ST518 - Notes packet # 1

Simple Linear Regression

Jason A. Osborne

N. C. State Univ.

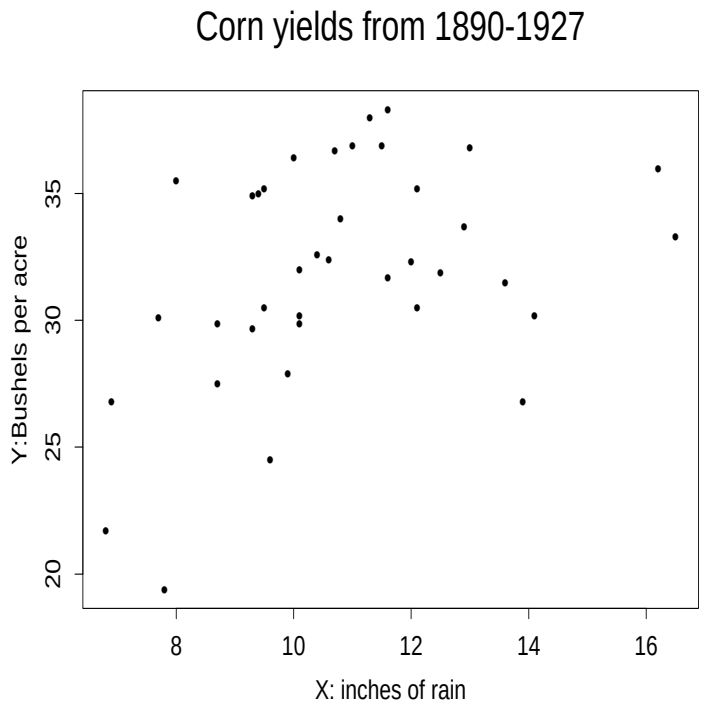
Simple linear regression(SLR)

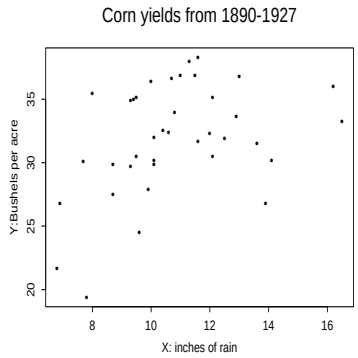
An example with the association between corn yield and rainfall. Yields y (in bushels/acre) on corn raised in six midwestern states from 1890 to 1927 recorded with rainfall x (inches/yr). ("`cornyields.txt`" on moodle)

$$y_1, \ldots, y_{38} \quad \text{and} \quad x_1, \ldots, x_{38}.$$

Year	1890	1891	1892	1893	1894	1895	1896	1897	1898	1899
Yield	24.5	33.7	27.9	27.5	21.7	31.9	36.8	29.9	30.2	32
Rainfall	9.6	12.9	9.9	8.7	6.8	12.5	13	10.1	10.1	10.1
Year	1900	1901	1902	1903	1904	1905	1906	1907	1908	1909
Yield	34	19.4	36	30.2	32.4	36.4	36.9	31.5	30.5	32.3
Rainfall	10.8	7.8	16.2	14.1	10.6	10	11.5	13.6	12.1	12
Year	1910	1911	1912	1913	1914	1915	1916	1917	1918	1919
Yield	34.9	30.1	36.9	26.8	30.5	33.3	29.7	35	29.9	35.2
Rainfall	9.3	7.7	11	6.9	9.5	16.5	9.3	9.4	8.7	9.5
Year	1920	1921	1922	1923	1924	1925	1926	1927		
Yield	38.3	35.2	35.5	36.7	26.8	38	31.7	32.6		
Rainfall	11.6	12.1	8	10.7	13.9	11.3	11.6	10.4		

A *scatterplot* provides indication of some association between y and x. In particular, yields seem to increase with rainfall.





$r = .4$

Some questions:

- How can we describe the association between yield and rainfall? Does it appear *linear*?
- How can we measure the strength of the linear association?
- To what degree is the variability in yield described or explained by its association with rainfall?
- How can we use this association to estimate average yield, given a certain level of rainfall? How can we model observed yield and predict yield given rainfall?

$\bar{x} = \frac{\sum x_i}{n}$

$\sum (x_i - \bar{x}) = 0$
$= \sum x_i - \sum \bar{x}$
$= \sum x_i - n\bar{x}$
$= \sum x_i - \sum x_i = 0$

For paired data, $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, the sample correlation coefficient r_{xy} defined by

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})/(n-1)}{\sqrt{\sum(x_i - \bar{x})^2/(n-1) * \sum(y_i - \bar{y})^2/(n-1)}} = \frac{s_{xy}}{s_x s_y}$$

s_{xy} is called the sample covariance of x and y:

$$s_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{\sum(x_i - \bar{x})y_i - \sum_{i=1}^{n}(x_i - \bar{x})\bar{y}}{n-1}$$

The sample correlation coefficient quantifies the linear association between two continuous variables.

Alternative expressions for $(n-1)s_{xy}$: $\sum(x_i - \bar{x})y_i$, $\sum x_i(y_i - \bar{y})$

Some properties of r_{xy}

- r_{xy} is a measure of the linear assn. between x and y in a dataset.
- correlation coefficients always between -1 and 1:

$$-1 \leq r_{xy} \leq 1$$

- The closer r_{xy} is to $(-1/+1)$, the stonger the (negative/positive) linear association between x and y
- If $|r_{xy}| = 1$, then x and y are said to be perfectly correlated.

Summary statistics for corn yields data:

$$\bar{x} = 10.8, \quad s_x^2 = 5.13 \quad s_x = 2.27$$

$$\bar{y} = 31.9, \quad s_y^2 = 19.0 \quad s_y = 4.44$$

$$s_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{147.3}{38-1} = 3.98(\text{units?})$$

Applying the formula for r_{xy}, we get

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{3.98}{\sqrt{5.13 \times 19.0}} = \frac{3.98}{9.87} = 0.40$$ unitless

The *population* correlation coefficient ρ

Just as $\bar{x}$ provides empirical information about a population mean μ_X, r_{xy} is an empirical estimate of the *population correlation coefficient* ρ_{xy}.

R_{xy} is an *estimator* of the unknown parameter, ρ_{XY}.

ρ is an unknown parameter to be estimated from data.

$$\rho_{XY} = E\left[\left(\frac{X-\mu_X}{\sigma_X}\right)\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\right] = \rho. \quad \frac{\sum(x_i-\bar{x})(y_i-\bar{y})/(n-1)}{s_x s_y} = r_{xy}$$

$E(\cdot)$ denotes *expectation* w.r.t the joint probability distribution, $f(x,y)$

Y refers to a theoretical quantity, which we model using the theory of random variables and their probability distribution. If (X,Y) is bivariate normal with mean vector (μ_X, μ_Y), variances σ_X^2, σ_Y^2, correlation ρ, then

$$E(Y|X) = \mu_Y + \rho\sigma_Y\left(\frac{X-\mu_X}{\sigma_X}\right)$$

Q: What is $E(Y)$? $= \mu_Y$

Recall that for the $n = 38$ bivariate rainfalls and corn yields, $r = 0.4$.

Consider $H_0 : \rho = 0$.

If $\{X_i, Y_i\}$ bivariate continuous *Fisher's transformation* of R approx. normal with mean given by Fisher's transformation of ρ:

$$Z(\rho) = (\frac{1}{2}\sqrt{n-3})\left(\log_e \frac{1+R}{1-R} - \log_e \frac{1+\rho}{1-\rho}\right) \dot{\sim} N(0,1)$$

$Z \dot{\sim} N(0,1)$, leading to tests,confidence intervals for ρ. Under $H_0 : \rho = 0$,

$$(\frac{1}{2}\sqrt{n-3}) \log \frac{1+R}{1-R} \dot{\sim} N(0,1).$$

A large-sample test of H_0 with level α then rejects H_0 whenever

$$\left|\frac{1}{2}\sqrt{n-3} \log\left(\frac{1+R}{1-R}\right)\right| > z_{\alpha/2}$$

where z_α satisfies $\alpha = \Pr(Z > z_\alpha)$ with $Z \sim N(0,1)$.

Corn yields example: $r = 0.4$, $z(0) = 0.5\sqrt{38-3}\log(1.4/0.6) = 2.5$.
Using level $\alpha = .05$, $z_{.025} = 1.96$. (or $\alpha = .01, z_{.005} = 2.58$)

Since $|z_{obs}| > 1.96$, $\rho = 0$ is not consistent with the observed data the sample correlation coefficient $r = 0.4$ can be declared "significantly different from zero at level $\alpha = .05$." (but not at $\alpha = .01$)

Q: Smallest α at which $H_0 : \rho = 0$ can be rejected (see next page)?

```
data one;
   input yield rain;
   cards;
24.5 9.6
... (abbreviated)
32.6 10.4
;
run;

proc corr data=one cov fisher(biasadj=no);
   var yield rain;
run;
```

```
The CORR Procedure

        Covariance Matrix, DF = 37

                     yield                    rain

yield          19.04190612            3.98025605
rain            3.98025605            5.13217639

Pearson Correlation Coefficients, N = 38
        Prob > |r| under H0: Rho=0

                yield                  rain

yield          1.00000                0.40263
                                       0.0122 <-- this is from a different test we'll get to later

rain           0.40263                1.00000
                0.0122

                     Pearson Correlation Statistics (Fisher's z Transformation)

           With                    Sample                                                  p Value for
Variable   Variable       N  Correlation  Fisher's z      95% Confidence Limits             H0:Rho=0

yield      rain          38      0.40263     0.42678       0.095199       0.639943          0.0116
```

$z(\rho=0) = \frac{1}{2}\sqrt{n-3}\log\left(\frac{1+R}{1-R}\right)$

$(obs) = \frac{1}{2}\sqrt{38-3}\log\left(\frac{1+.4}{1-.4}\right) = 2.5$

$H_0: \rho = 0$ is implausible

What procedure did the software use to obtain the 95% confidence interval (.095, .64)?

$100(1-\alpha)\%$ conf.interval obtained by finding values of ρ that satisfy

$$1-\alpha = \Pr\left(z_{1-\alpha/2} < \sqrt{n-3}\frac{1}{2}\left(\log\left(\frac{1+R}{1-R}\right) - \log\left(\frac{1+\rho}{1-\rho}\right)\right) < z_{\alpha/2}\right)$$

Algebraically messy. Let $\psi = \frac{1}{2}\log\left(\frac{1+\rho}{1-\rho}\right)$, then rearranging statement above gives an approximate $100(1-\alpha)\%$ confidence interval for ψ:

$$\frac{1}{2}\log\left(\frac{1+R}{1-R}\right) \pm \frac{z_{\alpha/2}}{\sqrt{n-3}}.$$

Note that ρ and ψ are related by

$$\rho = \frac{e^{2\psi}-1}{e^{2\psi}+1}$$

Evaluating ρ at the limits for ψ, denoting $z_{\alpha/2}$ by z, leads to the interval

$$\left(\frac{\frac{1+R}{1-R}e^{-2z/\sqrt{n-3}}-1}{\frac{1+R}{1-R}e^{-2z/\sqrt{n-3}}+1}, \frac{\frac{1+R}{1-R}e^{2z/\sqrt{n-3}}-1}{\frac{1+R}{1-R}e^{2z/\sqrt{n-3}}+1}\right).$$

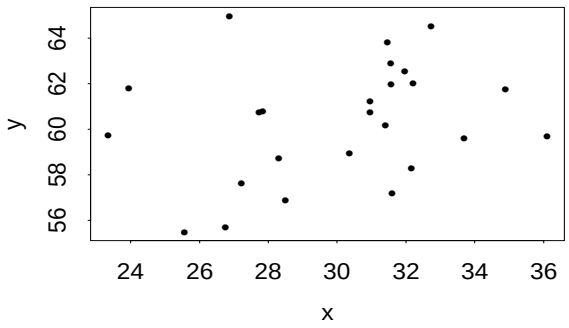
$z_{.025} \approx 1.96$
$z_{.05} \approx 1.645$

Fun exercises:

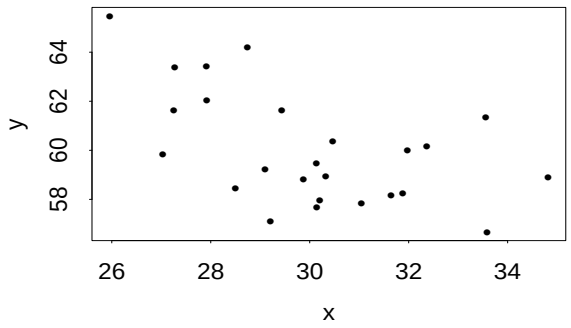
1. Consider a random sample of size $n = 30$ from a bivariate population with correlation $\rho = 0.6$. Approximate $P(R > 0.7)$.
2. Match the plots with the sample correlation coefficients $r_1 = 0.3, r_2 = 0.7, r_3 = 0.1, r_4 = -0.6$

apply Fisher transform to both sides using $\rho = .6$ then look up in normal table

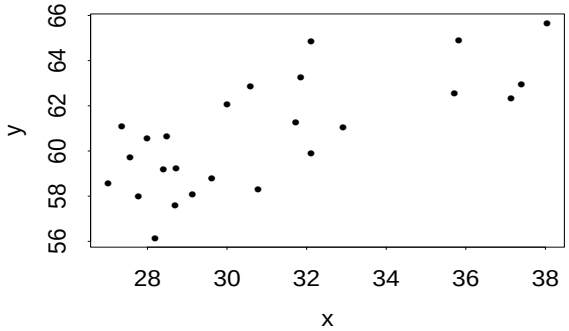
r = .3



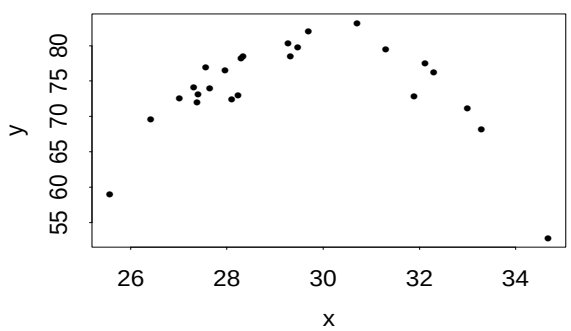
r = .6



r = .7



r = .1 ✓



Correlation does not imply causation

Famous examples of *spurious correlations*:

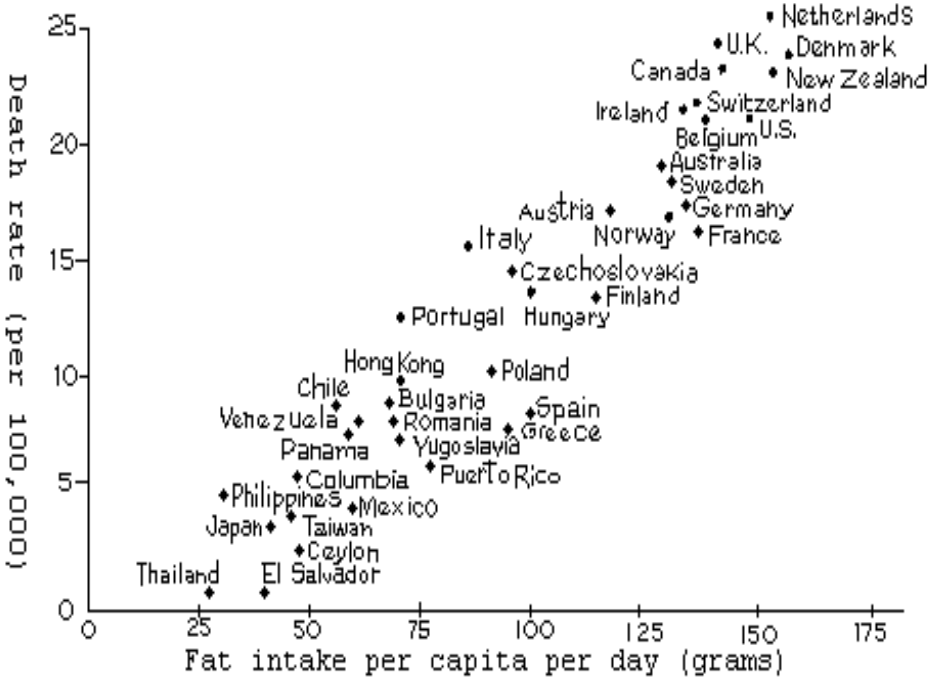
- A study finds a high positive correlation between coffee drinking and coronary heart disease. Newspaper reports say the fragrant essence of the roasted beans of *Coffea arabica* are a menace to public health.
- In a city, if you were to observe the amount of damage and the number of fire engines for enough recent fires, you would likely see a positive and significant correlation among these variables. Obviously, it would be erroneous to conclude that fire engines cause damage.
- *Lurking variable* - a third variable that is responsible for a correlation between two others. (A.k.a. confounding factor.) An example would be to assess the association between say the reading skills of children and other measurements taken on them, such as shoesize. There may be a statistically significant association between shoe size and reading skills, but that doesn't imply that one causes the other. Rather, both are positively associated with a third variable, *age*.
- Among 50 countries examined in a dietary study, high positive correlation among fat intake and cancer (see figure, next page). This example is taken from from *Statistics* by Freedman, Pisani and Purves.

In countries where people eat lots of fat like the United States rates of breast cancer and colon cancer are high. This correlation is often used to argue that fat in the diet causes cancer. How good is the evidence?

Discussion. If fat in the diet causes cancer, then the points in the diagram should slope up, other things being equal. So the diagram is some evidence for the theory. But the evidence is quite weak, because other things aren't equal. For example, the countries with lots of fat in the diet also have lots of sugar. A plot of colon cancer rates against sugar consumption would look just like figure 8, and nobody thinks that sugar causes colon cancer. As it turns out, fat and sugar are relatively expensive. In rich countries, people can afford to eat fat and sugar rather than starchier grain products. Some aspects of the diet in these countries, or other factors in the life-style, probably do cause certain kinds of cancer and protect against other kinds. So far, epidemiologists can identify only a few of these factors with any real confidence. Fat is not among them.

(p. 152, *Statistics* by Friedman, Pisani, Purves and Adhikari)

Figure 8. Cancer rates plotted against fat in the diet for a sample of countries



Source: K. Carroll. "Experimentalevidence of dietary factors and hormone-dependent cancers Cancer Research vol. 35 (1975) p.3379. Copyright by Cancer Research. Reproduced by permission

A pertinent quotation from G. B. Shaw's *The Doctor's Dilemma:*

> *Comparisons which are really comparisons between two social classes with different standards of nutrition and education are palmed off as comparisons between the results of a certain medical treatment and its neglect. Thus it is easy to prove the wearing of tall hats and the carrying of umbrellas enlarges the chest, prolongs life, and confers comparative immunity from disease; for the statistics show that the classes which use these articles are bigger, healthier, and live longer than the class which never dreams of possessing such things. It does not take much perspicacity to see that what really makes this difference is not the tall hat and the umbrella, but the wealth and nourishment of which they are evidence . . .*

Observe n independent pairs $(y_1, x_1), (y_2, x_2), \ldots, (y_n, x_n)$

A simple linear regression (probability) model for Y conditional on $X = x$:

$$Y_i = \underbrace{\beta_0 + \beta_1 x_i}_{\text{deterministic component}} + \underbrace{E_i}_{\text{random error component}}$$

where $E_1, \ldots, E_n$ are independent, identically distributed $N(0, \sigma^2)$

1. $\mu(x) = E(Y|X = x) = \beta_0 + \beta_1 x$ ("regression function")
2. $\text{Var}(Y|X = x) = \sigma^2$ ("error variance")

Definitions:

- response or dependent variable Y (left side of regression equation)
- independent variable or predictor variable X (right side)
- intercept term $\beta_0 = E(Y|X = 0)$ (where $\mu(x)$ crosses y-axis.)
- slope term β_1, average change in $E(Y|X = x)$ per unit increase in x

β_0, β_1 and σ^2 are fixed, unknown parameters which can be estimated from data using simple linear regression(SLR)

Fitting a linear model

Derive estimators of β_0 and β_1 to minimize error sum of squares:

$$SS[E] = \sum_{1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2$$

Take partial derivatives with respect to β_0 and β_1, set to 0, solve the resulting system of two equations and two unknowns to get "least squares" (LS) estimates.

$$\begin{aligned} \widehat{\beta}_1 &= \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} \qquad \text{(many ways to write)} \\ \widehat{\beta}_0 &= \bar{y} - \widehat{\beta}_1 \bar{x} \end{aligned}$$

Fitted value for i^{th} observation (revisit slide 7):

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i = \cdots (\text{derive}) \cdots = \bar{y} + r_{xy} s_y \left(\frac{x_i - \bar{x}}{s_x} \right)$$

Exercise: Show that $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are *unbiased* estimators.
That is, show that $E(\hat{\beta}_1|x_1,\ldots,x_n) = \beta_1$ and $E(\hat{\beta}_0|x_1,\ldots,x_n) = \beta_0$.

Consider estimation of the error variance, $\sigma^2 = \text{Var}(Y_i|x_i) = \text{Var}(E_i)$:
An unbiased estimator of the error variance given by

$$\widehat{\sigma}^2 = \frac{\sum(Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2}{n-2}$$

For the corn yields, (enter units in space)

$$\widehat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = r_{xy}\frac{s_y}{s_x} = 0.4\frac{4.44}{2.27} = 0.78(\qquad\qquad)$$

$$\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1\bar{x} = 31.9 - (0.78)(10.8) = 23.5(\qquad\qquad)$$

The LS regression line is then given by

$$\widehat{y} = 23.5 + .78x$$

Note that with LS estimates,

1. $\sum_1^n(y_i - \hat{y}_i) = 0$
2. $\sum_1^n(y_i - \hat{y}_i)^2$ is minimized

How much variability in yields is explained by linear dependence on rainfall?

Source	Sum of squares	df	Mean Square	EMS	F-Ratio
Reg.	$SS[R]$	1	$MS[R]$	$\sigma^2 + \beta_1^2 \sum (x_i - \bar{x})^2$	$F = \frac{MS[R]}{MS[E]}$
Error	$SS[E]$	$n-2$	$MS[E]$	σ^2	
Total	$SS[TOT]$	$n-1$			

$$
\begin{aligned}
SS[TOT] &= SS[R] + SS[E] \\
SS[TOT] &= \sum_{i=1}^{n} (y_i - \bar{y})^2 \\
SS[R] &= \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \\
SS[E] &= \sum (y_i - \hat{y}_i)^2
\end{aligned}
$$

If $\beta_1 = 0$ then $E(MS[R]) = E(MS[E])$ and $F \sim F_{1,n-2}$.
$SS[R]/SS[TOT] = r^2$ is called the _____

The yield on corn by rainfall example

Source	Sum of squares	df	Mean Square	F-Ratio
Regression	114	1	114	6.95
Error	591	36	16.4	
Total	705	37		

$$705 = 114 + 591$$

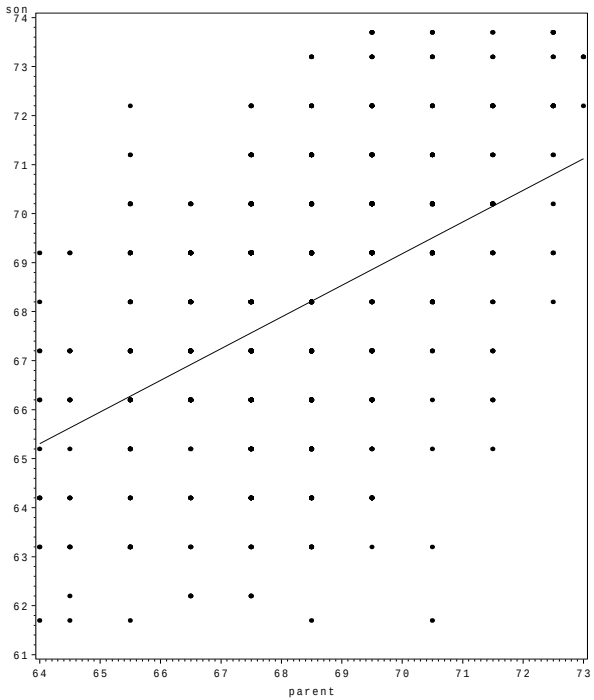
The coefficient of determination is $114/705 = 0.16 = (0.4)^2$.

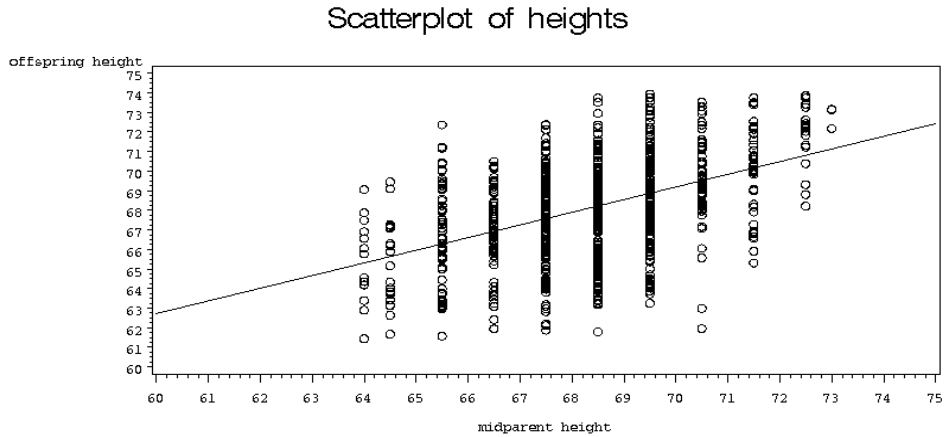
The table above is called an _____ table.

```
proc reg data=one;
   model yield=rain;
run;
```

```
The REG Procedure
                          Analysis of Variance

                                   Sum of           Mean
Source                     DF       Squares         Square    F Value    Pr > F
Model                       1     114.21474      114.21474       6.97    0.0122
Error                      36     590.33578       16.39822
Corrected Total            37     704.55053

Root MSE               4.04947    R-Square      0.1621
Dependent Mean        31.91579    Adj R-Sq      0.1388
Coeff Var             12.68799

                     Parameter       Standard
Variable      DF      Estimate          Error    t Value    Pr > |t|
Intercept      1      23.55210        3.23646       7.28      <.0001
rain           1       0.77555        0.29386       2.64      0.0122
```

A classical dataset: heights of adult males and parents





Model

$$Y_i = \qquad \text{for } i = 1, \ldots, 928 = n$$

where $E_1, \ldots, E_n$ are

- _____
- _____
- normally distributed random variables with mean 0 and _____ σ^2.

(Write $E_i \overset{iid}{\sim} N(0, \sigma^2)$.)

This implies

1. $\mu(x) = E(Y|X = x) =$ _____x
2. $\text{Var}(Y|X = x) =$ _____

(Parameters $\beta_0, \beta_1, \sigma^2$ characterize the whole population of interest.)

Question: Suppose we ignore midparent height x. Consider estimating the mean $E(Y)$. Propose a model and method for obtaining a confidence interval for the mean height in the population from which these data were randomly sampled.

Many questions to answer using regression analysis:

1. What is the meaning, in words, of β_1?
2. True/false: (a) β_1 is a statistic (b) β_1 is a parameter (c) β_1 is unknown.
3. What is the observed value of $\hat{\beta}_1$? Is $\hat{\beta}_1 = \beta_1$?
4. True/false: (a) $\hat{\beta}_1$ is a statistic (b) $\hat{\beta}_1$ is a parameter (c) $\hat{\beta}_1$ is unknown.
5. How much does $\hat{\beta}_1$ vary about β_1 from sample to sample? Report an estimate along with an expression for how to compute it.
6. What is a region of plausible values for β_1 suggested by the data?
7. What is the line that best fits these data, using the criterion that smallest sum of squared residuals is "best?"
8. How much of the observed variation in the heights of sons (the y-axis) is explained by this "best" line?
9. Estimate the mean height in the population with midparent height $x = 68$.
10. Under the model, what is the true average height of sons with midparent height $x = 68$?

11. Estimate the std dev. of heights in the subpopulation with midparent height $x = 68$? Would you call this std dev. a "standard error?"
12. What is the estimated standard deviation among the population of sons whose parents have midparent height $x = 72$? Bigger, smaller, or the same as that for $x = 68$?
13. What is the estimated standard error of the estimated average for sons with midparent height $x = 68$, $\hat{\mu}(68) = \hat{\beta}_0 + 68\hat{\beta}_1$? Provide an expression for this standard error.
14. Write the correct symbol ($< / = / >$): $\widehat{SE}\left(\hat{\mu}(72)\right)$ $\widehat{SE}\left(\hat{\mu}(68)\right)$
15. Let Y denote the height of a male randomly sampled from this population and X his midparent height. Let $\mu_Y, \sigma^2_Y, \mu_X, \sigma^2_X$ denote respective means and variances. Define the population correlation coefficient ρ

$$\rho = E[(\underline{\qquad\qquad}) \times (\underline{\qquad\qquad})]?$$

16. Report a test statistic, Z, for the hypothesis that $\rho = 0$.
17. Define $\mu_Y, \sigma_Y, \mu_X, \sigma_X, \rho$. Parameters or statistics?
18. Is $\boxed{E_1, \ldots, E_{928} \overset{iid}{\sim} N(0, \sigma^2)}$ a reasonable assumption?

```
data Galton;
 array cdata(14);
  if _n_ = 1 then input cdata1-cdata14 @ ;
  retain cdata1-cdata14;  drop cdata1-cdata14 i;
  input parent @;
    do i = 1 to 14; input count @ ; son=cdata(i);
   output;  end;
cards;
    61.7 62.2 63.2 64.2 65.2 66.2 67.2 68.2 69.2 70.2 71.2 72.2 73.2 73.7
73.0  0    0    0    0    0    0    0    0    0    0    1    3    0
72.5  0    0    0    0    0    0    0    1    2    1    2    7    2    4
                          (abbreviated)
64.5  1    1    4    4    1    5    5    0    2    0    0    0    0    0
64.0  1    0    2    4    1    2    2    1    1    0    0    0    0    0
 ;
proc print data=Galton(obs=100);
run;

data big; set galton; drop j count; do j=1 to count;output; end;
proc print data=big(obs=20);

proc means; var son parent;

data questions; /* these values used for prediction or estimation at x=68,x=72 */
   input parent son;
   cards;
   68 .
   72 .
;
run;

data big;
   set big questions;
run;
```

```
proc reg;
   model son=parent/clb;
   output out=out1 residual=r p=yhat ucl=pihigh lcl=pilow uclm=cihigh lclm=cilow
   stdp=stdmean;

data questions; set out1; if son=.;

proc print;
   title "questions regarding prediction, estimation when x=68, x=72";
run;
data fisherz;
   n=928;
   r=sqrt(0.2105);
   rratio=(1+r)/(1-r);
   z=probit(0.975);
   expon=probit(0.975)/sqrt(n-3);
   rlow=(rratio*exp(-2*expon)-1)/(rratio*exp(-2*expon)+1);
   rhigh=(rratio*exp(2*expon)-1)/(rratio*exp(2*expon)+1);
run;

symbol1 i=rl value=dot;

proc gplot; plot son*parent; run;
```

```
                        The MEANS Procedure

Variable      N            Mean         Std Dev         Minimum         Maximum
-------------------------------------------------------------------------------
son         928      68.0884698       2.5179414      61.7000000      73.7000000
parent      928      68.3081897       1.7873334      64.0000000      73.0000000
-------------------------------------------------------------------------------
```

The REG Procedure
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1236.93401	1236.93401	246.84	<.0001
Error	926	4640.27261	5.01109		
Corrected Total	927	5877.20663			

Root MSE	2.23855	R-Square	0.2105
Dependent Mean	68.08847	Adj R-Sq	0.2096
Coeff Var	3.28770		

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	23.94153	2.81088	8.52	<.0001
parent	1	0.64629	0.04114	15.71	<.0001

Parameter Estimates

Variable	DF	95% Confidence Limits	
Intercept	1	18.42510	29.45796
parent	1	0.56556	0.72702

questions regarding prediction, estimation when x=68, x=72 3

Obs	parent	son	yhat	stdmean	cilow	cihigh	pilow	pihigh	r
1	68	.	67.8893	0.07457	67.7429	68.0356	63.4936	72.2849	.
2	72	.	70.4745	0.16871	70.1434	70.8056	66.0688	74.8801	.

Obs	n	r	rratio	z	expon	rlow	rhigh
1	928	0.45880	2.69551	1.95996	0.064443	0.40645	0.50815

Answers to questions about Galton SLR

1. Change in average son's height (inches) per one inch increase in midparent height (in the whole population.)
2. β_1 is an unknown parameter.
3. $\hat{\beta}_1 = 0.65$ son inches/midparent inch (from output.) $P(\widehat{\beta}_1 = \beta_1) = 0$
4. $\hat{\beta}_1 = 0.65$ is an observed value of a statistic.
5. $\widehat{SE}(\hat{\beta}_1) = \sqrt{MS[E]/\sum(x_i - \overline{x})^2} = 0.04$ (from output.)
6. Add and subtract about 2 SE to get $(0.57, 0.73)$
7. $y = 23.9 + 0.65x$
8. $r^2 = 21\%$
9. $\hat{\mu}(68) = \hat{\beta}_0 + 68\hat{\beta}_1 = 67.9$ (from output also.)
10. $\mu(68) = \beta_0 + 68\beta_1$.

11. $\sqrt{MS[E]} = 2.24$. Not a SE.
12. $\sqrt{MS[E]} = 2.24$. (Assume homoscedasticity.)
13. $SE(\hat{\beta}_0 + 68\hat{\beta}_1) = 0.07$. Expressions given by

$$\begin{aligned} \widehat{SE}(\hat{\mu}(68)) &= \sqrt{MS[E]\left(\frac{1}{n} + \frac{(68-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right)} \\ &= \sqrt{(1,68)' MS[E] (X'X)^{-1} (1,68)} \end{aligned}$$

X a (928×2) *design matrix*.

14. $\widehat{SE}(\hat{\mu}(72)) > \widehat{SE}(\hat{\mu}(68))$
15. $\rho = E[(X-\mu_X)(Y-\mu_Y)]/(\sigma_X\sigma_Y)$
16. $z(\rho = 0) = \sqrt{928-3}\frac{1}{2}\log(1.4588/(1-.4588)) = 15.1 >> 1.96$
17. Parameters
18. Except for discreteness, very reasonable. (Residual plots not shown.)