Topic 9

Regression and Correlation

1. Definition of the Linear Regression Model

In the last unit, topic 8, the setting was that of two categorical (*discrete*) variables, such as smoking and low birth weight, and the use of chi-square tests of association and homogeneity.

In this unit, topic 9, our focus is in the setting of two continuous variables, such as age and weight. This topic is an introduction to simple linear regression and correlation.

Linear Regression

Linear regression models the mean μ of one random variable as a linear function of one or more other variables that are treated as fixed. The estimation and hypothesis testing involved are extensions of ideas and techniques that we have already seen. In linear regression,

- **we observe an outcome or dependent variable "Y" at several levels of the independent or predictor variable "X" (there may be more than one predictor "X" as seen later).**

- **A linear regression model assumes that the values of the predictor "X" have been fixed in advance of observing "Y".**

- **However, this is not always the reality. Often "Y" and "X" are observed jointly and are both random variables.**

Correlation

Correlation considers the association of two random variables.

- **The techniques of estimation and hypothesis testing are the same for linear regression and correlation analyses.**

- **Exploring the relationship begins with fitting a line to the points.**

- **We develop the linear regression model analysis for a simple example involving one predictor and one outcome.**

Example.
Source: Kleinbaum, Kupper, and Muller 1988
Available are pairs of observations of age and weight for n=11 chicken embryos.

WT=Y	AGE=X	LOGWT=Z
0.029	6	-1.538
0.052	7	-1.284
0.079	8	-1.102
0.125	9	-0.903
0.181	10	-0.742
0.261	11	-0.583
0.425	12	-0.372
0.738	13	-0.132
1.13	14	0.053
1.882	15	0.275
2.812	16	0.449

We'll use a familiar notation

- **The data are 11 pairs of (X_1, Y_1) where X=AGE and Y=WT**
  (X_1, Y_1) = (6, .029) $\cdots$ (X_{11}, Y_{11}) = (16 , 2.812) and

- **equivalently, 11 pairs of (X_1, Y_1) where X=AGE and Y=LOGWT**
  (X_1, Y_1) = (6, -1.538) $\cdots$ (X_{11}, Y_{11}) = (16 , 0.449)

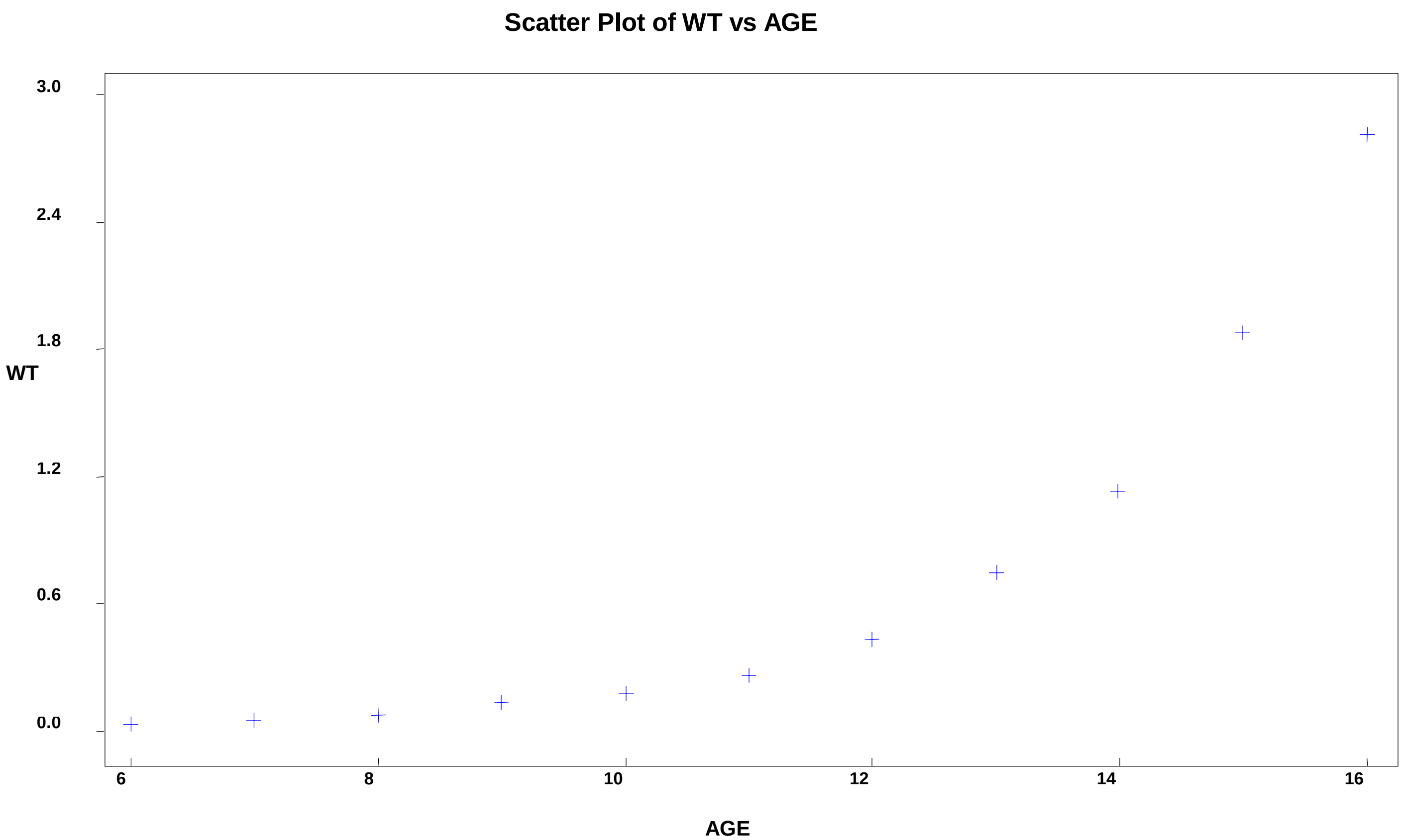
Though simple, it helps to be clear in the research question

Does weight change with age?

In the language of analysis of variance we are asking the following:
Can the variability in weight be explained, to a significant extent, by variations in age?

What is a "good" functional form that relates age to weight?

We begin with a plot of X=AGE versus Y=WT



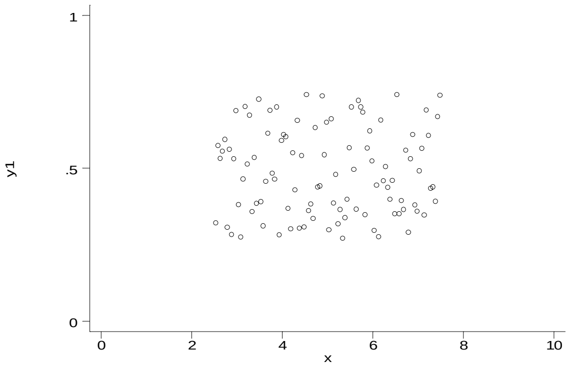
We check and learn about the following:

- **The average and median of X**
- **The range and pattern of variability in X**
- **The average and median of Y**
- **The range and pattern of variability in Y**
- **The nature of the relationship between X and Y**
- **The strength of the relationship between X and Y**
- **The identification of any points that might be influential**

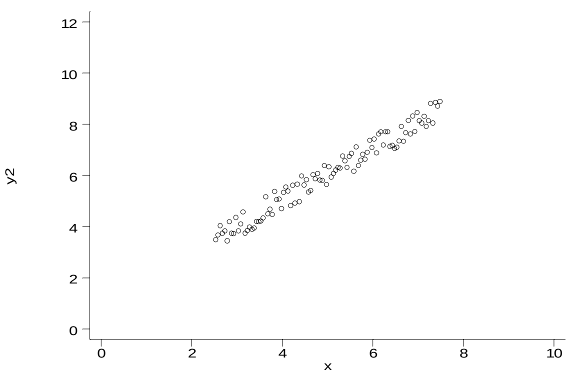
For these data:

- **The plot suggests a relationship between AGE and WT**
- **A straight line might fit well, but another model might be better**
- **We have adequate ranges of values for both AGE and WT**
- **There are no outliers**

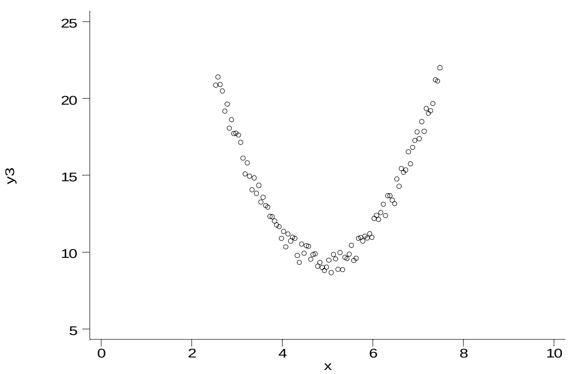
We might have gotten any of a variety of plots.



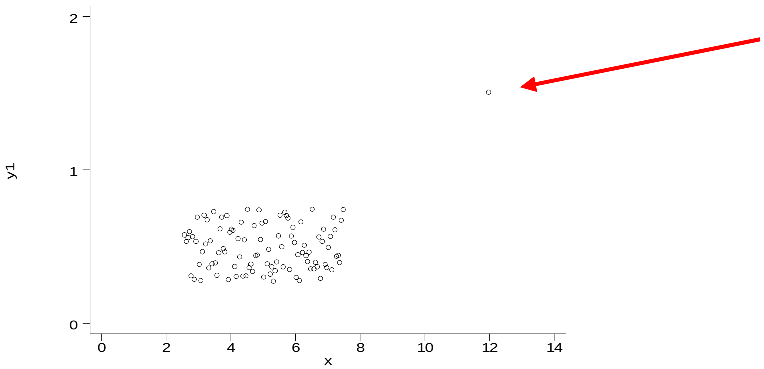
No relationship between X and Y



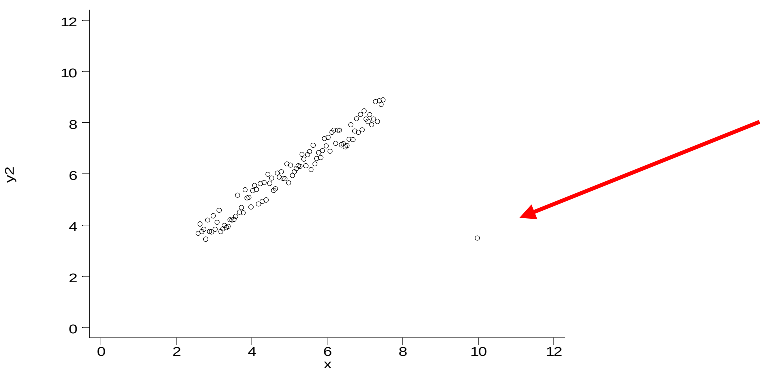
Linear relationship between X and Y



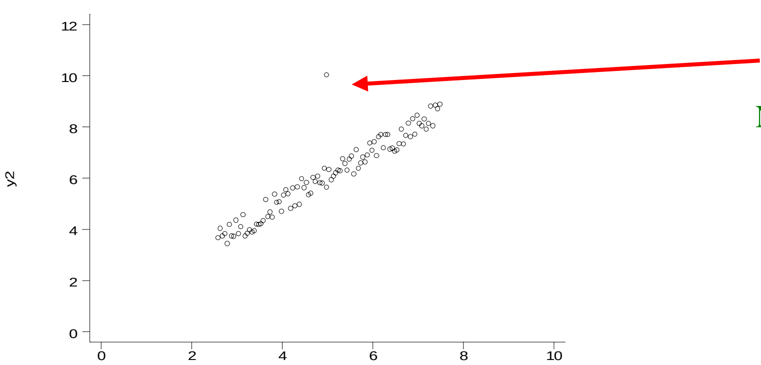
Non-linear relationship between X and Y



Note the arrow pointing to the outlying point
Fit of a linear model will yield estimated slope that is spuriously non-zero.

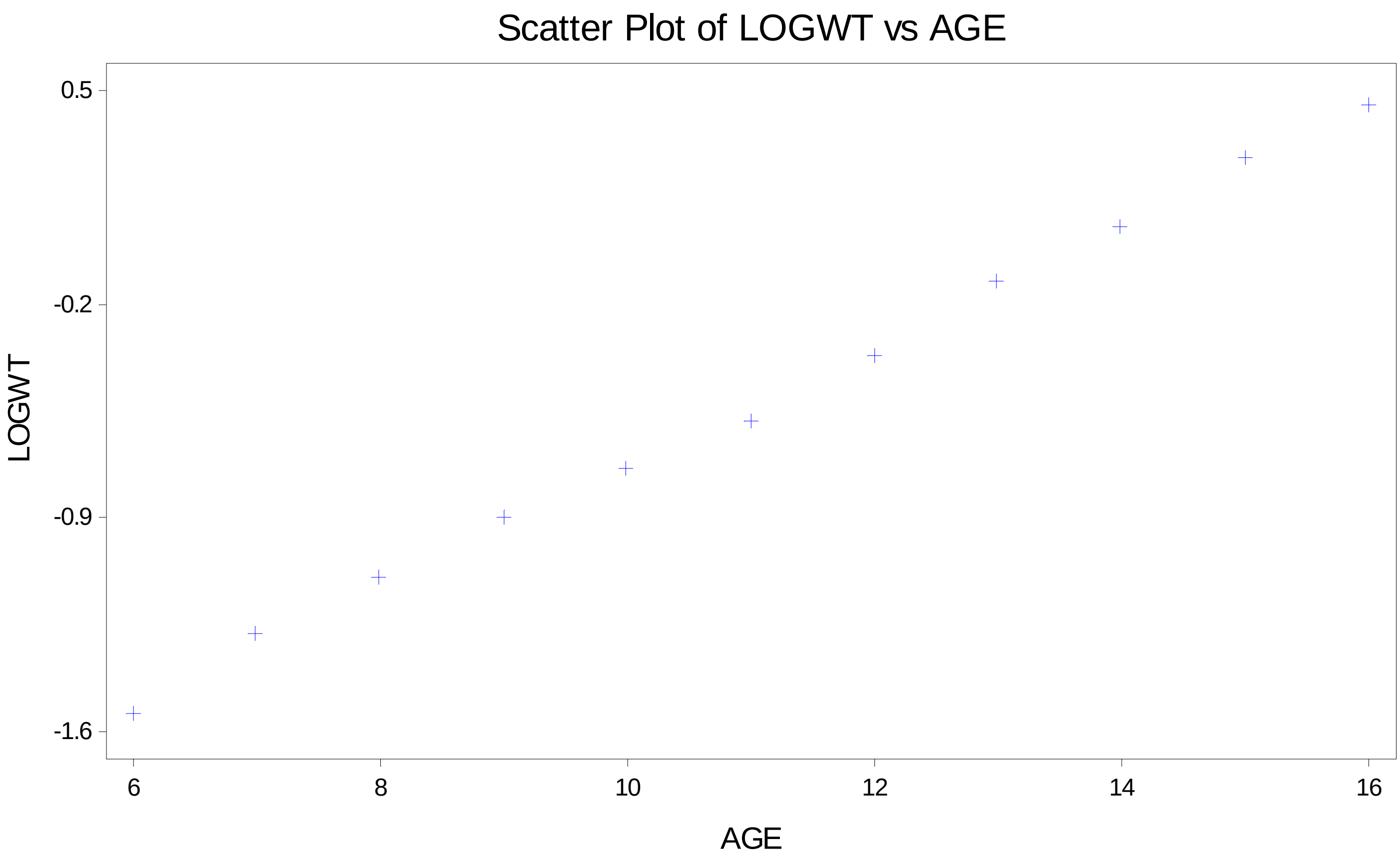


Note the arrow pointing to the outlying point
Fit of a linear model will yield an estimated slope that is spuriously near zero.



Note the arrow pointing to the outlying point
Fit of a linear model will yield an estimated slope that is spuriously high.

The "bowl" shape of our scatter plot suggests that perhaps a better model relates the logarithm of WT to AGE:

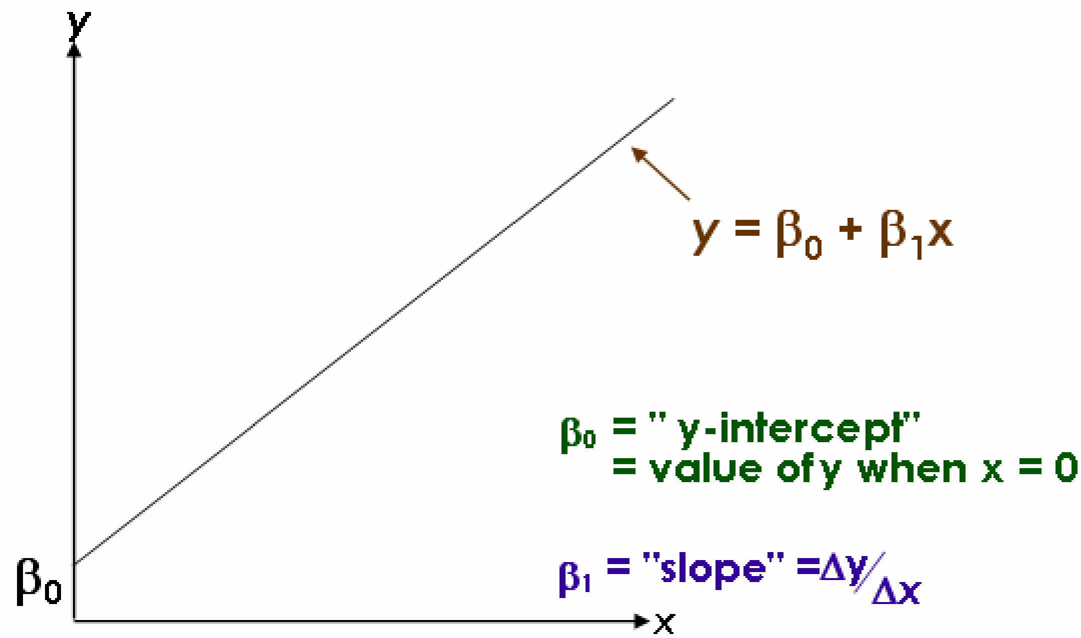


We'll investigate two models.

1) $\mathbf{WT = \beta_0 + \beta_1\ AGE}$

2) $\mathbf{LOGWT = \beta_0 + \beta_1\ AGE}$

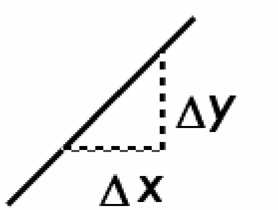
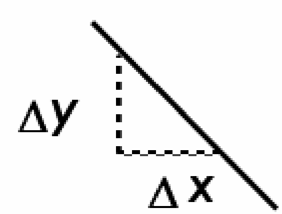
Recall what you might have learned in an old math class about the equation of a line



β_0 = "y-intercept" = value of y when x = 0

β_1 = "slope" = $\Delta y/\Delta x$ = (change in y)/(change in x)

You might recall, too, a feel for the slope

Slope > 0	Slope = 0	Slope < 0
		

Definition of the Straight Line Model

$$Y = \beta_0 + \beta_1 X$$

Population	Sample
$Y = \beta_0 + \beta_1 X + \varepsilon$	$Y = \hat{\beta}_0 + \hat{\beta}_1 X + e$
$Y = \beta_0 + \beta_1 X$ **is the relationship in the population. It is measured with error.**	$\hat{\beta}_0$, $\hat{\beta}_1$, **and** e **are our guesses of** β_0, β_1 **and** ε
ε = **measurement error**	e = **residual**
We do NOT know the value of β_0 **nor** β_1 **nor** ε	**We do have values of** $\hat{\beta}_0$, $\hat{\beta}_1$ **and** e
	The values of $\hat{\beta}_0$, $\hat{\beta}_1$ **and** e **are obtained by the method of <u>least squares estimation.</u>**
	To see if $\hat{\beta}_0 \approx \beta_0$ **and** $\hat{\beta}_1 \approx \beta_1$ **we perform <u>regression diagnostics</u>.** ***Note – This is not discussed in this course; see PubHlth 640, Intermediate Biostatistics***

<u>A little notation, sorry!</u>

Y = the outcome or dependent variable
X = the predictor or independent variable

μ_Y **= The expected value of Y for all persons in the population**
$\mu_{Y|X=x}$ **= The expected value of Y for the sub-population for whom X=x**

σ_Y^2 **= Variability of Y among all persons in the population**
$\sigma_{Y|X=x}^2$ **= Variability of Y for the sub-population for whom X=x**

2. Estimation

We will use the method of <u>least squares</u> to obtain guesses of β_0 and β_1.

<u>Goal</u>

From the many possible lines through the scatter of points, choose the one line that is "<u>closest</u>" to the data.

<u>What do we mean by "Close"?</u>

- **We'd like the vertical distance between each observed Y and its corresponding fitted $\hat{Y}$ to be as small as possible.**

- **It's not possible to choose $\hat{\beta}_0$ and $\hat{\beta}_1$ so that it minimizes**

$\left(Y_1 - \hat{Y}_1\right)^2$ **and minimizes individually**

$\left(Y_2 - \hat{Y}_2\right)^2$ **and minimizes individually**

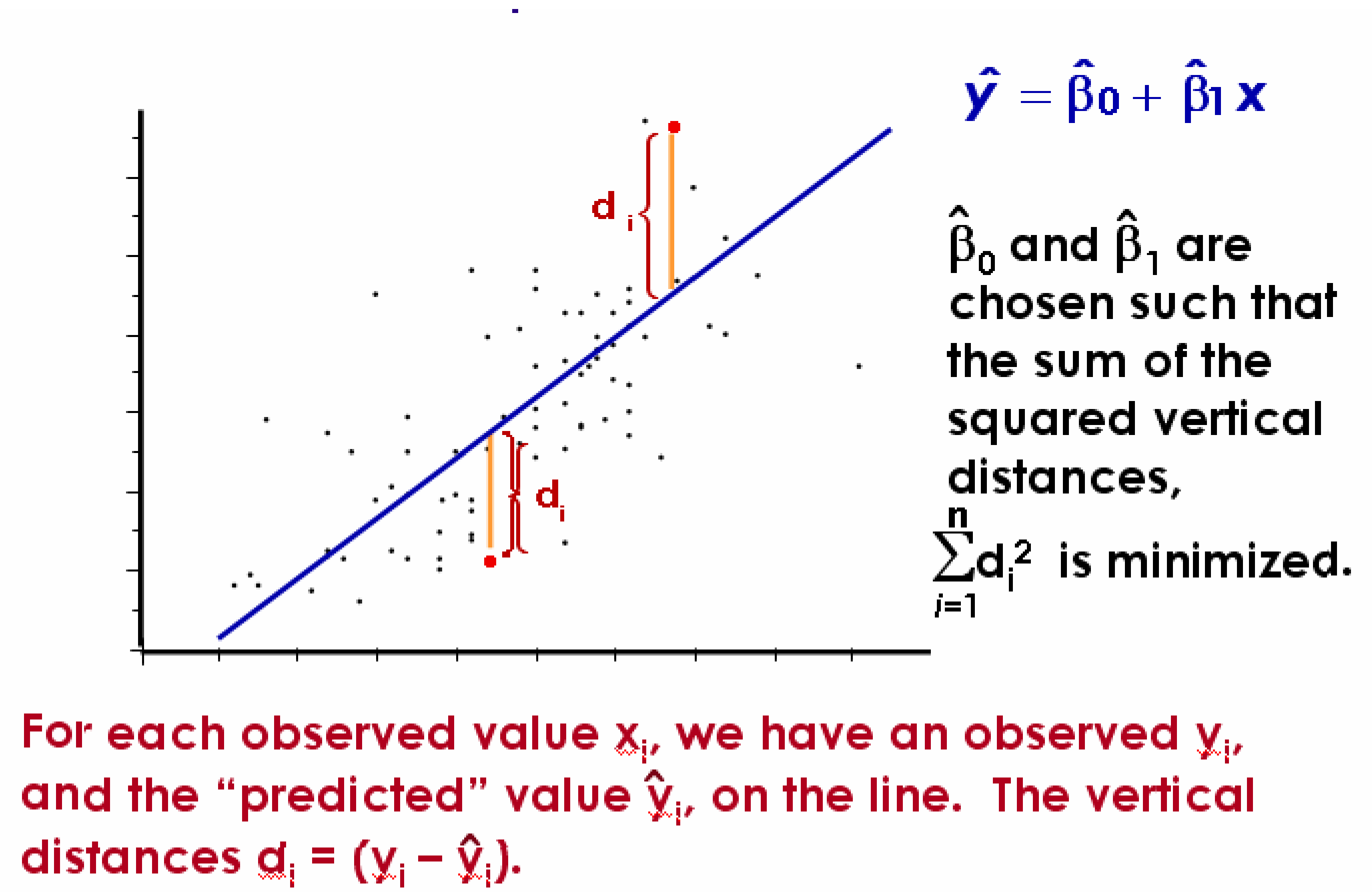
….

$\left(Y_n - \hat{Y}_n\right)^2$

- **So, instead, we choose $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimizes their total**

$$\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n}\left(Y_i - \left[\hat{\beta}_0 + \hat{\beta}_1 X_i\right]\right)^2$$

<u>A picture gives a feel for the fact that many lines are possible and that we seek the "closest" in the sense of vertical distances being as small as possible</u>



The expression to be minimized, $\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n}\left(Y_i - \left[\hat{\beta}_0 + \hat{\beta}_1 X_i\right]\right)^2$ **has a variety of names:**

- **residual sum of squares**
- **sum of squares about the regression line**
- **sum of squares due error (SSE)**
- $\hat{\sigma}^2_{Y|X}$

For the calculus lover, A little calculus yields the solution for the guesses $\hat{\beta}_0$ and $\hat{\beta}_1$

- **Consider SSE =** $\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n}\left(Y_i - \left[\hat{\beta}_0 + \hat{\beta}_1 X_i\right]\right)^2$

- ***Step #1:*** **Differentiate with respect to $\hat{\beta}_1$**
  Set derivative equal to 0 and solve.

- ***Step #2:*** **Differentiate with respect to $\hat{\beta}_0$**
  Set derivative equal to 0, insert $\hat{\beta}_1$ and solve.

For the non-calculus lover, here are the estimates of β_0 and β_1

β_1 **is the slope**

- **Estimate is denoted $\hat{\beta}_1$ or** b_1

β_0 **is the intercept**

- **Estimate is denoted $\hat{\beta}_0$ or** b_0

<u>Some very helpful preliminary calculations</u>

- $S_{xx} = \sum\left(X-\bar{X}\right)^2 = \sum X^2 - N\bar{X}^2$
- $S_{yy} = \sum\left(Y-\bar{Y}\right)^2 = \sum Y^2 - N\bar{Y}^2$
- $S_{xy} = \sum\left(X-\bar{X}\right)(Y-\bar{Y}) = \sum XY - N\bar{X}\bar{Y}$

Note - These expressions make use of a special notation called the "summation notation".

The capitol "S" indicates " summation".
In S_{xy}, the first subscript "x" is saying $(x-\bar{x})$.
The second subscript "y" is saying $(y-\bar{y})$.

$$S_{xy} = \underbrace{\sum}_{\textbf{S}}\underbrace{\left(X-\bar{X}\right)}_{\textbf{subscript x}}\underbrace{(Y-\bar{Y})}_{\textbf{subscript y}}$$

Slope	$\hat{\beta}_1 = \dfrac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sum_{i=1}^{n}(X_i-\bar{X})^2} = \dfrac{\hat{\text{cov}}(X,Y)}{\hat{\text{var}}(X)}$	$\hat{\beta}_1 = \dfrac{S_{xy}}{S_{xx}}$
Intercept	$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X}$	
Prediction of Y	$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$ $= b_0 + b_1 X$	

Do these estimates make sense?

Slope	$\hat{\beta}_1 = \dfrac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \dfrac{\hat{\text{cov}}(X,Y)}{\hat{\text{var}}(X)}$	**The linear movement in Y with linear movement in X is measured relative to the variability in X.** $\hat{\beta}_1 = 0$ **says:** **With a unit change in X, overall there is a 50-50 chance that Y increases versus decreases** $\hat{\beta}_1 \neq 0$ **says:** **With a unit increase in X, Y increases also (**$\hat{\beta}_1 > 0$**) or Y decreases (**$\hat{\beta}_1 < 0$**).**
Intercept	$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$	**If the linear model is incorrect, or, if the true model does not have a linear component, we obtain** $\hat{\beta}_1 = 0$ **and** $\hat{\beta}_0 = \overline{Y}$ **as our best guess of an unknown Y**

Illustration in SAS

Code.

```
data temp;
  input wt age logwt;
  label wt="Weight, Y"
        age="Age, X"
         logwt="Log(Weight),Y";
   cards;
   .029 6 -1.538
   .052 7 -1.284
   .079 8 -1.102
   .125 9 -0.903
   .181 10 -0.742
   .261 11 -0.583
   .425 12 -0.372
   .738 13 -0.132
   1.13 14 0.053
   1.882 15 0.275
   2.812 16 0.449
;
run;
quit;
proc reg data=temp simple;         /* option simple produces simple descriptives */
  title "Regression of Y=Weight on X=Age";
  model wt=age;
run;
quit;
```

Partial listing of output ...

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-1.88453	0.52584	-3.58	0.0059
age	Age, X	1	0.23507	0.04594	5.12	0.0006

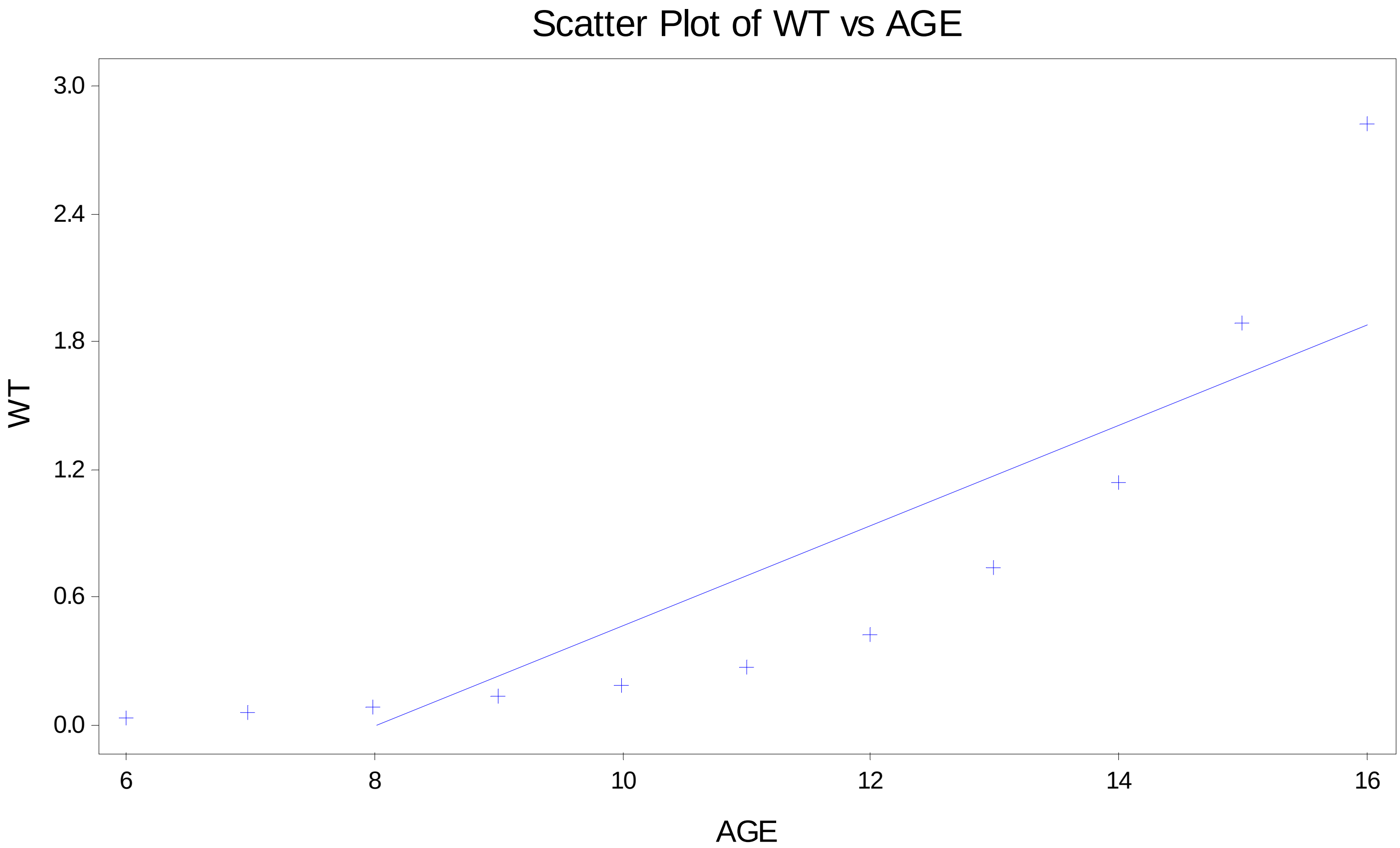
Annotated …

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-1.88453 = **intercept = β_0**	0.52584	-3.58	0.0059
age	Age, X	1	0.23507 = **slope = β_1**	0.04594	5.12	0.0006

The fitted line is therefore $\hat{WT} = -1.88453 + 0.23507 * AGE$

Let's overlay the fitted line on our scatterplot.



- **As we might have guessed, the straight line model may not be the best choice.**

- **The "bowl" shape of the scatter plot does have a linear component, however.**

- **Without the plot, we might have believed the straight line fit is okay.**

Let's try a straight line model fit to Y=LOGWT versus X=AGE.

Partial listing of output ...

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-2.68925	0.03064	-87.78	<.0001
age	Age, X	1	0.19589	0.00268	73.18	<.0001

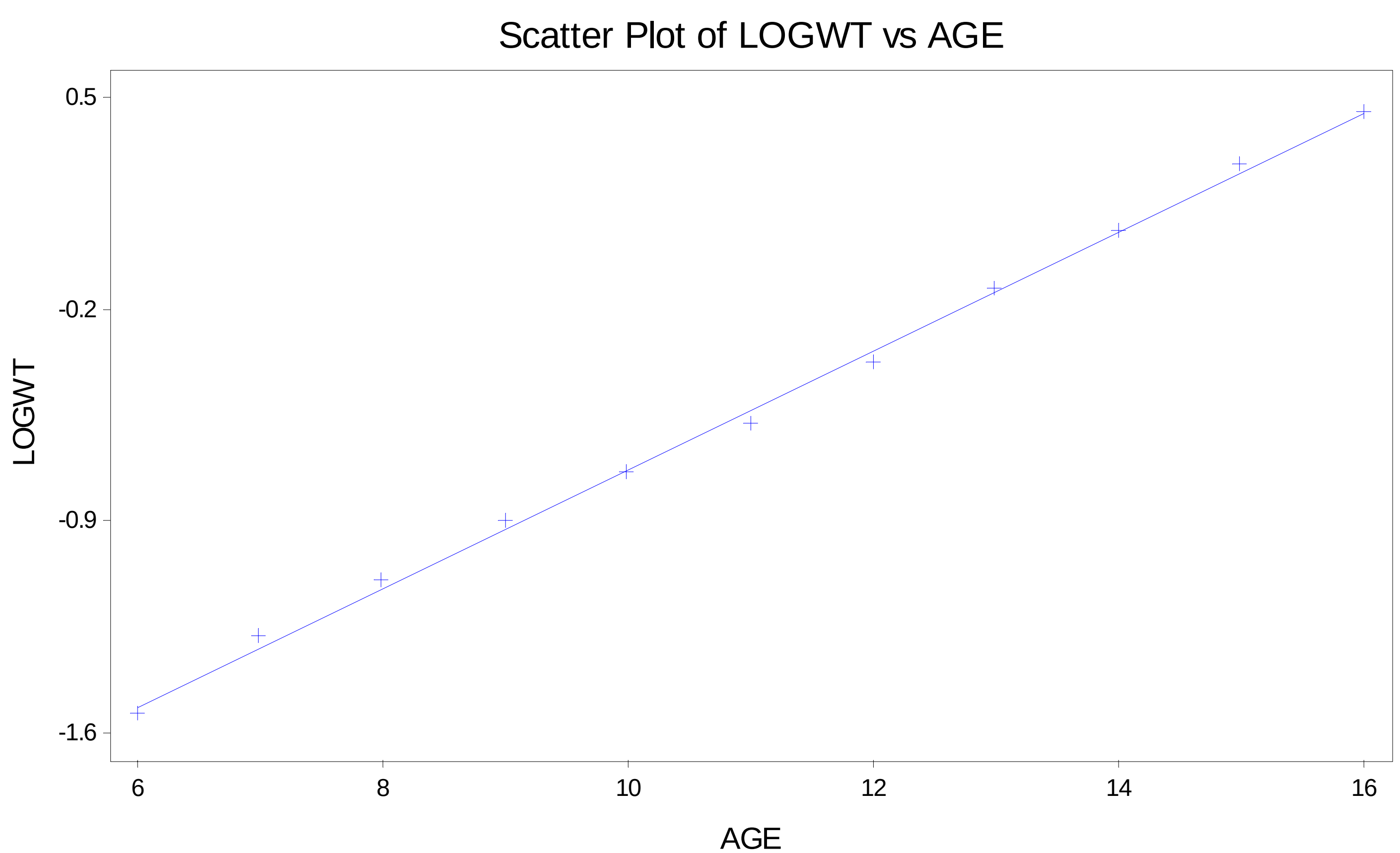
Annotated …

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-2.68925 = **intercept = β_0**	0.03064	-87.78	<.0001
Age	Age, X	1	0.19589 = **slope = β_1**	0.00268	73.18	<.0001

- **Thus, the fitted line is LOGWT = -2.68925 + 0.19589*AGE**

Now the scatterplot with the overlay of the fitted line looks much better.



Now You Try …

Prediction of Weight from Height
Source: Dixon and Massey (1969)

Individual	Height (X)	Weight (Y)
1	60	110
2	60	135
3	60	120
4	62	120
5	62	140
6	62	130
7	62	135
8	64	150
9	64	145
10	70	170
11	70	185
12	70	160

It helps to do the preliminary calculations

$\bar{X}=63.833$	$\bar{Y}=141.667$
$\sum X_i^2 = 49{,}068$	$\sum Y_i^2 = 246{,}100$
$\sum X_i Y_i = 109{,}380$	$S_{xx} = 171.667$
$S_{yy} = 5{,}266.667$	$S_{xy} = 863.333$

Slope	$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$	$\hat{\beta}_1 = \frac{863.333}{171.667} = 5.0291$
Intercept	$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$	$\hat{\beta}_0 = 141.667 - (5.0291)(63.833$ **= -179.3573**

3. The Analysis of Variance Table

In Topic 1, *Summarizing Data*, we learned that the numerator of the sample variance of the Y data is $\sum_{i=1}^{n}\left(\mathrm{Y}_{\mathrm{i}}-\overline{\mathrm{Y}}\right)^{2}$. In regression settings where Y is the outcome variable, this same quantity $\sum_{i=1}^{n}\left(\mathrm{Y}_{\mathrm{i}}-\overline{\mathrm{Y}}\right)^{2}$ is appreciated as the "total variance of the Y's". As we will see, other names for this are "total sum of squares", "total, corrected", and "SSY". ***(Note – "corrected" has to do with subtracting the mean before squaring.)***

An analysis of variance table is all about partitioning the total variance of the Y's (corrected) into two components:

1. **Due residual (the individual Y about the individual prediction $\hat{\mathrm{Y}}$)**
2. **Due regression (the prediction $\hat{\mathrm{Y}}$ about the overall mean $\overline{\mathrm{Y}}$)**

Aside - Why are we interested in such a partition?
We'd like to know if, within the data, there exists the suggestion of a linear relationship ("signal") that can be discerned from chance variability ("noise")

1) **the leftover variability of the observed Y_{i} about the predicted $\hat{\mathrm{Y}}_{\mathrm{i}}$ ("noise")**
2) **the explained variability of the predicted $\hat{\mathrm{Y}}_{\mathrm{i}}$ about the overall mean $\overline{Y}$ ("signal")**

Here is the partition (Note – Look closely and you'll see that both sides are the same)

$$\left(Y_i-\overline{Y}\right)=\left(Y_i-\hat{Y}_i\right)+\left(\hat{Y}_i-\overline{Y}\right)$$

Some algebra (not shown) reveals a nice partition of the total variability.

$$\sum(Y_i-\overline{Y})^2=\sum(Y_i-\hat{Y}_i)^2+\sum(\hat{Y}_i-\overline{Y})^2$$

Total Sum of Squares = Due Error Sum of Squares + Due Model Sum of Squares

A closer look…

Total Sum of Squares = Due Model Sum of Squares + Due Error Sum of Squares

$$\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2=\sum_{i=1}^{n}\left(\hat{Y}_i-\overline{Y}\right)^2+\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2$$

- $\left(Y_i-\overline{Y}\right)$ = **deviation of** Y_i **from** $\overline{Y}$ **that is to be explained**
- $\left(\hat{Y}_i-\overline{Y}\right)$ = **"due model", "signal", "systematic", "due regression"**
- $\left(Y_i-\hat{Y}_i\right)$ = **"due error", "noise", or "residual"**

In the world of regression analyses, we seek to *explain* the total variability $\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2$**:**

What happens when $\beta_1 \neq 0$?	**What happens when $\beta_1 = 0$?**
A straight line relationship is helpful	**A straight line relationship is not helpful**
Best guess is $\hat{Y}=\hat{\beta}_0+\hat{\beta}_1X$	**Best guess is** $\hat{Y}=\hat{\beta}_0=\overline{Y}$
Due model is LARGE because $\left(\hat{Y}-\overline{Y}\right)=(\left[\hat{\beta}_0+\hat{\beta}_1X\right]-\overline{Y})$ $=\overline{Y}-\hat{\beta}_1\overline{X}+\hat{\beta}_1X-\overline{Y}$ $=\hat{\beta}_1\left(X-\overline{X}\right)$	**Due error is nearly the TOTAL because** $\left(Y-\hat{Y}\right)=\left(Y-\left[\hat{\beta}_0\right]\right)=\left(Y-\overline{Y}\right)$
Due error has to be small	**Due regression has to be small**
$\dfrac{\text{due(model)}}{\text{due(error)}}$ **will be large**	$\dfrac{\text{due(model)}}{\text{due(error)}}$ **will be small**

How to Partition the Total Variance

1. **The "total" or "total, corrected" refers to the variability of Y about $\overline{Y}$**

   - $\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2$ **is called the "total sum of squares"**
   - **Degrees of freedom = df = (n-1)**
   - **Division of the "total sum of squares" by its df yields the "total mean square"**

2. **The "residual" or "due error" refers to the variability of Y about $\hat{Y}$**

   - $\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$ **is called the "residual sum of squares"**
   - **Degrees of freedom = df = (n-2)**
   - **Division of the "residual sum of squares" by its df yields the "residual mean square".**

3. **The "regression" or "due model" refers to the variability of $\hat{Y}$ about $\overline{Y}$**

   - $\sum_{i=1}^{n}\left(\hat{Y}_i - \overline{Y}\right)^2 = \hat{\beta}_1^2 \sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$ **is called the "regression sum of squares"**
   - **Degrees of freedom = df = 1**
   - **Division of the "regression sum of squares" by its df yields the "regression mean square" or "model mean square". This is an example of a variance component.**

Source	df	Sum of Squares	Mean Square
Regression	**1**	$SSR = \sum_{i=1}^{n}\left(\hat{Y}_i - \overline{Y}\right)^2$	**SSR/1**
Error	**(n-2)**	$SSE = \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$	**SSE/(n-2)**
Total, corrected	**(n-1)**	$SST = \sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2$	

Hint – Mean square = (Sum of squares)/(df)

Be careful! The question we may ask from an analysis of variance table is a limited one.

> **Does the fit of the straight line model explain a significant portion of the variability of the individual Y about $\bar{Y}$?**
>
> **Is this better than using $\bar{Y}$ alone?**

We are NOT asking:

Is the choice of the straight line model correct?
Would another functional form be a better choice?

We'll use a hypothesis test approach and the method of "proof by contradiction".

- **We begin with a null hypothesis that says $\beta_1 = 0$ ("no linear relationship")**
- **Evaluation will focus on the comparison of the due regression mean square to the residual mean square**
- **Recall that we reasoned the following:**

  If $\beta_1 \neq 0$ Then due(regression)/due(residual) will be LARGE
  If $\beta_1 = 0$ Then due(regression)/due(residual) will be SMALL

- **Our p-value calculation will answer the question:**
  If $\beta_1 = 0$ truly, what are the chances of obtaining an value of due(regression)/due(residual) as larger or larger than that observed?

To calculate "chances" we need a probability model.

So far, we have not needed one.

4. Assumptions for a Straight Line Regression Analysis

In performing least squares estimation, we did not use a probability model. We were doing geometry.

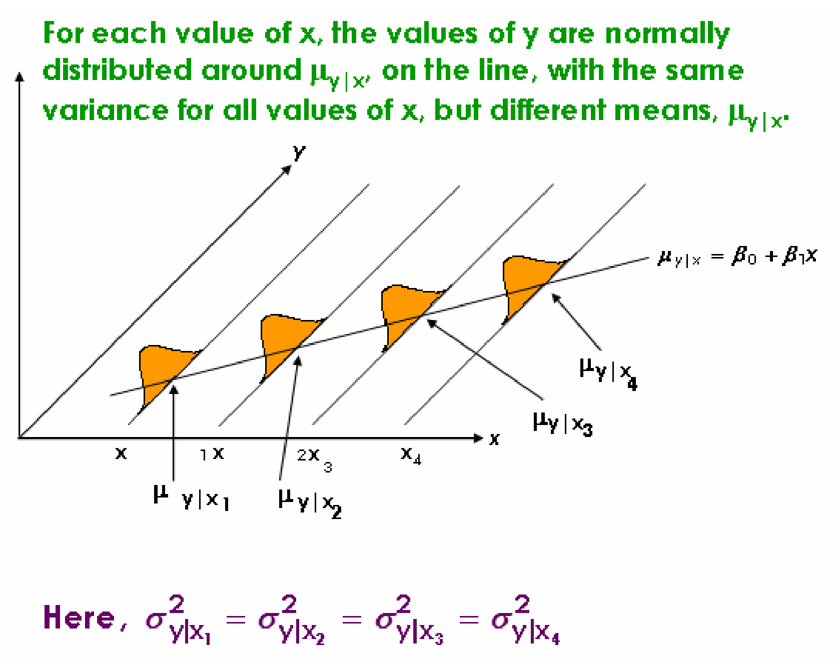
Hypothesis testing requires some assumptions and a probability model.

Assumptions

- **The separate observations $Y_1, Y_2, \cdots, Y_n$ are independent.**
- **The values of the predictor variable X are fixed and measured without error.**
- **For each value of the predictor variable X=x, the distribution of values of Y follows a normal distribution with mean equal to $\mu_{Y|X=x}$ and common variance equal to $\sigma_{Y|x}^2$.**
- **The separate means $\mu_{Y|X=x}$ lie on a straight line; that is –**

$$\mu_{Y|X=x} = \beta_0 + \beta_1 X$$

Schematically, here is what the situation looks like (courtesy: Stan Lemeshow)



With these assumptions, we can assess the significance of the variance explained by the model.

$$F = \frac{\text{msq(model)}}{\text{msq(residual)}} \quad \textbf{with df = 1, (n-2)}$$

$\beta_1 = 0$	$\beta_1 \neq 0$
Due model MSR has expected value $\sigma_{Y\|X}^2$	**Due model MSR has expected value** $\sigma_{Y\|X}^2 + \beta_1^2 \sum_{i=1}^{n} (X_i - \overline{X})^2$
Due residual MSE has expected value $\sigma_{Y\|X}^2$	**Due residual MSE has expected value** $\sigma_{Y\|X}^2$
F = (MSR)/MSE will be close to 1	**F = (MSR)/MSE will be LARGER than 1**

We obtain the analysis of variance table for the model of Y=LOGWT to X=AGE:

The following is in SAS with annotations in red.

```
                                    Analysis of Variance

                                    Sum of        Mean
Source                   DF        Squares       Square             F Value                Pr > F
                                                          = MSQ(Regression)/MSQ(Residual)

Model                     1        4.22106      4.22106             5355.60                <.0001
Error                     9        0.00709      0.00078816
Corrected Total          10        4.22815


                  Root MSE              0.02807     R-Square     0.9983 = SSQ(regression)/SSQ(total)
                  Dependent Mean       -0.53445     Adj R-Sq     0.9981 = R² adjusted for n and # predictors
                  Coeff Var            -5.25286
```

This output corresponds to the following.

Source	Df	Sum of Squares	Mean Square
Regression	**1**	$SSR = \sum_{i=1}^{n}\left(\hat{Y}_i - \overline{Y}\right)^2$ **= 4.22063**	**SSR/1 = 4.22063**
Error	**(n-2) = 9**	$SSE = \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$ **= 0.00705**	**SSE/(n-2) = 7/838E-04**
Total, corrected	**(n-1) = 10**	$SST = \sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2$ **= 4.22768**	

Other information in this output:

- **R-SQUARED = Sum of squares regression/Sum of squares total is the proportion of the "total" that we have been able to explain with the fit of the straight line model.**

  - *Be careful!* As predictors are added to the model, R-SQUARED can only increase. Eventually, we need to "adjust" this measure to take this into account. See ADJUSTED R-SQUARED.

- **We also get an overall F test of the null hypothesis that the simple linear model does not explain significantly more variability in LOGWT than the average LOGWT. F = MSQ (Regression)/MSQ (Residual)**

  = 4.22063/0.0007838
  = 5384.94 with df =1, 9

  Achieved significance < 0.0001. Reject H_O. Conclude that the fitted line is a significant improvement over the average LOGWT.

5. Hypothesis Testing

Straight Line Model: $Y = \beta_0 + \beta_1 X$

1) **Overall F-Test**
2) **Test of slope**
3) **Test of intercept**

1) Overall F-Test

Research Question: **Does the fitted model, the $\hat{Y}$ explain significantly more of the total variability of the Y about $\overline{Y}$ than does $\overline{Y}$?**

Assumptions: **As before.**

H_O and H_A:

$$H_o: \beta_1 = 0$$
$$H_A: \beta_1 \neq 0$$

Test Statistic:

$$F = \frac{msq(regresion)}{msq(residual)}$$
$$df = 1, (n-2)$$

Evaluation rule:

When the null hypothesis is true, the value of F should be close to 1. Alternatively, when $\beta_1 \neq 0$, the value of F will be LARGER than 1.

Thus, our p-value calculation answers: "What are the chances of obtaining our value of the F or one that is larger if we believe the null hypothesis that $\beta_1 = 0$"?

Calculations:

For our data, we obtain p-value =

$$pr\left[F_{1,(n-2)} \geq \frac{msq(\text{mod}\,el)}{msq(residual)} | \beta_1 = 0\right] = pr\left[F_{1,9} \geq 5384.94\right] << .0001$$

Evaluate:

Under the null hypothesis that $\beta_1 = 0$, the chances of obtaining a value of F that is so far away from the expected value of 1, with a value of 5394.94, is less than 1 chance in 10,000. This is a very small likelihood!

Interpret:

We have learned that, at least, the fitted straight line model does a much better job of explaining the variability in LOGWT than a model that allows only for the average LOGWT.

… later … (BE640, Intermediate Biostatistics), we'll see that the analysis does not stop here …

2) Test of the Slope, β_1

Some interesting notes!

- **The overall F test and the test of the slope are equivalent.**
  - **The test of the slope uses a t-score approach to hypothesis testing**
  - **It can be shown that { t-score for slope $\}^2$ = { overall F }**

Research Question: **Is the slope $\beta_1 = 0$?**

Assumptions: **As before.**

H_O and H_A:

$$H_O: \beta_1 = 0$$
$$H_A: \beta_1 \neq 0$$

Test Statistic:

To compute the t-score, we need an estimate of the standard error of $\hat{\beta}_1$

$$S\hat{E}\left(\hat{\beta}_1\right) = \sqrt{msq(residual)\left[\frac{1}{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}\right]}$$

Our t-score is therefore:

$$t - score = \left[\frac{(observed) - (expected)}{\hat{se}(expected)} \right] = \left[\frac{\left(\hat{\beta}_1\right) - (0)}{\hat{se}\left(\hat{\beta}_1\right)} \right]$$

$$df = (n-2)$$

We can find this information in our output:

The following is in SAS with annotations in red.

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value = Estimate/Error	Pr > |t|
Intercept	Intercept	1	-2.68925	0.03064	-87.78	<.0001
age	Age, X	1	0.19589	0.00268	73.18 = 0.19589/0.00268	<.0001

Recall what we mean by a t-score:

T=73.38 says "the estimated slope is estimated to be 73.38 standard error units away from its expected value of zero".

Check that { t-score }^{2} = { Overall F }:

[73.38]^2 = 5384.62 which is close.

Evaluation rule:

When the null hypothesis is true, the value of t should be close to zero. Alternatively, when $\beta_1 \neq 0$, the value of t will be DIFFERENT from 0.

Here, our p-value calculation answers: "What are the chances of obtaining our value of the t or one that is more far away from 0 if we believe the null hypothesis that $\beta_1 = 0$"?

Calculations:

For our data, we obtain p-value =

$$2\,pr\left[t_{(n-2)} \geq \left| \frac{\hat{\beta}_1 - 0}{\hat{se}\left(\hat{\beta}_1\right)} \right| \right] = 2\,pr\left[t_9 \geq 73.38 \right] << .0001$$

Evaluate:

Under the null hypothesis that $\beta_1 = 0$, the chances of obtaining a t-score value that is 73.38 or more standard error units away from the expected value of 0 is less than 1 chance in 10,000.

Interpret:

The inference is the same as that for the overall F test. The fitted straight line model does a much better job of explaining the variability in LOGWT than the sample mean.

3) Test of the Intercept, β_0

This pertains to whether or not the straight line relationship passes through the origin. It is rarely of interest.

***Research Question*: Is the intercept $\beta_0 = 0$?**

***Assumptions:* As before.**

H_O and H_A:

$$H_O: \beta_0 = 0$$
$$H_A: \beta_0 \neq 0$$

Test Statistic:

To compute the t-score for the intercept, we need an estimate of the standard error of $\hat{\beta}_0$

$$\hat{SE}\left(\hat{\beta}_0\right) = \sqrt{msq(residual)\left[\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}\right]}$$

Our t-score is therefore:

$$t - score = \left[\frac{(observed) - (expected)}{\hat{se}(expected)} \right] = \left[\frac{\left(\hat{\beta}_0\right) - (0)}{\hat{se}\left(\hat{\beta}_0\right)} \right]$$

$$df = (n-2)$$

We can find this information in our output:

The following is in SAS with annotations in red.

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value = Estimate/Error	Pr > |t|
Intercept	Intercept	1	-2.68925	0.03064	-87.78 = -2.68925/0.03064	<.0001
age	Age, X	1	0.19589	0.00268	73.18	<.0001

This t=-87.78 says "the estimated intercept is estimated to be 87.78 standard error units away from its expected value of zero".

Evaluation rule:

When the null hypothesis is true, the value of t should be close to zero. Alternatively, when $\beta_0 \neq 0$, the value of t will be DIFFERENT from 0.

Our p-value calculation answers: "What are the chances of obtaining our value of the t or one that is more far away from 0 if we believe the null hypothesis that $\beta_0 = 0$"?

Calculations:

p-value =

$$2\,pr\left[t_{(n-2)} \geq \left| \frac{\hat{\beta}_0 - 0}{\hat{se}\left(\hat{\beta}_0\right)} \right| \right] = 2\,pr\left[t_9 \geq 87.78 \right] << .0001$$

Evaluate:

Under the null hypothesis that $\beta_0 = 0$, the chances of obtaining a t-score value that is 87.78 or more standard error units away from the expected value of 0 is less than 1 chance in 10,000.

Interpret:

The inference is that the straight line relationship between Y=LOGWT and X=AGE does not pass through the origin.

6. Confidence Interval Estimation

Straight Line Model: $Y = \beta_0 + \beta_1 X$

Recall (*Topic 6, Estimation*) that there are 3 elements of a confidence interval:

1) **Best single guess (estimate)**
2) **Standard error of the best single guess (SE[estimate])**
3) **Confidence coefficient**

- **These will be percentiles from the t distribution with df=(n-2)**
- **For a 95% confidence interval, this will be a 97.5^{th} percentile**
- **For a $(1-\alpha)100\%$ confidence interval, this will be a $(1-\alpha/2)100^{th}$ percentile.**

The generic form of a confidence interval is then

Generic Form of Confidence Interval
Straight Line Model: $Y = \beta_0 + \beta_1 X$

Lower limit = (Estimate) - (confidence coefficient)*SE(estimate)
Upper limit = (Estimate) + (confidence coefficient)*SE(estimate)

We might want confidence interval estimates of the following 4 parameters:

(1) **Slope**
(2) **Intercept**
(3) **Mean of subset of population for whom $X=x_0$**
(4) **Individual response for person for whom $X=x_0$**

1) SLOPE **estimate =** $\hat{\beta}_1$

$$\hat{se}\left(\hat{b}_1\right) = \sqrt{\text{msq(residual)}\,\frac{1}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}}$$

2) INTERCEPT **estimate =** $\hat{\beta}_0$

$$\hat{se}\left(\hat{b}_0\right) = \sqrt{\text{msq(residual)}\left[\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}\right]}$$

3) MEAN at X=x_0 **estimate =** $\hat{Y}_{X=x_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0$

$$\hat{se} = \sqrt{\text{msq(residual)}\left[\frac{1}{n} + \frac{\left(x_0 - \overline{X}\right)^2}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}\right]}$$

4) INDIVIDUAL with X=x_0 **estimate =** $\hat{Y}_{X=x_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0$

$$\hat{se} = \sqrt{\text{msq(residual)}\left[1 + \frac{1}{n} + \frac{\left(x_0 - \overline{X}\right)^2}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}\right]}$$

Illustration for the model which fits Y=LOGWT to X=AGE.

Recall that we obtained the following fit:

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-2.68925	0.03064	-87.78	<.0001
age	Age, X	1	0.19589	0.00268	73.18	<.0001

95% Confidence Interval for the Slope, β_1

1) **Best single guess (estimate) =** $\hat{\beta}_1 = 0.19589$

2) **Standard error of the best single guess (SE[estimate]) =** $se\left(\hat{\beta}_1\right) = 0.00268$

3) **Confidence coefficient = 97.5th percentile of Student t =** $t_{.975,df=9} = 2.26$

95% Confidence Interval for Slope β_1 = Estimate $\pm$ (confidence coefficient)*SE
= 0.19589 $\pm$ (2.26)(0.00268)
= (0.1898, 0.2019)

95% Confidence Interval for the Intercept, β_0

1) **Best single guess (estimate) =** $\hat{\beta}_0 = -2.68925$

2) **Standard error of the best single guess (SE[estimate]) =** $se\left(\hat{\beta}_0\right) = 0.03064$

3) **Confidence coefficient = 97.5th percentile of Student t =** $t_{.975,df=9} = 2.26$

95% Confidence Interval for Slope β_0 = Estimate $\pm$ (confidence coefficient)*SE
= -2.68925 $\pm$ (2.26)(0.03064)
= (-2.7585,-2.6200)

Confidence Intervals for Predictions

Code.

```
proc reg data=temp alpha=.05;         /* alpha=.05 is type I error */
  title "Regression of Y=Weight on X=Age";
  model wt=age/cli clm;              /*cli for individual, clm for mean */
run;
quit;
```

Output.

Output Statistics

Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict		Residual
1	-1.5380	-1.5139	0.0158	-1.5497	-1.4781	-1.5868	-1.4410	-0.0241
2	-1.2840	-1.3180	0.0136	-1.3489	-1.2871	-1.3886	-1.2474	0.0340
3	-1.1020	-1.1221	0.0117	-1.1485	-1.0957	-1.1909	-1.0534	0.0201
4	-0.9030	-0.9262	0.0100	-0.9489	-0.9036	-0.9937	-0.8588	0.0232
5	-0.7420	-0.7303	0.008878	-0.7504	-0.7103	-0.7970	-0.6637	-0.0117
6	-0.5830	-0.5345	0.008465	-0.5536	-0.5153	-0.6008	-0.4681	-0.0485
7	-0.3720	-0.3386	0.008878	-0.3586	-0.3185	-0.4052	-0.2720	-0.0334
8	-0.1320	-0.1427	0.0100	-0.1653	-0.1200	-0.2101	-0.0752	0.0107
9	0.0530	0.0532	0.0117	0.0268	0.0796	-0.0156	0.1220	-0.000218
10	0.2750	0.2491	0.0136	0.2182	0.2800	0.1785	0.3197	0.0259
11	0.4490	0.4450	0.0158	0.4092	0.4808	0.3721	0.5179	0.004000

7. Introduction to Correlation

Definition of Correlation

A correlation coefficient is a measure of the association between two paired random variables (e.g. height and weight).

The Pearson product moment correlation, in particular, is a measure of the strength of the *straight line* relationship between the two random variables.

Another correlation measure (not discussed here) is the Spearman correlation. It is a measure of the strength of the *monotone increasing (or decreasing)* relationship between the two random variables. The Spearman correlation is a non-parametric (meaning model free) measure. It is introduced in PubHlth 640, Intermediate Biostatistics.

Formula for the Pearson Product Moment Correlation ρ

- **The population parameter designation is rho, written as ρ**
- **The estimate of ρ, based on information in a sample is represented using r.**
- **Some preliminaries:**

(1) Suppose we are interested in the correlation between X and Y

(2) $$\hat{\text{cov}}(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)} = \frac{S_{xy}}{(n-1)}$$ **This is the covariance(X,Y)**

(3) $$\hat{\text{var}}(X) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{(n-1)} = \frac{S_{xx}}{(n-1)}$$ **and similarly**

(4) $$\hat{\text{var}}(Y) = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{(n-1)} = \frac{S_{yy}}{(n-1)}$$

Formula for Estimate of Pearson Product Moment Correlation from a Sample

$$\hat{\rho} = r = \frac{\widehat{\text{cov}}(x,y)}{\sqrt{\widehat{\text{var}}(x)\widehat{\text{var}}(y)}}$$

$$= \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

If you absolutely have to do it by hand, an equivalent (more calculator friendly formula) is

$$\hat{\rho} = r = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sqrt{\left[\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right]}\sqrt{\left[\sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}\right]}}$$

- **The correlation r can take on values between 0 and 1 only**
- **Thus, the correlation coefficient is said to be dimensionless – it is independent of the units of x or y.**
- **Sign of the correlation coefficient (positive or negative) = Sign of the estimated slope $\hat{\beta}_1$.**

There is a relationship between the slope of the straight line, $\hat{\beta}_1$, and the estimated correlation r.

Relationship between slope $\hat{\beta}_1$ and the sample correlation r

Because $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$ **and** $r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$

A little algebra reveals that

$$r = \left[\frac{\sqrt{S_{xx}}}{\sqrt{S_{yy}}}\right]\hat{\beta}_1$$

Thus, beware!!!

- **It is possible to have a very large (positive or negative) r might accompanying a very non-zero slope, inasmuch as**

  - **A very large r might reflect a very large S_{xx} , all other things equal**

  - **A very large r might reflect a very small S_{yy} , all other things equal.**

8. Hypothesis Test of Correlation

The null hypothesis of zero correlation is equivalent to the null hypothesis of zero slope.

Research Question: **Is the correlation $\rho = 0$? Is the slope $\beta_1 = 0$?**

Assumptions: **As before.**

H_O and H_A:

$$H_O : \rho = 0$$
$$H_A : \rho \neq 0$$

Test Statistic:

A little algebra (not shown) yields a very nice formula for the t-score that we need.

$$t-score = \left[\frac{r\sqrt{(n-2)}}{\sqrt{1-r^2}} \right]$$
$$df = (n-2)$$

We can find this information in our output. Recall the first example and the model of Y=LOGWT to X=AGE:

The Pearson Correlation, r, is the $\sqrt{\text{R-squared}}$ in the output.

```
Root MSE              0.02807    R-Square     0.9983
Dependent Mean       -0.53445    Adj R-Sq     0.9981

Coeff Var            -5.25286

Pearson Correlation, r = √0.9983 = 0.9991
```

Substitution into the formula for the t-score yields

$$t-score=\left[\frac{r\sqrt{(n\text{-}2)}}{\sqrt{1-r^2}}\right]=\left[\frac{.9991\sqrt{9}}{\sqrt{1\text{-}.9983}}\right]=\left[\frac{2.9974}{.0412}\right]=72.69$$

Note: The value .9991 in the numerator is $r=\sqrt{R^2}=\sqrt{.9983}=.9991$

This is very close to the value of the t-score that was obtained for testing the null hypothesis of zero slope. The discrepancy is probably rounding error. I did the calculations on my calculator using 4 significant digits. SAS probably used more significant digits - cb.