18. Simple Linear Regression Example(s)

Readings: Kleinbaum, Kupper, Nizam, and Rosenberg (KKNR): Chs. 1-5, 7, 8

SAS: PROC REG

Homework: Homework 7 due by 11:59 pm on October 31
Homework 8 due by 11:59 pm on November 7
Bonus Matrix Algebra Homework due by 11:59 pm on November 14

Overview

A) Re/Preview of Topics and Motivating Example
B) Complete Interpretation and Decision Making
C) SLR Example with Continuous Predictor
D) SLR Example with Categorical Predictor
E) SLR with >1 Predictor?! (Cue Multiple Linear Regression!)

A. Review Lecture 15-17/ Current Lecture 18/ Preview Lecture 19

Lectures 15-17:

- Fit a line to data: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ and derived $\hat{\beta}_0, \hat{\beta}_1, Var(\hat{\beta}_0), Var(\hat{\beta}_1)$ by minimizing $SS_{Error} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$

- Partitioning out the Regression and Residual Components: $SS_{Total} = SS_{Model} + SS_{Error}$

- ANOVA vs parameter table for simple linear regression

- R^2: Proportion of the variance of *Y* that can be explained by the variable *X*

- Prediction Intervals (larger variance) and Confidence Intervals (smaller variance)

- Diagnostics and evaluating regression model assumptions

Lecture 18:

- Simple Linear Regression Examples (one covariate)

- Motivation for Multiple Linear Regression (more than one covariate)

Lecture 19:

- Multiple Linear Regression Example (more than one covariate)

Motivating Example: VO_2 max

An investigator is interested in studying

1) The difference between males and females in physical endurance during exercise, specifically, time required to complete a two-mile run
2) Determining whether a difference in VO_2 max explains the sex difference in endurance.

NOTE: VO_2 max is

- Maximum volume of oxygen that can be transported and utilized by the body during exercise
- Measure of the capacity of the body to generate the energy required for endurance activities
- An important factor in determining ability to exercise for longer than four to five minutes

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 (expressed as ml/kg/min) was measured using a treadmill test and then the time (in minutes) required to complete a two-mile run was recorded

Potential Scientific Questions:

Model 1. What is the relationship between VO_2 max and time to complete a two-mile run?
Model 2. What is the relationship between gender and time to complete a two-mile run?
Model 3. What is the relationship between VO_2 max and time to complete a two-mile run after adjusting for differences in sex?

SAS Code to Create VO_2 Max Data Example:

```
PROC FORMAT;
 VALUE sex 0 = "Female" 1 = "Male";
RUN;

DATA vo2max;
INPUT sex vo2max minutes;
/* Permanently assign format */
FORMAT sex sex.;

LABEL sex = "Male"
      vo2max = "VO2 Max (ml/kg/min)"
      minutes = "2-mile Run Time (min)";

DATALINES;
0     33.40      17.53
0     32.61      17.08
0     33.68      17.08
0     35.53      16.55
0     39.37      15.75
0     39.73      16.12
0     42.53      16.13
0     43.18      14.41
0     47.40      14.80
0     47.75      15.01
1     36.23      17.42
1     41.49      15.45
1     42.33      15.30
1     43.21      14.33
1     47.80      14.07
1     49.66      13.50
1     53.10      13.42
1     53.29      12.38
1     53.69      11.67
1     60.62      12.47
.     50.0       .     /* NOTE: This 21st pt added for example later to calculate CIs and PIs */
;
```

B. A Complete Interpretation and Decision Making Framework

A complete interpretation of study results includes:

1) A point estimate (observed magnitude of effect)
2) An interval estimate (range of true values that are consistent with the experimental results)
3) A decision (fail to reject/ reject the null hypothesis)
4) A measure of uncertainty in the decision (p-value)

Three steps in statistical decision-making:

1) Identify appropriate reference value(s) for β_1 (usually $\beta_1=0$).
2) Calculate the likelihood of the observed slope under the assumption the reference value is true.
3) If the observed slope is too unusual, then conclude that the reference value is not true (i.e., "reject" the value).

Decision-making with confidence intervals:

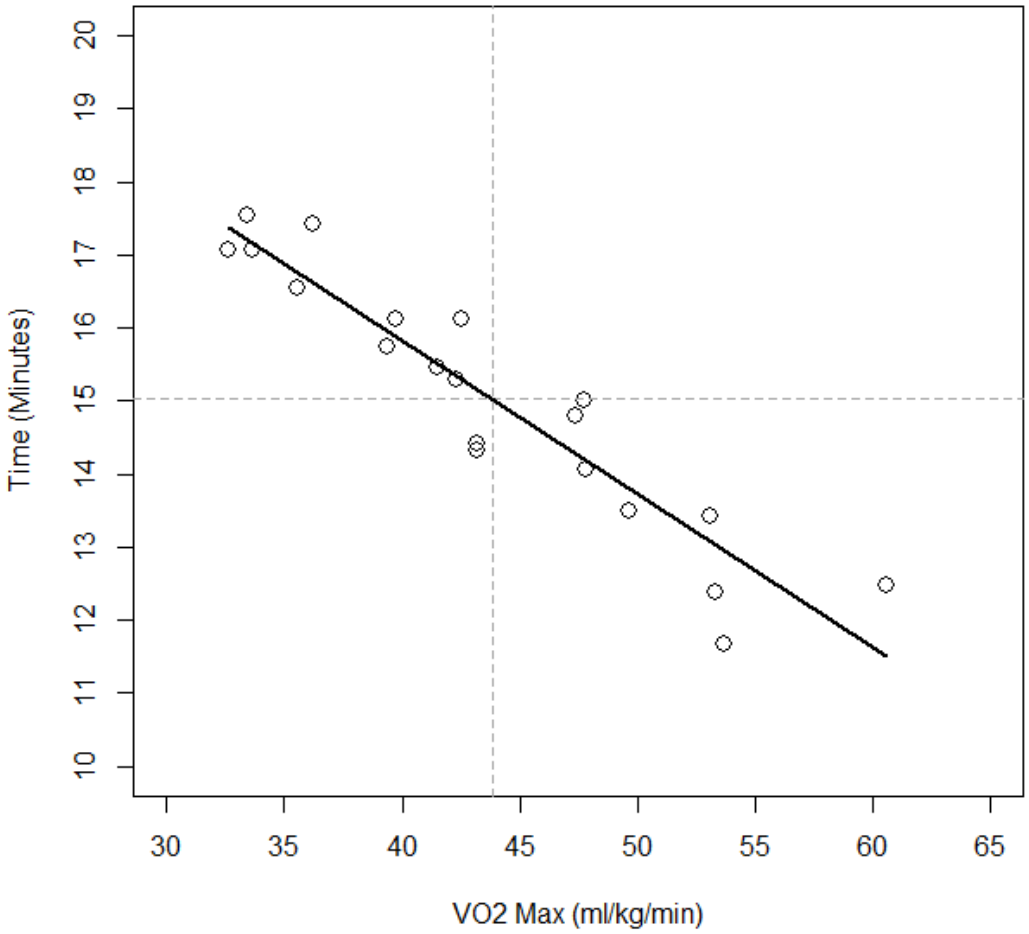
- Values of β_1 outside the 95% confidence interval are not consistent with the observed slope.
- Using this property:
  - Conclude that a reference β_1 is NOT consistent with the observed slope if it is outside of the 95% confidence interval.
  - Conclude that a reference β_1 is consistent with the observed slope if it is inside the 95% confidence interval.

C. Model 1: VO_2 max (continuous predictor)

Scientific Question: What is the relationship between VO_2 max (covariate) and time to complete a two-mile run (outcome)?

VO_2 max data:

		(X)	(Y)
ID	*Sex (0=F;1=M)*	*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
Average		43.8300	15.0235



The data can be summarized by the following:

$\sum X_i = 876.7$ $S_{XX} = \sum(X_i - \bar{X})^2 = 1143.6792$

$\sum Y_i = 300.47$ $S_{YY} = \sum(Y_i - \bar{Y})^2 = 57.49045$

$n = 20$ $S_{XY} = \sum(X_i - \bar{X})(Y_i - \bar{Y}) = -240.0608$

Model 1: VO_2 max

These summary statistics are all we need for SLR calculations. For example:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sum_{i=1}^{n}(X_i-\bar{X})^2} = \frac{S_{XY}}{S_{XX}} = -0.20990 \qquad \hat{\beta}_0 = \frac{\sum Y_i}{n} - \hat{\beta}_1 \frac{\sum X_i}{n} = 24.22351$$

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

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

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

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

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

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

And recall: $\hat{\sigma}^2_{Y|X} = MS_{Error}$

$$\begin{aligned}
SS_{Model} &= \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 \\
&= \sum_{i=1}^{n}(\hat{\beta}_0 + \hat{\beta}_1 X_i - \bar{Y})^2 \\
&= \sum_{i=1}^{n}(\bar{Y} - \hat{\beta}_1\bar{X} + \hat{\beta}_1 X_i - \bar{Y})^2 \\
&= \sum_{i=1}^{n}(\hat{\beta}_1(-\bar{X} + X_i))^2 \\
&= \hat{\beta}_1{}^2 \sum_{i=1}^{n}(X_i - \bar{X})^2) \\
&= \left(\frac{S_{XY}}{S_{XX}}\right)^2 S_{XX} = \frac{(S_{XY})^2}{S_{XX}}
\end{aligned}$$

A Complete Interpretation: (1) Point Estimate and (2) Interval Estimate

Point estimate:

$$\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}} = \frac{-240.0608}{1143.67920} = -0.20990$$

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.

Interval estimate:

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

Note: $t_{20\text{-}1\text{-}1,0.975} = t_{18,0.975} = 2.1009$

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.

A Complete Interpretation: (3) Decision

Reference value for β_1?

- Can be any hypothesized value.
- We are often most interested in testing H_0: $\beta_1 = 0$

Decision: Is our result consistent with this reference value?

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

Note: 2-sided p-value (almost always use 2-sided p-values)

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

Note: $t^2 = F$
$(-11.30)^2 = 127.73$

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 less than zero.

A Complete Interpretation: (4) Uncertainty

Uncertainty in the decision:

$p < 0.0001$ (from previous slide)

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.

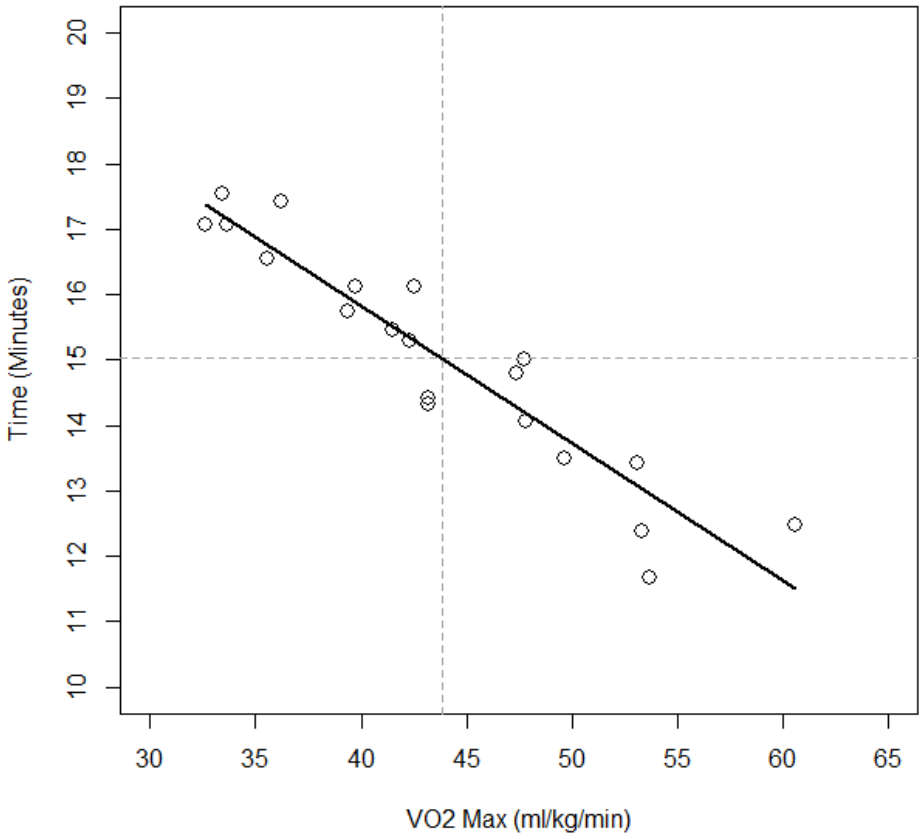
Summary (All 4 components of a complete interpretation):

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.

Predicted Values and Partitioning Variability

VO_2 max data (with predicted values):

ID	*VO_2 Max (X)*	*Minutes (Y)*	*Minutes ($\hat{Y}$)*
1	33.40	17.53	17.21
2	32.61	17.08	17.38
3	33.68	17.08	17.16
4	35.53	16.55	16.77
5	39.37	15.75	15.96
6	39.73	16.12	15.88
7	42.53	16.13	15.30
8	43.18	14.41	15.16
9	47.40	14.80	14.27
10	47.75	15.01	14.20
11	36.23	17.42	16.62
12	41.49	15.45	15.51
13	42.33	15.30	15.34
14	43.21	14.33	15.15
15	47.80	14.07	14.19
16	49.66	13.50	13.80
17	53.10	13.42	13.08
18	53.29	12.38	13.04
19	53.69	11.67	12.95
20	60.62	12.47	11.50
Average	43.8300	15.0232	15.0232



$$SS_{Total} = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 = S_{YY} = 57.490$$

$$SS_{Error} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = 7.101$$

$$SS_{Model} = SS_{Total} - SS_{Error} = 50.389$$

VO_2 max data (variance explained):

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

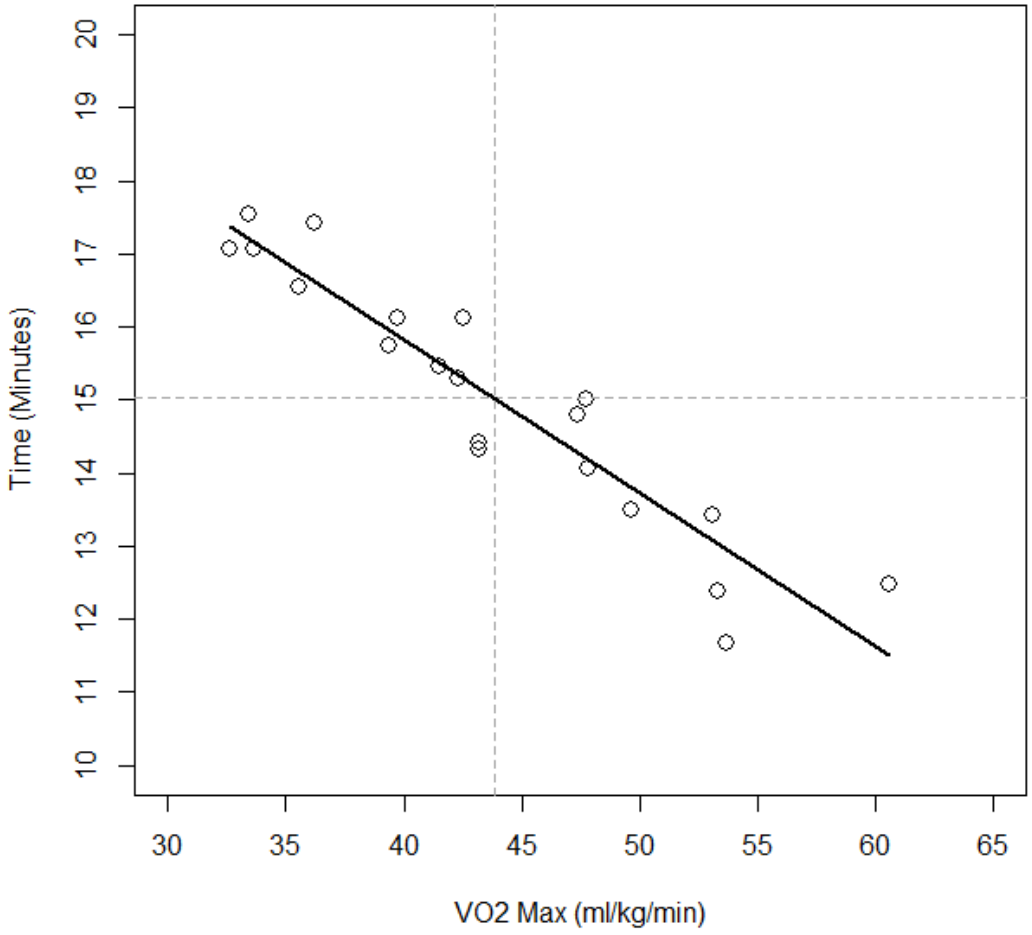
$$s_Y^2 = \frac{SS_{Total}}{n-1} = \frac{57.490}{19} = 3.026$$

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?

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

What proportion of the variation in the time required to complete a two-mile run can be explained by VO_2 max?

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



Confidence and Prediction Intervals

Scientific Question (1): 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?

Scientific Question (2): 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?

Application of Linear Regression: You seek a quantitative formula/equation to predict the dependent variable *Y* (required time to complete a 2-mile run) as a function of the independent variable (VO_2 max).

Prediction Equation:

$\hat{Y} = 24.22351 - 0.2099X$

$\hat{Y}_{50} = 24.22351 - 0.2099(50) = 13.73$

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. The predicted time to complete a 2-mile run for an individual with a VO_2 max of 50 ml/kg/min is also 13.73 minutes.

Confidence Interval (for VO_2 max = 50 ml/kg/min)**:**

$$SE\left(\hat{\mu}_{Y|X_0=50}\right) = \sqrt{\frac{\hat{\sigma}^2_{Y|X}}{n} + \frac{\hat{\sigma}^2_{Y|X}}{n-1}\left(\frac{(X_0 - \bar{X})^2}{\hat{\sigma}^2_X}\right)} = \sqrt{\frac{0.3945}{20} + \frac{0.3945}{19}\left(\frac{(50-43.83)^2}{\frac{1143.6792}{19}}\right)} = 0.181$$

Note: $\hat{\sigma}^2_X = \frac{S_{XX}}{n-1}$.

95% CI: $\hat{\mu}_{Y|X_0=50} \pm t_{18,0.975} SE\left(\hat{\mu}_{Y|X_0=50}\right) = 13.73 \pm 2.1009 \times 0.181 = (13.35, 14.11)$

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 (X). Note: does not include variability in X.

Prediction Interval (for VO_2 max = 50 ml/kg/min)***:***

$$SE\left(\hat{Y}\middle|X_0 = 50\right) = \sqrt{\hat{\sigma}^2_{Y|X} + \frac{\hat{\sigma}^2_{Y|X}}{n} + \frac{\hat{\sigma}^2_{Y|X}}{n-1}\left(\frac{(X_0 - \bar{X})^2}{\hat{\sigma}^2_X}\right)} = \sqrt{0.3945 + 0.03285646} = 0.654$$

$$95\%\ \text{PI: } \left(\hat{Y}\middle|X_0 = 50\right) \pm t_{18,0.975}\, SE\left(\hat{Y}\middle|X_0 = 50\right) = 13.73 \pm 2.1009 \times 0.654 = (12.36, 15.10)$$

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 predicting an individual's 2-mile run time for a given VO_2 max?

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

Example using SAS (Model 1)

$E[\text{Minutes}_i] = \beta_0 + \beta_1 \text{VO}_2\text{max}_i$

```
PROC REG data=vo2max;
    MODEL minutes = vo2max /clb cli clm;
RUN;
```

SAS Options:
clb: Confidence Limits for β
cli: Prediction Interval
clm: Confidence interval

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	50.38929	50.38929	127.73	<.0001
Error	18	7.10116	0.39451		
Corrected Total	19	57.49045			

Root MSE	0.62810	**R-Square**	0.8765
Dependent Mean	15.02350	**Adj R-Sq**	0.8696
Coeff Var	4.18078		

Parameter Estimates

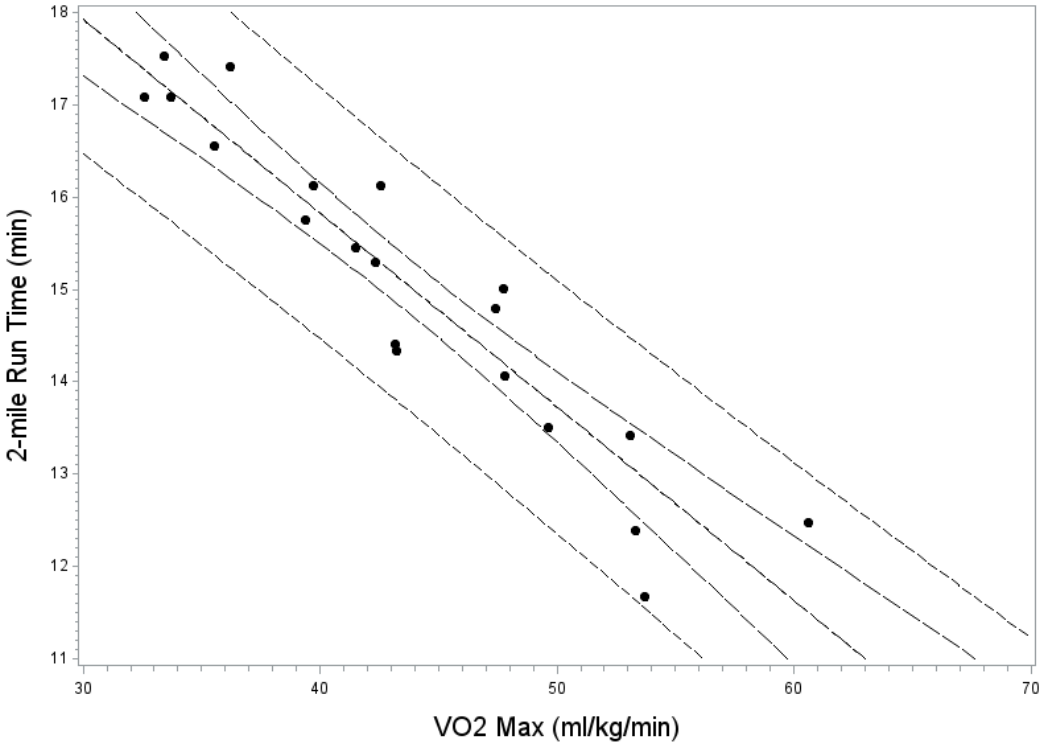
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	Intercept	**1**	24.22351	0.82607	29.32	<.0001	22.48800	25.95902
vo2max	VO2 Max (ml/kg/min)	**1**	-0.20990	0.01857	-11.30	<.0001	-0.24892	-0.17088

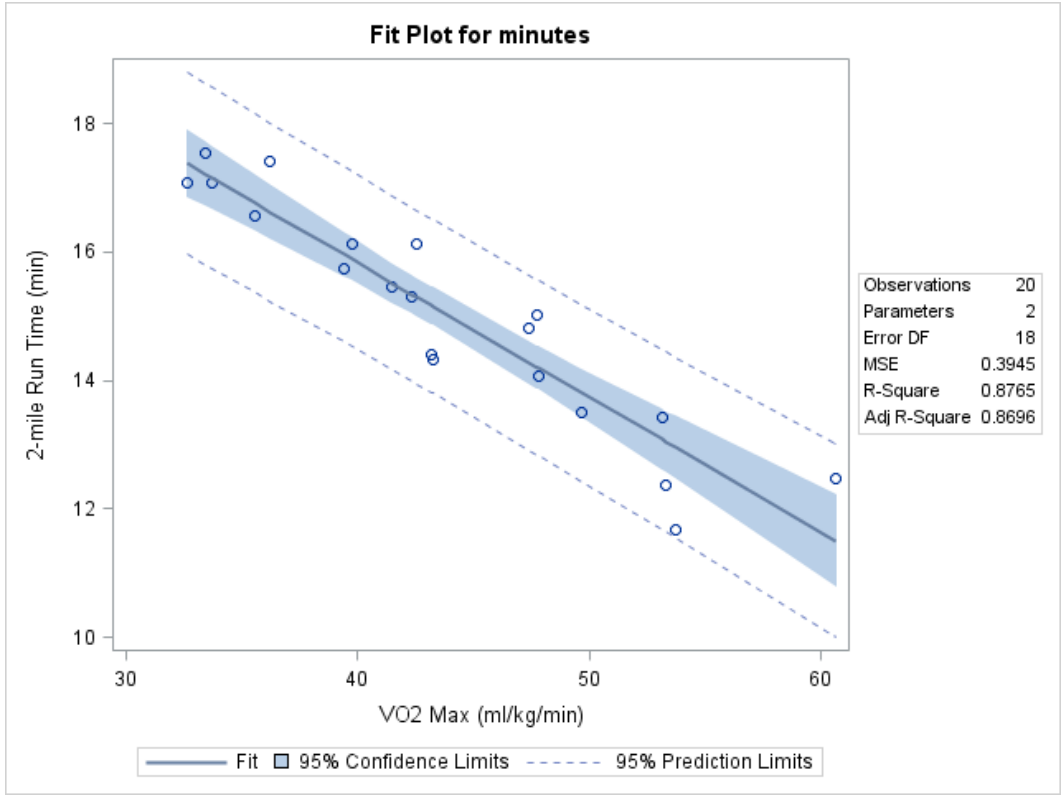
Output Statistics								
Obs	Dependent Variable (y_i)	Predicted Value $(\hat{y}_i)$	Std Error Mean Predict	95% CL Mean		95% CL Predict		Residual
1	17.5	17.2128	0.2393	16.7101	17.7155	15.8007	18.6249	0.3172
2	17.1	17.3786	0.2513	16.8506	17.9066	15.9573	18.7999	-0.2986
3	17.1	17.1540	0.2351	16.6601	17.6479	15.7450	18.5630	-0.0740
4	16.6	16.7657	0.2085	16.3276	17.2038	15.3753	18.1561	-0.2157
5	15.8	15.9597	0.1631	15.6171	16.3022	14.5963	17.3230	-0.2097
6	16.1	15.8841	0.1598	15.5485	16.2197	14.5225	17.2457	0.2359
7	16.1	15.2964	0.1425	14.9970	15.5958	13.9432	16.6495	0.8336
8	14.4	15.1599	0.1410	14.8638	15.4561	13.8075	16.5123	-0.7499
9	14.8	14.2741	0.1553	13.9479	14.6004	12.9148	15.6335	0.5259
10	15.0	14.2007	0.1582	13.8683	14.5330	12.8399	15.5615	0.8093
11	17.4	16.6188	0.1991	16.2004	17.0371	15.2344	18.0031	0.8012
12	15.5	15.5147	0.1470	15.2058	15.8235	14.1594	16.8699	-0.0647
13	15.3	15.3384	0.1432	15.0375	15.6392	13.9849	16.6918	-0.0384
14	14.3	15.1536	0.1409	14.8576	15.4497	13.8012	16.5060	-0.8236
15	14.1	14.1902	0.1586	13.8569	14.5234	12.8292	15.5512	-0.1202
16	13.5	13.7998	0.1773	13.4272	14.1723	12.4286	15.1709	-0.2998
17	13.4	13.0777	0.2222	12.6109	13.5445	11.6780	14.4774	0.3423
18	12.4	13.0378	0.2249	12.5653	13.5104	11.6362	14.4395	-0.6578
19	11.7	12.9539	0.2308	12.4690	13.4387	11.5480	14.3597	-1.2839
20	12.5	11.4992	0.3420	10.7807	12.2178	9.9967	13.0018	0.9708
21	.	13.7284	0.1813	13.3476	14.1092	12.3550	15.1018	.

Graphics using SAS

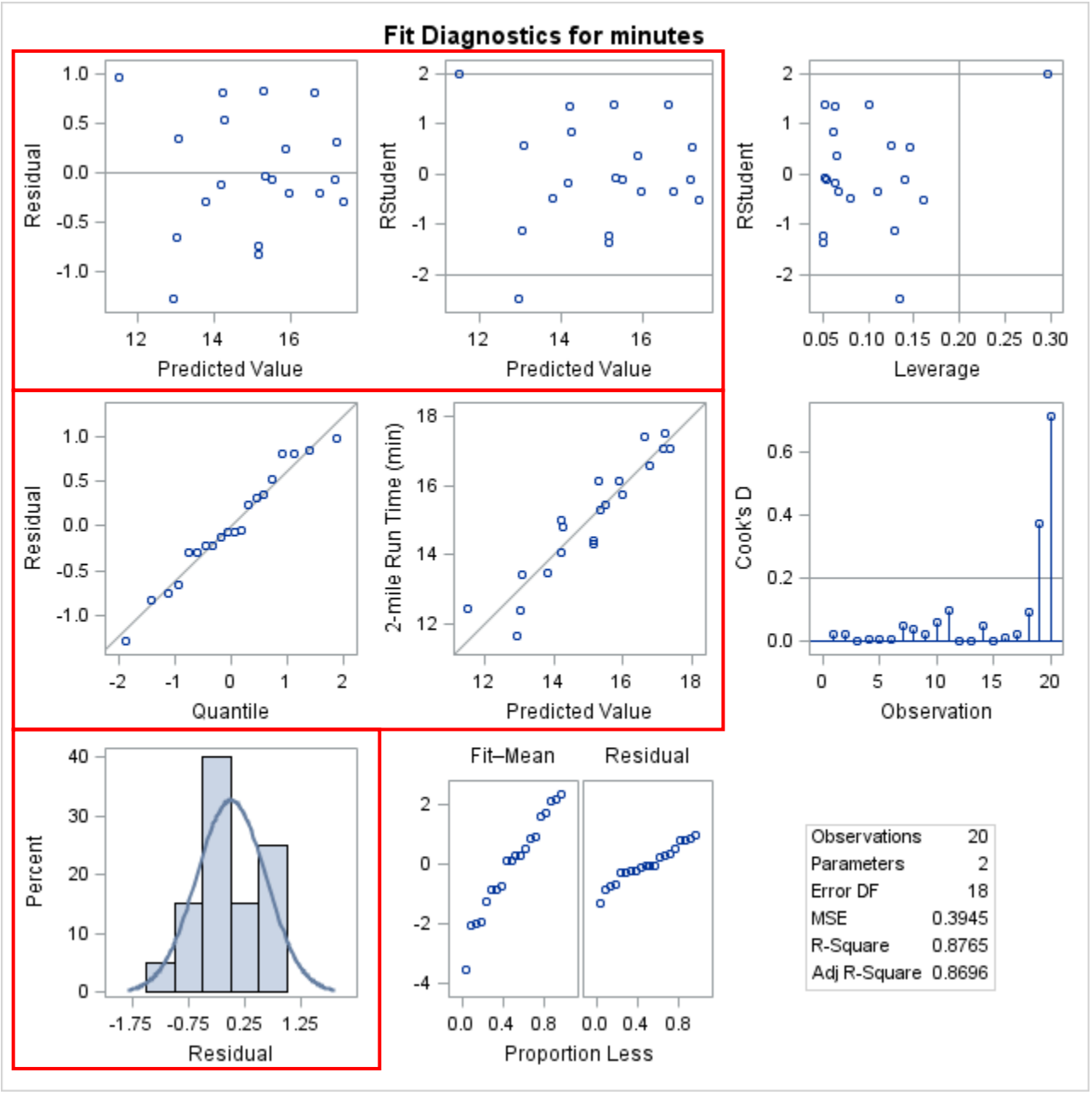
```
PROC GPLOT;
     PLOT minutes*vo2max minutes*vo2max /overlay VAXIS=axis1
     HAXIS=axis2;
     SYMBOL1 INTERPOL=rlcli COLOR=black VALUE=dot LINE=2;
     SYMBOL2 INTERPOL=rlclm COLOR=black VALUE=dot LINE=3;
     AXIS1 LABEL = (FONT=ARIAL HEIGHT= 1.8 ANGLE=90);
     AXIS2 LABEL = (FONT=ARIAL HEIGHT= 1.8);
RUN;
```

```
PROC REG data=vo2max;
     MODEL minutes = vo2max /clb cli
     clm;
RUN;
```





Automatically generated regression diagnostic figures:



Diagnostics Evaluation:

1. Scatterplot of residuals and predicted values are centered around 0 and have pretty equal spread.
2. Scatterplot of studentized residuals suggest similar spread (and can be used to indicate outliers, more on this later).
3. Q-Q plot follows the 45-degree reference line.
4. The predicted vs. dependent variable plot shows pretty similar variability across the predicted values and form an even band around the line.
6. Our data set is small, so the histogram is somewhat challenging to evaluate. Based on the other plots, it seems we may be okay.

D. Model 2: Categorical Explanatory Variable

$E(\text{Minutes}_i) = \beta_0 + \beta_1 \text{Sex}_i$, where Sex=0 for females

Scientific Question: After 16-weeks of exercise training, do males and females differ in time required to complete a two-mile run?

Application of Linear Regression*:* You wish to *characterize the relationship* between the dependent variable (time required to complete a 2-mile run) and the independent variable (sex) by determining the extent, direction, and strength of the association.

VO_2 max data (females):

		(X)	(Y)
ID	*Sex (0=F;1=M)*	*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
Average		39.518	16.046

VO_2 max data (males):

		(X)	(Y)
ID	*Sex (0=F;1=M)*	*VO_2 Max*	*Minutes*
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
Average		48.142	14.001

SAS Example: 2-sample t test

```
PROC TTEST data=vo2max;
     CLASS sex;
     VAR minutes;
RUN;
```

sex	N	Mean	Std Dev	Std Err	Minimum	Maximum
Female	10	16.0460	1.0539	0.3333	14.4100	17.5300
Male	10	14.0010	1.7186	0.5435	11.6700	17.4200
Diff (1-2)		2.0450	1.4256	0.6375		

sex	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
Female		16.0460	15.2921	16.7999	1.0539	0.7249	1.9241
Male		14.0010	12.7716	15.2304	1.7186	1.1821	3.1376
Diff (1-2)	**Pooled**	2.0450	0.7056	3.3844	1.4256	1.0772	2.1082
Diff (1-2)	**Satterthwaite**	2.0450	0.6856	3.4044			

Method	Variances	DF	t Value	Pr > |t|	
Pooled	Equal	18	3.21	0.0049	*This is what's reflected in our parameter estimates Pg. 22.*
Satterthwaite	Unequal	14.93	3.21	0.0059	*This is preferred per the Moser and Stevens paper.*

Equality of Variances				
Method	**Num DF**	**Den DF**	**F Value**	**Pr > F**
Folded F	9	9	2.66	0.1613

Categorical Explanatory Variables

Categorical explanatory variables (e.g., sex, smoking status, diabetes status) can also be used in a regression analysis.

We can create an indicator variable (or "dummy variable") that denotes the category. For example:

sex	*smoke*
0 = Female	0 = Non-Smoker
1 = Male	1 = Smoker

We can then use this indicator variable in the regression model:

$$Y = \beta_0 + \beta_1 X \qquad E[\text{minutes}|\text{females}] = \beta_0$$

$$E[\text{minutes}] = \beta_0 + \beta_1 \times \text{sex} \qquad E[\text{minutes}|\text{males}] = \beta_0 + \beta_1$$

$$E[\text{min}|\text{M}] - E[\text{min}|\text{F}] = (\beta_0 + \beta_1) - (\beta_0) = \beta_1$$

- The "0" category is called the "reference group."
- β_0 is the mean response for the reference group.
- β_1 is the difference in response between the two groups.
- It does not matter which category is chosen as the reference (as long as you get the correct interpretation).

Note: This approach gives the same result as a two-sample *t* test with equal variances.

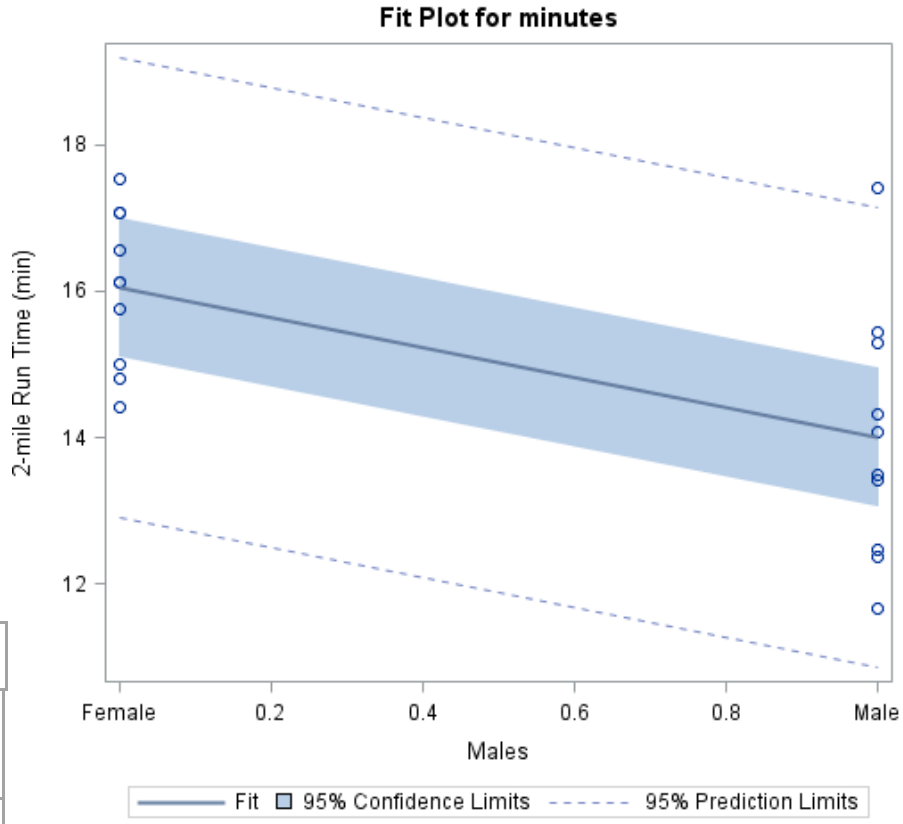
SAS Example: Linear regression model with sex as a covariate

```
PROC REG data=vo2max;
     MODEL minutes = sex; /* Female = 0; Male = 1 */
RUN;
```

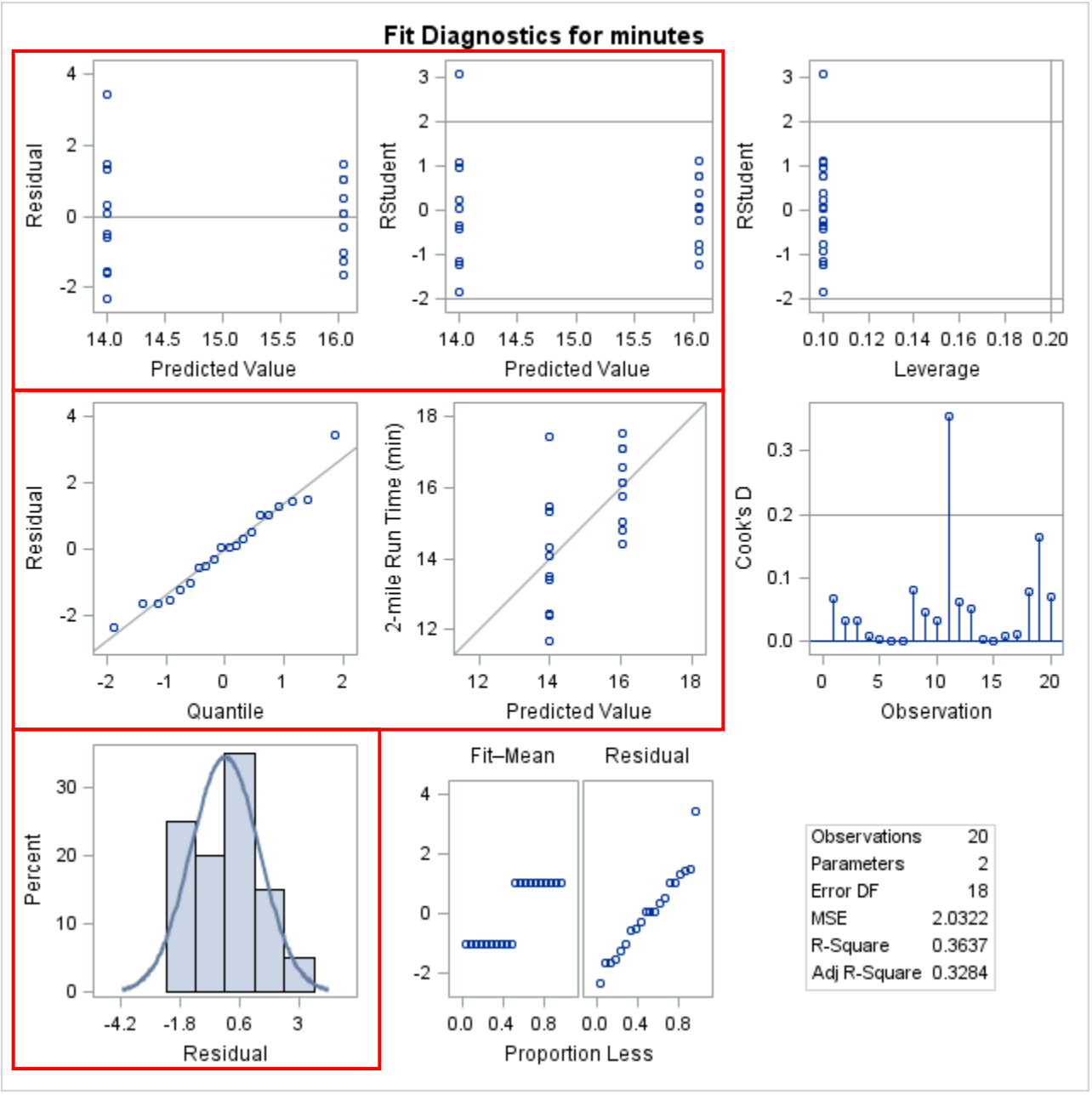
Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	1	20.91013	20.91013	10.29	0.0049
Error	18	36.58033	2.03224		
Corrected Total	19	57.49045			

Root MSE	1.42557	**R-Square**	0.3637
Dependent Mean	15.02350	**Adj R-Sq**	0.3284
Coeff Var	9.48891		

Parameter Estimates						
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	Intercept	**1**	16.04600	0.45080	35.59	<.0001
sex	Males	**1**	-2.04500	0.63753	-3.21	0.0049



Automatically generated regression diagnostic figures:



Diagnostics Evaluation:

1/2. Scatterplots of residuals and predicted values suggest some differences in variability for our two groups.

3. Q-Q plot follows the 45-degree reference line.

4. The predicted vs. dependent variable plot also shows some differences in variability between our two groups and doesn't follow the line (but with categorical data we wouldn't really expect that relationship).

5. Our data set is small, so the histogram is somewhat challenging to evaluate. Based on the other plots, it seems we may be okay.

Hand Calculations: $E(\text{Minutes}_i) = \beta_0 + \beta_1 \text{Sex}_i$, where Sex=0 for females

VO_2 max data:

	(X)		(Y)
ID	*Sex (0=F;1=M)*	*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
Average		43.8300	15.0235

The data can be summarized by the following:

$n = 20$

$\sum X_i = 10$

$\sum Y_i = 300.47$

$S_{XX} = \sum (X_i - \bar{X})^2 = 5$

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

$S_{XY} = \sum (X_i - \bar{X})(Y_i - \bar{Y}) = -10.225$

Hand Calculations: $E(\text{Minutes}_i) = \beta_0 + \beta_1 \text{Sex}_i$, where Sex=0 for females

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{S_{XY}}{S_{XX}} = -2.0450 \qquad \hat{\beta}_0 = \frac{\sum Y_i}{n} - \hat{\beta}_1 \frac{\sum X_i}{n} = 16.046$$

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

$$SS_{Model} = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 = \frac{(S_{XY})^2}{S_{XX}} = 20.91013$$

$$SS_{Error} = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = SS_{Total} - SS_{Model} = 36.58033$$

$$MS_{Model} = \frac{SS_{Model}}{1} = 20.91013$$

$$MS_{Error} = \frac{SS_{Error}}{n-2} = \frac{\left[S_{YY} - \left(\frac{(S_{XY})^2}{S_{XX}}\right)\right]}{n-2} = 2.03224$$

$$F = \frac{MS_{Model}}{MS_{Error}} = 10.29$$

A Complete Interpretation: (1) Point Estimate and (2) Interval Estimate

Point estimate:

$$\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}} = \frac{\text{-}10.2250}{5.0} = \text{-}2.0450$$

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

Interval estimate:

$$SE(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2_{Y|X}}{\sum_{i=1}^{n}(X_i - \bar{X})^2}} = \sqrt{\frac{\hat{\sigma}^2_{Y|X}}{S_{XX}}} = \sqrt{\frac{2.03224}{5.0}} = 0.63753$$

95% CI: $\text{-}2.045 \pm 2.1009 \times 0.63753 = (\text{-}3.384, \text{-}0.706)$

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.

A Complete Interpretation: (3) Decision

Reference value for β_1? No difference between males and females (H_0: $\beta_1 = 0$)

Decision: Is our result consistent with this reference value?

$$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)} = \frac{-2.04500}{0.63753} = -3.21 \sim t_{18} \rightarrow p = 0.0049$$

$$F = \frac{MS_{Model}}{MS_{Error}} = \frac{20.91013}{2.03224} = 10.29 \sim F_{1,18} \rightarrow p = 0.0049$$

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 and conclude that the slope is less than zero.

A Complete Interpretation: (4) Uncertainty

Uncertainty in the decision:

p = 0.0049 (from previous slide)

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.

Summary (including 4 components of a complete interpretation):

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

Reversed Coding for sex: $E(\text{Minutes}_i) = \beta_0^* + \beta_1^* \text{Sex\_Reverse}_i$, where Sex_Reverse=0 for males

```
DATA vo2max;
     SET vo2max;
     IF sex = 1 THEN sex_reverse = 0; /* Female = 1; Male = 0 */
     ELSE IF sex = 0 THEN sex_reverse = 1;
RUN;

PROC REG;
     MODEL minutes = sex_reverse;
RUN;
```

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	1	20.91012	20.91012	10.29	0.0049
Error	18	36.58033	2.03224		
Corrected Total	19	57.49045			

Root MSE	1.42557	**R-Square**	0.3637
Dependent Mean	15.02350	**Adj R-Sq**	0.3284
Coeff Var	9.48891		

Parameter Estimates						
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	Intercept	**1**	14.00100	0.45080	31.06	<.0001
sex_reverse	*Female*	**1**	2.04500	0.63753	3.21	0.0049

$$E[\text{Minutes}] = \beta_0 + \beta_1 \times \text{sex} = 16.046 - 2.045 \times \text{sex}$$

$$E[\text{Minutes}|\text{Female}] = \beta_0 = 16.046$$

$$E[\text{Minutes}|\text{Male}] = \beta_0 + \beta_1 = 16.046 - 2.045 = 14.001$$

$$E[\text{Minutes}|\text{Male}] - E[\text{Minutes}|\text{Female}] = (\beta_0 + \beta_1) - (\beta_0) = \beta_1 = -2.045$$

$$E[\text{Minutes}] = \beta_0^* + \beta_1^* \times \text{sex\_reverse} = 14.001 + 2.045 \times \text{sex\_reverse}$$

$$E[\text{Minutes}|\text{Female}] = \beta_0^* + \beta_1^* = 14.001 + 2.045 = 16.046$$

$$E[\text{Minutes}|\text{Male}] = \beta_0^* = 14.001$$

$$E[\text{Minutes}|\text{Female}] - E[\text{Minutes}|\text{Male}] = (\beta_0^* + \beta_1^*) - (\beta_0^*) = \beta_1^* = 2.045$$

E. Model 3: Adjusting for sex *and* VO_2max

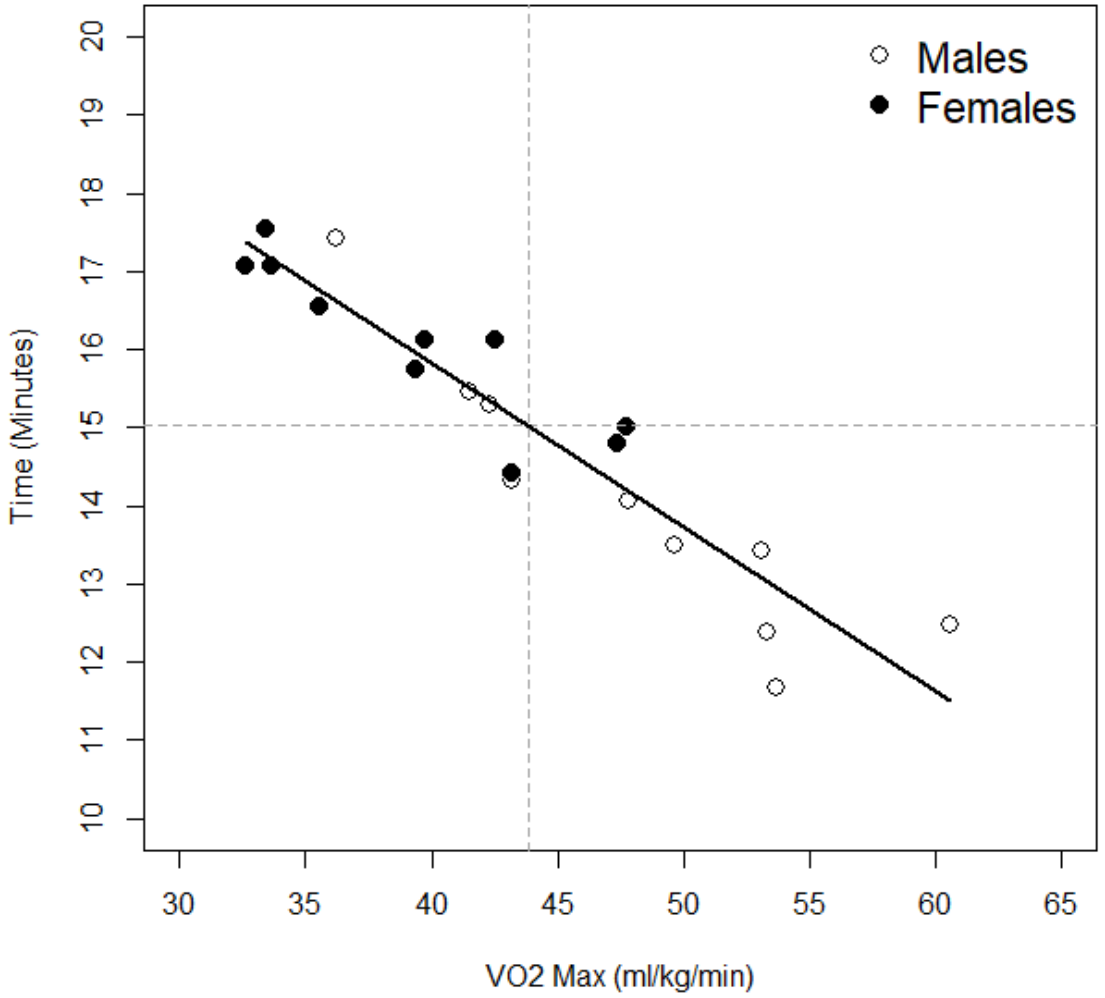
Scientific Question: Do men and women differ in the time required to complete a two-mile run after adjusting for sex differences in VO_2max?

Application of Linear Regression: You want to *describe quantitatively or qualitatively* the relationship between *X* (sex) and *Y* (time to complete 2-mile run) *but control for the effects of still other variable(s) C* (VO_2 max), which you believe have an important relationship with the dependent variable.

How can we address this question?

ANSWER: Multiple Linear Regression!!!

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 \times \text{sex} + \hat{\beta}_2 \times \text{VO}_2\text{max}$$



Example using SAS (Model 3)

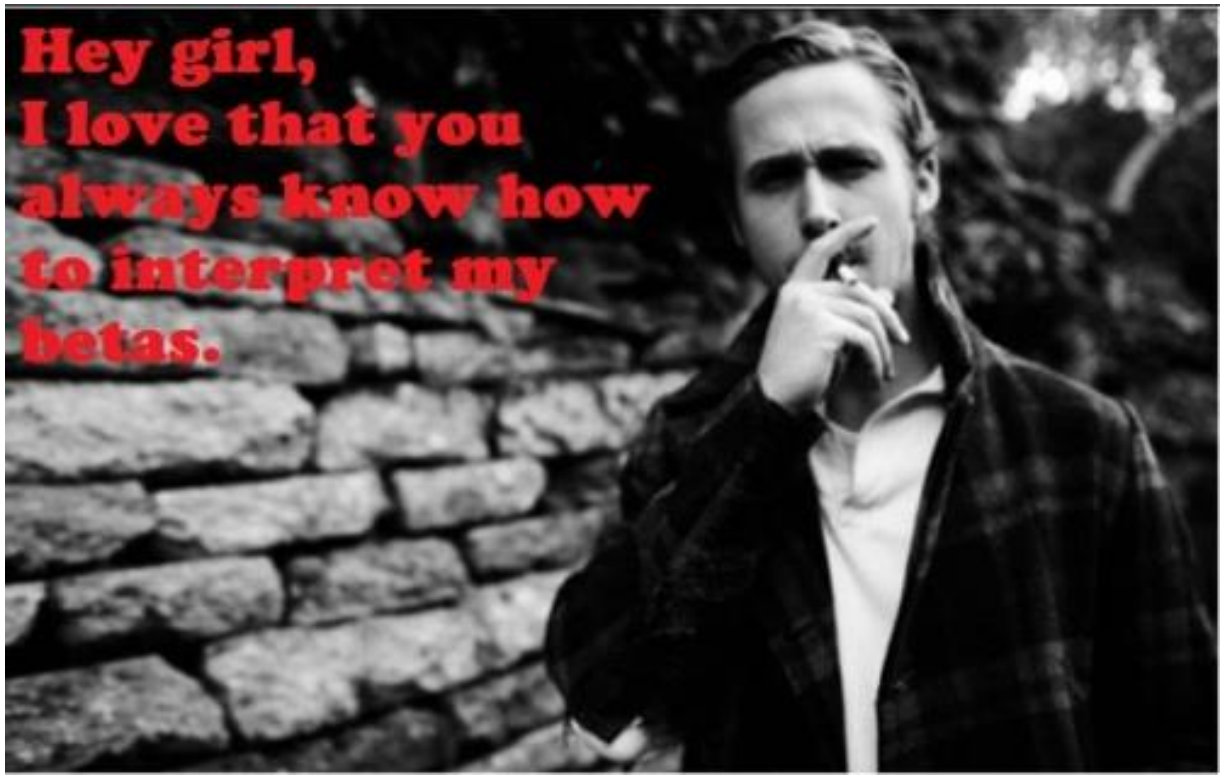
$$E[\text{Minutes}_i] = \beta_0 + \beta_1 \times \text{sex}_i + \beta_2 \text{VO}_2\text{max}_i$$

```
PROC REG;
    MODEL minutes = sex vo2max;
RUN;
```

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	2	50.79777	25.39889	64.52	<.0001
Error	17	6.69268	0.39369		
Corrected Total	19	57.49045			

Root MSE	0.62744	**R-Square**	0.8836
Dependent Mean	15.02350	**Adj R-Sq**	0.8699
Coeff Var	4.17642		

Parameter Estimates						
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	Intercept	**1**	23.82251	0.91430	26.06	<.0001
sex	Males	**1**	-0.34793	0.34158	-1.02	**0.3227**
vo2max	VO2 Max (ml/kg/min)	**1**	-0.19678	0.02258	-8.71	**<.0001**



http://biostatisticsryangoslingreturns.tumblr.com/