Goals: Understanding the Simple Regression Model

Simple Regression model: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$

- Y_i is the value of the response variable in the i^{th} trial,
- β_0 and β_1 are parameters
- The regressor variable X_i is assumed to be under the control of the experimenter, who can set their values. That is why X_i are considered as constants.
- ε_i is a random error term

Properties of residuals:

- mean zero $E\{\varepsilon_i\} = 0$
- constant variance $\sigma^2\{\varepsilon_i\} = \sigma^2$;
- ε_i and ε_j are independent $i \neq j$
- ε_i is normally distributed

It is "simple" in that there is only one predictor variable, "linear in the parameters," because no parameter appears as an exponent or is multiplied or divided by another parameter

Consequences of the Assumptions:

- $\mu_i = E(Y_i) = \beta_0 + \beta_1 X_i$
- Var(Y_i)= σ^2 is constant, all observations have the same precision
- Y_i and Y_j are independent $i \neq j$

Method of Least Squares

$$\boxed{\text{minimize} \quad S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2}$$

How to minimize S? take partial derivatives and set them equal to 0.

$$\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_1} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) x_i = 0$$

$$\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0 \quad \text{same as} \sum_{i=1}^{n} y_i = n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i$$

Dividing by n we get: $\beta_0 = \bar{y} - \beta_1 \bar{x}$

For the 2nd one:

$$\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) x_i = 0 \text{ same as} \sum_{i=1}^{n} y_i x_i = \beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} {x_i}^2$$

Replacing β_0 by $\bar{y} - \beta_1\bar{x}$

$$\sum_{i=1}^{n} y_i x_i = (\bar{y} - \beta_1\bar{x}) \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i{}^2 = \bar{y} \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i(x_i - \bar{x})$$

$$\sum_{i=1}^{n} y_i x_i - \bar{y} \sum_{i=1}^{n} x_i = \beta_1 \sum_{i=1}^{n} x_i(x_i - \bar{x})$$

$$\sum_{i=1}^{n} (y_i - \bar{y}) x_i = \beta_1 \sum_{i=1}^{n} x_i(x_i - \bar{x})$$

It turns out that:

$$\sum_{i=1}^{n} (y_i - \bar{y}) x_i = \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x}) \text{ because } \sum_{i=1}^{n} (y_i - \bar{y})\bar{x} = 0$$

And

$$\sum_{i=1}^{n} x_i(x_i - \bar{x}) = \sum_{i=1}^{n} (x_i - \bar{x})^2 \text{ because } \sum_{i=1}^{n} (x_i - \bar{x})\,\bar{x} = 0$$

And so we solve for β_1

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

In summary, we have obtained the estimates

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} \quad \text{and} \quad \hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{s_{xy}}{s_{xx}}$$

$\hat{y}_i = E(y_i) = \mu_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ predicted value = fitted value

Estimation of the Variance σ^2: Mean Square Error (MSE)

$$s^2 = \frac{S(\hat{\beta}_0, \hat{\beta}_1)}{n-2} = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-2}$$

Consequences:

1. $\sum_{i=1}^{n} e_i = 0$ because $e_i = \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)$, see derivative with respect to β_0
2. $\sum_{i=1}^{n} e_i x_i = 0$, see derivative with respect to β_1
3. $\sum_{i=1}^{n} \hat{y}_i e_i = 0$, because
$$\sum_{i=1}^{n} \hat{y}_i e_i = \sum_{i=1}^{n} (\hat{\beta}_0 + \hat{\beta}_1 x_i) e_i = \hat{\beta}_0 \sum_{i=1}^{n} e_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i\, e_i = 0$$
4. $(\bar{x}, \bar{y})$ is a point in the regression line $\hat{\mu} = \hat{\beta}_0 + \hat{\beta}_1 x$
5. $S\left(\hat{\beta}_0, \hat{\beta}_1\right) = \sum_{i=1}^{n} e_i^2$ is the minimum value of $S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$

Before continuing, a little review about tests of hypothesis and confidence intervals

Basic steps of hypothesis testing

1. H_0 Null hypothesis ("no effect")
   H_a Alternative hypothesis ("some effect")

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

Null: there is no useful linear relationship between X and Y

Alternative: there is a significant relationship between X and Y

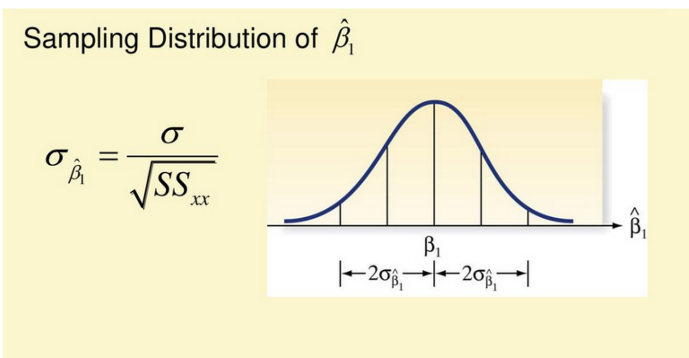
2. Test statistics (depends on the null hypothesis)

$$t = \frac{\hat{\beta}_1}{\text{se}(\hat{\beta}_1)}$$

3. Determine the "sampling distribution"

$$\hat{\beta}_1 \sim N(\beta_1, \sigma^2(\beta_1))$$

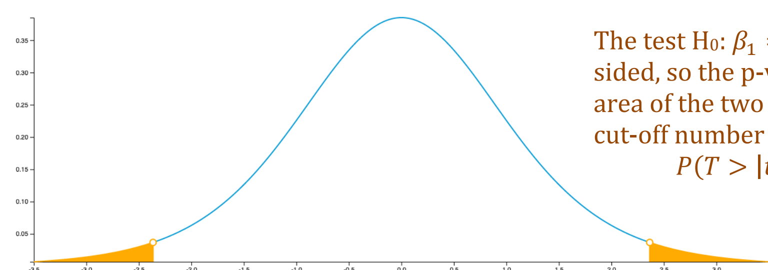
   If H_0 is true, and we "drew" many samples of size n from this population, calculating t for each sample, what would be distribution of these t values?



4. When model assumptions are true and H_0 is true, statistical theory says:

$$\hat{\beta}_1 \sim N(0, \sigma^2(\beta_1))$$

5. Find the p value: P-value is probability of observing a difference (t) at least as extreme as what was seen, just by chance, when H_0 is true.

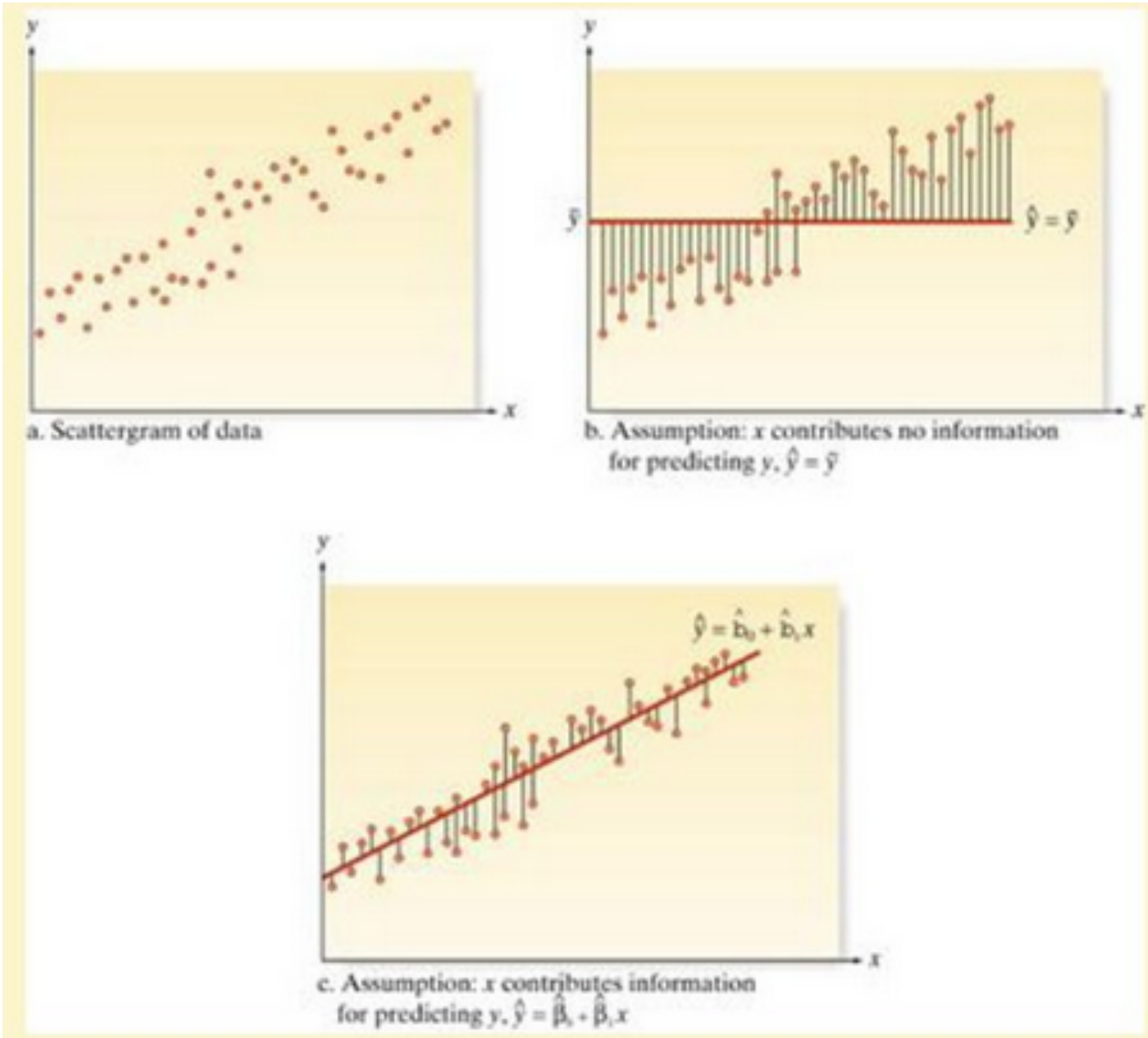


The test H_0: $\beta_1 = 0$ vs H_a: $\beta_1 \neq 0$ is two-sided, so the p-value for our t-test is the area of the two yellow regions where the cut-off number is the value we got for t

$$P(T > |t|) + P(T < -|t|)$$

6. Make conclusions in context (Is X useful in the model???)

In linear regression, we will run test of "utility" for our variables or models.



y
x
a. Scattergram of data
y
$\bar{y}$
$\hat{y} = \bar{y}$
x
b. Assumption: x contributes no information for predicting y, $\hat{y} = \bar{y}$
y
$\hat{y} = \hat{b}_0 + \hat{b}_1 x$
x
c. Assumption: x contributes information for predicting y, $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

Assessing the Utility of the Model
Making inferences about the slope

Inferences about the Regression Parameters
The sampling distribution for the Slope:

$\hat{\beta}_1$ is Normal

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{\sum (x_i - \bar{x})\ y_i}{\sum (x_i - \bar{x})^2} = \sum_{i=1}^{n} c_i y_i$$

where $c_i = {}^{(x_i - \bar{x})}/_{S_{xx}}$, that is, it is a linear combination of independent normal random variables. Hence, it is itself a normal random variable.
Mean:

$$E(\hat{\beta}_1) = \sum_{i=1}^{n} c_i E(y_i\,) = \sum_{i=1}^{n} c_i\ (\beta_0 + \beta_1 x_i) = \beta_1 \sum_{i=1}^{n} c_i\, x_i = \beta_1$$

Variance:

$$\mathrm{V}(\hat{\beta}_1) = \sum_{i=1}^{n} {c_i}^2\ \mathrm{V}(y_i\,) = \sum_{i=1}^{n} {c_i}^2\, \sigma^2\ = \sigma^2\ \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{{S_{xx}}^2} = \frac{\sigma^2}{S_{xx}}$$

$$\boxed{\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{S_{xx}})}$$

Estimation of the standard deviation of $\hat{\beta}_1$is the standard error:

$$se(\hat{\beta}_1) = \frac{s}{\sqrt{S_{xx}}}$$

T-distribution with n-2 df:

$$T = \frac{\hat{\beta}_1 - \beta_1}{se(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - \beta_1}{\frac{s}{\sqrt{S_{xx}}}} = \overbrace{\frac{\hat{\beta}_1 - \beta_1}{\sigma/\sqrt{S_{xx}}}}^{\text{Standard Normal}} \Bigg/ \underbrace{\sqrt{\frac{(n-2)s^2}{\sigma^2(n-2)}}}_{\text{Chi square with df=n-2}}$$

Test statistic for utility test: H₀: $\beta_1\ = 0$

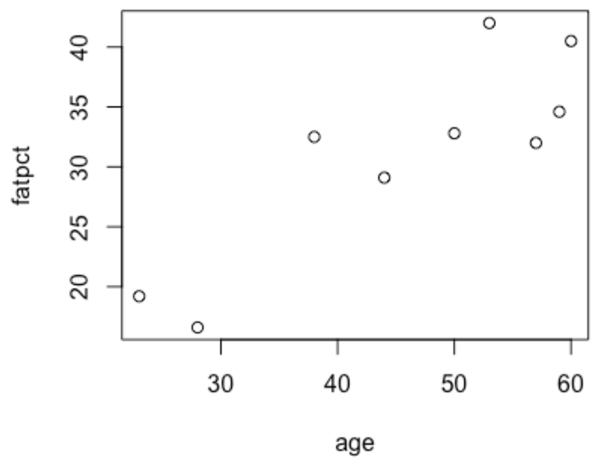
$$T = \frac{\hat{\beta}_1}{s/\sqrt{S_{xx}}}$$

under the Null Hypothesis this has a t-distribution with n-2 degrees of freedom.

(1-α)100% confidence interval for the slope: $\hat{\beta}_1 \pm t_{\alpha/2}\ se(\hat{\beta}_1)$
(estimate) $\pm$(t-critical value)(standard error of estimate)

Example: Study of percent of bodyfat and age

age	fatpct
23	19.2
28	16.6
38	32.5
44	29.1
50	32.8
53	42
57	32
59	34.6
60	40.5



Question: what is the relationship between age and fatness? (`plot(agefat)`)

```
model<-lm(fatpct~age,data=agefat)
summary(model)
```

Call:
lm(formula = fatpct ~ age, data = agefat)

Residuals:

Min	1Q	Median	3Q	Max
-5.1149	-3.5987	-0.5214	1.7593	7.0528

Coefficients:

	Estimate		Std. Error	t value	Pr(>|t|)
(Intercept)	$\boxed{6.2254}$	β_0	5.7114	1.09	0.31181
age	$\boxed{0.5419}$	β_1	0.1202	4.51	0.00277 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.61 on 7 degrees of freedom
Multiple R-squared: 0.744, Adjusted R-squared: 0.7074
F-statistic: 20.34 on 1 and 7 DF, p-value: 0.002765

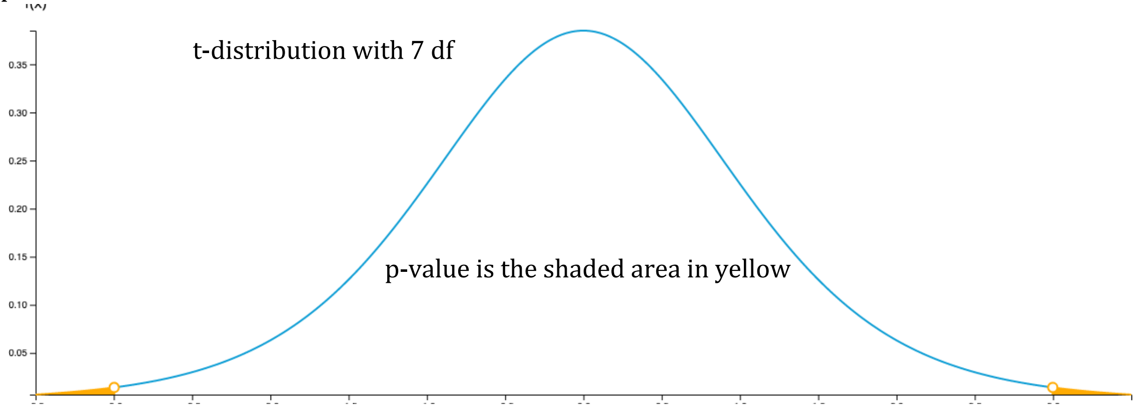
Estimated regression line:

$$\hat{Y}_i = 6.2254 + 0.5419\, X_i$$

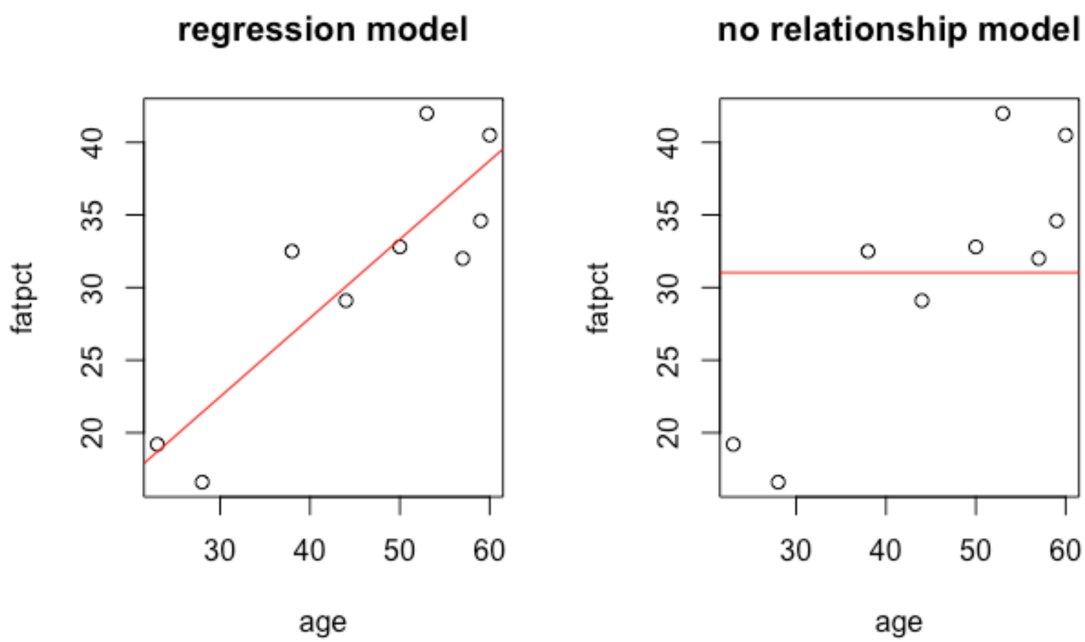
Utility Test: $H_0: \beta_1 = 0$ vs $H_a: \beta_1 \neq 0$

$$se(\hat{\beta}_1) = \frac{s}{\sqrt{S_{xx}}} = 0.1202 \qquad t = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)} = \frac{0.5419}{0.1202} = 4.51$$

p-value: 0.00277



Conclusion: there is a significant linear relationship between %fat and age (age is significant in the model)



Model:

$$\hat{Y}_i = 6.2254 + 0.5419\, X_i$$

A point estimate for the percent of bodyfat for a person that is 45 years old is:

$$\hat{Y}_i = 6.2254 + 0.5419(45) = 27.9\%$$

This is both an estimate for the percent of bodyfat for a 45 years old and the average percent of bodyfat for a 45 years old. Two interpretations that are quite different!

To create confidence intervals for the predictions we need a little more information.

Other inferences:
The Intercept:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \sum_{i=1}^{n} d_i y_i$$

where $d_i = \frac{1}{n} - \frac{(x_i - \bar{x})\bar{x}}{s_{xx}}$

Thus $\hat{\beta}_0$ is also normal with mean β_0 and variance $\left[\frac{1}{n} + \frac{\bar{x}^2}{s_{xx}}\right]\sigma^2$

$$se(\hat{\beta}_0) = s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{s_{xx}}}$$

$T = \frac{\hat{\beta}_0 - \beta_0}{se(\hat{\beta}_0)}$ also has a t-distribution with n-2 df.

Confidence Interval for β_0: $\hat{\beta}_0 \pm t_{\alpha/2}\ se(\hat{\beta}_0)$

Test of Hypothesis about β_0:

H$_0$: $\beta_0 = 0$ vs H$_a$: $\beta_0 \neq 0$

Test statistic: $T = \frac{\hat{\beta}_0}{se(\hat{\beta}_0)}$, under the Null Hypothesis has the t-distribution with n-2 df

Another utility test
ANOVA Table: F test for adequacy of model (overall fitness of the model)

Sums of squares:

$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$ Total variation (of the Y variable w/r to its mean)

$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2$ Sum of square errors

$SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ Regression sum of squares

Each one of these sums correspond to a Chi-square distribution with certain degrees of freedom

Fundamental Equation: $\boxed{SST=SSE+SSR}$

And the corresponding equation for their degrees of freedom $df_T = df_E + df_R$

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + 2\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y})$$

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = \sum_{i=1}^{n} e_i(\hat{y}_i - \bar{y}) = \sum_{i=1}^{n} e_i \underbrace{\hat{y}_i}_{\hat{\beta}_0 + \hat{\beta}_1 x_i} - \bar{y}\underbrace{\sum_{i=1}^{n} e_i}_{=0}$$

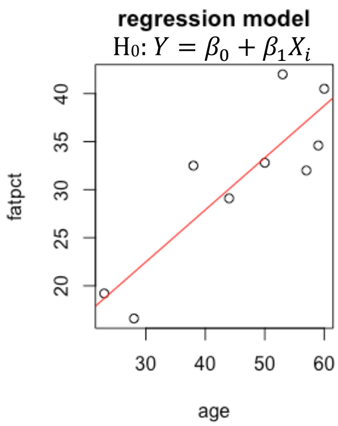
$$= \hat{\beta}_0 \underbrace{\sum_{i=1}^{n} e_i}_{=0} + \hat{\beta}_1 \underbrace{\sum_{i=1}^{n} e_i x_i}_{=0} = 0$$

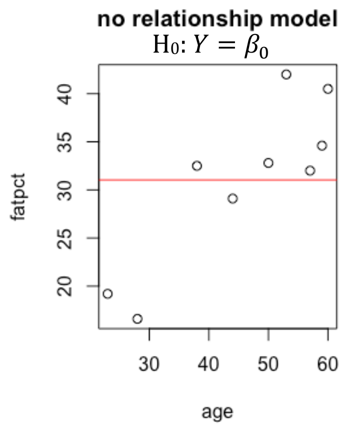
See the least square method where we got these relationships

ANOVA Table

Source	SS	df	MS = ss/df	F	p-value
Regression	SSR	1	MSR=SSR	MSR/MSE	
Errors	SSE	n-2	MSE=SSE/(n-2)		
Total	SST	n-1			

F-test: Reduced model H0: $Y = \beta_0 + \varepsilon$ (X has no linear association with Y)
Full model Ha: $Y = \beta_0 + \beta_1 X_i + \varepsilon$ (X has linear association with Y)

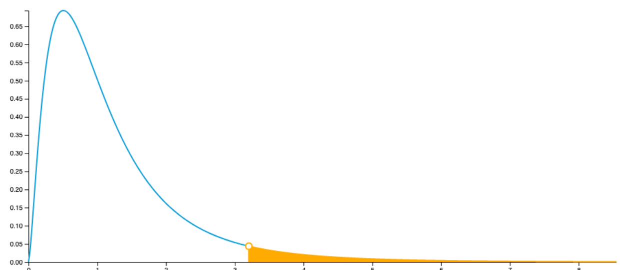




Test statistic: $F = \frac{MS_{model}}{MS_{error}} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2/1}{\sum_{i=1}^{n} e_i^2/n-2}$

F-statistic looks at change in SSerror between these two models

This is a one-sided test. P-value: $P(F > \text{value})$



In the age vs %bodyfat

F-statistic: 20.34 on 1 and 7 DF, p-value: 0.002765

Source	SS	df	MS = ss/df	F	p-value
Regression	432.16	1	432.16	20.339	0.002765
Errors	148.74	7	21.25		
Total	580.90	8			

```
nova(model)
```

Analysis of Variance Table

Response: fatpct

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
age	1	432.16	432.16	20.339	0.002765 **
Residuals	7	148.74	21.25		

Notice that the p-value is exactly the same as the test for the slope!

Fact: (statistic for test for the slope)^2 = statistic for F-test (under the null hypothesis)

Under H₀: $\beta_1 = 0$

Standard Normal

$$T = \frac{\hat{\beta}_1 - \beta_1}{se(\hat{\beta}_1)} = \frac{\hat{\beta}_1}{\frac{s}{\sqrt{S_{xx}}}} = \frac{\hat{\beta}_1}{\sigma/\sqrt{S_{xx}}} \Big/ \sqrt{\frac{(n-2)s^2}{\sigma^2(n-2)}}$$

Chi square with df=n-2

Work: Show T^2 has the F distribution

Basic Measures of Fit

1. Coefficient of determination: R^2

$$R^2 = \frac{SS_{model}}{SS_{total}} = 1 - \frac{SS_{error}}{SS_{total}}$$

Percent of variation in Y explained by the model

In the example of percent of bodyfat vs age:
Residual standard error: 4.61 on 7 degrees of freedom
Multiple R-squared: 0.744, Adjusted R-squared: 0.7074

$$R^2 = \frac{432.16}{580.90} = 0.7074$$

70.74% of variation in percent of bodyfat can be explained by its linear association with age

2. MSE = Mean Square Error

$$MSE = s^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2}$$

Inference about Predicted Mean for a new observation:

$$\hat{\mu}_p = \hat{\beta}_0 + \hat{\beta}_1 x_v$$

Thus $\hat{\mu}_v = \sum_{i=1}^{n} l_i y_i$, is normal, where $l_i = \frac{1}{n} + \frac{(x_i - \bar{x})(x_v - \bar{x})}{s_{xx}}$.
Mean: $E(\hat{\mu}_v) = \beta_0 + \beta_1 x_v$ and
Variance: $\mathrm{V}(\hat{\mu}_v) = \sigma^2 \left[\frac{1}{n} + \frac{(x_v - \bar{x})^2}{s_{xx}}\right]$
Standard Error: $\mathrm{se}(\hat{\mu}_v) = s\sqrt{\frac{1}{n} + \frac{(x_v - \bar{x})^2}{s_{xx}}}$
T-distribution with n-2 df: $T = \frac{\hat{\mu}_v - (\beta_0 + \beta_1 x_v)}{\mathrm{se}(\hat{\mu}_v)}$

Confidence Interval: $(\hat{\beta}_0 + \hat{\beta}_1 x_v) \pm t_{\alpha/2}\, \mathrm{se}(\hat{\mu}_v)$

Inference about Predicted Individual value for a new observation:

$$y_v = \beta_0 + \beta_1 x_v + \varepsilon_v$$

Because the new error is independent of the previous observations, y_v is also normal.
Mean: $E(y_v) = \beta_0 + \beta_1 x_v$ which is the same as the mean of $\hat{y}_v = \hat{\beta}_0 + \hat{\beta}_1 x_v$
Thus $E(y_v - \hat{y}_v) = 0$
Variance of y_v: $\mathrm{V}(y_v) = \mathrm{V}(\varepsilon_v) = \sigma^2$
Variance: $V(\hat{y}_{new}) = \sigma^2 \left[\frac{1}{n} + \frac{(x_v - \bar{x})^2}{s_{xx}} + 1\right]$
Standard Error: $se(\hat{y}_{new}) = s\sqrt{\frac{1}{n} + \frac{(x_v - \bar{x})^2}{s_{xx}} + 1}$

T-distribution with n-2 df: $T = \frac{y_v - \hat{y}_v}{\mathrm{se}(y_v - \hat{y}_v)}$
Confidence Interval: $(\hat{\beta}_0 + \hat{\beta}_1 x_v) \pm t_{\frac{\alpha}{2}} \mathrm{se}(\hat{y}_{new})$

Standard errors and confidence intervals are larger for Predictions that for Estimates

mean response	individual response
se for Estimate $$\mathrm{se}(\hat{\mu}_v) = s\sqrt{\frac{1}{n} + \frac{(x_v - \bar{x})^2}{s_{xx}}}$$	se for Prediction $$se(\hat{y}_{new}) = s\sqrt{\frac{1}{n} + \frac{(x_v - \bar{x})^2}{s_{xx}} + 1}$$
100(1-α)% confidence interval for Estimate $(\hat{\beta}_0 + \hat{\beta}_1 x_v) \pm t_{\alpha/2}\, \mathrm{se}(\hat{\mu}_v)$	100(1-α)% confidence interval for Prediction $(\hat{\beta}_0 + \hat{\beta}_1 x_v) \pm t_{\frac{\alpha}{2}}\, \mathrm{se}(\hat{y}_{new})$

In R: model<-lm(Y~X, data=dataname); summary(model); plot(model)

- confint(regmodel) #CIs for all parameters

  confint(model)

```
              2.5 %          97.5 %
(Intercept) -7.2798266      19.7306270
age          0.2577804       0.8260613
```

- predict.lm(regmodel, interval="confidence") #make prediction and give confidence interval for the mean response
- predict.lm(regmodel, interval="prediction") #make prediction and give prediction interval for the individual response

predict.lm(model, interval="confidence")

```
      fit      lwr      upr
1 18.68958 11.26740 26.11176
2 21.39918 15.17686 27.62151
3 26.81839 22.56577 31.07102
4 30.06992 26.40168 33.73816
5 33.32144 29.49521 37.14768
6 34.94721 30.77443 39.11998
7 37.11489 32.28079 41.94899
8 38.19873 32.97230 43.42517
9 38.74065 33.30638 44.17493
```

predict.lm(model, interval="prediction")

```
      fit       lwr      upr
1 18.68958  5.502619 31.87654
2 21.39918  8.848307 33.95006
3 26.81839 15.118306 38.51848
4 30.06992 18.569345 41.57049
5 33.32144 21.769504 44.87338
6 34.94721 23.275906 46.61851
```

7 37.11489 25.191142 49.03864
8 38.19873 26.110604 50.28686
9 38.74065 26.561220 50.92009

- newx=data.frame(age=40) #create a new data frame with one new x* value of 40
- predict.lm(regmodel, newx, interval="confidence") #get a CI for the mean at the value x*

newx=data.frame(age=40) #create a new data frame with one new x* value of 40
predict.lm(model, newx, interval="confidence")

```
       fit      lwr      upr
1 27.90223 23.91526 31.88921
```