Chapter 11: Linear Regression and Correlation

1. Simple Linear Regression
2. Fitted Regression Line
3. Checking Regression Assumptions
4. Inferences about the Slope, Intercept and Error Variance
5. Regression Special Cases
   A. Regression with zero intercept
   B. Regression with centered x's
6. Inference concerning average and new Y values at X = *x*.
7. Transformations
8. Checking for Outliers
9. F-test for "lack of fit"
10. Adding quadratic terms
11. Calibration Problem (inverse prediction)
12. Correlation
13. Sample Size and Power for Slope and Pearson Correlation
14. Additional comments and examples

Chapter 11 Examples

1. Corn Example: Regression
2. Intercept example
3. Stopping Distance: Transformation example
4. Florida Election: Transformation and Outliers
5. Corn Example: Lack of Fit test
6. Flow Rate: Calibration example
7. Fat States: Correlation example

0. Some Perspective

In STAT511 we focus on analyses with a single response (or dependent) variable (Y) and a single predictor (or independent) variable (X). In R: `lm(Y ~ X)`, `plot(Y ~ X)`

- **Continuous response** with a **categorical predictor** -> two-sample t-test (CH6) or one-way ANOVA (CH8)

  Example Rat lead: Y = amount of solution consumed (#),
  X = treatment group (Control or Deficient)
- **Categorical response** with **categorical predictor** -> chi-squared test, FET or Z-test for proportions (CH10)

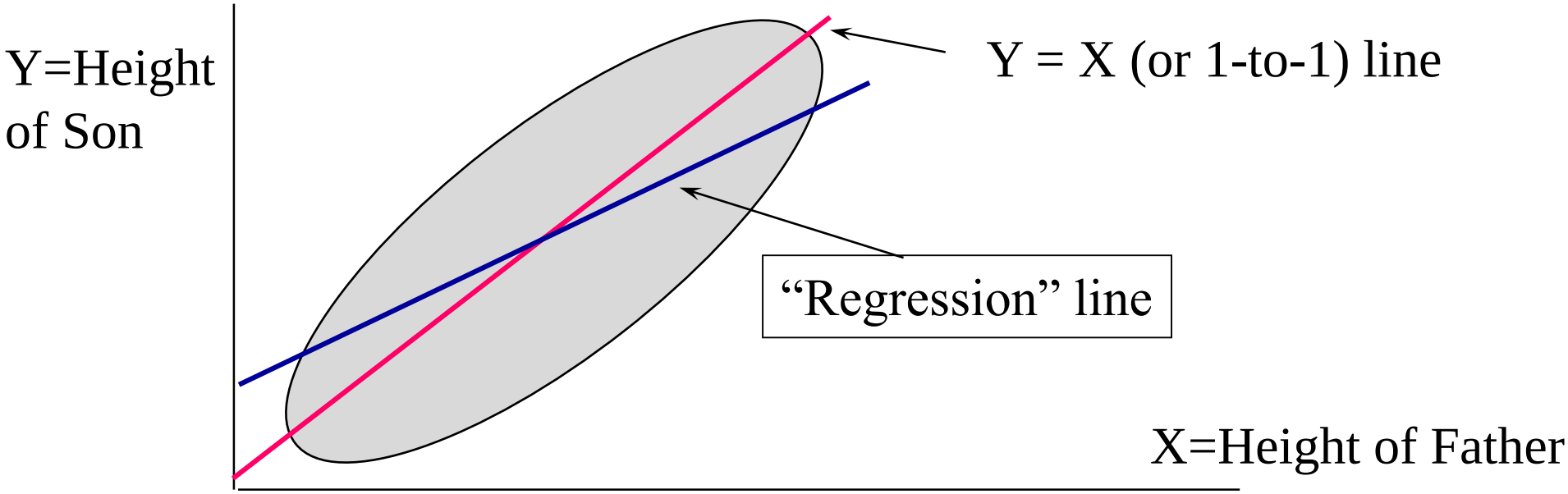
  Example French Skiers: Y = cold status (Yes or No),
  X = treatment group (Vitamin C or Placebo).
- **Continuous response** with **continuous predictor** -> consider simple linear regression (CH11)

  Example Corn Yield: Y = corn yield (#), X = fertilizer (#)
- **Categorical (binary) response** with **continuous predictor** -> logistic regression (Extra Topics 2)

  Example Beetle Kill: Y = status (dead or alive), X = pesticide dose (#)

1. Simple Linear Regression

History: Sir Francis Galton (1888) studied the relationship between heights of fathers and heights of sons for 1078 father-son pairs



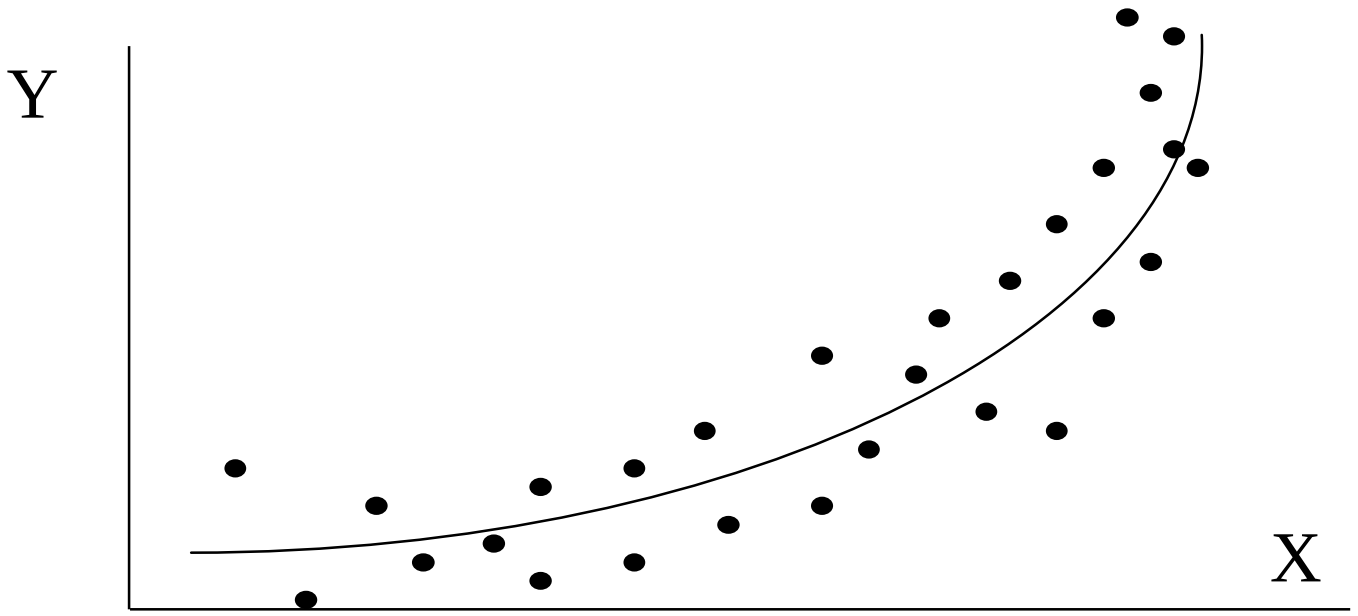
For each height of father, we mark the average height of sons having fathers that height. Then draw a line connecting these averages.

1) Among fathers of above average height, average height of sons is above average, but not by as much as the fathers.
2) Among fathers of below average height, average height of sons is below average, but not by as much as the fathers.

Galton wrote "this line represents a **regression toward mediocrity".**

Regression of Y on X

Definition: the **regression of Y on X** is a line (could be curved) that gives the **average value of Y for each given value of X**.



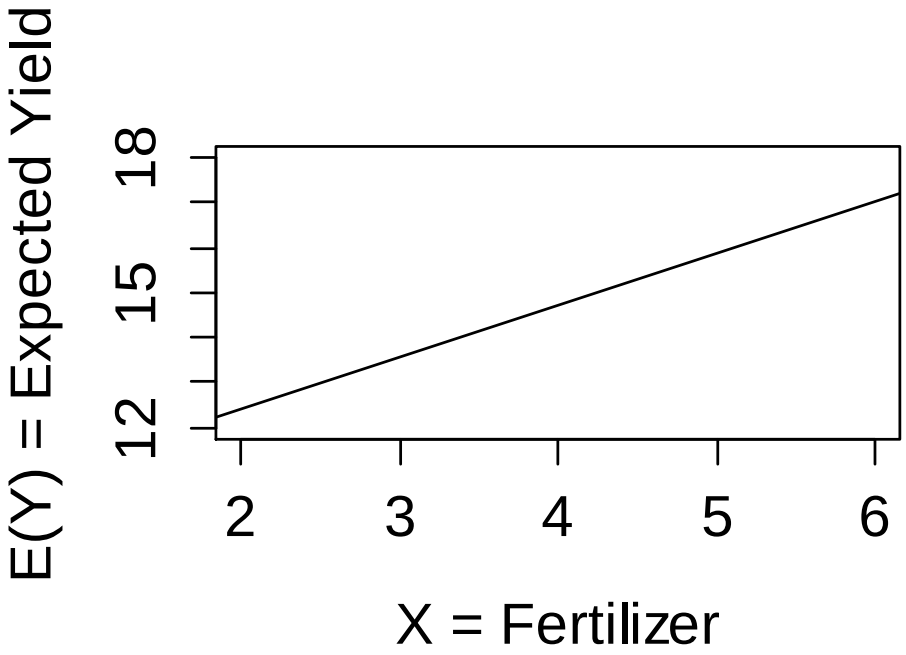
Note: In this group of notes, will primarily focus on studies in which the regression line can be adequately modeled by a straight line. In other words, linear regression. We will (briefly) discuss non-linear regression in the Extra Topics 2 notes.

Regression Model

Corn Example:
Response of corn yield (bu/plot) to fertilizer (lbs/plot). We assume that in the field in which we are to use, the yield of corn (Y) will be on average a linear function of the fertilizer applied (X).

$$E(Y) = \beta_0 + \beta_1 x$$

E(Y) is read as "the expected value of Y" and means the average value that would be seen if a very large number of plots with that x were used.



Parameter (Slope and Intercept) Interpretation

If we apply x=0 lbs per plot:

$$E(Y) = \beta_0 + \beta_1(0) = \beta_0$$

Therefore, β_0 = the predicted average yield if no fertilizer is applied.
= **intercept**

If we increase fertilizer from x lbs to (x+1) lbs the <u>change</u> in yield is:

$$\Delta E(Y) = \left(\beta_0 + \beta_1(x+1)\right) - \left(\beta_0 + \beta_1(x)\right) = \beta_1$$

Therefore, β_1 = the predicted increase (decrease, if negative) in average yield (Y) associated with a 1 unit increase in fertilizer applied (X).
= **slope (This is unit dependent – bu./lb.)**

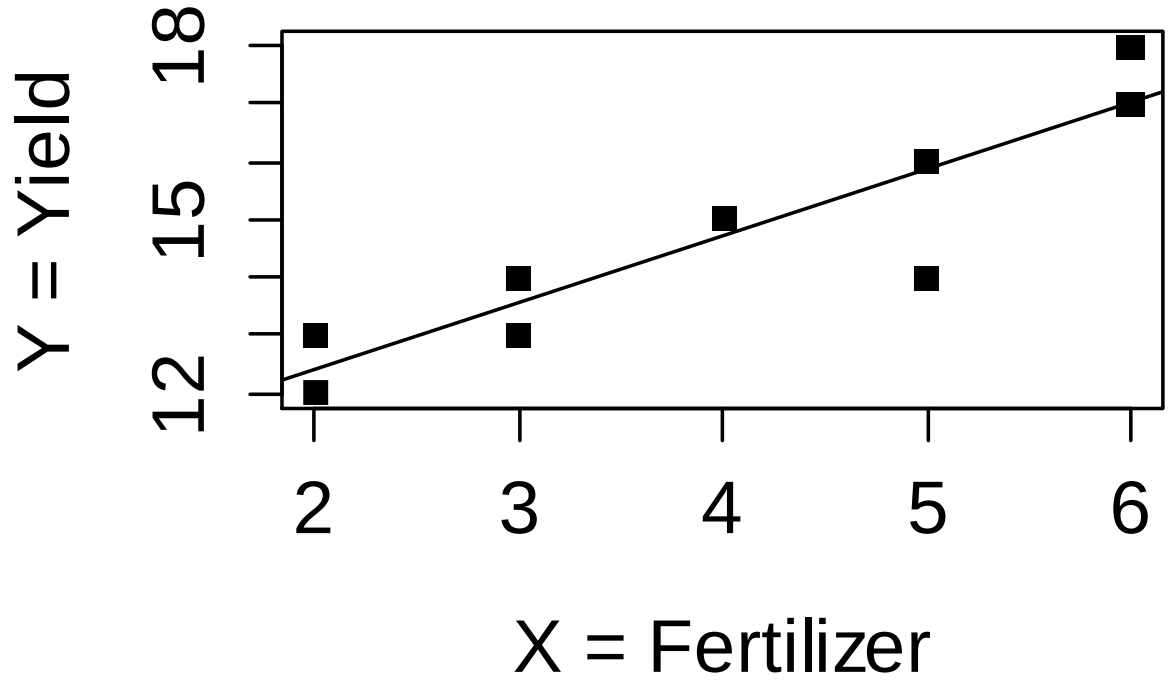
The Error Term in the Simple Linear Regression Model

When we conduct an experiment, we observe data on *n* individual points that do not equal the expected values, so we add a random "error" term to account for the difference. For the ith plot, we have

$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ i=1,2,...,n

Standard deviation of $\varepsilon_i = \sigma_\varepsilon$ (same for all x, equal variance)

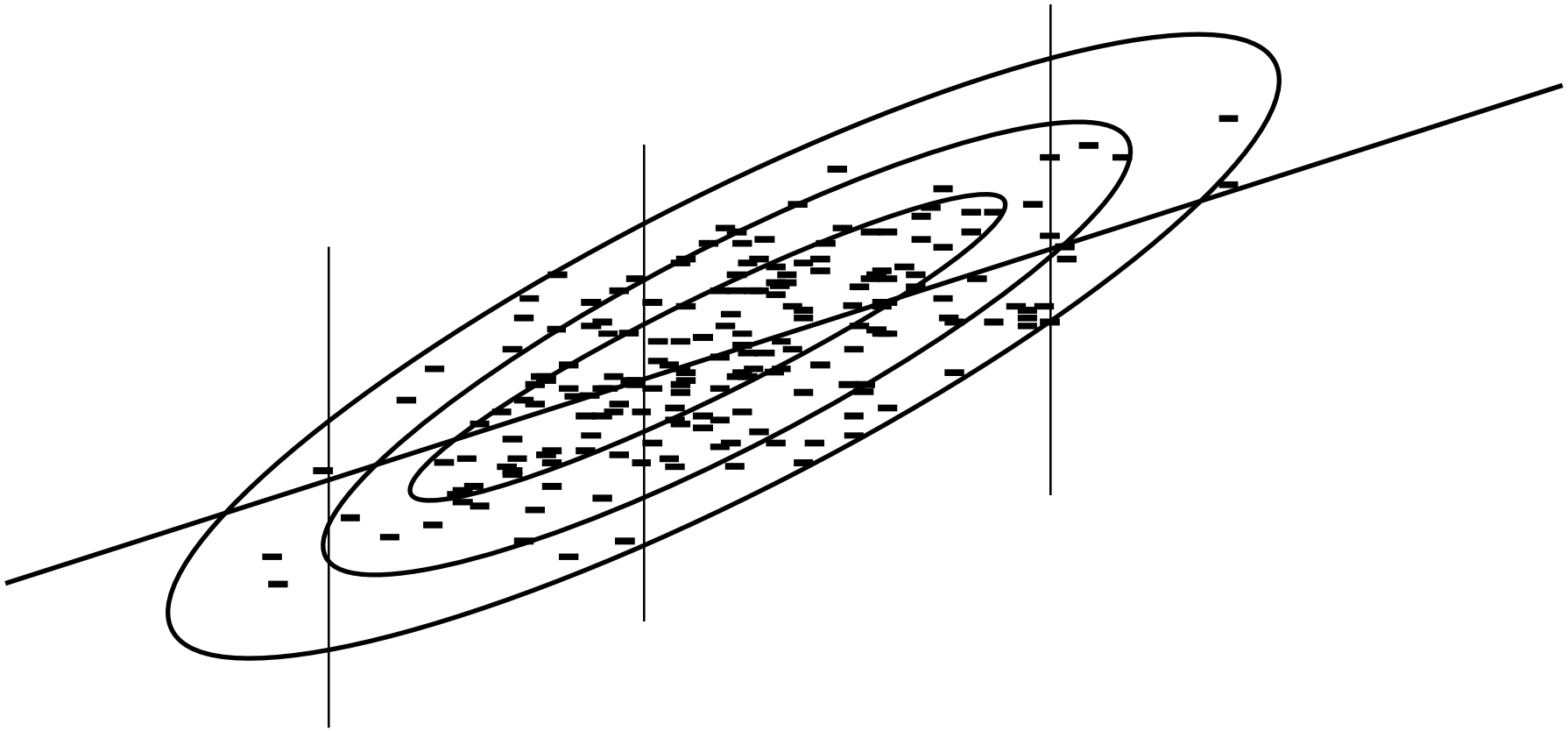
$E(\varepsilon_i) = 0$ (for each x) and ε_i's are independent.



Simple Linear Regression *(Notes)*

1. In regression, the x's are treated as if they were fixed, whether they were really fixed (as in this example) or random (as in the Father's heights)
2. The β_0 and β_1 values are unknown population parameters, hence the regression line ($\beta_0 + \beta_1 x$) is an unknown parameter. σ is also a parameter, called the "error standard deviation".
3. Galton (and others) noticed that for many types of bivariate data in the study of natural populations:
    - The regression line was straight.
    - The errors were homoscedastic (equal variance).
    - Contour lines of equal point frequency (or density) were concentric ellipses.
    - X and Y variables taken individually were normally distributed. Data of this type is "bivariate normal".

Bivariate Normal Distribution



2. The Fitted Regression Line

The **least squares estimate of the regression line** is a line (equation) that defines the linear relationship between the response variable (**y**) and the predictor variable (**x**):

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

So to obtain the **fitted or predicted value** of y (denoted $\hat{y}_i$) for a particular value of x, we just "plug in" to the equation:

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

A **residual** is the difference between the observed and predicted values:

$$e_i = y_i - \hat{y}_i$$

The **residual sum of squares (SSResid)** or **sum of squares error (SSE)** is defined as:

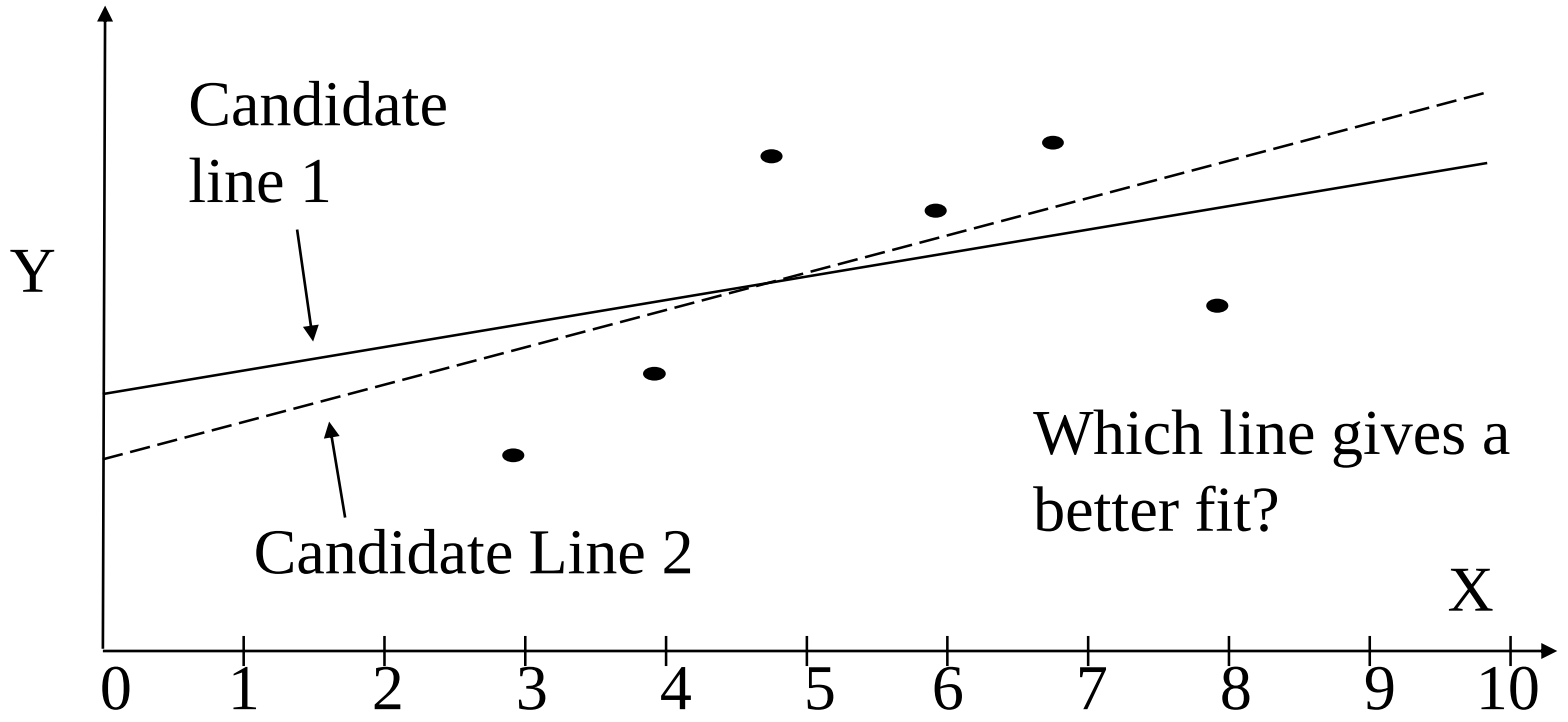
$$\text{SSResid} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

Least Squares Criteria

The "best" straight line is the one that minimizes the residual sum of squares (SSResid as defined on the previous slide).

The formulas for $\hat{\beta}_0$ and $\hat{\beta}_1$ were derived from the least-squares criteria by taking partial derivatives to solve the minimization problem.

Web Demo: https://www.desmos.com/calculator/zvrc4lg3cr



Formulas for least squares estimates

Define $S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - \frac{(\Sigma x_i)(\Sigma y_i)}{n}$

$S_{xx} = (n-1)s_x^2 = \sum (x_i - \bar{x})^2 = \sum x_i^{\,2} - \frac{(\Sigma x_i)^2}{n}$

$S_{yy} = (n-1)s_y^2 = \sum (y_i - \bar{y})^2 = \sum y_i^{\,2} - \frac{(\Sigma y_i)^2}{n}$

Estimate of Slope: $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$

Estimate of Intercept: $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

These formulas are derived by taking derivatives of SSResid with respect to the slope and intercept parameters, setting the derivatives equal to zero, then solving the two equations simultaneously.

In practice, we estimate the slope and intercept using lm().

Simple Linear Regression *(Example)*

Corn and fertilizer experiment with data given in the "**Corn Regression**" example.

Here we calculate the estimated slope and intercept "by hand", but the values are also given in the computer output.

$\bar{x} = 4.0 \qquad \bar{y} = 14.7$

$S_{xy} = 23 \qquad S_{xx} = 20$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{23}{20} = 1.15$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 14.7 - (1.15)(4.0) = 10.10$$

Interpretation of Estimates:

Intercept: Predicted yield (y) with no fertilizer applied (x=0) is estimated to average 10.10 bu/plot. (This prediction should be taken with caution, because 0 is beyond the range of the data).

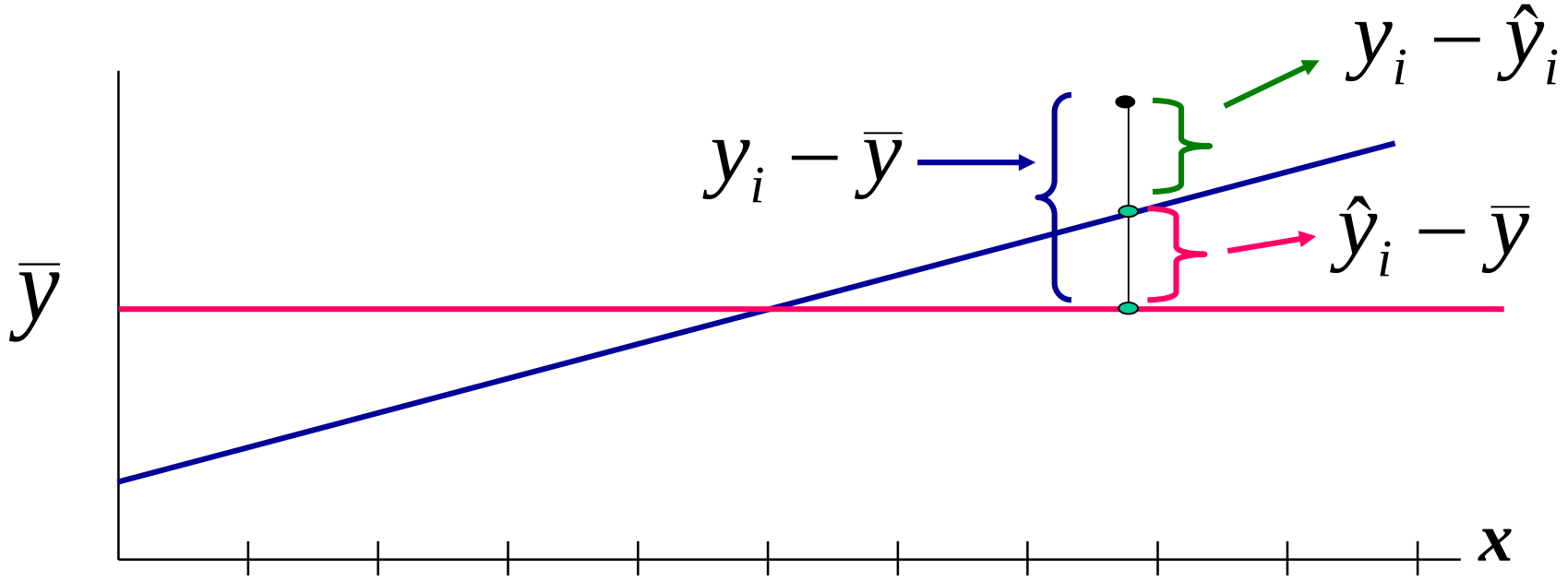
Slope: A one lb/plot increase in fertilizer (1 unit increase in x) is associated with a 1.15 bu/plot predicted increase in yield (y).

Partitioning variability and estimating σ

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

SSTotal = SSReg + SSResid

Variability in Y = Variability explained by X + Unexplained Variability



Estimating σ_ε^2

$e_i = y_i - \hat{y}_i$

= deviation from regression line

= residual

$\text{SSResid} = \sum e_i^2 =$ sum of squares of residuals (errors)

$$\hat{\sigma}_\varepsilon^2 = s_\varepsilon^2 = \frac{\sum (y_i - \hat{y}_i)^2}{n-2} = \frac{\text{SSResid}}{n-2} = \text{MSResid}$$

$$\hat{\sigma}_\varepsilon = s_\varepsilon = \sqrt{\frac{\text{SSResid}}{n-2}} = \sqrt{\text{MSResid}}$$

Can be computed as $\text{SSResid} = S_{yy} - \hat{\beta}_1 S_{xy}$

Example *(Corn Regression-- continued)*

$\text{SSResid} = S_{yy} - \hat{\beta}_1 S_{xy}$

$= 32.10 - (1.15)23 = 5.65$

$\text{MSResid} = \frac{\text{SSResid}}{n-2} = \frac{5.65}{10-2} = 0.706$

$s_\varepsilon = \sqrt{\text{MSResid}} = \sqrt{0.706}$

$= 0.840$

In R: s_ε= sqrt(MSResid) = "Residual Standard Error"

3. Checking Regression assumptions

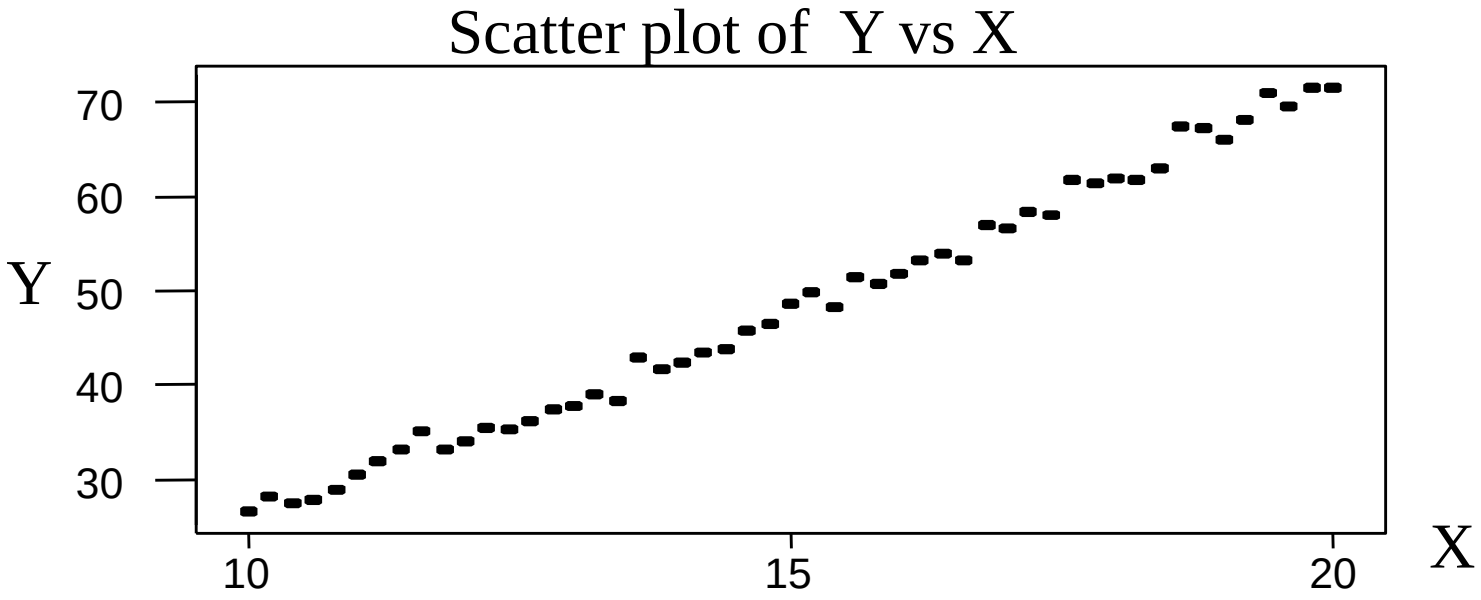
The simple linear regression model carries with it some assumptions:

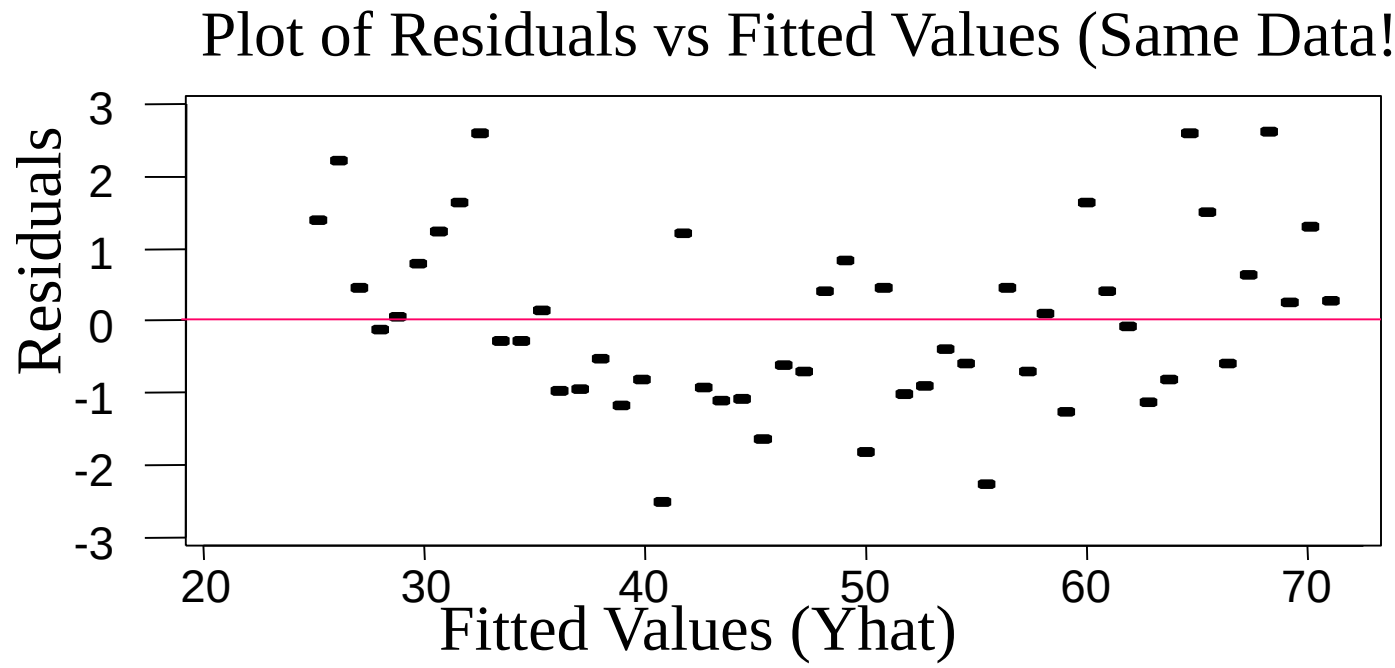
$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad i = 1,2,..., n$$

1. **Independence:** Observations (and ε_i's) are independent.
2. **Linear Response:** $E(\varepsilon_i) = 0$ for each x.
   - Scatter plot of Y vs X: should show linear trend.
   - Plot of residuals vs fitted values: should not show a trend.
3. **Equal Variance:** $Var(\varepsilon_i) = \sigma^2$ for each x.
   - Plot of residuals vs fitted values: should show equal scatter.
4. **Normality:** ε_i's are normally distributed.
   - QQ plot of residuals: should be linear.

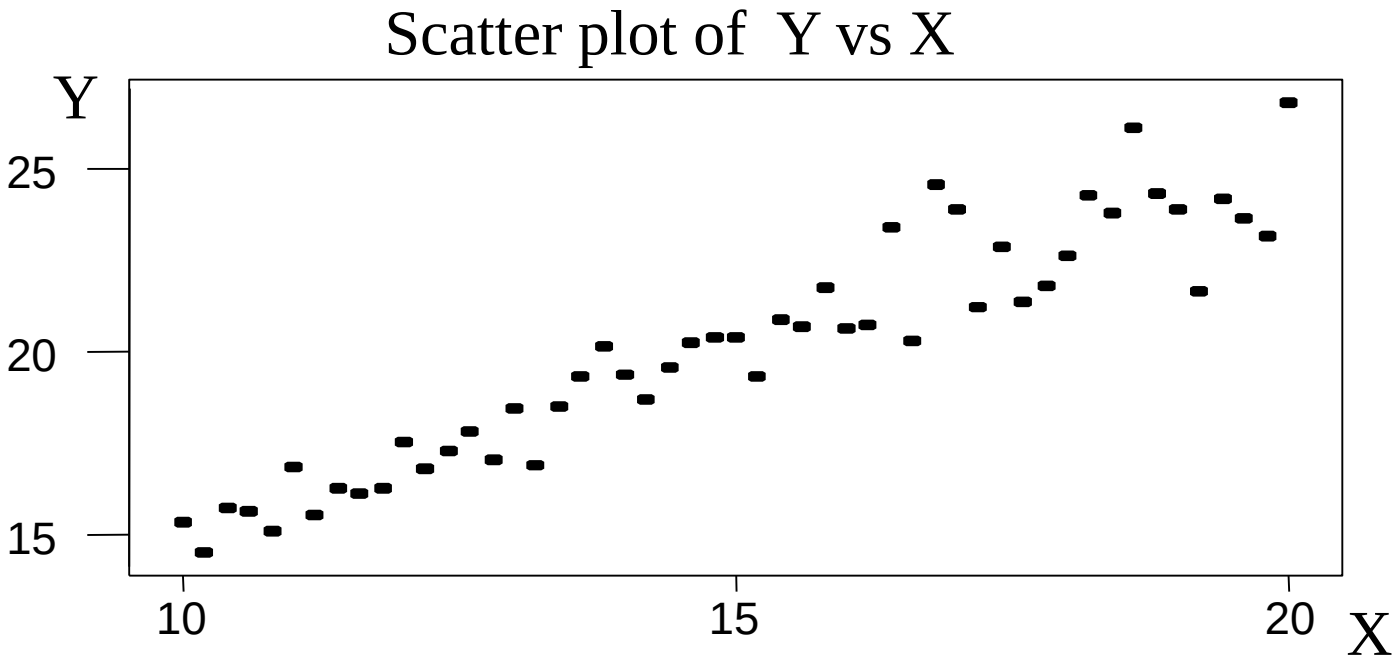
You should always look at a scatterplot of the raw data, but diagnostic plots (based on residuals) can be used to check assumptions.
See "**Corn Example**" example for code to produce diagnostic plots.

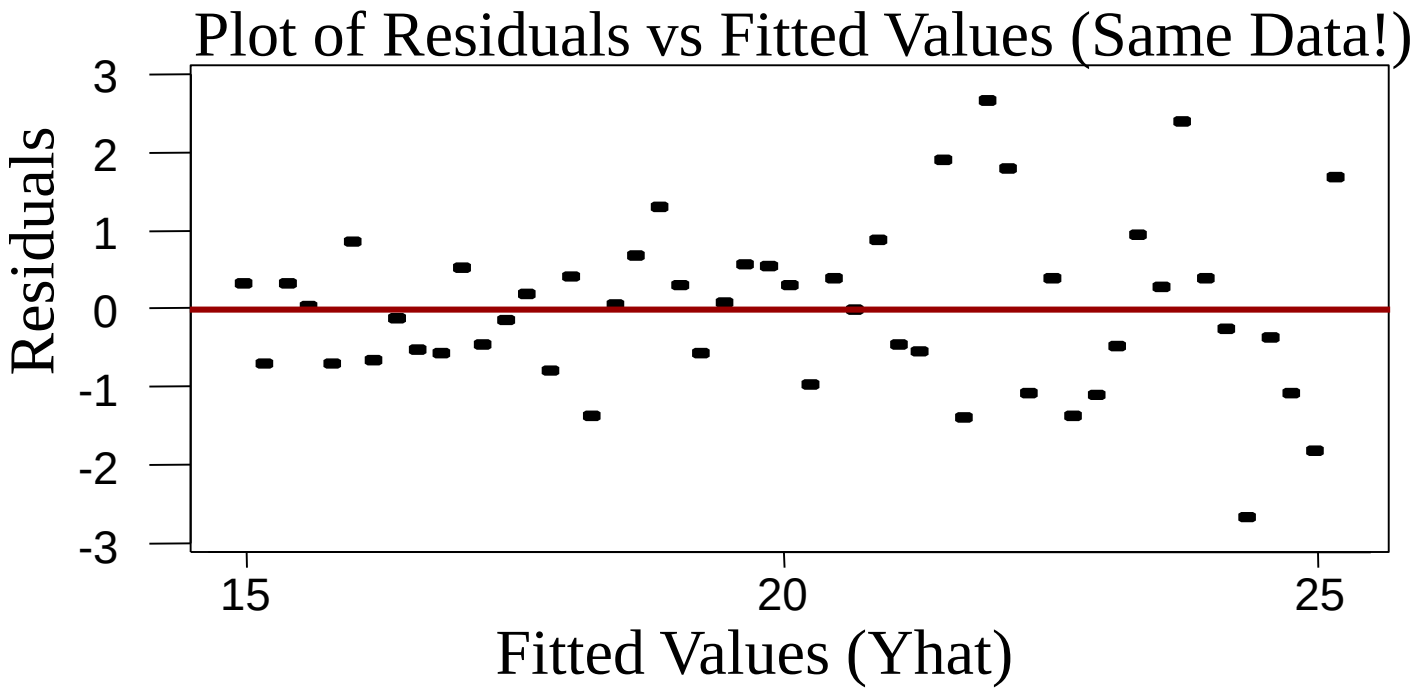
Checking Regression Assumptions: *Example #1*



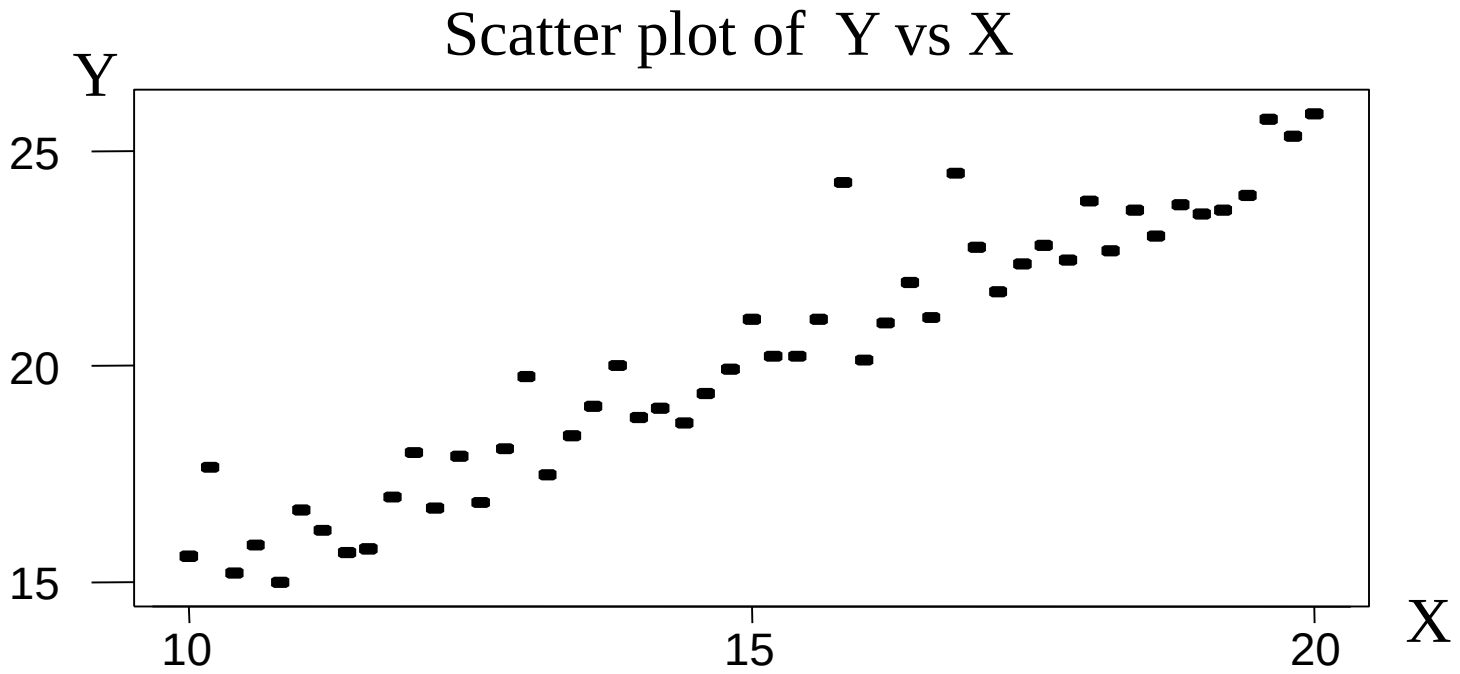


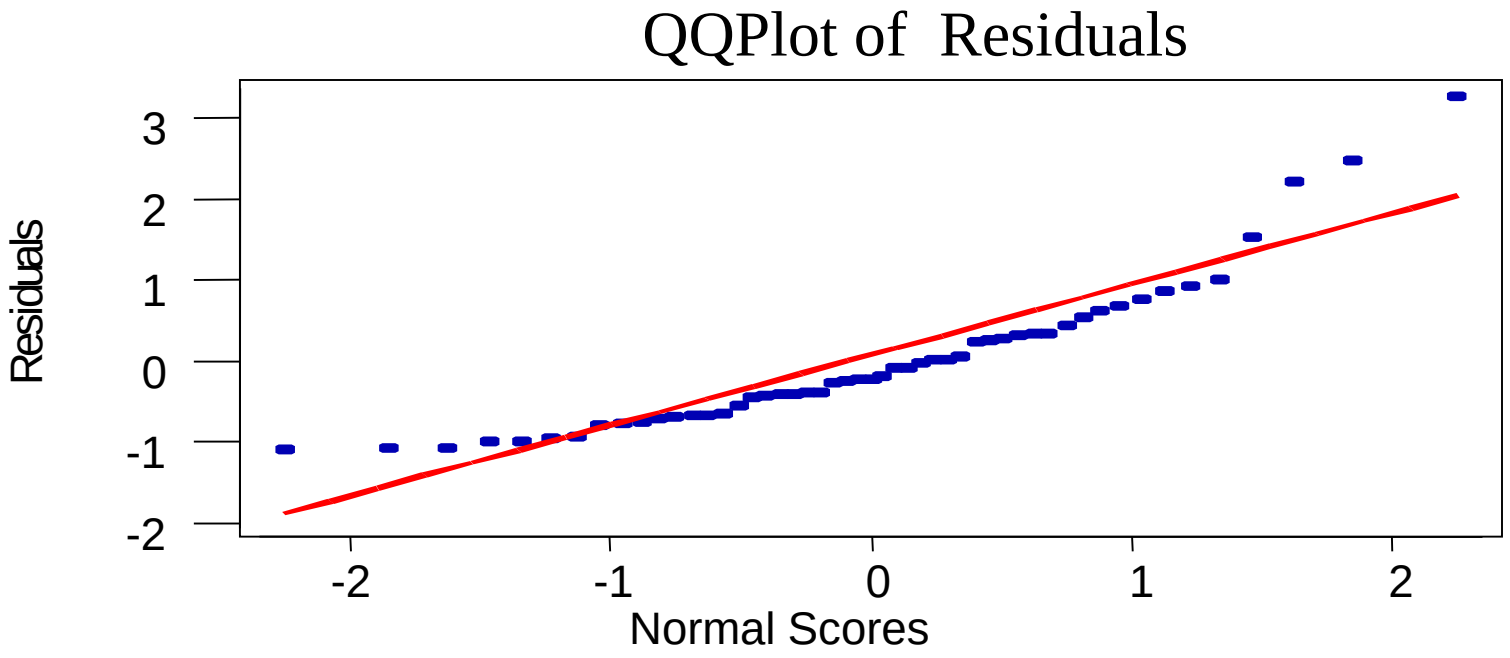
Checking Regression Assumptions: *Example #2*





Checking Regression Assumptions: *Example #3*





Standardized residuals

$$\text{"Raw" residual: } e_i = y_i - \hat{y}_i$$

$$\text{SE}(e_i) = s_\varepsilon \sqrt{1 - \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}} \right)}$$

$$\text{Standardized residual: } s_i = \frac{e_i}{\text{SE}(e_i)}$$

Standardized (or studentized) residuals are residuals that have been "standardized" by dividing each residual by it's SE.

- They have approximately a t-distribution, which is approximately normal for moderate sample sizes, so we expect that about 95% of the standardized residuals will be between -2 and +2. Values greater in absolute value than 3.5 are usually considered outliers.
- In R, standardized residuals can be found using `rstandard(ModelObject)`.

4. Inferences about the Slope, Intercept and Error Variance

The mean and the standard deviation for the sampling distributions of the intercept and slope estimators are as follows:

$$\mu_{\hat{\beta}_0} = \beta_0 , \quad \sigma_{\hat{\beta}_0} = \sigma_\varepsilon \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}$$

$$\mu_{\hat{\beta}_1} = \beta_1, \qquad \sigma_{\hat{\beta}_1} = \frac{\sigma_\varepsilon}{\sqrt{S_{xx}}}$$

Thus $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased estimator, and their standard errors can be calculated by replacing σ_ε by s_ε in the above formulas.

$$SE(\hat{\beta}_0) = s_\varepsilon \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}, \quad SE(\hat{\beta}_1) = \frac{s_\varepsilon}{\sqrt{S_{xx}}}$$

$$\text{Recall}: \; S_{xx} = \sum (x_i - \bar{x})^2, s_\varepsilon = \sqrt{\text{MSResid}}$$

Test and Confidence Interval for β_0 (Intercept)

Assumptions : Regression Assumptions.

Hypotheses:

H_0: $\beta_0 = b_0$ vs H_A: $\beta_0 \neq b_0$

Test Statistic: $t = \dfrac{\hat{\beta}_0 - b_0}{SE(\hat{\beta}_0)}$

Reject H_0 if $|t| > t_{\alpha/2}$ with df=n-2.

p-values and tests for one-sided alternatives handled in the usual way.

Confidence Interval: $\hat{\beta}_0 \pm t_{\alpha/2:n-2} SE(\hat{\beta}_0)$

NOTES: R provides $SE(\hat{\beta}_0)$ as well as the test statistic and p-value for testing H_0: $\beta_0 = 0$.

Use the `confint()` function to get a confidence interval.

Test and Confidence Interval for β_1 (Slope)

Assumptions : Regression Assumptions.

Hypotheses:

H_0: $\beta_1 = b_1$ vs H_A: $\beta_1 \neq b_1$

Test Statistic: $t = \dfrac{\hat{\beta}_1 - b_1}{SE(\hat{\beta}_1)}$

Reject H_0 if $|t| > t_{\alpha/2}$ with df=n-2.

p-values and tests for one-sided alternatives handled in the usual way.

Confidence Interval: $\hat{\beta}_1 \pm t_{\alpha/2:n-2} SE(\hat{\beta}_1)$

NOTES: R provides $SE(\hat{\beta}_1)$ as well as the test statistic and p-value for testing H_0: $\beta_1 = 0$.

Use the `confint()` function to get a confidence interval.

Regression in R and Rcmdr

- In R, use `lm(Y~X)`
- In Rcmdr, choose Statistics > Fit models > Linear regression.

For the Corn example:

```
> Fit<-lm(Yield~X,data=Corn)
> summary(Fit)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  10.1000     0.7973   12.67 1.42e-06
X             1.1500     0.1879    6.12 0.000283
```

Test and Confidence Interval for σ_ε^2

Tests and CI's can be obtained using minor modifications to the methods of Chapter 7. Use df = n-2, and use the same formulas:

$$H_0 : \sigma_\varepsilon^2 \le \sigma_0^2 \quad vs \quad H_A : \sigma_\varepsilon^2 > \sigma_0^2$$

Test Statistic: $\chi^2 = \dfrac{\text{SSResid}}{\sigma_0^2}$

Reject H_0 if $\chi^2 > \chi_\alpha^2$

A 95% confidence interval:

$$\sqrt{\frac{\text{SSResid}}{\chi^2_{0.025}}} < \sigma_\varepsilon < \sqrt{\frac{\text{SSResid}}{\chi^2_{0.975}}}$$

A 95% confidence interval in the Corn regression:

df=n-2=8

$$\sqrt{\frac{5.65}{17.53}} < \sigma_\varepsilon < \sqrt{\frac{5.65}{2.18}}$$

$$0.568 < \sigma_\varepsilon < 1.61$$

The ANOVA table and the F-test

The Analysis of Variance (ANOVA or AOV) table is a way of organizing Sum of Squares statistics for use in F-tests. It is not very useful in regression problems, because those F-tests can be done more easily using t-tests. However, the ANOVA table will be an important tool in STAT512.

Source	Sum of Squares	df	Mean Square (MS)	F
Regression	SSReg	1	MSReg = SSReg/df	F=MSReg/MSResid
Residual	SSResid	n-2	MSResid=SSResid/df	
Total	SSTotal	n-1		

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

$TS: \qquad F = \dfrac{\text{MSReg}}{\text{MSResid}}$

Reject H_0 if $F > F_\alpha$ $\quad df_1 = 1 \quad df_2 = n-2$

Note: The t-test for the slope is equivalent to the F-test:

$$t^2 = \left(\frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} \right)^2 = F$$

Ex: Corn Regression

$$t^2 = (6.12)^2 = 37.45$$

5A. Regression with zero intercept *(through the origin)*

In some problems it is reasonable to believe that the intercept is zero.

Example: Crystals are grown in solution for X hours; the weight of crystal produced (Y) is recorded. It is theoretically reasonable that zero time should result in zero growth, i.e. $\beta_0=0$. If we omit β_0 from the model, then:

$$Y_i = \beta_1 x_i + \varepsilon_i \qquad i = 1,2,...,n$$

The least squares estimate of β_1 is:

$$\hat{\beta}_1 = \frac{\sum x_i y_i}{\sum x_i^2} \quad \text{and}$$

$$\hat{\sigma}_\varepsilon^2 = s_\varepsilon^2 = \frac{\sum (y_i - \hat{y}_i)^2}{n-1}$$

The standard error of $\hat{\beta}_1$ is:

$$\mathrm{SE}(\hat{\beta}_1) = s_\varepsilon \Big/ \sqrt{\sum x_i^2}$$

A test concerning the slope:

$$H_0 : \beta_1 = b_1 \quad \text{vs} \quad H_A : \beta_1 \neq b_1$$

$$\textit{Test Statistic} \quad t = \frac{\hat{\beta}_1 - b_1}{SE(\hat{\beta}_1)}$$

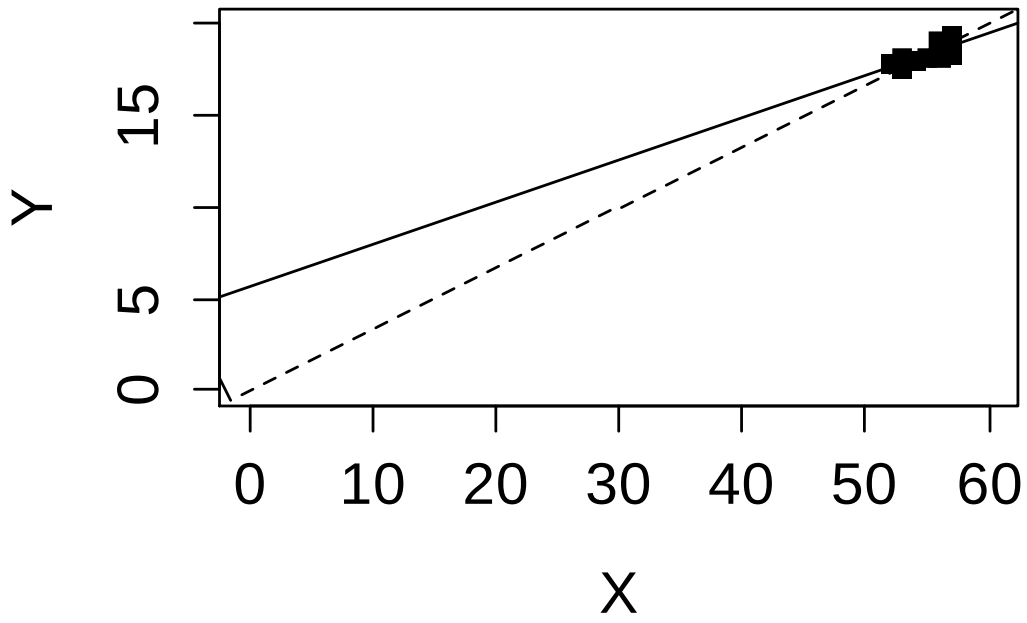
Reject H_0 if $\left| t \right| > t_{\alpha/2}$

Regression with zero intercept *(Notes)*

1. In R the zero intercept model is fit using `lm(Y ~ X - 1)`

See **"Intercept Example".**

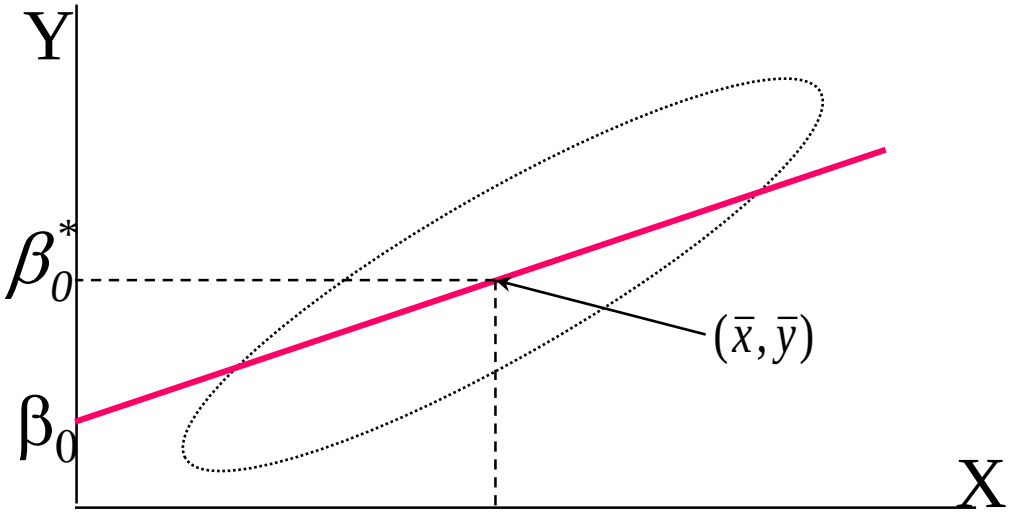
2. Some researchers fit a model with an intercept, test the hypothesis that the intercept is zero, and eliminate it if it is not significant. My opinion is that it is better to keep the intercept in the model, even when it is not significant, unless there is good reason to force the line through the origin. Often the *x* values are well above zero, and the model fits much better with a nonzero intercept, even though you don't have enough power to reject the hypothesis that the intercept is zero.



5B. Regression with "centered" X's

An alternate parameterization of the straight line regression model is sometimes useful:

$$Y_i = \beta_0^* + \beta_1(x_i - \bar{x}) + \varepsilon_i$$



Using the original model, β_0 is the intercept at x = 0.
Using the "centered" model, β_0^* is the height of the line, at the center of the X's, that is, at $x = \bar{x}$.
For both models, β_1 is the slope of the line.

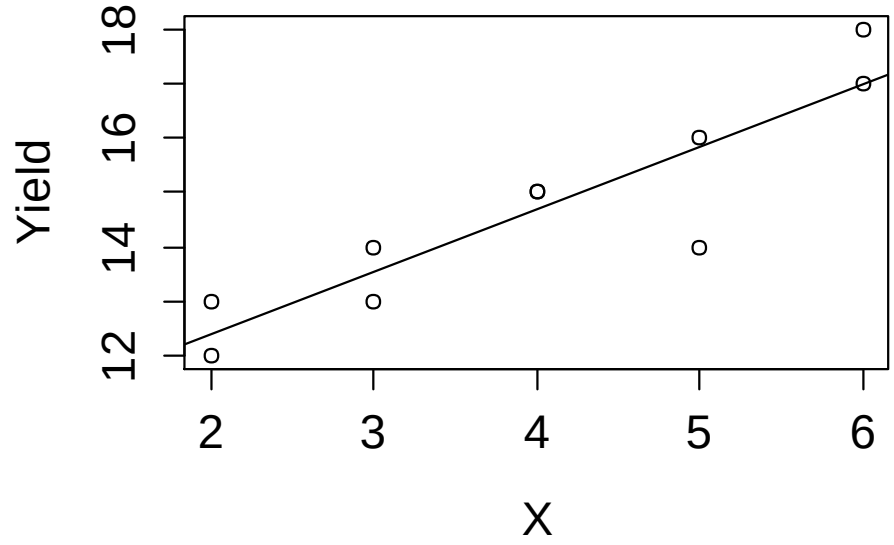
Use the centered model when the height of the line at the center of the X's is a more useful parameter than the height of the line when X=0.
Two advantages of the centered version are:

1) $\hat{\beta}_0^* = \bar{y}$ (a very simple formula), and
2) $\hat{\beta}_0^*$ is independent of $\hat{\beta}_1$.

Example of Centered Regression (Corn Data)

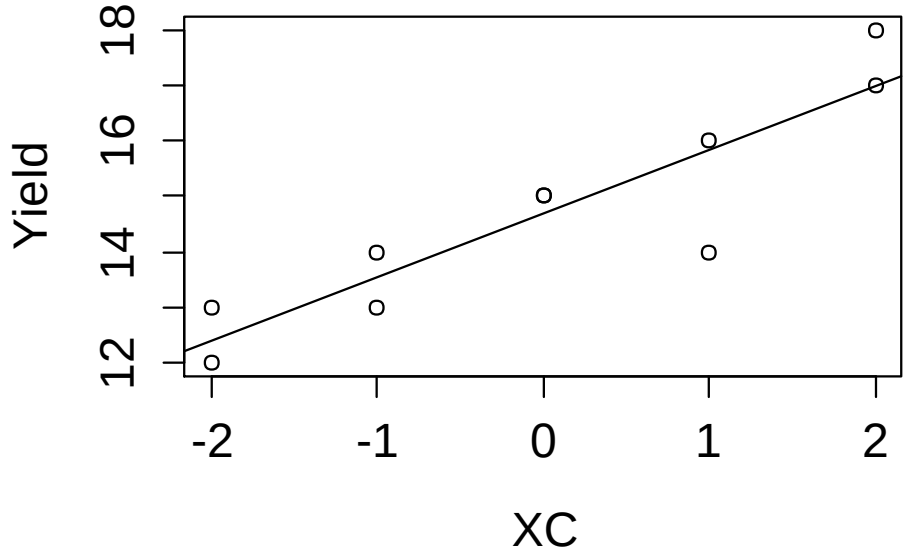
Original:

Yield = 10.1 + 1.15X, R2 = 0.824



Centered:

Yield = 14.7 + 1.15XC, R2 = 0.824



6. Inference for Means and New Observations

Let x* denote a specified value of the predictor variable x.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x^*$$

can be regarded as an estimate of:

1. The expected or true average value of Y when x=x* or
2. A prediction of the Y value that will result from a new observation made when x=x*.

While the estimate ($\hat{y}$) is the same for either of these scenarios, the standard error of the estimates are different (with a larger SE for the prediction case).

NOTE: Caution should be used when making predictions beyond the range of the data because the linear relationship might only hold in a certain range. (**Extrapolation**)

Inferences concerning Average Y at Given X

The expected value (average value) of Y for any specified value of X is given by the regression line. For a given x, the estimated mean response is: $\hat{y} = \hat{\beta}_0^* + \hat{\beta}_1(x - \bar{x})$

So, $$Var(\hat{y}) = Var\left(\hat{\beta}_0^* + \hat{\beta}_1(x-\bar{x})\right) = Var(\hat{\beta}_0^*) + (x-\bar{x})^2 Var(\hat{\beta}_1)$$

$$= \frac{\sigma_\varepsilon^2}{n} + \frac{(x-\bar{x})^2\sigma_\varepsilon^2}{S_{xx}} = \sigma_\varepsilon^2\left(\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}\right)$$

$$\mathrm{SE}(\hat{y}) = s_\varepsilon\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}$$

This SE was "derived" using regression of centered data, but can be used in non-centered regression.

A Confidence Interval for Average Y at Given X

The 100(1-α)% confidence interval for E(Y) at a given X is:

$$\hat{y} \pm t_{\alpha/2} s_\varepsilon \sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}} \quad ; s_\varepsilon = \sqrt{\text{MSResid}} \; ; \text{df} = \text{n} - 2$$

Assumptions: Regression Assumptions, X within range of observed data.

In R use the `predict( ,interval="confidence")` to get the CI for an average Y.

See "**Corn Regression**" example.
90% confidence interval for average Y when X=5.5: (15.70, 17.15)

Prediction interval for **New observation** Y for a given X

We often use the regression line to predict the Y value for a new observation at $X=x_0$. The predicted value is the height of the line at $X=x_0$ which is equal to $\hat{Y}_{new} = \hat{\beta}_0 + \hat{\beta}_1 x_0$.

Var(prediction error) = Var(new obs – height of true line)
+ Var(height of estimated line at center)
+ Var(error due to slope of estimated line)

$$= \sigma_\varepsilon^2 + \frac{\sigma_\varepsilon^2}{n} + \frac{(x-\bar{x})^2 \sigma_\varepsilon^2}{S_{xx}}$$

$$= \sigma_\varepsilon^2 \left(1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}} \right)$$

SE(prediction error for Y_{new})= $s_\varepsilon \sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}$

A Prediction Interval for a New observation Y at Given X

The 100(1-α)% prediction interval at a given X is:

$$\hat{y} \pm t_{\alpha/2} s_\varepsilon \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}} \quad ; s_\varepsilon = \sqrt{\text{MSResid}} \; ; \text{df} = \text{n} - 2$$

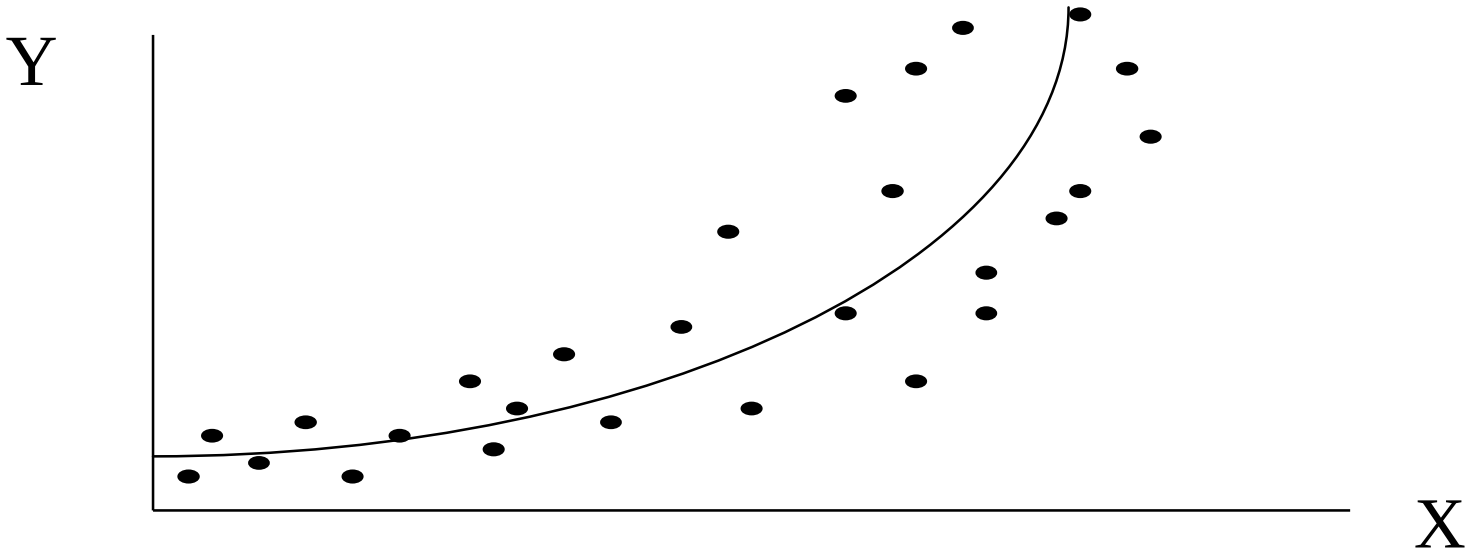
Assumptions: Regression Assumptions, X within range of observed data.

In R use the `predict( ,interval="prediction")` to get the prediction interval for a new Y.

See "**Corn Regression**" example.
90% prediction interval for new Y when X=5.5: (14.70, 18.15)

7. Transformations for linearity and homogeneity of errors

When the plot of Y vs X reveals that the relationship between Y and X is not linear, a linear relationship can sometimes be achieved by transformation of one (or both) of the Y and X.



Example: In the above plot Y looks like an exponential function of X, then we can:

1. Transform Y by regressing log(Y) on X
2. Transform X by regressing Y on e^X
3. Use non-linear regression (covered in "Extra Topics 2")

Deciding whether to transform Y or X

Choice of transforming Y or X depends on:

1. Does a transformation of Y also make error variance more homoscedastic? Sometimes transforming Y makes the line straighter **and** makes the variance more constant.
2. Does transformation of X or Y make the data look more like a "football shaped cloud"? This makes the correlation more meaningful, and reduces the influence of unusual data points on the regression equation.
3. Experimenter preference. Often, the experimenter does not want to interpret the results in some transformed scale, e.g log(dollars).

See "**Stopping Distance: Transformation Example**".

8. Checking for Outliers:

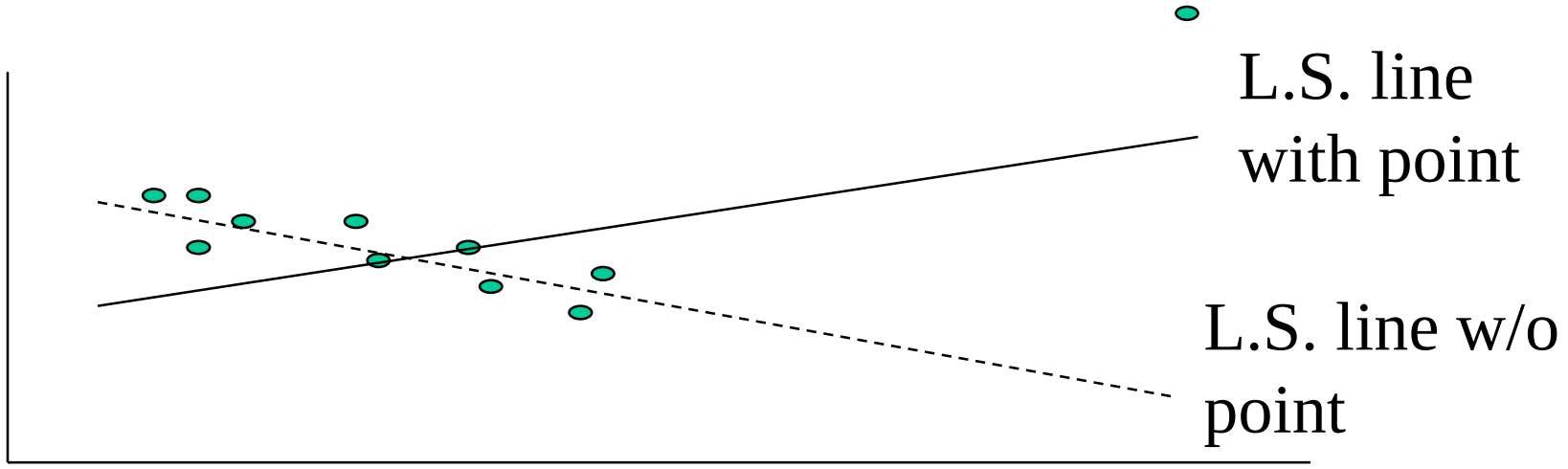
Outlier: An observation (data point) that does not follow the same distribution as the rest of the observations.

Strategies when you find an outlier:

1. Check for data errors. Correct or omit the error.
2. Consider transforming. Some outliers don't look like outliers after transformation.
3. Do the analysis **with and without** the outlier, to see if its presence makes a difference. If it doesn't make a difference, leave it in.
4. Omit the outlier, but comment on its omission in the text of your write-up.

A Test for Outliers: A naïve test for outliers would be to compare the standardized residuals to the t-distribution. This strategy has three problems:

1. An outlier can "pull" the regression line to itself, reducing the size of its residual.



2. An outlier can inflate the estimate of σ^2 so much that the standardized residual is too small.

3. When looking for outliers, we naturally select the point with the largest residual, out of many points. We need some kind of multiple comparison adjustment.

R-Student Residual:

The R-Student residual addresses (1) and (2) from the previous slide. It is appropriate residual for testing potential outliers. It has a Student's t distribution with df = n-3.

Let: $\hat{\beta}_{0,-i}$, $\hat{\beta}_{1,-i}$ and $\hat{\sigma}^2_{-i}$ be estimates of β_0, β_1, and σ^2 using all data points except the i^{th}.

Let: $\tilde{y}_i = \hat{\beta}_{0,-i} + \hat{\beta}_{1,-i} x_i$, the predicted value for the i^{th} point, based on the other points.

Let: $e_{-i} = y_i - \tilde{y}_i$

The "R-Student" residual is: $t_i = \dfrac{e_{-i}}{\mathrm{SE}(e_{-i})}$

(Note: $\mathrm{SE}(e_{-i})$ is based on $\hat{\sigma}^2_{-i}$)

Review of 3 Types of Residuals

Residual $= e_i = y_i - \hat{y}_i$ (the usual residual)

Standardized $= s_i = \dfrac{e_i}{\mathrm{SE}(e_i)} = \dfrac{y_i - \hat{y}_i}{\mathrm{SE}(y_i - \hat{y}_i)}$

Rstudent $= t_i = \dfrac{e_{-i}}{SE_{\text{-i}}(e_{-i})} = \dfrac{y_i - \tilde{y}_i}{SE_{\text{-i}}(y_i - \tilde{y}_i)}$

In R:
Residuals can be found using `residuals()` or `resid()`.
Standardized residuals can be found using `rstandard()`.
R-student residuals can be found using `rstudent()`.

H_0: Observation is NOT an outlier vs H_A: Observation is an outlier
The Rstudent residual is the test statistic!

Unadjusted Test :

Compare the Rstudent value to:

```
tcrit=qt(1-alpha/2, df = n-3)
```

Or compute a (two-sided) p-value :

```
pval=2*(1-pt(abs(rstudent),df = n-3))
```

Bonferroni Adjusted Test:

If we are testing the largest outlier because it is the most extreme out of n, then we are effectively doing n tests. Hence a Bonferroni adjustment is appropriate.

Compare the Rstudent value to:

```
tcrit=qt(1-alpha/(2*n),df = n-3)
```

Or compute a Bonferroni adjusted (two-sided) p-value :

```
pval=2*n*(1-pt(abs(rstudent), df = n-3))
```

Florida Election Example (Transformations and Outliers)

Example: In the 2000 presidential election, there was much controversy over the counting of ballots in Palm Beach County, Florida. A "butterfly ballot" format was used (despite being prohibited by state election laws). It was suspected that some voters intending to vote for Gore, voted for Pat Buchanan by mistake. (Gore: 269K, Bush: 154K, Buchanan: 3.4K) To investigate this, we do a regression by county:

Y = Votes for Buchanan X = Votes for Bush n=67 (# counties)

Using X as a measure of conservativeness of the county, we ask, "Does Buchanan have an unusually large number of votes in Palm Beach County?" See "**Florida Election**" Example.

Step 1: Regress Y on X, and check residual plot.

Step 2: Regress Log(Y) on Log(X) and check residual plot again.

H_0: PB county is NOT an outlier. vs H_A: PB county is an outlier.

We will run the test based on the log transformed model (since assumptions are better satisfied). The Rstudent value (test stat) is 4.0419 and n = 67.

We will consider several options for running the outlier test. See discussion on next slide.

Adjustment	Sides	R code	P-value
None	Two-sided	`2*(1-pt(4.0419,64))`	0.00015
None	One-sided	`1-pt(4.0419,64)`	0.00072
Bonferroni	Two-sided	`2*67*(1-pt(4.0419,64))`	0.0097
Bonferroni	One-sided	`67*(1-pt(4.0419,64))`	0.0048

Note: We can also use the `outlierTest()` function from the `car` package to test the largest residual. Both the unadjusted and Bonferonni adjusted two-sided p-values will be returned.

Conclusion: The two-sided Bonferoni adjusted p-value = 0.00097.

Note that this is the most conservative approach (largest p-value).

Hence, we reject H0 and conclude Palm Beach county is an outlier.

In my opinion: A test without Bonferroni adjustment is justified here, because Palm Beach County was the subject of many complaints **before the polls closed.** We did not select Palm Beach County for outlier testing because it looked large in the data set. It was the purpose of the analysis ***a priori***.

If the complaints involved the fear that voters may have voted for Buchanan by mistake, then we could justify the one-sided p-value.

Focusing on PB county, how many more votes did Buchanan receive than would have been expected (based on our model)?

The observed value (# votes) is 3407.

Using the log transformed model, the predicted value (log votes) is 6.4761.

Back transform by exponentiating: exp(6.4761) = 649.4

The predicted excess votes is 3407 – 649 = 2758.

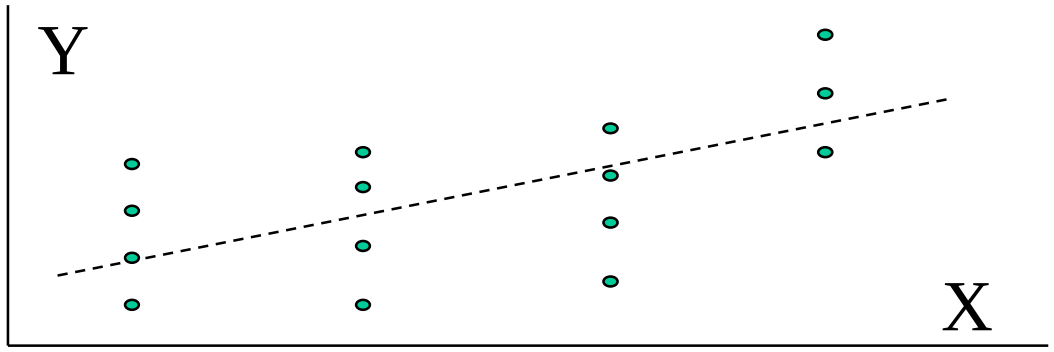
That is more than Bush's margin of victory.

9. An F-test for "lack of fit"

The easiest way to examine possible lack of fit is by plotting (1) scatter plot of Y versus X and (2) residuals versus fitted values. Curvature or other trend in either of these plots might suggest lack of fit.

In this section, we will (briefly) discuss a formal test for lack of fit. This test can be used when there are repeated Y observations at the same X value. In this case we can compare two different models for the data: linear regression (2 parameters) versus one-way ANOVA (t parameters). We compare the estimated σ^2 values from the two approaches.

The lack of fit test determines whether the group means are farther from the line than they should be if linear regression was appropriate.



An F-test for "lack of fit"

H_0: The linear regression model is appropriate.
H_A: The linear regression model is not appropriate.

Test Statistic:

$$F = \frac{\left(\text{SSResid}_{\text{Reg}} - \text{SSResid}_{\text{ANOVA}}\right)/(\text{dfResid}_{\text{Reg}} - \text{dfResid}_{\text{ANOVA}})}{\text{MSResid}_{\text{ANOVA}}}$$

Reject H_0 if $F > F_{\alpha,df1,df2}$
where $df1 = t - 2 = \text{dfResid}_{\text{Reg}} - \text{dfResid}_{\text{ANOVA}}$, $df2 = \text{dfResid}_{\text{ANOVA}}$

NOTES:

1. Values of X that are very close can be altered to be the same to get groups with repeats.

2. To do the test in R, run the analysis as a regression and a one-way ANOVA. Then compare the models using the anova() function. See "**Lack of Fit**" Example.

Corn Example: First fit the regression and one-way ANOVA analyses in R. See "**Lack of Fit**" Example.

$$F = \frac{\left(\text{SSResid}_{\text{Reg}} - \text{SSResid}_{\text{ANOVA}}\right)/(\text{dfResid}_{\text{Reg}} - \text{dfResid}_{\text{ANOVA}})}{\text{MSResid}_{\text{ANOVA}}}$$

$$= \frac{(5.65 - 3.5)/(8 - 5)}{0.70} = 1.0238$$

pvalue = 1- pf(1.0238, df1=3, df2=5)= 0.4564

Conclusion: Fail to Reject H0. There is no evidence of lack of fit for the corn regression.

10. Adding quadratic terms

This topic will be discussed further in STAT512.
We mention it here only because it is an approach sometimes used to deal with lack of fit. When a simple linear regression model does not fit due to curvature, one could consider a quadratic regression.

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$$

This model can be fit in R using:

```
lm(y ~ poly(x, 2, raw=TRUE))
lm(y ~ x + I(x^2))
```

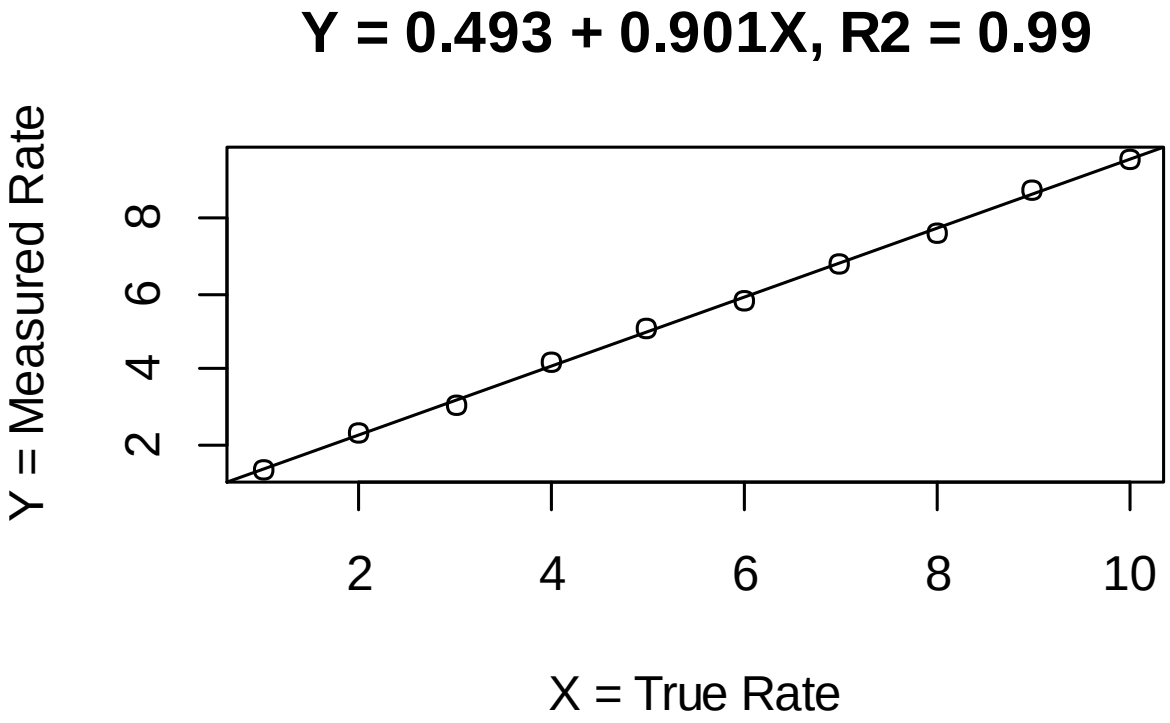
Or just create a new variable:

```
x2 <- x*x
lm(y ~ x + x2)
```

11. The "Calibration Problem"

(predicting x for a given value of y)

Flow Rate Example (from O&L 6th Edition): A new flow-rate meter is being tested. Flow rates (X) are **set by the experimenter** at 1,2,..10, and the meter is read (Y) for each x.



The "Calibration Problem" *(continued)*

The least squares line is:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = 0.4934 + 0.9012x$$

If we then use this instrument in the future, and observe a meter reading of y, we would solve:

$$y = \hat{\beta}_0 + \hat{\beta}_1 x$$

$$\hat{x} = \frac{y - \hat{\beta}_0}{\hat{\beta}_1} = \frac{y - 0.4934}{0.9012} = \frac{y}{0.9012} - \frac{0.4934}{0.9012} = -0.5474 + 1.1096y$$

The "Calibration Problem" *(Notes)*

1. In calibration problems, Y is the new method, and X is the "gold standard".

2. Formula for confidence interval for a calibrated value is given in older editions of O&L.

3. Calibration only makes sense when the relationship between y and x is very strong, otherwise your instrument isn't very useful anyway.

4. If the relationship is curvilinear you might try to linearize it with a transformation. Otherwise, there is such a thing as "quadratic" calibration, but it is much more complicated.

5. The most obvious question is: why not just regress X on Y?
The first thing to realize is that the predicted values will not be the same. See flow rate example on the next slide.

6. See older editions of O&L for further discussion.

The "Calibration Problem" *(Flow Rate Example)*

Fit1: lm(Y ~ X)

$$\hat{y} = 0.4933 + 0.9012x$$

$$\hat{x} = \frac{y - 0.4933}{0.9012} = -0.5474 + 1.1096y$$

When measured flow rate y=9 we find:

$$\hat{x}_{new} = 9.4391$$

Fit2: lm(X ~ Y)

$$\hat{x} = -0.5420 + 1.1086\,y$$

When measured flow rate y=9 we find:

$$\hat{x}_{new} = 9.4356$$

*In this case the difference is not very large/dramatic, but it can make a difference.

12. Correlation

Correlation measures the strength of the linear relation between two variables (x and y). Correlation is NOT causation.

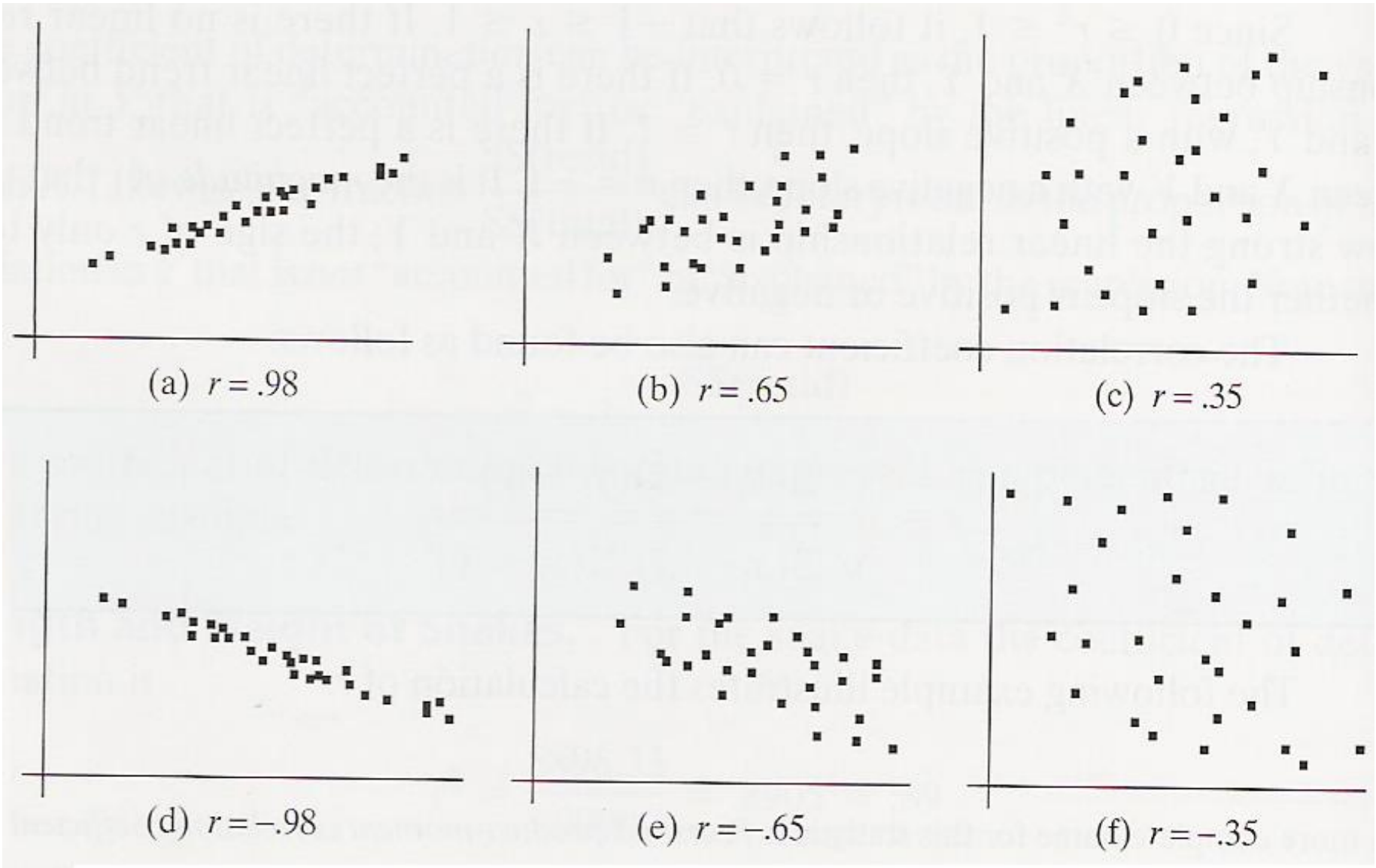
$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} = \frac{s_{xy}}{s_x s_y}$$

- r (or R) is called the sample correlation coefficient or the **Pearson** product-moment correlation.
- It is an estimate of the population correlation = ρ.

- In R, the value of R^2 appears in regression output.
- Correlation can also be computed using `cor()` or `cor.test()`.

See **Correlation Example**.

Some Correlation Examples



Correlation Notes and Warnings

- r is between -1 and +1.
- r does not depend on which of the two variables labeled x.
- r equals +1 only when all the points in a scatter plot of the data lie exactly on a straight line that slopes upward.
- r equals -1 only when all the points in a scatter plot of the data lie exactly on a downward-sloping line.
- r is a measure of the extent to which x and y are linearly related.
- r will have the same sign as the slope of the regression line.
- **Pearson correlation** is most useful in describing bivariate normal data.
- **Spearman** correlation is a rank-based alternative.

WARNINGS:

- Outliers can substantially inflate or deflate correlations.
- Groups combined inappropriately may mask relationships.
- Remember that legitimate correlation does not imply causation!!

The connection between regression and correlation

Parameters: $\rho = \beta_1 \frac{\sigma_x}{\sigma_y}$

Sample Statistics: $r = \hat{\beta}_1 \frac{s_x}{s_y}$

The correlation can be thought of as the slope of the regression line after X and Y have been standardized.

Two consequences of this relationship:

1. r will have the same sign as the slope of the regression line.
2. Test of H_0: β_1 (slope) = 0 is equivalent to test of
   H_0: ρ (correlation) = 0

Coefficient of Determination (r^2 or R^2)

$$r^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = \frac{\text{SSReg}}{\text{SSTotal}}$$

Hence the r^2 value can be interpreted as the proportion or variability in y that is explained by the linear regression on x.

r^2 takes values between 0 and 1. (Some programs express it as a percentage between 0 and 100%.)

In R, the value of R^2 appears in regression output.

The values of SSReg and SSTotal can be found in an ANOVA table. R does not show SSTotal, but it can be calculated as:
SSTotal = SSReg + SSResid.

Corn Example (see next slide): r^2 = (26.45)/(26.45 + 5.65) = 0.824

Example: Corn Regression

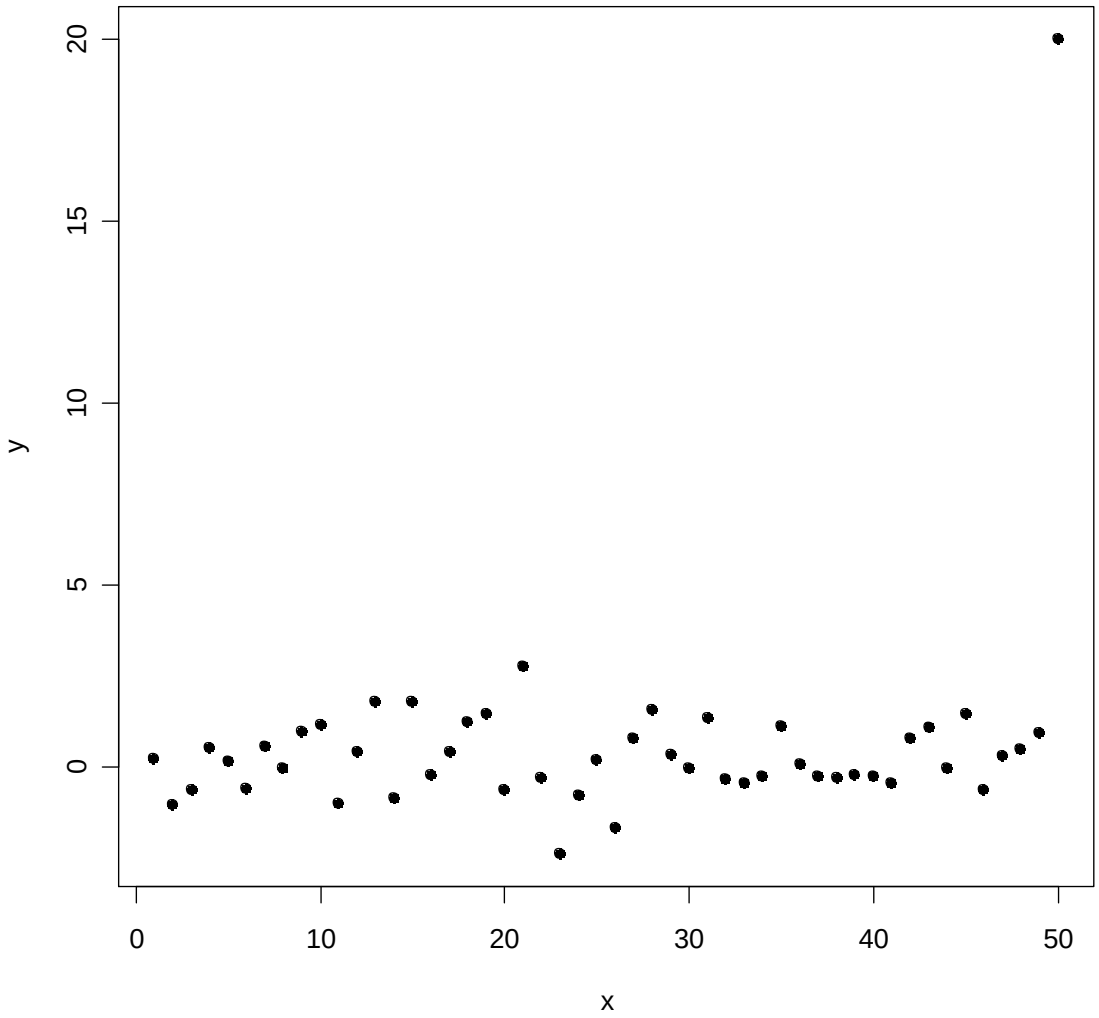
```
> Fit <- lm(Yield ~ X, data = Corn)
> summary(Fit)
Estimate Std. Error t value Pr(>|t|)
(Intercept)  10.1000     0.7973   12.67 1.42e-06 ***
X             1.1500     0.1879    6.12 0.000283 ***

Residual standard error: 0.8404 on 8 degrees of freedom
Multiple R-squared: 0.824,      Adjusted R-squared: 0.802
F-statistic: 37.45 on 1 and 8 DF,  p-value: 0.0002832
> anova(Fit)
          Df Sum Sq Mean Sq F value    Pr(>F)
X          1  26.45 26.4500  37.451 0.0002832 ***
Residuals  8   5.65  0.7062
```

Interpretation: 82.4% of the variability in yield (y) is explained by the linear regression on fertilizer (x).

Since the sign of the slope (+1.15) is positive, then

r (or R) = $+\sqrt{0.824} = +0.9077$

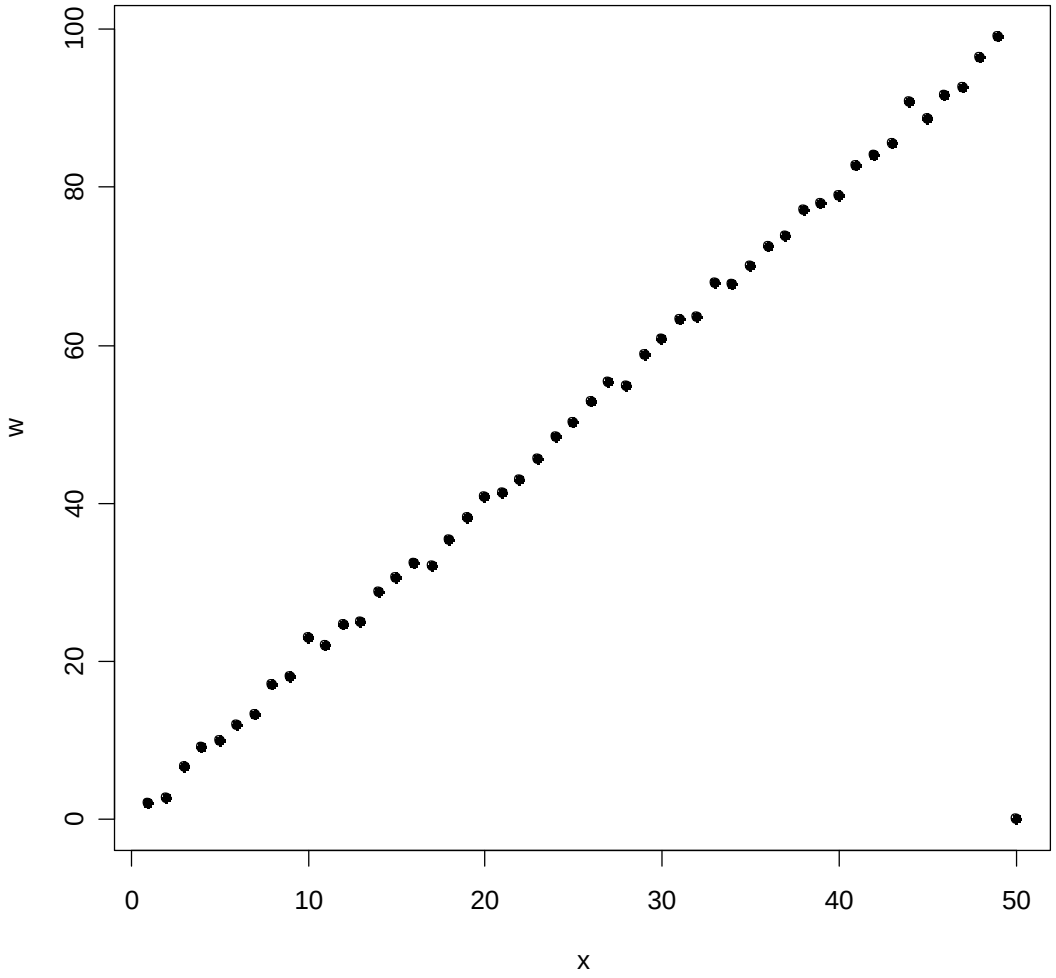
Correlation Example #1



All Observations: r = 0.244

Outlier Removed: r = 0.045

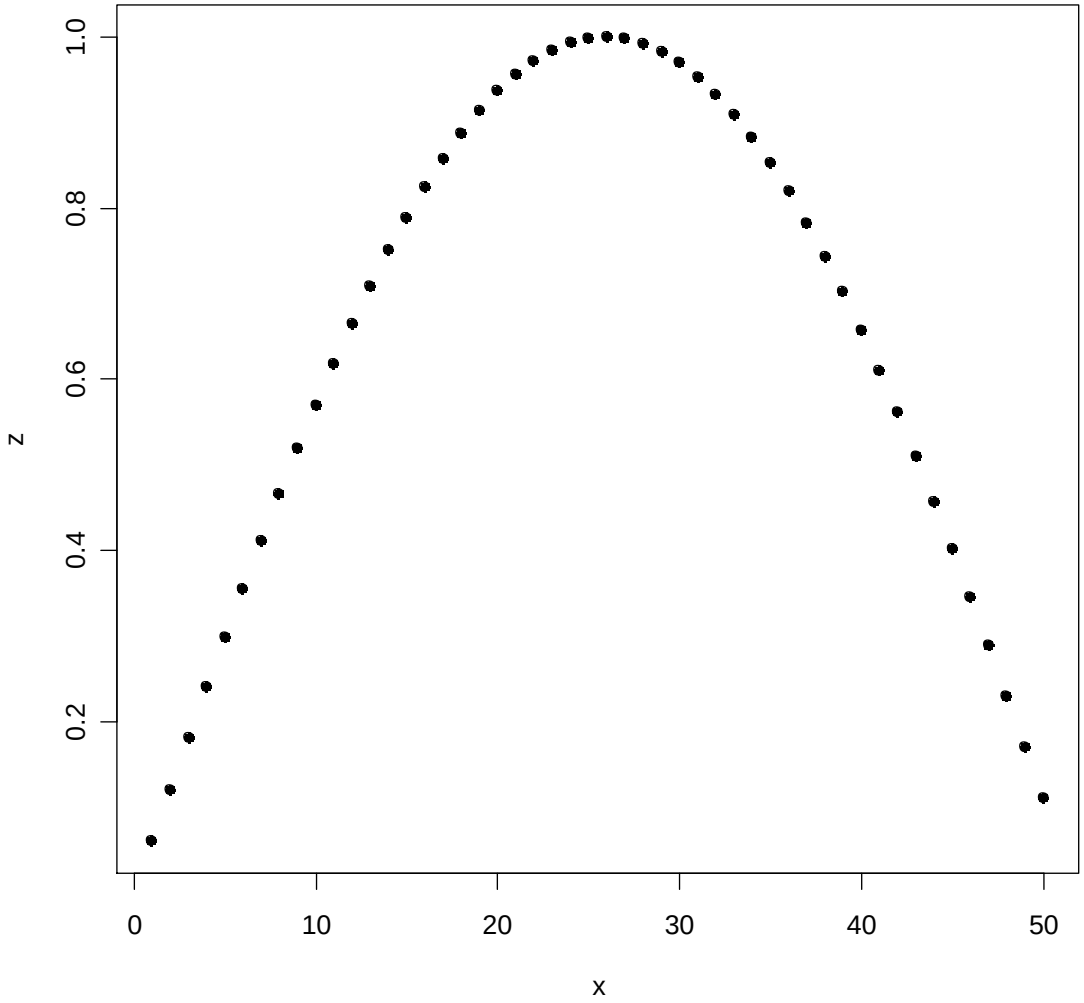
Correlation Example #2



All Observations: r = 0.882

Outlier Removed: r = 0.999

Correlation Example #3



r = 0.060

Correlation Example #4: Quality vs Productivity

There is a long standing perception that Japanese cars have higher manufacturing quality than American made cars.

We examine here data that come from 27 automotive plants in 1989. We consider 11 Japanese plants and 16 non-Japanese plants.

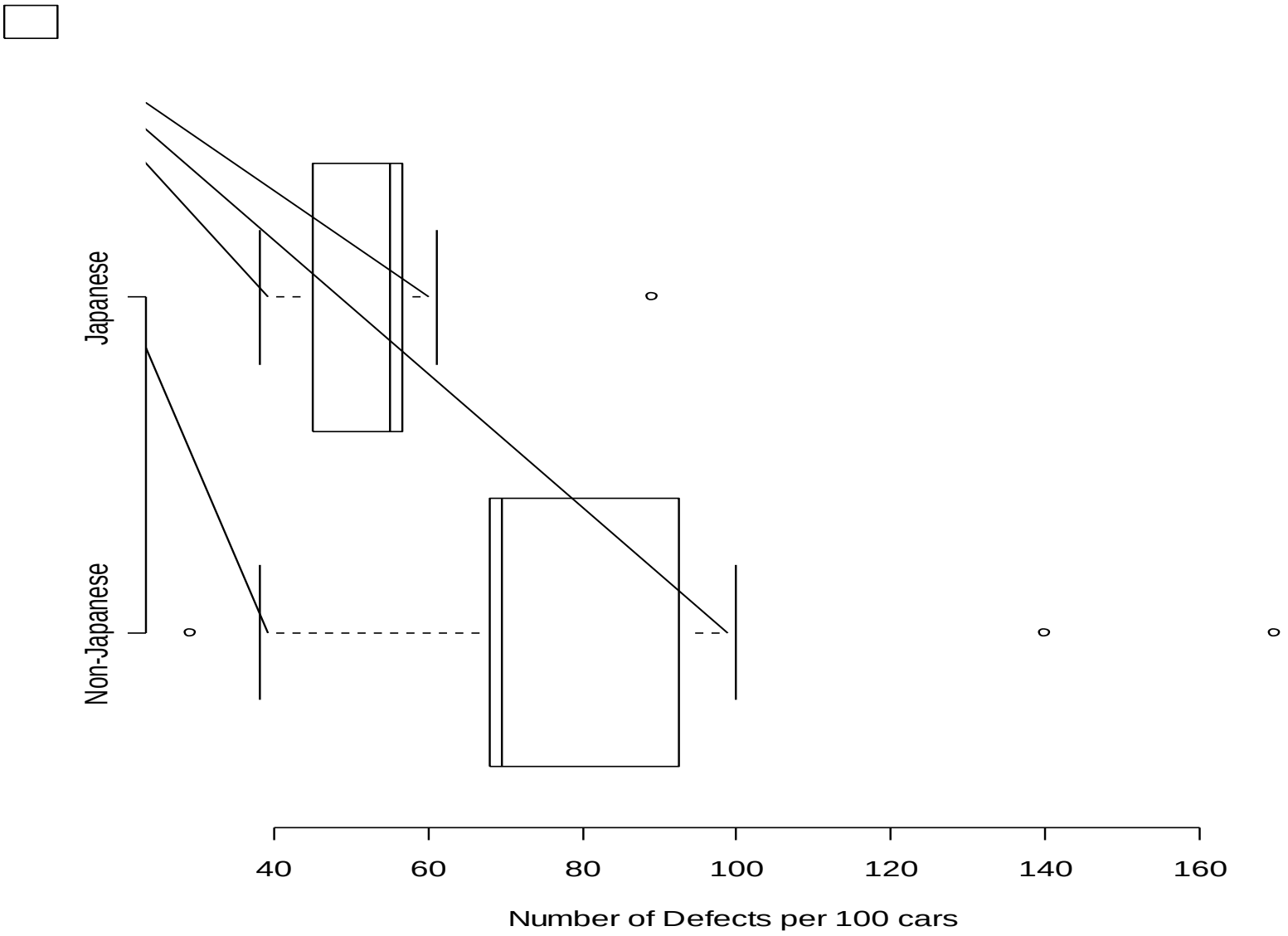
Number of assembly defects per 100 cars is used as a measure of Quality. Hours per car is used as a measure of Productivity.

We will see that when we combine locations (Japanese vs Non-Japanese) there appears to be a positive relationship between the two variables.

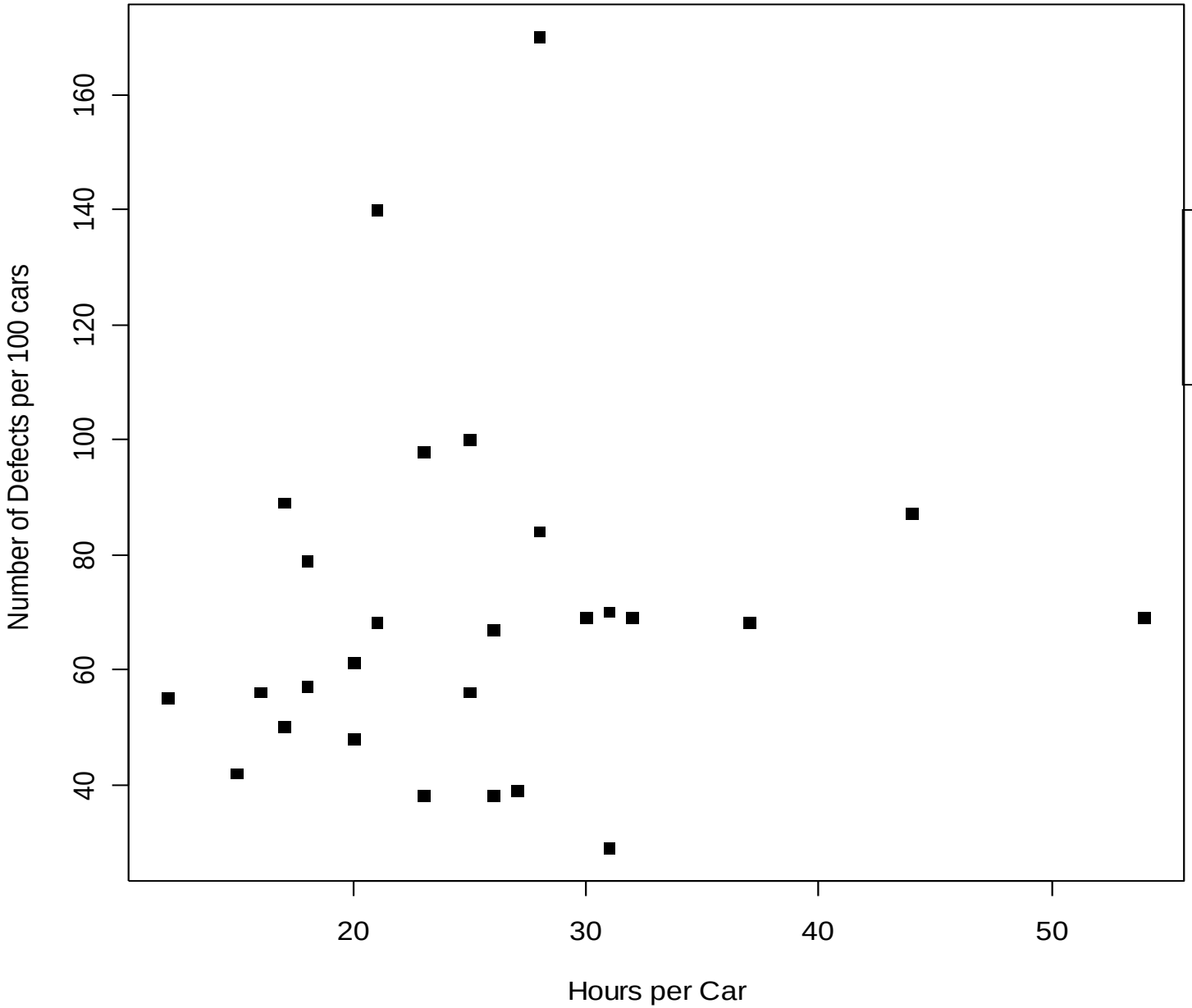
However, when examined separately by location, there is a negative relationship at each location.

This is an example of Simpson's paradox.

Boxplots of Defects for Japanese and Non-Japanese Manufacturers



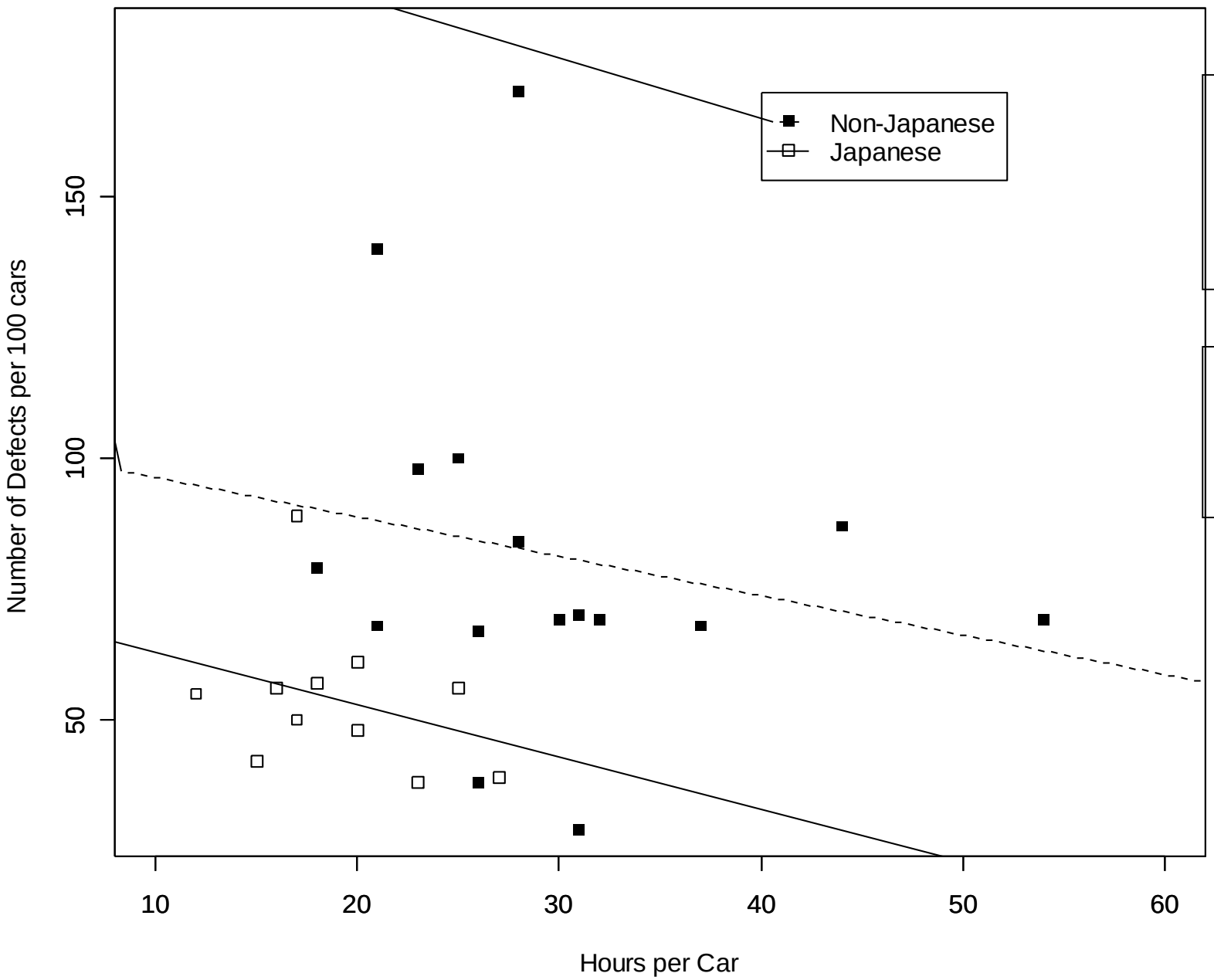
How are Quality and Productivity related?



$\hat{y}$ = 61.1+0.36x

r = +0.11

Identify Japanese and Non-Japanese



Non-Japanese:

$\hat{y} = 104 - 0.75x$

r = -0.20

Japanese:

$\hat{y} = 72.8 - 0.99x$

r = -0.32

t-test for H_0: ρ=0

Hypotheses:

H_0: ρ=0 vs H_A: ρ≠0

Test Statistic: $t = \dfrac{r\sqrt{n-2}}{\sqrt{1-r^2}}$

Reject H_0 if |t| > $t_{\alpha/2}$ with df=n-2.

p-values and tests for one-sided alternatives handled in the usual way.

In R: The p-value will be given in the `cor.test()` output.

Example: n=50 with r =0.527.

$$t = 0.527\frac{\sqrt{50-2}}{\sqrt{1-(0.527)^2}} = 4.296$$

Since 4.296 = t > $t_{\alpha/2}$=2.011 (df=48), we Reject H_0.

Approximate CI for ρ using Fisher's z-transform

Because the distribution of r is skewed, CIs for ρ are computed using a transformed r, and then the results are back-transformed.
(Similar to the CI for the odds-ratio (λ) in CH10.)

The "Fisher Z" transformation: $z = (0.5)\ln\left(\frac{1+r}{1-r}\right)$

z is approximately normal with std. dev. = $\frac{1}{\sqrt{\text{n-3}}}$

An approx. 95% C.I. using z is: $z \pm 1.96\frac{1}{\sqrt{\text{n-3}}}$

Endpoints are back-transformed using: $\frac{e^{2(\text{end pt.})}-1}{e^{2(\text{end pt.})}+1}$

In R: The CI for ρ can be obtained using `cor.test()`.

Example: A sample of n=50 pairs results in a sample correlation coefficient r =0.527

An approximate 95% C.I. is computed as follows:

$$z = (0.5)\ln\left(\frac{1+r}{1-r}\right) = (0.5)\ln\left(\frac{1+0.527}{1-0.527}\right) = 0.586$$

$$\text{A 95\% C.I. using z is : } \quad 0.586 \pm 1.96\frac{1}{\sqrt{50-3}}$$

$$(0.300, 0.872)$$

Back-transform the endpoints to get a C.I. for ρ

$$\text{LCL} = \frac{e^{2(0.300)}-1}{e^{2(0.300)}+1} = 0.291; \quad \text{UCL} = \frac{e^{2(0.872)}-1}{e^{2(0.872)}+1} = 0.702$$

Notes on Inference for ρ

1. The t-test for $\rho=0$ is exact, if the bivariate normal assumptions hold.
2. The tests and confidence intervals based on the Fisher – Z are approximate, but generally good when r is not too close to -1 or 1, and the sample size is relatively large.
3. For very small sample sizes ($n < 7$) CI charts based on the exact distribution of r are available (but not discussed in this course).

13. Sample Size and Power for Slope and Correlation

Case1: Compute power for test of slope (using Lenth)
For this calculation, we need:

- Conjecture for slope (β_1)
- Conjecture for standard deviation (σ_ε) from the regression. This is <u>not</u> typically found in publications. Can be calculated for pilot data.
- Calculated or conjectured value for σ_x the standard deviation of x. In some cases this may be known (ex: designed experiment with fixed values of x). Otherwise, use a conjectured value.

Case2: Compute power for test of correlation (using Lenth)
For this calculation, we need:

- Conjecture for correlation (ρ)

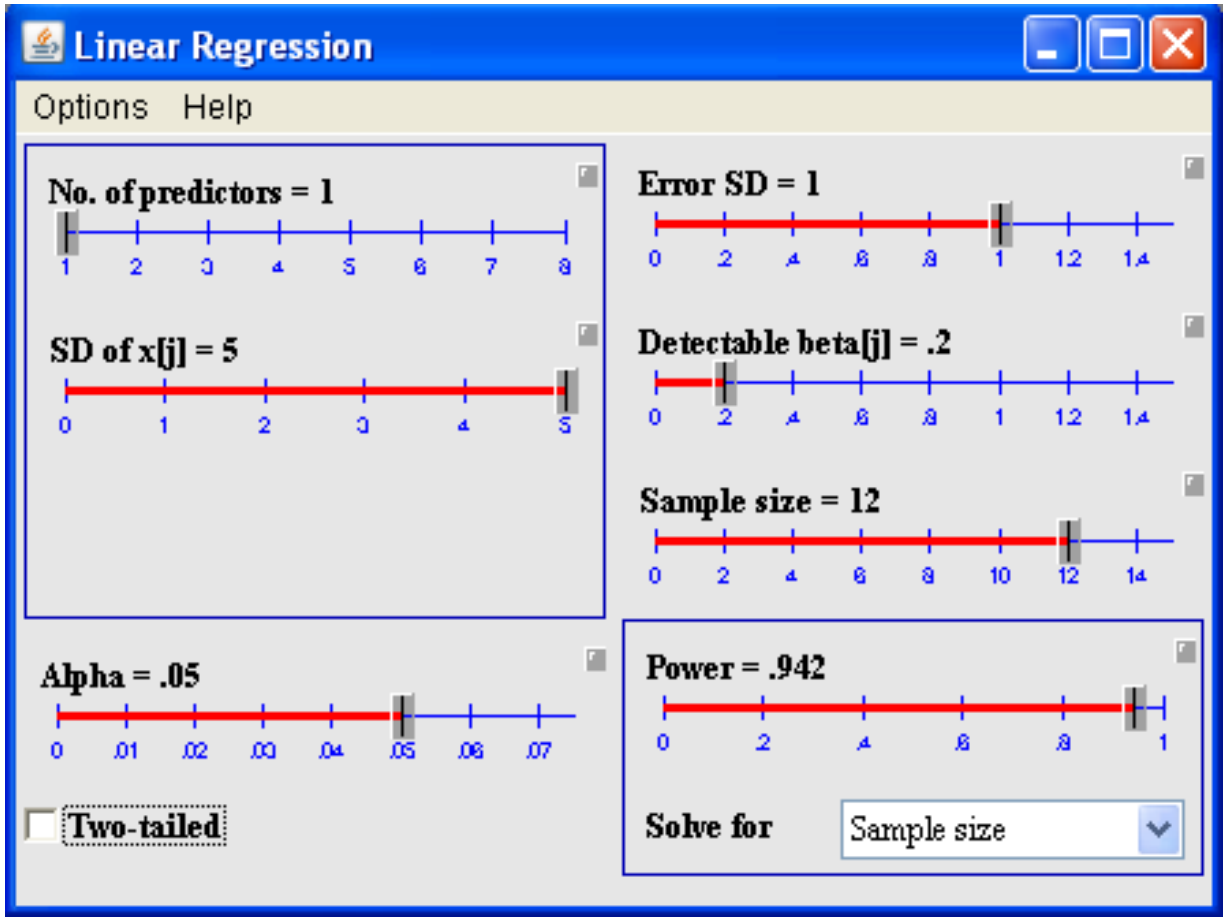
Note the short list of conjectures for power for correlation!

Case 1: Power for Test of Slope using Lenth

http://homepage.stat.uiowa.edu/~rlenth/Power/

- Choose Linear Regression.

- Here we calculate power for HA: $\beta 1>0$ using n=12 based on a conjectured values of $\beta 1=0.20$, $\sigma=1$, $s_x=5$.

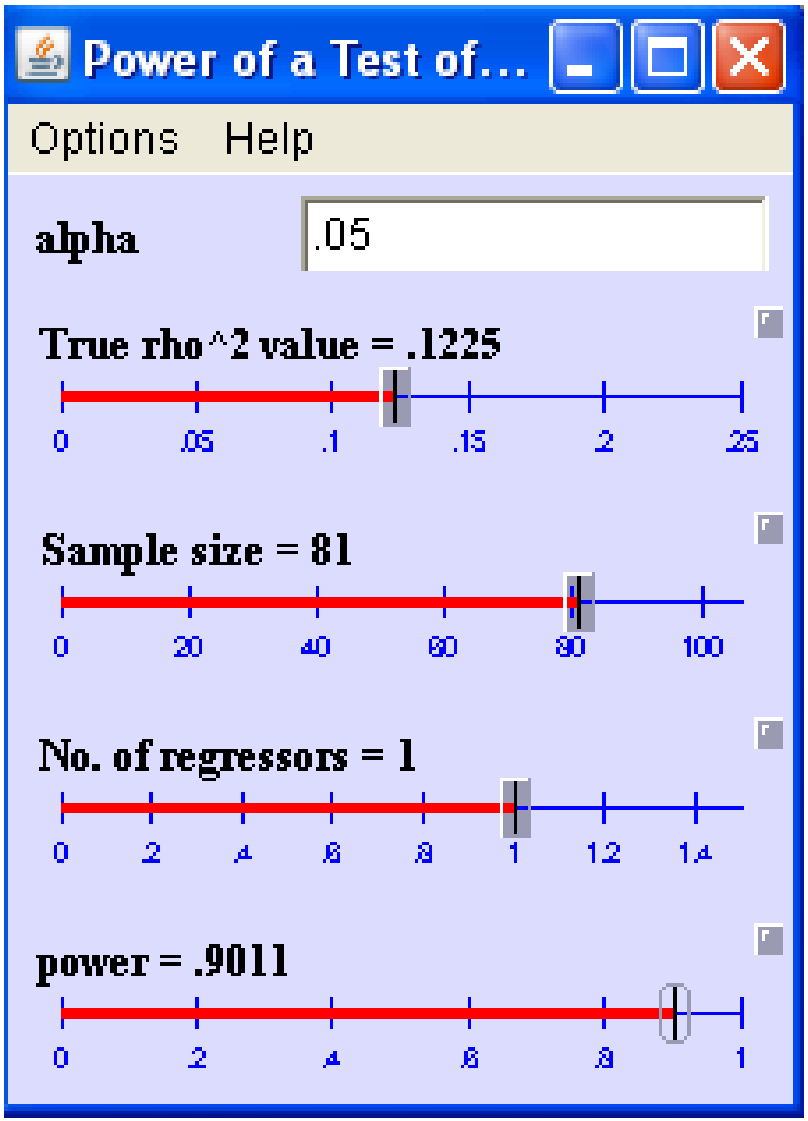


Case 2: Power for Test of ρ using Lenth

http://homepage.stat.uiowa.edu/~rlenth/Power/

- Choose R-square (multiple correlation)

- Here we calculate power for HA: $\rho \neq 0$ using n=81 based on a conjectured values of $\rho=0.35$. Hence $\rho^2 = 0.1225$.



14. Some Comments on Ecological Correlation & Regression

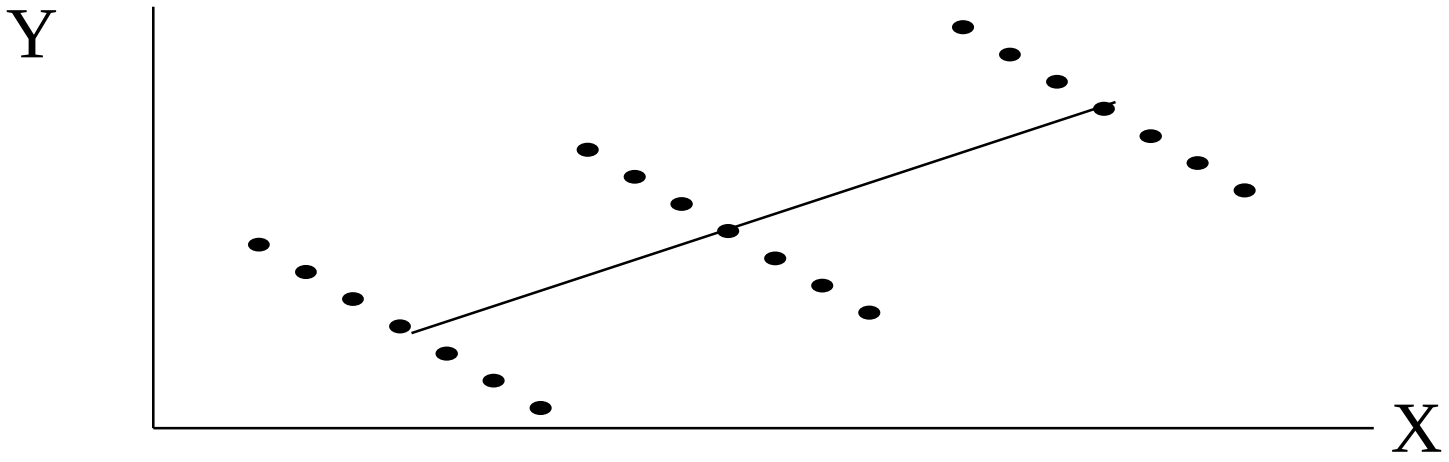
When correlations are computed between subgroup averages (rather than individuals), they are called "**ecological correlations**". It is common for coarser groupings of data to produce larger ecological correlations.

Example: For 13,820 adults in the U.S. the correlation between education level and income was 0.29. After averaging over county (64 counties or county groups), the correlation of the county average education level and county average income was 0.60. The correlation between state averages (for the 34 states) was 0.67. The correlation between regional averages (for 6 regions) was 0.99.

Interpreting ecological correlations as if they were relevant to individuals, is called the "**ecological fallacy**". Ecological correlations can be used to generate hypotheses for confirmation or rejection in more appropriate studies. The first researcher (Doll, 1955) to suggest the connection between smoking and lung cancer, computed the correlation between lung cancer rates and cigarette smoking rates in 11 countries as 0.7. Correlation between fat consumption and cancer rates.

Ecological Regression *(Notes)*

1. It is possible for the ecological correlation and the correlation between individuals within each group to have opposite sign (Simpson's paradox)

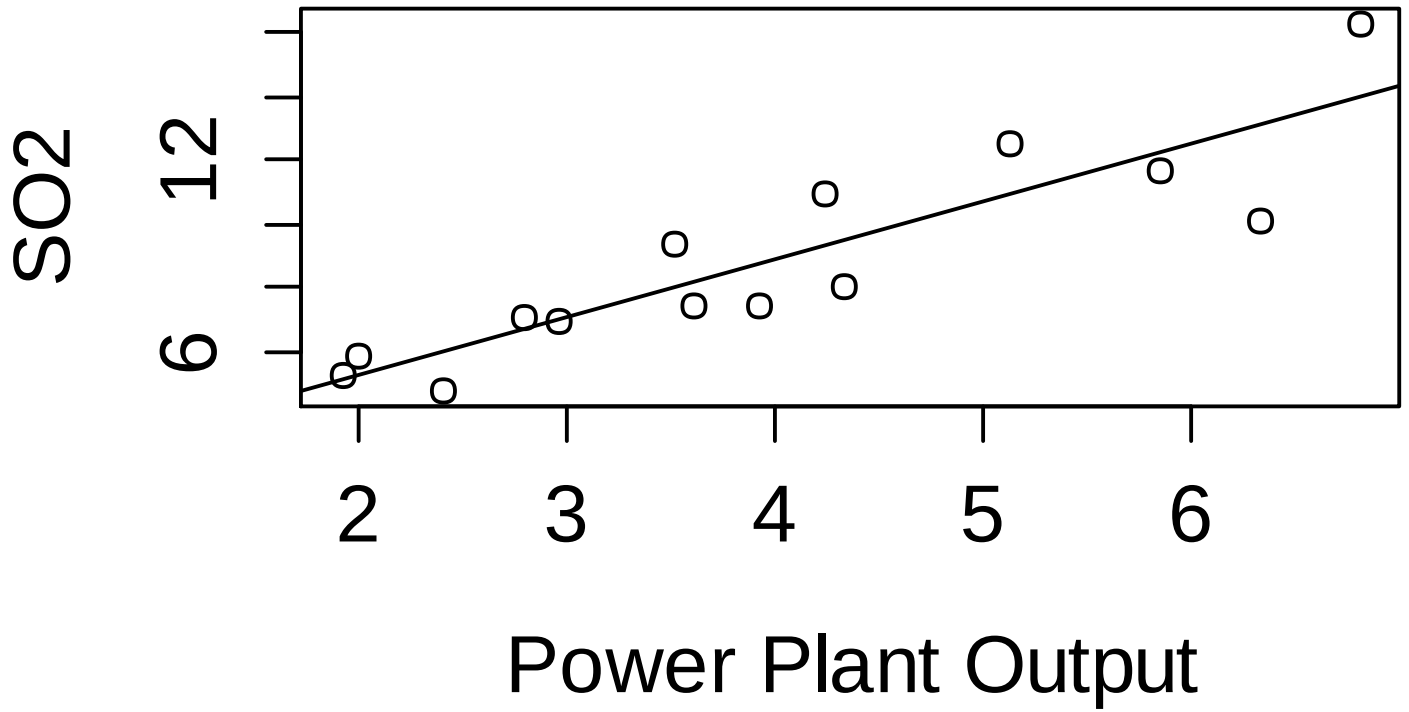


Within each group r = -1, but between group means r = 1.

2. Sometimes researchers report correlations of averages to exaggerate the strength of the relationship they are studying.
3. Sometimes ecological correlations are the only ones available: "Reconstruction" election data: regression by county of percent voting democrat versus percent minorities in the county population.

Correlation and Causation

Power Plant Example: A power plant is located 25 miles from a national park. Output of SO_2 (tons/hr) at the plant and air concentrations of SO_2 ($\mu g/m^3$) at the park were recorded at randomly selected times. What proportion of the air pollution at the park is attributable to the power plant?



Power Plant Example:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.5429     1.1429   1.350    0.202
x              1.8248     0.2684   6.799 1.91e-05

Multiple R-squared:  0.7939
```

Interpretation:

1. A 1 unit increase in power plant output (x) is associated with a 1.82 unit increase in SO2 concentration (y) at the park.
2. 79.4% of the variability in concentration at the park is explained by the linear relationship with output.
3. This is an observational study, so we cannot establish a causal relationship! We can discuss the association between power plant output and SO2 concentrations.

Power Plant Example continued:

Some other variables that might influence pollution levels at the park:

1. Temperature
2. Automobile Traffic
3. Other power plants in the area

Multiple regression can be used to "control" or "account" for other variables . This is a step towards inferring causality, but it is impossible to control for every candidate "confounding" variable.

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 w_i + \beta_3 z_i + \varepsilon_i$$

where : w = temperature , z = automobile traffic level

Multiple regression will be discussed briefly in the Extra Topics 2 notes and in much more detail in STAT512.