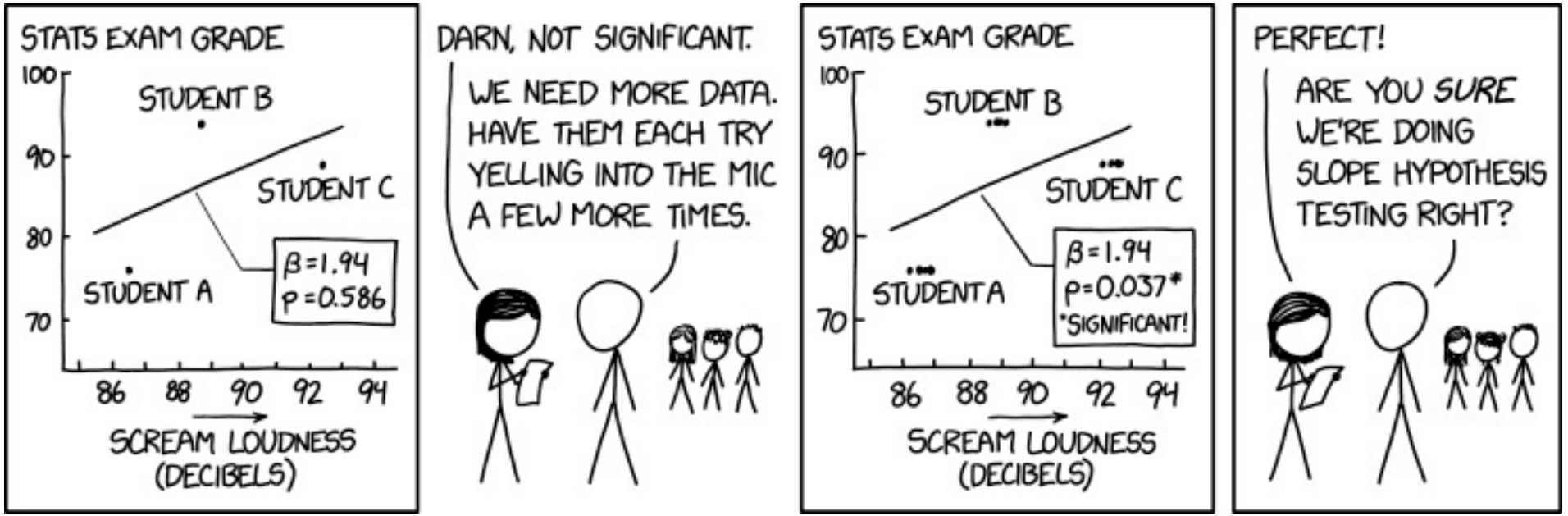
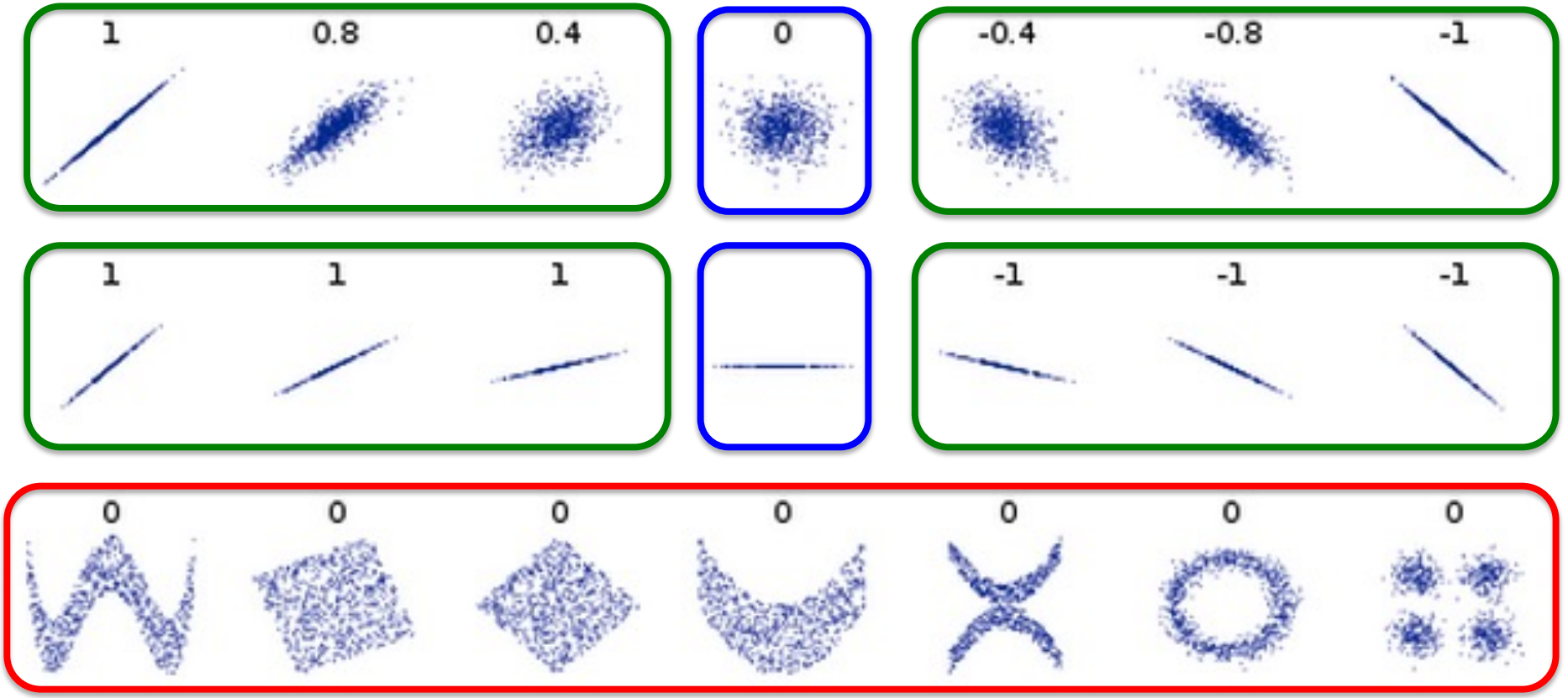
201ab Quantitative methods
L.09: Correlation, regression (2)



Linear relationship.



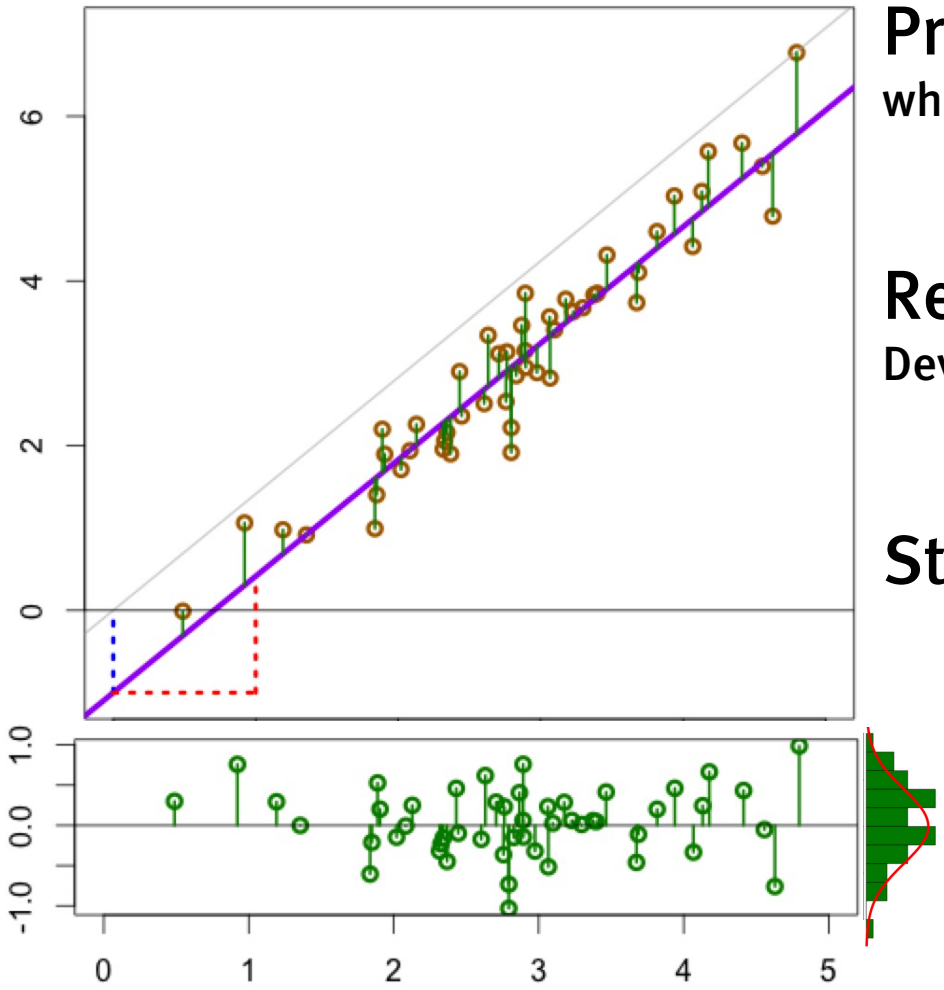
X and Y can be...

- Independent.
- Dependent, but not linearly (tricky to measure in general)
- Linearly dependent (this is what we are measuring)

Ordinary, least-squares regression

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Score on Y for the ith individual = Y Intercept + Slope (*Effect*) × Score on X for the ith individual + Error



Least squares estimates

$$\hat{\beta}_1 = r_{xy}\frac{s_y}{s_x} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$

Prediction (mean of y at each x)

where the estimated line passes at each x value

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Residuals (estimated error)

Deviation of real y value from line prediction

$$\hat{\varepsilon}_i = \left(y_i - \hat{y}_i\right)$$

Standard deviation of residuals

$$\hat{\sigma}_\varepsilon = s_r = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}$$

The sum of squared errors: SS[e]

df=n-2; we fit two parameters (Bo,B1)

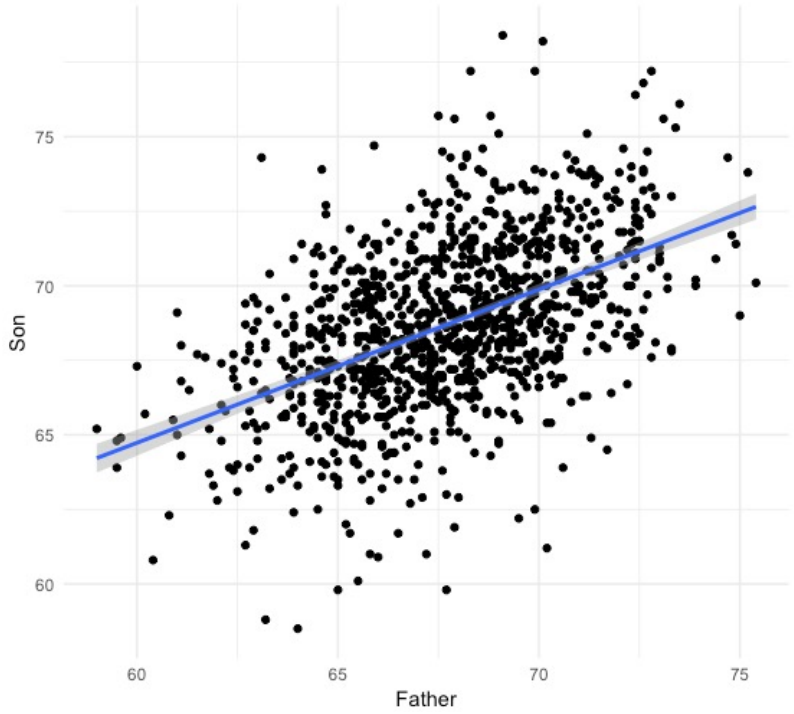
Regression in R

Karl Pearson's data on fathers' and (grown) sons' heights (England, c. 1900)

```
fs = read.csv(url('http://vulstats.ucsd.edu/data/Pearson.csv'))          f = fs$Father;  s = fs$Son
```

```
cor.test(f,s)
```

```
t = 18.997, df = 1076, p-value < 2.2e-16

95 percent confidence interval:
 0.4550726 0.5445746
sample estimates: cor  0.5011627
```




```
anova(lm(data = fs, Son~Father))
```

```
Analysis of Variance Table

Response: Son
            Df Sum Sq Mean Sq F value    Pr(>F)
Father       1 2145.4 2145.35   360.9 < 2.2e-16
Residuals 1076 6396.3    5.94
```

```
cov(f,s)
```

```
3.8733
```

```
cor(f,s)
```

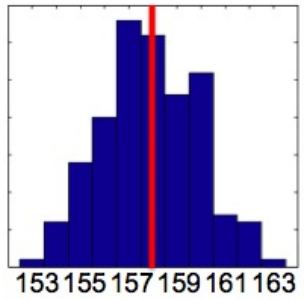
```
0.5011627
```

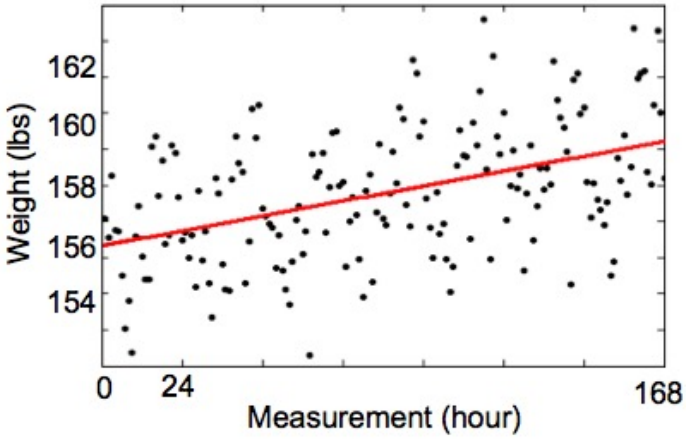
```
summary(lm(data = fs, Son~Father))
```

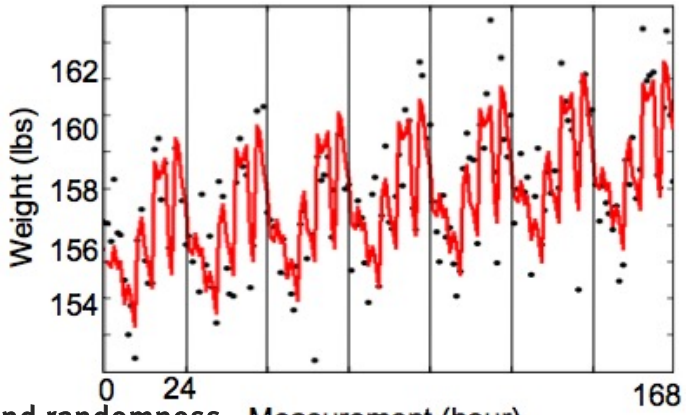
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 33.89280    1.83289   18.49   <2e-16
Father       0.51401    0.02706   19.00   <2e-16

Residual standard error: 2.438 on 1076 degrees of freedom
Multiple R-squared:  0.2512,     Adjusted R-squared:  0.2505
F-statistic: 360.9 on 1 and 1076 DF,  p-value: < 2.2e-16
```

Variation and randomness

- In regression, ANOVA, GLM, etc. we partition variance of an outcome measure into different sources.
  - Our null hypotheses are that a given source contributes zero variance.
  - If a source contributes non-zero variance then we can use it to improve predictions of the outcome.







Regression in R

Karl Pearson's data on fathers' and (grown) sons' heights (England, c. 1900)

```
fs = read.csv(url('http://vulstats.ucsd.edu/data/Pearson.csv'))                f = fs$Father;  s = fs$Son
```

```
summary(lm(data = fs, Son~Father))
```

```
Call:
lm(formula = Son ~ Father, data = fs)

Residuals:
    Min      1Q  Median      3Q     Max
-8.8910 -1.5361 -0.0092  1.6359  8.9894

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 33.89280    1.83289   18.49   <2e-16
Father       0.51401    0.02706   19.00   <2e-16

Residual standard error: 2.438 on 1076 degrees of freedom
Multiple R-squared:  0.2512,     Adjusted R-squared:  0.2505
F-statistic: 360.9 on 1 and 1076 DF,  p-value: < 2.2e-16
```

```
anova(lm(data = fs, Son~Father))
```

```
Analysis of Variance Table

Response: Son
            Df Sum Sq Mean Sq F value    Pr(>F)
Father       1 2145.4 2145.35   360.9 < 2.2e-16
Residuals 1076 6396.3    5.94
```

Where do all these numbers come from? What do they mean?

Sums of squares

Sums of squares are handy for doing calculations by hand (which was the only option when they were developed), because you don't have to divide or take square roots. As we have learned: they are a step along the way to getting sample variance (before we divide by the degrees of freedom).

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Sample variance of X

Degrees of freedom for estimate of variance of X

Sum of squares of X "SS[X]" or "SSX"

$$SS[x] = \sum_{i=1}^{n}(x_i - \bar{x})^2$$

Sums of squares

So, when we are dealing with analyses of sums of squares, just keep in mind that these sums of squares are just measuring variance components (scaled by sample size).

There are many things we can square and sum
(and estimate the variance of)

We are focused on the relationship between the last three:

$$SS[x] = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$SS[y] = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

SS[y] "Sum of squares of y".
Also called "SS total", SST, SSTO, ...

$$SS[e] = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

SS[e] "Sum of squares of the residuals".
Also called "SS error", SSE.

$$SS[\hat{y}_i] = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

SS[y.hat] "Sum of squares of the regression".
Also called "SS regression", SSR, and more.

Sums of squares

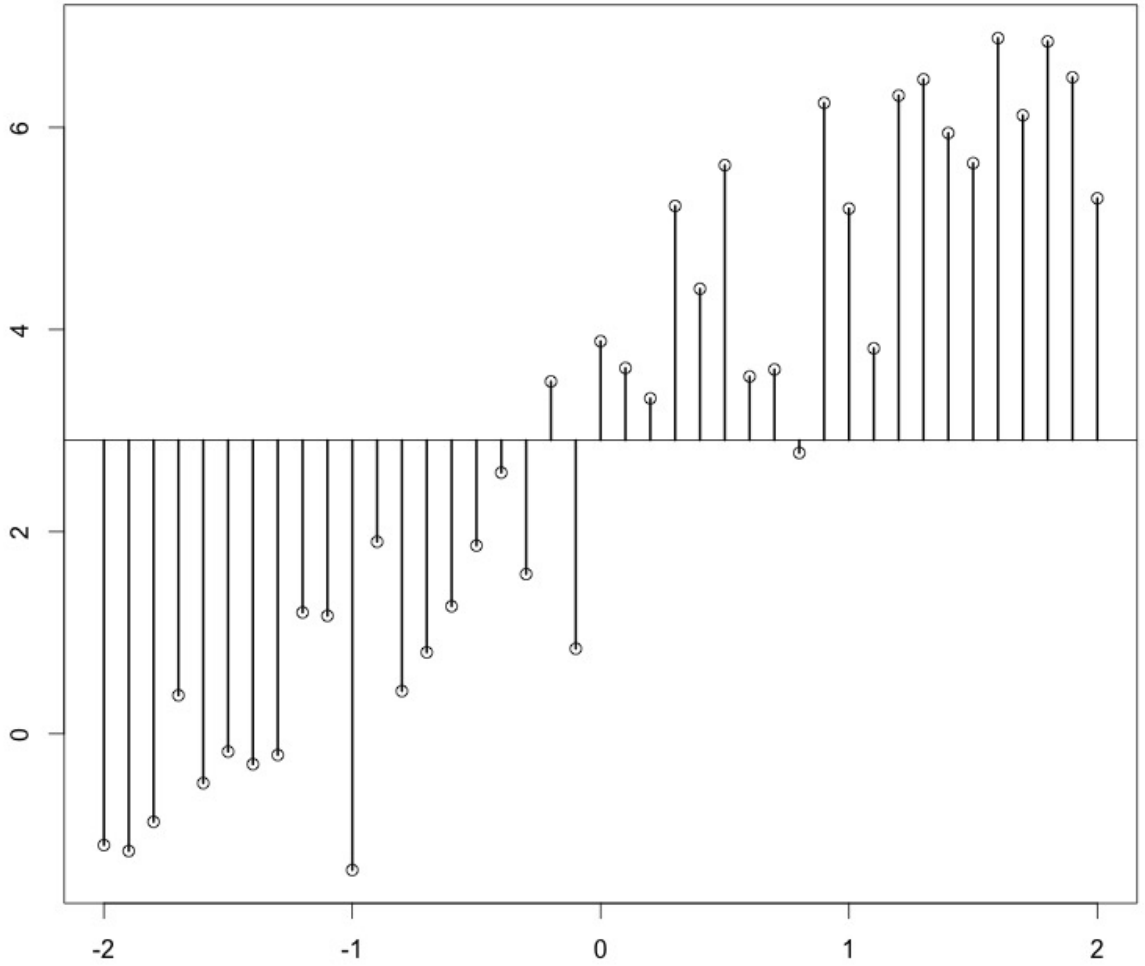
$$SS[y] = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

SS[y] “Sum of squares of y”.
Also called “SS total”, SST, SSTO, …

$$SS[e] = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$SS[\hat{y}_i] = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

“Sum of squares of y”
“Sum of squares total”
The net deviation of the ys from the mean of y



Regression in R

Karl Pearson's data on fathers' and (grown) sons' heights (England, c. 1900)

```
fs = read.csv(url('http://vulstats.ucsd.edu/data/Pearson.csv'))                    f = fs$Father;  s = fs$Son
```

```
summary(lm(data = fs, Son~Father))
```

```
Call:
lm(formula = Son ~ Father, data = fs)

Residuals:
    Min      1Q  Median      3Q     Max
-8.8910 -1.5361 -0.0092  1.6359  8.9894

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 33.89280    1.83289   18.49   <2e-16
Father       0.51401    0.02706   19.00   <2e-16

Residual standard error: 2.438 on 1076 degrees of freedom
Multiple R-squared:  0.2512,     Adjusted R-squared:  0.2505
F-statistic: 360.9 on 1 and 1076 DF,  p-value: < 2.2e-16
```

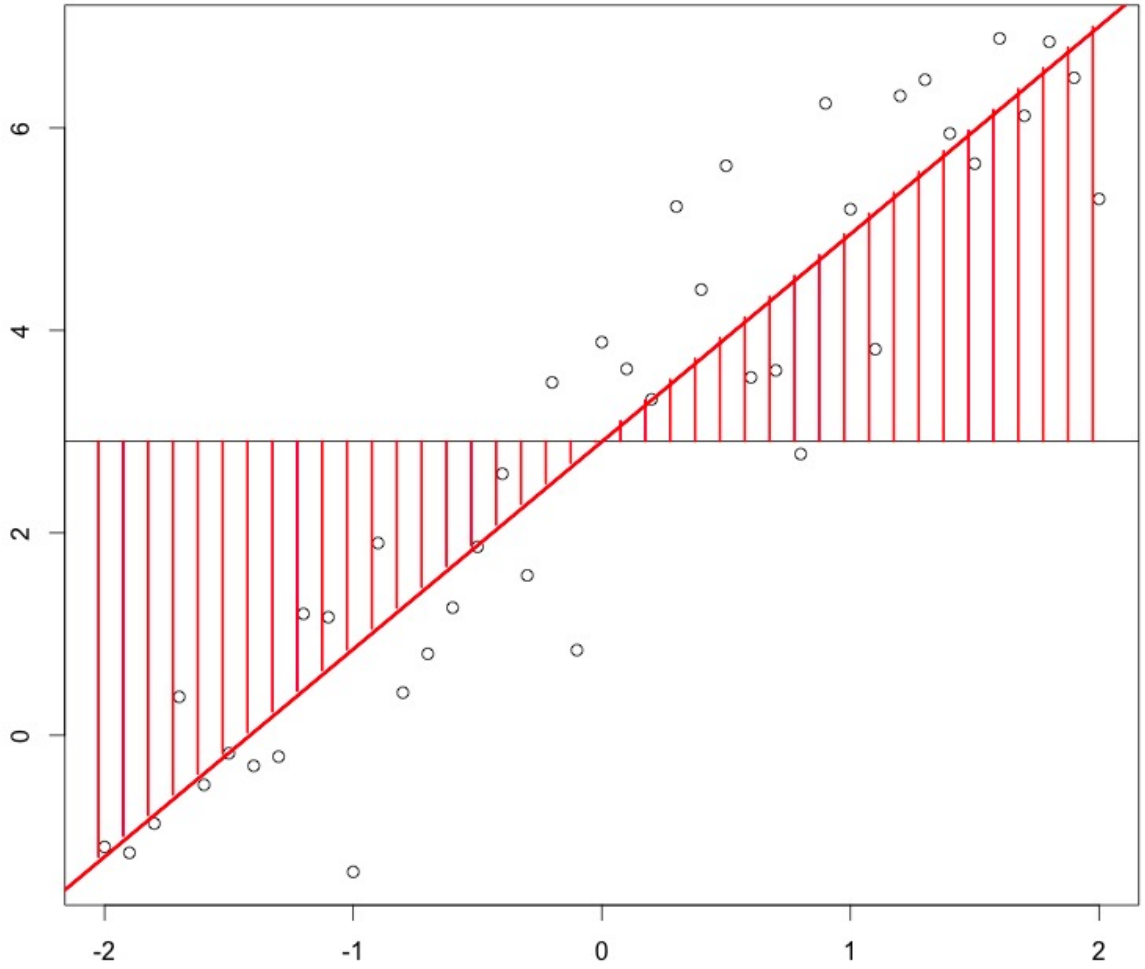
```
anova(lm(data = fs, Son~Father))
```

```
Analysis of Variance Table

Response: Son
            Df Sum Sq Mean Sq F value    Pr(>F)
Father       1 2145.4 2145.35   360.9 < 2.2e-16
Residuals 1076 6396.3    5.94
```

Where do all these numbers come from? What do they mean?

Sums of squares

Sum of squares regression. The net deviation of predicted ys from the mean of y. How much variability is captured by the regression line?



$$SS[y] = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$SS[e] = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$SS[\hat{y}_i] = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

SS[y.hat] “Sum of squares of the regression”. Also called “SS regression”, SSR, and more.

Regression in R

Karl Pearson's data on fathers' and (grown) sons' heights (England, c. 1900)

```
fs = read.csv(url('http://vulstats.ucsd.edu/data/Pearson.csv'))                f = fs$Father;  s = fs$Son
```

```
summary(lm(data = fs, Son~Father))
```

```
Call:
lm(formula = Son ~ Father, data = fs)

Residuals:
    Min      1Q  Median      3Q     Max
-8.8910 -1.5361 -0.0092  1.6359  8.9894

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 33.89280    1.83289   18.49   <2e-16
Father       0.51401    0.02706   19.00   <2e-16

Residual standard error: 2.438 on 1076 degrees of freedom
Multiple R-squared:  0.2512,     Adjusted R-squared:  0.2505
F-statistic: 360.9 on 1 and 1076 DF,  p-value: < 2.2e-16
```

```
anova(lm(data = fs, Son~Father))
```

```
Analysis of Variance Table

Response: Son
            Df Sum Sq Mean Sq F value    Pr(>F)
Father       1 2145.4 2145.35   360.9 < 2.2e-16
Residuals 1076 6396.3    5.94
```

Where do all these numbers come from? What do they mean?

Sums of squares

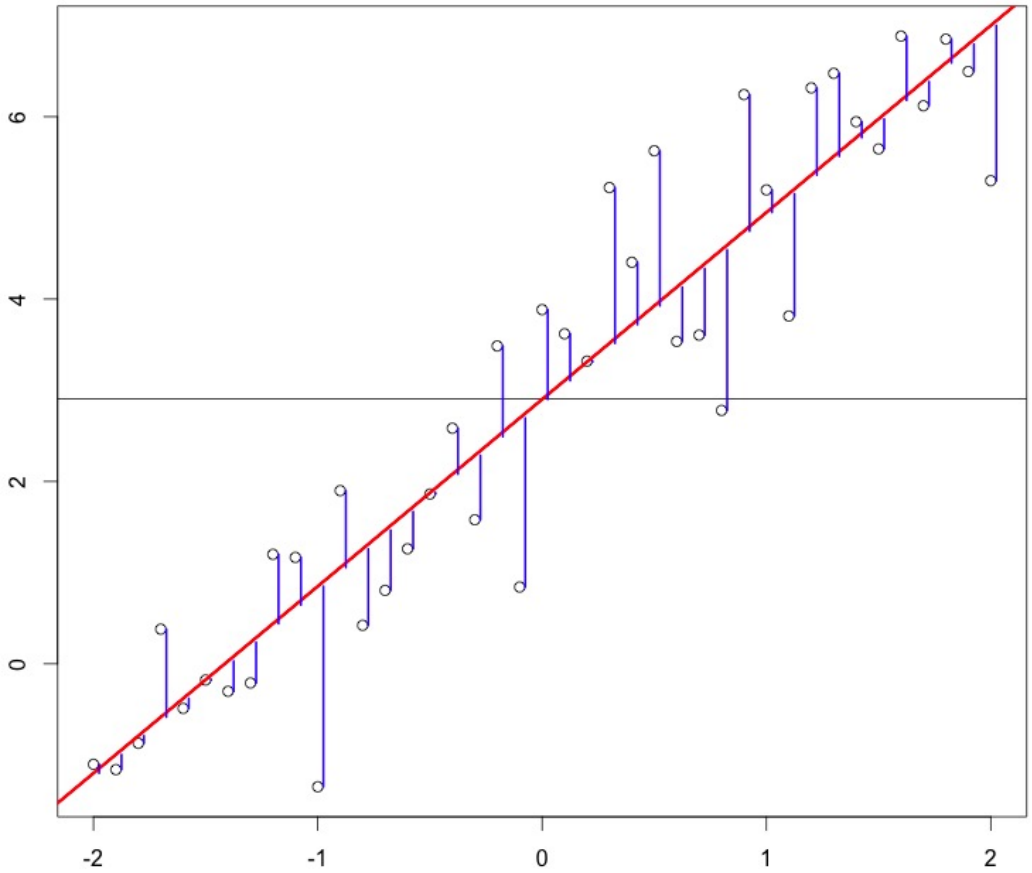
$$SS[y] = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$SS[e] = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

SS[e] "Sum of squares of the residuals". Also called "SS error", SSE.

$$SS[\hat{y}_i] = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

Sum of squares error. The net deviation of real ys from the predicted ys. How much variance is left over in the residuals?

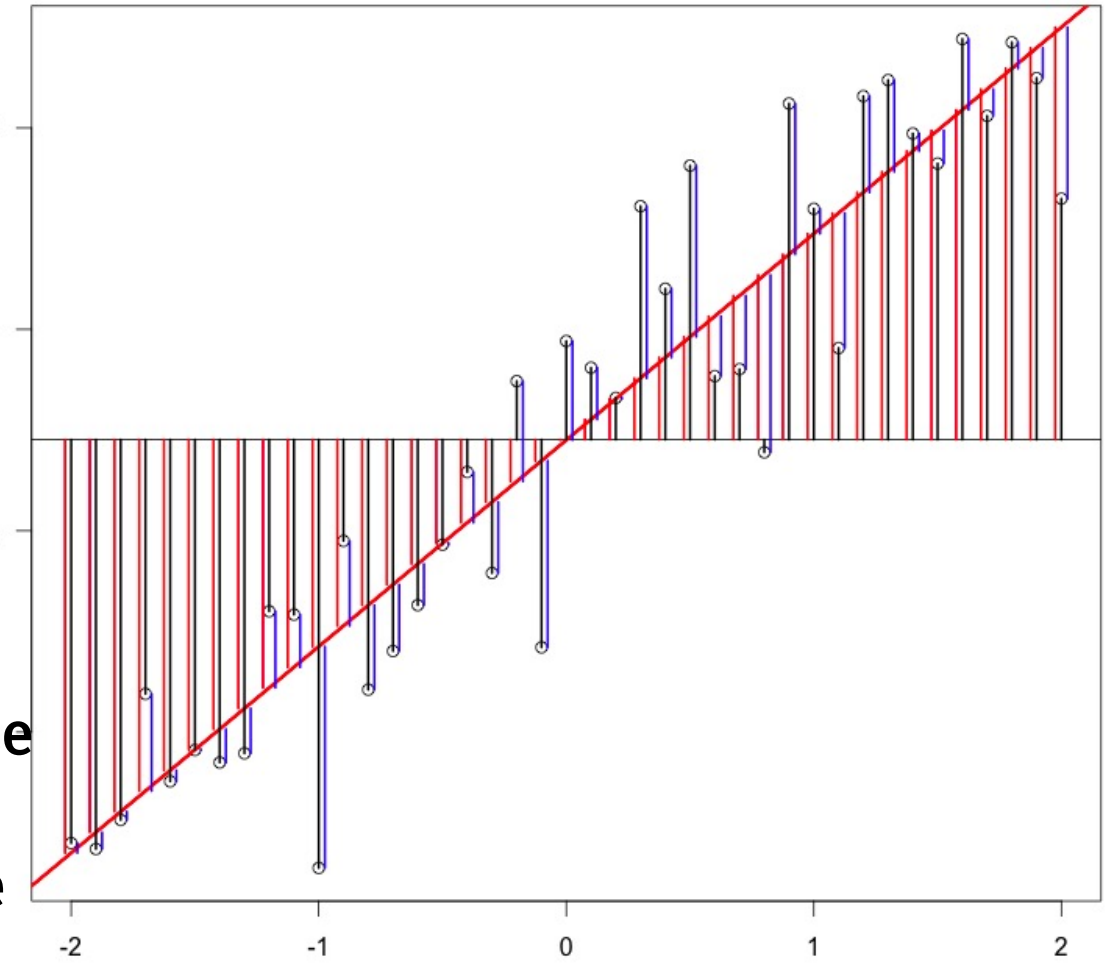


Sums of squares

$$SS[y] = \sum_{i=1}^{n} (y_i - \bar{y})^2$$ SS total

$$SS[e] = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$ SS error

$$SS[\hat{y}_i] = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$ SS regression



The deviation of y from the mean, should be equal to the deviation of the regression line from the mean, plus the deviation of y from the regression line.

$$y_i - \bar{y} = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)$$

Similarly:

SST = SSE+SSR

Regression in R

Karl Pearson's data on fathers' and (grown) sons' heights (England, c. 1900)

```
fs = read.csv(url('http://vulstats.ucsd.edu/data/Pearson.csv'))                    f = fs$Father;  s = fs$Son
```

```
summary(lm(data = fs, Son~Father))
```

```
Call:
lm(formula = Son ~ Father, data = fs)

Residuals:
    Min      1Q  Median      3Q     Max
-8.8910 -1.5361 -0.0092  1.6359  8.9894

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 33.89280    1.83289   18.49   <2e-16
Father       0.51401    0.02706   19.00   <2e-16

Residual standard error: 2.438 on 1076 degrees of freedom
Multiple R-squared:  0.2512,    Adjusted R-squared:  0.2505
F-statistic: 360.9 on 1 and 1076 DF,  p-value: < 2.2e-16
```

```
anova(lm(data = fs, Son~Father))
```

```
Analysis of Variance Table

Response: Son
            Df Sum Sq Mean Sq F value    Pr(>F)
Father       1 2145.4 2145.35   360.9 < 2.2e-16
Residuals 1076 6396.3    5.94
```

Where do all these numbers come from? What do they mean?

Coefficient of determination

$$SS[y] = \sum_{i=1}^{n} (y_i - \bar{y})^2$$ SS total

$$SS[e] = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$ SS error

$$SS[\hat{y}_i] = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$ SS regression

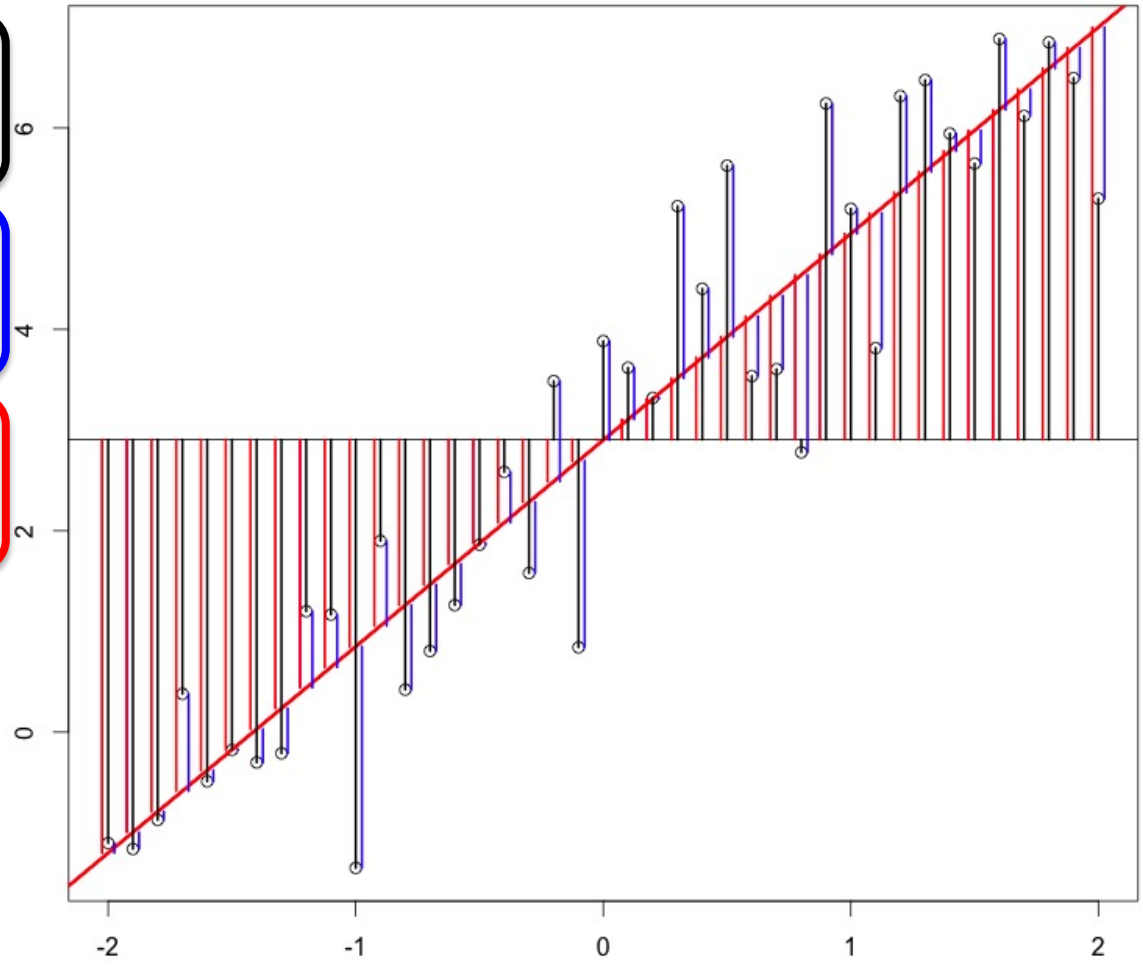
SST = SSE+SSR

So, proportion of total variance accounted for by the regression:

R^2 = SSR / SST

Proportion left to error:

1-R^2 = SSE/SST



(Yes, R^2 is just the correlation coefficient squared in this case.)

Regression in R

Karl Pearson's data on fathers' and (grown) sons' heights (England, c. 1900)

```
fs = read.csv(url('http://vulstats.ucsd.edu/data/Pearson.csv'))                    f = fs$Father;  s = fs$Son
```

```
summary(lm(data = fs, Son~Father))
```

```
Call:
lm(formula = Son ~ Father, data = fs)

Residuals:
    Min      1Q  Median      3Q     Max
-8.8910 -1.5361 -0.0092  1.6359  8.9894

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 33.89280    1.83289   18.49   <2e-16
Father       0.51401    0.02706   19.00   <2e-16

Residual standard error: 2.438 on 1076 degrees of freedom
Multiple R-squared:  0.2512,     Adjusted R-squared:  0.2505
F-statistic: 360.9 on 1 and 1076 DF,  p-value: < 2.2e-16
```

```
anova(lm(data = fs, Son~Father))
```

```
Analysis of Variance Table

Response: Son
             Df Sum Sq Mean Sq F value    Pr(>F)
Father        1 2145.4 2145.35   360.9 < 2.2e-16
Residuals  1076 6396.3    5.94
```

Where do all these numbers come from? What do they mean?

Analysis of variance via Sums of squares

Source of Variation	SS	df	MS
Regression	$SSR = \sum(\hat{Y}_i - \bar{Y})^2$	1	$MSR = \frac{SSR}{1}$
Error	$SSE = \sum(Y_i - \hat{Y}_i)^2$	$n-2$	$MSE = \frac{SSE}{n-2}$
Total	$SSTO = \sum(Y_i - \bar{Y})^2$	$n-1$	
Correction for mean	$SS(\text{correction for mean}) = n\bar{Y}^2$	1	
Total, uncorrected	$SSTOU = \sum Y_i^2$	n	

These are not included in the R anova table, as they are only useful for pedagogical reasons.

Analysis of variance via Sums of squares

$$SS[y] = \sum_{i=1}^{n} (y_i - \bar{y})^2$$ **SS total**

$$SS[e] = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$ **SS error**

$$SS[\hat{y}_i] = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$ **SS regression**

```
SST = sum((sons-mean(sons))^2)
```
```
[1] 8532.581
```

```
SSE = sum((sons-fathers*b1-b0)^2)
```
```
[1] 6388.001
```

```
SSR = sum((fathers*b1+b0-mean(sons))^2)
```
```
[1] 2144.580
```

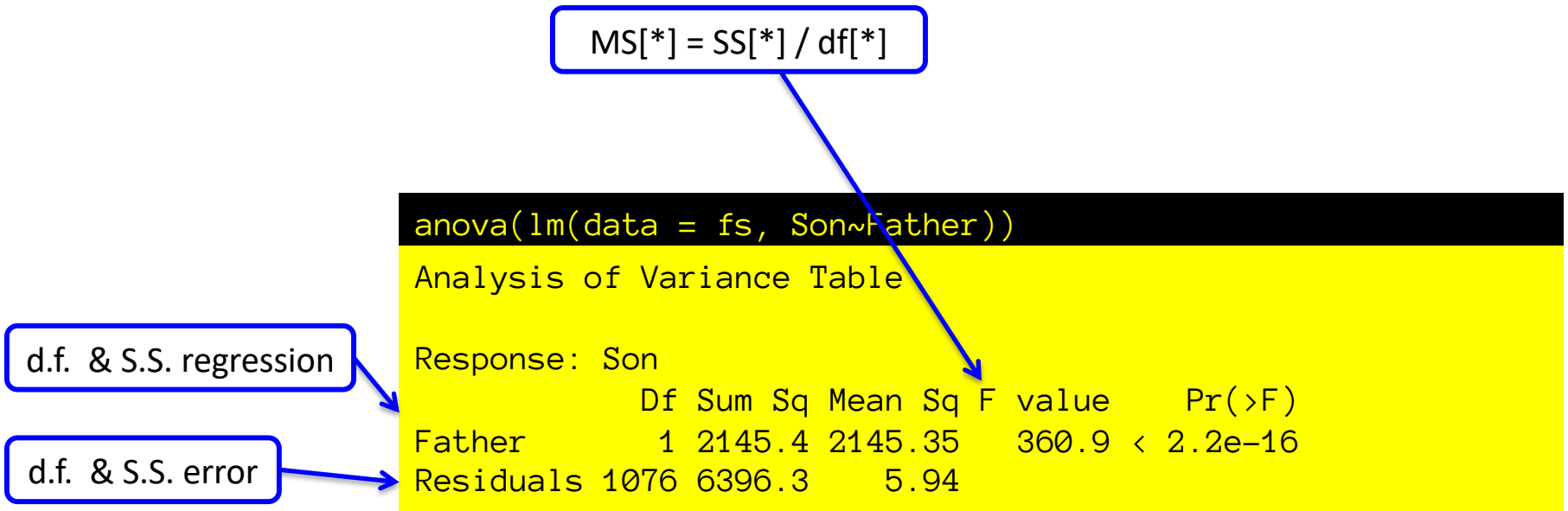
```
anova(lm(sons~fathers))
```
```
Analysis of Variance Table

Response: sons
            Df Sum Sq Mean Sq F value    Pr(>F)
fathers      1 2144.6 2144.58  361.23 < 2.2e-16
Residuals 1076 6388.0    5.94
```

```
SSR+SSE
```
```
[1] 8532.581
```

```
SSR/SST
```
```
[1] 0.2513401
```

```
summary(lm(lm(data = fs, Son~Father))
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 33.89280    1.83289   18.49   <2e-16 ***
Father       0.51401    0.02706   19.00   <2e-16 ***

Residual standard error: 2.438 on 1076 degrees of freedom
Multiple R-squared:  0.2512,   Adjusted R-squared:  0.2505
F-statistic: 360.9 on 1 and 1076 DF,  p-value: < 2.2e-16
```



```
Summary(lm(data = fs, Son~Father))
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 33.89280    1.83289   18.49   <2e-16 ***
Father       0.51401    0.02706   19.00   <2e-16 ***

Residual standard error: 2.438 on 1076 degrees of freedom
Multiple R-squared:  0.2512,   Adjusted R-squared:  0.2505
F-statistic: 360.9 on 1 and 1076 DF,  p-value: < 2.2e-16
```

SSR/(SSR+SSE)

Sd/var Of residuals

```
anova(lm(data = fs, Son~Father))
```

```
Analysis of Variance Table

Response: Son
             Df Sum Sq Mean Sq F value    Pr(>F)
Father        1 2145.4 2145.35   360.9 < 2.2e-16
Residuals  1076 6396.3    5.94
```

Regression in R

Karl Pearson's data on fathers' and (grown) sons' heights (England, c. 1900)

```
fs = read.csv(url('http://vulstats.ucsd.edu/data/Pearson.csv'))                f = fs$Father;  s = fs$Son
```

```
summary(lm(data = fs, Son~Father))
```

```
Call:
lm(formula = Son ~ Father, data = fs)

Residuals:
    Min      1Q  Median      3Q     Max
-8.8910 -1.5361 -0.0092  1.6359  8.9894

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 33.89280    1.83289   18.49   <2e-16
Father       0.51401    0.02706   19.00   <2e-16

Residual standard error: 2.438 on 1076 degrees of freedom
Multiple R-squared:  0.2512,    Adjusted R-squared:  0.2505
F-statistic: 360.9 on 1 and 1076 DF,  p-value: < 2.2e-16
```

F = MSR / MSE

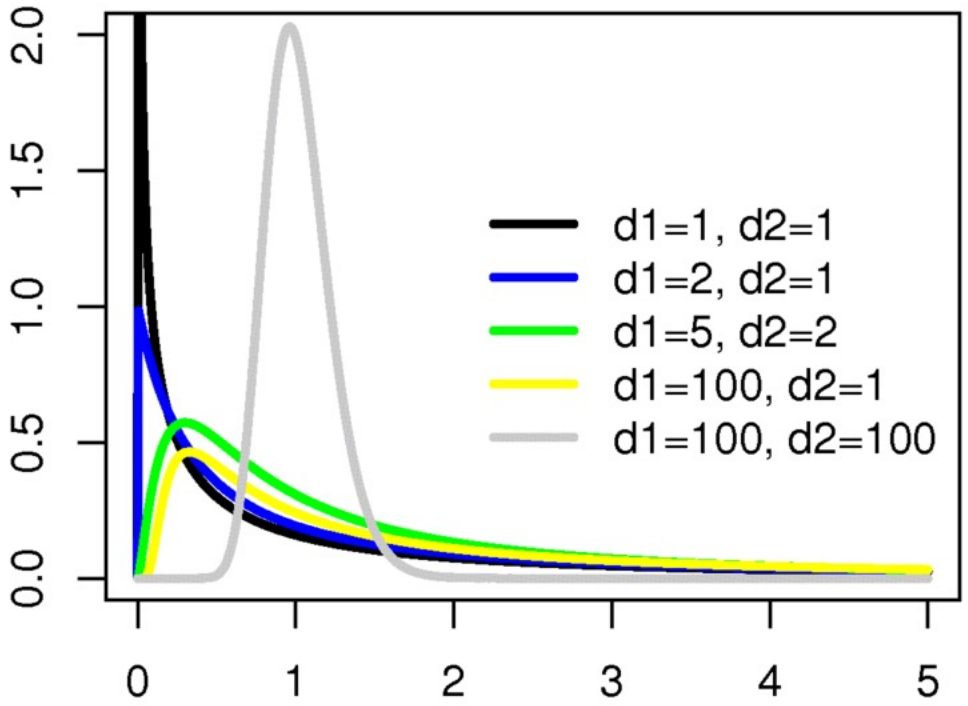
```
anova(lm(data = fs, Son~Father))
```

```
Analysis of Variance Table

Response: Son
            Df Sum Sq Mean Sq F value    Pr(>F)
Father       1 2145.4 2145.35   360.9 < 2.2e-16
Residuals 1076 6396.3    5.94
```

Where do all these numbers come from? What do they mean?

F statistic for OLS regression

$$F = \frac{MSR}{MSE} = \frac{SS[R]}{SS[E]/(n-2)} = \frac{R^2}{(1-R^2)}(n-2)$$



The F-statistic

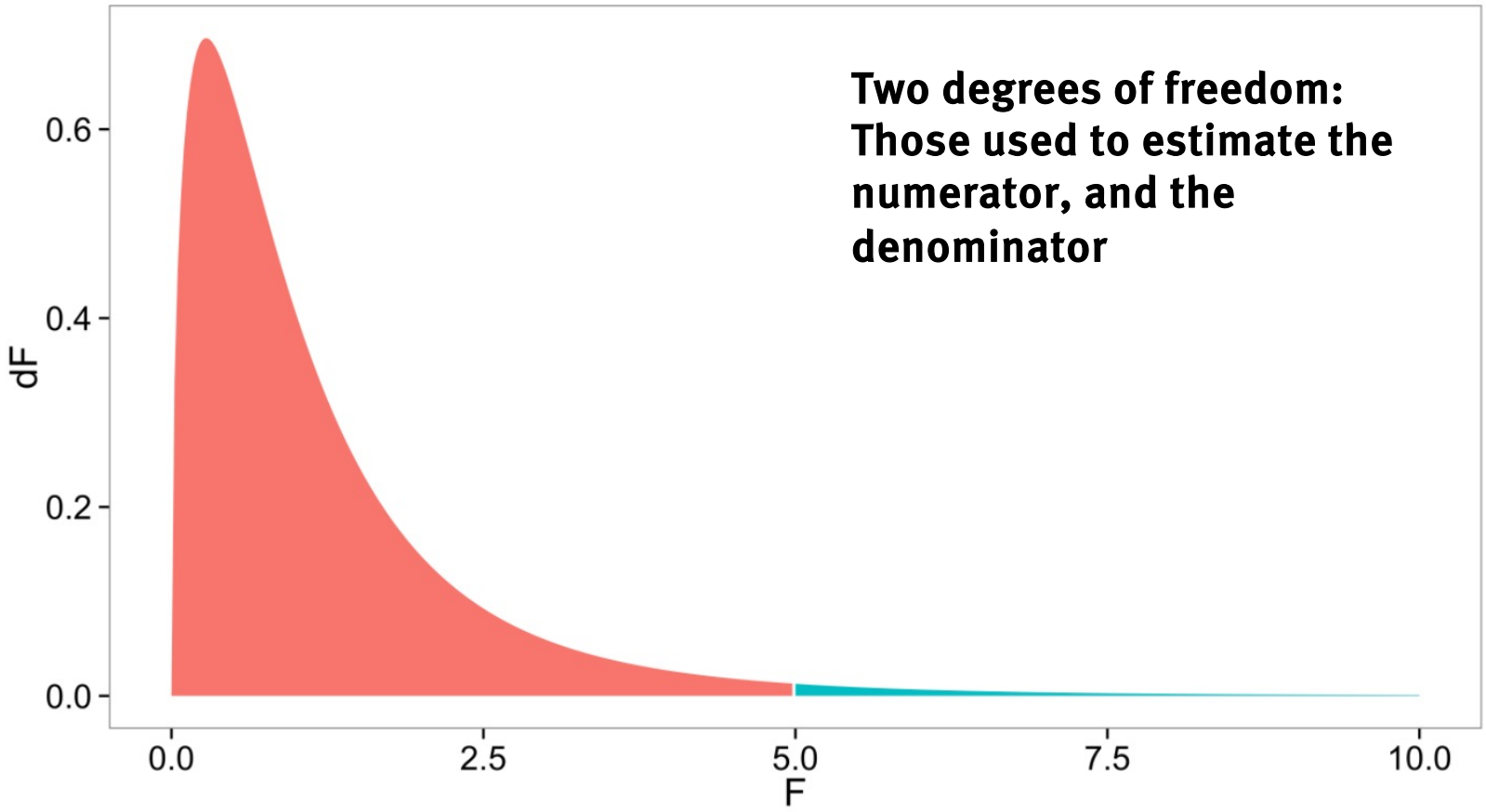
Under Ho: the ratio of two (identical) sample variances estimated with different degrees of freedom.

So, given random variation, even under Ho, we expect the regression to take up *some* variance, and our question is: does it account for more variance than expected by chance?

So, F-test is, like Chi-squared, one tailed (positive tail).

F statistic for OLS regression

$$F = \frac{MSR}{MSE} = \frac{SS[R]/1}{SS[E]/(n-2)} = \frac{R^2}{(1-R^2)}(n-2)$$



Two degrees of freedom: Those used to estimate the numerator, and the denominator

Equivalent tests for bivariate linear relation

T-test for slope

$$t_{b1} = \frac{\hat{\beta}_1}{s\{\hat{\beta}_1\}}$$

T-test for correlation

$$t_r = \frac{\hat{r}}{\sqrt{1-\hat{r}^2}}\sqrt{n-2}$$

F-test for regression

$$F = \frac{MSR}{MSE}$$

Exercise for the algebraically ambitious:
Convince yourself that t{b1}=t{r} and t{r}^2=F

Predicting mean(y)@x vs new y@x

Predicted y values

where the estimated line passes at each x value

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

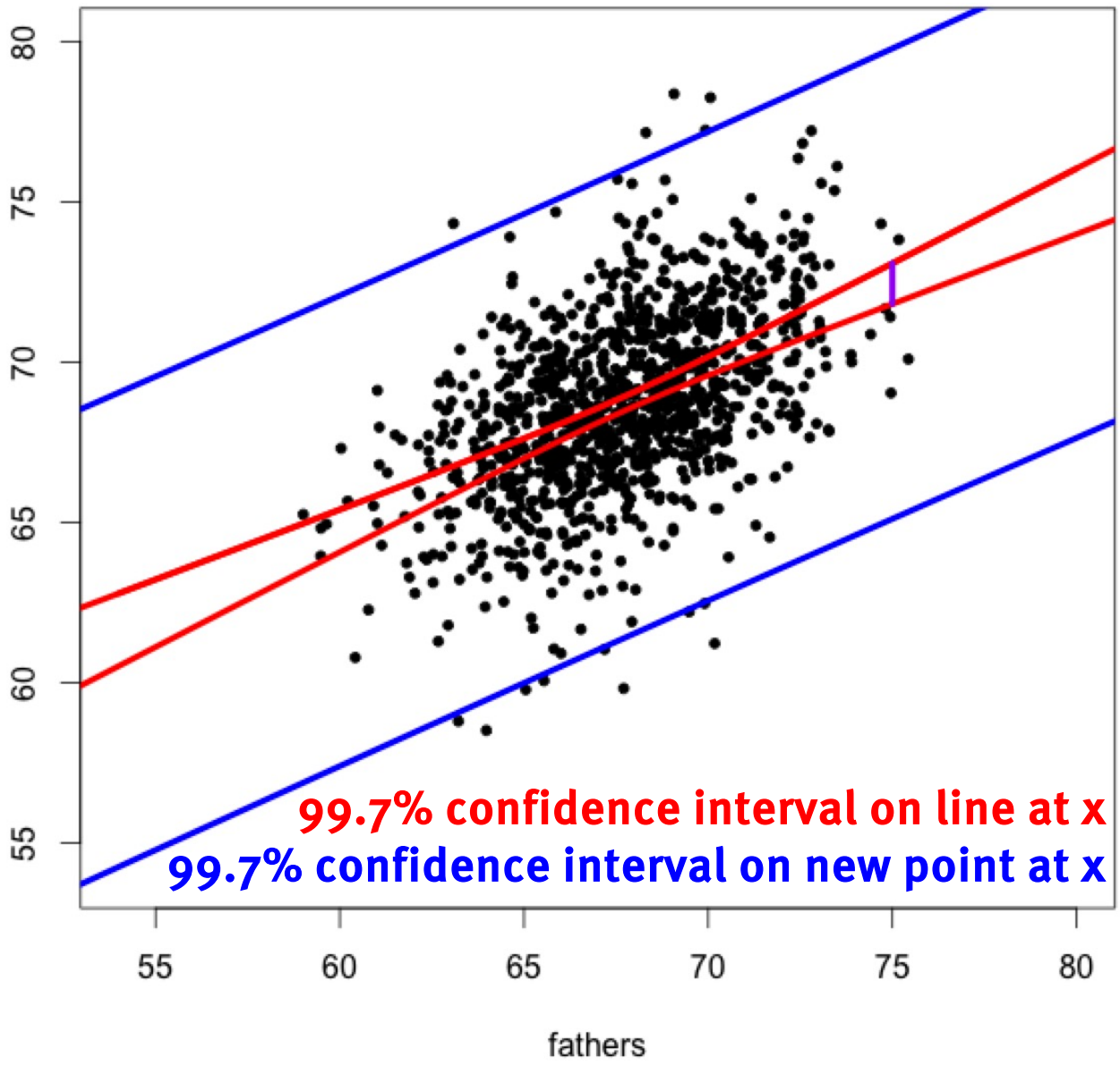
Standard error of predicted y mean

$$s\{\hat{y}_p\} = s_r \sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

sons

Standard error of predicted new y data point

$$s\{\hat{y}_p\} = s_r \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$



Predicting mean(y)@x vs new y@x

Predicted y values

where the estimated line passes at each x value

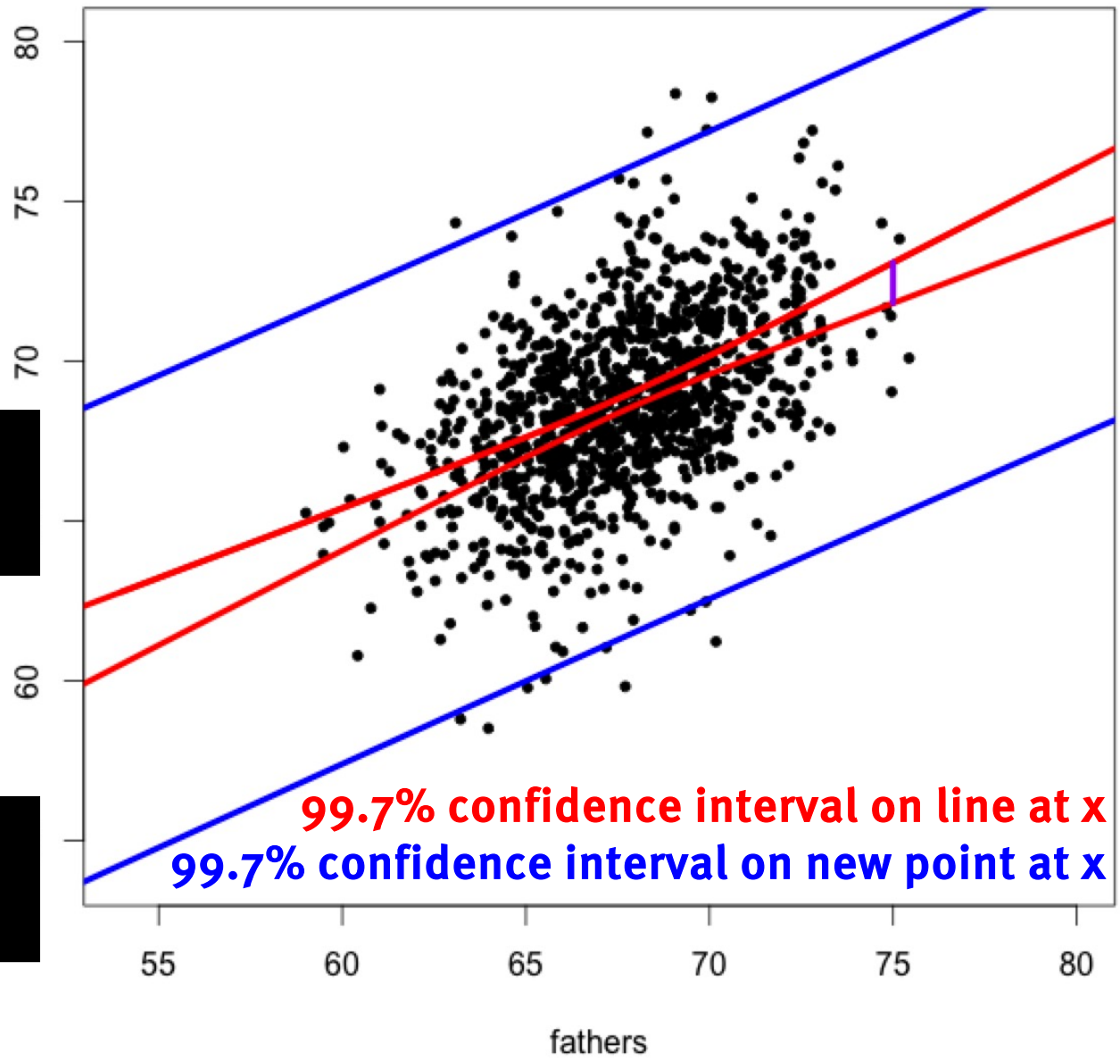
$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Confidence interval on mean(y) at a given x (the line)

```
predict.lm(
      model,
      newdata,
      interval='confidence')
```

Confidence interval on a *new y* at a given x

```
predict.lm(
      model,
      newdata,
      interval='prediction')
```



Regression safety tips.

Assumptions:

(1) Validity: Make sure your measures make sense, and map onto the substantive research questions you have.

(2) Additivity and linearity: The relationship between x and y may not be neatly linear, check scatterplots, residuals! Noise (and, later, other factors) should be additive.

(3) Errors should have equal variance and be normally distributed (could give whacky results if there are some outliers in both x and y – check robustness)

(4) Independence of errors: errors should not be correlated with each other, y, x, etc.

(5) Most error in y, not in x. (parameter estimates biased!)

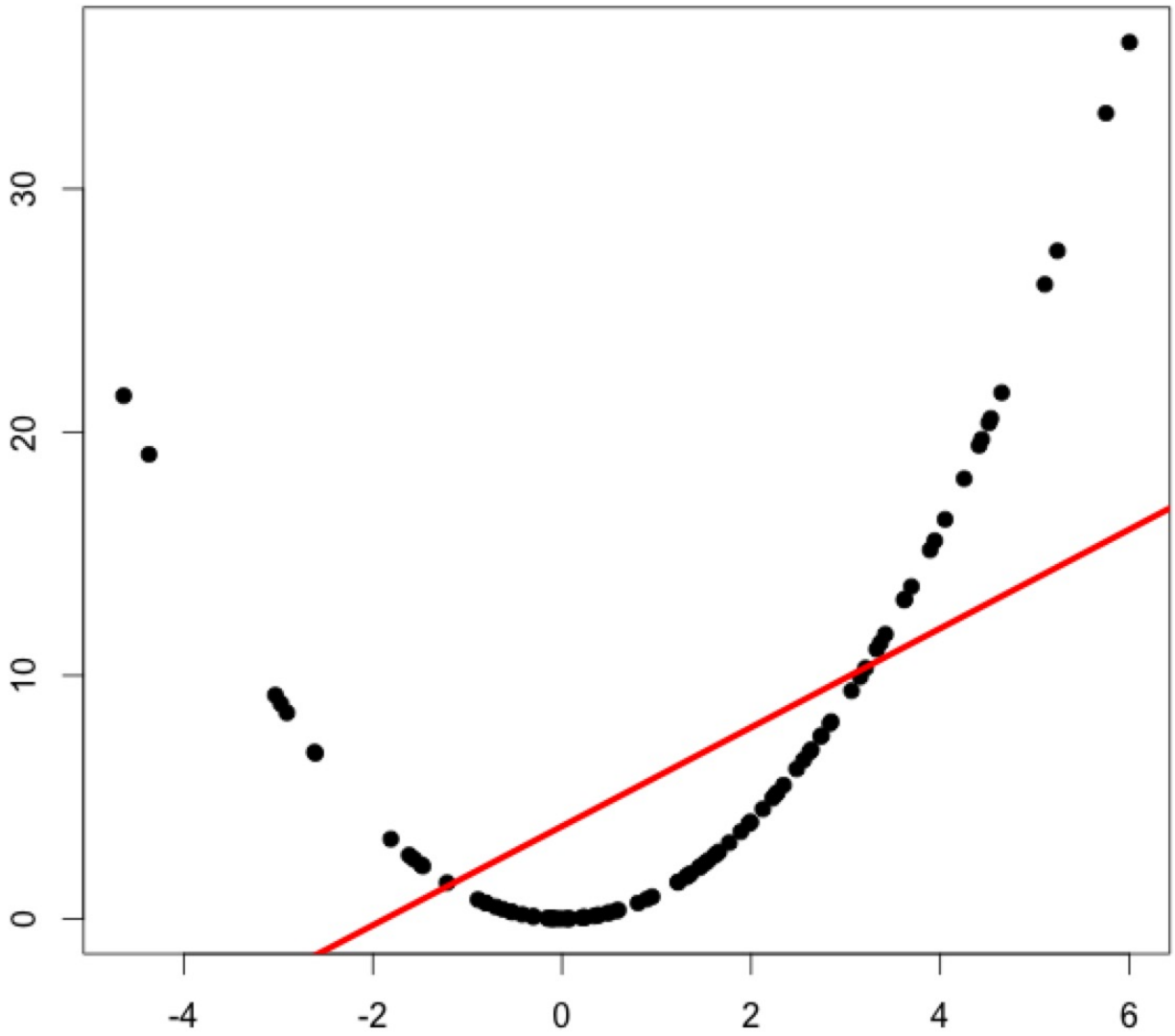
Safety tips:

(1) Don't trust extrapolation.

(2) Check for structure in the residuals.

(3) Be careful with causal interpretations.

Look at the scatterplot!



Regression safety tips.

Assumptions:

(1) Validity: Make sure your measures make sense, and map onto the substantive research questions you have.

(2) Additivity and linearity: The relationship between x and y may not be neatly linear, check scatterplots, residuals! Noise (and, later, other factors) should be additive.

(3) Errors should have equal variance and be normally distributed (could give whacky results if there are some outliers in both x and y – check robustness)

(4) Independence of errors: errors should not be correlated with each other, y, x, etc.

(5) Most error in y, not in x. (parameter estimates biased!)

Safety tips:

(1) Don't trust extrapolation.

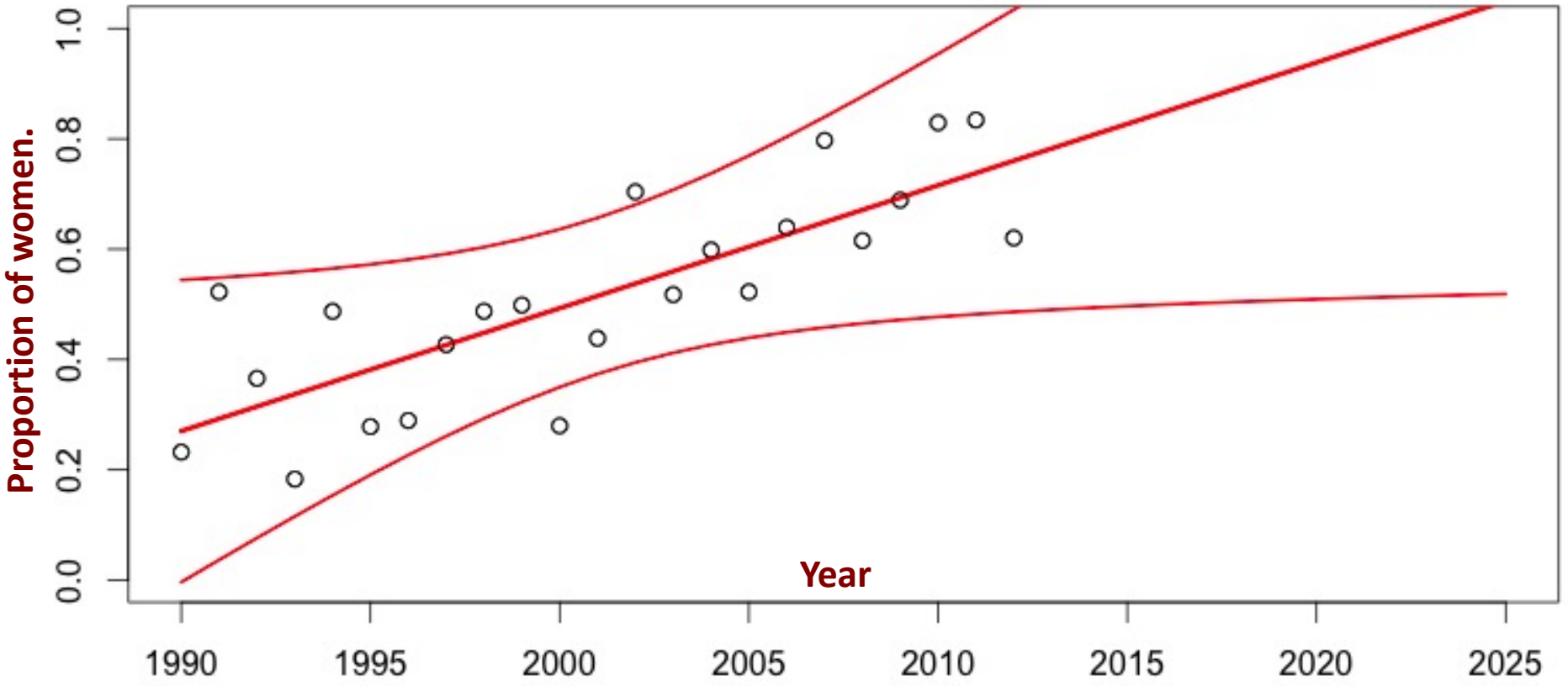
(2) Check for structure in the residuals.

(3) Be careful with causal interpretations.

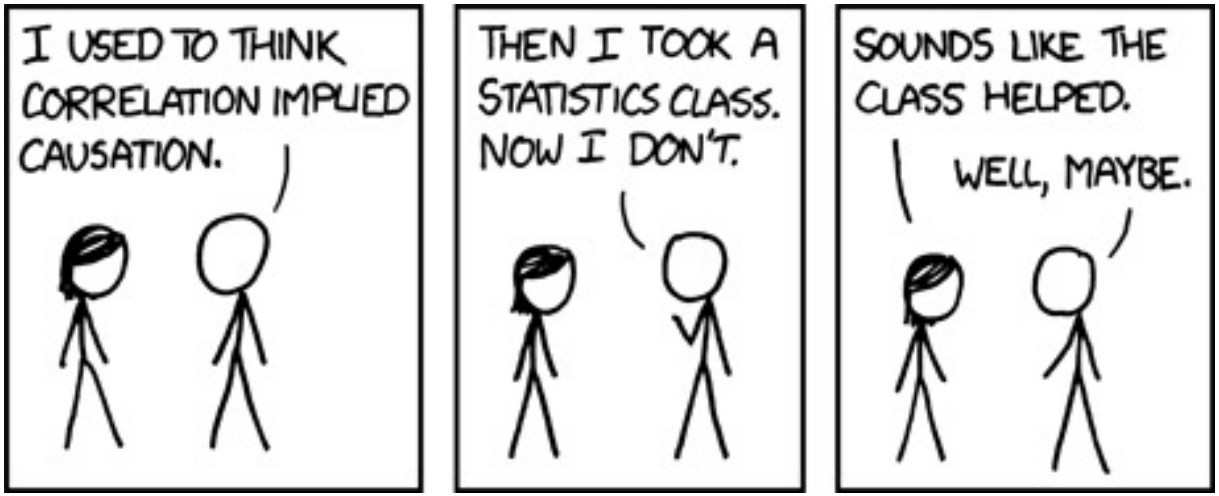
Perils of extrapolation.

(1) The further from the mean of x you extrapolate, the bigger your error!

(2) Relationship might be linear in a small range, but may not be linear forever… (indeed, it might be impossible)



Correlation is not causation



Why not?

Possibility of common or correlated causes, etc.
Correlation / Covariance / Regression line just measure statistical relation.

Intervention needed to ascertain causality
(ideally with random assignment)

Regression safety tips.

Assumptions:

(1) Validity: Make sure your measures make sense, and map onto the substantive research questions you have.

(2) Additivity and linearity: The relationship between x and y may not be neatly linear, check scatterplots, residuals! Noise (and, later, other factors) should be additive.

(3) Errors should have equal variance and be normally distributed (could give whacky results if there are some

(4)

(5)

What should you care about / do?

- Validity!
- Linearity, outliers – look at scatterplots!
- Consider alternate model formulations (more in 201b)

Safety tips:

(1) Don't trust extrapolation.

(2) Check for structure in the residuals.

(3) Be careful with causal interpretations.

```
load(url('http://vulstats.ucsd.edu/data/cal1020.cleaned.Rdata'))
glimpse(cal1020)
```

```
Observations: 3,252
Variables: 13
$ bib        (int) 1205, 9, 13, 15, 1303, 1213, 3, 1055, 12, 1351, 1054, 1216, 1352, 1218, 6, 1220, ...
$ name.first (fctr) Jordan, Macdonard, Sergio, Jamesom, Darren, Okwaro, Steven, Edwin, Lindsey, Dere...
$ name.last  (fctr) Chipangama, Ondara, Reyes, Mora, Brown, Raura, Underwood, Figueroa, Scherf, Brad...
$ City       (fctr) Flagstaff, Grand Prairie, Palmdale, Arroyo Grande, Solana Beach, Oceanside, Enci...
$ State      (fctr) AZ, TX, CA, CA, CA, CA, CA, CA, NY, CA, CA, CA, CA, CA, CA, CA, AZ, ?, CA, CA, C...
$ Division   (fctr) 10 Mile Overall, 10 Mile Overall, 10 Mile Overall, 10 Mile Overall, 10 Mile Over...
$ Age        (dbl) 25, 29, 32, 30, 28, 39, 26, 42, 27, 33, 60, 34, 33, 39, 26, 32, 41, 24, 42, 48, 5...
$ Zip        (fctr) 86004, 75054, 93551, 93420, 92075, 92057, 92024, 90040, 12440, 92024, 91016, 920...
$ time.sec   (dbl) 2880, 2885, 2970, 3062, 3083, 3206, 3222, 3241, 3289, 3318, 3320, 3363, 3388, 341...
$ corral     (fctr) 1, 0, 0, 0, 1, 1, 0, 1, 0, 1, 1, 1, 1, 1, 0, 1, 1, 0, 1, 1, 1, 0, 1, 0, 1, 1, 0,...
$ wheelchair (lgl) FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALS...
$ pace.sec   (dbl) 288.0, 288.5, 297.0, 306.2, 308.3, 320.6, 322.2, 324.1, 328.9, 331.8, 332.0, 336....
$ speed.mph  (dbl) 12.500000, 12.478336, 12.121212, 11.757022, 11.676938, 11.228946, 11.173184, 11.1...
```

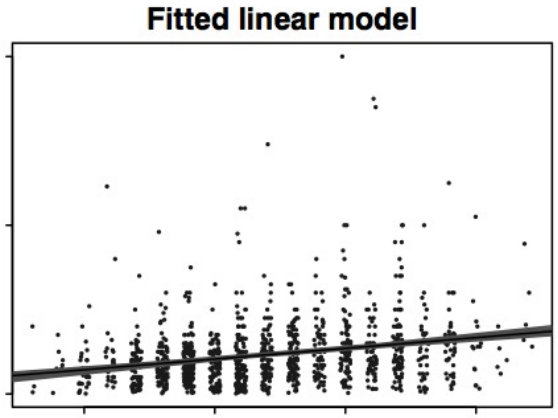
- What is the correlation, covariance, regression slope of speed ~ Age, speed ~ corral (as numeric). Significant?
- Find 95% confidence interval on the mean speed of 60 yo.s … on the speed of a single 60 yo
- Is anything worrisome about the speed ~ age regression?
- What happens if you do speed ~ sex ?
  How does it relate to a t-test comparing male/female speed?
- Make a plot of the speed-age relationship for difft corrals.
  Use facet_wrap and geom_smooth(method='lm').

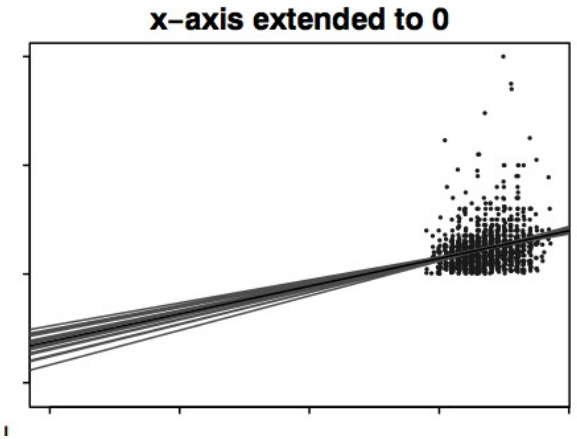
Why transform predictors?

earnings = −61000 + 51 · height (in millimeters) + error

earnings = −61000 + 81000000 · height (in miles) + error

- A few things here:
  - -$61000 is meaningless: income of person of height zero





Why transform predictors?

```
earnings = -61000 + 51 · height (in millimeters) + error
earnings = -61000 + 81000000 · height (in miles) + error
```

- A few things here:
  - -$61000 is meaningless: income of person of height zero

Center the predictor:

```
height.c = (height - mean(height))
```
(in mm or miles)

we get:

```
earnings = $27128+ $51 · height.c (in millimeters) + error
earnings = $27128+ $81000000 · height.c (in miles) + error
```

The intercept, $27128, now means:

earnings of a person of average height.

Why transform predictors?

```
earnings = $27128 + $51·height.c (in millimeters) + error
earnings = $27128 + $81000000·height.c (in miles) + error
```

- A few things here:
  - Slope of $51/height seems trivial, $81,000,000 huge.
    (but really they are the same: $51/mm = $81M/mile:
    1 mile = 1609344mm, $51 * 1609344 = 81000000)

We can ascertain the relative importance of predictors by multiplying the slope by the standard deviation of the predictor, to see how much influence they have:

sd(height) = 3.8 inches = 97 mm = 0.000061 miles.

51 $/mm * 97 mm = 81000000 $/mile * 0.000061 miles = **$4950**

4950 $ / sd(height) ‹- this is more useful!

Why transform predictors?

```
earnings = $27128 + $51·height.c (in millimeters) + error
earnings = $27128 + $81000000·height.c (in miles) + error
```

- A few things here:
  - Slope of $51/height seems trivial, $81,000,000 huge.

4950 $ / sd(height) <- this is more useful!

We can get this from the start by using z-score of height

```
z.height = ( height - mean(height) ) / sd(height)
earnings = $27128 + $4950 * z.height + error
```

But $/sd(height) is not a particularly intuitive measure of slope – we think of height in particular units

Why transform predictors?

earnings = $27128 + $51·height.c (in millimeters) + error

earnings = $27128 + $81000000·height.c (in miles) + error

- A few things here:
  - Slope of $51/height seems trivial, $81,000,000 huge.

Slopes: $51/mm, $510/cm, $1300/inch, $15600/ft, $51000/mile

Variation in heights on the order of inches (~4), or centimeters (~10), so those are better denominator units.

earnings = $27128 + $1300 height (inches) + error

earnings = $27128 + $510 height (cm) + error

Why transform predictors?

~~earnings = −61000 + 51 · height (in millimeters) + error~~

earnings = $27128 + $510 height (cm) + error

- A few things here:
  - -$61000 is meaningless: income of person of height zero
  - Slope of $51/height seems trivial, $81,000,000 huge.

We transform variables to get the coefficients and intercepts to be more interpretable: results don't change, but some units are more sensible than others.

Transforming response variables...

...To make coefficients more interpretable

```
earnings ($1) = $27128 + $1300 height (inches) + error
Earnings ($1000) = $27 + $1.3 height (inches) + error
```

If we predict (earnings/$1000), then our slope and intercept are of a more manageable magnitude.

This seems like the best setup for this regression, but other candidates are also reasonable.

Linearly transforming variables.

- When linearly transforming variables:
  $X' = aX + b$
  - the regression does not change:
    the same fit,
    the same correlation,
    etc.
  - But, it is gives us more interpretable coefficients

- We could always transform the coefficients ourselves after the fact, but it is easier to just set up the regression intuitively ahead of time.

Linearly transforming variables: w’ = a*w + b

- Centering: X’ = X-mean(X)
  makes the intercept mean: Y value at average X

Linearly transforming variables: w' = a*w + b

- Centering: X' = X-mean(X)
  makes the intercept mean: Y value at average X
- Z scoring ("standardizing"): X' = (X-mean(X))/sd(X)
  also makes the slope mean: change in Y/sd change in X
  this is gives a clearer sense of the importance of X
  useful for arbitrary scales of X (like personality score)
  less useful for real, physical quantities (e.g., height)

Linearly transforming variables: w' = a*w + b

- Centering: X' = X-mean(X)
  makes the intercept mean: Y value at average X
- Z scoring: X' = (X-mean(X))/sd(X)
  also makes the slope mean: change in X/sd change in Y
- Picking units of X (mm, cm, m, inches, feet, miles):
  use real units when you have a "real" measurement, but pick unit magnitude so units are of the same order of magnitude as the sd of X.
  You then get the best of both worlds: slope in terms of real units, and slope that gives a good sense of the importance of the predictor.

Linearly transforming variables: w' = a*w + b

- Centering: X' = X-mean(X)
  makes the intercept mean: Y value at average X
- Z scoring: X' = (X-mean(X))/sd(X)
  also makes the slope mean: change in X/sd change in Y
- Pick real units of X that are of the same order of magnitude as the sd of X.
- Scale dependent variable (Y' = Y*k)
  to make the numerical values of slope and intercept be of a more manageable magnitude

There will be some tradeoffs, and there isn't one 'right' answer (depends on question!) but a bit of scale/unit optimization will help a lot.

Making new variables

- Often it is useful to make new variables out of other variables, because we expect these derived quantities to behave more lawfully.
  - From city population and area, we can get population density.
  - From # of murders and population, we can get murder rate.
  - From hit rate and false alarm rate, we can calculate d' = qnorm(hit.proportion) – qnorm(miss.proportion)
  - From errors and RTs we can estimate 'evidence accumulation rate' and 'decision criterion'.
  - If we have mother's height and father's height, we can get average parents' height, and father-mother height difference
- The goal here is to find variables that behave nicely: are predictable, less susceptible to extraneous influence, are uncorrelated with each other, etc.

Linear transformation practice.

1) We find that Bo = 0; B1 = 0.1 in:
   z.extraversion ~ (height.in – mean(height))*B1 + Bo
   How do we expect extraversion to differ between a 5'9" and a 6'0" person?
2) We are trying to predict newborn weight based on the weights of the mother and the father.
   How would you set up this regression?
3) We find: gre.percentile ~ (income.percentile)*0.5-0.4
   What is wrong with extrapolation of this regression line?
4) We find: z.rt ~ –0.4*(z.iq).
   Mean(rt) = 400, sd(rt) = 150; mean(iq)=102; sd(iq)=14
   What is the predicted RT of someone with an IQ of 106?
5) We find: fat.percentage = 17 + 3800*(weight.lb / height.in^3)
   (weight.lb / height.in^3): mean = 0.0005. sd=0.0005
   What's a better way to have set up this regression?