Simple Linear Regression: Continuous Predictor Example

BIOS 6611

CU Anschutz

Week 9

1. Motivating Example

2. SLR by Hand

3. SLR in `R`

Motivating Example

Motivating Example

An investigator is interested in studying endurance and wants to:

1. Determine if VO_2 max explains endurance during exercise, specifically, the time required to complete a two-mile run.
2. Determine if there is a difference between males and females in physical endurance during exercise.

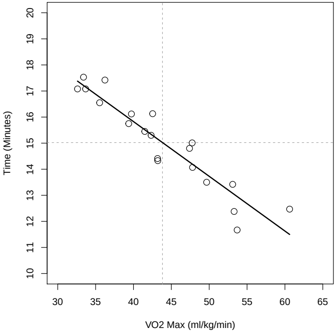
VO_2 max (expressed as ml/kg/min) is the maximum volume of oxygen that can be transported and utilized by the body during exercise.

Twenty subjects (10 males and 10 females) participated in a 16-week exercise study

- 30 minutes of aerobic activity, 5 days a week
- At the end of the study, VO_2 max was measured using a treadmill test and then the time (in minutes) required to complete a two-mile run was recorded

Our VO_2 Max Dataset

ID	Male	VO_2 Max	Minutes
1	0	33.40	17.53
2	0	32.61	17.08
3	0	33.68	17.08
4	0	35.53	16.55
5	0	39.37	15.75
6	0	39.73	16.12
7	0	42.53	16.13
8	0	43.18	14.41
9	0	47.40	14.80
10	0	47.75	15.01
11	1	36.23	17.42
12	1	41.49	15.45
13	1	42.33	15.30
14	1	43.21	14.33
15	1	47.80	14.07
16	1	49.66	13.50
17	1	53.10	13.42
18	1	53.29	12.38
19	1	53.69	11.67
20	1	60.62	12.47



The data can be summarized by:

$n = 20$

$\sum X_i = 876.7$

$\sum Y_i = 300.47$

$S_{XX} = \sum \left(X_i - \bar{X}\right)^2 = 1143.6792$

$S_{YY} = \sum \left(Y_i - \bar{Y}\right)^2 = 57.49045$

$S_{XY} = \sum \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right) = -240.0608$

SLR by Hand

Using Our Summary Formulas

The summary statistics are all we need to estimate our beta coefficients:
$n = 20$, $\sum X_i = 876.7$, $\sum Y_i = 300.47$

$S_{XX} = \sum \left(X_i - \bar{X}\right)^2 = 1143.6792$

$S_{YY} = \sum \left(Y_i - \bar{Y}\right)^2 = 57.49045$

$S_{XY} = \sum \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right) = -240.0608$

With this information we can calculate our slope to be

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{S_{XY}}{S_{XX}} = \frac{-240.0608}{1143.6792} = -0.20990$$

Then, using our estimated slope, we can calculate our intercept to be

$$\hat{\beta}_0 = \frac{\sum Y_i}{n} - \hat{\beta}_1 \frac{\sum X_i}{n} = 15.0234 + (-0.20990)(43.83) = 24.22351$$

Our predicted regression equation is
$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 = 24.22 + (-0.21)X_1 = 24.22 - 0.21X_1$.

ANOVA Table Calculations

We can also calculate the various components of the ANOVA table:

$SS_{Total} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = S_{YY} = 57.49045$

$SS_{Model} = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 = \frac{(S_{XY})^2}{S_{XX}} = 50.3893$

$SS_{Error} = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = SS_{Total} - SS_{Model} = 7.1012$

$MS_{Model} = \frac{SS_{Model}}{1} = 50.3893$

$MS_{Error} = \hat{\sigma}^2_{Y|X} = \frac{SS_{Error}}{n-2} = \frac{\left[S_{YY} - \left(\frac{(S_{XY})^2}{S_{XX}}\right)\right]}{n-2} = 0.39451$

$F = \frac{MS_{Model}}{MS_{Error}} = 127.726$

These are useful to conduct hypothesis tests or calculate different intervals (e.g., confidence or prediction). We will apply these calculations to the brief, but complete, framework for summarizing the result for our slope in the following slides.

Complete Interpretation: Point Estimate

Our point estimate for the slope was

$$\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}} = \frac{-240.0608}{1143.6792} = -0.20990$$

Interpretation: On average, the time required to complete a two-mile run decreases by 0.21 minutes (12.6 seconds) for every 1 ml/kg/min increase in VO_2 max.

Complete Interpretation: Interval Estimate

We first need to calculate the standard error for $\hat{\beta}_1$ to be able to calculate the 95% confidence interval:

$$SE(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2_{Y|X}}{\sum_{i=1}^n (X_i - \bar{X})^2}} = \sqrt{\frac{\hat{\sigma}^2_{Y|X}}{S_{XX}}} = \sqrt{\frac{0.39451}{1143.6792}} = 0.01857$$

For our problem, we will use $t_{20-1-1,0.975} = t_{18,0.975} = 2.1009$ (i.e., `qt(df=18,0.975)` in `R`) to calculate the 95% CI:

$$-0.2099 \pm 2.1009 \times 0.01857 = (-0.249, -0.171)$$

Interpretation: We are 95% confident that the time required to complete a two-mile run decreases by between 0.171 and 0.249 minutes (10.4 to 14.8 seconds) for every 1 ml/kg/min increase in VO_2 max.

Complete Interpretation: Decision

Before we can make a decision about our slope, we need to identify the null or reference value. In practice it can be any value, but most often we are interested in testing $H_0 : \beta_1 = 0$.

For our simple linear regression model, we can evaluate the significance of VO_2 max as a predictor using either a $t-$ or $F-$statistic:

$$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)} = \frac{-0.20990}{0.01857} = -11.30 \sim t_{18} \rightarrow p < 0.0001$$

$$F = \frac{MS_{Model}}{MS_{Error}} = \frac{50.38929}{0.39451} = 127.73 \sim F_{1,18} \rightarrow p < 0.0001$$

Interpretation: A true slope of zero is not consistent with our observed slope since our t-statistic is more extreme than our critical value ($t_{18,0.025} = -2.1009$). We thus reject the null hypothesis and conclude that the slope is not zero.

Additionally, we could note that since $p < 0.05$, we reject the null hypothesis (reject $\beta_1 = 0$) and conclude that the slope is not equal to zero.

Complete Interpretation: Uncertainty

From the previous slide we saw $p < 0.0001$.

Interpretation: If the null hypothesis is true and there is no association between VO_2 max and the time required to complete a 2-mile run (i.e., if $\beta_1 = 0$), then the probability of observing a slope of -0.210 (or something more extreme) is less than 0.0001.

Complete Interpretation: Putting it All Together

There is a significant decrease $\{$**decision**$\}$ in the time required to complete a two-mile run with increasing levels of VO_2 max ($p < 0.0001$) $\{$**uncertainty**$\}$. On average, the time required to complete a two-mile run decreases by 12.6 seconds $\{$**point estimate**$\}$ (95% CI: 10.4 to 14.8 seconds) $\{$**interval estimate**$\}$ for every 1 ml/kg/min increase in VO_2 max.

Amount of Variation Explained by VO$_2$ Max

What is an estimate of the variation in time required to complete a two-mile run for a randomly selected individual?

$$s_Y^2 = \frac{SS_{Total}}{n-1} = \frac{57.490}{19} = 3.026$$

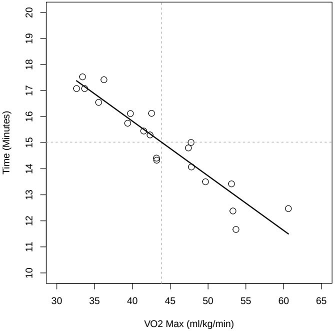
What is an estimate of the variation in the time required to complete a two-mile run for an individual with a given VO$_2$ max?

$$S_{Y|X}^2 = \frac{SS_{Error}}{n-2} = \frac{7.101}{18} = 0.3945$$

So the amount of variation explain by VO$_2$ max is

$$R^2 = \frac{SS_{Model}}{SS_{Total}} = \frac{50.389}{57.490} = 0.8765$$

87.65% of the variation in two-mile run time can be explained by VO$_2$ max.



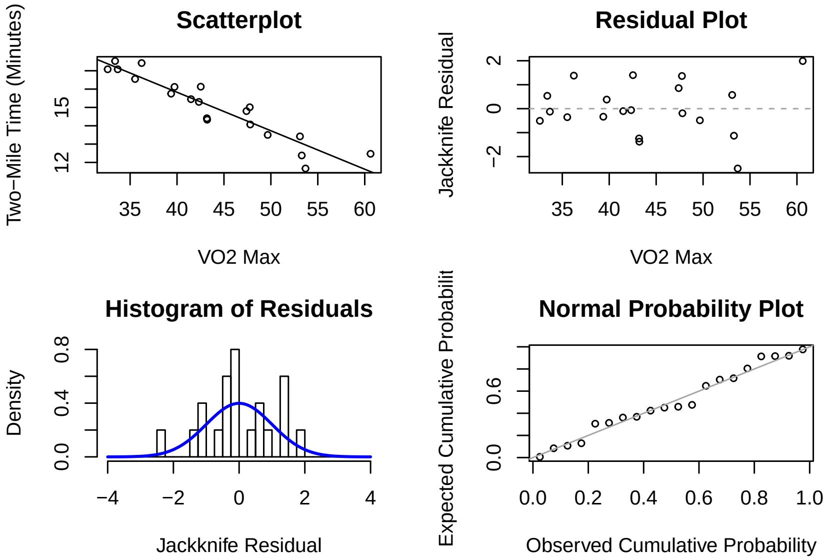
SLR in R

SLR in R

Let's read in our data, fit a simple linear regression model, and generate four diagnostic plots to see if our assumptions appear to be met:

```r
# Code to generate figures
vo2 <- read.csv('vo2max_slr_example.csv')
mod1 <- glm(minutes ~ vo2_max, data=vo2)

par(mfrow=c(2,2), mar=c(4.1,4.1,3.1,2.1))
plot(x=vo2$vo2_max, y=vo2$minutes, xlab='VO2 Max', ylab='Two-Mile Time (Minutes)',
     main='Scatterplot', cex=0.7); abline( mod1 )

plot(x=vo2$vo2_max, y=rstudent(mod1), xlab='VO2 Max', ylab='Jackknife Residual',
     main='Residual Plot', cex=0.7); abline(h=0, lty=2, col='gray65')

hist(rstudent(mod1), xlab='Jackknife Residual',
     main='Histogram of Residuals', freq=F, breaks=seq(-4,4,0.25));
  curve( dnorm(x,mean=0,sd=1), lwd=2, col='blue', add=T)

plot( ppoints(length(rstudent(mod1))), sort(pnorm(rstudent(mod1))),
      xlab='Observed Cumulative Probability',
      ylab='Expected Cumulative Probability',
      main='Normal Probability Plot', cex=0.7);
  abline(a=0,b=1, col='gray65', lwd=1)
```

VO_2 Max Diagnostic Plots



Coefficient Summary Table

```r
summary(mod1)$coefficients
```

```
##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 24.2235142 0.82607068  29.32378 1.199053e-16
## vo2_max     -0.2099022 0.01857275 -11.30162 1.316718e-09
```

We can note from this output that we have the same fitted regression equation as before:

$$\hat{Y} = 24.2235 + (-0.20990)X_1$$

and that $t = -11.30$ with $p < 0.0001$ like we calculated by hand.

Confidence Interval Estimate

What is the average time required to complete a two-mile run for *a large group of individuals* with VO_2 max of 50 ml/kg/min? What is an interval estimate for this average time?

We know we can use our regression equation to predict the average time required:

$$\hat{\mu}_{Y|X_0=50} = 24.2235 + (-0.20990)(50) = 13.73$$

Interpretation: The average time to complete a 2-mile run for a large group of individuals with VO_2 max of 50 ml/kg/min is expected to be 13.73 minutes.

Confidence Interval Estimate

We can also calculate the standard error of this estimate by hand to use in calculating the 95% confidence interval. However, it is extremely easy to use R or SAS to calculate this:

```r
mod1lm <- lm(minutes ~ vo2_max, data=vo2) # need to use lm instead of glm
predict(mod1lm, newdata = data.frame(vo2_max=50), interval='confidence')
```

```
##       fit      lwr      upr
## 1 13.7284 13.34758 14.10923
```

Interpretation: We are 95% confident that the average time to complete a 2-mile run over a large group of individuals with VO$_2$ max of 50 ml/kg/min will be between 13.35 minutes and 14.11 minutes.

What source(s) of variability contribute to our uncertainty in estimating the average 2-mile run time for a given VO$_2$ max?

Error estimating β_0 and β_1 and distance of X_0 from the mean (of X). Note: does not include variability in X.

Prediction Interval Estimate

What is the predicted time required to complete a two-mile run for *an individual* with VO_2 max of 50 ml/kg/min? What is an interval estimate for this predicted time?

We know we can use our regression equation to predict the average time required:

$$\hat{Y}_{50} = 24.2235 + (-0.20990)(50) = 13.73$$

Interpretation: The average time to complete a 2-mile run for an individual with VO_2 max of 50 ml/kg/min is expected to be 13.73 minutes.

Prediction Interval Estimate

We can also calculate the standard error of the predicted time for an individual by hand to use in calculating the 95% prediction interval. However, it is extremely easy to use R or SAS to calculate this:

```
mod1lm <- lm(minutes ~ vo2_max, data=vo2) # need to use lm instead of glm
predict(mod1lm, newdata = data.frame(vo2_max=50), interval='prediction')
```

```
##       fit      lwr      upr
## 1 13.7284 12.35496 15.10184
```

Interpretation: We are 95% confident that an individual with a VO$_2$ max of 50 ml/kg/min (*assuming NO error in measuring VO$_2$ max*) will complete a 2-mile run in between 12.36 minutes and 15.10 minutes.

What source(s) of variability contribute to our uncertainty in estimating the average 2-mile run time for a given VO$_2$ max?

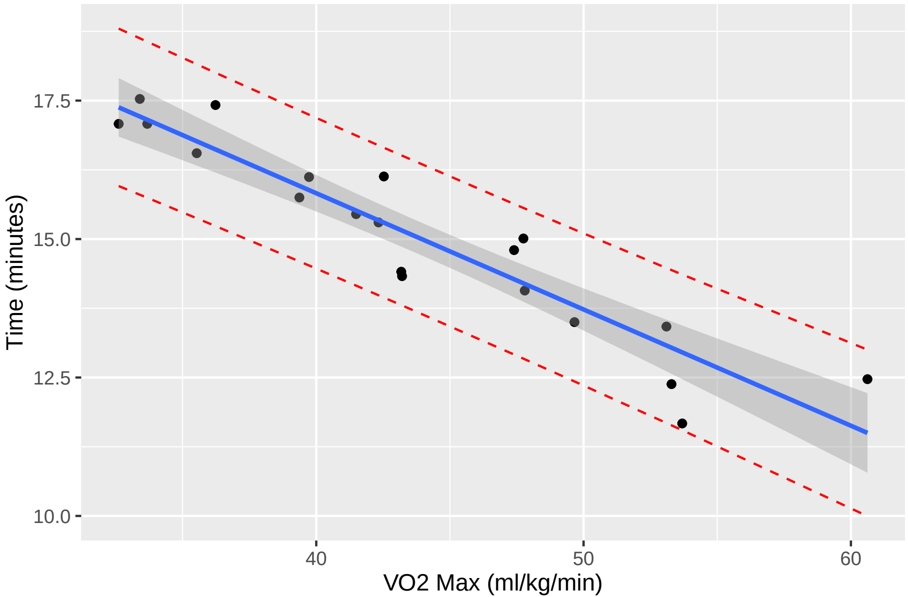
Error estimating β_0 and β_1 and distance of X_0 from the mean (of X), *and* individual variability around the mean of $Y|X$.

Figure of CI and PI

It is also useful to make a scatterplot of our data, with the 95% confidence and prediction intervals. Here we will use ggplot:

```
library(ggplot2)

# we can specify the predicted values at our data frame values without
# providing a new grid of values as well:
minutes_pred <- predict(mod1lm, interval='predict')
vo2fig <- cbind(vo2, minutes_pred)

ggplot(vo2fig, aes(x=vo2_max, y=minutes))+
    geom_point()+
    geom_line(aes(y=lwr), color = "red", linetype = "dashed")+
    geom_line(aes(y=upr), color = "red", linetype = "dashed")+
    geom_smooth(method=lm, se=TRUE) +
    labs(x='VO2 Max (ml/kg/min)', y='Time (minutes)')
```

Figure of CI and PI


Time (minutes)
VO2 Max (ml/kg/min)