Multiple Regression - Part 1

1. The Multiple Regression Model: terminology, interpretation and estimation
2. Model assumptions and diagnostic plots
3. Basic hypothesis testing (single parameter)
4. Confidence and prediction intervals
5. More hypothesis testing
6. Parameter interpretation and causality
7. Transformations
8. Outliers

Examples:

1. Rice Example

2. Fuel Example

3. Weight Loss Example

1. The Multiple Regression Model

The objective of multiple regression is to relate a response variable (Y) simultaneously to multiple predictor variables (X_1, X_2 , X_3 , etc.)

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \ldots + \beta_k X_{ik} + \varepsilon_i$$

k = # predictor variables, n = # observations

The β’s are called “partial regression coefficients”.

ε is called the “error”.

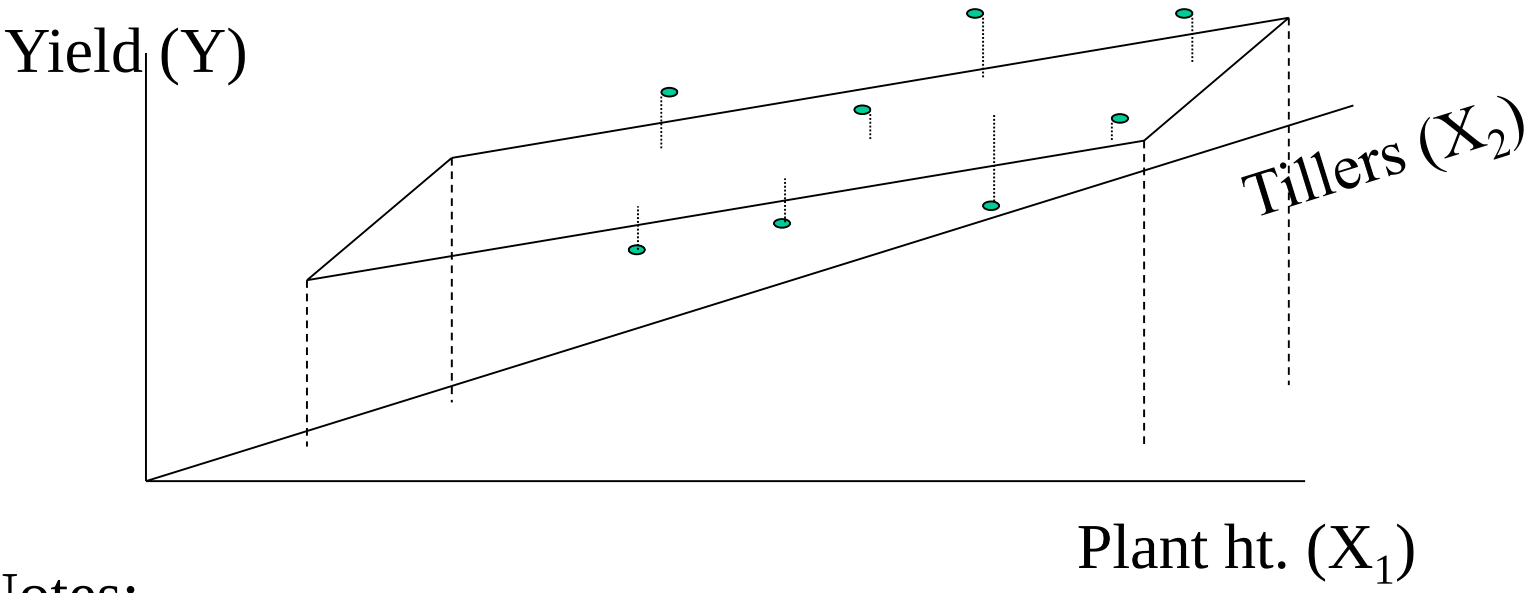
Parameter Interpretation: β_1 is the slope in the X_1 direction. Change in mean response corresponding to a one-unit change in X_1, with all other X’s held constant.

Rice Example (Gomez and Gomez): Response variable Y = yield, predictor variables X_1 = height and X_2 = tillers with n=8 varieties of rice.

Model: $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$, for $i = 1, 2, ..., n = 8$

In their respective simple linear regression models, tillers and height are each significant predictors of yield.

In the multiple regression with both tillers and height, neither is significant but the overall model is.

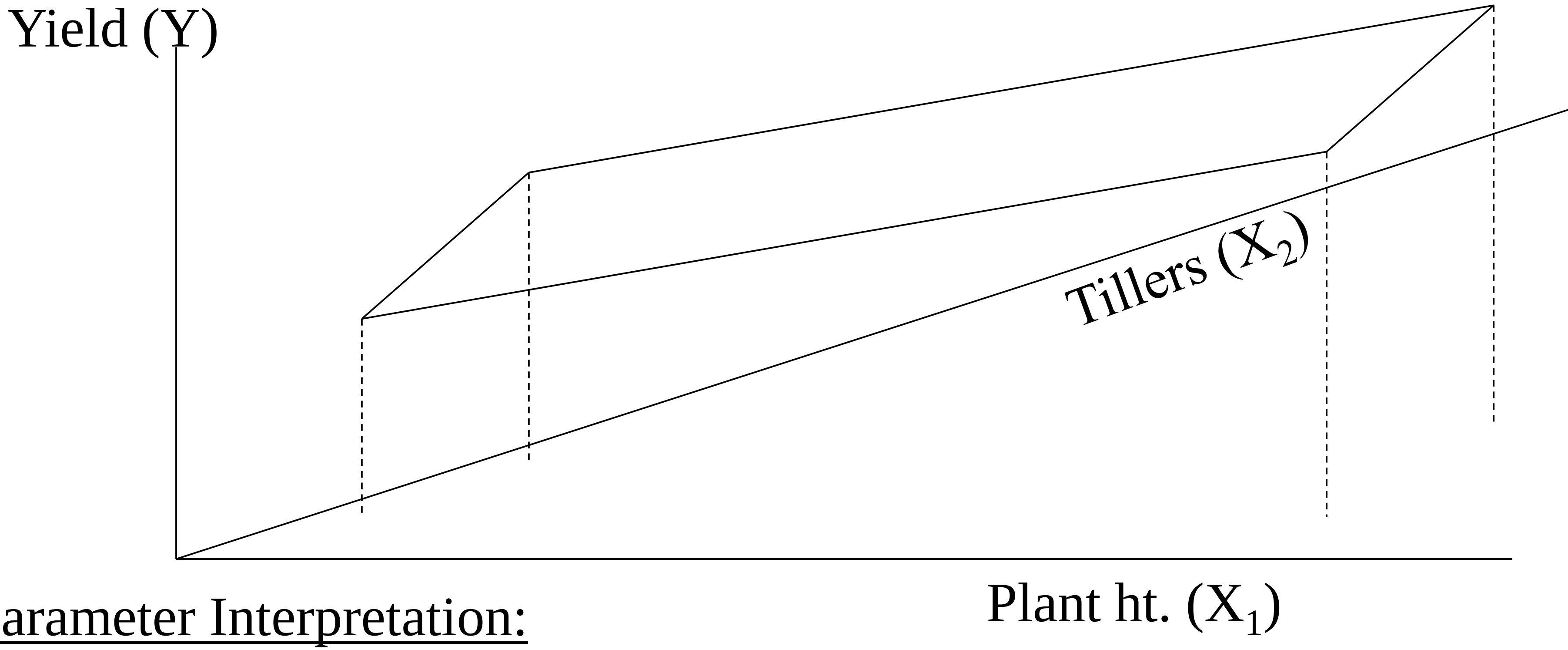


<u>Notes</u>:

1. The “general linear model” allows any number of predictors:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \ldots + \beta_k X_{ik} + \varepsilon_i$$

2. The “simple linear model” is a special case (*k*=1).
3. Since the *β*’s are unknown parameters, the height of the plane is an unknown parameter (and so in some sense is ε)
4. In the “general linear model” the *X*’s are taken as fixed, even if they were not sampled that way.



Parameter Interpretation:

β_0 is the intercept: mean yield if all X's are zero.

β_1 is the slope in the X_1 direction: change in mean yield for a unit change in X_1 (X_2 held constant).

β_2 is the slope in the X_2 direction (change in mean yield for a unit change in X_2 (X_1 held constant).

The "partial" in "partial regression coefficients" is because the other X's are held constant.

Terminology:

1. y_i is the "observed" value for the i^{th} data point.

2. $\mathrm{E(y}_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$ is the "expected" (average) response for the i^{th} data point. (height of the plane)

3. $\varepsilon_i = y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2})$ is the "error".

4. $\text{SSResid} = \text{SSE} = \sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}) \right)^2$

Note that β_0, β_1, β_2 are unknown population parameters. Their corresponding estimated values are denoted $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$.

Terminology (continued):

5. $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2}$ is called the least squares "fitted value" (or "predicted value") for the i^{th} data point.

The part of y_i that is explained by the model.

6. $\hat{\varepsilon}_i = e_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2})$ is called the "residual" (i.e. the estimated error).

The part of y_i that is unexplained by the model.

Then $\sum_{\mathrm{i}=1}^{\mathrm{n}} \left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2}) \right)^2 = \sum_{i=1}^{n} \left(e_i \right)^2 = \mathrm{SSResid} = \mathrm{SSE}$

Estimation of β_0, β_1, β_2 and σ^2

Least Squares Method: Find $\widehat{\beta_0}$, $\widehat{\beta_1}$ and $\widehat{\beta_2}$ that minimize SSResid.

The least squares estimates of β_0, β_1 and β_2 are obtained by:

1. Differentiating SSResid with respect to β_0, β_1 and β_2 ,
2. Setting the derivatives equal to zero, and
3. Solving the resulting system of linear equations (called the "normal" equations).

In practice, we will use the `lm()` function to do the estimation.

Parameter Estimates using `lm()`

The basic regression code will produce parameter estimates:

```
Model <- lm(yield ~ ht + tillers,
             data = Rice)

summary(Model)
```

See the **Rice Example**.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.33560    2.94293   2.153   0.0839
ht          -0.02375    0.01290  -1.842   0.1249
tillers      0.15031    0.11207   1.341   0.2375
```

Parameter Interpretation for the Rice Example:

$\widehat{\beta_0}$ = 6.336 the y-intercept of the plane; the estimated average yield when both height and tillers are zero.

$\widehat{\beta_1}$ = -0.0237 the slope in the height direction; the estimated change in yield for a unit change in height, among varieties with the same number of tillers.

$\widehat{\beta_2}$ = 0.1503 the slope in the tillers direction; the estimated change in yield for a unit change in tillers, among varieties with the same height.

$$\hat{y} = \widehat{\beta_0} + \widehat{\beta_1} * ht + \widehat{\beta_2} * tillers$$
$$\hat{y} = 6.336 - 0.0237 * ht + 0.1503 * tillers$$

Estimation of "error variance": Var(ε_i) = σ^2

(the same for all data points: homogeneity of error variance)

$$SS\,\mathrm{Re}\,sid = SSE = \sum_{\mathrm{i=1}}^{\mathrm{n}} \left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2}) \right)^2$$

$$MS\,\mathrm{Re}\,sid = MSE = \frac{SS\,\mathrm{Re}\,sid}{df_{\mathrm{Re}\,sid}} = \frac{SS\,\mathrm{Re}\,sid}{n-k-1}$$

$$\boxed{\hat{\sigma}^2 = s_e^2 = MS\,\mathrm{Re}\,sid}$$

df_{Resid} *=n–k–1 = #data points – #predictors – 1*

The "*1*" is for the intercept.

Error variance (σ^2) estimation for the Rice Example:

$$SS\,\mathrm{Re}\,sid = 0.57946 \qquad n = 8 \qquad k = 2$$

$$MS\,\mathrm{Re}\,sid = \frac{0.57946}{8-2-1} = \frac{0.57946}{5} = 0.11589$$

$$\hat{\sigma}^2 = 0.11589$$

R^2 Coefficient of Determination

R^2 = proportion of variability in y explained by the linear regression on the x's.

$$= \frac{\text{variability explained by the model}}{\text{total variability}}$$

$$= \frac{\sum\left(\hat{y}_i - \overline{y}_.\right)^2}{\sum\left(y_i - \overline{y}_.\right)^2} = \frac{\text{SSModel}}{\text{SSTotal}}$$

$$= \frac{\text{SSTotal - SSResid}}{\text{SSTotal}}$$

Notes about R^2:

1. $0 \le R^2 \le 1$ or $0\% \le R^2 \le 100\%$
2. R^2 is shown in the `summary()` output for an lm model object. Labeled "Multiple R-squared".
3. In simple regression (k=1), the squared sample correlation coefficient, $r^2 = R^2$.
4. R^2 must go up (or in rare cases stay the same) when an additional predictor (X) is added to the model.
5. R^2 = the squared Pearson correlation between the observed data values (y_i's) and the fitted values ($\hat{y}_i$'s) .

R^2 for the Rice Example:

82% of the variability in yield is explained by the linear regression on height and tillers.

To do the calculation "by hand" need to look at the `anova()` output.

$$
\begin{aligned}
R^2 &= \frac{\text{SSModel}}{\text{SSTotal}} \\
&= \frac{(2.424 + 0.208)}{(2.424 + 0.208 + 0.580)} = 0.819 \quad \text{(i.e. 81.9\%)}
\end{aligned}
$$

2. Model assumptions and diagnostic plots

Assumptions for multiple regression model:

1. **Independence:** Observations (and $\varepsilon_i' s$) are independent.
2. **Linear Response:** $E(\varepsilon_i) = 0$ for all i.
   Plot of residuals vs fitted values: should not show a trend.
3. **Equal Variance:** $\mathrm{Var}(\varepsilon_i) = \sigma^2$ same for all i.
   Plot of residuals vs fitted values: should show equal scatter.
4. **Normality:** $\varepsilon_i' s$ are normally distributed.
   QQ plot of residuals: should be linear.

The assumptions are about the residuals (not Y or X's). We look to the data for evidence to support or contradict these assumptions. The primary tools are **residual diagnostic plots**.

(Residual) Diagnostic plots

1. Scatterplot of Y vs X’s: This is a good starting point prior to formal model fitting. Gives some information about linearity and outliers.
2. Residuals versus fitted values: This is the primary diagnostic plot for assessing linearity and constant variance. Curvature, unequal variance (megaphone) or outliers indicate problems.
3. Normal QQplot of residuals: This plot is used to assess normality of residuals. Curvature and outliers indicate problems. Note: The QQ plot is not useful until variance is approximately equal.
4. Residuals versus individual X’s (optional): look primarily for curvature, but also for non-constant variance.

Residuals

Most of the assumptions are about ε_i. The problem is that they are unknown. To check assumptions about ε_i we first estimate them.

A residual is calculated as the difference between the observed and predicted (fitted) values for an observation:

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

Plotting residuals often allows us to see failures of model assumptions better than we can see by looking at the Y vs X scatterplots.

In R, after model fitting, residual diagnostic plots can be easily obtained using `plot(ModelObject)`.

“Standardized” residuals

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

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

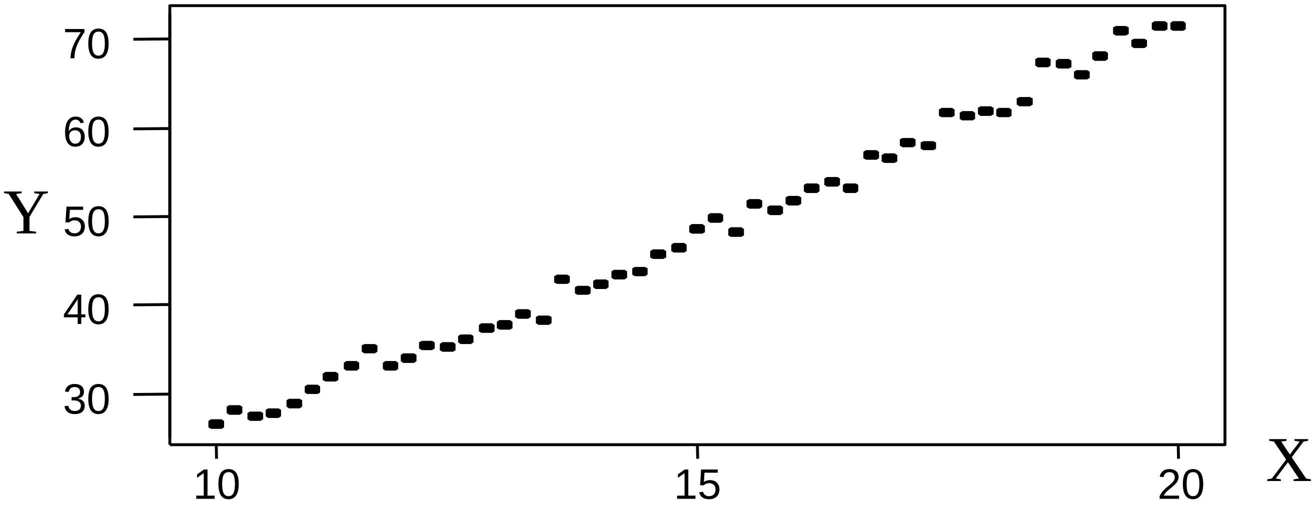
$$\text{Standardized residual: } s_i = \frac{e_i}{\text{SE}(e_i)}$$

Standardized (or studentized) residuals are residuals that have been “standardized” by dividing each residual by it’s SE.

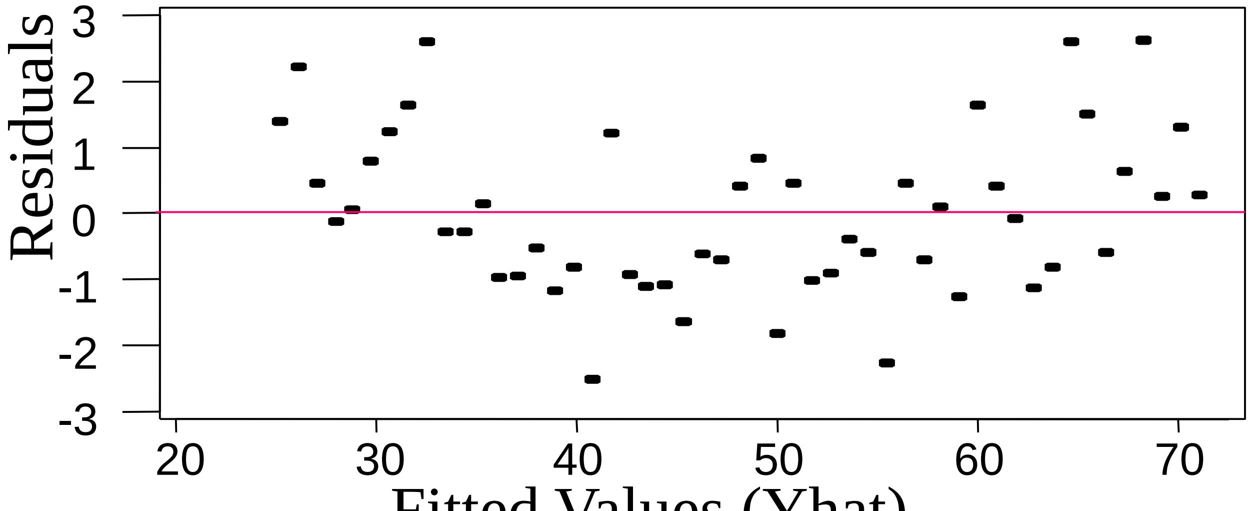
- They have approximately a t-distribution, which is approximately normal for moderate sample sizes, so we expect that about 95% of the standardized residuals will be between -2 and +2. Values greater in absolute value than 3.5 are usually considered outliers.
- In R, standardized residuals can be found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

Checking Regression Assumptions: *Example #1*

Scatter plot of Y vs X

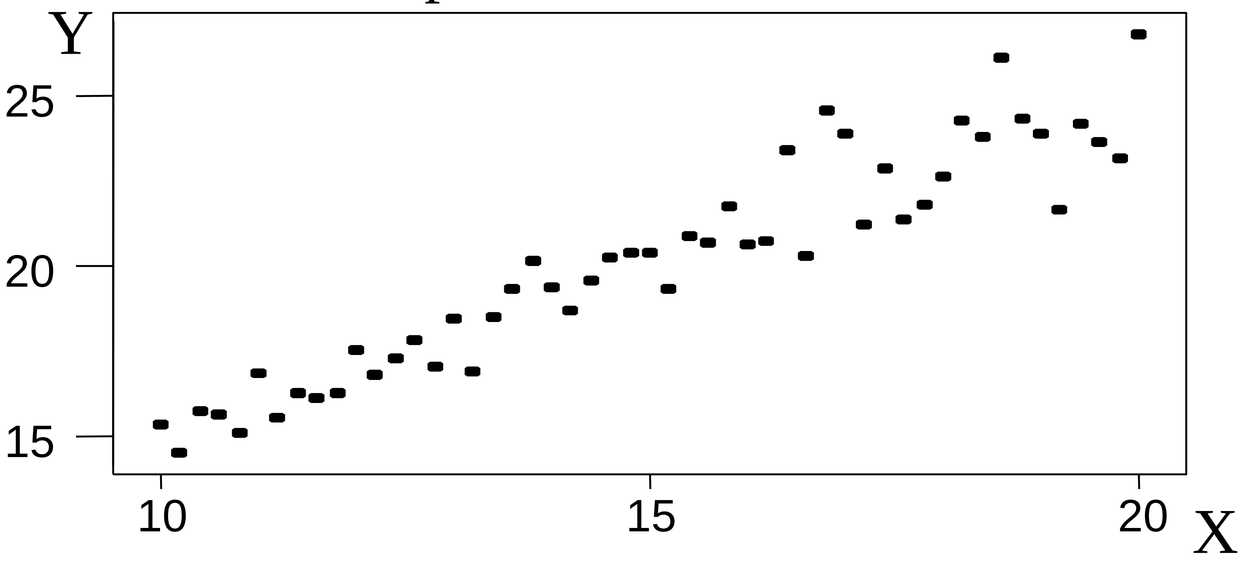


Plot of Residuals vs Fitted Values (Same Data!)

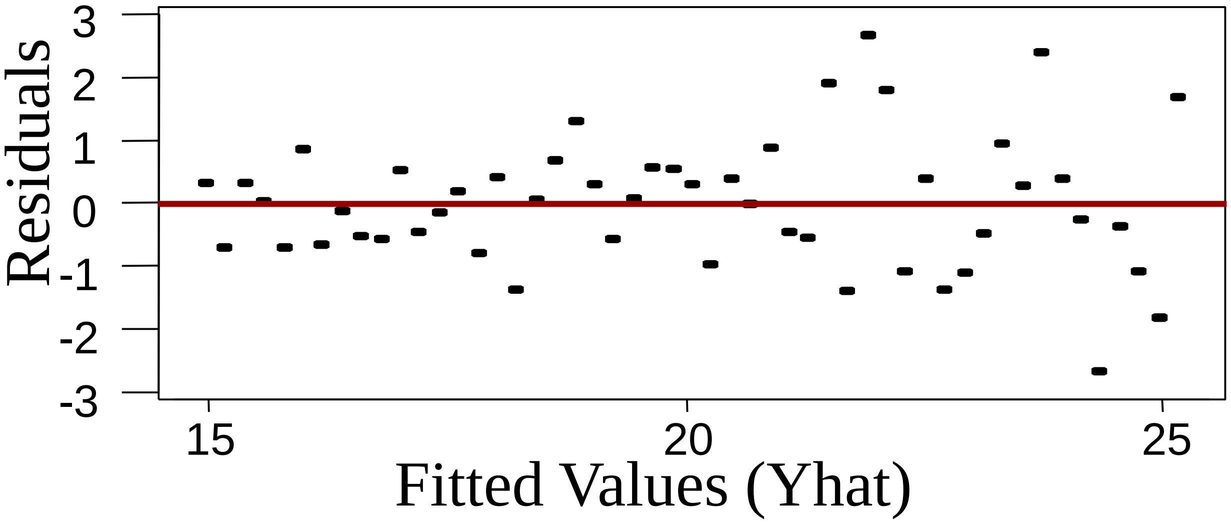


Checking Regression Assumptions: *Example #2*

Scatter plot of Y vs X

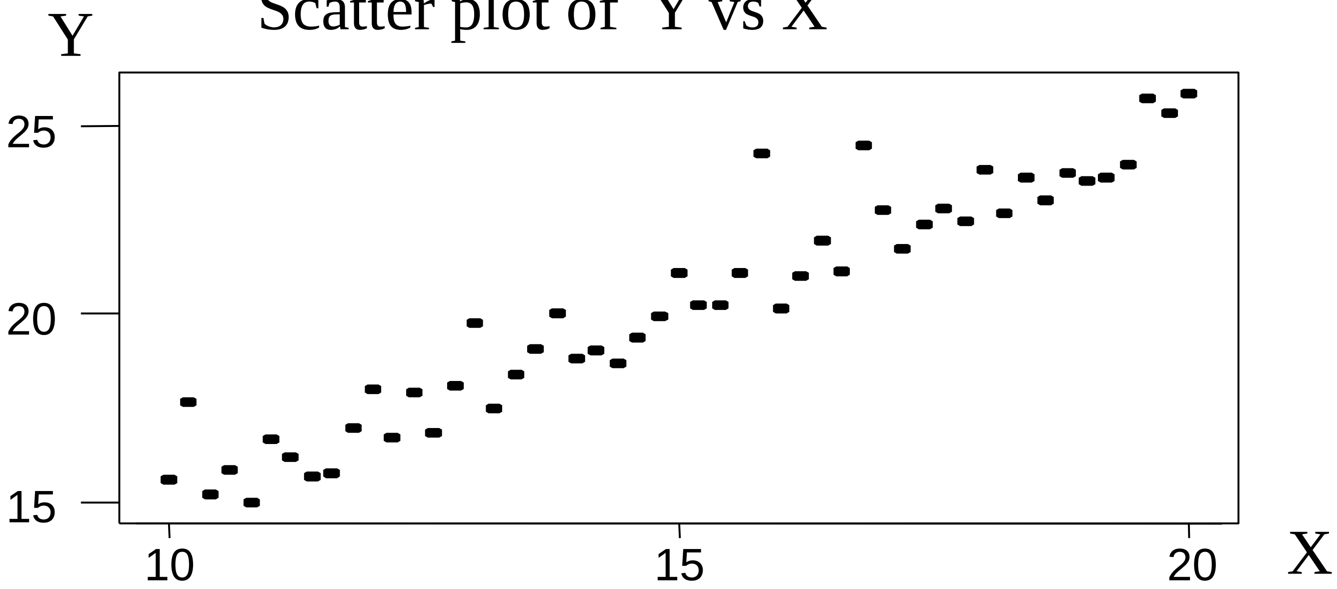


Plot of Residuals vs Fitted Values (Same Data!)

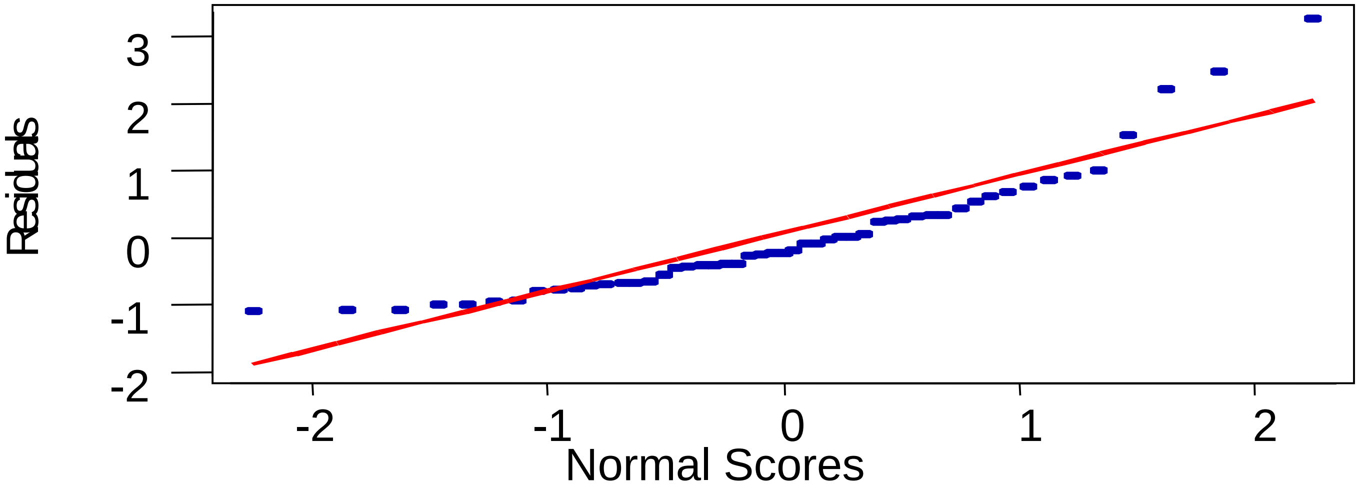


Checking Regression Assumptions: *Example #3*

Scatter plot of Y vs X

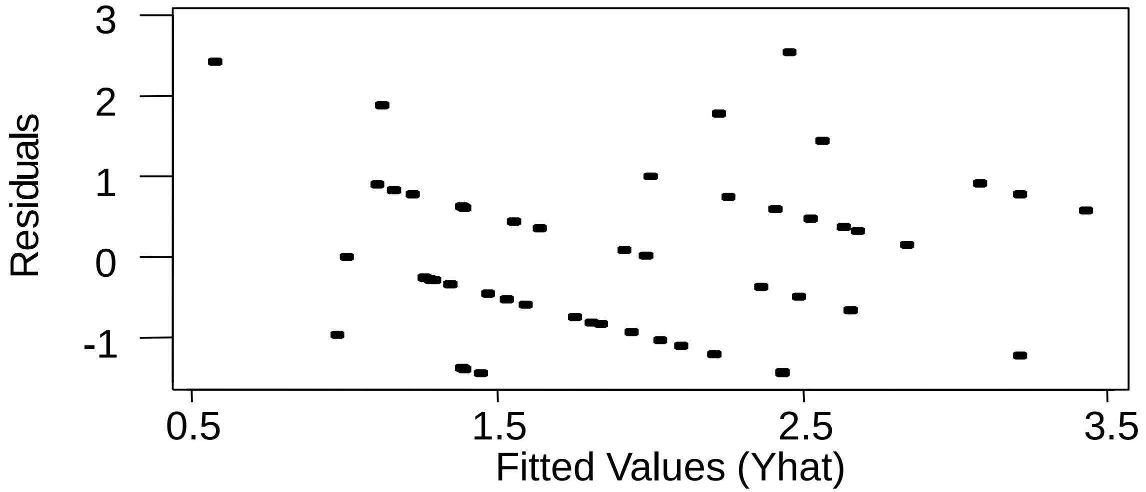


QQPlot of Residuals

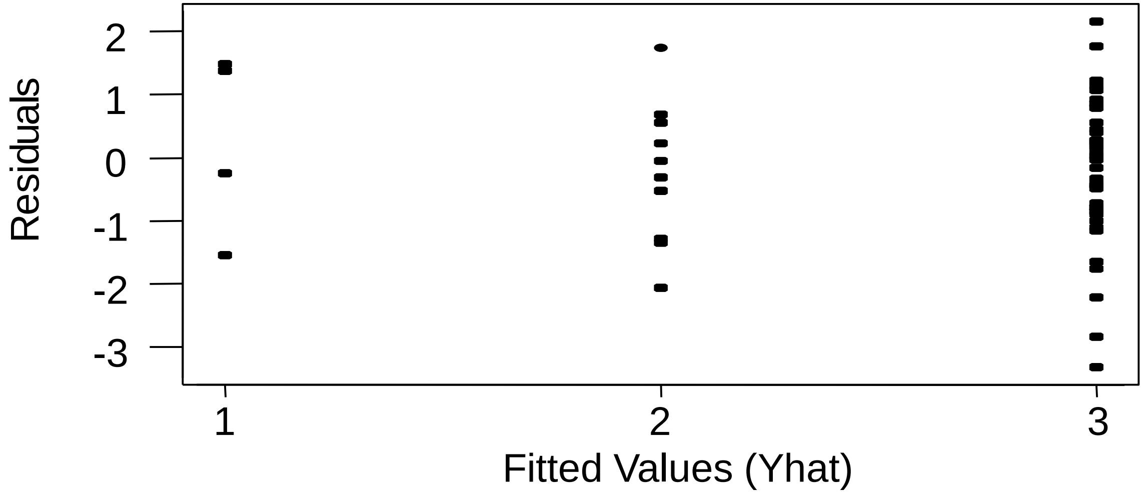


Checking Regression Assumptions: *More Examples*

What does this plot suggest about the data?



Does this plot suggest increasing variance?



3. Basic Hypothesis Testing (for a single β)

$$H_0 : \beta_i = \beta_{i,0} \quad vs \quad H_A : \beta_i \neq \beta_{i,0}$$

$$t = \frac{\hat{\beta}_i - \beta_{i,0}}{SE(\hat{\beta}_i)} \quad \text{with } df = n - k - 1 = \text{dfResid}$$

$$\text{Reject } H_0 \text{ if } |t| > t_{\alpha/2} \text{ or p-value} < \alpha$$

Typically interested in H_0: $\beta_i = 0$ vs H_A: $\beta_i \neq 0$.
For this “default” scenario, estimate, SE, test statistic and p-value are all given in the summary output.
For other scenarios, can use the `lht()` function from the `car` package.

Note: Formula for SE is too complicated to write down without matrix notation, but easily calculated with R!

Important Reminder: The hypotheses and α can (and should!) be stated in advance, before looking at the data.

To conduct a hypothesis test, we compare a test statistic (calculated based on observed data) to a table value or critical value (from a statistical table).

Another approach is to compare the p-value to stated alpha. For fixed alpha, the approaches are equivalent.

Recall the definition of the p-value:

p-value is probability of observing a value of the test statistic **as or more supportive of H_A** than the actual observed value**, given H_0 is true.**

We calculate a p-value based on the observed test statistic and the H_A. If p-value < α, then Reject H_0.

Tests of Individual Parameters using lm()

The basic regression code will produce 2-sided tests of individual parameters with H_0: $\beta_i=0$:

```
Model3 <- lm(yield ~ ht + tillers,
             data = Rice)
```

```
summary(Model3)
```

See the **Rice Example**.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.33560    2.94293   2.153   0.0839
ht          -0.02375    0.01290  -1.842   0.1249
tillers      0.15031    0.11207   1.341   0.2375
```

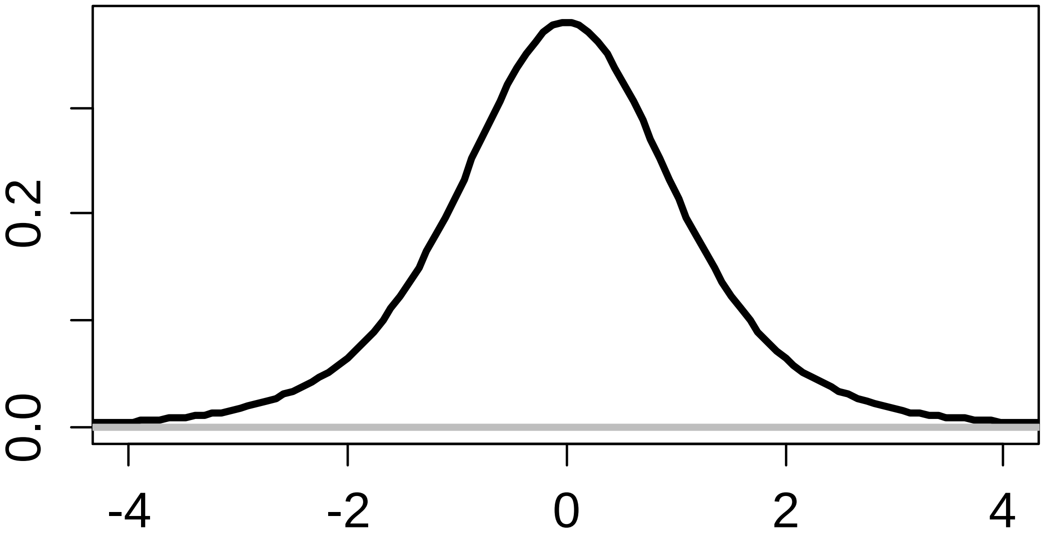
For the Rice Example:

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

$$t = \frac{-0.0237 - 0}{0.0129} = \text{-}1.842 \quad \text{df} = n \text{-} k \text{-} 1 = 5$$

Reject H_0 if $|t| > t_{\alpha/2} = 2.571$ for $\alpha = 0.05$

p-value = 0.1249 > α = 0.05, so we Fail to Reject H_0.

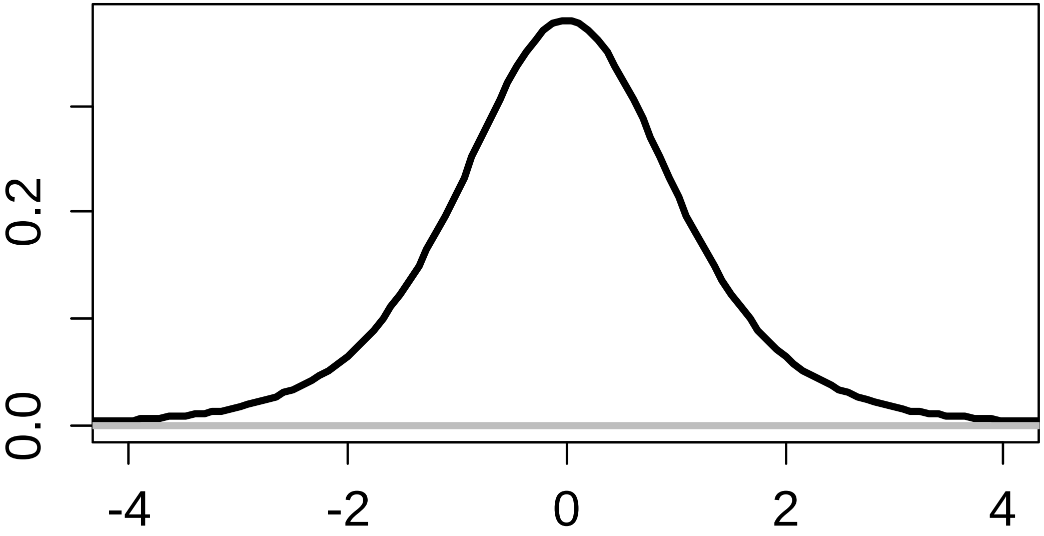


<u>One-sided p-values:</u> $H_0 : \beta_1 \geq 0$ $H_A : \beta_1 < 0$

For the Rice Example: *t* = -1.842; two-sided p = 0.1240

One-sided p-value = the area under the curve, in the direction supporting the alternative.

$$p = 0.1240 / 2 = 0.0620$$



Hypothesis Tests using lht() from the car package

lht stands for linear hypothesis test.

The `lht()` function from the car package is a function to test additional hypotheses (beyond the default tests shown in the summary output).

We need to specify 3 pieces of information:

```
lht(model, hypothesis.matrix, rhs)
```

model is a fitted model object (result of lm() function).
hypothesis.matrix specifies linear combination(s) of coefficients.
rhs gives the “right-hand side” for the hypothesis.

Rice Example: H_0: $\beta_2 = 0.1$ vs H_A: $\beta_2 \neq 0.1$
"By Hand": $t = (0.15031 – 0.1)/0.11207 = 0.4489$

H_0 can also be written as: H_0: $0 \cdot \beta_0 + 0 \cdot \beta_1 + 1 \cdot \beta_2 = 0.1$

Using the `lht()` function from the `car` package.

```
#Test1: B2 = 0.1
> c1 <- c(0, 0, 1)
> lht(Model3, c1, rhs=c(0.1))

Hypothesis:
tillers = 0.1
  Res.Df RSS      Df  Sum of Sq        F  Pr(>F)
1      6 0.60281
2      5 0.57946  1    0.023358 0.2015  0.6723
```

Note: $F = 0.2015 = (0.4489)^2 = t^2$

4. Confidence and Prediction Intervals

The estimate and SE used for hypothesis testing can be used to construct confidence and prediction intervals.

A. Confidence intervals for individual β's.

B. Confidence interval for mean response.

C. Prediction interval for a new observation.

D. Confidence intervals for σ^2 (or σ).

A. **Confidence intervals for any individual $\boldsymbol{\beta_i}$**

A 95% confidence interval for any β_i is:

$$\hat{\beta}_i \pm t_{0.025} \ \mathrm{SE}(\hat{\beta}_i) \qquad \mathrm{df} = \mathrm{dfResid} = \text{n-k-1}$$

In the Rice Example, a 95% CI for β_1 is:

$$\hat{\beta}_1 \pm t_{0.025} SE(\hat{\beta}_1)$$

$$-0.0237 \pm 2.571 * 0.0129$$

$$(-0.0569, 0.0094)$$

That this interval includes zero is consistent with the previous result that $\mathrm{H}_0 : \beta_1 = 0$ was not rejected in a two-sided test.

In R, these CIs can be computed using the `confint()` function.

B. Confidence Interval for Mean Response

$$\mathrm{E(y)} = \beta_0 + \beta_1 \mathrm{x}_1 + \beta_2 \mathrm{x}_2$$

$$\text{Estimate: } \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

$$\text{95\%CI: } \hat{\mathrm{y}} \pm t_{\alpha/2} \mathrm{SE}(\hat{\mathrm{y}}) \qquad \mathrm{df} = \mathrm{dfResid} = \text{n-k-1}$$

Rice Example:

Give a 95% CI for E(y) when ht $= 80$ (x_1) and tillers $= 17$ (x_2):

$$\hat{\mathrm{y}} = \hat{\beta}_0 + \hat{\beta}_1 * 80 + \hat{\beta}_2 * 17 = 6.9911$$

95% CI for Mean Response: (6.425802, 7.556324)

In R use the `predict(Model3 , interval="confidence")` to get the CI for the mean response.

C. Prediction Interval for a New Observation

A 95% prediction interval for y_{new} is given by:

$$\hat{y} \pm t_{\alpha/2} SE(\text{prediction error})$$

$$\hat{y} \pm t_{\alpha/2} SE(\hat{y} - y_{new})$$

$$where\ SE(\hat{y} - y_{new}) = \sqrt{\hat{\sigma}^2 + (SE(\hat{y}))^2}$$

The extra $\hat{\sigma}^2$ insures that the prediction interval will be wider than a corresponding CI for E(y) at the same x's.

Rice Example (ht = 80, tillers = 17):

$\hat{y} = 6.99,$ 95% Prediction Interval (5.949, 8.033)

In R use the `predict(Model3, interval="prediction")` to get the prediction interval for a new Y.

D. Confidence intervals for σ^2

Using the methods of Chapter 7, intervals are based on the Chi-square statistic. The only difference is that df = dfResid = *n-k-1*, rather than *n-1*. A 95% confidence interval for σ^2 is:

$$\boxed{\frac{SS\,\mathrm{Re}\,sid}{\chi^2_{0.025}} < \sigma^2 < \frac{SS\,\mathrm{Re}\,sid}{\chi^2_{0.975}}}$$

Rice Example:

$$\frac{0.57946}{12.83} < \sigma^2 < \frac{0.57946}{0.8312}$$

$$0.0452 < \sigma^2 < 0.6971$$

Notes:

1. A confidence interval for σ can be obtained by taking the square root of each endpoint.

2. A 100(1-α)% confidence interval can be obtained by using $\chi^2_{\alpha/2}$ and $\chi^2_{1-\alpha/2}$.

5. More Hypothesis Testing

These scenarios are less common than inference for a single β.

Remember that your individual research questions should drive the analysis!

A. Hypothesis tests involving combinations of β's

B. Hypothesis tests comparing a large model to a smaller sub-model.

C. Anova() vs anova()

D. Tests involving σ^2.

A. Tests of Combinations of β's

Examples:

H_0: $\beta_1 = 0$ and $\beta_2=0$ (or H_0: $\beta_1= \beta_2= 0$)

H_0: $\beta_1 = \beta_2$ (or H_0: $\beta_1 - \beta_2= 0$)

Another Example: Test the hypothesis that the mean response (height of the plane) is 7 when ht = x1 = 80 and tillers = x2 =17.

$$\mathrm{H}_0 : \ \beta_0 + \beta_1 80 + \beta_2 17 = 7.0$$

$$\mathrm{H}_\mathrm{A} : \ \beta_0 + \beta_1 80 + \beta_2 17 \neq 7.0$$

All of these tests can be done using the `lht()` function from the `car` package. See the **Rice Example.**

Tests of Combinations of β's using lht()

Suppose we want to test H0: $\beta_1=0$ and $\beta_2=0$.

Use the `lht()` function from the `car` package.

```
#Test2: B1 = B2 = 0
c2 <- matrix(c(0, 1, 0,
               0, 0, 1), nrow=2, byrow=TRUE)
lht(Model3, c2, rhs=c(0, 0))
```

See the **Rice Example**.

```
Hypothesis:
ht = 0
tillers = 0
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1      7 3.2115
2      5 0.5795  2    2.6321 11.356 0.01383
```

B. Comparing a Larger Model to a Sub-Model

Say we want to test the hypothesis that some of the β's are simultaneously zero. For ease of notation, we will assume that there are k predictor variables, and the ones we think may be zero are listed last. That is, the first g of the β's are not zero, and the last (*k*-g) are to be tested. We want to test:

$$\mathrm{H}_0 : \ \beta_{g+1} = \beta_{g+2} = ...\beta_k = \ 0$$

H_A : At least one of those $\beta_i\,'s$ is not zero.

Use the "Principle of Conditional Error" (Sec 12.5):

The "full" model is:

$$\mathrm{y_i} = \beta_0 + \beta_1 x_{i1} + ... + \beta_g x_{ig} + \beta_{g+1} x_{i(g+1)} ... + \beta_k x_{ik} + \varepsilon_\iota$$

Form a "reduced" model (the model if H_0 is true):

$$\mathrm{y_i} = \beta_0 + \beta_1 x_{i1} + ... + \beta_g x_{ig} + \varepsilon_\iota \qquad (g < k)$$

F-test for comparing a “full” to a “reduced” model:

$$F = \frac{(\text{SSResid(Red)} - \text{SSResid(Full)}) / (\text{dfResid(Red)} - \text{dfResid(Full)})}{\text{MSResid(Full)}}$$

$$= \frac{(\text{SSResid(Red)} \ - \ \text{SSResid(Full)}) / (k - g)}{\text{MSResid(Full)}}$$

$$= \frac{(\text{SSModel(Full} \ ) \ - \ \text{SSModel(Red} \ )) / (\text{dfModel(Full} \ ) - \ \text{dfModel(Red} \ ))}{\text{MSResid(Full)}}$$

$$\text{df}_1 = \text{df}_{\text{num}} = \text{dfResid(Red)} - \text{dfResid(Full)} = \text{k-g}$$

$$\text{df}_2 = \text{df}_{\text{den}} = \text{dfResid(Full)}$$

Notes: The various forms of the test statistic are all equivalent because SStotal = SSModel + SSResid.

To do the test in R, fit the full and reduced models and then use the `anova()` function or use `lht()` with full model.

Rice Example #1:

$\mathrm{H}_0:\ \beta_1 = \beta_2 = \ 0$

$\mathrm{H}_\mathrm{A}:$ At least one of those $\beta_i\,'s$ is not zero.

The "full" model is:

$$\mathrm{y_i} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i \quad \left(\mathrm{k} = 2\right)$$

The "reduced" model (the model if H_0 is true) is:

$$\mathrm{y_i} = \beta_0 + \varepsilon_i \qquad (\mathrm{g} = 0)$$

$$F = \frac{(\mathrm{SSResid(Red)} - \mathrm{SSResid(Full)}) / (\mathrm{dfResid(Red)} - \mathrm{dfResid(Full)})}{\mathrm{MSResid(Full)}}$$

$$= \frac{(3.212 - 0.579) / (7-5)}{0.116} = 11.36$$

$df\,1 = 2,\ \ df\,2 = 5$

Rice Example #2:

$\mathrm{H_0}: \ \beta_1 = \ 0$

$\mathrm{H_A}: \ \beta_1 \neq \ 0$

The "full" model is:

$$\mathrm{y_i} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i \quad (k = 2)$$

The "reduced" model (the model if $\mathrm{H_0}$ is true) is:

$$\mathrm{y_i} = \beta_0 + \beta_2 x_{i2} + \varepsilon_i \qquad (\mathrm{g} = 1)$$

$$F = \frac{(\mathrm{SSResid(Red)} - \mathrm{SSResid(Full)}) / (\mathrm{dfResid(Red)} - \mathrm{dfResid(Full)})}{\mathrm{MSResid(Full)}}$$

$$= \frac{(0.972 - 0.579) / (6 - 5)}{0.116} = 3.388$$

$df1 = 1, \ df2 = 5$

Rerun both examples using R:
Recall Model3 is the “full” model with both ht and tillers.

Example #1 H_0: $\beta_1 = \beta_2 = 0$

```
anova(Model0, Model3)
lht(Model3, c2, rhs = c(0, 0))
```

This test also appears in the `summary(Model3)` output labeled “F-statistic”.

Example #2 H_0: $\beta_1 = 0$

```
anova(Model2, Model3)
```

This test also appears in the `summary(Model3)` output labeled “ht”.

Notes about the test:

1. A test comparing full and reduced models can be done using `anova()` or `lht()`.

2. The test that ALL β's are simultaneously zero can be formed this way. That test is given automatically in the `summary()` output for an lm model object, labeled "F-statistic". See Example #1.

3.When the models only differ by one parameter, this F-test is equivalent to the t-test shown in the summary output. ($t^2 = F$; k-g = 1). See Example #2. This is important for interpreting the default parameter tests shown in the output! See next slide.

4. We can compare any full model to any in reduced model that can be formed by making linear restrictions on the parameters. We used this test when we did the "lack of fit" for simple linear regression.

Very Important: Notice that the estimate, interpretation and test for a predictor variable depends on the other predictors in the model!

For the Rice Example:

Model 1 (simple linear regression with ht only):
$\widehat{\beta_1}$ = -0.037, p-value = 0.0051

Model 3 (multiple regression with both ht and tillers):
$\widehat{\beta_1}$ = -0.024, p-value = 0.1249

Using Model3, we are testing H_0: $\beta_1 = 0$ (corresponding to ht) given that tillers is already in the model!
In this case, the strong correlation between the predictors (and small sample size) explains why neither predictor is significant in the multiple regression model.

C. anova() vs Anova() in R

The `anova()` function is part of the `stats` package (base R). The `Anova()` function is part of the `car` package.

`anova()` can be used to compare a reduced vs full model.

With only a single predictor variable or factor, there is no difference between `anova()` and `Anova()`.

However, when there are multiple predictors or factors, there can be differences. **We are generally interested in the `Anova()` results!**

When applied to an individual lm object, the `anova()` function will produce a sequential (or Type I) ANOVA table. The resulting table will show tests produced by fitting a sequence of models to the data. The results depend on the order the variables are listed.

The `Anova()` function will produce unique or marginal (or Type III) ANOVA table. The resulting table will show tests for adding one of the predictors to a model that includes all the others. The results do NOT depend on the order the variables are listed.

The `Anova()` function can produce Type II (default) or Type III tests. This only makes a difference when the model contains interactions.

anova() vs Anova() Example

```
> Model3 <- lm(yield ~ ht + tillers, data = Rice)

> Anova(Model3, type = 3)
Anova Table (Type III tests)
             Sum Sq Df F value  Pr(>F)
(Intercept) 0.53711  1  4.6346 0.08395 .
ht          0.39304  1  3.3914 0.12489
tillers     0.20848  1  1.7989 0.23754
Residuals   0.57946  5

> anova(Model3)
Analysis of Variance Table
Response: yield
          Df  Sum Sq Mean Sq F value   Pr(>F)
ht         1 2.42357 2.42357 20.9125 0.005985 **
tillers    1 0.20848 0.20848  1.7989 0.237538
Residuals  5 0.57946 0.11589
```

anova() vs Anova() Example

Note that the `Anova()` results match the tests in the `summary()` output.

The test from `anova()` for ht (first variable listed) matches the regression with just ht.

If we had reversed the order of the of the predictors:

```
lm(yield ~ tillers + ht, data = Rice)
```

The `anova()` results would have changed, but the `Anova()` results would not. Not shown.

For this example (no interaction), the results would have been the same if we used the (default) `Anova( , type = 2)`. Not shown.

D. Tests involving σ^2

These tests follow the same method used in Chapter 7. The only difference is that the df = dfResid = *n-k-1*, rather than *n-1*. Tests are based on the Chi-square, which is formed by multiplying the estimate of σ^2 by its df then dividing by its hypothesized value.

In the Rice Example:

$H_0 : \sigma^2 \le 0.10 \qquad (\sigma_0^2 = 0.10)$

$H_A : \sigma^2 > 0.10$

$$\chi^2 = \frac{(n-k-1)MS\,\mathrm{Re}\,sid}{\sigma_0^2} = \frac{SS\,\mathrm{Re}\,sid}{\sigma_0^2} = \frac{0.57946}{0.10} = 5.7946$$

Reject H_0 if $\chi^2 \;>\; \chi_\alpha^2$ with df = dfResid $=$ n-k-1 $=$ 8-2-1 $=$ 5

From Table 5: $\chi_{0.05}^2 = 11.07$ Fail to Reject H_0.

From R: pvalue = 1 - pchisq(5.7946, df = 5)

From R: pvalue=0.327

6. Interpretation and Causality

We can use simple linear regression to establish an association between X and Y.

But even a statistically significant association does NOT imply a causal relationship. Correlation is not causation!

If association has been shown, to show causality we must show **both**:

1. Y does not cause X.
2. The relationship between X and Y is not the result of a mutual relationship with other (confounding) variables.

Usually, we argue from the physical situation that Y does not cause X. For example: yield does not cause ht.

To show that the relationship between X and Y is not the result of a mutual relationship with other (confounding) variables, it is sufficient to show that X **or** Y is not related to other variables.

For experiments, random assignments of X to experimental units assures that X is not related to other variables.

For observational studies without randomization, it is very difficult to record and/or rule out all possible confounding variables to establish a casual conclusion. However, multiple regression can be a stronger approach towards causality.

The strategy for showing that the relationship between X_1 and Y is not the result of a mutual relationship with other variables is to include those other variables additional predictors.

Multiple regression estimates a relationship between X_1 and Y, “controlling for” or “adjusting for” their mutual relationship to other variables X_2, … X_k

Recall: $\beta_1 =$ change in the mean of Y, associated with a unit change in X_1, all other X’s in the model held constant (or controlled for or adjusted for).

But it still remains very difficult (or impossible) to measure and/or rule out all possible confounding variables to establish a casual conclusion.

Fuel Example (Weisberg):

In this observational study, data for 48 states relating various state road and income variables to per capita fuel consumption. Of particular interest is the relationship between fuel tax and fuel consumption.

Question: How much would fuel consumption change in response to a change in the fuel tax? (i.e. Does increased fuel tax cause decreased fuel consumption.)

Fuel (Y) = fuel consumption (gallons per capita)

Tax (X_1) = motor fuel tax (cents per gallon)

Road (X_2) = federal primary-aid highways (miles)

Dlic (X_3) = percent of population with driver's licenses.

Inc (X_4) = per capita income.

$$\text{Model1}: \hat{y} = 984.0 - 53.11 * \text{Tax}$$

$$\begin{aligned} \text{Model2}: \hat{y} &= 377.3 - 34.8 * \text{Tax} + 13.4 * \text{Dlic} \\ &\quad - 66.6 * \text{Inc} - 2.43 * \text{Road} \end{aligned}$$

Model1: A penny increase in the fuel tax is associated with a 53.11 gal/capita predicted decrease in fuel consumption (without controlling for other variables).

Model2: Dlic, Inc and Road held constant, a penny increase in the fuel tax is associated with a 34.8 gal/capita predicted decrease in fuel consumption.

To say that a change in tax would cause a change in consumption requires more information: Have other relevant variables (that might be related to both fuel consumption and fuel tax) been omitted?

Additional question: To what population is inference being made here?

Weight Loss Example (Ott and Longnecker):

Y=Wt_loss = weight loss of a compound

X_1=Time = time exposed to air

X_2= Humidity = humidity of environment

n = 12 combinations of time and humidity randomly assigned.

$$\text{Model1}: \hat{y} = -1.73 + 1.32 * \text{Time}$$

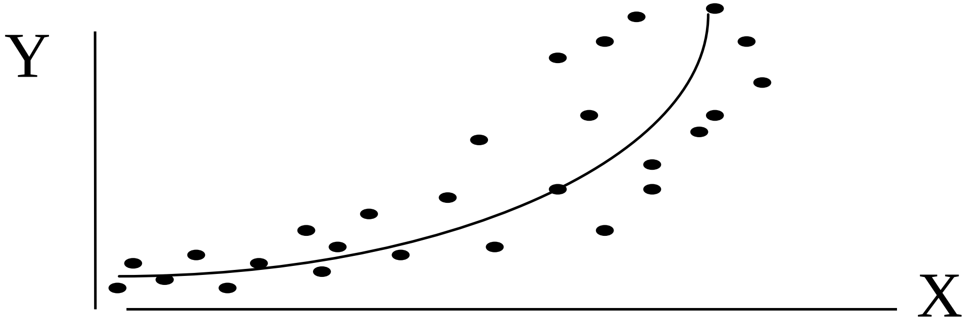
$$\text{Model2}: \hat{y} = 7.91 - 8.00 * \text{Humid}$$

$$\text{Model3}: \hat{y} = 0.67 + 1.32 * \text{Time -} 8.00 * \text{Humid}$$

1. Can we conclude that changing time or humidity causes a change in weight loss?
2. To which population are we making inference?
3. Why are the slope estimates in the simple regressions the same as the slope estimates in the multiple regression?

7. Transformations for linearity and homogeneity of variance.

When the residual plots indicate non-linearity, one approach is to consider transformation of Y or one or more of the X’s. The same basic ideas discussed for simple linear regression also apply to the case of multiple regression; however, the decisions are considerably more complicated because a transformation intended to correct a nonlinear relationship between Y and one of the X’s, also affects the other X’s.



If the relationship between Y and the X’s has a particular shape, we can transform X’s by that shape, or Y by the inverse of that shape. (e.g. if Y looks like an <u>exponential</u> function of X, then we can transform X by the exponential function, or transform Y by the inverse of the exponential function, the <u>log</u> function).

Some practical considerations:

1. Try to find the appropriate scale for Y before putting a lot of effort into selecting the X’s.

2. Highly skewed variables (X’s or Y’s) often need transformation.

3. If the Y data varies over orders of magnitude, it will usually need a transformation.

4. Transforming Y will affect the residual variance, transforming X will not.

5. Transformation of Y will affect the linearity of the relationship between Y and all the X’s.

6. Transforming Y will often solve a linearity and a variance problem at the same time.

8. Checking for Outliers

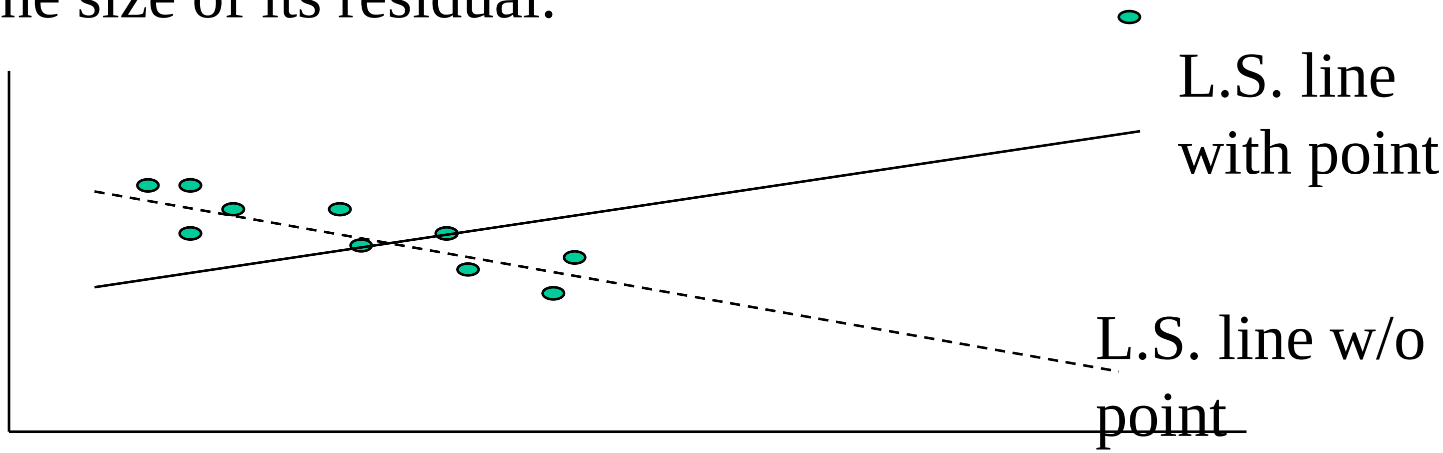
Outlier: An observation (data point) that does not follow the same distribution as the rest of the observations.

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 studentized 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

An appropriate statistic to address problems (1) and (2) from the previous slide.

Idea of the R-Student Residual: For each point (*i)*, estimate the regression line without that point; use that regression line to estimate y_i and σ^2; compute the studentized residual using those estimates. Note: –i indicates without observation i! Note: df = n – k – 2.

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: $t_i = \dfrac{e_{-i}}{\mathrm{SE}(\mathrm{e}_{-i})}$

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

Review of 3 Types of Residuals

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

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

$$\text{Rstudent } = t_i = \frac{e_{-i}}{\text{S}E_{\text{-i}}(e_{-i})} = \frac{y_i - \tilde{y}_i}{\text{S}E_{\text{-i}}(y_i - \tilde{y}_i\ )}$$

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()`.

Outlier Test and Multiple Testing Adjustment

H_0: Observation i is NOT an outlier
H_A: Observation i is an outlier
Test statistic = t = Rstudent residual
df = n – k – 2.

Multiple Testing Adjustment: We recognize that we are testing that particular data point because it is the largest out of n. By testing the largest outlier, we are effectively doing n tests. A Bonferroni adjustment for multiple comparisons involves comparing the Rstudent to the $\alpha/(2n)$ value of the t-distribution with df = n – k - 2 (for alpha=0.05)

```
tcrit = qt(1-0.05/(2*n),df = n-k-2)
```

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

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

Fuel Example: Wyoming had the largest residual:

$s=3.734$ (Standardized residual)

$t=4.490$ (Rstudent residual)

The critical value for $\alpha=0.05$ is obtained from R:

```
tcrit = qt(1-0.05/(2*48), df = 48-4-2)
pval = 2*48*(1-pt(4.490, df = 48-4-2))
```

Results: $\alpha=0.05$ gives tcrit=3.52

p-value = 0.0026 $< \alpha$

Conclusion: Reject H0 and conclude Wyoming is an outlier.

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.