Math 3080 § 1. **Second Midterm Exam** Name: Solutions
Treibergs $\mathscr{A}\mathscr{T}$ March 19, 2014

(1.) *The article "Withdrawl Strength of Threaded Nails," in Journal of Structural Engineering, 2001, describes an experiment to relate the diameter of a threaded nail (x in mm) to its ultimate withdrawal strength from Douglas Fir lumber (y in N/mm). Obtain the equation of the least-squares line. What proportion of observed variation in withdrawal strength can be attributed to the linear relationship between withdrawal strength and diameter? Values of summary quantities are*

```
> c( length(x),  sum(x),   sum(y), sum(x^2),  sum(x*y),   sum(y^2) )
[1]         10     37.3     659.9     145.6     2533.0     44906.0
```

We compute intermediate quantities

$$\begin{aligned}
\overline{x} &= \frac{1}{n}\sum x_i = \frac{37.3}{10} = 3.73, \qquad \overline{y} = \frac{1}{n}\sum y_i = \frac{659.9}{10} = 65.99, \\
S_{xx} &= \sum x_i^2 - \frac{1}{10}\left(\sum x_i\right)^2 = 145.6 - \frac{1}{10}\left(145.6\right)^2 = 6.471; \\
S_{xy} &= \sum x_i y_i - \frac{1}{10}\left(\sum x_i\right)\left(\sum y_i\right) = 2533 - \frac{1}{10}\left(37.3\right)\left(659.9\right) = 71.573; \\
S_{yy} &= \sum x_i^2 - \frac{1}{10}\left(\sum x_i\right)^2 = 44906 - \frac{1}{10}\left(659.9\right)^2 = 1359.199
\end{aligned}$$

Thus the least-squares line is $\boxed{y = \widehat{\beta_0} + \widehat{\beta_1} x}$ where

$$\widehat{\beta_1} = \frac{Sxy}{S_{xx}} = \frac{71.573}{6.471} = \boxed{11.06058} \qquad \widehat{\beta_0} = \overline{y} - \widehat{\beta_1}\overline{x} = 65.99 - (11.06058)(3.73) = \boxed{24.73404}$$

The proportion on of observed variation in withdrawal strength attributed to the linear relationship between withdrawal strength and diameter is R^2, which may be obtained by

$$R^2 = \frac{S_{xy}^2}{S_{xx}S_{yy}} = \frac{(71.573)^2}{(6.471)(1359.199)} = \boxed{0.5824303}$$

(2.) *Given n data points (x_i, y_i), suppose that we try to fit a linear equation $y = \beta_0 + \beta_1 x$. For the simple regression model, what are the assumptions on the data $\{(x_i, y_i)\}$? Find the variance $V(Y_i)$ for the y-coordinate of the ith observation. Explain. What is the fitted value $\hat{Y}_i$? Show that it may be expressed as a linear combination $\sum_{j=1}^{n} c_j Y_j$, where the c_j's don't depend on the Y_i's. Starting from the regression model, derive the expected value $E\left(\sum_{i=1}^{n} Y_i^2\right)$.*

For the simple regression model, we assume that x_i's are fixed and that

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad \text{where } \epsilon_i \sim \mathrm{N}(0, \sigma^2) \text{ are I.I.D.}$$

Hence, the y-coordinate of random observation at x_i has

$$V(Y_i) = V(\beta_0 + \beta_1 x_i + \epsilon_i) = V(\epsilon_i) = \sigma^2$$

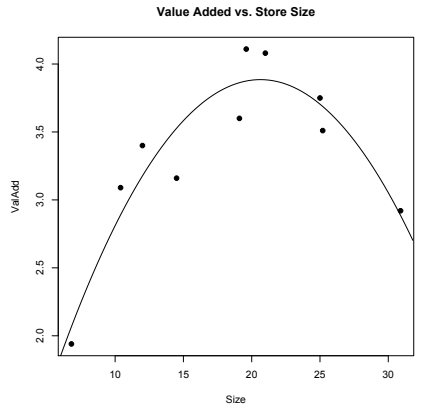
because Y_i and ϵ_i differ by a constant so have the same variance. The fitted value is the point predicted by the regression line, namely

$$\begin{aligned}\widehat{Y_i} &= \widehat{\beta_0} + \widehat{\beta_1} x_i = \overline{Y} - \widehat{\beta_1}\overline{x} + \widehat{\beta_1} x_i = \overline{Y} + \widehat{\beta_1}(x_i - \overline{x}) \\ &= \sum_{j=1}^{n} \frac{Y_j}{n} + (x_i - \overline{x})\frac{S_{xy}}{S_{xx}} = \sum_{j=1}^{n} \frac{Y_j}{n} + \frac{(x_i - \overline{x})}{S_{xx}} \sum_{j=1}^{n} (x_j - \overline{x})(Y_j - \overline{Y}) \\ &= \sum_{j=1}^{n} \left\{ \frac{1}{n} + \frac{(x_i - \overline{x})(x_j - \overline{x})}{S_{xx}} \right\} Y_j\end{aligned}$$

because $\sum_{j=1}^{n}(x_j - \overline{x})\overline{Y} = 0$. Now, use the fact that for any random variable, $E(Z^2) = V(Z) + E(Z)^2$ so that

$$\begin{aligned}E\left(\sum_{i=1}^{n} Y_i^2\right) &= \sum_{i=1}^{n} E\left(Y_i^2\right) = \sum_{i=1}^{n} \left\{ V\left(Y_i\right) + E\left(Y_i\right)^2 \right\} = \sum_{i=1}^{n} \left\{ \sigma^2 + \left(\beta_0 + \beta_1 x_i\right)^2 \right\} \\ &= n\sigma^2 + \sum_{i=1}^{n} \left(\beta_0 + \beta_1 x_i\right)^2 .\end{aligned}$$

(3.) *A 1984 study in Journal of Retailing looked at variables that contribute to productivity in the retail grocery trade. A measure of productivity is "value added per work hour," (in \$) which is the surplus money generated by the business available to pay for labor, furniture and fixtures, and equipment. How does value added depend on store size (in* 1000 ft^2*)? State the model and the assumptions on the data. Is there strong evidence that* $\beta_2 < -.006$*? State the null and alternative hypotheses. State the test statistic and rejection region. Carry out the test at significance level .05.*




```
> ValAdd = c( 4.08, 3.40, 3.51, 3.09, 2.92, 1.94, 4.11, 3.16, 3.75, 3.60 )
> Size   = c( 21.0, 12.0, 25.2, 10.4, 30.9,  6.8, 19.6, 14.5, 25.0, 19.1 )
> m1=lm(ValAdd ~ Size + I(Size^2))
> plot(ValAdd~Size, pch=19, main="Value Added vs. Store Size")
> lines(seq(0,32,.3),predict(m1,data.frame(Size=seq(0,32,.3))))
> summary(m1); anova(m1)


lm(formula = ValAdd ~ Size + I(Size^2))
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept) -0.159356   0.500580  -0.318 0.759512
Size         0.391931   0.058006   6.757 0.000263
I(Size^2)   -0.009495   0.001535  -6.188 0.000451
Residual standard error: 0.2503 on 7 degrees of freedom
Multiple R-squared: 0.8794,Adjusted R-squared: 0.845
F-statistic: 25.53 on 2 and 7 DF,  p-value: 0.0006085

Analysis of Variance Table
Response: ValAdd
          Df  Sum Sq Mean Sq F value    Pr(>F)
Size       1 0.80032 0.80032  12.774 0.0090466
I(Size^2)  1 2.39858 2.39858  38.286 0.0004507
Residuals  7 0.43855 0.06265}
```

The quadratic regression model assumes that sizes x_i are fixed and the value added y_i satisfies

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i, \qquad \text{where } \epsilon_i \text{ are I.I.D. } \mathrm{N}(0, \sigma^2)$$

for some constants β_i. This is a quadratic regression so the number of variables is $k = 2$.

We test the hypotheses

$$\mathcal{H}_0 : \beta_2 = \beta_{20} = -.006 \text{ vs. } \mathcal{H}_a : \beta_2 < \beta_{20} = -.006.$$

The test statistic is

$$t = \frac{\widehat{\beta_2} - \beta_{20}}{s_{\widehat{\beta_2}}}$$

which is a random variable distributed according to the t-distribution with $\nu = n - k - 1$ degrees of freedom. The null hypothesis is rejected if $t \leq -t_{\alpha, n-k-1}$.

When $\alpha = 0.05$ and the number of observations is $n = 10$, by reading the summary table, and looking at Table A5 from the text,

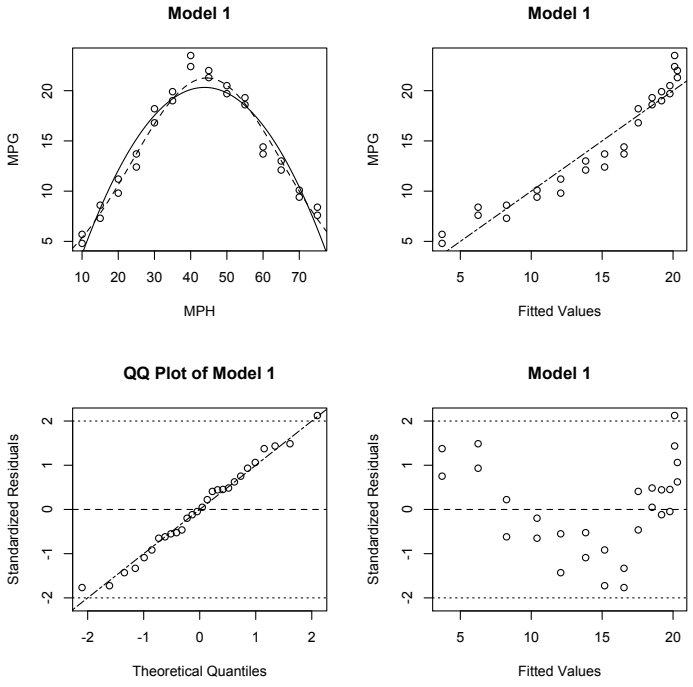
$$t = \frac{\widehat{\beta_2} - \beta_{20}}{s_{\widehat{\beta_2}}} = \frac{-.009495 - (-.006)}{.001535} = -2.276873 < -t_{\alpha, n-k-1} = -t_{.05,7} = 1.895.$$

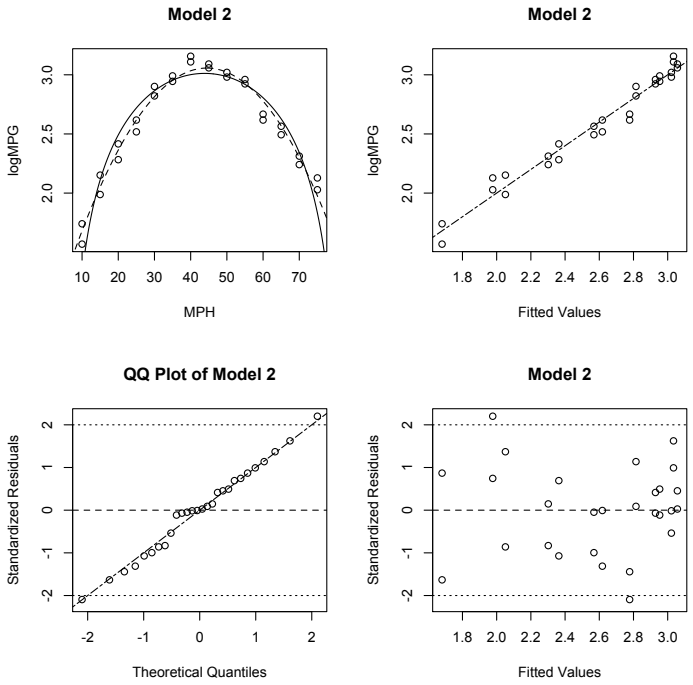
Thus we reject the null hypothesis: there is significant evidence that $\beta_2 < -.006$.

(4.) *To develop a model to predict miles per gallon based on highway speed, a test car was driven during two trial periods at speeds ranging from 10 to 75 miles per hour. Consider two models to explain the data: first, a quadratic regression* $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2$ *(the solid lines), and second a transformed version,* $\log(y_i) = \beta_0 + \beta_1\, x_i + \beta_2\, x_i^2$ *(the dashed lines).* **R**© *was used to generate tables and plots. What is the equation of the dashed line (for Model 2) in the first panel scatterplot in* (MPH, MPG) *coordinates? Discuss the two models with regard to quality of fit and whether the model assumptions are satisfied. Compare at least five features in the tables and plots. Which is the better model?*

```
MPH 10.0 10.0 15.0 15.0 20.0 20.0 25.0 25.0 30.0  30.0  35.0  35.0  40.0  40.0
MPG  4.8  5.7  8.6  7.3  9.8 11.2 13.7 12.4 18.2  16.8  19.9  19.0  22.4  23.5

MPH 45.0 45.0 50.0 50.0 55.0 55.0 60.0 60.0 65.0  65.0  70.0  70.0  75.0  75.0
MPG 21.3 22.0 20.5 19.7 18.6 19.3 14.4 13.7 12.1  13.0  10.1   9.4   8.4   7.6

lm(formula = MPG ~ MPH + I(MPH^2))
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -7.5555495  1.4241091  -5.305 1.69e-05
MPH          1.2716937  0.0757321  16.792 3.99e-15
```

```
I(MPH^2)    -0.0145014  0.0008719 -16.633 4.97e-15

Residual standard error: 1.663 on 25 degrees of freedom
Multiple R-squared: 0.9188,Adjusted R-squared: 0.9123
F-statistic: 141.5 on 2 and 25 DF,  p-value: 2.338e-14

Analysis of Variance Table
Response: MPG
          Df Sum Sq Mean Sq  F value    Pr(>F)
MPH        1  17.37   17.37   6.2775   0.01911
I(MPH^2)   1 765.46  765.46 276.6417 4.973e-15
Residuals 25  69.17    2.77
=========================================================
lm(formula = log(MPG) ~ MPH + I(MPH^2))
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.664e-01  6.768e-02   11.32 2.47e-11
MPH          1.030e-01  3.599e-03   28.62  < 2e-16
I(MPH^2)    -1.158e-03  4.144e-05  -27.96  < 2e-16

Residual standard error: 0.07906 on 25 degrees of freedom
Multiple R-squared: 0.9704,Adjusted R-squared: 0.968
F-statistic: 409.7 on 2 and 25 DF,  p-value: < 2.2e-16

Analysis of Variance Table
Response: log(MPG)
          Df Sum Sq Mean Sq F value    Pr(>F)
MPH        1 0.2362  0.2362  37.792 1.992e-06
I(MPH^2)   1 4.8850  4.8850 781.587 < 2.2e-16
Residuals 25 0.1563  0.0063
```





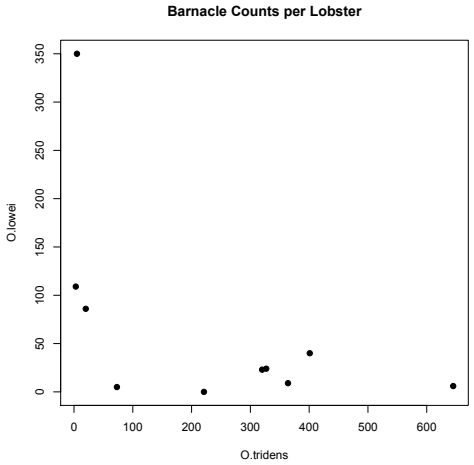
The equation of the dashed line in the first figure is found from the second model.

$$y = \exp\left(\widehat{\beta_0} + \widehat{\beta_0}\, x + \widehat{\beta_0}\, x^2 +\right) = \exp\left(0.774 + 0.1030\, x - .001158\, x^2\right)$$

The second model seems to perform better. First of the two intrinsically linear versions, the coefficient of multiple determination in the first model $R^2 = 0.9188$ is smaller than $R^2 = 0.9704$ for the second model. Thus the second model accounts for a larger percentage of variation relative to the intrinsic model. Because both models have two independent variables, we do not need to compare adjusted R^2 which accounts for different numbers of variables. Adjusted R^2 also shows that the second model accounts for more variability. However, the second model has dimensions $\log y$, and the first y, this comparison is limited. Second, looking at the scatterplots Panels 1 and 5, both show downward arc indicating that the quadratic term is useful. Indeed the quadratic coefficients are significantly nonzero in the tables. The dashed fitted line of the second model seems to capture the points slightly better than the solid line of the first model. Third, comparing the observed values vs. the fitted values, Panels 2 and 6, Model 1 seems to have a lot more upward bowing in the points than Model 2, indicating that the first model still needs some nonlinear adjustment, whereas the second has points lining up beautifully. Fourth, looking at the normal QQ-plots of the standardized residuals, Panels 3 and 7, both models line up very well with the 45^{deg}-line indicating that the residuals are equally plausibly normally distributed. Fifth, comparing the plots of standardized residuals vs. fitted values, Panels 4 and 8, Model 1 shows a marked U-shape showing dependence on fitted values and a missing nonlinear explanatory variables, whereas the plot for Model 2 is more uniformly scattered, looking like the "stars in the night sky."

(5.) *In the 1984 article "Diversity and Distribution of the Pedunculate Barnacles," in* Crustaceana, *researchers gave counts of the Octolasmis tridens and the* O. lowei *barnacles on each of 10 lobsters. Does it appear that the barnacles compete for space on the surface of a lobster? If they do compete, do you expect that the number of* O. tridens *and the number of* O. lowei *to be positively or negatively correlated? Run a test to see if the barnacles compete for space. State the assumptions on the data. State the null and alternative hypotheses. State the test statistic and the rejection region. Compute the statistic and state your conclusion.*

```
> O.tridens = c(  645, 320, 401, 364, 327,  73,  20, 221,   3,   5   )
> O.lowei   = c(    6,  23,  40,   9,  24,   5,  86,   0, 109, 350   )
> cor(O.tridens, O.lowei)
[1] -0.551912
> plot(O.tridens, O.lowei, pch = 19, main = "Barnacle Counts per Lobster")
```



The scatterplot show that when there are few *O. tridens* then there are many *O. lowei* and vice versa. There seems to be a competition, which we expect will yield a negative correlation between the counts of the two barnacle species.

To do a test of correlation, it is assumed that the data points (x_i, y_i) some from a bivariate normal distribution. Since the scatterplot is "L"-shaped, this may not hold, although it is hard to tell from such a small sample. The null and alternative hypotheses are on the correlation coefficient.

$$\mathcal{H}_0 : \rho = 0; \qquad \mathcal{H}_a : \rho < 0.$$

To test against zero, we use the Pearson's correlation test. The test statistic is

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

Under the null hypothesis, it is distributed according to the t-distribution with $n-2$ degrees of freedom. Thus the rejection region is $t \leq -t_{\alpha,n-2}$ for the one-tailed test. There are $n = 10$ data points. Computing, we find

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{(-0.551912)\sqrt{8}}{\sqrt{1-(-0.551912)^2}} = -1.871973$$

At the significance level $\alpha = .05$ we find $t_{\alpha,n-2} = t_{.05,8} = 1.860$ from Table A5 in the text. Thus we reject the null hypothesis. There is significant evidence that the correlation is negative.