

Statistics/BioSci 141, Fall 2006
Lab 4: Correlation, Regression and Contingency Tables
December 1, 2006

1 Correlation and Regression

Throughout this lab, we will be working with the dataset called babies. The data is a collection of variables taken for each new mother in a Child and Health Development Study.

```
> babies = read.table("http://www-stat.stanford.edu/~rag/stat141/exs/babies.dat", header = T)
> names(babies)
```

Here is a description of the variables we are interested in:

gestation length of gestation in days

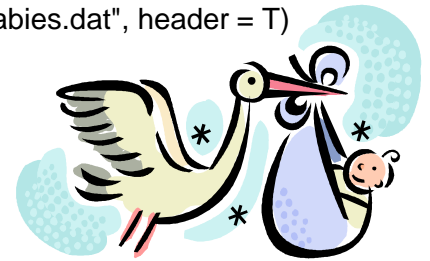
wt birth weight in ounces, 999 unknown

age mother's age in years at termination of pregnancy, 99=unknown

ht mother's height in inches to the last completed inch, 99=unknown

wt1 mother pre-pregnancy weight in pounds, 999=unknown

smoke does mother smoke? 0=never, 1= smokes now, 2=until current pregnancy, 3=once did, not now, 9 = unknown



We now need to create a subset of this data which excludes the observations with unknown values and only includes the variables we are interested in.

```
> babies = subset(babies, subset = gestation < 999 & wt1 < 999 & ht < 99
& smoke < 9 & age < 99, select = c("gestation", "smoke", "wt1", "wt",
"ht", "age"))
> attach(babies)
```

Look at your data:

```
> hist(wt, col = "yellow", border = "blue", xlab
="Birth Weight")
> hist(smoke)
> hist(wt1)
> plot(wt ~ factor(smoke), data=babies)
> plot(wt, age)
> plot(wt, gestation)
```

```
> plot( 0:25, pch = 0:25)
```

This is just to get a picture of point options for scatterplots.

Researchers are interested in seeing if there is a linear relationship between the baby's birth weight and the weight of the mother at birth. Create a plot of birth weight on the x-axis and weight of mother on the y-axis. Based on visual inspection, do birth weight and mother's weight appear to be positively correlated, negatively correlated, or neither?

```
> plot(wt, wt1)
```

It is the magnitude of the correlation coefficient, r , that tells how strong the linear relationship is between the two variables. Compute the Pearson correlation coefficient. Compare this to the Spearman-rank correlation.

```
> cor(wt, wt1, method = "pearson")
> cor(wt, wt1, method = "spearman")
> cor.test(wt, wt1)
```

Cor.test: tests for association between paired samples.

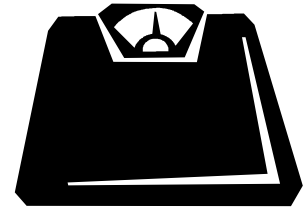
1.1 Regression

Compute the linear least squares fit, predicting birth weight as a linear function of mother's weight. Notice that our response and predictor variables are both continuous.

```
> reg = lm(wt ~ wt1)
> coef(reg)
> summary(reg)
```

Coefficients:

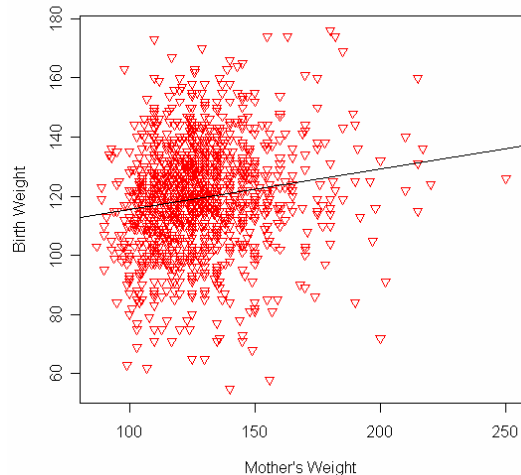
	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	101.75393	3.31927	30.655	< 2e-16	***
wt1	0.13783	0.02551	5.404	7.89e-08	***



What does this mean?

Equation: Birth Weight = 101.74 + .138 (Weight of Mother)

```
> plot(wt1, wt, xlab = "Mother's Weight", ylab = "Birth Weight", col =
"Red", pch = 25)
> abline(reg)
```



Remember: the least squares criterion is that the “best” straight line is the one that minimizes the residual sum of squares.

Look at the residual plot:

```
> plot(fitted(reg), residuals(reg), pch = 18)
```

Are the residuals normally distributed?

```
> qqnorm(residuals(reg))
```

What if we wanted to include age as a predictor of birth weight?

```
> reg2 = lm(wt ~ wt1 + age)
> summary(reg2)
```

Using the information from our model, what is the predicted weight of the baby given the mother's weight = 180 pounds and age = 29?

```
> predict.lm(reg2, newdata = data.frame(wt1 = 180, age = 29))
[1] 126.5592
```

2 Logistic Regression

Regression and correlation are used to analyze the relationship between two quantitative variables. Sometimes data arise in which a quantitative variable X is used to predict the response of a categorical variable Y. For example, we might wish to use X = cholesterol level as a predictor of whether or not a person has heart disease. When the response variable is dichotomous, a technique known as logistic regression can be used to model the relationship.

Create a variable to indicate whether or not the baby is premature (a birth is considered premature if the gestation period is less than 37 full weeks). Notice that now our response variable is categorical and our predictor variable is continuous. Use logistic regression to model how the probability of a premature baby depends on the baby's weight at birth.

```
> preemie = as.numeric(gestation < 7*37)
> babies = cbind(babies, preemie)
> table(preemie)
preemie
  0    1
1079  96

> lreg = glm(preemie ~ wt, family = binomial, data = babies)
> summary(lreg)
```

What does this mean?

$$P(\text{Baby is premature}) = P(\text{preemie} = 1) = \frac{e^{(5.017 - .067 \text{ weight})}}{1 + e^{(5.017 - .067 \text{ weight})}}$$

```
> plot(wt, preemie, xlab = "Baby's weight", ylab = "Predicted probability
of a premature baby", xlim = range(0, 180))
> curve(exp(5.017 - .067*x)/(1+exp(5.017 - .067*x)), add = TRUE, 1, 180)
```

Using the information from our model, what is the probability that the baby is premature given that we know the birth weight = 100?

```
> predict.glm(lreg, type="response", newdata = data.frame(wt = 100))
[1] 0.1559374
```

Now say we are interested in using logistic regression to model how the probability of having a premature baby depends on the mother's smoking habits.

```
> lreg1 = glm(preemie ~ factor(smoke), family = binomial, data = babies)
> summary(lreg1)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-2.52059	0.16644	-15.144	<2e-16	***
factor(smoke)1	0.17160	0.23471	0.731	0.465	
factor(smoke)2	0.29897	0.38841	0.770	0.441	
factor(smoke)3	0.08917	0.40459	0.220	0.826	

Why do you have to use factor(smoke) instead of smoke?

Now let's try multiple logistic regression. We shall use the body mass index of the mother as a measure of malnutrition. The BMI is weight in kilograms / height in meters squared.

```
> babies$BMI = with(babies,
  (wt1 / 2.2) / (ht*2.54/100) ^2)
> hist(BMI)
> lreg2 = glm(preemie ~ factor(smoke)
babies)
> summary(lreg2)
```

with(data, expr, ...): Evaluate an R expression in an environment constructed from data.
 data: data to use for constructing an environment.
 expr: expression to evaluate.

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-3.42944	0.71126	-4.822	1.42e-06	***
factor(smoke)1	0.19629	0.23572	0.833	0.405	
factor(smoke)2	0.31378	0.38897	0.807	0.420	
factor(smoke)3	0.10121	0.40499	0.250	0.803	
BMI	0.04035	0.03042	1.327	0.185	



Using the information from our model, what is the probability that a baby is premature if the mother never smoked and her BMI =20?

```
> predict.glm(lreg2, type="response", newdata = data.frame(smoke = 0, BMI
= 20))
[1] 0.06771174
```

3 Contingency Table Analysis

The focus of interest in a contingency table is the dependence or association between the column variable and row variable. Let's create a 2x2 table. We will only look at observations in which the mother either smokes or does not smoke i.e. smoke = 0 or 1.

```
> babiesnew = subset(babies, subset = smoke <2)
> tab2 = table(babiesnew$smoke, babiesnew$preemie)
> tab2
```

	0	1
0	485	39
1	419	40

```
> chisq.test(tab2)
```

Our p-value = 0.5391, so we cannot reject the null hypothesis that these two variables are independent.

3.1 General r x k Tables for Tests of Independence

We now consider a contingency table with r rows and k columns, an r x k contingency table. Does whether or not one has a premature baby appear to be independent of smoking habits (including all valid observations i.e. smoke = 0,1,2 or 3)?

```
> tab = table(preemie, smoke)
> tab
```

	smoke			
preemie	0	1	2	3

```
      0 485 419 83 91
      1 39 40 9 8
> chisq.test(tab)
```

We cannot reject the null hypothesis that the two factors are independent.



3.2 Odds Ratio

The odds of an event E is defined to be the ratio of the probability that E occurs to the probability that E does not occur. The odds ratio is the ratio of two odds under two conditions.

```
> install.packages("colorspace")
> install.packages("vcd")
> library("vcd")

> or = oddsratio(tab2, log = FALSE)
> summary(or)
      Odds Ratio
[1,]      1.1872
```

So the odds of having a preemie baby are about 1.18 as great for smokers as for nonsmokers.

```
> confint(or)
      lwr      upr
[1,] 0.7514444 1.875639
```

Notice this confidence interval contains 1.