

Stat 111 Spring 2011 Week 7 Problems

1. Suppose we observe pairs of data values x_i and y_i for $i = 1, \dots, n$ individuals (for example, x_i could be the height and y_i the weight for a sample of soccer players). It is common to assume a linear model for the relationship between the mean of a random random variable Y_i and some observed *covariate* x_i :

$$\mu_i = E(Y_i | X_i = x_i) = \beta_0 + \beta_1 x_i.$$

- a) The *least squares* estimates for β_0 and β_1 are the values $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize the sum of squared deviations between the observed y_i 's and the fitted means, $\hat{\mu}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$. Derive expressions for $\hat{\beta}_0$ and $\hat{\beta}_1$.
- b) A *Normal linear model* assumes $Y_i | X_i = x_i \stackrel{\text{indep}}{\sim} N(\mu_i, \sigma^2)$. Show that the least squares estimates are also the maximum likelihood estimates (assuming the x_i 's are observed and hence known) under this model. Also derive the MLE for σ^2 .
- c) Find an unbiased estimate for σ^2 and show that it is independent of $\hat{\beta}$.
- d) Fit a simple linear regression for the soccer height and weight data and interpret the results. Use the **lm** (linear model) function. Here are commands to get a summary, a graph of the data and a residual plot.

```
out = lm(weight ~ height); summary(out)
par(mfrow=c(2,1))   ### this allows two graphs in the same window
plot(height, weight); abline(out)
plot(out$fitted, out$resid)
```

2. Suppose X and Y follow a bivariate Normal distribution.

- a) Derive the conditional distribution of $Y | X = x$ and relate this to the linear regression prediction equations. By what fraction is the variance of Y reduced when $X = x$ is observed?
- b) For the linear model $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$, with errors ϵ_i having mean 0, show that the least squares estimates are unbiased for β_0 and β_1 , even if the ϵ_i 's are not Normal.
- c) Show that the least squares estimates (regardless of whether or not you specify Normal errors) can be expressed in terms of the sample means, standard deviations and the sample correlation coefficient r_{xy} :

$$r_{xy} = \frac{1}{n-1} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right).$$

Write out the relationship between the *standardized* fitted means, $(\hat{\mu}_i - \bar{y})/s_y$, and the *standardized* x values, $(x_i - \bar{x})/s_x$. Use this expression to justify the term “regression to the mean.” Show how to get the regression estimates for predicting weight from height and for predicting height from weight from the averages, sample standard deviations and the correlation coefficient (use **cor(x,y)** to get the correlation).

- d) In a certain hypothetical community, the correlation between years of education for husband and wife pairs is 0.5, and the average and standard deviation are 12 and 3 years, respectively, for both husbands and wives. Write out the least squares equation for predicting a husband's education from his wife's education. What years of education would you predict for the husband of a wife with 18 years of education? What years of education would you predict for the wife of a husband with 15 years of education? Apparently well-educated women tend to marry men with less education, who in turn tend to marry women with even less education. Explain this apparent paradox. Give a graphical as well as a theoretical explanation.
3. Suppose $\theta_1, \dots, \theta_k \stackrel{\text{i.i.d.}}{\sim} N(\mu, A)$ and $X_i | \theta_i \stackrel{\text{indep}}{\sim} N(\theta_i, V)$, $i = 1, \dots, k$. For example, the θ_i 's might represent the mean scores on a test for the entire i th school district, and X_i is the average for a random sample of n scores from that district. If the distribution of scores for individual students in district i is $N(\theta_i, \sigma^2)$, then the average X_i has variance $V = \sigma^2/n$.
- What is the marginal distribution of X_i ? What is the correlation between X_i and θ_i ?
 - We are most interested in the θ_i 's but typically only observe the X_i 's. Assuming V , μ and A are known, what is the conditional distribution of $\theta_i | X_i = x_i$? Relate this to the simple regression equation.
 - In a simple regression situation, we observe $\{x, y\}$ pairs and fit a line that summarizes their relationship. But here we wish to fit a line based only on the x_i 's. What other information can we take advantage of? Demonstrate the James-Stein estimate for a data set of your choosing (or I can help you find one). Assume V is known but that μ and A must be estimated.

Problem to turn in: No-Intercept Linear Regression.

Suppose we have n data points $\{x_i, y_i\}$ from the following model:

$$Y_i = \beta x_i + \epsilon_i, \quad \epsilon_i \stackrel{\text{i.i.d.}}{\sim} N(0, \sigma^2).$$

The x_i 's are assumed to be observed without error and may be treated as constants.

- Write down the likelihood function $L(\beta, \sigma^2)$. Find the MLE values $\hat{\beta}$ and $\hat{\sigma}^2$, and the Information matrix: the expectation of the negative second derivative matrix for $L(\beta, \sigma^2)$.
- Find the sampling distribution of $\hat{\beta}$.
- Find an unbiased estimate for σ^2 and its sampling distribution.
Hint: Write down an expression that involves β and follows a Chi-square distribution. Add and subtract a term with $\hat{\beta}$ in place of β . Follow the progression we used for finding the distribution of $\sum (X_i - \bar{X})^2$ and s^2 for a sample of independent Normal variables.
- Find an expression for a $(1 - \alpha)100\%$ t confidence interval for β .

Additional Problem not to turn in (from a 2009 Stat 11 Final)

Baby walkers are seats hanging from frames that allow babies to sit upright with their feet touching the floor. Walkers have wheels on their legs that allow the infant to propel the walker around long before he or she can walk or even crawl. Because most walkers have tray tables in front that block babies views of their feet, child psychologists have begun to question whether walkers affect infants cognitive development. One study compared the mental skills of a random sample of those who used walkers with a random sample of those who had never used walkers. The summary results on the mental skill scores for the study are in the following table:

Group	n	\bar{x}	s
Used walkers	54	113	12
No walkers	55	123	15

- a) (2 pts) Find a 95% confidence interval for the mean test score of the “used walkers” group.
- b) (2 pts) Which of the following is the best interpretation of the interval in part a? (Choose one)
- i) 95% of babies who used walkers have a test score in this range.
 - ii) 95% of the sampled babies who used walkers have a test score in this range.
 - iii) 95% of samples of 54 babies who used walkers will have an average test score in this range.
 - iv) 95% of samples of 54 babies who use walkers would give an interval that contains the mean test score for all babies who use walkers.
 - v) There is probability 0.95 that the mean test score for all babies who use walkers is in this range.
- c) (3 pts) Carry out a test to determine whether there is any difference in test scores for the two groups. State hypotheses and report the p -value.
- d) (2 pts) Let “treatment” refer to the distinction between babies who use walkers and those who don’t in the population. Which of the following is the best explanation of the p -value in b (Choose one)
- i) The probability that the mean test scores differ for the two treatments.
 - ii) The probability that the mean test score is the same for the two treatments.
 - iii) The probability, assuming the mean test scores differ for the two treatments, of observing average levels as different as we did.
 - iv) The probability, assuming the mean test score is the same for the two treatments, of observing average levels as different as we did.
 - v) The probability, assuming the mean test score differs for the two treatments, of observing average levels as similar as we did.