

Stat 111 Spring 2011 Week 8 - Multiple Regression

See Section 14.4 of Rice for details about vector random variables.

1. The *Normal multiple linear regression* model can be written as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I}),$$

where $\mathbf{0}$ represents an $n \times 1$ vector of 0's and \mathbf{I} is an $n \times n$ identity matrix. The $n \times p$ matrix \mathbf{X} contains p covariates for the i th individual in the i th row, and $\boldsymbol{\beta}$ is a $p \times 1$ ($p < n$) vector of regression coefficients. If an intercept is included in the model, the first column of \mathbf{X} is all 1's and the first element of $\boldsymbol{\beta}$ corresponds to the intercept, β_0 .

- a) Write out the likelihood function for $\boldsymbol{\beta}$ and σ^2 . Find the maximum likelihood estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$. Rather than performing vector calculus to find $\hat{\boldsymbol{\beta}}$, evaluate the log-likelihood function at $\hat{\boldsymbol{\beta}}$ and at $\hat{\boldsymbol{\beta}} + \boldsymbol{\delta}$, where $\boldsymbol{\delta}$ is an arbitrary $p \times 1$ vector. Find conditions on $\hat{\boldsymbol{\beta}}$ such that $L(\hat{\boldsymbol{\beta}}, \sigma^2) \geq L(\hat{\boldsymbol{\beta}} + \boldsymbol{\delta}, \sigma^2)$, $\forall \boldsymbol{\delta} \in R^p$.
 - b) Find the sampling distribution of $\hat{\boldsymbol{\beta}}$, given $\boldsymbol{\beta}$ and σ^2 .
 - c) The usual simple linear regression model has $p = 2$, with the i th row of \mathbf{X} given by $(1, x_i)$ and $\boldsymbol{\beta} = (\beta_0, \beta_1)'$. Write out the sampling distribution of $\hat{\boldsymbol{\beta}}$ in terms of the x_i 's and σ^2 . Discuss the formulae for the variances and covariance of $\hat{\beta}_0$ and $\hat{\beta}_1$, and the choices for the x_i 's that will make these estimates most precise. Generalize this discussion to the multivariate case.
2. The *fitted values* for a regression model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ are given by $\hat{\boldsymbol{\mu}} = \mathbf{X}\hat{\boldsymbol{\beta}}$, and the *fitted residuals* by $\hat{\boldsymbol{\epsilon}} = \mathbf{Y} - \hat{\boldsymbol{\mu}}$.
 - a) Show that $\hat{\boldsymbol{\mu}} = \mathbf{H}\mathbf{Y}$, where \mathbf{H} is an $n \times n$ *projection matrix*: $\mathbf{H}' = \mathbf{H} = \mathbf{H}\mathbf{H}$. This implies that the vector of fitted values $\hat{\boldsymbol{\mu}}$ is the projection of \mathbf{Y} onto the subspace spanned by the columns of \mathbf{X} . Note that this will not work unless the columns of \mathbf{X} are linearly independent.
 - b) Show that, if $\boldsymbol{\epsilon}$ has mean 0 and covariance matrix $\sigma^2\mathbf{I}$, then $\hat{\boldsymbol{\epsilon}}$ is uncorrelated with $\hat{\boldsymbol{\mu}}$ (i.e., the cross-covariance matrix $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\epsilon}}\hat{\boldsymbol{\mu}}}$ is a matrix of 0's) but correlated with \mathbf{Y} . This has implications for how we construct our residual plots.
 - c) If $\boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I})$, find the sampling distributions of $\hat{\boldsymbol{\mu}}$ and of $\hat{\boldsymbol{\epsilon}}$.
 3. The *mean squared error* for a regression is defined as $s^2 = \hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}}/(n - p)$.
 - a) Assuming that the errors are independent with mean 0 and standard deviation σ , show that s^2 is an unbiased estimate for σ^2 .
 - b) If $\boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I})$, show that s^2 is independent of $\hat{\boldsymbol{\beta}}$ and derive the sampling distribution of s^2 .
 - c) For the Normal model, argue that $(\hat{\beta}_j - \beta_j)/s_{\hat{\beta}_j} \sim t_{(n-p)}$, $j = 0, \dots, p$, where $s_{\hat{\beta}_j}^2$ is an estimate of $Var(\hat{\beta}_j)$. Use this to identify a $1 - \alpha$ level CI for β_j .

Problem to turn in: Download the height and shoe length data and read it into R:

```
x = scan("/ ...specify path to directory.../height.txt")
xmat = matrix(x, ncol=3, byrow=T)
male = xmat[,1]; height = xmat[,2]; shoe = xmat[,3]
```

Plot the data, using different colors for the males and females:

```
plot(shoe, height, type="n") # this will generate an empty plot.
points(shoe[male==0], height[male==0], col="red")
points(shoe[male==1], height[male==1], col="blue")
```

Fit a simple regression of height on shoe length, and a multiple regression of height on shoe length and male.

```
out1 = lm(height ~ shoe); out2 = lm(height ~ shoe + male)
```

- a) Typing **summary(out1)** (e.g.) will give a summary of the regression fit (out1 is a list that contains a lot of other information). How much more of the variability in heights is explained by including gender in the model?
- b) Add lines to your graph corresponding to the overall fit from the simple regression, and for the predictions for males and for females from the multiple regression (using the same color scheme). The R function **abline** allows you to specify a line by giving its intercept and slope. For the simple regression you can type **abline(out1)**, because the first two elements of this list are the intercept and slope for the simple regression. To add the lines for males and females you must figure out the corresponding slopes and intercepts. Then type **abline(a, b, col="red")**, where a and b are the intercept and slope for women, with a similar command for men. Print your graph (note that you can control the size either in your pdf command, or by importing it into another programs such as Word and resizing). Explain why the slopes differ in the way that they do (or do not).
- c) Make residual plots for your two regressions. You can put two graphs in the same window (and pdf) by typing the following:

```
par(mfrow=c(2,1))
plot(out1$fitted, out1$resid, main="Simple Regression",
     xlab="fitted", ylab="residual")
plot(out2$fitted, out2$resid, main="Multiple Regression",
     xlab="fitted", ylab="residual")
```

Compare the residual plots and explain why one model seems more appropriate.

- d) Use the output to construct a 95% confidence interval for the difference in mean heights for men and women with the same shoe length.