San José State University

Math 261A: Regression Theory & Methods

Simple Linear Regression

Dr. Guangliang Chen

This lecture is based on the following textbook sections:

- **Chapter 2: 2.1 - 2.6**

Outline of this presentation:

- The simple linear regression problem
- Least-square estimation
- Inference

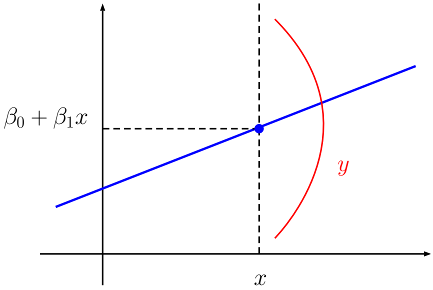
The simple linear regression problem

Consider the following (population) regression model

$$y = \beta_0 + \beta_1 x + \epsilon$$

where

- x: predictor (fixed)
- y: response (random)
- ϵ: random error/noise



β_0: intercept, β_1: slope

Sample regression model

Given a set of locations $x_1, \ldots, x_n$, let the corresponding responses be

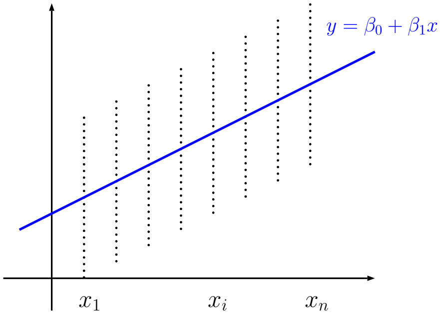
$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \; i = 1, \ldots, n$$

where the errors ϵ_i have mean 0 and variance σ^2:

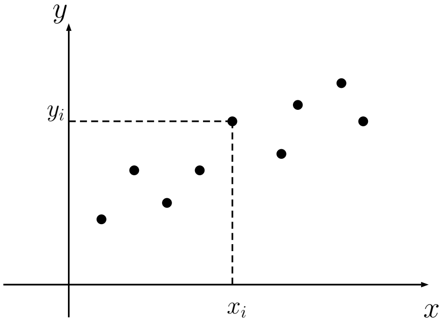
$$\mathrm{E}(\epsilon_i) = 0, \quad \mathrm{Var}(\epsilon_i) = \sigma^2,$$

and additionally are uncorrelated:

$$\mathrm{Cov}(\epsilon_i, \epsilon_j) = 0, \; i \neq j$$



In those same locations, let the observations of the responses also be $y_1, \ldots, y_n$ (this is an abuse of notation) such that we have a data set $\{(x_i, y_i) \mid 1 \le i \le n\}$.



The goal is to use the sample to estimate β_0, β_1 in some way (so as to fit a line to the data) .

Remark. Depending on the context, the notation y_i can denote either a random variable, or an observed value of it.

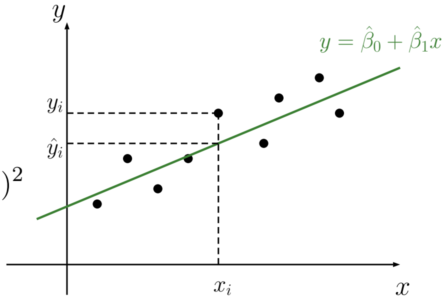
Least-squares (LS) estimation

To estimate the regression coefficients β_0, β_1, here we adopt the least squares criterion:

$$\min_{\hat{\beta}_0, \hat{\beta}_1} S(\hat{\beta}_0, \hat{\beta}_1) \stackrel{\text{def}}{=} \sum_{i=1}^{n} (y_i - (\underbrace{\hat{\beta}_0 + \hat{\beta}_1 x_i}_{\hat{y}_i}))^2$$

The corresponding minimizers are called **least squares estimators**.

Remark. Another way is to maximize the likelihood of the sample (Sec 2.11).



y_i: observation, $\hat{y}_i$: fitted value

Notation: To solve the problem, we need to define some quantities first:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

and

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$$
$$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

It can be shown that

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2,$$
$$S_{xy} = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}$$

Verify:

Theorem 0.1. The LS estimators of the intercept and slope in the simple linear regression model are

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \quad \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

Proof. Taking partial derivatives of

$$S(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

and setting them to zero gives that

$$\frac{\partial S}{\partial \hat{\beta}_0} = -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$
$$\frac{\partial S}{\partial \hat{\beta}_1} = -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)\, x_i = 0$$

which can then be simplified to

$$\sum y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum x_i$$
$$\sum x_i y_i = \hat{\beta}_0 \sum x_i + \hat{\beta}_1 \sum x_i^2$$

The first equation can be rewritten as

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

from which we obtain that

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Plugging it into the second equation yields that

$$\sum x_i y_i = (\bar{y} - \hat{\beta}_1 \bar{x})\, n\bar{x} + \hat{\beta}_1 \sum x_i^2$$

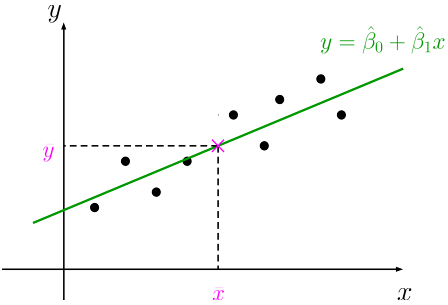
and further that

$$\underbrace{\sum x_i y_i - n\bar{x}\bar{y}}_{S_{xy}} = \hat{\beta}_1 \underbrace{\left(\sum x_i^2 - n\bar{x}^2\right)}_{S_{xx}}$$

This thus completes the proof. $\square$

Remark. We make the following observations:

- The LS regression line always passes through the centroid $(\bar{x}, \bar{y})$ of the data: $\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$.



- Alternative forms of the equation of the LS regression line are

$$y = \underbrace{(\bar{y} - \hat{\beta}_1 \bar{x})}_{\hat{\beta}_0} + \hat{\beta}_1 x = \bar{y} + \hat{\beta}_1 (x - \bar{x})$$

To study the effect of different samples on the regression coefficients, we regard the y_i as random variables (in this case $\bar{y}, \hat{\beta}_0, \hat{\beta}_1$ are also random variables). It can be shown that (homework problem: 2.25)

$$\mathrm{Cov}\left(\bar{y}, \hat{\beta}_1\right) = 0, \quad \mathrm{Cov}\left(\hat{\beta}_0, \hat{\beta}_1\right) = -\sigma^2 \frac{\bar{x}}{S_{xx}}$$

That is, $\bar{y}, \hat{\beta}_1$ are uncorrelated, but $\hat{\beta}_0, \hat{\beta}_1$ are not.

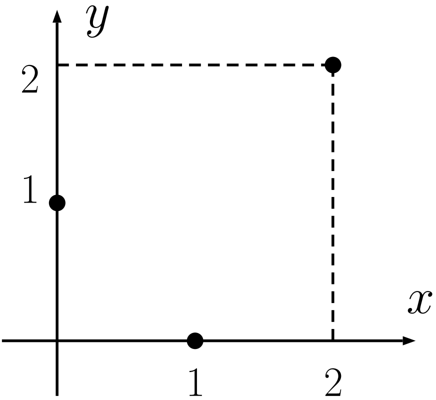
- The residuals of the model are

$$e_i = y_i - \hat{y}_i = y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right) = y_i - \left(\bar{y} + \hat{\beta}_1 (x_i - \bar{x})\right).$$

- $\sum e_i = 0$. This implies that $\sum y_i = \sum \hat{y}_i$, and thus $\{\hat{y}_i\}$ and $\{y_i\}$ have the same mean.
  Proof:

- $\sum x_i e_i = 0$, and $\sum \hat{y}_i e_i = 0$
  Proof:

Example 0.1 (Toy data)**.** Given a data set of 3 points: $(0,1),(1,0),(2,2)$, find the least-squares regression line.



Solution. First, $\bar{x} = 1 = \bar{y}$, and

$$S_{xx} = \sum x_i^2 - n\bar{x}^2 = 5 - 3 = 2, \quad S_{xy} = \sum x_i y_i - n\bar{x}\bar{y} = 4 - 3 = 1.$$

It follows that

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{1}{2}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \frac{1}{2}.$$

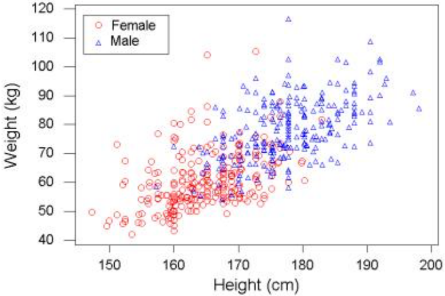
Thus, the regression line is given by

$$y = \hat{\beta}_0 + \hat{\beta}_1 x = \frac{1}{2} + \frac{1}{2}x.$$

The fitted values and their residuals are

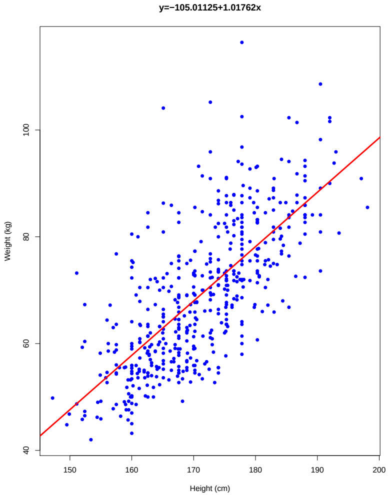
$$\hat{y}_1 = \frac{1}{2},\ \hat{y}_2 = 1,\ \hat{y}_3 = \frac{3}{2} \quad \text{and } e_1 = \frac{1}{2},\ e_2 = -1,\ e_3 = \frac{1}{2}$$

Example 0.2 (R demonstration)**.** Consider the dataset that contains weights and heights of 507 physically active individuals (247 men and 260 women).[1] We fit a regression line of weight (y) versus height (x) by R.



[1] http://jse.amstat.org/v11n2/datasets.heinz.html

Simple Linear Regression




```
> # linear regression (mydata is a data frame)
> mymodel<-lm(weight~height, data=mydata )
> summary(mymodel)

Call:
lm(formula = weight ~ height, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max
-18.743  -6.402  -1.231   5.059  41.103

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -105.01125    7.53941  -13.93   <2e-16 ***
height         1.01762    0.04399   23.14   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.308 on 505 degrees of freedom
Multiple R-squared:  0.5145,    Adjusted R-squared:  0.5136
F-statistic: 535.2 on 1 and 505 DF,  p-value: < 2.2e-16

> # plot the regression line on top of data
> plot(mydata$height, mydata$weight,
+      xlab="Height (cm)", ylab="Weight (kg)",
+      pch=16, col="blue",
+      main="y=-105.01125+1.01762x")
> abline(mymodel, col="red",lwd=3)
```

Inference in simple linear regression

- **Model parameters**: β_0 (intercept), β_1 (slope), σ^2 (noise variance)
- **Inference tasks** (for each parameter above): point estimation, interval estimation*, hypothesis testing*
- **Inference of the mean response** at any location x_0:

$$\mathrm{E}(y \mid x_0) = \beta_0 + \beta_1 x_0$$

**To perform the last two inference tasks, we will additionally assume that the model errors ϵ_i are normally and independently distributed with mean 0 and variance σ^2, i.e., $\epsilon_1, \ldots, \epsilon_n \overset{iid}{\sim} N(0, \sigma^2)$.*

Point estimation in regression

Theorem 0.2. The LS estimators $\hat{\beta}_0, \hat{\beta}_1$ are unbiased linear estimators of the model parameters β_0, β_1, that is,

$$\mathrm{E}(\hat{\beta}_0) = \beta_0, \quad \mathrm{E}(\hat{\beta}_1) = \beta_1$$

Furthermore,

$$\mathrm{Var}(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right), \quad \mathrm{Var}(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$$

Remark. The Gauss-Markov Theorem stats that the LS estimators $\hat{\beta}_0, \hat{\beta}_1$ are the best linear unbiased estimators in that they have the smallest possible variance (among all linear unbiased estimators of β_0, β_1).

Proof. Write

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum(x_i - \bar{x})y_i}{S_{xx}} = \sum c_i y_i, \quad c_i = \frac{x_i - \bar{x}}{S_{xx}}$$

It follows that

$$\mathrm{E}(\hat{\beta}_1) = \sum c_i \mathrm{E}(y_i) = \sum c_i(\beta_0 + \beta_1 x_i) = \beta_0 \underbrace{\sum c_i}_{=0} + \beta_1 \underbrace{\sum c_i x_i}_{=1} = \beta_1$$

and

$$\mathrm{Var}(\hat{\beta}_1) = \sum c_i^2 \underbrace{\mathrm{Var}(y_i)}_{=\sigma^2} = \sigma^2 \sum c_i^2 = \sigma^2 \cdot \frac{1}{S_{xx}} = \frac{\sigma^2}{S_{xx}}$$

For $\hat{\beta}_0$, it is unbiased for estimating β_0 because

$$\mathrm{E}(\hat{\beta}_0) = \mathrm{E}(\bar{y} - \hat{\beta}_1 \bar{x}) = \mathrm{E}(\bar{y}) - \mathrm{E}(\hat{\beta}_1)\bar{x} = (\beta_0 + \beta_1 \bar{x}) - \beta_1 \bar{x} = \beta_0.$$

Using the formula

$$\mathrm{Var}(X - Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) - 2\mathrm{Cov}(X, Y),$$

we obtain that

$$\begin{aligned}
\mathrm{Var}(\hat{\beta}_0) &= \mathrm{Var}(\bar{y}) + \mathrm{Var}(\hat{\beta}_1 \bar{x}) - 2\,\mathrm{Cov}(\bar{y}, \hat{\beta}_1 \bar{x}) \\
&= \frac{1}{n^2} \sum \mathrm{Var}(y_i) + \bar{x}^2\,\mathrm{Var}(\hat{\beta}_1) - 2\bar{x}\,\underbrace{\mathrm{Cov}(\bar{y}, \hat{\beta}_1)}_{=0} \\
&= \frac{1}{n^2}\, n\sigma^2 + \bar{x}^2\,\frac{\sigma^2}{S_{xx}} = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right).
\end{aligned}$$

To estimate the noise variance σ^2, we need to define

- **Total Sum of Squares**

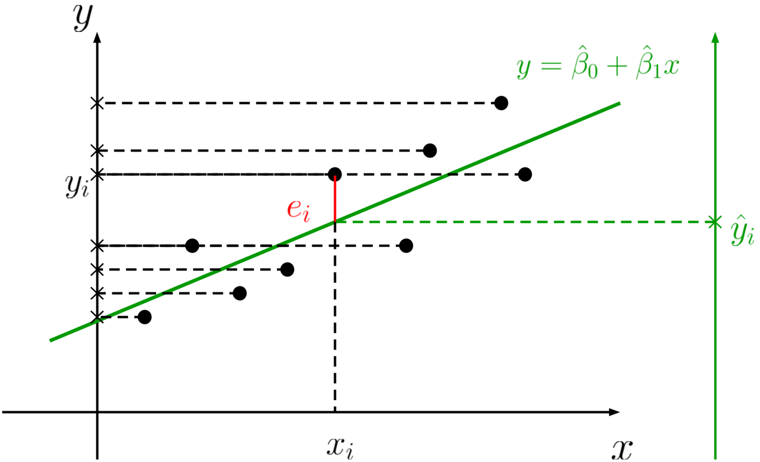
$$SS_T = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

- **Regression Sum of Squares**

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

- **Residual Sum of Squares**

$$SS_{Res} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$


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

It can be shown that

$$SS_T = SS_R + SS_{Res}$$

Proof:

$$\begin{aligned}
SS_T &= \sum (y_i - \bar{y})^2 \\
&= \sum (y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2 \\
&= \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \bar{y})^2 + 2 \sum (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) \\
&= SS_{Res} + SS_R + 2 \underbrace{\sum e_i \hat{y}_i}_{=0}
\end{aligned}$$

Another useful result is

$$SS_R = \hat{\beta}_1^2 S_{xx}$$

Proof.

$$\begin{aligned}
SS_R &= \sum (\hat{y}_i - \bar{y})^2 \\
&= \sum \Big(\underbrace{(\hat{\beta}_0 + \hat{\beta}_1 x_i)}_{\hat{y}_i} - \underbrace{(\hat{\beta}_0 + \hat{\beta}_1 \bar{x})}_{\bar{y}}\Big)^2 \\
&= \sum \hat{\beta}_1^2 (x_i - \bar{x})^2 \\
&= \hat{\beta}_1^2 \, S_{xx}.
\end{aligned}$$

The following theorem indicates how to use the residual sum of squares to estimate the error variance σ^2 when it is unknown.

Theorem 0.3. We have

$$\mathrm{E}(SS_{Res}) = (n-2)\sigma^2$$

This implies that the residual mean square

$$MS_{Res} = \frac{SS_{Res}}{n-2}$$

is an unbiased estimator for σ^2.

Proof. Write

$$SS_{Res} = SS_T - SS_R = \left(\sum y_i^2 - n\bar{y}^2\right) - \hat{\beta}_1^2\, S_{xx}$$

Using the formula $\mathrm{E}(X^2) = \mathrm{E}(X)^2 + \mathrm{Var}(X)$, we have

$$\begin{aligned}
&\mathrm{E}(SS_{Res}) = \sum \mathrm{E}\left(y_i^2\right) - n\,\mathrm{E}\left(\bar{y}^2\right) - \mathrm{E}\left(\hat{\beta}_1^2\right)\, S_{xx} \\
&= \sum \left[(\beta_0 + \beta_1 x_i)^2 + \sigma^2\right] - n\left[(\beta_0 + \beta_1 \bar{x})^2 + \frac{\sigma^2}{n}\right] - \left(\beta_1^2 + \frac{\sigma^2}{S_{xx}}\right) S_{xx} \\
&= (n-2)\sigma^2.
\end{aligned}$$

This implies that $\mathrm{E}(MS_{Res}) = \mathrm{E}(SS_{Res})/(n-2) = \sigma^2$. $\square$

Another way to use the sums of squares is to define a measure of the **goodness of fit** of the regression line.

Def 0.1 (Coefficient of determination)**.**

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_{Res}}{SS_T}$$

Remark. The quantity $0 \leq R^2 \leq 1$ indicates the proportion of variation of the response that is explained by the regression line.

Example 0.3 (Toy data)**.** Consider again the toy data set that consists of 3 points: $(0,1),(1,0),(2,2)$. We have fitted the LS regression line earlier.

It is straightforward to obtain that

$$SS_{Res} = \sum e_i^2 = \left(\frac{1}{2}\right)^2 + (-1)^2 + \left(\frac{1}{2}\right)^2 = \frac{3}{2}.$$

Accordingly, a point estimate of σ^2 is

$$MS_{Res} = SS_{Res}/(n-2) = 1.5$$

To compute the coefficient of determination, we also need to compute $SS_T = \sum (y_i - \bar{y})^2 = 2$. It follows that

$$R^2 = 1 - \frac{SS_{Res}}{SS_T} = 1 - \frac{1.5}{2} = 0.25$$

Example 0.4 (weight-height). From the R output:

- The residual standard error is $\hat{\sigma} = 9.308$;
- The residual mean square is $MS_{Res} = 9.308^2 = 86.639$.
- The coefficient of determination is $R^2 = 0.5145$ (meaning that the LS regression line only captures 51.45% of the total variation).

```
> # linear regression (mydata is a data frame)
> mymodel<-lm(weight~height, data=mydata )
> summary(mymodel)

Call:
lm(formula = weight ~ height, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max 
-18.743  -6.402  -1.231   5.059  41.103 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -105.01125    7.53941  -13.93   <2e-16 ***
height         1.01762    0.04399   23.14   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’

Residual standard error: 9.308 on 505 degrees of freedom
Multiple R-squared:  0.5145,	Adjusted R-squared:  0.5136 
F-statistic: 535.2 on 1 and 505 DF,  p-value: < 2.2e-16

> # plot the regression line on top of data
> plot(mydata$height, mydata$weight,
+      xlab="Height (cm)", ylab="Weight (kg)",
+      pch=16, col="blue",
+      main="y=-105.01125+1.01762x")
> abline(mymodel, col="red",lwd=3)
```

Summary: Point estimation in simple linear regression

Model parameters	Point estimators	Properties: Bias	Properties: Variance
β_0	$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$	unbiased	$\sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right)$
β_1	$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$	unbiased	$\frac{\sigma^2}{S_{xx}}$
σ^2	$MS_{Res} = \frac{SS_{Res}}{n-2}$	unbiased	

Remark. For the mean response at x_0:

$$\mathrm{E}(y \mid x_0) = \beta_0 + \beta_1 x_0,$$

it is easy to see that $\hat{\beta}_0 + \hat{\beta}_1 x_0$ is an unbiased point estimator.

Next

We consider the following inference tasks in regression:

- **Hypothesis testing**

- **Interval estimation**

The χ^2, t and F distributions

First, we need to review/introduce the following distributions:

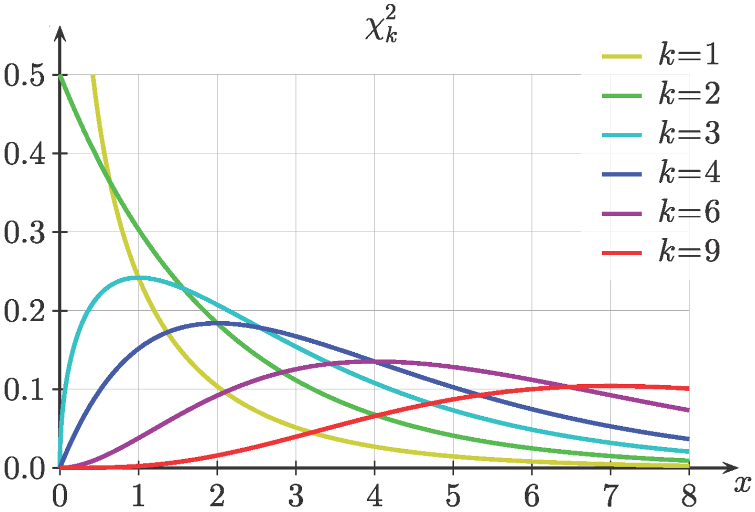
- χ^2
- t
- F

The χ^2 distribution

χ^2 is a special instance of Gamma: $\chi^2_k = \text{Gamma}(\alpha = \frac{k}{2}, \lambda = \frac{1}{2})$, where k is a positive integer and commonly referred to as the *degree of freedom* of the distribution. It can be shown that χ^2_k is the distribution of $X = Z_1^2 + \cdots + Z_k^2$ for $Z_1, \ldots, Z_k \overset{iid}{\sim} N(0,1)$.

Below are some known results about $X \sim \chi^2_k$ (inferred from Gamma):

- Density: $f(x) = \frac{1}{2^{k/2}\,\Gamma(k/2)} \left(\frac{x}{2}\right)^{\frac{k}{2}-1} e^{-\frac{x}{2}},\ x > 0$
- Properties: $\mathrm{E}(X) = k,\ \mathrm{Var}(X) = 2k$


χ_k^2
0.5
0.4
0.3
0.2
0.1
0.0
0
1
2
3
4
5
6
7
8
x
$k=1$
$k=2$
$k=3$
$k=4$
$k=6$
$k=9$

Student's t distribution

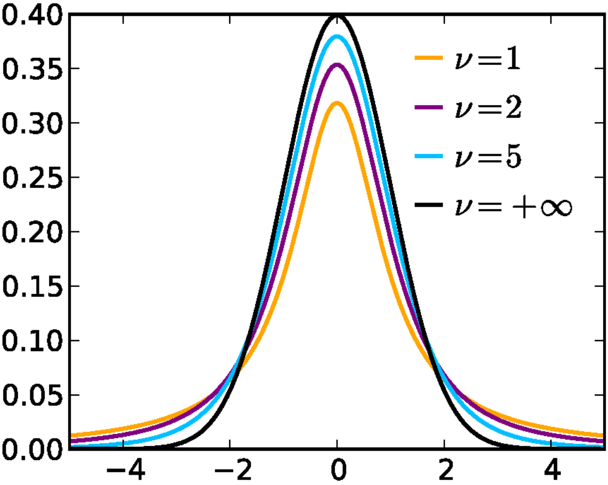
This is the distribution of a random variable of the form

$$T = \frac{Z}{\sqrt{X/\nu}}, \quad \text{where } Z \sim N(0,1), X \sim \chi^2_\nu \text{ are independent.}$$

Similarly, ν is referred to as the *degree of freedom* of the t distribution.

Density curves of the t-family are all unimodal, symmetric and bell-shaped, like those of the normal distributions. Below are some results about $T \sim t(\nu)$:

- Density: $f(x) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}, \ -\infty < x < \infty$
- Properties: $\mathrm{E}(T) = 0, \ \mathrm{Var}(T) = \frac{\nu}{\nu-2}$ (when $\nu > 2$).


0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
−4
−2
0
2
4
$\nu=1$
$\nu=2$
$\nu=5$
$\nu=+\infty$

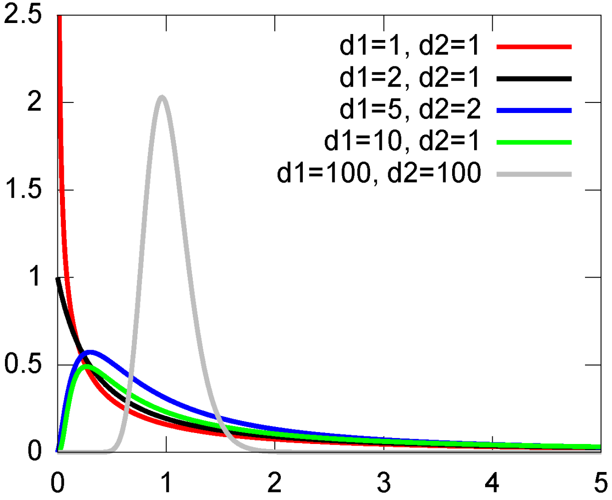
Snedecor's F distribution

This is the distribution of a random variable of the form

$$X = \frac{X_1/d_1}{X_2/d_2}, \quad \text{where } X_1 \sim \chi^2_{d_1}, X_2 \sim \chi^2_{d_2} \text{ are independent.}$$

What we know about $X \sim \mathrm{F}(d_1, d_2)$:

- Density: $f_X(x) = \frac{1}{B(\frac{d_1}{2},\frac{d_2}{2})} \left(\frac{d_1}{d_2}\right)^{\frac{d_1}{2}} x^{\frac{d_1}{2}-1} \left(1 + \frac{d_1}{d_2}x\right)^{-\frac{d_1+d_2}{2}}, \ x > 0$
- $\mathrm{E}(X) = \frac{d_2}{d_2-2}$ (if $d_2 > 2$), and $\mathrm{Var}(X) = \frac{2d_2^2(d_1+d_2-2)}{d_1(d_2-2)^2(d_2-4)}$ (if $d_2 > 4$)


d1=1, d2=1
d1=2, d2=1
d1=5, d2=2
d1=10, d2=1
d1=100, d2=100
2.5
2
1.5
1
0.5
0
0
1
2
3
4
5

Additional normality assumption on the errors

To perform the hypothesis testing and interval estimation tasks in regression, we need to assume additionally that the errors ϵ_i are iid $N(0, \sigma^2)$. This implies that

$$y_i \sim \mathrm{N}(\beta_0 + \beta_1 x_i,\, \sigma^2),\ i = 1, \ldots, n$$

and they are independent (but not identically distributed).

Since $\hat{\beta}_1$ is a linear combination of the random variables y_i, under the additional assumption we have

$$\hat{\beta}_1 \sim \mathrm{N}\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right).$$

Hypothesis testing in regression

Consider first the following hypothesis test about the slope parameter:

$$H_0 : \beta_1 = \beta_{10}, \qquad \text{vs} \quad H_1 : \beta_1 \neq \beta_{10}$$

where β_{10} represents a particular value (e.g., 0) that β_1 might take.

Under the normality assumption on the errors, we have the following result.

Theorem 0.4. At level α, a rejection region of the above test is

$$\begin{cases} \frac{\left|\hat{\beta}_1 - \beta_{10}\right|}{\sqrt{\sigma^2/S_{xx}}} > z_{\alpha/2}, & \text{if } \sigma^2 \text{ known;} \\ \frac{\left|\hat{\beta}_1 - \beta_{10}\right|}{\sqrt{MS_{Res}/S_{xx}}} > t_{\alpha/2,n-2}, & \text{if } \sigma^2 \text{ unknown.} \end{cases}$$

Proof. When H_0 is true, the distribution of $\hat{\beta}_1$ is

$$\hat{\beta}_1 \sim \mathrm{N}\left(\beta_{10}, \frac{\sigma^2}{S_{xx}}\right).$$

Therefore, we can write down the following decision rule (at level α):

$$\frac{\left|\hat{\beta}_1 - \beta_{10}\right|}{\sqrt{\sigma^2/S_{xx}}} > z_{\alpha/2}$$

When σ^2 is unknown, we need to use its estimator MS_{Res} instead. This leads to a t test:

$$\frac{\left|\hat{\beta}_1 - \beta_{10}\right|}{\sqrt{MS_{Res}/S_{xx}}} > t_{\alpha/2,n-2}$$

Remark. $\sqrt{\sigma^2/S_{xx}}$ is the standard deviation of $\hat{\beta}_1$, while $\sqrt{MS_{Res}/S_{xx}}$ is called the standard error of $\hat{\beta}_1$:

$$\mathrm{Std}(\hat{\beta}_1) = \sqrt{\sigma^2/S_{xx}}, \qquad se(\hat{\beta}_1) = \sqrt{MS_{Res}/S_{xx}}.$$

Depending on whether σ^2 is given, the test statistic needed is

$$Z_0 = \frac{\hat{\beta}_1 - \beta_{10}}{\mathrm{Std}(\hat{\beta}_1)} \ (\sigma^2 \text{ known}), \qquad t_0 = \frac{\hat{\beta}_1 - \beta_{10}}{se(\hat{\beta}_1)} \ (\sigma^2 \text{ unknown})$$

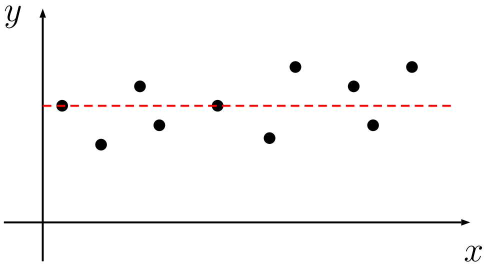
with corresponding decision rule:

$$|Z_0| > z_{\alpha/2} \ (\sigma^2 \text{ known}), \qquad |t_0| > t_{\alpha/2,n-2} \ (\sigma^2 \text{ unknown})$$

Remark. An important special case of the above hypothesis test is when $\beta_{10} = 0$, which concerns the **significance of regression**:

$$H_0 : \beta_1 = 0 \text{ (There is no linear relationship between } y \text{ and } x)$$

$$H_1 : \beta_1 \neq 0 \text{ (There is a linear relationship between } y \text{ and } x)$$



Example 0.5 (weight-height). From the R output, we see that

- The value of the t statistic for testing $H_0 : \beta_1 \neq 0$ against $H_0 : \beta_1 = 0$ is $t_0 = 23.14$;
- The p-value of the test is less than $2e$-16.

Thus, we can reject H_0 (at level 1%) and correspondingly conclude that there is a significant linear relationship between x and y.

```
> # linear regression (mydata is a data frame)
> mymodel<-lm(weight~height, data=mydata )
> summary(mymodel)

Call:
lm(formula = weight ~ height, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max
-18.743  -6.402  -1.231   5.059  41.103

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -105.01125    7.53941  -13.93   <2e-16 ***
height         1.01762    0.04399   23.14   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '

Residual standard error: 9.308 on 505 degrees of freedom
Multiple R-squared:  0.5145,    Adjusted R-squared:  0.5136
F-statistic: 535.2 on 1 and 505 DF,  p-value: < 2.2e-16

> # plot the regression line on top of data
> plot(mydata$height, mydata$weight,
+      xlab="Height (cm)", ylab="Weight (kg)",
+      pch=16, col="blue",
+      main="y=-105.01125+1.01762x")
> abline(mymodel, col="red",lwd=3)
```

Another approach to testing the significance of regression is through the **Analysis of Variance (ANOVA)**:

$$SS_T = SS_R + SS_{Res}, \quad \text{with d.o.f.:} \quad n-1 = 1 + (n-2)$$

We have previously defined the residual mean square

$$MS_{Res} = \frac{SS_{Res}}{n-2} \qquad \text{with} \quad \mathrm{E}(MS_{Res}) = \sigma^2$$

Define also the regression mean square

$$MS_R = SS_R/1.$$

It can be shown that

$$\mathrm{E}(MS_R) = \sigma^2 + \beta_1^2 S_{xx}$$

Observation: MS_R contains information about β_1.

- $\mathrm{E}(MS_R) = \mathrm{E}(MS_{Res})$ if $\beta_1 = 0$;
- $\mathrm{E}(MS_R) > \mathrm{E}(MS_{Res})$ if $\beta_1 \neq 0$.

As a result, large values of their ratio

$$F_0 = \frac{MS_R}{MS_{Res}} = \frac{SS_R/1}{SS_{Res}/(n-2)} \qquad \left(\overset{H_0 \text{ true}}{\sim} F_{1,n-2}\right)$$

are evidence against $H_0 : \beta_1 = 0$.

Therefore, we have the following significance of regression test:

$$\boxed{\text{Reject } H_0 : \beta_1 = 0 \text{ if and only if } F_0 > F_{\alpha,1,n-2}}$$

The ANOVA procedure is summarized in the following able.

Source of variation	Sum of squares	Degrees of freedom	Mean square	Test statistic
Regression	$SS_R = \hat{\beta}_1^2 S_{xx}$	1	MS_R	$F_0 = \frac{MS_R}{MS_{Res}}$
Residual	SS_{Res}	$n-2$	MS_{Res}	
Total	SS_T	$n-1$		

Example 0.6 (weight-height). From the R output, we see that

- The F statistic for testing $H_0 : \beta_1 = 0$ against a two-sided alternative is $F_0 = 535.2$ with 1 and 505 degrees of freedom;
- The p-value of the test is less than $2.2e$-16.

Thus, we can conclude that $\beta_1 \neq 0$, i.e., there is a significant linear relationship between x and y.

```
> # linear regression (mydata is a data frame)
> mymodel<-lm(weight~height, data=mydata )
> summary(mymodel)

Call:
lm(formula = weight ~ height, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max
-18.743  -6.402  -1.231   5.059  41.103

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -105.01125    7.53941  -13.93   <2e-16 ***
height         1.01762    0.04399   23.14   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '

Residual standard error: 9.308 on 505 degrees of freedom
Multiple R-squared:  0.5145,	Adjusted R-squared:  0.5136
F-statistic: 535.2 on 1 and 505 DF,  p-value: < 2.2e-16

> # plot the regression line on top of data
> plot(mydata$height, mydata$weight,
+      xlab="Height (cm)", ylab="Weight (kg)",
+      pch=16, col="blue",
+      main="y=-105.01125+1.01762x")
> abline(mymodel, col="red",lwd=3)
```

A more direct way of performing ANOVA in R is to use the `anova` function:

```
> anova(mymodel)
Analysis of Variance Table

Response: weight
           Df Sum Sq Mean Sq F value    Pr(>F)
height      1  46370   46370  535.21 < 2.2e-16 ***
Residuals 505  43753      87
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Remark. The ANOVA F test is equivalent to the (two-sided) t test regarding whether $\beta_1 = 0$ or not:

$$t_0^2 = \frac{\hat{\beta}_1^2}{MS_{Res}/S_{xx}} = \frac{\hat{\beta}_1^2 S_{xx}}{MS_{Res}} = \frac{SS_R/1}{SS_{Res}/(n-2)} = F_0$$

However, when one-sided alternatives such as

$$H_0 : \beta_1 = 0 \quad \text{vs} \quad H_1 : \beta_1 > 0$$

are used, only the t test can be used:

$$t_0 = \frac{\hat{\beta}_1 - 0}{\sqrt{MS_{Res}/S_{xx}}} > t_{\alpha,n-2} \quad (\sigma^2 \text{ unknown})$$

For the hypothesis test about the intercept parameter β_0,

$$H_0: \ \beta_0 = \beta_{00}, \quad \text{vs} \quad H_1: \ \beta_0 \neq \beta_{00}$$

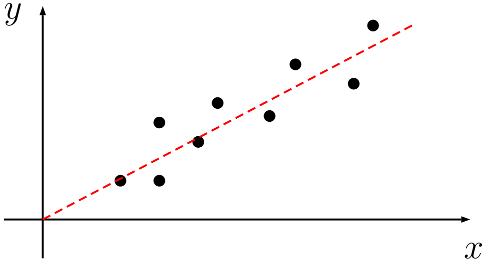
we have the following result.

Theorem 0.5. At level α, a rejection region of the test is

$$\begin{cases} \frac{\left|\hat{\beta}_0 - \beta_{00}\right|}{\sqrt{\sigma^2\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right)}} > z_{\alpha/2}, & \text{if } \sigma^2 \text{ known;} \\ \frac{\left|\hat{\beta}_0 - \beta_{00}\right|}{\sqrt{MS_{Res}\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right)}} > t_{\alpha/2,n-2}, & \text{if } \sigma^2 \text{ unknown;} \end{cases}$$

Remark. The previous R output also contains the results of the corresponding t-test for

$$H_0 : \beta_0 = 0 \text{ (The regression line passes through the origin)}$$
$$H_1 : \beta_0 \neq 0 \text{ (The regression line does not pass through the origin)}$$



Summary: hypothesis testing in regression

We covered the following tests with corresponding decision rules:

- $H_0 : \beta_1 = \beta_{10}$ vs $H_1 : \beta_1 \neq \beta_{10}$: $\quad \frac{\left|\hat{\beta}_1 - \beta_{10}\right|}{\sqrt{MS_{Res}/S_{xx}}} > t_{\alpha/2,n-2}$
- Significance of regression test ($H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$)
  - t-test: $\frac{\left|\hat{\beta}_1\right|}{\sqrt{MS_{Res}/S_{xx}}} > t_{\alpha/2,n-2}$
  - ANOVA F-test: $\frac{MS_R}{MS_{Res}} > F_{\alpha,1,n-2}$
- $H_0 : \beta_0 = \beta_{00}$ vs $H_1 : \beta_0 \neq \beta_{00}$: $\quad \frac{\left|\hat{\beta}_0 - \beta_{00}\right|}{\sqrt{MS_{Res}(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}})}} > t_{\alpha/2,n-2}$

Interval estimation in regression

Under the normality assumptions, the $1-\alpha$ CIs for β_0, β_1 are

- $\hat{\beta}_0 \pm t_{\alpha/2,n-2}\sqrt{MS_{Res}\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right)}$
- $\hat{\beta}_1 \pm t_{\alpha/2,n-2}\sqrt{MS_{Res}/S_{xx}}$

This is implemented in R through the CONFINT function:

```
> confint(mymodel, level=0.95)
                   2.5 %     97.5 %
(Intercept) -119.8237251 -90.198783
height         0.9311971   1.104036
```

We next construct a $1-\alpha$ confidence interval for the noise variance σ^2.

Theorem 0.6. Under the normality assumptions, a level $1-\alpha$ confidence interval for σ^2 is

$$\left(\frac{(n-2)MS_{Res}}{\chi^2_{\frac{\alpha}{2},n-2}}, \frac{(n-2)MS_{Res}}{\chi^2_{1-\frac{\alpha}{2},n-2}} \right)$$

Proof. It can be shown that

$$\frac{SS_{Res}}{\sigma^2} = \frac{(n-2)MS_{Res}}{\sigma^2} \sim \chi^2_{n-2}.$$

Thus,

$$1-\alpha = P\left(\chi^2_{1-\frac{\alpha}{2},n-2} < \frac{(n-2)MS_{Res}}{\sigma^2} < \chi^2_{\frac{\alpha}{2},n-2} \right).$$

Solving the inequalities for σ^2 yields the desired result.

Example 0.7 (weight-height). A 95% confidence interval for σ^2 is

$$\left(\frac{505MS_{Res}}{\chi^2_{.025,\,505}}, \frac{505MS_{Res}}{\chi^2_{.975,505}}\right) = \left(\frac{505 \cdot 9.308^2}{569.1608}, \frac{505 \cdot 9.308^2}{444.6268}\right) = (76.87,\, 98.40)$$

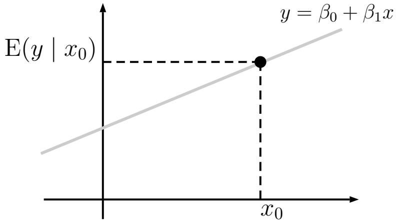
R commands:

```
> qchisq(.975, 505)
[1] 569.1608
> pchisq(569.1608,505)
[1] 0.9750001
> qchisq(.025, 505)
[1] 444.6268
```

The mean response

A major use of a regression model is to estimate the mean response at a particular location $x = x_0$

$$\mathrm{E}(y \mid x_0) = \beta_0 + \beta_1 x_0$$



Under the normality assumption, we obtain the following result.

Theorem 0.7. A $1-\alpha$ confidence interval for $\mathrm{E}(y \mid x_0)$ is

$$(\hat{\beta}_0 + \hat{\beta}_1 x_0) \pm t_{\alpha/2,n-2}\sqrt{MS_{Res}\left(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right)}$$

Proof. The point estimator of the mean response, $\hat\beta_0 + \hat\beta_1 x_0$, is a linear combination of the responses y_i, thus having a normal distribution with mean

$$\mathrm{E}(\hat\beta_0 + \hat\beta_1 x_0) = \beta_0 + \beta_1 x_0$$

and variance

$$\begin{aligned}
\mathrm{Var}(\hat\beta_0 + \hat\beta_1 x_0) &= \mathrm{Var}\left(\bar{y} + \hat\beta_1 (x_0 - \bar{x})\right) \\
&= \mathrm{Var}\left(\bar{y}\right) + \mathrm{Var}\left(\hat\beta_1\right)(x_0 - \bar{x})^2 \\
&= \frac{\sigma^2}{n} + \frac{\sigma^2}{S_{xx}}(x_0 - \bar{x})^2 \\
&= \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)
\end{aligned}$$

It follows that

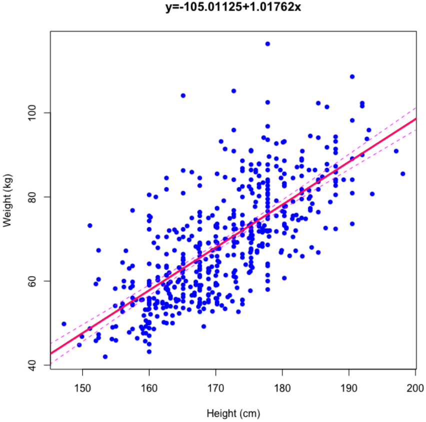
$$\frac{(\hat{\beta}_0 + \hat{\beta}_1 x_0) - (\beta_0 + \beta_1 x_0)}{\sqrt{MS_{Res}\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}} \sim t_{n-2}$$

and consequently we can use the following equality

$$1 - \alpha = P\left(-t_{\frac{\alpha}{2}, n-2} < \frac{(\hat{\beta}_0 + \hat{\beta}_1 x_0) - (\beta_0 + \beta_1 x_0)}{\sqrt{MS_{Res}\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}} < t_{\frac{\alpha}{2}, n-2}\right)$$

to construct a level $1 - \alpha$ confidence interval for $\beta_0 + \beta_1 x_0$. $\square$

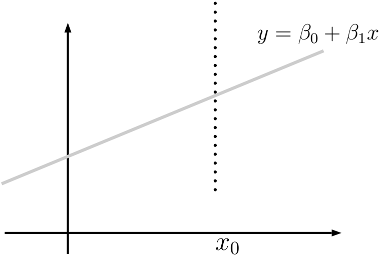
Remark. The confidence interval for the mean response is the shortest at the location $x_0 = \bar{x}$ and becomes wider as x moves away from $\bar{x}$ in either direction.



Prediction of new observations

Another way of using a regression model is to develop a **prediction interval** for the future observation at some specified location $x = x_0$:

$$y_0 = \beta_0 + \beta_1 x_0 + \epsilon_0, \quad \epsilon_0 \sim \mathrm{N}(0, \sigma^2)$$



Theorem 0.8. A $1-\alpha$ prediction interval for the response y_0 at $x = x_0$ is

$$(\hat{\beta}_0 + \hat{\beta}_1 x_0) \pm t_{\alpha/2,n-2}\sqrt{MS_{Res}\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}$$

Proof. First, note that a point estimator for the fixed component of y_0 (i.e., $\beta_0 + \beta_1 x_0$) is

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

Let $\Psi = y_0 - \hat{y}_0$ be the difference between the true response and the point estimator for its fixed part. Then Ψ (as a linear combination of $y_0, y_1, \ldots, y_n$) is normally distributed with mean

$$\Psi = \mathrm{E}(y_0) - \mathrm{E}(\hat{y}_0) = (\beta_0 + \beta_1 x_0) - (\beta_0 + \beta_1 x_0) = 0$$

and variance

$$\mathrm{Var}(\Psi) = \mathrm{Var}(y_0) + \mathrm{Var}(\hat{y}_0) = \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)$$

We then have

$$\frac{y_0 - \hat{y}_0}{\sqrt{\sigma^2 \left(1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right)}} \sim N(0,1)$$

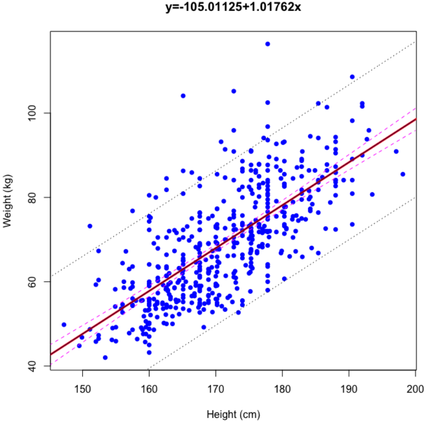
and correspondingly,

$$\frac{y_0 - \hat{y}_0}{\sqrt{MS_{Res} \left(1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right)}} \sim t_{n-2}$$

Accordingly, a $1-\alpha$ prediction interval on a future observation y_0 at x_0 is

$$(\hat{\beta}_0 + \hat{\beta}_1 x_0) \pm t_{\alpha/2,\, n-2} \sqrt{MS_{Res} \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}$$

Remark. The prediction interval for the response at all locations has a similar pattern to the confidence interval for the mean response but is much wider.



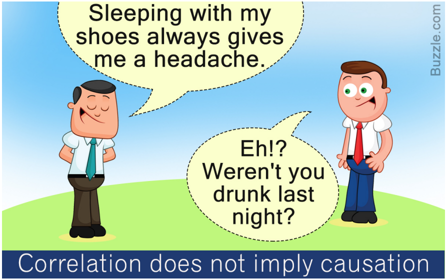
Summary: interval estimation in regression

- β_0 (intercept): $\hat{\beta}_0 \pm t_{\alpha/2,n-2}\sqrt{MS_{Res}\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right)}$
- β_1 (slope): $\hat{\beta}_1 \pm t_{\alpha/2,n-2}\sqrt{MS_{Res}/S_{xx}}$
- σ^2 (error variance): $\left(\frac{(n-2)MS_{Res}}{\chi^2_{\frac{\alpha}{2},n-2}}, \frac{(n-2)MS_{Res}}{\chi^2_{1-\frac{\alpha}{2},n-2}}\right)$
- $\mathrm{E}(y \mid x_0)$: $(\hat{\beta}_0+\hat{\beta}_1 x_0) \pm t_{\alpha/2,n-2}\sqrt{MS_{Res}\left(\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right)}$
- y_0 (response): $(\hat{\beta}_0+\hat{\beta}_1 x_0) \pm t_{\alpha/2,n-2}\sqrt{MS_{Res}\left(1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right)}$

Some considerations in the use of regression

Read Section 2.9 to understand the following issues (they will be covered in more depth later in this course):

- Extrapolation
- Influential points
- Outliers
- Correlation does not imply causation



Further learning

- 2.10 Regression Through the Origin
- 2.11 Maximum Likelihood Estimation
- 2.12 Case Where the Regressor x Is Random
- Linear regression via gradient descent
- Weighted least squares

$$S(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n} w_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$