i) = \beta_0 + \beta_1 X_i - E(b_0) - E(b_1) X_i = \beta_0 + \beta_1 X_i - \beta_0 - \beta_1 X_i = 0$.

5. (a) See figure 1
   (b) See top part of figure 2.

From R

```
Shapiro-Wilk normality test
W = 0.9644,  p-value = 0.3806.
```

From SAS

The UNIVARIATE Procedure

Test	--Statistic---		-----p Value------	
Shapiro-Wilk	W	0.964448	Pr < W	0.3806
Kolmogorov-Smirnov	D	0.10541	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.054882	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.381878	Pr > A-Sq	>0.2500

(c) c) i) The linear regression model looks pretty reasonable. The residuals look centered around 0 over the values of ni or predicted (which as noted in this case the two pictures are essentially the same)

ii) There is a bit of an indication that there is an increase in variability as intake increases. This is indicated both by the change in spread in the residuals over ni (or predicted) and by some signs of an increase in the mean absolute residual as a function of intake. This change in variance does not look very serious and when examined through various tests (later in this problem and then in group context as done in class) we never reject the hypothesis of equal variances.

iii) The normality is a bit tricky to assess visually. How straight is straight in a normal probability plot? The shape of the histogram is influenced by how the data is grouped, but the smoothing helps with this. The smoothed version of the histogram (a density estimate) does not look too bad and the Shapiro-Wilk test for normality (and other tests from proc univariate in SAS) are all non-significant. But, as noted these may not be very powerful. All in all, though there is no indication of gross violations of the normality assumption and proceeding as if it were true would be reasonable. This is especially true for inference for coefficients, and functions of them, since these will be pretty robust to the normality assumption for the errors because of the sample size of 31. This is because the inferences for the coefficients depend on the normality of the estimated coefficients, which holds for even moderate n in most cases, regardless of the shape of the distribution of the individual error/observations. The normality assumption is more critical for prediction intervals (and consequently inverse prediction) since the prediction intervals depend on the normality of the individual observations.

e) The regression of $log(e_i^2)$ on X_i (ni) is given below with a plot in bottom part of Figure 2. An approximate test of constant variance of the errors in the original model is given by the test of equal slope in the regression of $log(e_i^2)$ on X_i. This assumes the model is reasonable, which is somewhat plausible from the plot, but may not be a great model. The p-value is .1629 so not enough evidence to conclude unequal variances.

```
lm(formula = loge2 ~ x)
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.74056    1.26113  -0.587    0.562
x            0.02965    0.02071   1.432    0.163
```

```
                    Dependent Variable: logr2
                    Parameter       Standard
 Variable     DF       Estimate          Error    t Value    Pr > |t|
 Intercept     1       -0.74056        1.26113      -0.59      0.5616
 ni            1        0.02965        0.02071       1.43      0.1629
```

e)This proceeds as in d) but using the regression of $log(|e_i|)$ on $log(|\hat{Y}_i|)$. Linearity doesn't look too bad; in fact, looks a little better here than in d). The test for unequal variance ($H_0 : \theta_2 = 0 =$ variance is constant in i) has a p-value of 0.1678, leading to not rejecting equal variance.

```
lm(formula = logabresid ~ logfit)

            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.8389     0.3268   2.567   0.0157 *
logfit       -0.2139     0.1512  -1.415   0.1678
```

```
                    Dependent Variable: labsr
                     Parameter       Standard
 Variable     DF       Estimate          Error    t Value    Pr > |t|
 Intercept     1        0.83891        0.32683       2.57      0.0157
 lpred         1       -0.21394        0.15121      -1.41      0.1678
```

Both approaches in parts d) and e) lead to the same conclusion; namely there is not enough evidence to reject the assumption of constant variance. This agrees with residual plots which show that whatever change in variance may be there would appear to be minor.

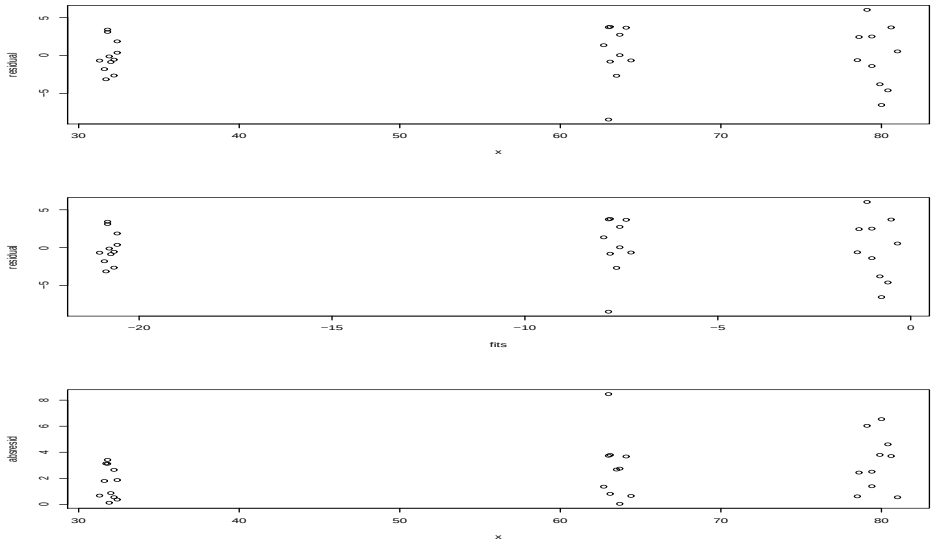

Figure 1: Plots for problem 5, part a)

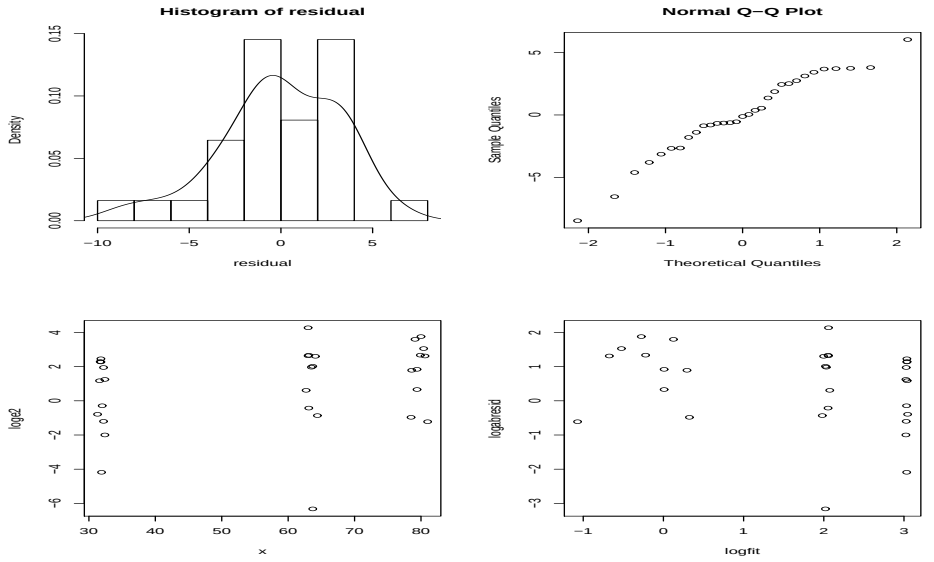

Figure 2: Top panels for problem 5, part b; bottom for d) and e)