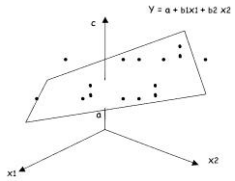
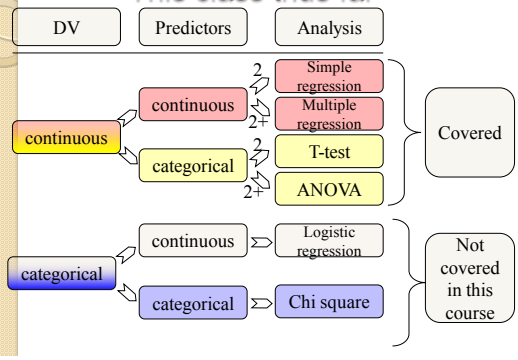
Simple and Multiple Regression



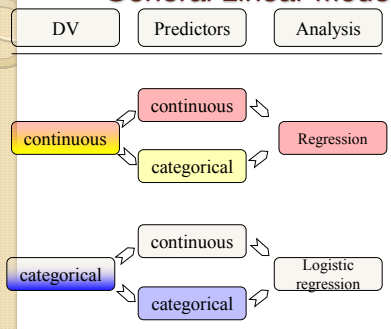
Lectures 32-35

Readings: GW 17 & SS 12.3-12.11

Analysis flowchart – This class thus far



Analysis flowchart – General Linear Model

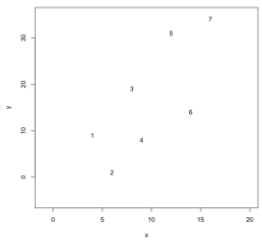


Simple Linear Regression: Examining Relationship between Two Continuous Variables

- Can we predict Exam 2 performance by knowing Exam 1 performance?
- To what extent does SAT performance predict college GPA's?
- Do wealthier people have more conservative political attitudes?
- Is there a relationship between the percentage of adults in a state who smoke and the mean educational level in the state?
- Are courses where students receive higher grades evaluated more positively?
- Is reading ability related to shoe size in children?

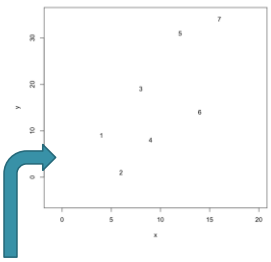
Simple Linear Regression example – best fit line

```
> dat1
   x  y
1  4  9
2  6  1
3  8 19
4  9  8
5 12 31
6 14 14
7 16 34
```



Simple Linear Regression example – best fit line

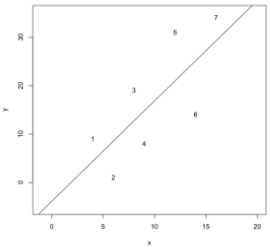
```
> dat1
   x  y
1  4  9
2  6  1
3  8 19
4  9  8
5 12 31
6 14 14
7 16 34
```



Find the line that minimizes the square vertical difference between the line and each point

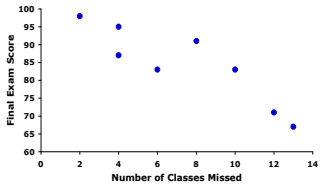
Simple Linear Regression example – best fit line

```
> dat1
   x  y
1  4  9
2  6  1
3  8 19
4  9  8
5 12 31
6 14 14
7 16 34
```

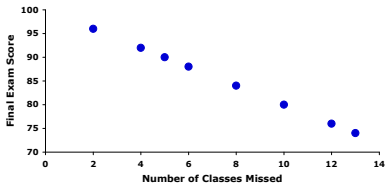


Predicting Exam Score from Attendance

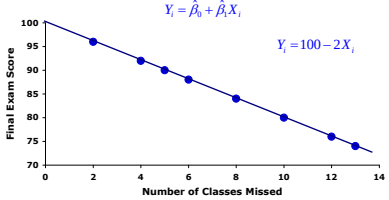
Classes Cut	Final Exam
4	95
6	83
12	71
2	98
4	87
10	83
13	67
8	91



If a perfect relationship



If a perfect relationship



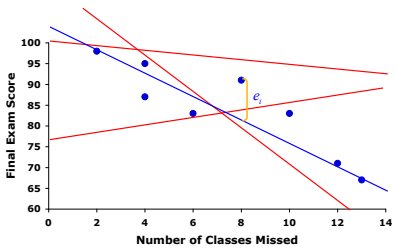
$\hat{\beta}_0$: Intercept: value of Y when X equals zero

If no classes missed 100 final exam score

$\hat{\beta}_1$: Slope: difference in Y as a function of a 1 unit difference in X

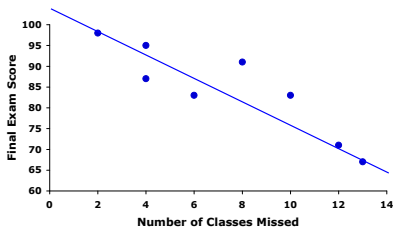
Each class missed associated with 2 fewer points on final

Finding the "best" line



$$Y_i = \beta_0 + \beta_1 X_i + e_i \qquad \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \qquad e_i = Y_i - \hat{Y}_i = Y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 X_i\right)$$

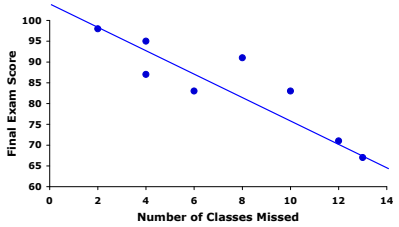
Finding the "best" line



$$Y_i = \beta_0 + \beta_1 X_i + e_i \qquad \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \qquad e_i = Y_i - \hat{Y}_i = Y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 X_i\right)$$

Minimize $\sum_i e_i = \sum_i \left(Y_i - \hat{Y}_i\right)$??

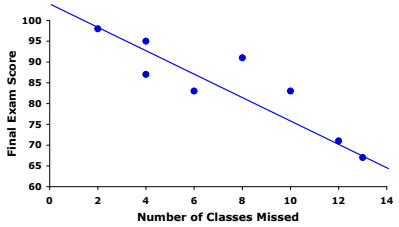
Finding the "best" line



$$Y_i = \beta_0 + \beta_1 X_i + e_i \qquad \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \qquad e_i = Y_i - \hat{Y}_i = Y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 X_i\right)$$

$$\text{Minimize } \sum_i e_i^2 = \sum_i \left(Y_i - \hat{Y}_i\right)^2$$

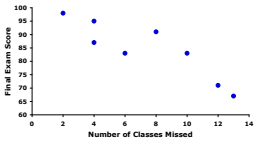
Finding the "best" line



$$\beta_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})/n-1}{\sum(X_i - \bar{X})^2/n-1}$$

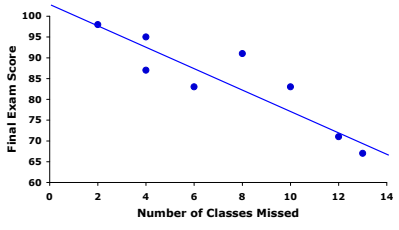
$$\hat{\beta}_0 = \bar{Y} - \beta_1 \bar{X}$$

Predicting Final Exam Score from Classes Missed



Classes Cut	Final Exam	$(X - \bar{X})$	$(Y - \bar{Y})$	$(X - \bar{X})(Y - \bar{Y})$	$(X - \bar{X})^2$
4	95	-3.375	10.625	-35.859	11.391
6	83	-1.375	-1.375	1.891	1.891
12	71	4.625	-13.375	-61.859	21.391
2	98	-5.375	13.625	-73.234	28.891
4	87	-3.375	2.625	-8.859	11.391
10	83	2.625	-1.375	-3.609	6.891
13	67	5.625	-17.375	-97.734	31.641
8	91	0.625	6.625	4.141	0.391
$\bar{X}$ = 7.375	$\bar{Y}$ = 84.375			$\sum$ -275.125	$\sum$ 113.875

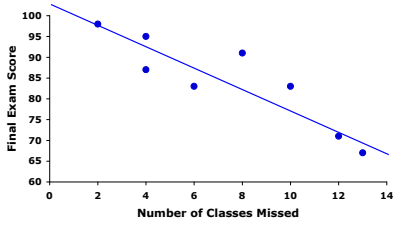
Finding the "best" line



$$\hat{\beta}_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} = \frac{-275.125}{113.875} = -2.42$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = 84.375 - (-2.42)7.375 = 102.2$$

Finding the "best" line

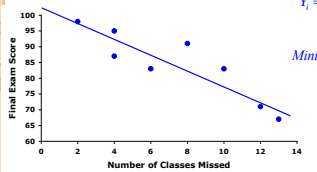


$$\text{Minimize } \sum_i e_i^2 = \sum_i \left(Y_i - \hat{Y}_i\right)^2$$

$$\hat{\beta}_1 = -2.42 \qquad \hat{\beta}_0 = 102.2$$

$$Y_i = -2.42 X_i + 102.2 + e_i$$

$$\hat{Y}_i = -2.42 X_i + 102.2$$

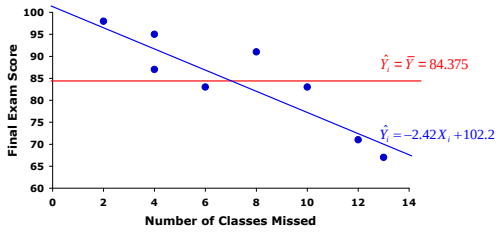


$$Y_i = -2.42 X_i + 102.2 + e_i$$

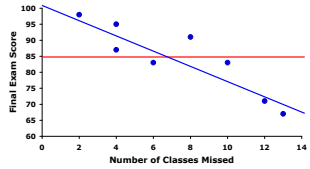
$$\text{Minimize } \sum_i e_i^2 = \sum_i \left(Y_i - \hat{Y}_i\right)^2$$

Classes Cut	Final Exam	$\hat{Y}_i$ (-2.42X_i +102.22)	$Y_i - \hat{Y}$	$\left(Y_i - \hat{Y}_i\right)^2$
4	95	92.54	2.46	6.052
6	83	87.70	-4.70	22.090
12	71	73.18	-2.18	4.752
2	98	97.38	0.62	0.384
4	87	92.54	-5.54	30.692
10	83	78.02	4.98	24.800
13	67	70.76	-3.76	14.138
8	91	82.86	8.14	66.260
$\bar{X}$ =7.375	$\bar{Y}$ =84.375			$\sum$ 169.168

If X is a useful predictor of Y, how much better is it than just predicting the mean?

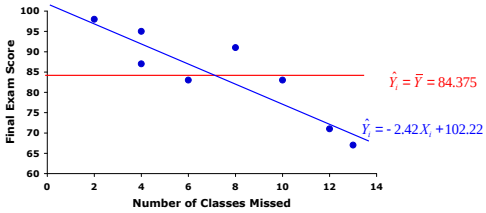


If X is a useful predictor of Y, how much better is it than just predicting the mean?



Classes Cut	Final Exam	$\hat{Y}_i$ $(-2.42X_i + 102.22)$	$Y_i - \hat{Y}_i$	$(Y_i - \hat{Y}_i)^2$	$Y_i - \bar{Y}$	$(Y_i - \bar{Y})^2$
4	95	92.54	2.46	6.052	10.625	112.891
6	83	87.70	-4.70	22.090	-1.375	1.891
12	71	73.18	-2.18	4.752	-13.375	178.891
2	98	97.38	0.62	0.384	13.625	185.641
4	87	92.54	-5.54	30.692	2.625	6.891
10	83	78.02	4.98	24.800	-1.375	1.891
13	67	70.76	-3.76	14.138	-17.375	301.891
8	91	82.86	8.14	66.260	6.625	43.891
$\bar{X}$ =7.375	$\bar{Y}$ =84.375			å 169.168		å 833.875

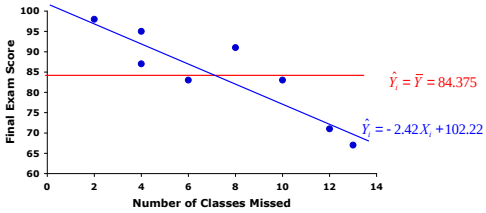
If X is a useful predictor of Y, how much better is it than just predicting the mean?



$\sum (Y_i - \bar{Y})^2 = 833.875$ $\quad\quad$ $\sum (Y_i - \hat{Y}_i)^2 = 169.168$

$$r^2 = \frac{\sum (Y_i - \bar{Y})^2 - \sum (Y_i - \hat{Y}_i)^2}{\sum (Y_i - \bar{Y})^2} = \frac{833.875 - 169.168}{833.875} = \frac{644.707}{833.875} = .797$$

If X is a useful predictor of Y, how much better is it than just predicting the mean?

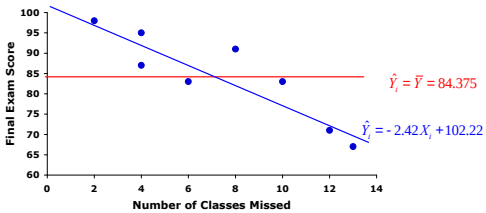


$\sum (Y_i - \bar{Y})^2 = 833.875$ $\quad\quad$ $\sum (Y_i - \hat{Y}_i)^2 = 169.168$

$$r^2 = \frac{\sum (Y_i - \bar{Y})^2 - \sum (Y_i - \hat{Y}_i)^2}{\sum (Y_i - \bar{Y})^2} = .797$$

Using Number of Classes Missed to predict Final Exam Scores improves our predictions by 79% relative to predicting the mean Final Exam Score

If X is a useful predictor of Y, how much better is it than just predicting the mean?



$r^2 = .797$

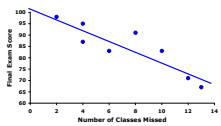
Using Number of Classes Missed to predict Final Exam Scores improves our predictions by 79% relative to predicting the mean Final Exam Score

Of the total variation (sum of squares) around the mean, 79% can be explained by differences due to the number of classes missed.

Linear Regression Questions

$\hat{\beta}_1 = -2.42$

$\hat{\beta}_0 = 102.19$

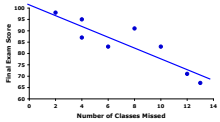


1) A student misses 3 classes. What is her predicted final exam score?
2) What is the interpretation of the intercept?
3) Bob has a "residual" (e_i) of +10. What does this mean in words?

Inference in linear regression

$\hat{\beta}_1 = -2.42$

$\hat{\beta}_0 = 102.19$



Question: If the null were true (that there was no relationship between classes missed & final exam score), how often would we observe a relationship as strong or stronger than the one observed?

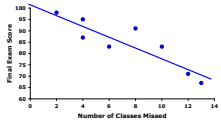
Null hypothesis $\beta_1 = 0$

Alt. hypothesis $\beta_1 \neq 0$

Inference in linear regression

$\hat{\beta}_1 = -2.42$

$\hat{\beta}_0 = 102.19$



$$se(\beta_1) = \sqrt{\frac{V(Y)(1-r^2)}{V(X)(n-2)}}$$

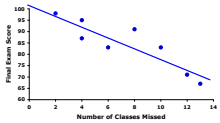
$$t(n-2) = \frac{\beta_1}{se(\beta_1)}$$

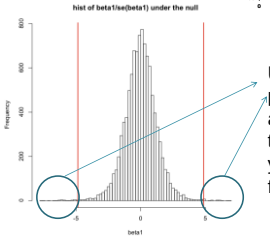
Thought experiment: Simulated data under the null hypothesis. Take many samples of beta1 and divide by the se(beta1). This distribution follows a t-distribution. How often will you observe a beta1 more extreme (in either direction) than the one observed?

Inference in linear regression

$\hat{\beta}_1 = -2.42$

$\hat{\beta}_0 = 102.19$



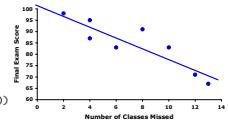


Under the "null", the proportion of beta1's that are more extreme than the observed beta1 is your p-value. (Similarly for beta0)

Inference in linear regression

$\hat{\beta}_1 = -2.42$

$\hat{\beta}_0 = 102.19$




```
> summary(lm(final.score ~ classes.missed))

Call:
lm(formula = final.score ~ classes.missed)

Residuals:
    Min      1Q  Median      3Q     Max
-5.529 -4.013 -0.781  3.095  8.135

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    102.1932     4.1228  24.792 2.83e-07 ***
classes.missed  -2.4160     0.4976  -4.856  0.00284 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.31 on 6 degrees of freedom
Multiple R-squared: 0.7971,     Adjusted R-squared: 0.7633
F-statistic: 23.58 on 1 and 6 DF,  p-value: 0.002836
```

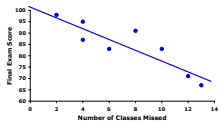
Assumptions

- **Independence**: data are independent observations
- **Linearity**: a straight-line relationship is adequate
- **Normality**: errors are normally distributed
- **Equal variance**: errors have the same variance across values of x

4 sentence summary of linear regression

$\hat{\beta}_1 = -2.42$

$\hat{\beta}_0 = 102.19$



It is commonly assumed that missing more classes will result in lower grades in college classes. To test this, we related the number of classes missed among 8 students taking an introduction to statistics class to their final exam score. We found that each class missed was associated with a predicted 2.4 lower final exam grade (beta1 = -2.42, t(6) = -4.86, p=.003). This evidence is consistent with the hypothesis that missing classes leads to lower grades in college.

How to perform a t-test with regression

- Usually, when you have a continuous dependent variable (e.g., height) and a discrete predictor (e.g., gender), you would perform a t-test.
- But if you code your discrete predictor (using "Contrast Codes"), you can use regression instead!

```
> ht2
   gender height weight
1    Male   66.0    140
14   Male   67.0    145
27 Female   68.0    130
40   Male   72.0    215
53   Male   73.0    155
66   Male   69.0    136
79 Female   65.0    122
92 Female   61.8    108
```

"Contrast Code" – just converting males to 1 and females to 0

How to perform a t-test with regression

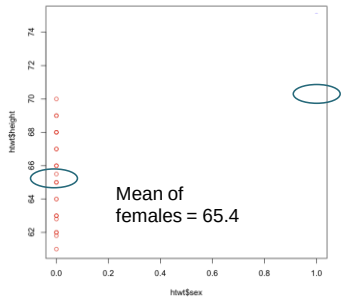
```
> t.test(height~gender,var.equal=TRUE,data=htwt)

        Two Sample t-test

data:  height by gender
t = -9.6823, df = 90, p-value = 1.307e-15
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -6.449594 -4.253464
sample estimates:
mean in group Female   mean in group Male
            65.40286            70.75439
```

70.75 – 65.4 = 5.35

```
> summary(lm(height~sex,data=htwt))

Call:
lm(formula = height ~ sex, data = htwt)

Residuals:
    Min      1Q  Median      3Q     Max
-4.7544 -1.9165  0.2456  2.2456  4.5971

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  65.4029     0.4351 150.332  < 2e-16 ***
sex           5.3515     0.5527   9.682 1.31e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.574 on 90 degrees of freedom
Multiple R-squared: 0.5102,     Adjusted R-squared: 0.5048
F-statistic: 93.75 on 1 and 90 DF,  p-value: 1.307e-15
```

Contrast coding in regression

```
> plot(htwt$sex,htwt$height,col=col)
```

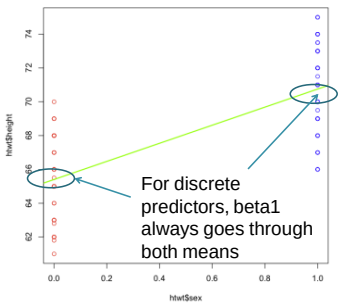


Mean of males = 70.75

Mean of females = 65.4

Contrast coding in regression

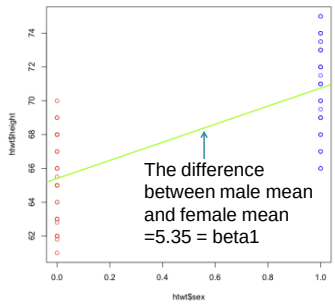
```
> plot(htwt$sex,htwt$height,col=col)
```



For discrete predictors, beta1 always goes through both means

Contrast coding in regression

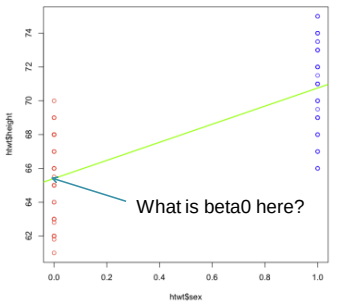
```
> plot(htwt$sex,htwt$height,col=col)
```



The difference between male mean and female mean =5.35 = beta1

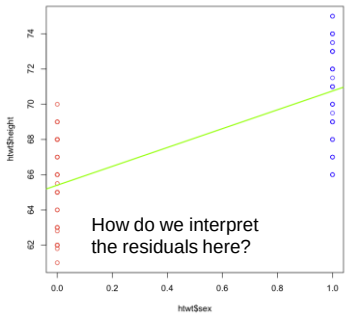
Contrast coding in regression

```
> plot(htwt$sex,htwt$height,col=col)
```



What is beta0 here?

Residuals again



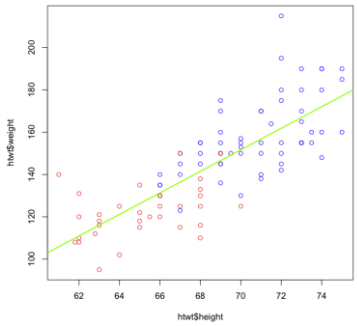
t-test with regression

```
> summary(lm(height~sex,data=htwt))

Call:
lm(formula = height ~ sex, data = htwt)

Residuals:
    Min      1Q  Median      3Q     Max
-4.7544 -1.9165  0.2456  2.2456  4.5971

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  65.4029     0.4351 150.332  < 2e-16 ***
sex           5.3515     0.5527   9.682 1.31e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.574 on 90 degrees of freedom
Multiple R-squared: 0.5102,     Adjusted R-squared: 0.5048
F-statistic: 93.75 on 1 and 90 DF,  p-value: 1.307e-15
```

Reversing the coding (making males 0)

```
> htwt$sex2 <- (htwt$gender=="Female")*1


> summary(lm(height ~ sex2,data=htwt))

Call:
lm(formula = height ~ sex2, data = htwt)

Residuals:
    Min      1Q  Median      3Q     Max
-4.7544 -1.9165  0.2456  2.2456  4.5971

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  70.7544     0.3409 207.545  < 2e-16 ***
sex2         -5.3515     0.5527  -9.682 1.31e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.574 on 90 degrees of freedom
Multiple R-squared: 0.5102,     Adjusted R-squared: 0.5048
F-statistic: 93.75 on 1 and 90 DF,  p-value: 1.307e-15
```

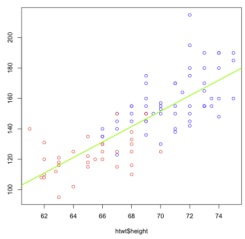
Scatterplot – height & weight



Regression – height & weight

```
> summary(lm(weight ~ height,data=htwt))

Call:
lm(formula = weight ~ height, data = htwt)

Residuals:
    Min      1Q  Median      3Q     Max
-31.492 -11.327  -1.115   8.626  53.131

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -204.918     29.177  -7.023 3.98e-10 ***
height         5.094      0.424  12.015  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.79 on 90 degrees of freedom
Multiple R-squared: 0.616,      Adjusted R-squared: 0.6117
F-statistic: 144.4 on 1 and 90 DF,  p-value: < 2.2e-16
```

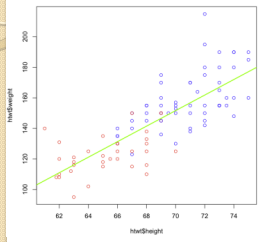
Scatterplot – height & weight



As we increase 1 on height, we are also increasing on some other variable.

What is it??

Scatterplot – height & weight



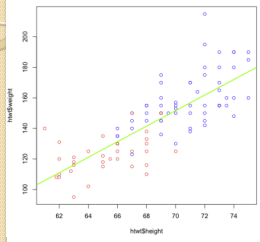
As we increase 1 on height, we are also increasing on some other variable.

What is it??

It is also becoming more likely the person is male!

In other words, part of the relationship between height and weight may simply be because males are heavier than females!

Scatterplot – height & weight

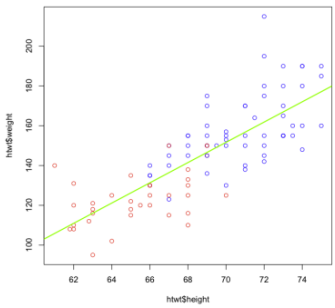


GOAL: We might want to know what the relationship between height and weight is – INDEPENDENT of gender.

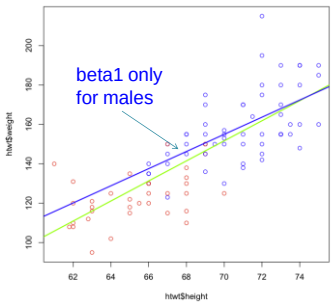
I.E. – What is the relationship between height and weight "controlling for gender"?

This is accomplished using MULTIPLE REGRESSION!

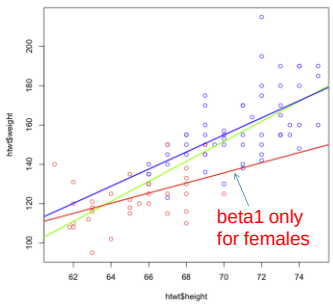
Multiple regression



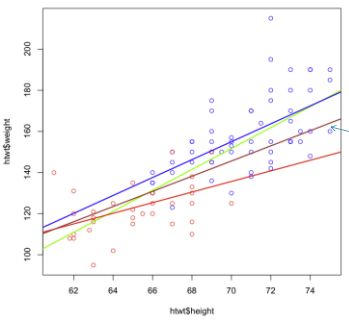
Multiple regression



Multiple regression



Multiple regression



The AVERAGE of the blue and red lines is the slope of height on weight INDEPENDENT of gender

Multiple regression

```
> summary(lm(weight ~ height + sex,data=htwt))

Call:
lm(formula = weight ~ height + sex, data = htwt)

Residuals:
    Min      1Q  Median      3Q     Max
-25.477  -8.594  -1.395   7.992  52.137

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -117.7104    37.5254  -3.137 0.002316 **
height         3.6927     0.5726   6.449 5.69e-09 ***
sex           14.7018     4.2902   3.427 0.000926 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.98 on 89 degrees of freedom
Multiple R-squared: 0.6607,     Adjusted R-squared: 0.6531
F-statistic: 86.67 on 2 and 89 DF,  p-value: < 2.2e-16
```

Multiple regression

```
> summary(lm(weight ~ height + sex,data=htwt))

Call:
lm(formula = weight ~ height + sex, data = htwt)

Residuals:
    Min      1Q  Median      3Q     Max
-25.477  -8.594  -1.395   7.992  52.137

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -117.7104    37.5254  -3.137 0.002316 **
height         3.6927     0.5726   6.449 5.69e-09 ***
sex           14.7018     4.2902   3.427 0.000926 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.98 on 89 degrees of freedom
Multiple R-squared: 0.6607,     Adjusted R-squared: 0.6531
F-statistic: 86.67 on 2 and 89 DF,  p-value: < 2.2e-16
```

In multiple regression, we add 2+ predictors

Multiple regression

```
> summary(lm(weight ~ height + sex,data=htwt))

Call:
lm(formula = weight ~ height + sex, data = htwt)

Residuals:
    Min      1Q  Median      3Q     Max
-25.477  -8.594  -1.395   7.992  52.137

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -117.7104    37.5254  -3.137 0.002316 **
height         3.6927     0.5726   6.449 5.69e-09 ***
sex           14.7018     4.2902   3.427 0.000926 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.98 on 89 degrees of freedom
Multiple R-squared: 0.6607,     Adjusted R-squared: 0.6531
F-statistic: 86.67 on 2 and 89 DF,  p-value: < 2.2e-16
```

In multiple regression, we add 2+ predictors and look at the UNIQUE effects of each, holding "constant" the other(s)

Multiple regression

```
> summary(lm(weight ~ height + sex,data=htwt))

Call:
lm(formula = weight ~ height + sex, data = htwt)

Residuals:
    Min      1Q  Median      3Q     Max
-25.477  -8.594  -1.395   7.992  52.137

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -117.7104    37.5254  -3.137 0.002316 **
height         3.6927     0.5726   6.449 5.69e-09 ***
sex           14.7018     4.2902   3.427 0.000926 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.98 on 89 degrees of freedom
Multiple R-squared: 0.6607,     Adjusted R-squared: 0.6531
F-statistic: 86.67 on 2 and 89 DF,  p-value: < 2.2e-16
```

Holding gender constant, for each 1" increase in height, weight is predicted to increase 3.69 lbs. I.e., the "unique" effect of height on weight is 3.69

Multiple regression

```
> summary(lm(weight ~ height + sex,data=htwt))

Call:
lm(formula = weight ~ height + sex, data = htwt)

Residuals:
    Min      1Q  Median      3Q     Max
-25.477  -8.594  -1.395   7.992  52.137

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -117.7104    37.5254  -3.137 0.002316 **
height         3.6927     0.5726   6.449 5.69e-09 ***
sex           14.7018     4.2902   3.427 0.000926 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.98 on 89 degrees of freedom
Multiple R-squared: 0.6607,     Adjusted R-squared: 0.6531
F-statistic: 86.67 on 2 and 89 DF,  p-value: < 2.2e-16
```

Holding height constant (for equally tall males and females), males are predicted to weight 14.7 lbs more

Multiple regression

```
> summary(lm(weight ~ height + sex,data=htwt))

Call:
lm(formula = weight ~ height + sex, data = htwt)

Residuals:
    Min      1Q  Median      3Q     Max
-25.477  -8.594  -1.395   7.992  52.137

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -117.7104    37.5254  -3.137 0.002316 **
height         3.6927     0.5726   6.449 5.69e-09 ***
sex           14.7018     4.2902   3.427 0.000926 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.98 on 89 degrees of freedom
Multiple R-squared: 0.6607,     Adjusted R-squared: 0.6531
F-statistic: 86.67 on 2 and 89 DF,  p-value: < 2.2e-16
```

Predicted score of a female (coded 0) who is 0" tall

Multiple regression

```
> summary(lm(weight ~ height + sex,data=htwt))

Call:
lm(formula = weight ~ height + sex, data = htwt)

Residuals:
    Min      1Q  Median      3Q     Max
-25.477  -8.594  -1.395   7.992  52.137

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -117.7104    37.5254  -3.137 0.002316 **
height         3.6927     0.5726   6.449 5.69e-09 ***
sex           14.7018     4.2902   3.427 0.000926 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.98 on 89 degrees of freedom
Multiple R-squared: 0.6607,     Adjusted R-squared: 0.6531
F-statistic: 86.67 on 2 and 89 DF,  p-value: < 2.2e-16
```

Together, height and gender explain 66.1% of the variation in weight. (Knowing a person's gender and height, you can guess their weight 66.1% better than just using the mean)

Multiple regression

```
> summary(lm(weight ~ height + sex,data=htwt))

Call:
lm(formula = weight ~ height + sex, data = htwt)

Residuals:
    Min      1Q  Median      3Q     Max
-25.477  -8.594  -1.395   7.992  52.137

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -117.7104    37.5254  -3.137 0.002316 **
height         3.6927     0.5726   6.449 5.69e-09 ***
sex           14.7018     4.2902   3.427 0.000926 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.98 on 89 degrees of freedom
Multiple R-squared: 0.6607,     Adjusted R-squared: 0.6531
F-statistic: 86.67 on 2 and 89 DF,  p-value: < 2.2e-16
```

It is very unlikely that the relationship between height & gender with weight is due to chance. I.e., we would observe an R^2 of .66 or higher under the null < 2.2 times out of 10 quadrillion.

Multiple regression

```
> summary(lm(weight ~ height + sex,data=htwt))

Call:
lm(formula = weight ~ height + sex, data = htwt)

Residuals:
    Min      1Q  Median      3Q     Max
-25.477  -8.594  -1.395   7.992  52.137

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -117.7104    37.5254  -3.137 0.002316 **
height         3.6927     0.5726   6.449 5.69e-09 ***
sex           14.7018     4.2902   3.427 0.000926 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.98 on 89 degrees of freedom
Multiple R-squared: 0.6607,     Adjusted R-squared: 0.6531
F-statistic: 86.67 on 2 and 89 DF,  p-value: < 2.2e-16
```

The max residual is 52.1. This is someone who is much (52.1 lbs) heavier than would be predicted given their gender and height