Simple Linear Regression: Categorical Predictor Example

BIOS 6611

CU Anschutz

Week 9

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.

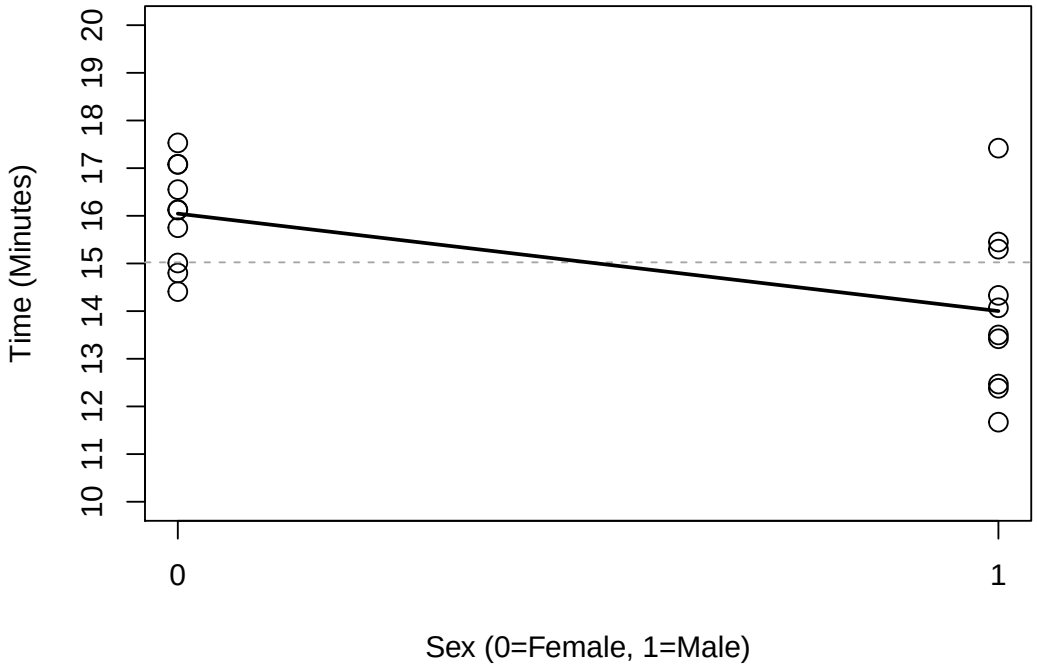
VO_2 max (expressed as ml/kg/min) is the maximum volume of oxygen that can be transported and utilized by the body during exercise.

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



t-test in R

t-test in R

Before learning about simple linear regression, to answer the question *"After 16-weeks of exercise training, do males and females differ in time required to complete a two-mile run?"* we might have conducted a two-sample t-test:

```
vo2 <- read.csv('vo2max_slr_example.csv')
t.test(minutes ~ male, data=vo2, var.equal=TRUE) # assume equal variances
```

```
##
##  Two Sample t-test
##
## data:  minutes by male
## t = 3.2077, df = 18, p-value = 0.004879
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.7055932 3.3844068
## sample estimates:
## mean in group 0 mean in group 1
##          16.046          14.001
```

Since $p < 0.05$, we reject the null hypothesis that the two-mile run time is equal between males and females.

t-test with Unequal Variances

We may instead want to assume unequal variances:

```
t.test(minutes ~ male, data=vo2, var.equal=FALSE) # assume unequal variances
```

```
##
##  Welch Two Sample t-test
##
## data:  minutes by male
## t = 3.2077, df = 14.93, p-value = 0.0059
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.6855794 3.4044206
## sample estimates:
## mean in group 0 mean in group 1
##          16.046          14.001
```

Similar to the previous slide, since $p < 0.05$, we reject the null hypothesis that the two-mile run time is equal between males and females.

t-test Limitation

A t-test only allows us to compare the means of our two groups, which in some cases may be what we are primarily interested in.

Simple linear regression, on the other hand, allows us to characterize the relationship between a dependent variable (e.g., time required to complete a 2-mile run) and an independent variable (e.g., sex) by determining the extent, direction, and strength of the association.

SLR in R

Categorical Explanatory Variables

We can use either continuous or categorical independent variables in a regression analysis.

For categorical variables we can create an indicator (or "dummy") variable that denotes the category for our independent variable X:

Sex	**Smoking Status**
0 = Female	0 = Non-Smoker
1 = Male	1 = Smoker

The "0" category is called the **reference group**. It does not matter which category we choose as the reference, as long as we get the correct interpretation.

Categorical Explanatory Variables

Once we define an indicator variable, we can use it in the regression model like we would any other independent variable. For example, with sex we have:

$$E(\text{minutes}) = \beta_0 + \beta_1 \times \text{sex}$$

Since $X = 1$ for males, we can further write out the conditional expectations:

$$E[Y|X = 0] = E[\text{minutes}|\text{females}] = \beta_0 + \beta_1 \times 0 = \beta_0$$
$$E[Y|X = 1] = E[\text{minutes}|\text{males}] = \beta_0 + \beta_1 \times 1 = \beta_0 + \beta_1$$

The intercept (β_0) in this model represents the mean response for the reference group. The slope (β_1) is the difference in response between the two groups:

$$E[Y|X = 1] - E[Y|X = 0] = (\beta_0 + \beta_1) - (\beta_0) = \beta_1$$

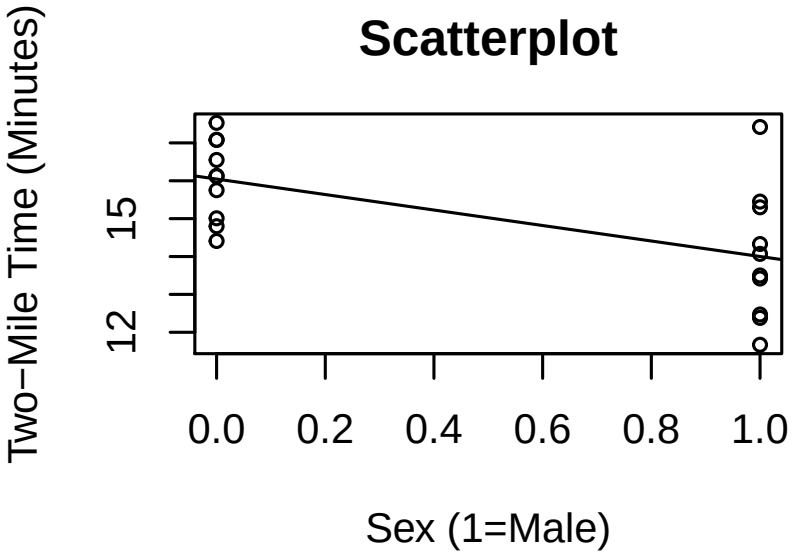
This approach to specifying an indicator variable will have an identical conclusion to the two-sample t-test with equal variances.

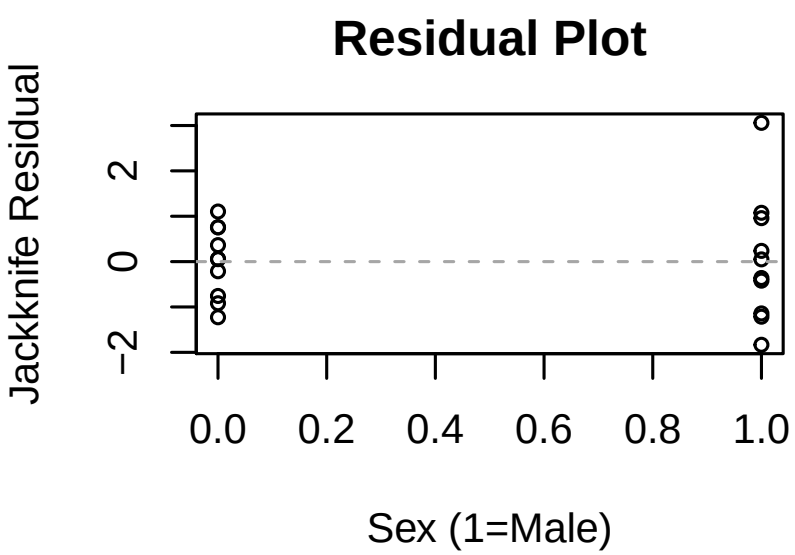
SLR in R

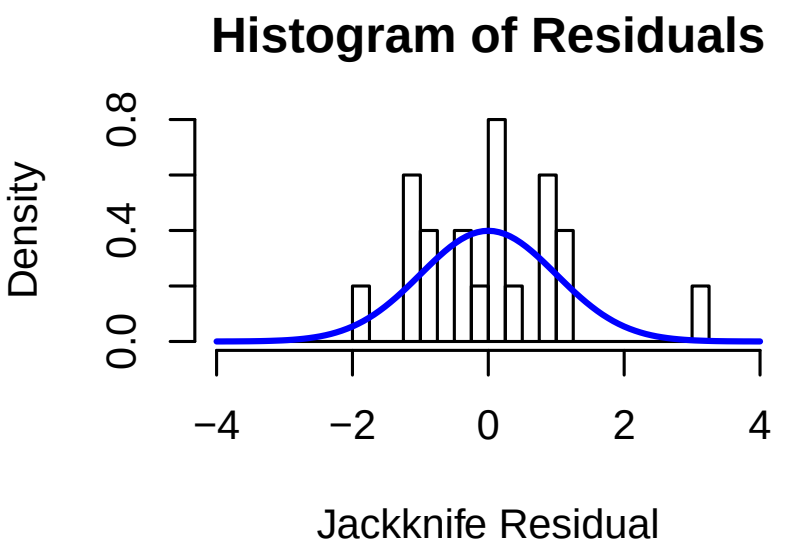
Let's read in our data, fit a simple linear regression model, and generate four diagnostic plots to see if our assumptions appear to be met:

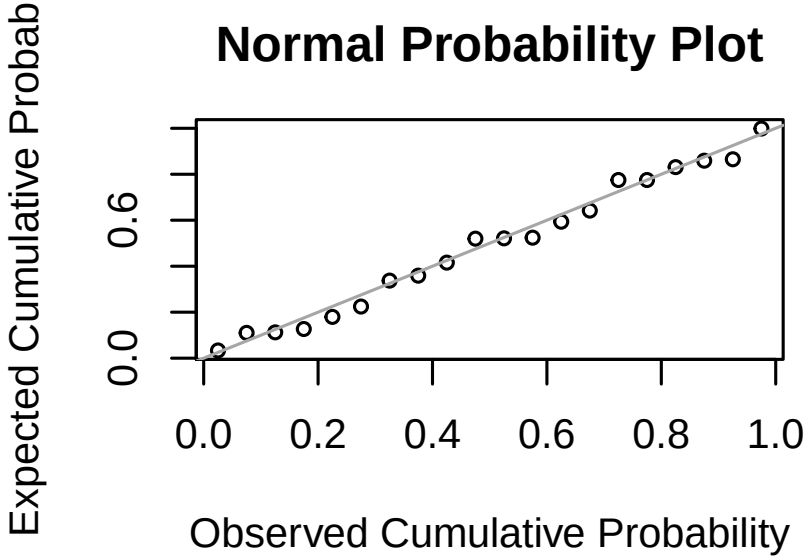
```
# Code to generate figures
mod1 <- glm(minutes ~ male, data=vo2)

par(mfrow=c(2,2), mar=c(4.1,4.1,3.1,2.1))
plot(x=vo2$male, y=vo2$minutes, xlab='Sex (1=Male)', ylab='Two-Mile Time (Minutes)'
     cex=0.7); abline( mod1 )

plot(x=vo2$male, y=rstudent(mod1), xlab='Sex (1=Male)', 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);
  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

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

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)      16.046  0.4508038 35.594200 3.877179e-18
## male             -2.045  0.6375328 -3.207678 4.879420e-03
```

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

$$\hat{Y} = 16.046 + (-2.045) \times \text{male}$$

and that $|t| = 3.2077$ with $p = 0.004879$ matches out earlier t-test output.

Let's walk through the complete interpretation of our slope parameter (i.e., sex effect).

Complete Interpretation: Point Estimate

Our point estimate for the slope was

$$\hat{\beta}_1 = -2.0450$$

Interpretation: On average, the time required to complete a two-mile run is 2.045 minutes shorter for males compared to females.

Complete Interpretation: Interval Estimate

From our output the standard error for $\hat{\beta}_1$ is:

$$SE(\hat{\beta}_1) = 0.6375328$$

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:

$$-2.045 \pm 2.1009 \times 0.6375328 = (-3.384, -0.706)$$

Interpretation: We are 95% confident that the time required to complete a two-mile run is between 0.706 and 3.384 minutes shorter for males compared to females.

We could also calculate this in `R`, but differs slightly because `confint` for `glm` assumes a normal distribution instead of a t-distribution (which `lm` *does* assume):

```
cbind( round(confint(mod1),3), '<-glm() / lm()->', round(confint(lm(minutes~male,data=vo2)),3) )
```

```
##             2.5 %    97.5 %                       2.5 %    97.5 %
## (Intercept) "15.162" "16.93"  "<-glm() / lm()->" "15.099" "16.993"
## male        "-3.295" "-0.795" "<-glm() / lm()->" "-3.384" "-0.706"
```

Complete Interpretation: Decision

Based on our immediate output, we will evaluate the t-statistic and its p-value for the reference that there is no difference between males and females ($H_0 : \beta_1 = 0$):

$$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)} = \frac{-2.045}{0.63753} = -3.21 \sim t_{18} \rightarrow p = 0.0049$$

Interpretation: A true difference of zero is not consistent with our observed difference since our t-statistic is smaller 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 $p = 0.0049$.

Interpretation: If the null hypothesis is true and there is no association between sex and time required to complete a 2-mile run (i.e., if $\beta_1 = 0$), then the probability of observing a difference of 2.045 (or something more extreme) is 0.0049.

Complete Interpretation: Putting it All Together

There is a significant difference {**decision**} in the average time required to complete a two-mile run for males versus females ($p = 0.0049$) {**uncertainty**}. On average, the time required to complete a two-mile run is 2.05 minutes {**point estimate**} (95% CI: 0.71 to 3.38 minutes) {**interval estimate**} shorter for males compared to females.

Amount of Variation Explained by Sex

```
summary(lm(minutes ~ male, data=vo2))
```

```
##
## Call:
## lm(formula = minutes ~ male, data = vo2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.3310 -1.0885  0.0715  1.0340  3.4190
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  16.0460     0.4508  35.594  < 2e-16 ***
## male         -2.0450     0.6375  -3.208  0.00488 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.426 on 18 degrees of freedom
## Multiple R-squared:  0.3637, Adjusted R-squared:  0.3284
## F-statistic: 10.29 on 1 and 18 DF,  p-value: 0.004879
```

36.37% of the variation in two-mile run time can be explained by sex.

What if I reverse my coding for my indicator variable?

If we reverse our coding for our indicator variable, our estimates change with respect to what the reference group is:

```
vo2$female <- abs(vo2$male-1) #define indicator variable for female=1
summary(glm(minutes ~ female, data=vo2))$coefficients
```

```
##             Estimate Std. Error   t value    Pr(>|t|)
## (Intercept)   14.001  0.4508038 31.057858 4.34495e-17
## female         2.045  0.6375328  3.207678 4.87942e-03
```

In this model, β_0 is the average two-mile time for *males*, whereas β_1 is the difference for females compared to males.

Further, β_1 is the same (absolute) value with either reference, but with a different direction in the sign (i.e., $+$ becomes - or - becomes $+$).

Noting our reference category is therefore important for the conclusions we draw, otherwise we might interpret the opposite of what the association is.