Linear Regression Review

FW8051 Statistics for Ecologists

Department of Fisheries, Wildlife and Conservation Biology

Objectives

Review important statistical concepts within the context of linear regression using simulated data:

- Sampling Distributions
- T-tests for regression coefficients
- Confidence intervals
- P-values
- How to check assumptions

Why Consider Simulated Data (Kery p. 7):

- Truth is known (can compare estimates to truth).
- Provides a check on coding errors.
- Facilitates understanding of **sampling distributions**.
- Provides a means to study properties of an estimator (mean, variance)
- . . . and, an assessment of the effect of assumption violations.
- Facilitates (and provides a check on) our understanding of statistical models

Why Real Data

- Real data almost never meet all assumptions of common statistical models.

Sustainable trophy hunting of African Lions

Whitman et al. 2014 Nature 428:175-178

Important to know the age of male lions to help manage trophy hunting

- Removing males over 6 has little effect on social structure
- Removing younger males is more disruptive

How, from afar, can we tell the age of a lion?

Its in the Nose!

After about 3 years of age, the fleshy part of the nose begins to freckle or become liver spotted. As the lion ages, more pigmentation appears until the nose is entirely black by about 8 years. A general rule of thumb is that by 6 years noses are >50% black.

Data are contained in `abd` library of Program R:

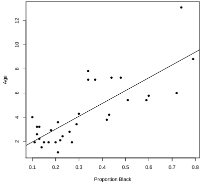
```
install.packages("abd")   # only if not installed (do once)
```

```
library(abd) # Each time you want to accesss the data
data(LionNoses)
head(LionNoses)
```

```
  age proportion.black
1 1.1             0.21
2 1.5             0.14
3 1.9             0.11
4 2.2             0.13
5 2.6             0.12
6 3.2             0.13
```

Lion's Noses

```
with(LionNoses,plot(proportion.black,age,  xlab="Proportion Black",
                    ylab="Age", pch=16))
lm.nose<-lm(age~proportion.black, data=LionNoses)
abline(lm.nose)
```



Linear Regression Assumptions

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

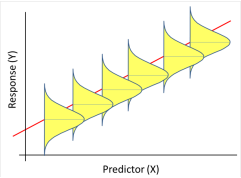
$$\epsilon_i \sim N(0, \sigma^2)$$

Assumptions (HILE Gauss):

- **H**omogeneity of variance (constant scatter about the line); $var(\epsilon_i) = \sigma^2$
- **I**ndependence: Correlation$(\epsilon_i, \epsilon_j) = 0$
- **L**inearity: $E[Y_i \mid X_i] = \beta_0 + X_i\beta_1$
- **E**xistence (we observe random variables that have finite variance; we won't worry about this one)
- **Gauss**: ϵ_i come from a Normal (Gaussian) distribution

Regression Assumptions

- We have a model for the mean $\mu_i = \beta_0 + \beta_1 X_i$
- We specify a probability distribution for $Y_i | X_i \sim N(\mu_i, \sigma^2)$



- How are these assumptions reflected in the figure? How can we evaluate the assumptions with our data?

Residuals Versus Fitted

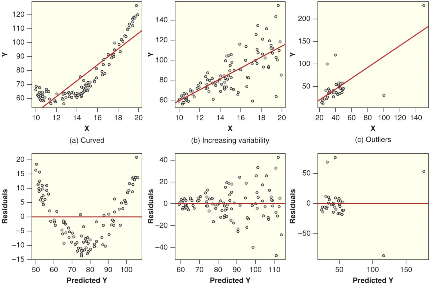
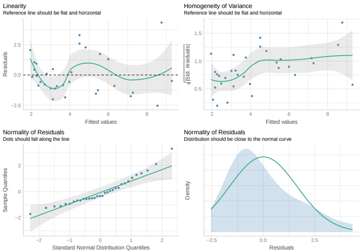


Figure 10.2

Graphical Check

```
performance::check_model(lm.nose,
  check = c("linearity", "homogeneity", "qq", "normality"))
```



Interpretation: Intercept, Slope, t-tests and p-values, Residual Standard Error ($\hat{\sigma}$), R^2

```
summary(lm.nose)
```

```
Call:
lm(formula = age ~ proportion.black, data = LionNoses)

Residuals:
    Min      1Q  Median      3Q     Max
-2.5449 -1.1117 -0.5285  0.9635  4.3421

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)        0.8790     0.5688   1.545    0.133
proportion.black  10.6471     1.5095   7.053 7.68e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.669 on 30 degrees of freedom
Multiple R-squared:  0.6238,    Adjusted R-squared:  0.6113
F-statistic: 49.75 on 1 and 30 DF,  p-value: 7.677e-08
```

Interpretation: Intercept and Slope

$$\widehat{age} = 0.879 + 10.65\text{Proportion.black}$$

Intercept: Estimate of the average age of lions that have no black pigmentation on their noses ($E[Y|X = 0]$).

Slope = Predicted change in age per unit increase in proportion black pigmentation

$$10.65 = \frac{\triangle \text{ age}}{\triangle \text{ Proportion.Black}}$$

But, proportion black < 1 for all lions.

$$? = \frac{\triangle age}{\triangle 0.1 Proportion.Black} = 1.065.$$

$H_0 : \beta_1 = 0$ vs. $H_A : \beta_1 \neq 0$?

```
summary(lm.nose)
```

```
Call:
lm(formula = age ~ proportion.black, data = LionNoses)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.5449 -1.1117 -0.5285  0.9635  4.3421 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)        0.8790     0.5688   1.545    0.133    
proportion.black  10.6471     1.5095   7.053 7.68e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.669 on 30 degrees of freedom
Multiple R-squared:  0.6238,	Adjusted R-squared:  0.6113 
F-statistic: 49.75 on 1 and 30 DF,  p-value: 7.677e-08
```

SE, t-value, p-value

Need to understand the concept of a **Sampling Distribution** of a statistic.

A sampling distribution is the distribution of sample statistics computed for different samples of the same size from the same population.

A sampling distribution shows us how the sample statistic varies from sample to sample

See: AP stats guy videos in Ch. 1 of the textbook

SE $\hat{\beta}$

Think of many repetitions of:

- Collecting a new data set (of the same size)
- Fitting the same regression model
- Calculating $\hat{\beta}$

Sampling distribution of $\hat{\beta}$ is the distribution of $\hat{\beta}$ values across the different samples.

SE = standard deviation of the sampling distribution!

Lets explore through simulation!

Simulation

Lets first generate a single data set consistent with our fitted model using the following code.

```
# Sample size of simulated observations
  n<-32

# Use the observed proportion.black to simulate obs.
  p.black<-LionNoses$proportion.black

# Use the estimated parameters to simulate data.
#   - can get these from the regression output
#   sigma<-summary(lm.nose)$sigma # residual variation about the line
#   betas<-coef(lm.nose) # Regression coefficients
  sigma<-1.67 # residual variation
  betas<-c(0.88, 10.65) #betas

# Create random errors (epsilons) and random responses
 epsilon<-rnorm(n,0, sigma) # Errors
 y<-betas[1] + p.black*betas[2] + epsilon # Response
```

Linear regression using `lm` function

```
lmfit<-lm(y~p.black)
summary(lmfit)
```

```
Call:
lm(formula = y ~ p.black)

Residuals:
    Min      1Q  Median      3Q     Max
-3.2401 -0.8812 -0.3871  0.9053  3.2192

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.7028     0.5627   3.026  0.00505 **
p.black       8.9392     1.4934   5.986 1.45e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.651 on 30 degrees of freedom
Multiple R-squared:  0.5443,    Adjusted R-squared:  0.5291
F-statistic: 35.83 on 1 and 30 DF,  p-value: 1.45e-06
```

Sampling Distribution

Use a for loop to:

- Generate 5000 data sets using the same code
- Fit a linear regression model to each data set
- For each fit, store $\hat{\beta}$

In-class exercise

Sampling distribution of $\hat{\beta}$

Because the sampling distribution of $\hat{\beta}$ depends on σ, we usually work with the distribution of:

$$\frac{\hat{\beta}_1 - \beta_1}{\widehat{SE}(\hat{\beta}_1)} \sim t_{n-2}$$

which does not!

Sampling distribution of the t-statistic

$$\frac{\hat{\beta}_1 - \beta_1}{\widehat{SE}(\hat{\beta}_1)} \sim t_{n-2}$$

Think of many repetitions of:

- Collecting a new data set (of the same size)
- Fitting the regression model
- Calculating: $t = \frac{\hat{\beta}_1 - \beta_1}{\widehat{SE}(\hat{\beta}_1)}$

A histogram of the different t values should be well described by a Student's t-distribution with $n - 2$ degrees of freedom.

Sampling Distribution

Use a for loop to:

- Generate 5000 data sets using the same code
- Fit a linear regression model to each data set
- For each fit, store $\hat{\beta}$ and calculate: $t = \frac{\hat{\beta}_1 - \beta_1}{\widehat{SE}(\hat{\beta}_1)}$

Helpful hints:

- β_1 = true value used to simulate the data, `coef(lm.nose)[2]` = 10.6471
- $\hat{\beta}_1$ is extracted using: `coef(lmfit)[2]`
- $\widehat{SE}(\hat{\beta}_1)$ is extracted using `sqrt(vcov(lm.temp)[2,2])`

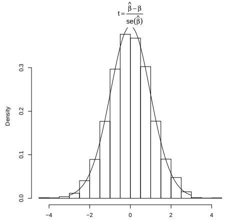
In-class exercise

My code

```
nsims<-5000 # number of simulations
tsamp.dist<-matrix(NA, nsims,1) # matrix to hold results
for(i in 1:5000){
  epsilon<-rnorm(n,0,sigma) # Errors
  y<-betas[1] + betas[2]*p.black + epsilon # Response
  lm.temp<-lm(y~p.black) # lm

# Here is our statistic, calculated for each sample
  tsamp.dist[i]<-(coef(lm.temp)[2]-betas[2])/sqrt(vcov(lm.temp)[2,2])
}
# Plot results
hist(tsamp.dist, xlab="",
     main=expression(t==frac(hat(beta)-beta, se(hat(beta)))), freq=FAL
tvalues<-seq(-3,3, length=1000) # xvalues to evaluate t-distribution
lines(tvalues,dt(tvalues, df=30)) # overlay t-distribution
```

Sampling distribution of t-statistic



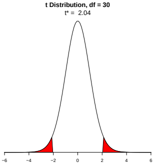
Confidence Interval

A confidence interval for a parameter is an interval computed from sample data by a method that will capture the parameter for a specified proportion of all samples

- The success rate (proportion of all samples whose intervals contain the parameter) is known as the confidence level or coverage rate
- A 95% confidence interval should contain the true parameter for 95% of all samples
- The parameter is *fixed*, but the endpoints of the interval are *random*

Confidence Interval

$$\frac{\hat{\beta}_1 - \beta_1}{\widehat{SE}(\hat{\beta}_1)} \sim t_{n-2}$$



Note:
$t_{0.025,n-2}$ = `qt(p=0.025, df=30)` = -2.04
$t_{0.975,n-2}$ = `qt(p=0.975, df=30)` = 2.04

Confidence Interval

$$P(t_{0.025,n-2} < \frac{\hat{\beta} - \beta}{\widehat{SE}(\hat{\beta})} < t_{0.975,n-2}) = 0.95$$

$$P(-2.04 < \frac{\hat{\beta} - \beta}{\widehat{SE}(\hat{\beta})} < 2.04) = 0.95$$

$$P(-2.04\widehat{SE}(\hat{\beta}) < \hat{\beta} - \beta < 2.04\widehat{SE}(\hat{\beta})) = 0.95$$

$$P(-\hat{\beta} - 2.04\widehat{SE}(\hat{\beta}) < -\beta < -\hat{\beta} + 2.04\widehat{SE}(\hat{\beta})) = 0.95$$

$$P(\hat{\beta} + 2.04\widehat{SE}(\hat{\beta}) > \beta > \hat{\beta} - 2.04\widehat{SE}(\hat{\beta})) = 0.95$$

So, take $(\hat{\beta} - 2.04\widehat{SE}(\hat{\beta}), \hat{\beta} + 2.04\widehat{SE}(\hat{\beta}))$ as the the 95% confidence interval.

`confint` function

```
confint(lm.nose)
```

```
                   2.5 %    97.5 %
(Intercept)     -0.2826733  2.040686
proportion.black  7.5643082 13.729931
```

What is wrong with the following interpretation?

$$P(7.56 \leq \beta \leq 13.72) = 0.95$$

- β is either in this particular interval (P = 1) or it is not (P = 0)
- $P(L \leq \beta \leq U) = 0.95$, where L and U are random variables that determine the upper and lower limits of the 95% confidence interval

We are 95% sure that the true slope (relating proportion of nose that is black and age) falls between 7.56 and 13.73.

Explore CIs through simulation

Simulate another 5000 data sets in R:

Determine 95% confidence limits for each data set and examine whether or not the CI contains the true β.

My Code

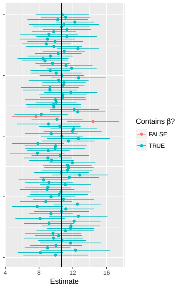
```
nsims<-5000 # number of simulations
Limits<-matrix(NA,nsims,2) # to hold results
beta.hats<-matrix(NA,nsims,1) # to hold estimates
for(i in 1:nsims){
  epsilon<-rnorm(n,0, sigma) # Errors
  y<-betas[1] + betas[2]*p.black + epsilon # Response
  lm.temp<-lm(y~p.black)

# Beta.hat & Confidence limits
  beta.hats[i]<-coef(lm.temp)[2]
  Limits[i,]<-confint(lm.temp)[2,]
}
# True parameter in interval?
I.in<-betas[2]>=Limits[,1] & betas[2] <= Limits[,2]

# Coverage
sum(I.in)/nsims
```

```
[1] 0.9444
```

First 100 simulations



```
summary(lm.nose)
```

```
Call:
lm(formula = age ~ proportion.black, data = LionNoses)

Residuals:
    Min      1Q  Median      3Q     Max
-2.5449 -1.1117 -0.5285  0.9635  4.3421

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)        0.8790     0.5688   1.545    0.133
proportion.black  10.6471     1.5095   7.053 7.68e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.669 on 30 degrees of freedom
Multiple R-squared:  0.6238,    Adjusted R-squared:  0.6113
F-statistic: 49.75 on 1 and 30 DF,  p-value: 7.677e-08
```

P-values

If the null hypothesis, $H_0 : \beta_1 = 0$, is true, then:

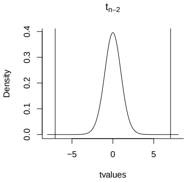
$$t = \frac{\hat{\beta}_1 - 0}{\widehat{SE}(\hat{\beta}_1)} \sim t_{n-2}$$

Is the value we get for $t = \frac{\hat{\beta}_1 - 0}{\widehat{SE}(\hat{\beta}_1)} = 7.053$ consistent with $H_0 : \beta_1 = 0$?

- Overlay $t = \frac{\hat{\beta}_1 - 0}{\widehat{SE}(\hat{\beta}_1)} = 7.053$ on a t_{n-2} distribution
- Determine the probability of getting a t-statistic as or more extreme as the one we observed.

Hypothesis test

- t-distribution tells about the sampling distribution of $t = \frac{\hat{\beta}_1 - 0}{\widehat{SE}(\hat{\beta}_1)} \sim t_{n-2}$ *when the null hypothesis is true*
- our t-statistic falls in the tail of this distribution (so, the Null hypothesis is unlikely to be true!)



R^2

```
summary(lm.nose)
```

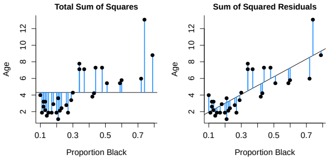
```
Call:
lm(formula = age ~ proportion.black, data = LionNoses)

Residuals:
    Min      1Q  Median      3Q     Max
-2.5449 -1.1117 -0.5285  0.9635  4.3421

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)        0.8790     0.5688   1.545    0.133
proportion.black  10.6471     1.5095   7.053 7.68e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.669 on 30 degrees of freedom
Multiple R-squared:  0.6238,    Adjusted R-squared:  0.6113
F-statistic: 49.75 on 1 and 30 DF,  p-value: 7.677e-08
```

Sum of Squares



SST (Total sum of squares) = $\sum_i^n (Y_i - \bar{Y})^2$

SSE (Sum of Squares Error) = $\sum_i^n (Y_i - \hat{Y})^2$

SSR (sum of squares regression) = SST - SSE = $\sum_i^n (\hat{Y}_i - \bar{Y})^2$

$R^2 = \frac{SST - SSE}{SST} = \frac{SSR}{SST}$ = proportion of the variation explained by the linear model.

Residual Standard Error

```
summary(lm.nose)
```

```
Call:
lm(formula = age ~ proportion.black, data = LionNoses)

Residuals:
    Min      1Q  Median      3Q     Max
-2.5449 -1.1117 -0.5285  0.9635  4.3421

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)        0.8790     0.5688   1.545    0.133
proportion.black  10.6471     1.5095   7.053 7.68e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.669 on 30 degrees of freedom
Multiple R-squared:  0.6238,    Adjusted R-squared:  0.6113
F-statistic: 49.75 on 1 and 30 DF,  p-value: 7.677e-08
```

We expect 95% of the observations to fall within 2*1.669 of the regression line.

Residual standard error = $\hat{\sigma}$

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$
$$\epsilon_i \sim N(0, \sigma^2)$$

- σ describes the amount of variability about the regression line

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-p}} = \sqrt{\frac{SSE}{n-p}} = \sqrt{MSE}$$

- Listed as `Residual Standard Error` in R output from `summary` function
- $n - p$ since we lose one degree of freedom for each parameter we estimate