

R Computing, Stat 141, Fall 2007
Lab 4: Correlation, Regression, Contingency Tables
Nov 30, 2007

1 Correlation

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

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

The variables we are interested in are:

- 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 but not now (9 = unknown)

To start, we would like to create a subset of the data that 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)
plot( 0:25, pch = 0:25) # make a plot of available point characters
hist(wt, col = "yellow", border = "blue", xlab="Birth Weight")
hist(smoke)
hist(wt1)
plot(wt ~ factor(smoke), data=babies)
plot(babies)
```

We 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)
```

The magnitude of the correlation coefficient, r , tells how strong the linear relationship is between 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)
```

Results are as follows:

```
[1] 0.1559233
[1] 0.1816586
      Pearson's product-moment correlation
data:  wt and wt1
t = 5.404, df = 1172, p-value = 7.887e-08
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.09959872 0.21125178
sample estimates:
      cor
0.1559233
```

When would we use Pearson over Spearman? Which test is more powerful?

Note that `cor.test()` can use one of three types of correlation, with Pearson's as the default. Note that a t -statistic is given. Recall that

$$t_s = \frac{b_1}{SE_{b_1}} = r \sqrt{\frac{n-2}{1-r^2}}.$$

2 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)
attributes(reg) # print out available attributes
summary(reg)
coef(reg) # or reg$coef
```

The result is:

```

$names
 [1] "coefficients" "residuals"      "effects"        "rank"
 [5] "fitted.values" "assign"          "qr"             "df.residual"
 [9] "xlevels"       "call"           "terms"          "model"
$class
 [1] "lm"

Call:
lm(formula = wt ~ wt1)
Residuals:
    Min       1Q   Median       3Q      Max
-66.0505 -10.9155  0.3278  11.0265  56.0845
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 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 18.11 on 1172 degrees of freedom
Multiple R-Squared:  0.02431,    Adjusted R-squared:  0.02348
F-statistic: 29.2 on 1 and 1172 DF,  p-value: 7.887e-08

(Intercept)      wt1
101.7539279    0.1378329

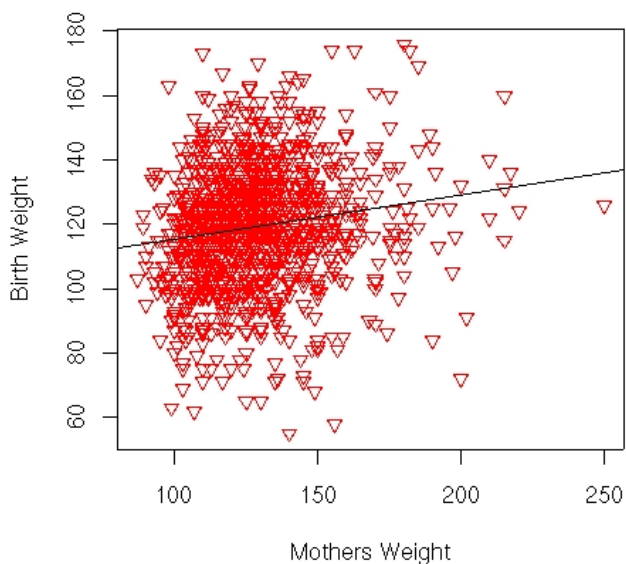
```

What does this mean? The OLS equation that best summarizes the data is $wt = 101.74 + 0.138 wt1$.

```

plot(wt1, wt, xlab = "Mothers Weight", ylab = "Birth Weight", col = "Red", pch = 25)
abline(reg)

```



Remember: the least squares method finds the “best” straight line 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)
```

The result is:

```
Call:
lm(formula = wt ~ wt1 + age)
Residuals:
    Min       1Q   Median       3Q      Max
-66.1448 -10.9632  0.3379  11.0477  56.0388
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  101.47076    3.88425   26.124 < 2e-16 ***
wt1           0.13730    0.02580    5.322 1.23e-07 ***
age           0.01292    0.09194    0.141  0.888
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 18.12 on 1171 degrees of freedom
Multiple R-Squared:  0.02433,    Adjusted R-squared:  0.02266
F-statistic:  14.6 on 2 and 1171 DF,  p-value: 5.461e-07
```

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

```
predict.lm(reg2, newdata = data.frame(wt1 = 180, age = 29))
```

Do we gain by using the second predictor?

```
anova(reg, reg2)
```

The result is:

```
Analysis of Variance Table
Model 1: wt ~ wt1
Model 2: wt ~ wt1 + age
  Res.Df  RSS   Df Sum of Sq    F Pr(>F)
1   1172 384477
2   1171 384471    1      6 0.0197 0.8883
```

The residual sum of squares measures the variation between data and the model. The null hypothesis here is that the normalized differences in the RSS for the two models is zero. Put another way, the null hypothesis is that the coefficient for the extra parameter is zero. We assume errors for both models are distributed as $N(0, \sigma^2)$. Denote the number of parameters for Model 1 by k and that for Model 2 by p . The F statistic used is:

$$F = \frac{\frac{RSS(k) - RSS(p)}{p - k}}{\frac{RSS(p)}{n - p - 1}} = \frac{RSS(k) - RSS(p)}{\hat{\sigma}^2}.$$

This is called the partial F -test. If Model 2 had more than two predictors, the null hypothesis would be that the coefficients for all of the extra predictors are zero. The alternative hypothesis is that at least one of them is not zero. In general, what does $\frac{SS}{df}$ give?

3 Logistic Regression

Regression and correlation are used to analyze the relationship between two quantitative variables. Sometimes data arise in which a quantitative variable(s) is used to predict the response of a categorical variable. For example, we might wish to use 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 a baby is premature (a birth is considered premature if the gestation period is less than 37 full weeks). This binary variable will be our response. Our predictor variable will be weight. 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)
table(preemie)
lreg = glm(preemie ~ wt, family = binomial)
summary(lreg)
```

The result is:

```
preemie
  0    1
1078  96

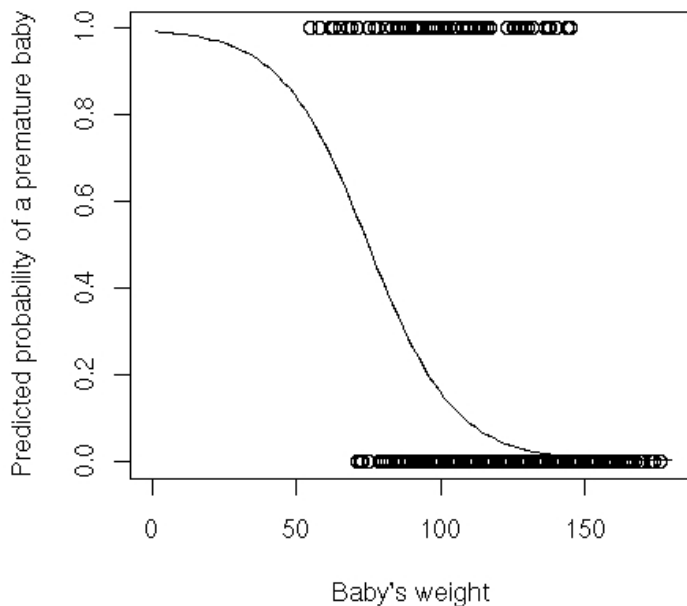
Call:
glm(formula = preemie ~ wt, family = binomial)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2879 -0.3985 -0.2784 -0.1810  3.0710

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  5.017338   0.717952   6.988 2.78e-12 ***
wt          -0.067061   0.006808  -9.851 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)
Null deviance: 664.66  on 1173  degrees of freedom
Residual deviance: 545.31  on 1172  degrees of freedom
AIC: 549.31
Number of Fisher Scoring iterations: 6
```

What do the results mean?

$$P(\text{preemie} = 1) = \frac{\exp(5.0 - 0.67wt)}{1 + \exp(5.0 - 0.67wt)} = \frac{1}{1 + \exp(-5.0 + 0.67wt)}$$

```
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 is 100?

```
predict.glm(lreg, type="response", newdata = data.frame(wt = 100))
```

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)
```

The results are:

```
Call:
glm(formula = preemie ~ factor(smoke), family = binomial, data = babies)
Deviance Residuals:
```

```

      Min       1Q   Median       3Q      Max
-0.4538 -0.4270 -0.4105 -0.3933  2.2794
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
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)
      Null deviance: 664.66  on 1173  degrees of freedom
Residual deviance: 663.80  on 1170  degrees of freedom
AIC: 671.8
Number of Fisher Scoring iterations: 5

```

Why do we 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.

```

BMI = (wt1 / 2.2) / (ht*2.54/100) ^2
lreg2 = glm(preemie ~ factor(smoke) + BMI, family=binomial)
summary(lreg2)

```

The results are:

```

Call:
glm(formula = preemie ~ factor(smoke) + BMI, family = binomial)
Deviance Residuals:
      Min       1Q   Median       3Q      Max
-0.6316 -0.4264 -0.4041 -0.3809  2.3897
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
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)
      Null deviance: 664.66  on 1173  degrees of freedom
Residual deviance: 662.14  on 1169  degrees of freedom
AIC: 672.14
Number of Fisher Scoring iterations: 5

```

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))

```

Which is the “best” model? One way to judge this is by using the AIC (Akaike Information Criterion) values. The better model tends to have the lower AIC, which would mean that the best model in our case is the simplest one. We might suspect troubles with the more complicated models because only the intercept terms are significant.

4 Contingency Tables

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

```
bnew = data.frame(preemie,smoke)
bnew = subset(bnew, subset = smoke <2)
tab2 = table(bnew$smoke, bnew$preemie)
chisq.test(tab2)
```

The results are:

```
      0  1
0 485  39
1 419  40

      Pearson's Chi-squared test with Yates' continuity correction
data:  tab2
X-squared = 0.3773, df = 1, p-value = 0.5391
```

The p -value is 0.54, so we cannot reject the null hypothesis that these two variables are independent.

We now consider a contingency table with r rows and k columns. We can ask whether or not having a premature baby is independent of smoking habits, where we include all valid observations (i.e., `smoke = 0,1,2` or `3`).

```
bnew3 = data.frame(preemie,smoke)
tab3 = table(bnew3$smoke, bnew3$preemie)
chisq.test(tab3)
```

The results are:

```
      0  1
0 485  39
1 419  40
2  83   9
3  91   8

      Pearson's Chi-squared test
data:  tab3
X-squared = 0.87, df = 3, p-value = 0.8327
```

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

5 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. For example, E could be the event of having a premature baby for mothers that smoke as compared to mothers that do not smoke.

```
fisher.test(tab2)
```

The result is:

```
      Fisher's Exact Test for Count Data
data:  tab2
p-value = 0.4823
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.7291067 1.9338259
sample estimates:
odds ratio
 1.186987
```

So the odds of having a premie baby are about 1.18 as great for smokers as for nonsmokers. Notice this confidence interval contains 1. The odds is defined as

$$\text{odds of } E_1 = \frac{P\{E_1\}}{1 - P\{E_1\}}.$$

The odds ratio is defined as the odds of E_1 divided by the odds of E_2 .