Math 3080 § 1. **Second Midterm Exam** Name: Solutions
Treibergs March 23, 2016

1. *The article "Cotton Square Damage by the Plant Bug* Lygus hesperus*...," in* J. Econ. Entom.*, 1988, describes an experiment to relate the age x of cotton plants (in days) to the percentage of damaged squares y. Obtain the equation of the least-squares line.*

```
> rbind(x,y)
x    9   12   12   15   18   18   21   21   27   30   30   33
y   11   12   23   30   29   52   41   65   60   72   84   93

> c(  sum(x),   sum(y), sum(x^2),  sum(x*y),   sum(y^2) )
[1]      246       572      5742      14022        35634
```

There are $n = 12$ points. We compute

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{246}{12} = 20.5$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{572}{12} = 47.667$$

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2 = 5742 - \frac{246^2}{12} = 699$$

$$S_{xy} = \sum_{i=1}^{n} x_i\, y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right) = 14022 - \frac{246 \cdot 572}{12} = 2296$$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{2296}{699} = 3.285$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = 47.667 - 3.285 \cdot 20.5 = -19.676$$

Thus the least squares line is

$$y = -19.676 + 3.285\,x.$$

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 $\{(x_i, y_i)\}$? Show that $\hat{\beta}_1$, the least squares estimator for β_1, can 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. Show that $\hat{\beta}_1$ is normally distributed with mean β_1.*

We assume in linear regression that

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

Note that $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$. Thus

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{1}{S_{xx}}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) =$$

$$= \sum_{i=1}^{n}\frac{x_i - \bar{x}}{S_{xx}} y_i - \frac{\bar{y}}{S_{xx}}\sum_{i=1}^{n}(x_i - \bar{x}) = \sum_{i=1}^{n}\frac{x_i - \bar{x}}{S_{xx}} y_i = \sum_{i=1}^{n} c_i y_i$$

where $c_i = (x_i - \bar{x})/S_{xx}$. It follows that $\hat{\beta}_1$ is a normal random variable as it is the linear combination of independent normal random variables. Its mean

$$\begin{aligned}
\mathbb{E}(\hat{\beta}_1) &= \sum_{i=1}^{n} \frac{x_i - \bar{x}}{S_{xx}} \mathbb{E}(Y_i) \\
&= \sum_{i=1}^{n} \frac{x_i - \bar{x}}{S_{xx}} (\beta_1 + \beta_1 x_i + 0) \\
&= \sum_{i=1}^{n} \frac{x_i - \bar{x}}{S_{xx}} (\beta_1 + \beta_1 \bar{x} + \beta_1 (x_i - \bar{x})) \\
&= \frac{\beta_1 + \beta_1 \bar{x}}{S_{xx}} \sum_{i=1}^{n} (x_i - \bar{x}) + \frac{\beta_1}{S_{xx}} \sum_{i=1}^{n} (x_i - \bar{x})^2 \\
&= 0 + \beta_1.
\end{aligned}$$

3. *The effect of manganese on wheat growth was studied in "Manganeses Deficiency and Toxicity Effects on Growth ... in Wheat,"* Agronomy Journal, 1984. *A quadratic regression model was used to relate plant height (in cm) and* $\log_{10}$*(added Mn), added Mn (in* μM*). State the model and the assumptions on the data. Is there strong evidence that* $\beta_2 < -2$*? State the null and alternative hypotheses. State the test statistic and rejection region. Carry out the test at significance level .05. Predict the Height of the next observation when* $\log_{10}$*(added Mn)* $= 3$*. What is the standard error of your prediction? [Hint: the estimated standard deviation of* $\hat{\beta}_1 + 3\hat{\beta}_2 + 9\hat{\beta}_2$ *is* $s_{Y\cdot 3} = 0.850$*.]*

```
logMn    -1.0  -0.4     0   0.2   1.0   2.0   2.8   3.2   3.4   4.0
Height     32    37    44    45    46    42    42    40    37    30

lm(formula = Height ~ logMn + I(logMn^2))

Residuals:
    Min      1Q  Median      3Q     Max
-3.4555 -0.6675  0.1139  1.1278  2.2578

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  41.7422     0.8522  48.979 3.87e-10
logMn         6.5808     1.0016   6.570 0.000313
I(logMn^2)   -2.3621     0.3073  -7.686 0.000118

Residual standard error: 1.963 on 7 degrees of freedom
Multiple R-squared: 0.898,Adjusted R-squared: 0.8689
F-statistic: 30.81 on 2 and 7 DF,  p-value: 0.000339

Analysis of Variance Table

Response: Height
           Df  Sum Sq Mean Sq F value    Pr(>F)
logMn       1   9.822   9.822  2.5483 0.1544437
I(logMn^2)  1 227.698 227.698 59.0770 0.0001176
Residuals   7  26.980   3.854
```

The model and assumptions on the data is that there are β_0, β_1 and β_2 such that for all i,

$$y_i = \beta_0 + \beta_1 \log_{10} x_i + \beta_2 (\log_{10} x_i)^2 + \epsilon_i$$

where the $\epsilon_i \sim N(0, \sigma^2)$ are IID normal variables.

We test

$$\mathcal{H}_0 : \beta_2 = -2 \qquad \text{versus} \qquad \mathcal{H}_a : \beta_2 < -2$$

The test statistic is

$$t = \frac{\hat{\beta}_2 - (-2)}{s_{\hat{\beta}_2}}$$

which is distributed as a t-distribution with $n - k - 1$ degrees of freedom. The rejection region at the α-level of significance is that we reject $\mathcal{H}_0$ in favor of $\mathcal{H}_a$ if $t < t(\alpha, n-k-1) = t(.05, 10 - 2 - 1) = -1.895$. Computing using the output

$$t = \frac{-2.3621 + 2}{0.3073} = -1.1783$$

Thus, we cannot reject the null hypothesis. Equivalently, the p-value is $\mathbb{P}(T < -1.1783) = .056$ from Table A8 with $\nu = 7$. The evidence that $\beta_2 < -2$ is not significant at the $\alpha = .05$ level.

If the next $\log_{10}(x_{n+1}) = 3$ then the predicted value of Height is

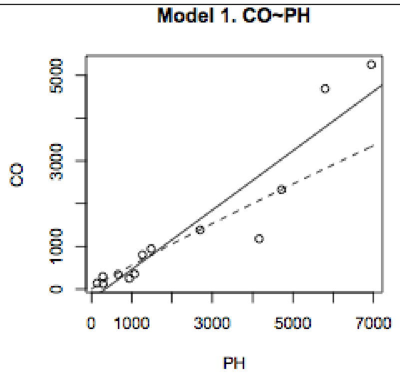
$$\begin{aligned} y_{n+1} &= \hat{\beta}_0 + \hat{\beta}_1 \log_{10} x_i + \hat{\beta}_2 (\log_{10} x_i)^2 = \beta_0 + 3\hat{\beta}_1 + 9\hat{\beta}_2 \\ &= 41.7422 + 3 \cdot 6.5808 + 9 \cdot (-2.3621) = \boxed{40.2257} \end{aligned}$$

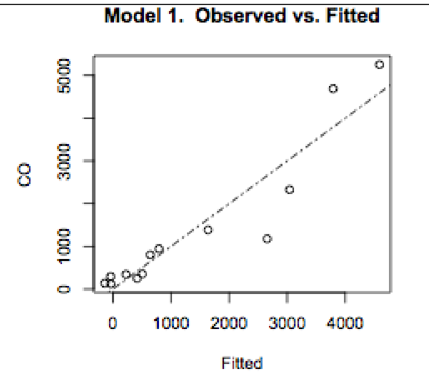
The standard error of the prediction is

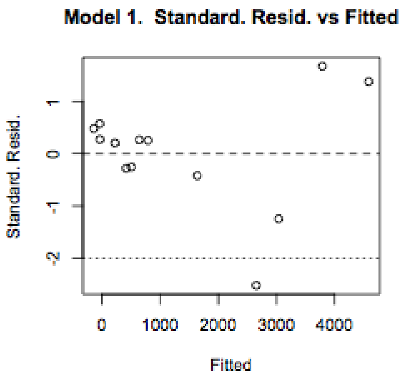
$$s_{y_{n+1}} = \sqrt{s^2 + s^2_{Y \cdot 3}} = \sqrt{(1.963)^2 + (0.850)^2} = \boxed{2.139}$$

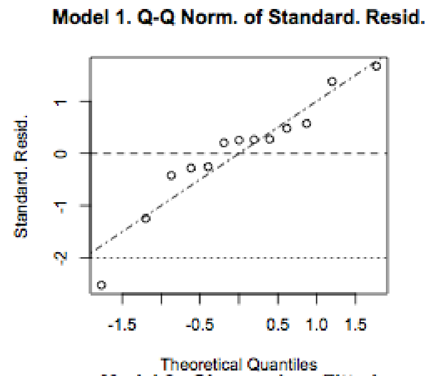
4. *A study "Forcasting Engineering Manpower..." in Journal of Management Engineering, 1995, presented data on construction costs (in \$ 1000) and person hours of labor required for several projects. Consider Model 1:* $CO = \beta_0 + \beta_1 PH$ *(solid line) and Model 2:* $\log(CO) = \beta_0 + \beta_1 \log(PH)$ *(dashed line).* **R**© *was used to generate tables and plots. 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 and why?*

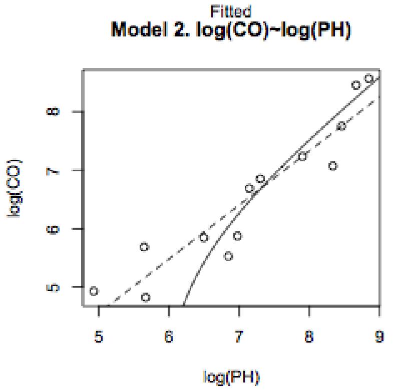
```
PH    939  5796   289   283   138  2698   663  1069  6945  4159  1266  1481  4716
CO    251  4690   124   294   138  1385   345   355  5253  1177   802   945  2327

lm(formula = CO ~ PH)

Residuals:
    Min      1Q  Median      3Q     Max
-1476.5  -165.9   151.6   277.5   899.4

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -235.32047  253.19026  -0.929    0.373
PH             0.69461    0.07854   8.844 2.49e-06

Residual standard error: 627.4 on 11 degrees of freedom
Multiple R-squared:  0.8767,Adjusted R-squared:  0.8655
F-statistic: 78.21 on 1 and 11 DF,  p-value: 2.488e-06

Analysis of Variance Table
Response: CO
          Df   Sum Sq  Mean Sq F value    Pr(>F)
PH         1 30782367 30782367  78.208 2.488e-06
Residuals 11  4329561   393596
========================================================

lm(formula = log(CO) ~ log(PH))

Residuals:
     Min       1Q   Median       3Q      Max
-0.73469 -0.34931 -0.00141  0.44173  0.53340

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.07446    0.75726  -0.098    0.923
log(PH)      0.92546    0.10417   8.884 2.38e-06

Residual standard error: 0.4519 on 11 degrees of freedom
Multiple R-squared:  0.8777,Adjusted R-squared:  0.8666
F-statistic: 78.93 on 1 and 11 DF,  p-value: 2.378e-06

Analysis of Variance Table
Response: log(CO)
          Df  Sum Sq Mean Sq F value    Pr(>F)
log(PH)    1 16.1185 16.1185  78.932 2.378e-06
Residuals 11  2.2463  0.2042
```

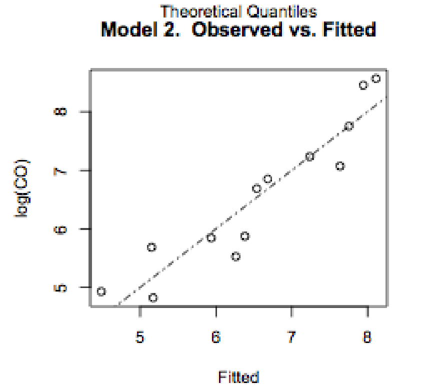

Model 1. CO~PH
CO
0 1000 3000 5000
0 1000 3000 5000 7000
PH

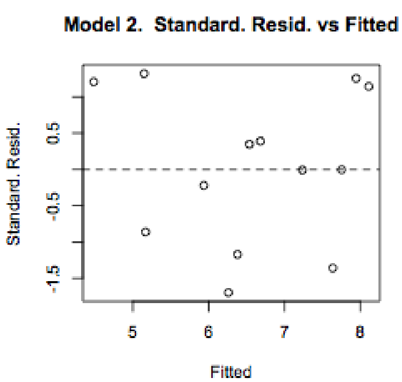

Model 1. Observed vs. Fitted
CO
0 1000 3000 5000
0 1000 2000 3000 4000
Fitted

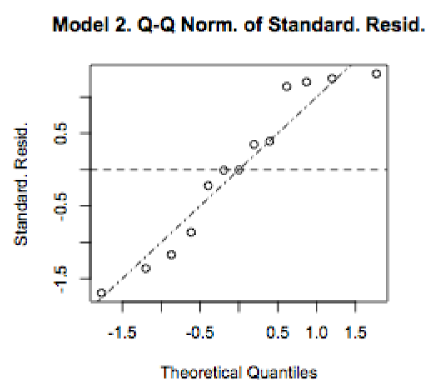

Model 1. Standard. Resid. vs Fitted
Standard. Resid.
-2 -1 0 1
0 1000 2000 3000 4000
Fitted


Model 1. Q-Q Norm. of Standard. Resid.
Standard. Resid.
-2 -1 0 1
-1.5 -0.5 0.5 1.0 1.5
Theoretical Quantiles


Model 2. log(CO)~log(PH)
log(CO)
5 6 7 8
5 6 7 8 9
log(PH)


Model 2. Observed vs. Fitted
log(CO)
5 6 7 8
5 6 7 8
Fitted


Model 2. Standard. Resid. vs Fitted
Standard. Resid.
-1.5 -0.5 0.5
5 6 7 8
Fitted


Model 2. Q-Q Norm. of Standard. Resid.
Standard. Resid.
-1.5 -0.5 0.5
-1.5 -0.5 0.5 1.0 1.5
Theoretical Quantiles

The **R**© output on the exam was incorrect since it was analysis of a different data set. The correct analysis is given here. However the diagnostic graphs given on the exam were correct ones.

First of all, the coefficient of determination, which gives the fraction of the variation accounted for the model is $R^2 = 0.8767$ for the first model and $R^2 = 0.8777$ for the second, which is a hair better but virtually the same. (On the original exam, we had $R^2 = .9188$ on the first model and $R^2 = .9704$ on the second model, which is slightly better. The adjusted R_a^2 tells us the same information as R^2 because both models have the same number of variables.)

Second, looking at the scatter plots (Panels 1 and 5) we see that in Model 1, many points bunch up an the line near zero in Model 1 whereas they have a more uniform looking spread around the dashed line in Model 2.

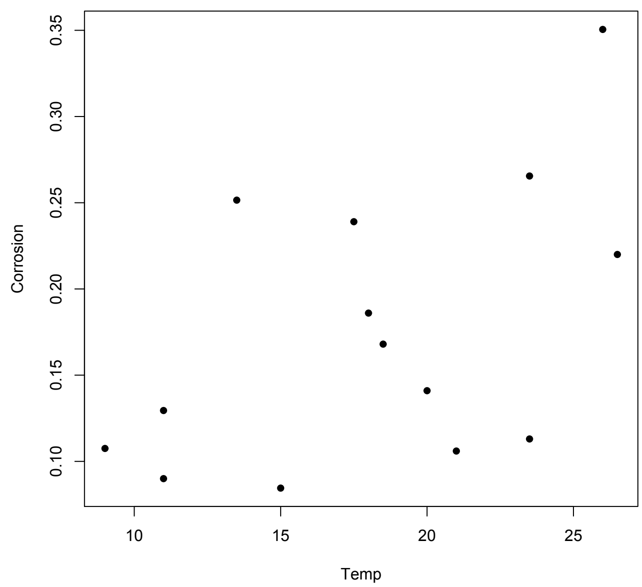
Third, comparing observed versus fitted (Panels 2 and 6), points bunch near the origin in Model 1 but have an even spread in Model 2. This suggests that the variability depends on y a lot more in Model 1.

Fourth, comparing standardized residuals versus fitted values (Panels 3 and 7), the distribution has a funnel shape for Model 1 but is much more uniform in Model 2. Also, there are no standardized residuals larger than two standard deviation in Model 2 but there are such for Model 1.

Fifth, comparing QQ-normal plots (Panels 4 and 8), both look pretty good, although there is a slight downward bow in Model 1 which is not there in Model 2. However, Model 2 has a slight "S" shape, indicating a lighter tailed distribution than normal.

In the three features, Model 2 has the advantage and they're equally good in the first and fifth, so I would recommend Model 2 over Model 1.

5. *In "The Effect of Temperature on the Marine Immersion Corrosion of Carbon Steels,"* Corrosion, *2002, corrosion loss (in mm) and mean water temperature (in C°) were measured for of copper bearing steel specimens immersed in seawater in 14 random locations after one year of immersion. Is there a linear relationship between corrosion and mean temperature? State the assumptions on the data. Is there strong evidence that $\rho > \frac{1}{3}$, where ρ is the population correlation coefficient between corrosion and temperature? State the test statistic and rejection region. Carry out the test at significance level $\alpha = .05$.*



Temp	Corrosion	Temp	Corrosion
23.5	0.2655	26.5	0.2200
18.5	0.1680	15.0	0.0845
23.5	0.1130	18.0	0.1860
21.0	0.1060	9.0	0.1075
17.5	0.2390	11.0	0.1295
20.0	0.1410	11.0	0.0900
26.0	0.3505	13.5	0.2515

```
> cor(Temp, Corrosion)
[1] 0.5230942

> plot(Corrosion~Temp,pch=16)
```

The assumption for correlation tests is that the data (x_i, y_i) are IID bivariate normal variables.

The null and alternate hypotheses are

$$\mathcal{H}_0 : \rho = \rho_0 = \frac{1}{3} \qquad \text{versus} \qquad \mathcal{H}_a : \rho > \rho_0 = \frac{1}{3}$$

The test statistic uses Fisher's z-transform

$$f(\rho) = \frac{1}{2} \ln \left(\frac{1+\rho}{1-\rho} \right).$$

For bivariate normal data with correlation ρ, then $f(\hat{\rho})$ is approximately distributed as normal random variable with mean $f(\rho)$ and variance $1/(n-3)$, where $\hat{\rho}$ is the sample correlation. Hence we reject $\mathcal{H}_0$ in favor of $\mathcal{H}_a$ if

$$z = \frac{f(\hat{\rho}) - f(\rho_0)}{1/\sqrt{n-3}} > z_\alpha$$

Since the critical z-value is $z_{.05} = 1.645$, we compute

$$Z = \frac{f(.523) - f(\frac{1}{3})}{1/\sqrt{14-3}} = \frac{\frac{1}{2} \ln \left(\frac{1+.523}{1-.523} \right) - \frac{1}{2} \ln \left(\frac{1+\frac{1}{3}}{1-\frac{1}{3}} \right)}{1/\sqrt{11}} = 0.776.$$

Since $Z < z_\alpha$ we cannot reject the null hypothesis. Equivalently, the p-value is $\mathbb{P}(Z > 0.776) = \Phi(-0.776) = .2165$. The evidence that $\rho > 1/3$ is not significant at the $\alpha = .05$ level.