<<Lect16_SS_Regression.pdf>>

Birge 145
5:05 – 7:05pm

Stat 324 –
Introduction to Statistics for Engineers

LECTURE 16: REGRESSION WITH 1 QUANTITATIVE VARIABLE;

SECTIONS 11.1-11.5 IN OTT AND LONGDECKER

Relating two Quantitative Variables

Sir Francis Galton (1822-1911) was interested in how children resemble their parents. One simple measure of this is height. So Galton (actually his disciple, Karl Pearson) measured the heights of father son pairs (in inches) at maturity. In the actual study, 1078 pairs were measured. For convenience, we will use a small subsample of n= 14 pairs:

Family	Father's Height	Son's Height
1	71.3	68.9
2	65.5	67.5
3	65.9	65.4
4	68.6	68.2
5	71.4	71.5
6	68.4	67.6
7	65.0	65.0
8	66.3	67.0
9	68.0	65.3
10	67.3	65.5
11	67.0	69.8
12	69.3	70.9
13	70.1	68.9
14	66.9	70.2

Relating two Quantitative Variables cont.

Often two (bivariate) or more variables (multivariate) are observed for each experimental unit in order to determine:

1. Whether the variables are related.
2. How strong the relationships appear to be
3. Whether one variable of primary interest can be predicted from observations on the others.

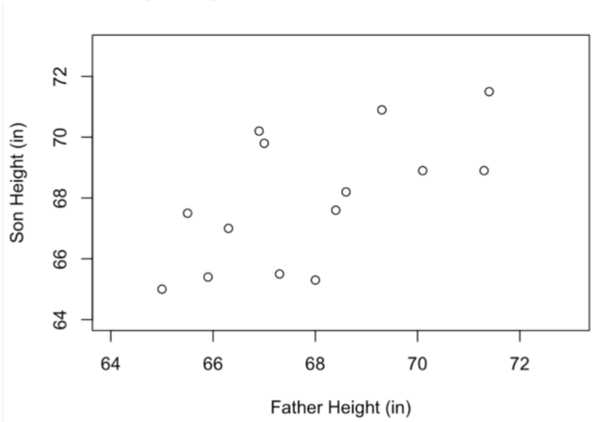
In ANOVA problems, the treatment was recorded, as was the measurement. (1 Quant and 1 Cat).

When there appeared to be a difference in the means of the groups, we used a model that allowed for different group means.

Often, when plotting bivariate quantitative data we see trends between the data, and we try to model those trends with linear, quadratic, exponential, sinusoidal, functions.

Relating two Quantitative Variables cont.

Plotting the bivariate data in a **scatter plot** is the first step in understanding what type of relationship may exist between the two variables.



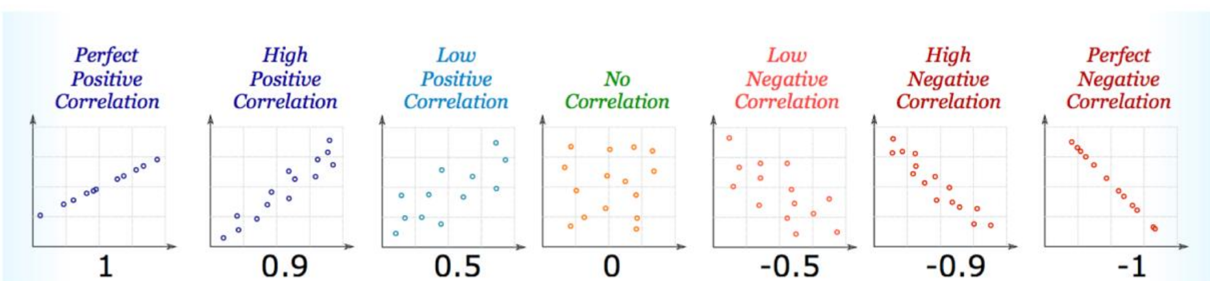
Family	Father's Height	Son's Height
1	71.3	68.9
2	65.5	67.5
3	65.9	65.4
4	68.6	68.2
5	71.4	71.5
6	68.4	67.6
7	65.0	65.0
8	66.3	67.0
9	68.0	65.3
10	67.3	65.5
11	67.0	69.8
12	69.3	70.9
13	70.1	68.9
14	66.9	70.2

From the graph, we can see a shape, direction and strength of the relationship:

Pearson's Sample Correlation

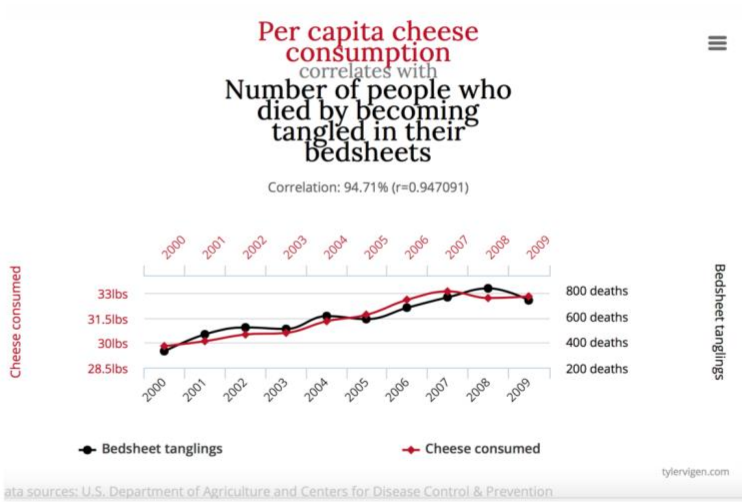
There is a mathematical computation for the
"closeness to a straight line" which is applicable when the shape of the cloud has a linear trend.

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$



Spurious Correlation

http://www.tylervigen.com/spurious-correlations



Sample Correlation for Father, Son Height Data

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} = \frac{1}{n-1}\sum\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

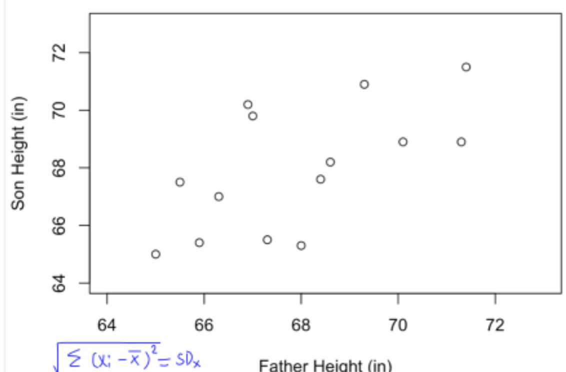
$$= \frac{35.40857}{\sqrt{54.249 * 61.104}}$$

$$r = 0.615$$

$$\sum(x_i - \bar{x})(y_i - \bar{y}) = 35.40857$$

$sd_X = 2.043,$ $sd_Y = 2.17$

$n_X = 14$ $n_Y = 14$



$$\frac{\sqrt{\sum(x_i - \bar{x})^2}}{\sqrt{n-1}} = SD_x$$

$$SD_x^2 * (14-1) = 2043^2 * 13 = 26 = \sum(x_i - \bar{x})^2$$

$$SD_y^2 * (14-1) = 2.17^2 * 13 = 61.22 = \sum(y_i - \bar{y})^2$$

Modeling (x,y) values

A regression line describes how dependent variable y changes as the variable x changes

from a data set $(x_1, y_1), (x_2, y_2) \dots (x_n, y_n)$

Because we see a linear pattern, we can express our model as:

$$Son's\ Height = \beta_0 + \beta_1 * Father's\ Height + Random\ Error$$

$$y_i = \beta_0 + \beta_1 * x_i + \epsilon_i$$

Where

y_i is the height of son i, x_i is the height of father i

β_0: average value of sons' height when fathers height is zero

β_1: average amount of change in sons' height for one inch change in father's height

ϵ_i is the error of individual obs from group mean $\epsilon \sim N(0, \sigma^2)$

Modeling (x,y) values: Linear Regression with LSE

Our goal is to estimate β_0 and β_1 such that the line is as close to all of the values on average

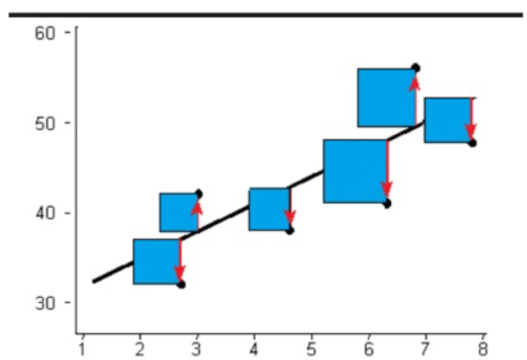
If $\widehat{y_i} = \widehat{\beta_0} + \widehat{\beta_1} * x_i$ is the equation of the regression line,

each value has a "vertical error" of $e_i = \mathrm{y_i} - \widehat{y_i}$ "Residule".

The Least Squares Regression Line minimizes the

Sum of squared [vertical] errors:

$SSE = \sum e_i^2 = \sum (y_i - \widehat{y_i})^2 = \sum\left[\left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 * x_i)\right)^2\right]$



*This SSE is similar to that calculated in ANOVA

Modeling (x,y) values: Linear Regression with LSE

The Least Squares Regression Line minimizes the sum of squared [vertical] errors: (Using some Calculus....)

x: Fathers height

y: Sons height

When Slope is estimated to be:

$$\widehat{\beta_1} = \frac{\sum_{i=1}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}(x_i-\bar{x})^2} = \frac{r s_y}{s_x}$$

```
> mean(father); sd(father)
[1] 67.92857
[1] 2.042784
> mean(son); sd(son)
[1] 67.97857
[1] 2.168012
> cor(father, son)
[1] 0.6150083
```

$$\widehat{\beta_1} = \frac{35.40857}{54.24857} = \frac{0.6150083 * 2.168012}{2.042784}$$

$\hat{\beta}_1 = 0.653$

And Y Intercept: $\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\,\bar{x}$ (or solve using y=mx+b0)

$y = mx + b$

$\widehat{\beta_0} = 67.98 - 0.653(67.93) = 23.62$

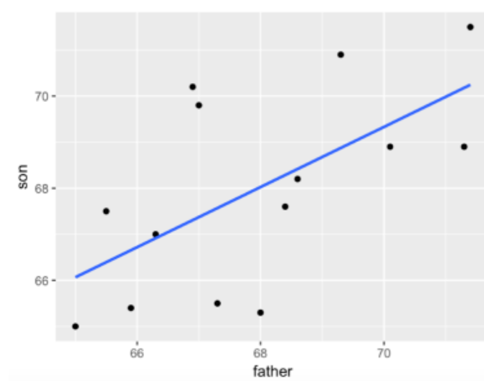
$67.98 = 0.653(67.93) + b$

So LSRL: $\widehat{Son's\ Height} = 23.62 + 0.653(\text{Fathers Height})$

Modeling (x,y) values: Linear Regression with LSE

So LSRL: $\widehat{Son's\ Height} = 23.62 + 0.653(Father's\ Height)$

(notice, the model's predicted son's height is giving the average son's height for all father's with a given height)



What is the residual for the father with height 71.3?
Identify the value and the residual on the graph.

residule = obs − EV = $y_i - \hat{y}_i = 68.9 - 70.18 = -1.28$

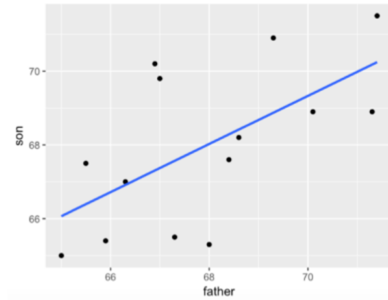
observed sons height 68.9

Expected sons height ; $\hat{y} = 23.62 + 0.653(71.3) = 70.18$

The points seem to be evenly spread around the regression line, but to better assess the assumptions (similar to those from ANOVA), we can look at the residuals strategically.

Point has a positive residule then it is above line of fit and line of fit under Predicted the value

Linear Regression Assumptions Check



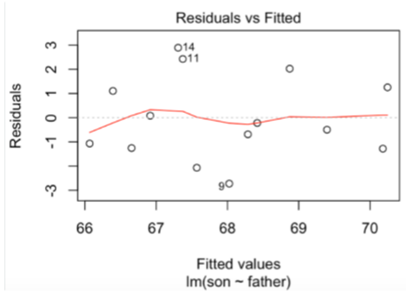
Assumption 1: The model is linear (a straight line makes sense for the data) * Check original and residual plot

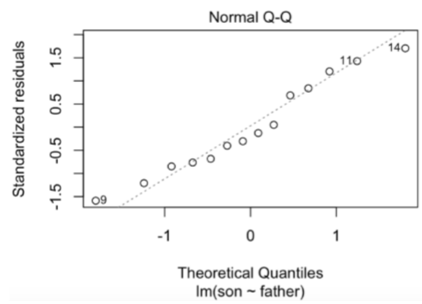
Assumption 2: The observations are independent *Check data collection/science of question

Assumption 3: The variance around the true line is constant for all values of x. * Check residual plot of observed residuals

Assumption 4: The random error around the true line is normal.
* Check qqplot of observed residuals

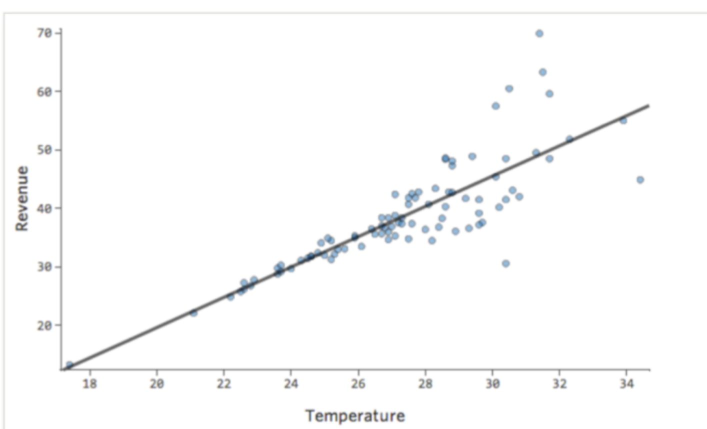
We summarize Assumptions 2-4 with $\epsilon \sim iid\ N(0, \sigma^2)$

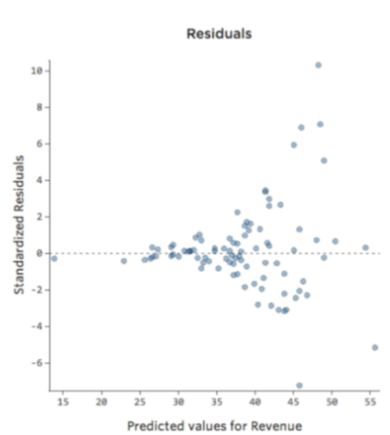




Least Squares Regression Assumptions Check

Patterns in Residuals

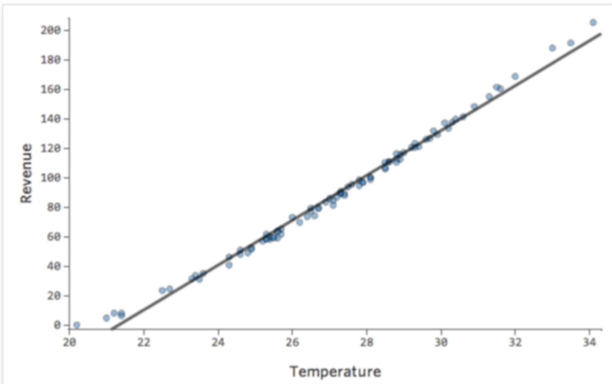


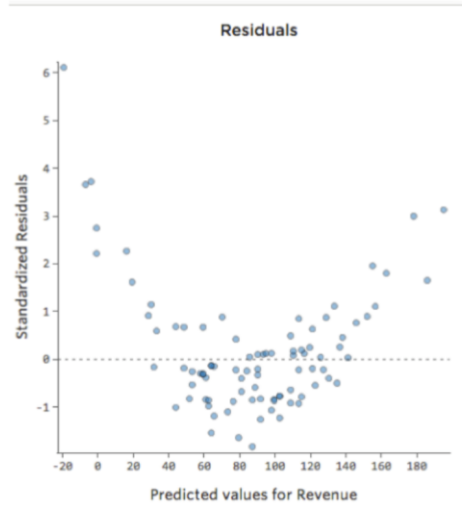


https://www.qualtrics.com/support/stats-iq/analyses/regression-guides/interpreting-residual-plots-improve-regression/

Least Squares Regression Assumptions Check

Patterns in Residuals





https://www.qualtrics.com/support/stats-iq/analyses/regression-guides/interpreting-residual-plots-improve-regression/

Statistics of Linear Regression

Is the slope of the line significantly different from zero

$H_0: \beta_1 = 0$ vs $H_A: \beta_1 \neq 0$

Find the standard error of the estimator: $\widehat{\beta_1}$:

$$\mathrm{SE}(\widehat{\beta_1}) = \frac{\sigma}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$$

$\sum (y_i - \hat{y}_i)^2$

To estimate σ, we use the residuals (SSE).

$$\widehat{\sigma^2} = \frac{SSE}{n-2} = MSE$$

H_0

Then, we get a t statistic: $T = \frac{\widehat{\beta_1}-\beta_1^0}{SE(\widehat{\beta_1})} \sim \boxed{T_{n-2}}$

*Using, but not proving that $\widehat{\beta_1} \sim N(\beta_1, \frac{\sigma^2}{\sum(x_i-\bar{x})^2})$

```
> SSE=sum((mod$residuals)^2)
> MSE=SSE/(length(son)-2)
> sigma.hat=sqrt(MSE)
>
> SSE; MSE; sigma.hat
[1] 37.99205
[1] 3.166004
[1] 1.779327
```

```
> sqrt(var(father)*(14-1))
[1] 7.365363
> sqrt(sum((father-mean(father))^2))
[1] 7.365363
```

Is the slope of the line significantly different from zero in our father/son example? (Ie, is knowing the father's height more useful than just guessing the mean son's height?)

$H_0: \beta_1 = 0$ vs $H_A: \beta_1 \neq 0$

To estimate σ:

$$\widehat{\sigma^2} = \frac{SSE}{n-2} = \frac{37.99205}{14-2} = 3.166 \qquad \hat{\sigma} = \sqrt{3.166} = 1.78$$

$$\mathrm{SE}(\widehat{\beta_1}) = \frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}} = \frac{1.78}{7.365} = 0.242$$

Then, we get a t statistic: $T = \frac{0.653 - 0}{0.242} = \boxed{2.7}$

Pvalue $= 2 * P(t_{12} > 2.7) = 0.02 < p < 0.04$ one side

$0.01 < p < 0.02$ two

We have sufficient evidence at 5% level to reject the null; evidence suggests father's height help predicts son's height

Statistics of Linear Regression (in R)

Is the slope of the line significantly different from zero in our father/son example? (Ie, is knowing the father's height more useful than just guessing the mean son's weight?) ($H_0: \beta_1 = 0$ vs $H_A: \beta_1 \neq 0$)

σ: $\widehat{\sigma^2} = \frac{SSE}{n-2} = \frac{37.99205}{17-2} = \boxed{3.166}$, $\hat{\sigma} = \boxed{1.78}$

$SE(\widehat{\beta_1}) = \frac{\sigma}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} = \frac{\boxed{1.78}}{\sqrt{54.24857}} = \boxed{0.2416718}$

Then, we get a t statistic: $T = \frac{0.653}{\boxed{0.2416718}} = \boxed{2.70} \sim T_{12}$

$P(T_{12} > 2.70) = 0.0096$ so pvalue= $\boxed{0.0193}$

mod=lm(son~father)
summary(mod)

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  23.6409    16.4171   1.440   0.1754
father        0.6527     0.2416   2.702   0.0192 *
Residual standard error: 1.779 on 12 degrees of freedom
Multiple R-squared:  0.3782,    Adjusted R-squared:  0.3264
F-statistic:   7.3 on 1 and 12 DF,  p-value: 0.01924
```

```
> anova(mod)
Analysis of Variance Table

Response: son
          Df Sum Sq Mean Sq F value  Pr(>F)
father     1 23.112  23.111  7.2999 0.01924 *
Residuals 12 37.992   3.166
```

$\widehat{Son's\ Height} = \boxed{23.62} + \boxed{0.653}(Father's\ Height)$

Statistics of Linear Regression (in R)

If we define $SSTot = \sum_{i=1}^{n}(y_i - \bar{y})^2$ we can create a quantity called R^2

$R^2 = \frac{SSTot - SSE}{SSTot} = \frac{(23.112+37.992)-37.992}{(23.112+37.992)} = \frac{23.112}{61.104} =$ 0.3782404

37.8 % of the variability in sons' heights can be explained by fathers' heights.

The remaining variability is due to other factors that are not in our model. In general, R^2 can be interpreted as the fraction of the total sum of squares that is explained by the regression line. R^2 is a good measure of how well x explains y (when the model is linear)

mod=lm(son~father)
summary(mod)

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  23.6409    16.4171   1.440   0.1754
father        0.6527     0.2416   2.702   0.0192 *

Residual standard error: 1.779 on 12 degrees of freedom
Multiple R-squared:  0.3782,    Adjusted R-squared:  0.3264
F-statistic:   7.3 on 1 and 12 DF,  p-value: 0.01924
```

```
> anova(mod)
Analysis of Variance Table

Response: son
          Df Sum Sq Mean Sq F value  Pr(>F)
father     1 23.112  23.111  7.2999 0.01924 *
Residuals 12 37.992   3.166
```

(Notice, $\sqrt{R^2} = \sqrt{0.3782404} =$ 0.615

Predicting New Values with LS Regression

To predict an [average] value of y for a given x, use our regression line:

$$\hat{y}(x^*) = \hat{y}|x^* = \widehat{\beta_0} + \widehat{\beta_1}x^*$$

But we know better than a point estimate is a point estimate with a measure of accuracy.

a. Predicting the position of the fitted line, or the average of infinite predictions at the same x^*.

$$SE\big(E(\hat{y}|x^*)\big) = \sigma\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^\wedge 2}}$$

Notice the further x is from $\bar{x}$, the larger the SE.

Predicting New Values with LS Regression

From R: $\widehat{Son's\ Height} = 23.64 + 0.65(Father's\ Height)$

a. Suppose we want to predict the average son's height when the father is $x^* =$ 70 inches tall.

$$\hat{y}(x^* = 70) = 23.64 + 0.65(70) = 69.14$$

mean X

$$SE\big(E(\hat{y}|x^*)\big) = 1.76\sqrt{\underbrace{\frac{1}{14}}_{\frac{1}{n}} + \frac{(70 - 67.928)^2}{54.25}} = 0.69$$

So a 95% CI for the **average son's height** when the father is 70 inches tall is:

$$\hat{y}|x^* \pm \underbrace{t_{n-2,.025}}_{12} * SE\big(E(\hat{y}|x^*)\big) = 69.14 \pm 2.18 * 0.69$$

$$= (67.64, 70.64)$$

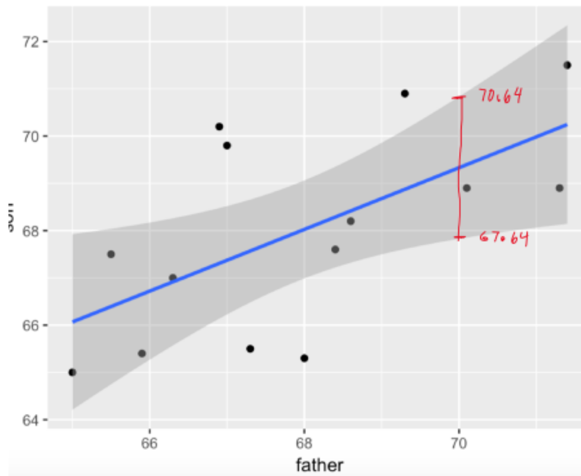
Predicting New Values with LS Regression

So a 95% CI for the average son's height when the father is 70 inches tall is:

$69.14 \pm 2.18 * 0.69 = (67.64, 70.64)$

Don't extrapolate!

lL only



Predicting New Values with LS Regression

b. Suppose we want to predict **the actual value of son's height** when the father is $x^* = 70$ inches tall.

$$\hat{y}(x^* = 70) = 23.64 + 0.65 * 70 = 69.14$$

$SE\left((\hat{y}|x^*)\right) = \sigma\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$ *we need to include the additional random error of each observation around the mean

$SE\left((\hat{y}|x^* = 70)\right) = 1.78\sqrt{1 + \frac{1}{14} + \frac{(70 - 67.93)^2}{54.25}} = 1.91$

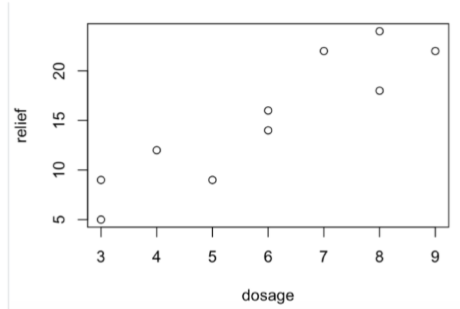
So a 95% CI for the **son's height** when the father is 70 inches tall is:
(Prediction Interval)

$\hat{y}|x^* \pm t_{n-2,.025} * SE\left((\hat{y}|x^*)\right) = 69.14 \pm 2.18 * 1.91 = (64.98, 73.30)$ *prediction interval wider

Regression Example Take 2:

An experiment is conducted to study how different dosages of the drug affect the duration of relief from the allergic symptoms. Ten patients are included in the experiment. Each patient receives a specified dosage of the drug and is asked to report back as soon as the protection of the drug seems to wear off. The observations are recorded in Table 1 which shows the dosage and duration of relief for the 10 patients.

Dosage	Duration of Relief
3	9
3	5
4	12
5	9
6	14
6	16
7	22
8	18
8	24
9	22



From the plot we see:

Regression Example Take 2:

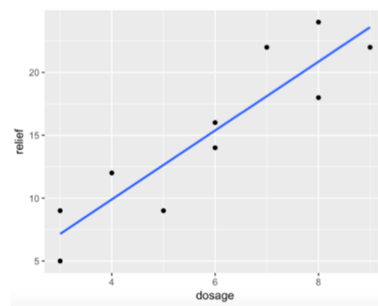
Use your calculator to calculate the (a) correlation and (b) LSRL between dosage and duration of relief. Why does it make sense to do so?

Dosage	Duration of Relief
3	9
3	5
4	12
5	9
6	14
6	16
7	22
8	18
8	24
9	22

$\overline{dosage} = 5.9, \qquad \overline{relief} = 15.1$

$s_{dos} = 2.132, \qquad s_{relief} = 6.42$

```
> sum((dosage-mean(dosage))*(relief-mean(relief)))
[1] 112.1
```

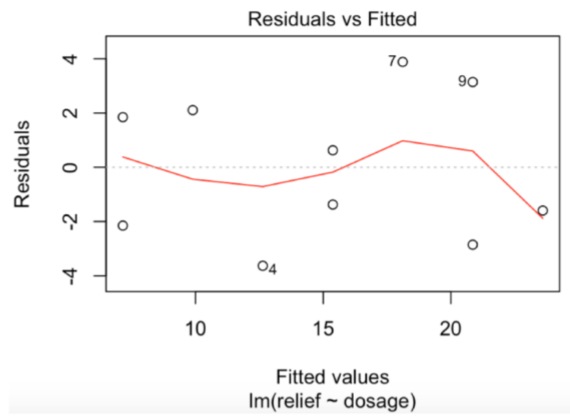


Regression Example Take 2:

Use your calculator to calculate the (c) the residuals for the LSRL. Sketch a residual plot (predicted x, residual on y). What does the residual plot tell us?

$$\widehat{relief} = -1.07 + 2.74 * dosage$$

Dosage	Observed Relief	Predicted Relief	Residual: Obs-Pred
3	9	7.15	
3	5		-2.15
4	12		2.11
5	9	12.63	-3.63
6	14	15.37	-1.37
6	16		
7	22	18.11	3.89
8	18		-2.85
8	24	20.85	3.15
9	22		-1.59



Regression Example Take 2:

Do the data give strong evidence that the mean duration of relief increases with higher dosages of the drug?

Dosage	Duration of Relief
3	9
3	5
4	12
5	9
6	14
6	16
7	22
8	18
8	24
9	22

$$\overline{dosage} = 5.9, \qquad \overline{relief} = 15.1$$

$$s_{dos} = 2.132, \qquad s_{relief} = 6.42$$

$$r = 0.9101$$

$$\widehat{relief} = -1.071 + 2.741 * dosage$$

$$H_o: \beta_1 = 0, \qquad H_a: \beta_1 > 0$$

$$\hat{\sigma} = \sqrt{MSE} = 2.821$$

Regression Example Take 2:

a. Calculate a 95% confidence interval for the (i) expected duration of relief when the dosage is $x^* = 6$ and (ii) predicted duration of relief for a single new patient with dosage $x^* = 6$

$\overline{dosage} = 5.9,\qquad \overline{relief} = 15.1$

$s_{dos} = 2.132,\qquad s_{relief} = 6.42$

$r = 0.9101$

$\widehat{relief} = -1.071 + 2.741 * dosage$

$\hat{\sigma} = \sqrt{MSE} = 2.821$ and