Stat 451
5-29-14

Multiple regression

(1)

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \dots + \varepsilon_i$$

x_1, x_2, x_3 are all predictors
(independent variables)

y is the outcome variable
(dependent variable)

HW#9 due 6/5
p. 533 #27
p. 552 #39bc, 41

(2)

Polynomial regression

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \dots + \varepsilon_i$$

Mixture of new predictors + their powers

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{1i}^2 + \beta_3 x_{2i} + \beta_4 x_{1i} x_{2i} + \dots + \varepsilon_i$$

③

The output will be

① A prediction model (i.e., each β_i will be estimated by b_i)

② A hypothesis test for each coefficient

$$H_0: \beta_i = 0$$
$$H_a: \beta_i \neq 0$$

③ An ANOVA that tests the entire model

$$H_0: \text{All } \beta_i = 0$$
$$H_a: \text{At least one } \beta_i \neq 0$$

④

Bayes' Rule example

<u>Quality testing</u>

Assume that 3% of the products are defective

If Good, the test certifies it 95% of the time.

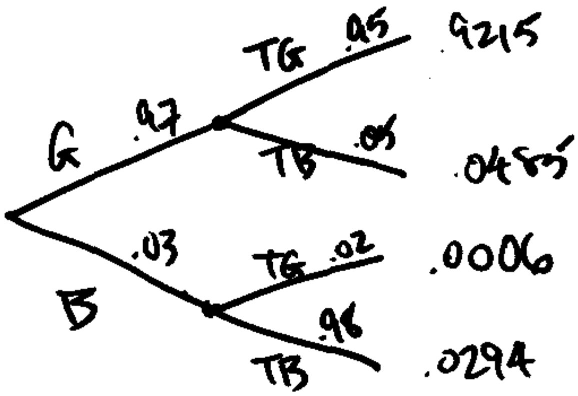
If bad, the test flags it 98% of the time.

2% false positives

5% false negatives

(1) If the test flags it as bad, what is the prob that it is actually defective.

(2) If the test certifies it, what is the prob. that it is actually good?



	TG	TB	
G	.9215	.0485	.97
B	.0006	.0294	.03
	.9221	.0779	1

(1) $P(B \mid TB) = \dfrac{P(B \cap TB)}{P(TB)} = \dfrac{.0294}{.0779} = .38$

(2) $P(G \mid TG) = \dfrac{P(G \cap TG)}{P(TG)} = \dfrac{.9215}{.9221} = .999$

b. What proportion of observed variation in stress can be attributed to the approximate linear relationship between the two variables?

c. Determine a 90% confidence interval for the true average threshold stress of all similar steel specimens whose yield strength is 800 MPa.

d. Determine a 90% prediction interval for the threshold stress of a single steel specimen whose yield strength is 800 MPa.

27. Milk is an important source of protein. How does the amount of protein in milk from a cow vary with milk production? The article "Metabolites of Nucleic Acids in Bovine Milk" (*J. of Dairy Science*, 1984: 723–728) reported the accompanying data on x = milk production (kg/day) and y = milk protein (kg/day) for Holstein-Friesan cows:

x:	42.7	40.2	38.2	37.6	32.2	32.2	28.0
y:	1.20	1.16	1.07	1.13	.96	1.07	.85
x:	27.2	26.6	23.0	22.7	21.8	21.3	20.2
y:	.87	.77	.74	.76	.69	.72	.64

Relevant calculated values include $S_{xx} = 762.012$, $b = .024576$, $a = .175576$, SSTo $= .48144$, and SSResid $= .02120$.

a. Does the simple linear regression model specify a useful relationship between production and protein?

b. Estimate true average protein for all cows whose production is 30 kg/day; use a confidence interval with a confidence level of 99%. Does the resulting interval suggest that this mean value has been precisely estimated? Explain your reasoning.

c. Calculate a 99% prediction interval for the protein from a single cow whose production is 30 kg/day.

28. Obtain an expression for s_a, the estimated standard deviation of the intercept of the least squares line. Then use the fact that $t = (a - \alpha)/s_a$ has a t distribution with $n - 2$ df to test H_0: $\alpha = 0$ for the data in Exercise 27 (this null hypothesis says that the population regression line passes through the origin). *Hint:* When $x = 0$, $\hat{y} = a + b(0) = a$, and we have a general expression for $s_{\hat{y}}$.

11.4 MULTIPLE REGRESSION MODELS

The regression models considered thus far have involved relating the dependent or response variable y to a single independent or predictor variable x. But it is virtually always the case that a model relating y to two or more predictors will explain more variation and provide better predictions than will a model with just a single predictor. For example, we should be able to predict fuel efficiency of a car more precisely from knowing *both* engine size and weight of the car than from knowing only one of these variables. Let k denote the number of predictor variables to be used in a model, and denote the predictors themselves by $x_1, x_2, \ldots, x_k$ (previously $x_1, x_2, \ldots$ represented various values of the single variable x, whereas now they represent different variables). For example, let y be the concentration of a certain chemical contaminant in an industrial worker's bloodstream. Then we might use the four predictors

x_1 = number of years of exposure to the contaminant
x_2 = number of years since the last exposure
x_3 = age of the worker
x_4 = a quantitative index of body mass

It is almost never true that the value of y is completely and uniquely determined by values of $x_1, \ldots, x_k$. A probabilistic relationship is obtained by starting with some deterministic function $f(x_1, \ldots, x_k)$ and adding (or perhaps multiplying by) a random deviation e to incorporate uncertainty due to various other factors.

c. When $x_1 = 11.5$ and $x_2 = 40$, the estimated standard deviation of $\hat{y}$ is $s_{\hat{y}} = .02438$. Calculate a 95% confidence interval for true average deposition rate for the given values of x_1 and x_2.

d. Calculate a 95% prediction interval for the deposition rate resulting from a single experimental run with $x_1 = 11.5$ and $x_2 = 40$.

36. Exercise 37 of Section 3.5 gave R output for a regression of y = deposition over a specified time period on two complex predictors x_1 and x_2 defined in terms of PAH air concentrations for various species, total time, and total amount of precipitation. Use the output in that exercise to answer the following:

a. Does there appear to be a useful linear relationship between y and at least one of the predictors?

b. The estimated standard deviation of $\hat{y}$ when x_1 is 20,000 and x_2 is .002 is $s_{\hat{y}} = 21.7$. Calculate a 95% confidence interval for the mean value of deposition under these circumstances.

c. Fitting the model with predictors x_1 and x_2 gave SSResid = 27,454, whereas fitting with x_1, x_2, and $x_3 = x_1x_2$ resulted in SSResid = 20519. Using $\alpha = .01$, can we conclude that the x_1x_2 term adds useful information to a 'reduced' model containing only x_1 and x_2? Note: when $g = 1$, the resulting F test gives the same conclusion as the t-test for whether a single variable (here, x_1x_2) contributes useful information to a model.

37. The article "Analysis of the Modeling Methodologies for Predicting the Strength of Air-Jet Spun Yarns" (*Textile Res. J.*, 1997: 39–44) reported on a study carried out to relate yarn tenacity (y, in g/tex) to yarn count (x_1, in tex), percentage polyester (x_2), first nozzle pressure (x_3, in kg/cm^2), and second nozzle pressure (x_4, in kg/cm^2). The estimate of the constant term in the corresponding multiple regression equation was 6.121. The estimated coefficients for the four predictors were $-.082$, .113, .256, and $-.219$, respectively, and the coefficient of multiple determination was .946.

a. Assuming that the sample size was $n = 25$, state and test the appropriate hypotheses to decide whether the fitted model specifies a useful linear relationship between the dependent variable and at least one of the four model predictors.

b. Again using $n = 25$, calculate the value of adjusted R^2.

c. Calculate a 99% confidence interval for true mean yarn tenacity when yarn count is 16.5, yarn contains 50% polyester, first nozzle pressure is 3, and second nozzle pressure is 5 if the estimated standard deviation of predicted tenacity under these circumstances is .350.

38. A regression analysis carried out to relate y = repair time for a water filtration system (hr) to x_1 = elapsed time since the previous service (months) and x_2 = type of repair (1 if electrical and 0 if mechanical) yielded the following model based on $n = 12$ observations: $\hat{y} = .950 + .400x_1 + 1.250x_2$. In addition, SSTo = 12.72, SSResid = 2.09, and $s_{b_2} = .312$.

a. Does there appear to be a useful linear relationship between repair time and the two model predictors? Carry out a test of the appropriate hypotheses using a significance level of .05.

b. Given that elapsed time since the last service remains in the model, does type of repair provide useful information about repair time? State and test the appropriate hypotheses using a significance level of .01.

c. Calculate and interpret a 95% confidence interval for β_2.

d. The estimated standard deviation of a prediction for repair time when elapsed time is 6 months and the repair is electrical is .192. Predict repair time under these circumstances by calculating a prediction interval with a 99% prediction level. Does the resulting interval suggest that the estimated model will give an accurate prediction? Why or why not?

39. The accompanying data on x = frequency (MHz) and y = power (W) for a certain laser configuration was read from a graph in the article "Frequency Dependence in RF Discharge Excited Waveguide CO_2 Lasers" (*IEEE J. of Quantum Electronics*, 1984: 509–514):

x:	60	63	77	100	125	157	186	222
y:	16	17	19	21	22	20	15	5

Fitting a quadratic regression model to this data yielded the following summary quantities: $a = -1.5127$, $b_1 = .391902$, $b_2 = -.00163141$, SSResid = .29, SSTo = 202.87, and $s_{b_2} = .00003391$.

a. Why is b_2 negative rather than positive?

b. What proportion of observed variation in output power can be attributed to the model relationship between power and frequency?

c. Carry out a test of hypotheses to decide whether the quadratic regression model is useful.

d. Carry out a test of hypotheses to decide whether the quadratic predictor should be retained in the model.

e. When $x = 150$, the estimated standard deviation of $\hat{y}$ is $s_{\hat{y}} = .1410$. Calculate a 99% confidence interval for true average power when frequency is 150, and also a 99% prediction interval for a single output power observation to be made when frequency is 150.

40. The article "Sensitivity Analysis of a 2.5 kW Proton Exchange Membrane Fuel Cell Stack by Statistical Method" (*J. of Fuel Cell Sci. and Tech.*, 2009: 1–6) used regression methodology to investigate the relationship between fuel cell power (W) and the independent variables x_1 = H_2 pressure (psi), x_2 = H_2 flow (stoc), x_3 = air pressure (psi), and x_4 = airflow (stoc).

Here is the Minitab output from fitting the model with the aforementioned independent variables as predictors (also fit by the authors of the cited article):

```
Predictor      Coef   SE Coef      T      P
Constant     1507.3     206.8   7.29  0.000
x1           -4.282     4.969  -0.86  0.407
x2             7.46     62.11   0.12  0.907
x3          -0.9162    0.6227  -1.47  0.169
x4            90.60     24.84   3.65  0.004

S = 4.6885   R-sq = 59.6%   R-sq(adj) = 44.9%

SOURCE       DF     SS     MS     F      P
Regression    4  40048  10012  4.06  0.029
Res.Error    11  27158   2469
Total        15  67206
```

a. Does there appear to be a useful relationship between power and at least one of the predictors? Carry out a formal test of hypotheses.

b. Fitting the model with predictors x_3, x_4, and the interaction x_3x_4 gave $R^2 = .834$. Does this model appear to be useful? Can an F test be used to compare this model to the model of part (a)? Explain.

c. Fitting the model with all 4 predictors as well as all second-order interactions gave $R^2 = .960$ (this model was also fit by the investigators). Does it appear that at least one of the interaction predictors provides useful information about power over and above what is provided by the first-order predictors? State and test the appropriate hypotheses using a significance level of .05.

41. The article "The Undrained Strength of Some Thawed Permafrost Soils" (*Canadian Geotechnical J.*, 1979: 420–427) reported the following data on undrained shear strength of sandy soil (y, in kPa), depth (x_1, in m), and water content (x_2, in %):

Obs	Depth	Watcont	Shstren
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	16.0
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	37.8	24.9
13	2.9	34.6	7.3
14	2.0	36.4	12.8

Fitting the model with predictors x_1 and x_2 only gave SSResid = 894.95, whereas fitting the complete second-order model with predictors x_1, x_2, x_1^2, x_2^2, and x_1x_2 resulted in SSResid = 390.64. Carry out a test at significance level .01 to decide whether at least one of the second-order predictors provides useful information about shear strength.

42. Soluble dietary fiber (SDF) can provide health benefits by lowering blood cholesterol and glucose levels. The article "Effects of Twin-Screw Extrusion on Soluble Dietary Fiber and Physicochemical Properties of Soybean Residue" (*Food Chemistry*, 2013: 884–889) reported the following data on y = SDF content (%) in soybean residue and the three predictors x_1 = extrusion temperature (in °C), x_2 = feed moisture (in %), and x_3 = screw speed (in rpm) of a twin-screw extrusion process.