Announcements:

- Wendy's new office hours: Mon 11-12:30, Tues 3:30-5, 2032 DBH.
- No office hours on Thursday for any of us now, but lots on Monday. See website for full schedule.

TODAY: Sections 2.2 and 2.3

Some of today's lecture will be on the white board, not in these notes.

- ANOVA = Analysis of Variance, partitioning variability into explainable and unexplainable parts
- R^2 = "Coefficient of determination" = $(\text{Correlation})^2$
- Another way to test the hypotheses that $\beta_1 = 0$ vs $\beta_1 \neq 0$

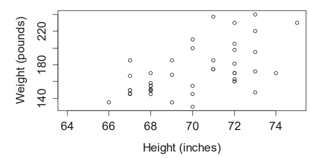
Example:
n = 43 male college students
Y = weight (in pounds)
X = height (in inches)

Data available on class website (HtWt.txt) in list of data sets

Goals:

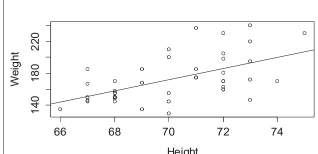
- Predict weight from height.
- Estimate average weight at any given height.

CHOOSE:
It looks like a linear model could be appropriate.



Plot the data with regression line:

FIT (output on next page):
Regression line shown:
$\hat{Y} = -317.92 + 6.996X$
Slope interpretation??
Example:
X = 75 inches
$\hat{Y} \approx -318 + 7(75)$
= 207 pounds



ASSESS:
Stem and leaf plot of standardized residuals looks good.

```
> stem(HtWt$StResids)

  The decimal point is at the |

  -2 | 0
  -1 | 8
  -1 | 331110
  -0 | 97775
  -0 | 4433322221100
   0 | 0113
   0 | 55789
   1 | 022
   1 | 569
   2 | 0
   2 | 5
```

- Some (partial) results from R; things you should already know in boxes
- *Interpretation of s = $\hat{\sigma}_\varepsilon$ = residual standard error of 24 (pounds)??*
- Thing shown in bold red explained today.

```
> Mod<-lm(Weight~Height)
> summary(Mod)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -317.919    110.922  -2.866  0.00653 **
Height         6.996      1.581   4.425 6.98e-05

Residual standard error: 24 on 41 degrees of freedom
Multiple R-squared:  0.3232,  Adjusted R-squared:  0.3067
F-statistic: 19.58 on 1 and 41 DF,  p-value: 6.978e-05

> anova(Mod)
Analysis of Variance Table

Response: Weight
          Df Sum Sq Mean Sq F value    Pr(>F)
Height     1  11277   11277  19.578 6.978e-05 ***
Residuals 41  23617     576
```

Create the same plot with a dotted line at the mean weight, which is 172.6 pound, for reasons that will become clear.

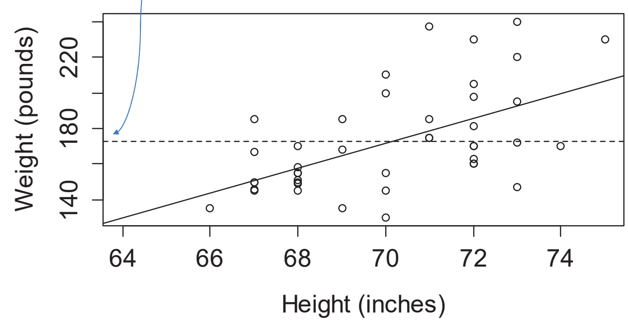
R Commands:
Create plot, add the line for the model called "Mod", then add horizontal line at the mean weight:

```
> plot(x=HtWt$Height, y = HtWt$Weight, xlab =
"Height (inches)", ylab = "Weight (pounds)", xlim
= c(64, 75), ylim = c(130, 240), type = "p")
> abline(Mod)
> abline(h=mean(HtWt$Weight), lty = 2)
```

In the abline command

- "Mod" is the name of the regression model used previously, and this tells R to add the regression line.
- "h" means to insert a horizontal line at the value given, using line type 2 (lty = 2), which is a dotted line.

If we didn't know height, our best guess for any individual's weight would be $\bar{Y}$ = 172.6 pounds.



Let's compare how well we can predict weight for our sample:

- If we don't use height, Predicted weight = $\bar{Y}$ = 172.6 pounds for all
- If we do use height, Predicted weight = $\hat{Y} = -317.919 + 6.996X$

How much better (or worse) can we predict weight using height info, using the "Least squares" criterion?

X = Ht	Y = Wt	$\bar{Y}$= Pred. Wt. w/o Height	$\hat{Y}$ = Pred. Wt. with Height	$(Y-\bar{Y})$ = "total deviation"	$(Y-\hat{Y})$ = residual	$(\hat{Y}-\bar{Y})$ "explained" using height	Squared total deviation	Squared residual
73	195	172.6	192.8	22.4	2.2	20.2	501.76	4.84
69	135	172.6	164.8	-37.6	-29.8	-7.8	1413.76	888.04
70	145	172.6	171.8	-27.6	-26.8	-0.8	761.76	718.24
69	168	172.6	164.8	-4.6	3.2	-7.8	21.16	10.24
68	155	172.6	157.8	-17.6	-2.8	-14.8	309.76	7.84
71	185	172.6	178.8	12.4	6.2	6.2	153.76	38.44
71	175	172.6	178.8	2.4	-3.8	6.2	5.76	14.44
68	158	172.6	157.8	-14.6	0.2	-14.8	213.16	0.04
Etc	Etc	Etc	Etc	Etc	Etc	Etc	Etc	Etc

Without using height: Sum = $\sum(Y-\bar{Y})^2$ = SSY = SSTO = Total SS = 34894
With using height: Sum = $\sum(Y-\hat{Y})^2$ = SSE = Error SS = Residual SS = 23617

The sum is *smaller* using the predicted values from the regression line than using the mean weight as the prediction for everyone.

ANALYZING DIFFERENT SOURCES OF VARIABILITY ("Analysis of variance")

For each individual: $(Y-\bar{Y})$ = $(\hat{Y}-\bar{Y})$ + $(Y-\hat{Y})$
Total deviation = explained by regression + unexplained residual

Let's compare the sum of squares for these:

Without using height: Sum = $\sum(Y-\bar{Y})^2$ = SSY = SSTO = Total SS = 34894
With using height: Sum = $\sum(Y-\hat{Y})^2$ = SSE = Error SS = Residual SS = 23617
Difference: Sum = $\sum(\hat{Y}-\bar{Y})^2$ = SSR = SS due to Regression = 11277
(Called SS Model in the book.)

See picture on white board.

By algebraic magic, SSTO = SS Model + SSE.
Degrees of freedom: $(n-1)$ = (1) + $(n-2)$

SSTO measures the total variability of the Y values around their mean.

SS Model = SSR measures the amount of that total variability that's "explained" by using the explanatory variable, sometimes noted as "explained" by the regression equation.

SSE is the part of that variability that is still unexplained, even after using X.

<u>Definition</u> of the "coefficient of determination," R^2

$$R^2 = \frac{Variability\ explained\ by\ the\ model}{Total\ variability\ in\ Y} = \frac{SS\ Model}{SSTO}$$

This is the proportion of the total variability in the Y values that can be "explained" by using the model.

Note: It can be written as $R^2 = \frac{SSTO-SSE}{SSTO} = 1 - \frac{SSE}{SSTO}$

Height/Weight example, partial output from before:

```
> summary(Mod)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -317.919    110.922  -2.866  0.00653 **
Height         6.996      1.581   4.425 6.98e-05

Residual standard error: 24 on 41 degrees of freedom
Multiple R-squared:  0.3232,  Adjusted R-squared:  0.3067
```

$R^2 = \frac{11277}{34894} = 0.3232$

<u>Interpretation</u>: About 32% of variation in male weights (at least for this sample) is explained by knowing heights.

Notes about $R^2 = \frac{SS\ Model}{SSTO} = 1 - \frac{SSE}{SSTO}$

1. It is the correlation coefficient (*r*) squared.
2. $0 \le R^2 \le 1$
   - When $R^2 = 0$ it means SSE = SSTO, and SS Model = 0, so *no* additional variability is explained by using X
   - When $R^2 = 1$ it means SSE = 0, so all points fall on the line and Y is <u>completely</u> predicted by knowing X.
3. It is often expressed as a percent rather than a proportion.

Analysis of Variance Table, generic:

Source		Degrees of freedom	Sum of Squares	Mean Square = SS/df	Test statistic F	p-value for F; df = 1, *n - 2*
Book	R	Df	Sum Sq	Mean Sq	F value	Pr (> F)
Model	X name	1	SS Model	MS Model	$\frac{MSModel}{MSE}$	Area above F
Error	Residuals	$n-2$	SSE	MSE		
Total	Not shown	$n-1$	SSTO	Not used		

Height/Weight example:

```
> anova(Mod)
Analysis of Variance Table
Response: Weight
          Df Sum Sq Mean Sq F value    Pr(>F)
Height     1  11277   11277  19.578 6.978e-05 ***
Residuals 41  23617     576
```

New example: Skin cancer and latitude/longitude (separate file)

Derivation and explanation of the F test, shown on white board.