STA 113: Probability and Statistics in Engineering
Homework # 9: Solutions and Quiz Answers

0.1 Problem 1

The following is the input and output from my minitab session in which I read in the data:

```
MTB > read into c1
DATA> 3320
DATA> 2655
DATA> 3855
DATA> 3195
DATA> end.
      4 rows read.

MTB > read into c2
DATA> 20
DATA> 26
DATA> 18
DATA> 18
DATA> end.
      4 rows read.
```

a.) I used minitab as follows:

```
MTB > plot y in c2 vs x in c1

         -     *
         -
     25.0+
         -
 C2      -
         -
         -
     22.5+
         -
         -
         -
         -
     20.0+                           *
         -
         -
         -
         -                      *                         *
          --------+---------+---------+---------+---------+--------C1
                2750      3000      3250      3500      3750
```

b.) b0 and b1 can be obtained from the regression equation or from the Coef column in the printout below:

```
MTB > regress c2 on 1 predictors c1

The regression equation is
C2 = 41.4 - 0.00643 C1

Predictor       Coef       Stdev    t-ratio        p
Constant      41.449       9.767       4.24    0.051
C1         -0.006434    0.002974      -2.16    0.163

s = 2.537       R-sq = 70.1%     R-sq(adj) = 55.1%

Analysis of Variance

SOURCE       DF          SS          MS         F        p
Regression    1      30.125      30.125      4.68    0.163
Error         2      12.875       6.437
Total         3      43.000
```

c.) I did this in minitab by creating a vector of x values spanning the range of weights and then plugging each x value into the regression equation. Then I plotted both the data and regression line (as a series of points). There may be a much easier way to do this in minitab but I don't know of it. Hand-drawn is fine as well.

```
MTB > read into c4
DATA> 0
DATA> 250
DATA> 500
and so forth...
DATA> 4250
DATA> 4500
DATA> 4750
DATA> end.
     20 rows read.
MTB > let c5 = 41.449 + c4*-.006434
```

```
MTB > mplot c2 vs c1, c5 vs c4

         - B
       40+    B
         -      B
         -         B B
         -            B
         -              B
       30+                 B
         -                   B
         -                      B B A
         -                           B
         -                             B
       20+                                2
         -                               A  B  BA
         -                                       B
         -                                          B
         -                                            B
       10+                                               B
           +---------+---------+---------+---------+---------+------
           0      1000      2000      3000      4000      5000
        A = C2 vs. C1             B = C5 vs. C4
```

The B's are the regression line and the A's are the data points. The 2 indicates overlap between the regression line and one of the data points.

d.) and e.)

```
MTB > read into c3
DATA> 3000
DATA> 5000
DATA> end.
      2 rows read.

MTB > regress c2 on 1 predictors c1;
SUBC> predict for c3.

The regression equation is
C2 = 41.4 - 0.00643 C1

Predictor        Coef       Stdev    t-ratio        p
Constant       41.449       9.767       4.24    0.051
C1          -0.006434    0.002974      -2.16    0.163

s = 2.537       R-sq = 70.1%     R-sq(adj) = 55.1%
```

```
Analysis of Variance

SOURCE       DF          SS          MS         F        p
Regression    1      30.125      30.125      4.68    0.163
Error         2      12.875       6.437
Total         3      43.000

     Fit  Stdev.Fit         95% C.I.             95% P.I.
   22.15       1.48   (   15.78,   28.52)  (    9.51,   34.79)
    9.28       5.34   (  -13.69,   32.25)  (  -16.15,   34.71) XX
```

The predicted mileage for the 3000 lb car is 22.15 under the Fit column. The predicted mileage for the 5000 lb car is 9.28.

0.2 Problem 2

This problem is basically the same as problem 1, so I've just put in the minitab output:
a.)

```
MTB > read into c7
 DATA> 5
 DATA> 3.846
 DATA> 5.556
 DATA> 5.556
 DATA> end.

MTB > plot c7 vs c1

          -
          -                            *                          *
      5.40+
          -
  C7      -
          -                                 *
          -
      4.80+
          -
          -
          -
          -
      4.20+
          -
          -
          -     *
          -
            --------+---------+---------+---------+---------+--------C1
                 2750      3000      3250      3500      3750
```

b.)

```
MTB > regress c7 on 1 predictors c1

 The regression equation is
 C7 = 0.55 + 0.00136 C1

 Predictor       Coef       Stdev    t-ratio        p
 Constant       0.548       2.101       0.26    0.819
 C1         0.0013639   0.0006397       2.13    0.167

 s = 0.5457      R-sq = 69.5%     R-sq(adj) = 54.2%

 Analysis of Variance

 SOURCE       DF          SS          MS         F        p
 Regression    1      1.3540      1.3540      4.55    0.167
 Error         2      0.5956      0.2978
 Total         3      1.9495
```

c.)

```
MTB > mplot c7 vs c1, c9 vs c4

   -
7.5+
   -                                                 B
   -                                            B B
   -                                         B
   -                                 A  B  BA
5.0+                                  2
   -                             B B
   -                          B A
   -                     B  B
   -                   B
2.5+              B B
   -           B
   -         B
   -    B B
   - B
0.0+
     +---------+---------+---------+---------+---------+------
     0      1000      2000      3000      4000      5000
  A = C7 vs. C1             B = C9 vs. C4
```

d.) and e.) The predicted value for the 3000 lb car is 4.64 gal per 100 miles with 7.368 the predicted gal per 100 mil for the 5000 car. Remember to calculate mileage from the fuel values.

```
      Fit  Stdev.Fit         95% C.I.            95% P.I.
    4.640      0.318   (   3.270,   6.010)  (   1.922,   7.358)
    7.368      1.148   (   2.427,  12.309)  (   1.898,  12.838) XX

MTB > let c10=100/4.64
 MTB > print(c10)

 C10
   21.5517

 MTB > let c10=100/7.368
 MTB > print(c10)

 C10
   13.5722
```

0.3 Problem 3

Predictions from Model 1 are: $\begin{array}{ll} 3000: & 22.15 \\ 5000: & 9.28 \end{array}$

Predictions from Model 2 are: $\begin{array}{ll} 3000: & 21.55 \\ 5000: & 13.57 \end{array}$

Robert made an argument in class in favor of Model 2 because we might expect the amount of fuel used to be linearly related to car weight based on physical principles (F=ma) whereas we have no reason to think that mileage is linearly related to car weight. With only 4 data points it's tough to check whether the relationship is linear in the two cases. One other possibility is to compare the R^2 values but these are almost identical. Comparing the SSE or s values from the models won't work because the transformation between fuel and mileage means that these values are not directly comparable.

0.4 Problem 4

The estimates of b0 and b1 minimize the sum of squared deviations between the data points and the regression line. The deviations are the vertical distances (parallel to the y axis) between each data point and the regression line. Note that minimizing the sum of squared deviations also maximizes the likelihood function with respect to b0, b1 and σ^2.

0.5 Problem 5

Minitab appears to be able to give a 95% prediction interval, but not a 90% prediction interval. One could use a more sophisticated statistical package or do it by hand. For a 90% interval, $\alpha = .10$:

$$\hat{y}_p \pm t_{\alpha/2} s \sqrt{1 + \frac{1}{n} + \frac{(x_p - \overline{x})^2}{SS_{xx}}}$$

where t is based on n-2 df = 2 df (n=4), $t_{.05,2} = 2.92$, $s = \sqrt{\frac{SSE}{n-2}} = 0.5457$, $x_p = 3000$ (or 5000), $\overline{x} = 3256.25$, and $SS_{XX} = 727818.8$.

For the 3000 lb car, the interval is 4.64 ± 1.84 in terms of gal consumed per 100 miles. Converting to mileage, we get $(15.43, 35.71)$. Note that you need to get upper and lower bounds in terms of fuel used and then convert those bounds to mileage. Don't divide 100 by 1.84 as the transformation is not linear.

For the 5000 lb car, the interval is 7.37 ± 3.71 in terms of gal consumed per 100 miles. Converting to mileage, we get $(9.03, 27.32)$.

The interval for the heavier car in terms of fuel used is a wider interval because the uncertainty in the mean fuel used by a 5000 lb car is larger than the uncertainty in the mean fuel used by a 3000 lb car. This occurs because 5000 is much farther from the average weight of the 4 cars than is 3000. Changes in the slope have a minimal effect in the vicinity of $\overline{x}$ but a large effect far from $\overline{x}$. The full interval is partly a result of this parameter uncertainty, which varies with x, causing uncertainty in $E[y \mid x]$ (mean value of y for a given x) and partly due to variability in the individual y's about $E[y \mid x]$. This variability of the individual y's about the mean y is constant with respect to x, and is estimated by s^2. The intervals for the mileage are affected by the transformation done in converting fuel used to mileage, which is why the interval for the lighter car is larger in terms of mileage.

Quiz Answers

```
 The regression equation is
 I = - 0.07 + 1.19 V

 Predictor       Coef       Stdev    t-ratio        p
 Constant     -0.0698      0.1892    -0.3687   0.7219
 V             1.1898      0.0544    21.8852   0.0000

R-squared: 98.7%
```

a) Give a 95% CI for your resistance R.

Since $R = 1/\beta_1$, then getting a 95% C.I. for R would be the same thing as getting a 95% C.I. for β_1. Since β_1 follows a student-t distribution centered at MLE (in this case $\hat{\beta}_{MLE} = 1.1898$) with $n-2$ d.f., **standard deviation given by the computer output (0.0544)**, the 95% C.I. would be

$\beta_1 = \hat{\beta}_1 \pm t_{10-2,.025}\, std = 1.1898 \pm 2.306 * 0.0544 = (1.065, 1.315)$

Then the CI for R is simply **the reciprocal of this interval**, from 1/1.315 to 1/1.065, which would be

$(0.76, 0.94)$

b) from this data, would you say that $\beta_0 \neq 0$? Explain.

Tesing hypotheses (two-tailed)

$H_0 : \beta_0 = 0$

$H_a : \beta_0 \neq 0$

Note: The p-values at the end of each line in the computer output is the p-value of the hypotheses if the corresponding coefficient equals 0.

For this data, the p-value of β_0 being 0 is 0.7219. Given thus a big p-value, we **don't** reject β_0 being 0.

c) What is the residual of $V = 2.10, I = 2.62$?

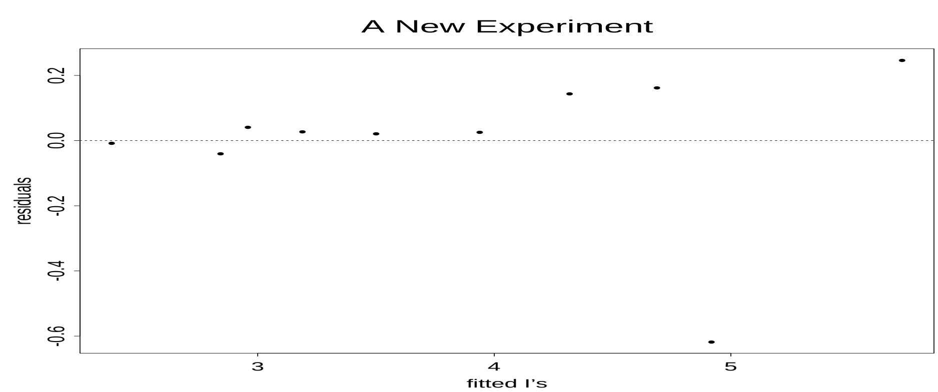
The predicted value (or fitted value) at $V = 2.10$ is

$\hat{I} = -0.07 + 1.19 * 2.10 = 2.43$

Residual$= I - \hat{I} = 2.62 - 2.429 = 0.19$

d) What is the problem with this residual plot?

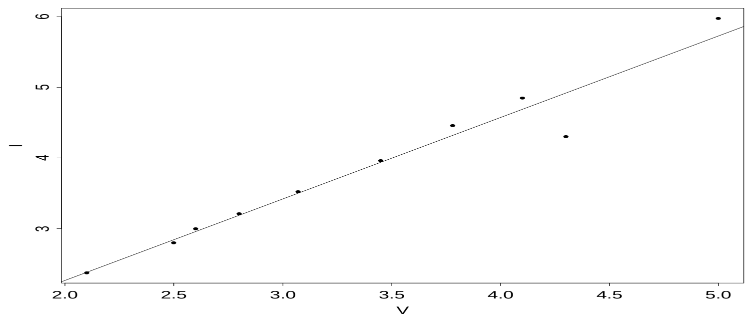
There seem to be an outlier, far down at -0.6. The rest of the residuals form a positive slope. So actually this outlier drags down the actual slope (from $beta_1 = 1.25$ to$beta_1 = 1.15$. In a real experiment, the true value of R is probably 0.80 ohm, which gives our data a 1.25 slope. But because of one bad measurement with our 9th data point, we have to give a biased conclusion of $R = 1/1.15 = 0.87 ohm$. In order to avoid this kind of bias, we throw out the 9th data point and see how the rest of the data goes. (Note: in reality, the case is more complicated, it's not obvious that if a datum is an outlier or not.)



Usually we exclude this data point and redo the regression.

]

Before:



]

After removing the outlier: we can see that te regression has improved