

Handed out: Wednesday October 28, 2009

Due: Wednesday November 11, 2009

Reading:	Oct. 28	Finish simple regression (Chap 1-4 as listed on HW 4)
	Oct. 30	No lecture (out of town)
	Nov. 2	Bivariate normal / correlation (Section 2.11)
	Nov. 4	MIDTERM (thru simple regression)
	Nov. 6	No lecture (out of town)
	Nov. 9 - Nov. 16	Multiple regression basics (Chap 6-7, review Chap 5)
	Nov. 10	Note: lecture during discussion

NOTE - Midterm exam is in class on Wednesday November 4, 2009. You should bring a calculator and some blank paper. No notes or books are permitted. This HW includes questions on simple regression; the topics covered by these questions (questions 1-6 but especially questions 1-3) are included on the exam. We will have discussion next week (Tuesday November 3 from 4-6pm – please make sure you look at this HW and last year’s exam before that discussion section.

1. **Basics of regression I - baseball:** The baseball World Series (an admittedly exaggerated name since it only contains US teams) begins tonight (Wed Oct 28). There is concern that teams with more money to spend on players have more success than teams that have less money to spend. The payroll (X) (in millions of dollars) and the number of games won in the season (out of 162) (Y) are provided in the table below for the 30 major league teams.

team	payroll	wins	team	payroll	wins	team	payroll	wins
nyy	201	103	atl	97	86	kc	71	65
nym	149	70	chw	96	79	tex	68	87
chc	135	84	sf	83	88	bal	67	64
bos	122	95	cle	82	65	min	65	86
det	115	86	tor	81	75	tb	63	84
laa	114	97	mil	80	80	oak	62	75
phi	113	93	stl	78	91	was	60	59
hou	103	74	col	75	92	pit	49	62
lad	100	95	cin	74	88	sd	44	75
sea	99	85	arz	74	70	fla	37	87

From these data we compute the following summaries:

$$\sum_{i=1}^{30} x_i = 2657 \quad \sum_{i=1}^{30} x_i^2 = 268469 \quad \sum_{i=1}^{30} (x_i - \bar{x})^2 = 33147.37 \quad \sum_{i=1}^{30} x_i y_i = 220490$$

$$\sum_{i=1}^{30} y_i = 2430 \quad \sum_{i=1}^{30} y_i^2 = 200616 \quad \sum_{i=1}^{30} (y_i - \bar{y})^2 = 3786 \quad \sum_{i=1}^{30} (x_i - \bar{x})(y_i - \bar{y}) = 5273$$

- Find the least squares regression line for using payroll to predict the number of wins.
 - The sum of squared residuals for the best fitting regression line is 2947.18. Find the t-statistic for testing the hypothesis that the slope is equal to 0. What conclusion do you draw from this?
 - Give a 95% CI for the expected number of wins for a team with \$100 million payroll. Give a 95% PI for the number of wins for a team with \$100 million payroll. Explain the difference between these two intervals.
 - The league commissioner notes that the Florida Marlins with the lowest payroll won 87 games and the NY Mets with the second highest payroll won 70 games. This, he argues, proves there is no advantage to teams with a higher payroll. Comment on his argument.
2. **Basics of regression II - temperature and latitude:** A climate researcher tries to relate the average April temperature of 20 U.S. cities to the latitude of the city. She finds that a 5-degree change in latitude is associated with an 8 degree (Fahrenheit) drop in average April temperature. Her results also show that the predicted average April temperature for L.A. (latitude = 34 degrees north) is 63 degrees (Fahrenheit). About 95% of the cities in the sample have average April temperatures within +/- 6 degrees Fahrenheit of the predictions from the regression lines.
- From the information provided infer the equation of the regression line for predicting temperature from latitude and the (approximate) residual standard deviation of the regression.
 - The s.d. of the average April temperatures in this collection of cities is 9.5 degrees (Fahrenheit). What is R^2 (approximately) for the regression?

3. **SAS: Simple linear regression** – One key issue in health care reform is to determine what exactly is driving up the costs of health care. The text provides a data set from one health insurance company regarding the total costs for services provided to 788 subscribers with ischemic (coronary) heart disease during 1998 and 1999. Each line in the data set (ischemic.txt, ischemic.xls, ischemic.csv) provides an identification number and nine other variables:

id = identification number
 cost = total cost of claims in dollars
 age = age of subscriber
 gender = gender of subscriber (1 if male; 0 if female)
 interv = number of interventions or procedures
 drugs = number of drugs prescribed
 ervisit = emergency room visits
 complic = number of complications
 comorb = number of other diseases that the subscriber had
 dur = number of days of duration of treatment

It is almost certain that more than one predictor is needed to explain total cost. We start this week however using just one predictor to get used to performing regression analyses in SAS. (A sample SAS regression program is on the course website. It shows how to save and plot residuals which will come in handy for this question. It also includes options that give confidence and prediction intervals for each case.)

- Plot the cost (Y) vs duration (X). Does a linear model seem appropriate? Explain.
- Regardless of your answer, find the least squares regression line that relates cost to duration.
- Plot the residuals (studentized or regular) vs the predicted value of cost. Also obtain a normal probability plot of the residuals. Do any of the usual regression assumptions appear to be violated? If so, explain which ones.
- Rerun the regression with logarithm of cost as the response. Do the usual regression assumptions seem appropriate now? Discuss.
- The remainder of the problem uses the log transformed cost as the response variable. Give a 95% confidence interval for the slope of the regression and indicate whether there is a significant effect of duration on cost. Explain how to interpret the slope. What is the impact of an additional day of treatment on the expected log(cost)? the expected cost?
- Predict the total cost for a subscriber with dur = 182 (six months of treatment) and give a 95% confidence interval for the predicted cost. (SAS Hint: You can do this easily by adding a line to the data set with dur = 182 and cost and other predictors “missing” (indicated by putting a “.”. SAS will not use this point for the regression calculations but will provide a prediction analysis for it.)
- Summarize your regression results in a paragraph. Your discussion should (as usual) include what you are trying to learn, what you did learn from the statistical analysis, and what limitations if any are associated with your analysis.

4. **Regression/ANOVA/Model Checking** – Checking regression models based on traditional test statistics can be difficult because any evidence of an incorrect model (e.g., a quadratic trend in the data) is mixed together with individual variation in the error term. There is one special case in which we can formally check our simple linear regression model with a very convincing statistical test. Suppose that there are repeated observations at one or more of the x -values – in that case we can use the repeated observations to provide a fairly precise estimate of the variation that is due to individual variation. We illustrate a formal model checking approach here with some data from small study of chemical decay over time. Fifteen samples of a chemical solution are prepared with the same initial concentration. The fifteen samples are randomly assigned to 5 conditions which correspond to 1-hour, 3-hour, 5-hour, 7-hour and 9-hour waiting times. At the end of the waiting time the concentration is measured again. There is interest in a linear model for the decay in concentration over time. The data are provided below.

Time	Concentration(s)
1	2.57, 2.84, 3.10
3	1.07, 1.15, 1.22
5	0.49, 0.53, 0.58
7	0.16, 0.17, 0.21
9	0.07, 0.08, 0.09

- We fit a linear regression to these data and find the slope estimate is $\hat{\beta}_1 = -0.324$ with a standard error of 0.043. The ANOVA table for the linear regression model is:

Source	df	SS
Model	1	12.60
Error	13	2.92
Total	14	15.52

- i. Interpret the estimated regression coefficient.
 - ii. Show that time is a significant predictor of concentration.
- (b) We can also do an ANOVA to compare concentrations across the five experimental conditions. If we do that, then we get the ANOVA table:

Source	df	SS
Model	4	15.36
Error	10	0.16
Total	14	15.52

Are there significant differences among the mean ending concentration for the five conditions? Explain.

- (c) Notice that the two models have the same SS(Total) (because the same Y) but different SS(Error) and hence different estimates of the residual standard error (square root of the MSE). The ANOVA residual standard error can be thought of as providing an estimate that measures “pure error” while the regression residual standard error combines “pure error” and “error due to lack of fit”. Explain by considering the difference between the two models.
- (d) The difference between the SS(Error) in the two models (which is equal to the difference between the SS(Model) in the two models) can be thought of as a SS(Lack-of-fit) with 3 d.f. measuring variation in ending chemical concentrations due to lack of fit of the linear model. Carry out a statistical test comparing the lack of fit mean square to the pure error mean square. What is your conclusion?

NOTE: This procedure is of limited use in multiple regression because we don’t usually have exact replicates with the same X values for every variable.

5. Theory behind residual analysis

- (a) Recall that we find the least squares (or maximum likelihood) estimates of the regression parameters by minimizing $\sum_i (Y_i - \beta_0 - \beta_1 X_i)^2$. Show that the first-order conditions for minimizing this quantity guarantee that $\sum_i e_i = \sum_i e_i X_i = 0$ where $e_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i$.
- (b) Use the results of part (a) to argue that $\sum_i e_i \hat{Y}_i = 0$.
- (c) Show that $\sum_i e_i Y_i = (1 - r^2) \sum_i (Y_i - \bar{Y})^2$ where $r = \sum_i (X_i - \bar{X})(Y_i - \bar{Y}) / \sqrt{\sum_i (X_i - \bar{X})^2 \sum_i (Y_i - \bar{Y})^2}$. (Hints: It may help to note that $\sum_i e_i Y_i = \sum_i e_i (Y_i - \bar{Y})$. Why? It may also help to remember that $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$.)
- (d) The results in (b) and (c) help explain why we plot the residuals against \hat{Y} rather than Y to detect problems with the model fit. Explain. (Hint: The sums in (b) and (c) are the numerators of the respective correlations $Corr(e_i, \hat{Y}_i)$ and $Corr(e_i, Y_i)$. What pattern would you expect when plotting e_i vs \hat{Y}_i and when plotting e_i vs Y_i .)

6. **Regression theory with no intercept** – Regression theory for the simple linear regression model is covered in detail in the textbook. In this question we consider modifying the argument for the regression model without an intercept. The relevant model is $Y_i = \beta x_i + \epsilon_i$ with the ϵ_i ’s assumed to be independent normal random variables with mean 0 and variance σ^2 .

- (a) Describe an example where you might expect Y to be linearly related to x with no intercept.
- (b) Find the least squares estimator of β in this model.
- (c) Show that the least squares estimator is unbiased and find its variance.

7. **Correlation** – We can reuse the data from problem 1 to illustrate statistical methods for drawing inferences about correlations.

- (a) Find the correlation between team payroll and the number of wins. (SAS Note: For future reference you can get correlations from SAS using PROC CORR; VAR X1 X2 X3;)
- (b) Use Fisher’s z-transformation to give a 95% confidence interval for the population correlation coefficient.
- (c) An earlier independent study in 2005 found $r = 0.70$. Is there evidence that the two “population” correlations are different? (Hint: You need to develop a two-sample test for this situation. Use the idea behind the one sample test (see class lecture notes) to construct an appropriate test based on Fisher’s Z-transformations.)