1. **1a; problem 1.22** Below is the summary information (edited) from the regression (using R output); code at end of solution as is code and output for SAS.

   a) The estimated regression function is $E(Y) = 168.60000 + 2.03438 * X$, where Y is hardness and X is time. The plot of data and fitted line is below. There are two ways to interpret the question of whether the linear regression supplies a good fit. One is whether a straight line does a good job of modeling the expected value. The answer to that seems to be yes from graphical inspection. A second way, and different way, to view the question is how "tight" the fit is around the line. There is clearly variability in hardness at a fixed time. The question of how this variability relates to how good the fit is will depend on how the fit will be used.

   b) This is the fitted value at $X = 40$ is $b_0 + b_1 40$, which equals 249.975. You can get this directly or use the predicted value in the output for a case with $X = 40$.

   c) This is just β_1 which is estimated by 2.03438

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 168.60000    2.65702   63.45  < 2e-16 ***
time          2.03437    0.09039   22.51 2.16e-12 ***
---
Residual standard error: 3.234 on 14 degrees of freedom
```

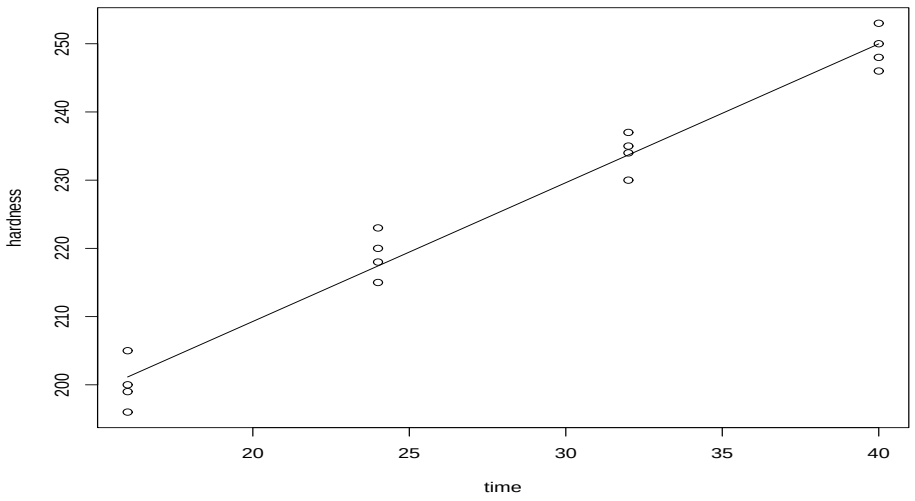

Figure 1: Plot of data and fitted line.

1b; problem 1.26 a.) In SAS you can get the residuals using the p option in proc reg; in R using the residuals function. Easy to see that they add to 0. **Whenever we fit a regression model with an overall intercept the residuals will add to 0. This is not the case if we fit with no intercept.**

```
     hardness time  fits resids
1       199   16 201.150 -2.150
2       205   16 201.150  3.850
```

```
3       196   16 201.150 -5.150
4       200   16 201.150 -1.150
5       218   24 217.425  0.575
6       220   24 217.425  2.575
7       215   24 217.425 -2.425
8       223   24 217.425  5.575
9       237   32 233.700  3.300
10      234   32 233.700  0.300
11      235   32 233.700  1.300
12      230   32 233.700 -3.700
13      250   40 249.975  0.025
14      248   40 249.975 -1.975
15      253   40 249.975  3.025
16      246   40 249.975 -3.975
```

b) The estimate of σ^2 is $\hat{\sigma}^2 = MSE = 10.45893$ (10.5 in the R output).

The estimate of σ is the square root of this so $\hat{\sigma} = 3.23403$ (also root MSE on SAS output, Residual standard error in R output)).

1c) The (estimated) standard error for b_0 is 2.657 and for b_1 is .09039. The CI for b_0 is computed using $168.6 \pm t(.975, 14)2.657$ and for b_1 using $2.03435 \pm t(.975, 14).09039$, where $t(.975, 14) = 2.145$. Using the confints in R or clb in SAS yields 95% confidence intervals of

(Intercept),β_0: [162.9013, 174.29875]
time, β_1: [1.8405, 2.22825]

1d, problems 2.7 a and b:

a) Asking for a 99% confidence interval for β_1, which is (1.7653, 2.30346). The interval can also be computed directly using $b_1 \pm t(.995, 14)s\{b_1\}$ of obtained in either R (via confint with level = .99) or SAS (using clb with alpha = .01).

Using R

```
> confint(regout,level=.99)     #99% confidence intervals
                 0.5 %     99.5 %
(Intercept) 160.690457 176.509543
time          1.765287   2.303463
```

b) Testing $H_0 : \beta_1 = 2$ versus $H_A : \beta_1 \neq 2$.

The t-statistic is $t^* = (2.03438 - 2)/0.09039 = .38035$. With $t(.975, 14) = 2.145$, you do not reject H_0 since $.38 < 2.145$. The P-value is the sum of the area to the right of .38 and the left of -.38 under the t distribution with 14 degrees of freedom. This actually equals .7094. Just using the t-tables, you can see that the area to the right of .38 is somewhere between .3 and .4, so from the tables you know the p-value is between .6 and .7.

Note that we also accept accept H_0 since the P-value is $> .01$.

- You can also test this directly using the 99% CI for β_1 and reject H_0 if 2 is not in the interval. Since 2 is in the interval we do not reject H_0. This is equivalent to doing the t-test.

1d: problem 2.16 You can get a) and b) in SAS directly from the output using the clm and cli option if you include a new case in the data with Y missing (.) and $X = 30$. In R and for the other parts, the intervals can be computed in various ways using either SAS or R as a calculator (as demonstrated for the Kishi example). See code and output at end of solution that corresponds to this.

a) [227.4569, 231.8056] $= 229.6313 \pm (2.264)0.8285$, where $s\{\hat{\mu}(30)\} = 0.8285$ is the standard error associated with the estimated mean and $t(.99, 14) = 2.624$.

Note that $s^2\{\hat{\mu}(30)\} = 0.8285 = 7.06 + 30^2(.0082) + 2 * 30 * (-.2288)$, where $s^2\{b_0\} = 7.06$, $s^2\{b_1\} = .0082$ and $s\{b_0, b_1\} = -.2288$ from the variance-covariance matrix of the coefficients

```
>  vcov(regout)                  #this gives the variance-covariance
            (Intercept)         time
(Intercept)   7.0597768 -0.228789063
time         -0.2287891  0.008171038
```

b) $(220.8695, 238.3930) = 229.6313 \pm (2.624)(10.45893 + 0.8285^2)^{1/2}$.

c) $229.6313 \pm (2.264)((10.45893/10) + 0.8285^2)^{1/2} = [226.2, 233.1]$.

d) The interval in c) is smaller since you are trying to predict the mean of 10 values which has less variability ($\sigma^2/10$) than one value.

e) $229.6313 \pm (2 * 5.24)^{1/2} * 0.8285 = [226.95, 232.32]$, where $F(.99, 2, 14) = 5.24$ (obtained exactly using SAS or R you could approximate using the entries for $F(.975, 2, 12)$, $F(.975, 2, 15)$, $F(.99, 2, 12)$ and $F(.99, 2, 15)$ from the F table.)

1f.

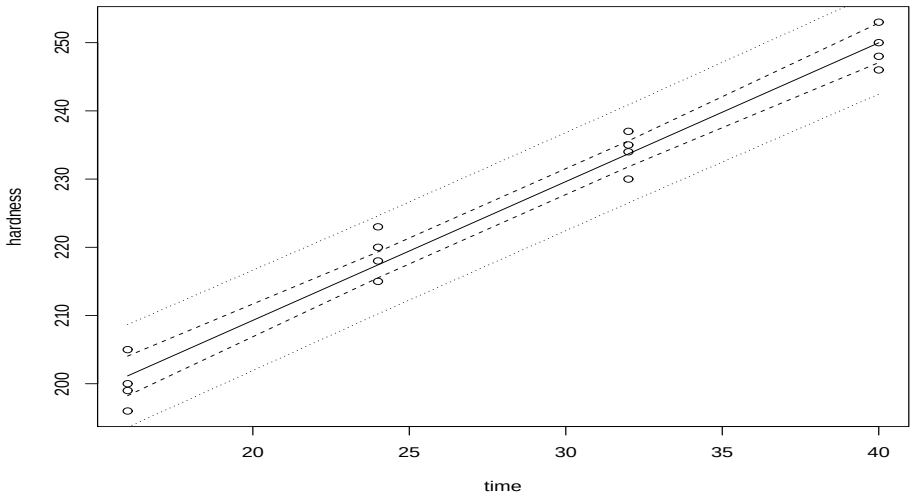

Figure 2: Individual CIs for means and Prediction intervals

1g.

1g: Problem 4.9

a) and b) $X_j = 20, 30$ and 40.

For Bonferroni, use $b_0 + b_1 X_j \pm t(1-.10/6, 14) s\{\hat{\mu}\{X_j\}\}$, with $t(1-.10/6, 14) = 2.360$ The Working-Hotelling/Scheffe intervals use the same form but with the t value replaced by $(2F(.90, 2, 14))^{1/2} = 2.3352$. The Working-Hotelling intervals will be smaller and more efficient since $2.335 < 2.360$, but the difference is minor.

```
     Scheffe intervals      Bonferroni intervals  estimate    SE
X=20  206.755    211.821   206.728    211.847     209.288    1.08473
X=30  227.697    231.566   227.676    231.586     229.631    0.82847
X=40  246.816    253.134   246.783    253.168     249.975    1.35289
```

c). Now doing prediction intervals for two future values at $X_1 = 30$ and $X_2 = 40$. Use $b_0 + b_1 X_j \pm t(1 - .10/4, 14)s\{pred_j\}$, with $t(1 - .10/4, 14) = 2.145$ and $s^2\{pred_j\} = MSE + s^2\{\hat{\mu}(X_j)\}$. The Scheffe intervals use the same form but with the t value replaces by $(2F(.90, 2, 14))^{1/2} = 2.3352$. The Bonferroni intervals should be used here since shorter.

	Bonferroni	Scheffe	SEpredj
X=30	222.471, 236.792	221.836, 237.427	3.33846
X=40	242.456, 257.494	241.789, 258.161	3.50560

2. **Problem 2**. a) The estimate of β_0 is .07471, of β_1 is 2.10983, of σ^2 is .06307, and of σ is .2511 = $(.06307)^{1/2}$.

   b) $2.10983 \pm 1.96(.01194) = [2.08643, 2.13323]$

   c) $H_0 : \beta_1 = 0$. $t^* = 2.10983/.01194 = 176.67$. The p-value for this is the area to the right of 176.67 plus the area to the left of - 176.67 under the standard normal (use this since the degrees of freedom is 3164). Since the area to the right of 3.291 is equal to .0005 (see table) we know the P-value is less than 2*.0005 = .0001.

   As a probability, the P-value is the probability of getting a value of the $|t^*|$ greater than or equal to the value of the observed absolute value, *under the null hypothesis*. Here, it is the probability that $|t^*| > 176.67$ where t^* is distributed t with 3164 degrees of freedom, which is for practical purposes the standard normal distribution.

   d) $\hat{Y}_h = 3.5917966$, $s\{\hat{Y}_h\} = 0.0066024 = (s^2\{b_0\} + X_h^2 s^2\{b_1\} + 2X_h s\{b_0, b_1\})^{1/2}$, $s\{pred\} = 0.2512242 = (MSE + s^1 2\{\hat{Y}_h\})^{1/2}$.

   Use $z = 1.96$ in getting the intervals.

   Confidence interval for E(Y) at X = 1.667 is $[3.5788559, 3.6047373]$

   Prediction interval for Y at X = 1.667, is $[3.0993972, 4.084196]$.

3. **Problem 3.** Below are results for designs 1 and 2. Here I've given parts of the SAS output. Similar results apply for R. The first part of each program gives you the true standard errors of the estimated coefficients. The expected values of b_0, b_1 and MSE are known to be exactly β_0 and β_1 and σ^2 (0, 1 and .0225) respectively.

   To choose between designs we can do that in terms of the true standard errors of the estimated coefficients go. The second design is better because the standard errors are smaller. We know the estimators are unbiased for both unbiased so we can just compare variance or standard errors to make that decision.

   - The second part of the output gives summary statistics over the thousands of simulated values. The fact that the means of the three variables are not exactly the true parameters is because we are running a limited number of simulations. The histograms represent the sampling distribution of these estimators (this is not the exact sampling distribution because of a limited number of simulations but it will be close).

   - If you thought of which design is better for estimating σ^2 you'd have to rely on the simulation results since I didn't tell you the variance of MSE. From the simulation results it looks like design 2 is a little better. **In fact the two designs are equivalent from this perspective**. It can be shown that the standard deviation of MSE is $(2\sigma^4/(n-2))^{1/2}$ (= .01006 here), which only depends on the design through n. All designs with the same n are equally good for estimating σ^2. This may surprise you.

```
                Homework 2, number 3 Design 1
coefficients:  beta0 =          0  beta1 =          1
          sigma2    0.0225  sigma =       0.15
            number of simulations       1000
          n =         12 xvalues =       7.67
                                         6.31
                                         6.14
                                         7.07
```

```
                                                        6.39
                                                        5.95
                                                        6.53
                                                        6.55
                                                        5.34
                                                        5.74
                                                        4.94
                                                        7.07
                              using normal errors
                    THEORETICAL VARIANCES AND STANDARD DEVIATION
                   variance of b0 =  0.1413792 sd of b0 =  0.3760042
                   variance of b1  =  0.0035056 sd of b1 =  0.0592078
                              The MEANS Procedure
Variable       N            Mean         Std Dev         Minimum         Maximum
------------------------------------------------------------------------------
B0          1000       0.0228333       0.3734848      -1.2547479       1.3339013
B1          1000       0.9966975       0.0588648       0.7961849       1.1947983
MSE         1000       0.0226470       0.0102871       0.0033132       0.0705808
------------------------------------------------------------------------------
```

And here are results from design 2.

```
                     Homework 2, number 3 Design 2                          1
                coefficients:  beta0 =          0  beta1 =          1
                         sigma2    0.0225  sigma =       0.15
                           number of simulations       1000
                         n =         12 xvalues =        5.3
                                                         5.3
                                                         5.3
                                                         5.3
                                                         5.3
                                                           6
                                                           6
                                                           7
                                                           7
                                                           7
                                                           7
                                                           7
                              using normal errors
                    THEORETICAL VARIANCES AND STANDARD DEVIATION
                   variance of b0 =  0.1181024 sd of b0 =  0.3436603
                   variance of b1  =  0.0030981 sd of b1 =  0.0556606
                              The MEANS Procedure
Variable       N            Mean         Std Dev         Minimum         Maximum
------------------------------------------------------------------------------
B0          1000       0.0081894       0.3404172      -1.0216963       1.1265641
B1          1000       0.9987300       0.0549423       0.8055471       1.1717619
MSE         1000       0.0225139       0.0098290       0.0037846       0.0690602
------------------------------------------------------------------------------
```