Introduction to R for Epidemiologists

Jenna Krall, PhD

Thursday, March 5, 2015

Linear regression

Simple linear regression:

$E(y|x) = \beta_0 + \beta_1 x$

where

- $E(y|x)$ is the expectation or mean of y
- β_0 is the intercept, $E(y|x)$ when $x = 0$
- β_1 is the slope, the change in $E(y|x)$ for a one unit change in x

Assumptions

1. Linearity
2. Independence
3. Normality
4. Equal variances

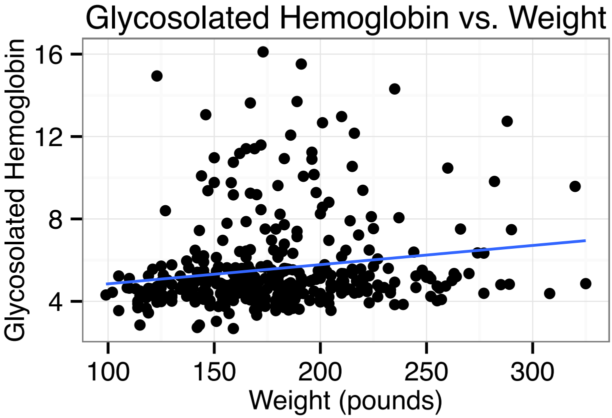
Linear regression

```r
library(dplyr)
diab <- read.csv("diabetes.csv")
diab <- dplyr::select(diab, chol, weight, glyhb, diab1, gender)
diab <- diab[complete.cases(diab), ]
head(diab)
```

```
##   chol weight glyhb diab1 gender
## 1  203    121  4.31     0 female
## 2  165    218  4.44     0 female
## 3  228    256  4.64     0 female
## 4   78    119  4.63     0   male
## 5  249    183  7.72     1   male
## 6  248    190  4.81     0   male
```

Linear regression

```
ggplot(diab, aes(x = weight, y = glyhb)) + geom_point() +
    xlab("Weight (pounds)") + ylab("Glycosolated Hemoglobin") +
    ggtitle("Glycosolated Hemoglobin vs. Weight") +
    theme_bw() + theme(title = element_text(size = 10)) +
    geom_smooth(method = "lm", se = F)
```



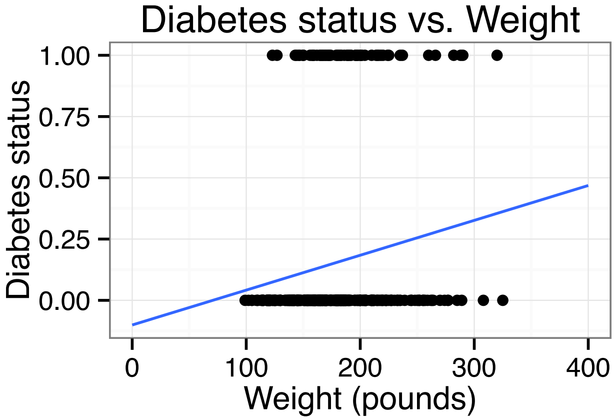
Linear regression

```r
lm1 <- lm(glyhb ~ weight, data = diab)
summary(lm1)$coef
```

```
##                Estimate  Std. Error  t value     Pr(>|t|)
## (Intercept) 3.918552481 0.500749206 7.825379 4.908947e-14
## weight      0.009315444 0.002750545 3.386763 7.799073e-04
```

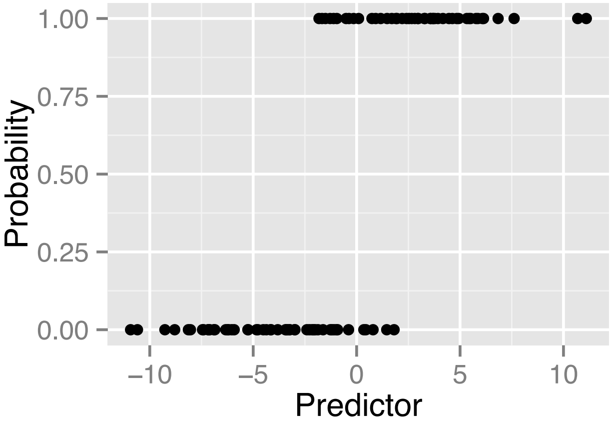
Linear regression

```
ggplot(diab, aes(x = weight, y = diab1)) + geom_point() +
    xlim(c(0, 400)) + xlab("Weight (pounds)") +
    ylab("Diabetes status") + ggtitle("Diabetes status vs. Weight") +
    theme_bw() + geom_smooth(method = "lm", se = F, fullrange = T)
```



Logistic regression

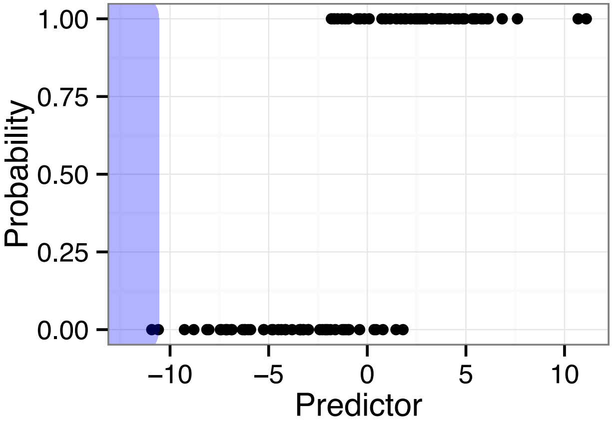
Simulated data



Logistic regression

Simulated data

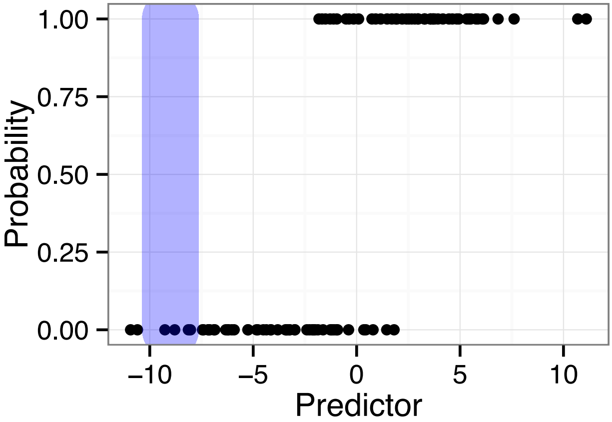
```
size1 <- 10
alpha1 <- 0.3
g1 <- ggplot(dat, aes(x = x, y = y)) + geom_point() +
    theme_bw() + ylab("Probability") + xlab("Predictor")
g1 + geom_line(aes(x = -12), color = "blue", alpha = alpha1, size  = size1)
```



Logistic regression

Simulated data

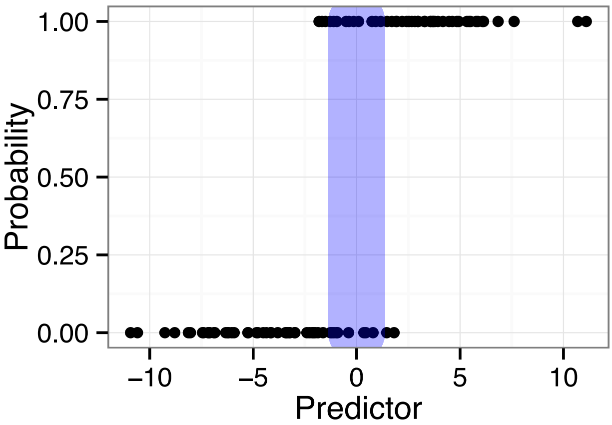
```
g1 + geom_line(aes(x = -9), color = "blue", alpha = alpha1, size  = size1)
```



Logistic regression

Simulated data

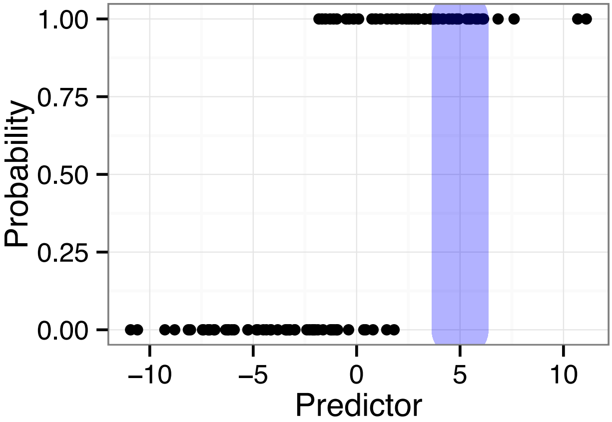
```
g1 + geom_line(aes(x = 0), color = "blue", alpha = alpha1, size  = size1)
```



Logistic regression

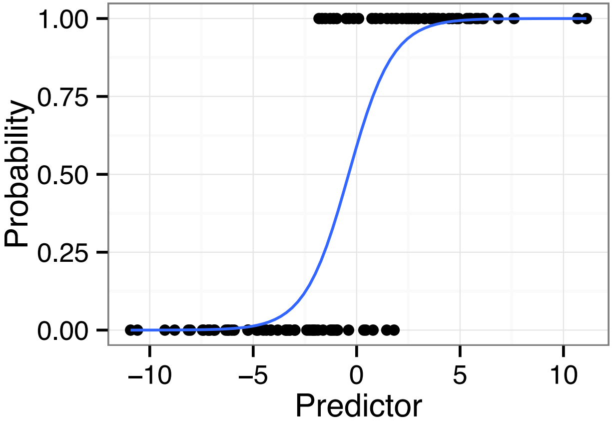
Simulated data

```
g1 + geom_line(aes(x = 5), color = "blue", alpha = alpha1, size  = size1)
```



Logistic regression

Simulated data



Logistic regression

$$\text{logit}(E(y|x)) = \beta_0 + \beta_1 x$$

$$\text{logit}(p) = \beta_0 + \beta_1 x$$

- Recall that when y is binary (1 or 0), $E(y|x)$ is simply the probability (e.g. of diabetes)
- Logit is the (natural) log odds
  - Recall: $odds = p/(1-p)$
  - So $log(odds) = log(p/(1-p))$

Logistic regression: interpretation

- β_0 is the "left hand side" when $x = 0$
  - For linear regression: "left hand side" is mean outcome
  - For logistic regression: "left hand side" is log odds of outcome
- β_1 is the change in "left hand side" for a one unit change in x
  - For linear regression: E($y|x = 1$) - E($y|x = 0$)
  - For logistic regression: logit($p|x = 1$) - logit($p|x = 0$)
  - For logistic regression: difference in log odds of outcome for a unit change in x
  - The log odds ratio is the same as the difference in log odds
  - β_1 still quantifies the association between x and y

Logistic regression

```
glm1 <- glm(diab1 ~ gender, data = diab, family = "binomial")
glm1
```

```
##
## Call:  glm(formula = diab1 ~ gender, family = "binomial", data = diab)
##
## Coefficients:
## (Intercept)   gendermale
##     -1.7415       0.0551
##
## Degrees of Freedom: 387 Total (i.e. Null);  386 Residual
## Null Deviance:       330.8
## Residual Deviance: 330.7     AIC: 334.7
```

```
summary(glm1)$coef
```

```
##                Estimate Std. Error    z value     Pr(>|z|)
## (Intercept) -1.74149763  0.1859204 -9.3668982 7.469540e-21
## gendermale   0.05509868  0.2863107  0.1924437 8.473947e-01
```

Logistic regression interpretation

β_0

$\beta_0 + \beta_1$

β_1

Logistic regression interpretation

- β_0 is the log odds of diabetes for females

```
# Restrict data to females
female <- filter(diab, gender == "female")
# Find proportion of diabetics among females
prop.table(table(female$diab1))
```

```
##
##         0         1
## 0.8508772 0.1491228
```

```
p1 <- prop.table(table(female$diab1))[2]
# Compute odds
fodds <- p1 / (1 - p1)
```

Logistic regression interpretation

- β_0 is the log odds of diabetes for females

```
# Compute log odds
log(fodds)
```

```
##         1
## -1.741498
```

```
# Same as regression!
glm1$coef
```

```
## (Intercept)  gendermale
## -1.74149763  0.05509868
```

Logistic regression interpretation

- $\beta_0 + \beta_1$ is the log odds for males

```
# Restrict data to males
male <- filter(diab, gender == "male")
# Find proportion of diabetics among males
prop.table(table(male$diab1))
```

```
##
##       0       1
## 0.84375 0.15625
```

```
p1 <- prop.table(table(male$diab1))[2]
# Compute odds
modds <- p1 / (1 - p1)
```

Logistic regression interpretation

- $\beta_0 + \beta_1$ is the log odds for males

```
# Compute log odds
log(modds)
```

```
##         1
## -1.686399
```

```
# Same as regression!
glm1$coef[1] + glm1$coef[2]
```

```
## (Intercept)
##   -1.686399
```

Logistic regression interpretation

- β_1 is the difference in log odds of diabetes comparing males to females
- β_1 is the log odds ratio of diabetes comparing males to females

```
# Compute difference in log odds
log(modds) - log(fodds)
```

```
##          1
## 0.05509868
```

- Recall $log(a) - log(b) = log(a/b)$

```
log(modds / fodds)
```

```
##          1
## 0.05509868
```

Logistic regression interpretation

```
# Log odds ratio
log(modds / fodds)
```

```
##          1
## 0.05509868
```

```
# Same as regression!
glm1$coef[2]
```

```
## gendermale
## 0.05509868
```

Logistic regression interpretation

We generally are not interested in log odds or log odds ratios, so exponentiate results from logistic regression:

```
# Odds for females
fodds
```

```
##         1
## 0.1752577
```

```
exp(glm1$coef[1])
```

```
## (Intercept)
##   0.1752577
```

```
# Odds for males
modds
```

```
##         1
## 0.1851852
```

```
exp(glm1$coef[1] + glm1$coef[2])
```

```
## (Intercept)
##   0.1851852
```

Logistic regression interpretation

We generally are not interested in log odds or log odds ratios, so exponentiate results from logistic regression:

```
# Odds ratio comparing males to females
modds / fodds
```

```
##        1
## 1.056645
```

```
exp(glm1$coef[2])
```

```
## gendermale
##   1.056645
```

Confidence intervals

- `confint` in R computes confidence intervals on the LOG scale!
- We need to exponentiate the lower and upper bounds to get the confidence interval for the odds or odds ratio

```
glm1$coef
```

```
## (Intercept)  gendermale
## -1.74149763  0.05509868
```

```
confint(glm1)
```

```
##                  2.5 %     97.5 %
## (Intercept) -2.1226520 -1.3915117
## gendermale  -0.5139659  0.6129859
```

Confidence intervals

```r
exp(glm1$coef)
```

```
## (Intercept)  gendermale
##   0.1752577   1.0566449
```

```r
exp(confint(glm1))
```

```
##                 2.5 %    97.5 %
## (Intercept) 0.1197137 0.2486991
## gendermale  0.5981188 1.8459350
```

Inference

Is gender associated with diabetes?

- Null hypothesis: $\beta_1 = 0$
- Alternative hypothesis: $\beta_1 \neq 0$

```
summary(glm1)$coef
```

```
##                Estimate Std. Error    z value     Pr(>|z|)
## (Intercept) -1.74149763  0.1859204 -9.3668982 7.469540e-21
## gendermale   0.05509868  0.2863107  0.1924437 8.473947e-01
```

```
confint(glm1)
```

```
##                  2.5 %     97.5 %
## (Intercept) -2.1226520 -1.3915117
## gendermale  -0.5139659  0.6129859
```

Since 0 falls within the 95% confidence interval for the log odds ratio, we fail to reject the null hypothesis that the log odds ratio is equal to 0.

Inference

Is gender associated with diabetes?

- Null hypothesis: $OR = exp(\beta_1) = 1$
- Alternative hypothesis: $OR = exp(\beta_1) \neq 1$

```
exp(confint(glm1))
```

```
##                 2.5 %    97.5 %
## (Intercept) 0.1197137 0.2486991
## gendermale  0.5981188 1.8459350
```

Since 1 falls within the 95% confidence interval for the odds ratio, we fail to reject the null hypothesis that the odds ratio is equal to 1.

Multiple logistic regression

```
glm2 <- glm(diab1 ~ gender + chol + weight, data = diab, family = "binomial")
glm2
```

```
##
## Call:  glm(formula = diab1 ~ gender + chol + weight, family = "binomial",
##     data = diab)
##
## Coefficients:
## (Intercept)   gendermale         chol       weight
##   -6.273769     0.007973     0.012363     0.010276
##
## Degrees of Freedom: 387 Total (i.e. Null);  384 Residual
## Null Deviance:       330.8
## Residual Deviance: 305.2     AIC: 313.2
```

Multiple logistic regression

```
summary(glm2)$coef
```

```
##                 Estimate Std. Error     z value     Pr(>|z|)
## (Intercept) -6.273769295 0.998715467 -6.28183853 3.345923e-10
## gendermale   0.007973252 0.298553803  0.02670625 9.786940e-01
## chol         0.012362651 0.003144352  3.93170056 8.434709e-05
## weight       0.010275624 0.003492660  2.94206226 3.260344e-03
```

Multiple logistic regression: interpretation

Text

- β_0
- β_1 (gender)
- β_2 (chol)
- β_3 (weight)

Multiple logistic regression: interpretation

1. What is the odds ratio of diabetes comparing males to females, adjusting for weight and cholesterol?
2. What is the odds ratio of diabetes corresponding to a 1 pound increase in weight, controlling for gender and cholesterol?

Multiple logistic regression: interpretation

```
exp(glm2$coef["gendermale"])
```

```
## gendermale
##   1.008005
```

```
exp(glm2$coef["weight"])
```

```
##   weight
## 1.010329
```

Confidence interval and inference

Is weight associated with diabetes, after controlling for cholesterol and gender?

- Null hypothesis: $\beta_3 = 0$
- Alternative hypothesis: $\beta_3 \neq 0$

```
summary(glm2)$coef
```

```
##                 Estimate  Std. Error     z value     Pr(>|z|)
## (Intercept) -6.273769295 0.998715467 -6.28183853 3.345923e-10
## gendermale   0.007973252 0.298553803  0.02670625 9.786940e-01
## chol         0.012362651 0.003144352  3.93170056 8.434709e-05
## weight       0.010275624 0.003492660  2.94206226 3.260344e-03
```

For a hypothesis test with $\alpha = 0.05$, since the p-value is less than 0.05, we reject the null hypothesis that the log odds ratio is equal to 0.

Confidence interval and inference

Is weight associated with diabetes, after controlling for cholesterol and gender?

- Null hypothesis: $\beta_3 = 0$
- Alternative hypothesis: $\beta_3 \neq 0$

```r
confint(glm2)
```

```
##                    2.5 %      97.5 %
## (Intercept) -8.316662724 -4.38806199
## gendermale  -0.585330243  0.58975335
## chol         0.006321040  0.01870552
## weight       0.003391084  0.01715061
```

Since 0 does not fall within the 95% confidence interval for the log odds ratio, we reject the null hypothesis that the log odds ratio is equal to 0.

Confidence interval and inference

Is weight associated with diabetes, after controlling for cholesterol and gender?

- Null hypothesis: $OR = exp(\beta_3) = 1$
- Alternative hypothesis: $OR = exp(\beta_3) \neq 1$

```
exp(confint(glm2))
```

```
##                     2.5 %     97.5 %
## (Intercept) 0.0002444102 0.01242479
## gendermale  0.5569219117 1.80354351
## chol        1.0063410600 1.01888157
## weight      1.0033968404 1.01729853
```

Since 1 does not fall within the 95% confidence interval for the odds ratio, we reject the null hypothesis that the odds ratio is equal to 1.

Conclusions

- Logistic regression is used for a *binary* outcome
- Coefficients from logistic regression are log odds and log odds ratios
- Results from R need to be exponentiated to report the odds and odds ratios