PLSC 503 – Spring 2017
Model Fit

January 26, 2017

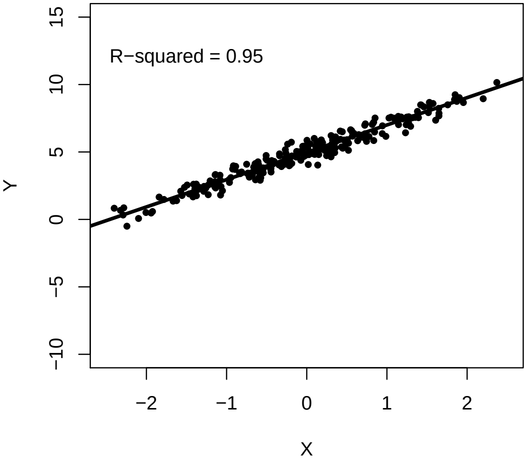
A (Simulated) Example

```
> X<-rnorm(250)
> Y1<-5+2*X+rnorm(250,mean=0,sd=sqrt(0.2))
> Y2<-5+2*X+rnorm(250,mean=0,sd=sqrt(20))
> fit<-lm(Y1~X)
> summary(fit)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.97712    0.02846  174.86   <2e-16 ***
X            2.02529    0.02785   72.73   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.4491 on 248 degrees of freedom
Multiple R-squared:  0.9552,Adjusted R-squared:  0.955
F-statistic:  5290 on 1 and 248 DF,  p-value: < 2.2e-16
```

Regression of $Y_i = 5 + 2X_i + u_i$ ($R^2 = 0.95$)



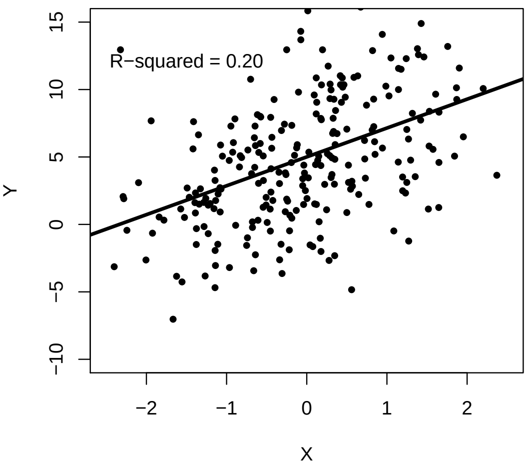
Same Slope/Intercept, Different R^2

```
> fit2<-lm(Y2~X)
> summary(fit2)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.0048     0.2757  18.151  < 2e-16 ***
X             2.1402     0.2697   7.934 7.29e-14 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 4.351 on 248 degrees of freedom
Multiple R-squared:  0.2024,Adjusted R-squared:  0.1992
F-statistic: 62.95 on 1 and 248 DF,  p-value: 7.288e-14
```

Regression of $Y_i = 5 + 2X_i + u_i$ ($R^2 = 0.20$)



Variation in Y

$$
\begin{aligned}
\mathrm{Var}(Y) &= \mathrm{Var}(\hat{Y} + \hat{u}) \\
&= \mathrm{Var}(\hat{Y}) + \mathrm{Var}(\hat{u}) + 2\,\mathrm{Cov}(\hat{Y}, \hat{u}) \\
&= \mathrm{Var}(\hat{Y}) + \mathrm{Var}(\hat{u})
\end{aligned}
$$

TSS	$=$	**MSS**	$+$	**RSS**
("Total")		("Estimated," or "Model")		("Residual")

R^2 Introduced

$$
\begin{aligned}
R^2 &= \frac{\text{MSS}}{\text{TSS}} \\
&= \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2} \\
&= 1 - \frac{\text{RSS}}{\text{TSS}} \\
&= 1 - \frac{\sum \hat{u}_i^2}{\sum(Y_i - \bar{Y})^2}
\end{aligned}
$$

R^2 Is...

R-squared:

- is "the proportion of variance explained"
- $\in [0, 1]$
  - $R^2 = 1.0 \equiv$ a "perfect (linear) fit"'
  - $R^2 = 0 \equiv$ no (linear) $X - Y$ association

For a single X,

$$\begin{aligned} R^2 &= \hat{\beta}_1^2 \frac{\sum(X_i - \bar{X})^2}{\sum(Y_i - \bar{Y})^2} \\ &= r_{XY}^2 \end{aligned}$$

R^2 is Also an *Estimate*...

Luskin: Population analogue "P^2":

$$P^2 = 1 - \frac{\sigma^2}{\sigma_Y^2}$$

Then $\hat{P}^2 = R^2$ has variance:

$$\widehat{\text{Var}(R^2)} = \frac{4R^2(1 - R^2)^2(N - k)^2}{(N^2 - 1)(N + 3)}$$

and standard error:

$$\widehat{\text{s.e.}(R^2)} = \sqrt{\frac{4R^2(1 - R^2)^2(N - k)^2}{(N^2 - 1)(N + 3)}}.$$

Adjusted R^2

$$R^2_{adj.} = 1 - \frac{(1 - R^2)(N - c)}{(N - k)}$$

where $c = 1$ if there is a constant in the model and $c = 0$ otherwise.

$R^2_{adj.}$:

- $R^2_{adj.} \to R^2$ as $N \to \infty$
- $R^2_{adj.}$ can be > 1, or < 0...
- $R^2_{adj.}$ increases with model "fit," but
- The extent of that increase is discounted by a factor proportional to the number of covariates.

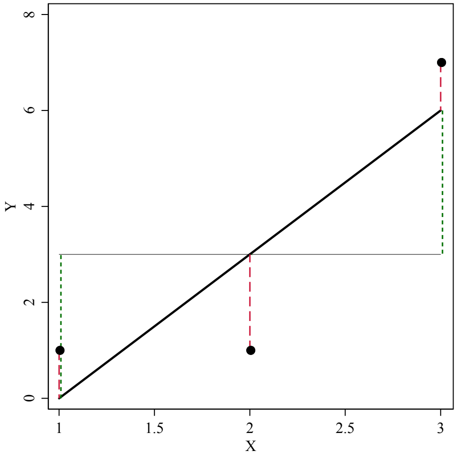
The World's Simplest Regression

Data:

```
  X Y
1 1 1
2 2 1
3 3 7
```

	X_i	Y_i	$X_i - \bar{X}$	$Y_i - \bar{Y}$	$(X_i - \bar{X})^2$	$(Y_i - \bar{Y})^2$	$(X_i - \bar{X})(Y_i - \bar{Y})$
	1	1	-1	-2	1	4	2
	2	1	0	-2	0	4	0
	3	7	1	4	1	16	4
$\sum_{i=1}^{3}(\cdot) =$	6	9	0	0	2	24	6

The World's Simplest Regression



The World's Simplest Regression

```
> X<-c(1,2,3)
> Y<-c(1,1,7)
> WSR<-lm(Y~X)
> summary(WSR)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -3.000      3.742  -0.802    0.570
X              3.000      1.732   1.732    0.333

Residual standard error: 2.449 on 1 degrees of freedom
Multiple R-squared:   0.75,Adjusted R-squared:    0.5
F-statistic:     3 on 1 and 1 DF,  p-value: 0.3333
```

R^2 Alternatives

- Standard Error of the Estimate:

$$\text{SEE} = \sqrt{\frac{\text{RSS}}{N - k}}$$

- F-tests (later...)

- ROC / AUC (later...)

- Graphical methods

Caution: Different Ways to get $R^2 = 0$