Probability and Statistics for Engineers Homework Four
TMATH 390

Ben Foster*
Instructor: Julia Eaton

May 8, 2015

*Institute of Technology, University of Washington Tacoma

Contents

Chapter 3 questions: all but one should be done in R.

Problem 1

(Interpreting output from statistical software) Here is a relatively standard-looking output from some statistical software. The data deals with predicting concrete strength from its modulus of elasticity.

Predictor	Coeff	Stdev	t-ratio	p
Constant	3.2925	0.6008	5.48	0.000
mod elas	0.10748	0.01280	8.40	0.000

s = 0.8657 R-sq = 73.8% R-sq (adj) = 72.8%

Analysis of Variance:

SOURCE	DF	SS	MS	F	p
Regression	1	52.870	52.870	70.55	0.000
Error	25	18.736	0.749		
Total	26	71.605			

User your knowledge of regression and the various quantities that arise within regression to:
(a) Identify the estimated intercept and slope in the regression equation, and interpret them.
(b) Identify SS_{total}, $SS_{explained}$, and $SS_{unexplained}$.
(c) Identify R^2 and state what it means (it is a percentage of...).

Answer to a From looking at the given data, the slope is $\alpha = 0.10748$ and the intercept is $\beta = 3.2925$. This shows that the line with closest fit doesn't have a very large slope and seems almost uniform.

$$\hat{y} = 3.2925 + 0.10748(x)$$

Answer to b In order to find SS_{total}, $SS_{explained}$, and $SS_{unexplained}$, all we have to do is look at the given data.
$SS_{explained} = 52.870$
$SS_{unexplained} = 18.736$
$SS_{total} = 71.605$.

Answer to c From looking at the given data, $R^2 = 73.8\%$ or 0.738. This is a percentage of how well the regression line approximates the actual data points given.

Problem 2

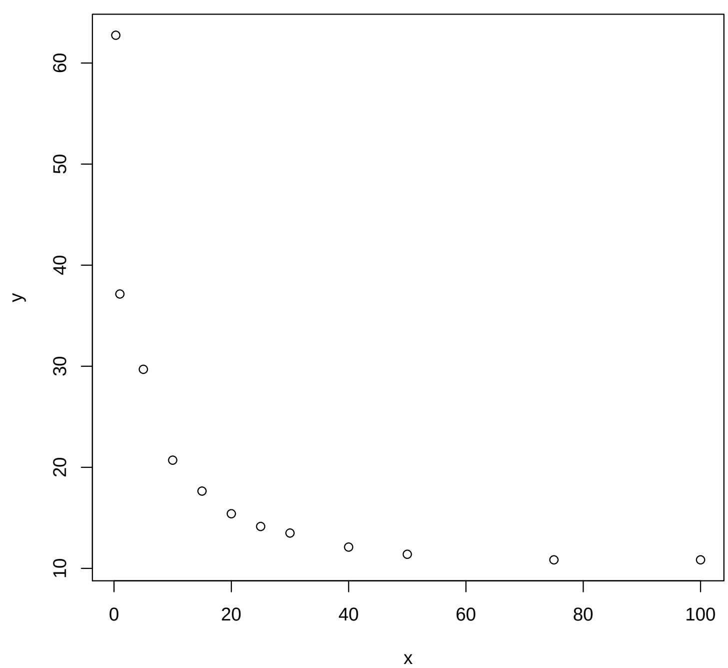
(Transforming data) The article ”Reduction in Soluble Protein and Chlorophyll Contents in a Few Plants as Indicators of Automobile Exhaust Pollution” (Intl. J. of Environ. Studies, 1983: 239-244) reported the accompanying data on x distance from a highway (meters) and y lead content of soil at a distance (parts per million, or ppm):

x:	0.3	1	5	10	15	20	25	30	40	50	75	100
y:	62.75	37.15	29.70	20.71	17.65	15.41	14.15	13.50	12.11	11.40	10.85	10.85

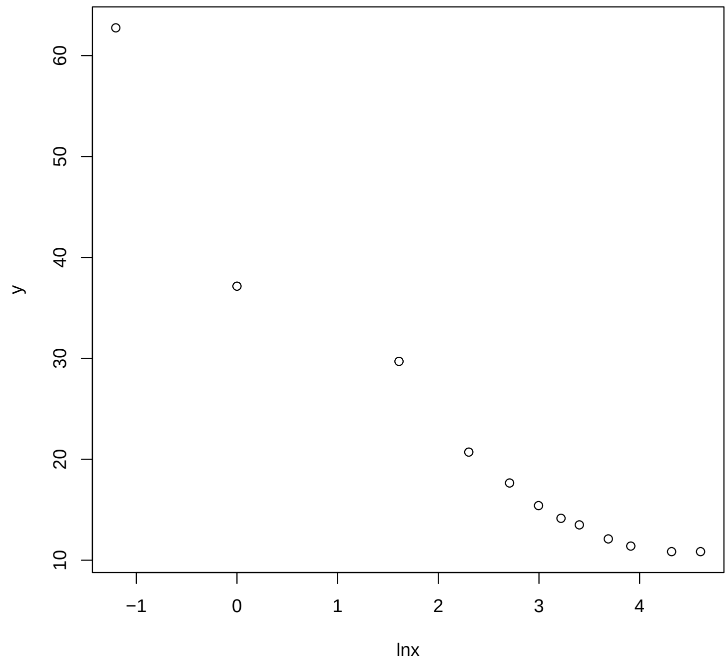
(a) Construct scatter plots of y versus x, y versus $ln(x)$, $ln(y)$ versus $ln(x)$, and $\frac{1}{y}$ versus $\frac{1}{x}$.
(b) Based on the results of part (a), which transformation does the best job of producing an approximate linear relationship?
(c) Use the selected transformation to predict lead content when distance is 45 meters.

Answer to a

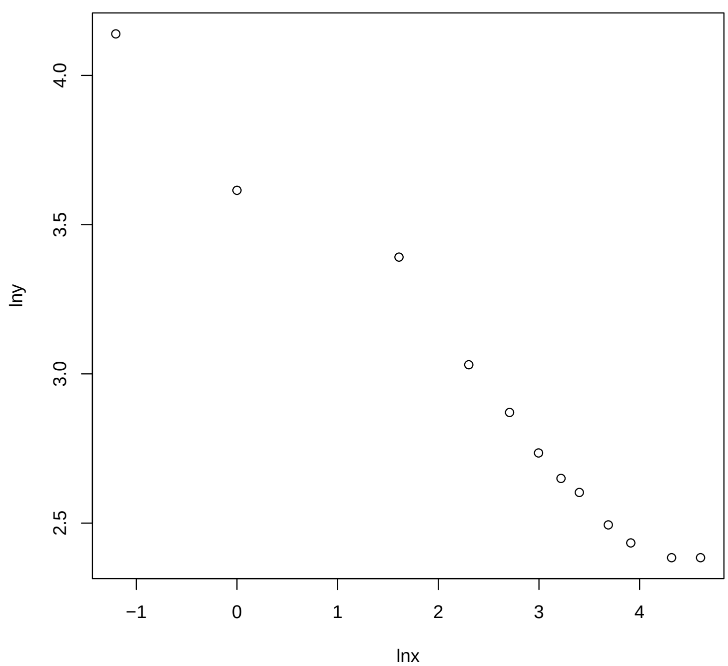
Graph of y versus x



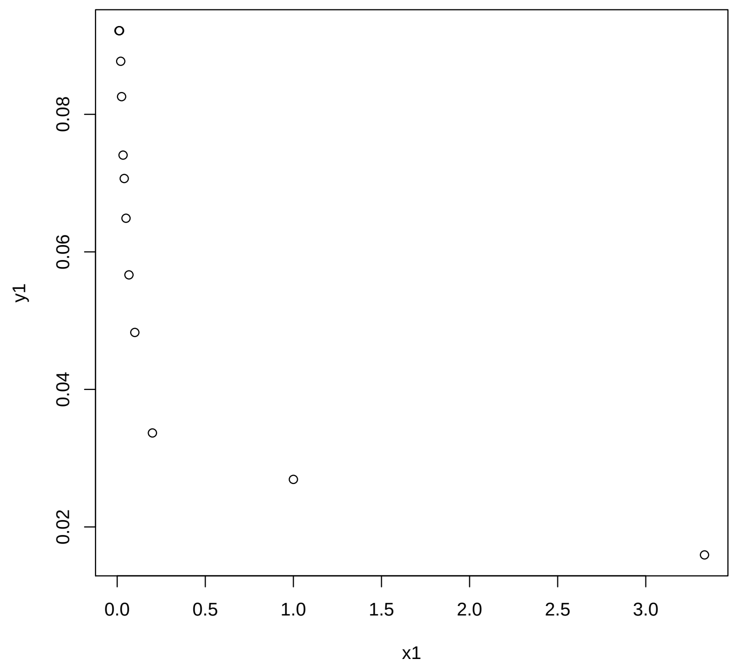
Graph of y versus $ln(x)$



Graph of $ln(y)$ versus $ln(x)$



Graph of $\frac{1}{y}$ versus $\frac{1}{x}$



Answer to b Based on the results of part(a), the transformation that does the best job of producing an approximate linear relationship is the transformation $ln(y)$ versus $ln(x)$.

Answer to c Using the selected transformation, when the distance is 45 meters, the corresponding lead content will be about 12.46 ppm.

$$ln(\hat{y}) = 3.72433 - (0.3157)ln(x)$$
$$\hat{y} = e^{2.523}$$
$$\hat{y} = 12.461$$

Problem 3

(Polynomial regression) One frequently encountered problem in crop production is deciding when to harvest to maximize yield. Data on the time to harvesting (number of days after flowering) and the yield (kg/ha) of paddy—a grain farmed in India—appeared in the article "Determination of Biological Maturity and Effect of Harvesting and Drying Conditions on Milling Quality of Paddy" (J. of Agric. Engr., 1975: 353-361), and appears below.

(time to harvest):	16	18	20	22	24	26	28	30
(paddy yield):	2508	2518	3304	3423	3057	3190	3500	3883
(time to harvest):	32	34	36	38	40	42	44	46
(paddy yield):	3823	3646	3708	3333	3517	3241	3103	2776

(a) Is it possible to transform this data as described in this section so that there is an approximate linear relationship between the transformed variables? Why or why not? (Think about the goal of the researchers: to maximize yield.)
(b) Use a statistical computer package to fit a quadratic function to this data, and then predict yield when time to harvesting is 25 days. Asses the fit of the quadratic data, i.e., interpret the R^2 and the standard deviation about regression, s_e. Remember if you want to do quadratic regression of y versus x you should use this: `lm( y ~ x + I(x^2))` in R.

Answer to a The data appears to fit a quadratic relationship more and therefore it is not possible to transform the given data to approximate a linear relationship.

Answer to b Using the code above and the command `summary(lm~x + I(x^2))`, we get the following data:

Residuals:	Min	1Q	Median	3Q	Max
	-303.96	-118.11	13.86	115.67	319.06

Coefficients:	Estimate	Std. Error	t value	Pr(<|t|)
(Intercept)	-1070.3977	617.2527	-1.734	0.107
x	293.4829	42.1776	6.958	9.94e-06
I(x^2)	-4.5358	0.6744	-6.726	1.41e-05

Residual standard error: 203.9 on 13 degrees of freedom
Multiple R-squared: 0.7942, Adjusted R-squared: 0.7625
F-statistic: 25.08 on 2 and 13 DDF, p-value: 3.452e-05

From the data, we can see that $R^2 = 0.7942$ and $s_e = 203.9$. This means that the regression line determined has about an 80% fit. This is pretty good. We are also able to grab the equation:

$$\hat{y} = -(4.5358)x^2 + (293.4829)x - 1070.3977$$

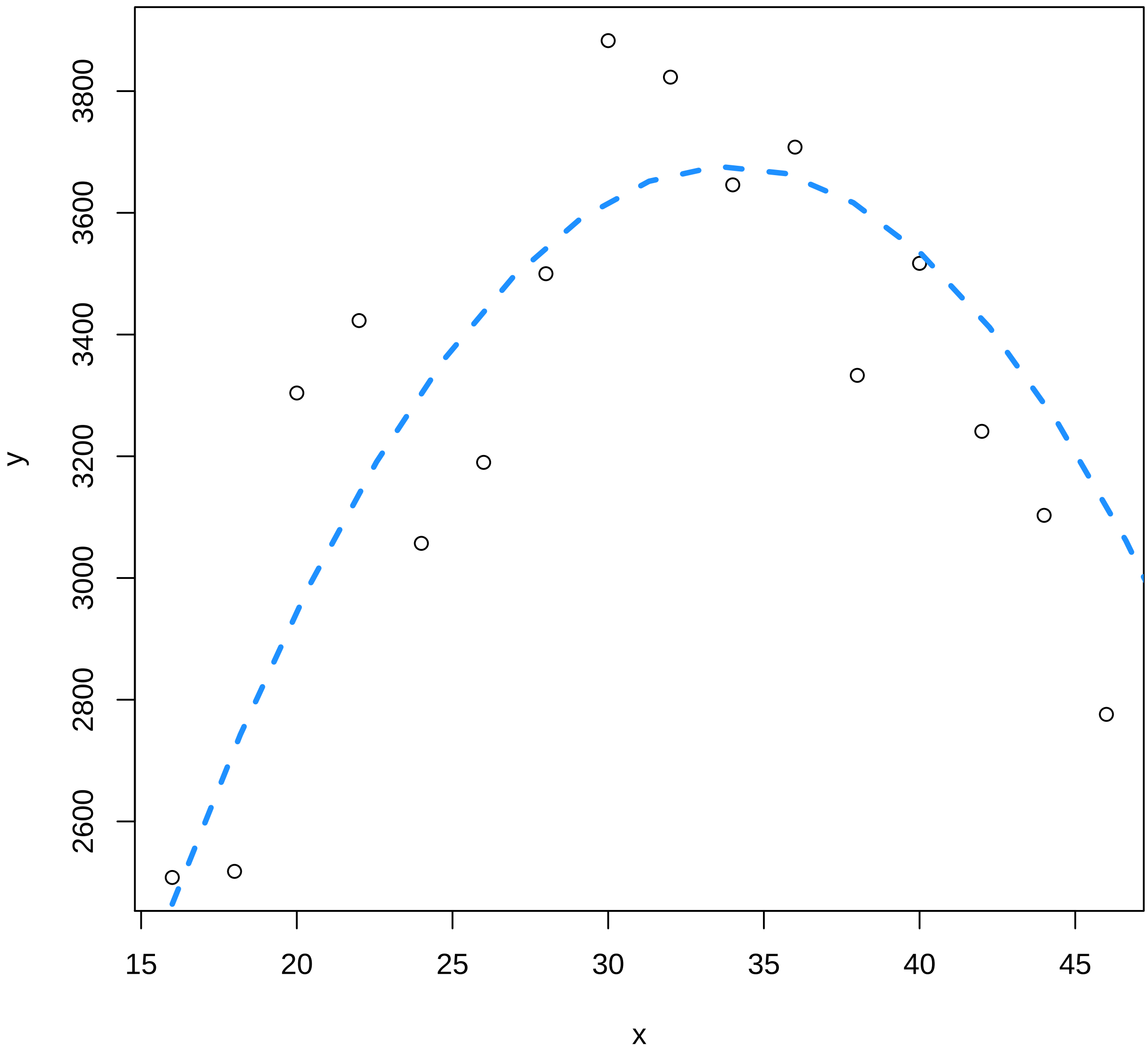
We can then use this formula to predict the yield when time to harvesting is 25 days:

$$\hat{y} = -(4.5358)(25)^2 + (293.4829)(25) - 1070.3977$$

$$\hat{y} = -2834.875 + 7337.0725 - 1070.3977$$

$$\hat{y} = 3431.7998$$

Graph of data with quadratic line fit



Problem 4

(Polynomial regression two independent variables) The article "The Undrained Strength of Some Thawed Permafrost Soils" (Canadian Geotech. J., 1979: 420-427) contained the accompanying data on y shear strength of sandy soil (kPa), x_1 depth (m), and x_2 water content (%).

Obs	Depth	Water	Strength
1	8.9	31.5	14.7
2	36.6	27.0	48.0
3	36.8	25.9	25.6
4	6.1	39.1	10.0
5	6.9	39.2	16.9
6	6.9	38.3	16.8
7	7.3	33.9	20.7
8	8.4	33.8	38.8
9	6.5	27.9	16.9
10	8.0	33.1	27.0
11	4.5	26.3	16.0
12	9.9	37.0	24.9
13	2.9	34.6	7.3
14	2.0	36.4	12.8

(a) Perform regression to predict y from x_1, x_2, x_1^2, x_2^2. Remember to put $I()$ around any terms you're squaring. You don't need it around "`x1 * x2`". Write down the coefficients of the various terms.
(b) Compute the R^2 and explain what it says about goodness-of-fit.
(c) Now perform regression to predict y from x_1 and x_2 only.
(d) Compute R^2 and explain what it says about goodness-of-fit.
(e) Compare the above two R^2 values. Does the comparison suggest that at least one of the higher order terms in the regression equation provides useful information about strength?

Note: the data for this problem are posted. Below is code that loads the file.

```
dat = read.table("Hwk4_prob_4.dat", sep="&", header=T)
y = dat$Strength
x1 = dat$Depth
x2 = dat$Water
```

Answer to a Using the command `summary(lm(y~ x1 + x2 + I(x1^2) + I(x2^2) + I(x1 * x2))` we get the following table:

Coefficients:	Estimate	Std. Error	t value	Pr(<|t|)
(Intercept)	-140.22976	136.13743	-1.030	0.3331
x1	-16.47521	9.07116	-1.816	0.1069
x2	12.82710	8.25854	1.553	0.1590
x1ˆ2	0.09555	0.07206	1.326	0.2214
x2 ˆ2	-0.24339	0.12744	-1.910	0.0925
x1 * x2	0.49864	0.23543	2.118	0.0670

From this table, we get the coefficients:

(Intercept)	x1	x2	I(x1ˆ2)	I(x2ˆ2)	x1 * x2
-140.22976	-16.47521	12.82710	0.09555	-0.24339	0.49864

Therefore:

$$\hat{y} = -140.23 - 16.48x_1 + 12.83x_2 + 0.096x_1^2 - 0.243x_2^2 + 0.499(x_1)(x_2)$$

Answer to b From the same command as above, we also get the following data:
Residual standard error: 7.023 on 8 degrees of freedom
Multiple R-squared: 0.7561, Adjusted R-squared: 0.6037
F-statistic: 4.961 on 5 and 8 DF, p-value: 0.02307

So from this we get $R^2 = 0.7561$. This means that 75.61% of variability in strength can be attributed to variation in depth and water content.

Answer to c By typing in code similar to above: `summary(lm(y~ x1 + x2))`, we can get the coefficients for predicting y from x_1 and x_2:

Coefficients:	Estimate	Std. Error	t value	Pr(<|t|)
(Intercept)	14.8893	23.2447	0.641	0.5349
x1	0.6607	0.2737	2.414	0.0344
x2	-0.0284	0.6423	-0.044	0.9655

From this table, we get the coefficients:

(Intercept)	x1	x2
14.8893	0.6607	-0.0284

Therefore:

$$\hat{y} = 14.89 + 0.66x_1 - 0.028x_2$$

Answer to d From the previous command, we get the following data:
Residual standard error: 9.019 on 11 degrees of freedom
Multiple R-squared 0.447, Adjusted R-squared: 0.3465
F-statistic: 4.446 on 2 and 11 DF, p-value: 0.03845

So from this we get $R^2 = 0.447$. This shows that this way of determining y only from x_1 and x_2 is not a very good way of predicting y. It's goodness of fit isn't very good.

Answer to e Comparing the two R^2 values above, (0.7561 and 0.4470), we can see that determining y from more predictors is more accurate. So, yes, at least one of the higher order terms in the regression equation provides useful information about strength.

Problem 5

(Interpreting more output) An experiment carried out to study the effect of the mole contents of cobalt (x_1) and the calcination temperature (x_2) on the surface area of an iron cobalt hydroxide catalyst (y) resulted in the following data ("Structural Changes and Surface Properties of CoxFe3xO4 Spinels," J. of Chemical Tech. and Biotech., 1994: 161-170):

x_1:	0.6	0.6	0.6	0.6	0.6	1.0	1.0	1.0	1.0	1.0
x_2:	200	250	400	500	600	200	250	400	500	600
y:	90.6	82.7	58.7	43.2	25.0	127.1	112.3	19.6	17.8	9.1

x_1:	2.6	2.6	2.6	2.6	2.6	2.8	2.8	2.8	2.8	2.8
x_2:	200	250	400	500	600	200	250	400	500	600
y:	53.1	53.0	43.4	42.4	31.6	40.9	37.9	27.5	27.3	19.0

A request to the SAS package to fit $\alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$ yielded the following output:

Dependent Variable: SURFAREA

Analysis of Variance:

Souce	DF	Sum of Squares	Mean Square	F Value	Prob >F
Model	3	15223.52829	5074.50943	18.924	0.0001
Error	16	4290.53971	268.15873		
Total	19	19514.06800			

Root MSE 16.47555 R-Square 0.7801
Dep Mean 48.06000 Adj R-sq 0.7389
C.V 34.07314

Parameter Estimates:

Variable	DF	Parameter Estimate	Standard Error	T	Prob <|T|
INTERCEP	1	185.486740	21.19747682	8.750	0.0001
COBCON	1	-45.969466	10.61201173	-4.332	0.0005
TEMP	1	-0.301503	0.05074421	-5.942	0.0001
CONTEMP	1	0.088801	0.02540388	3.496	0.0030

(a) Interpret the value of the coefficient of determination R^2.
(b) Predict the value of surface area when cobalt content is 2.6 and temperature is 250.
(c) Since β_1 is about -46.0, is it legitimate to conclude that if cobalt content increases by 1 unit while the values of the other predictors remain fixed, surface area can be expected to decrease by 46 units? Explain your reasoning.

Answer to a From the data provided above, we have a R^2 value of 0.7801. This means that 78.01% of the variance in y can be explained by the equation for $\hat{y}$ shown above.

Answer to b Using the following equation:

$$\hat{y} = 185.49 - 45.97x_1 - 0.302x_2 + 0.0888(x_1)(x_2)$$

Then we can predict what the value of surface area will be when the cobalt content is 2.6 (x1) and the temperature is 250 (x2).

$$\hat{y} = 185.49 - 45.97(2.6) - 0.302(250) + 0.0888(2.6)(250)$$
$$\hat{y} = 48.311 \text{ units}^2$$

Answer to c No because the cobalt content predictor is present in more than one term of the polynomial. Since the coefficient β_3 is also being multiplied by cobalt content while the other predictor remains the same, then our final value will be altered by more or less than 46 units. If β_3 were zero or if the cobalt content predictor only showed up in one term, then this would be true.