Winter Term

V. Zinde-Walsh

Topic 4. Simple Linear Regression.

A relation between an explanatory variable and the dependent variable of interest:

A. May be given by a theoretical model, esp. in science.

Example. Boyle's law that relates the pressure and volume of an ideal gas at a constant temperature.

B. May be inferred from observation, e.g. it is observed that taller parents generally have taller children; IQ and other test scores generally predict better grades; GDP increases over the years, etc.

In either case the relation is not exactly observed, due to possible measurement errors, other unaccounted for factors, etc.

F. Galton in "Regression towards mediocrity in hereditary stature" (1886) observed that extreme characteristics (e.g., height) in parents are not passed on completely to their offspring. Rather, the characteristics in the offspring regress towards a "mediocre point"; we say regression towards the mean.

Two issues to examine: (a) the underlying theoretical probability model of the dependence between two variables and (b) what kind of a linear dependence can fit the data best. Then follows the issue of inference about the true relationship between the variables from the data.

1. The linear regression model (simple regression: one explanatory variable).

The **dependent** (stochastic) variable Y depends on the **explanatory variable** (predictor), X.

Assumption "average linear dependence". $E(Y|X = x) = f(x)$, where $f(x) = \beta_0 + \beta_1 x$.

Note that some non-linear cases can be approximated: $f(x) = f(x_0) + f'(x_0)(x - x_0) + R(x, x_0)$, where R is the remainder term; if R is small the linear approximation may not be bad. Some non-linear cases transform into linear: $Y = AX^{\beta}$, then $\ln Y = \ln A + \beta \ln X$.

From the assumption it follows that for any observed (X_i, Y_i) we can write

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,$$

where the error term ε_i has the property that $E\varepsilon_i|X_i = 0$.

The error term characterizes the deviations of individual randomly drawn observations X_i, Y_i from the ones that lie on the true regression line: $(X_i, E(Y_i|X_i))$.

The case of **non-random (non-stochastic, fixed)** X. Then the Assumption is $E(Y_i) = \beta_0 + \beta_1 X_i$, implying in the model $E(\varepsilon_i) = 0$.

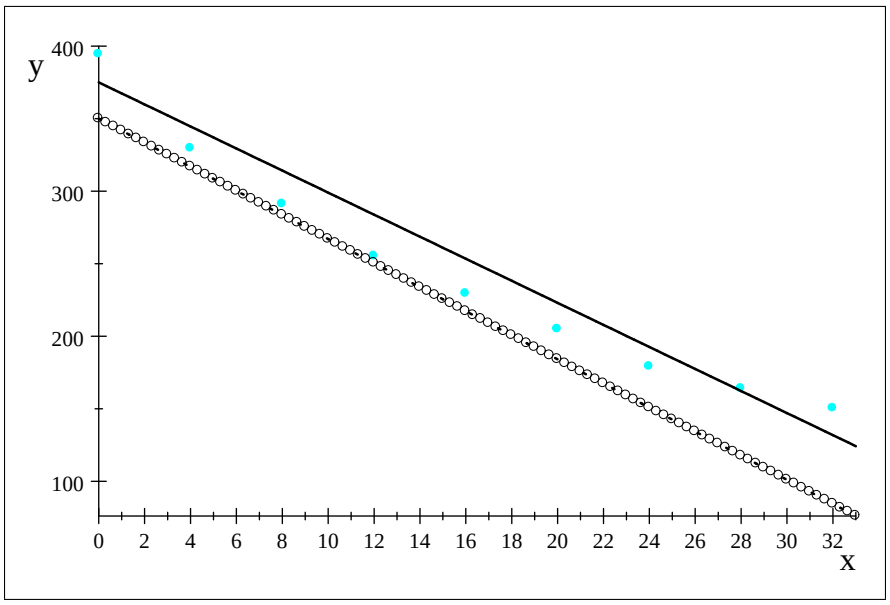
If the error ε_i comes from a normal distribution $N(0, \sigma^2)$, then Y_i is normal $N(\beta_0 + \beta_1 X_i, \sigma^2)$. Show.

2. Least squares estimator (OLS).

Gauss proposed the least squares criterion to fit the regression line to the data.

Example. Tire tread wear vs mileage. Lab test data.

Mileage(in 1000 miles)	Groove depth of tire (in .001 inches)
0	394.33
4	329.5
8	291.00
12	255.17
16	229.33
20	204.83
24	179.00
28	163.83
32	150.33



Different straight lines can be fitted. Question: what is the line of best fit? We measure fit by **sum of squared vertical deviations** and **minimize** over all lines to get the least squares line.

Consider any line $\hat{y} = a + bX$; deviations from this line of observed Y_i are $e_i = Y_i - \hat{y}_i = Y_i - a - bX_i$.

Sum of squared deviations: $SS = \Sigma e_i^2 = \Sigma_{i=1}^n (Y_i - a - bX_i)^2$.

Definition of the OLS (ordinary least squares) estimator.

OLS estimator of the parameters is the argument of the function

$$SS(a, b) = \Sigma_{i=1}^n (Y_i - a - bX_i)^2$$

at which this function is minimized.

Denote this value $\left(\hat{\beta}_0, \hat{\beta}_1\right)$. Then the definition:

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg\min_{a,b} \Sigma_{i=1}^n (Y_i - a - bX_i)^2.$$

Solving for OLS estimators.
To solve minimize $\Sigma_{i=1}^n (Y_i - a - bX_i)^2$ with respect to the two values.
First-order condition (FOC).

$$\begin{aligned} \frac{\partial SS}{\partial a} &= -2\Sigma(Y_i - a - bX_i) = 0; \quad (1) \\ \frac{\partial SS}{\partial b} &= -2\Sigma(Y_i - a - bX_i)X_i = 0. \quad (2) \end{aligned}$$

This is equivalent to

$$\left\{ \begin{array}{ll} \Sigma Y_i = & an + b\Sigma X_i \\ \Sigma Y_i X_i = & a\Sigma X_i + b\Sigma X_i^2. \end{array} \right.$$

Then dividing the first equation by n we get the so-called normal equations:

$$\left\{ \begin{array}{ll} \bar{Y} = & a + b\bar{X} \\ \Sigma Y_i X_i = & a\Sigma X_i + b\Sigma X_i^2. \end{array} \right. \quad (3)$$

These are two linear equations with two unknowns, a, b.
Solving two linear equations with two unknowns.
Recall how you solve a system
$\left\{ \begin{array}{ll} B_1 = & A_{11}x + A_{12}y \\ B_2 = & A_{21}x + A_{22}y \end{array} \right.$ or, equivalently in matrix form: $\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} B_1 \\ B_2 \end{pmatrix}$.
The inverse matrix is $\frac{1}{A_{11}A_{22} - A_{12}A_{21}} \begin{pmatrix} A_{22} & -A_{12} \\ -A_{21} & A_{11} \end{pmatrix}$; solution

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \frac{A_{22}B_1 - A_{12}B_2}{A_{11}A_{22} - A_{12}A_{21}} \\ \frac{-A_{21}B_1 + A_{11}B_2}{A_{11}A_{22} - A_{12}A_{21}} \end{pmatrix}.$$

Solution for first order conditions for OLS.
So solution to (3) denoted $(\hat{\beta}_0, \hat{\beta}_1)$ is

$$\begin{aligned} \hat{\beta}_0 &= \frac{(\Sigma X_i^2)\bar{Y} - \bar{X}(\Sigma Y_i X_i)}{\Sigma X_i^2 - (\Sigma X_i)\bar{X}}; \\ \hat{\beta}_1 &= \frac{(\Sigma X_i Y_i) - n\bar{X}\bar{Y}}{\Sigma X_i^2 - (\Sigma X_i)\bar{X}}. \end{aligned}$$

Denote $\Sigma(X_i - \bar{X})\Sigma(Y_i - \bar{Y}) = (\Sigma X_i Y_i) - n\bar{X}\bar{Y}$ by S_{xy}; $\Sigma(X_i - \bar{X})^2 = \Sigma X_i^2 - n\bar{X}^2$ by S_{xx}; $\Sigma(y_i - \bar{Y})^2 = \Sigma Y_i^2 - n\bar{Y}^2$ by S_{yy}.

Then $\hat{\beta}_1$ can be expressed as:

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}.$$

Also

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X};$$

In the example. (complete the calculation)

i	X_i	Y_i	X_i^2	X_iY_i
1	0	394.33		
2	4	329.5		
3	8	291.00		
4	12	255.17		
5	16	229.33		
6	20	204.83		
7	24	179.00		
8	28	163.83		
9	32	150.33		
Σ	144	2197.32	3264	28167.72

$\bar{X} = 16; \bar{Y} = 244.15;\ S_{xy} = 28167.72 - \frac{1}{9}(144 \times 2197.32) = -6989.4; S_{xx} = 3264 - \frac{1}{9}(144)^2 = 960.$

$\hat{\beta}_1 = \frac{-6989.4}{960} = -7.281;$

$\hat{\beta}_0 = 244.15 + 7.281 \times 16 = 360.65$

The OLS line is $y = 360.65 - 7.281x$.

Fitted values and residuals.

The values on the regression line $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ are fitted values (predictors of Y_i), and the differences $e_i = Y_i - \hat{Y}_i$ are regression residuals.

We can compute them for the example (finish the computation).

i	X_i	Y_i	X_i^2	X_iY_i	$\hat{Y}_i$	$\hat{e}_i$
1	0	394.33			360.65	33.68
2	4	329.5			331.53	−2.03
3	8	291.00			302.4	
4	12	255.17			273.28	
5	16	229.33				
6	20	204.83				
7	24	179.00				
8	28	163.83				
9	32	150.33				
Σ	144	2197.32	3264	28167.72		0

Properties:

Sum of regression residuals is zero: $\Sigma e_i = 0$ (from FOC(1)).

Regression line passes through the point of means $(\bar{X}, \bar{Y})$, (from normal equations (3)).

Then the average fitted value is the same as the average of observed: $\overline{\hat{Y}} = \bar{Y}$.

3. **Goodness of fit of the regression line.**

In the example SSresiduals=$\Sigma e_i^2 = \Sigma(Y_i - \hat{Y}_i)^2 = 2531.53$.

This is a minimal possible value for any straight line.

How good is it? One way to evaluate this is to compare with just evaluating the dependent variable by its expectation, $\bar{Y}$, without conditioning on X. Of course, the sum of squared deviations will be larger (why?), but will the difference be sufficient to justify the extra complication?

We want to compare the TSS (total sum of squares): $TSS = \Sigma(Y_i - \bar{Y})$ with the SSresiduals: $\Sigma(Y_i - \hat{Y})^2$ and see (a) what the difference represents; (b) how large is it.

(a) $Y_i - \bar{Y} = (\hat{Y}_i - \bar{Y}) + (Y_i - \hat{Y}_i)$; total variation is decomposed into variation coming from the regression (explained, "between") and variation of the residuals ("within").

Square and sum:

$$TSS = S_{YY} = \Sigma\left(Y_i - \bar{Y}\right)^2 = \Sigma(\hat{Y}_i - \bar{Y})^2 + \Sigma(Y_i - \hat{Y}_i)^2,$$

since $2\Sigma(\hat{Y}_i - \bar{Y})(Y_i - \hat{Y}_i) = 0$ (from $\Sigma(\hat{Y}_i - \bar{Y})(Y_i - \hat{Y}_i) = \left(\hat{\beta}_0 - \bar{Y}\right)\Sigma(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) + \hat{\beta}_1 \Sigma X_i(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0$ from FOC).

So TSS=SSresiduals+SSexplained by regression.

Coefficient of determination

$$R^2 = \frac{\text{SSexplained by regression}}{TSS} = 1 - \frac{SSresiduals}{TSS}.$$

Note that $0 \leq R^2 \leq 1$.

In example TSS=53418.73; SSexplained=TSS-SSresiduals=53418.73-2531.53=50887.2.

$R^2 = \frac{50887.2}{53418.73} = 0.952\,61$.

SSexplained=$\Sigma(\hat{Y}_i - \bar{Y})^2 = \Sigma(\hat{\beta}_0 + \hat{\beta}_1 X_i - (\hat{\beta}_0 + \hat{\beta}_1 \bar{X}))^2 = \hat{\beta}_1^2 S_{XX}$.

Thus

$$R^2 = \frac{\text{SSexplained by regression}}{TSS} = \frac{\hat{\beta}_1^2 S_{XX}}{S_{YY}} = \frac{S_{XY}^2}{S_{XX}^2}\frac{S_{XX}}{S_{YY}} = \frac{S_{XY}^2}{S_{XX}S_{YY}}.$$

Correlation coefficient.

Recall correlation between X and Y.

$$r = \frac{S_{xy}/n-1}{\sqrt{\frac{S_{xx}}{n-1}}\sqrt{\frac{Syy}{n-1}}} = \frac{S_{XY}}{\sqrt{S_{XX}S_{YY}}}.$$

There is a relation between correlation of X, Y and regression. The formulas imply that

$$\begin{aligned} r &= \hat{\beta}_1\sqrt{\frac{S_{XX}}{S_{YY}}}; \\ r^2 &= R^2. \end{aligned}$$

In the example $r^2 = 0.952\,61$, the sign of r is the same as that of $\hat{\beta}_1$, -ve. so $r = -\sqrt{0.952\,61} = -0.976\,02$.

There is a straightforward relation. The stronger the linear relation, the higher the $|r|$ and R^2.

4. Estimation of variance, σ^2.

$\sigma^2 = E\varepsilon_i^2$, where $\varepsilon_i = Y_i - \beta_0 - \beta_1 X_i$.

The errors are not observed, we only observe the regression residuals: $e_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i$.

We estimate the population moment of the unobservable errors, $E\varepsilon_i^2$, by the sample moments of the regression residuals; we divide by d.f. whai here is $n-2$, since 2 parameters, β_0 and β_1 were estimated:

$$\hat{\sigma}^2 = \frac{1}{n-2}\Sigma e_i^2 = \frac{1}{n-2}\Sigma(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2.$$

This is similar to the way the variance of a population was estimated based on the average of squared deviations from the estimated mean $\hat{\sigma}^2 = \frac{1}{n-1}\Sigma(Y_i - \bar{Y})$, for an unbiased estimator we divide by d.f., where for the mean one was lost because we use the estimated by $\bar{Y}$ population mean.

In the example SSresiduals=2531.53, n=9, so $\hat{\sigma}^2 = \frac{2531.53}{7} = 361.\,65$.

Topic 4. Simple linear regression. Part 2.

5. Inference in the simple linear regression model.

Assumptions.

$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$,

where X is non-stochastic (non-random), and

$$\begin{aligned} E\varepsilon_i &= 0; \\ E(\varepsilon_i^2) &= \sigma^2; \\ E(\varepsilon_i \varepsilon_j) &= 0 \text{ for } i \neq j. \end{aligned}$$

5.1. **Properties and the distribution of the OLS estimators.**

5.1.1 **Linearity**.

Recall that

$$\begin{aligned} \hat{\beta}_0 &= \bar{Y} - \hat{\beta}_1 \bar{X}; \\ \hat{\beta}_1 &= \frac{\Sigma Y_i X_i - n\bar{Y}\bar{X}}{\Sigma X_i^2 - n\bar{X}^2} \\ &\quad \text{and write} \\ &= \frac{\Sigma Y_i (X_i - \bar{X})}{\Sigma (X_i - \bar{X})^2} = \Sigma w_i Y_i \end{aligned}$$

with $w_i = \frac{X_i - \bar{X}}{\Sigma (X_i - \bar{X})^2}$.

The estimators are **linear** functions of the random Y_i with non-random coefficients. (Show for $\hat{\beta}_0$, too).

Lemma. Properties of the weights: a. $\Sigma w_i = 0$; b. $\Sigma w_i X_i = 1$; c. $\Sigma w_i^2 = \frac{1}{\Sigma (X_i - \bar{X})^2}$.

Proof.
a. $\Sigma w_i = \Sigma \frac{X_i - \bar{X}}{\Sigma (X_i - \bar{X})^2} = \frac{\Sigma \left(X_i - \bar{X}\right)}{\Sigma (X_i - \bar{X})^2} = 0$;
b. $\Sigma w_i X_i = \frac{\Sigma (X_i - \bar{X}) X_i}{\Sigma (X_i - \bar{X})^2} = \frac{\Sigma X_i^2 - \bar{X} \Sigma X_i}{\Sigma (X_i - \bar{X})^2} = \frac{\Sigma \left(X_i - \bar{X}\right)^2}{\Sigma \left(X_i - \bar{X}\right)^2} = 1$
since $\Sigma \left(X_i - \bar{X}\right)^2 = \Sigma X_i^2 - 2\Sigma X_i \bar{X} + n\bar{X}^2 = \Sigma X_i^2 - \Sigma X_i \bar{X}$.
c. $\Sigma w_i^2 = \Sigma \frac{\left(X_i - \bar{X}\right)^2}{\left(\Sigma (X_i - \bar{X})^2\right)^2} = \frac{1}{\Sigma (X_i - \bar{X})^2}$.

■

5.1.2. **Unbiasedness**.
The estimators are **unbiased**.
Theorem. $E\left(\hat{\beta}_0\right) = \beta_0; E(\hat{\beta}_1) = \beta_1$.
Proof.
$E\hat{\beta}_1 = \Sigma w_i E Y_i = \Sigma w_i (\beta_0 + \beta_1 X_i) = \beta_0 \Sigma \omega_i + \beta_1 \Sigma w_i X_i$.
Then from properties a. and b. of the Lemma $E\hat{\beta}_1 = \beta_1$.
$E\hat{\beta}_0 = E\left(\bar{Y} - \hat{\beta}_1 \bar{X}\right) = E\bar{Y} - \bar{X} E\hat{\beta}_1 = \frac{1}{n}(n\beta_0 + \beta_1 \Sigma X_i) - \bar{X}\beta_1 = \beta_0$.
■

5.1.3. **Variance and MSE**.
Variance of the estimators.
Theorem. $var\left(\hat{\beta}_1\right) = \sigma^2 \frac{1}{\Sigma (X_i - \bar{X})^2}$;
$var\left(\hat{\beta}_0\right) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{X}}{\Sigma (X_i - \bar{X})^2}\right)$.
Proof.
$var\hat{\beta}_1 = \Sigma w_i^2 var Y_i = \sigma^2 \Sigma w_i^2 = \frac{\sigma^2}{\Sigma (X_i - \bar{X})^2}$ from c. of the Lemma.
$var\hat{\beta}_0 = var\bar{Y} - 2cov\left(\bar{Y}, \hat{\beta}_1\right)\bar{X} + var(\hat{\beta}_1)\bar{X}^2 = \frac{\sigma^2}{n} - 2\bar{X}\Sigma w_i cov(\bar{Y}, Y_i) + \sigma^2 \frac{\bar{X}^2}{\Sigma (X_i - \bar{X})^2}$;
note that $cov(\bar{Y}, Y_i)$ is the same for all i, then by a. of Lemma we find that the middle term is zero and
$var\left(\hat{\beta}_0\right) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{\Sigma (X_i - \bar{X})^2}\right)$.■

Recall MSE (mean squared error) of an estimator $\hat{\beta}$ is defined as $\text{MSE}(\hat{\beta}) = E\left(\hat{\beta} - \beta\right)^2$ and thus (adding and subtracting $E\left(\hat{\beta}\right)$) :

$$\begin{aligned} MSE\left(\hat{\beta}\right) &= E\left(\hat{\beta} - E\left(\hat{\beta}\right)\right)^2 + \left(E\left(\hat{\beta}\right) - \beta\right)^2 \\ &= var\left(\hat{\beta}\right) + \left(bias\left(\hat{\beta}\right)\right)^2. \end{aligned}$$

Since the estimators are unbiased the variance and MSE are the same.
5.1.4. **Gauss-Markov Theorem.**
Theorem. The OLS estimators in the linear model

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,$$

where $X's$ are non-random and $E\varepsilon_i = 0; E\varepsilon_i\varepsilon_j = \sigma^2$, if $i = j$, zero otherwise, have the **smallest variance** (and MSE) among all unbiased linear estimators.

Proof.

The proof is optional for your reading.

The proof here is provided for $var\left(\hat{\beta}_1\right)$, variance of the slope (the intercept is similar); the proof uses constrained optimization via a Lagrangian. Consider an unbiased linear estimator for the slope: $\tilde{\beta} = \Sigma\tilde{w}_i Y_i$ (linear form); since unbiased $E\tilde{\beta} = \beta_1$, where

$$\begin{aligned} E\tilde{\beta} &= \Sigma\tilde{w}_i EY_i = \Sigma\tilde{w}_i\left(\beta_0 + \beta_1 X_i\right) \\ &= \beta_0\Sigma\tilde{w}_i + \beta_1\Sigma\tilde{w}_i X_i, \end{aligned}$$

so $\Sigma\tilde{w}_i = 0$; $\Sigma\tilde{w}_i X_i = 1$.

Variance of this linear estimator is $var\left(\tilde{\beta}\right) = \Sigma\tilde{w}_i^2 varY_i = \sigma^2\Sigma\tilde{w}_i^2$ since $cov\left(Y_i, Y_j\right) = cov\left(\varepsilon_i, \varepsilon_j\right) = 0$ for $i \neq j$ by the assumptions of the model.

The linear unbiased estimator with the smallest variance will have weights, $\{\tilde{w}_i\}_{i=1}^n$ such that $\Sigma\tilde{w}_i^2$ is minimized, and the conditions $\Sigma\tilde{w}_i = 0$; $\Sigma\tilde{w}_i X_i = 1$ are satisfied.

Consider the Lagrangian

$$L = \Sigma w_i^2 - \mu\Sigma w_i - \lambda(\Sigma w_i X_i - 1).$$

The FOC for constrained minimization are

$$\begin{aligned} \frac{\partial L}{\partial w_i} &= 2w_i - \lambda X_i - \mu = 0;\ i = 1, ..., n; \\ (*)\ \frac{\partial L}{\partial \lambda} &= \Sigma w_i X_i - 1 = 0; \\ (**)\ \frac{\partial L}{\partial \lambda} &= \Sigma w_i = 0. \end{aligned}$$

To solve, first sum the first set of n equations:

$$2\Sigma w_i - \lambda\Sigma X_i - n\mu = 0,$$

then substituting from $(**)$ express:

$$(***)\ \mu = -\lambda\bar{X}.$$

Next multiply each of the first n equations by corresponding X_i and sum:

$$2\Sigma w_i X_i - \lambda\Sigma X_i^2 - \mu\Sigma X_i = 0,$$

From $(*)$ and $(***)$ we get $\left(\Sigma X_i = n\bar{X}\right)$:

$$2 - \lambda\left(\Sigma X_i^2 - n\bar{X}^2\right) = 0.$$

Then $\lambda = 2\frac{1}{\Sigma X_i^2 - n\bar{X}^2}; \mu = -2\frac{\bar{X}}{\Sigma X_i^2 - n\bar{X}^2}$ and substituting into the each of the first n equation from the FOC the weights are $w_i = \frac{X_i - \bar{X}}{\Sigma X_i^2 - n\bar{X}^2}$.

These solutions are the weights of the linear unbiased OLS estimator. Then any $\tilde{\beta}$ with weights $\tilde{w}$ that do not coincide with these cannot be a solution to the minimization, and thus would have a larger variance.■

This property of the OLS estimator is **BLUE** (Best Linear Unbiased Estimator).

This means that for any other linear estimator $\tilde{\beta}_0, \tilde{\beta}_1$ of β_0, β_1 such that it is unbiased $var\left(\tilde{\beta}_0\right) \geq var\left(\hat{\beta}_0\right)$ and $var\left(\tilde{\beta}_1\right) \geq var\left(\hat{\beta}_1\right)$.

For example, the sample mean $\hat{\beta} = \bar{Y}$ is the least squares estimator of the expectation β for the model $Y_i = \beta + \varepsilon_i$ (without any X).

Example. In a sample of n observations of an i.i.d. variable Y with variance σ^2 consider an estimator for the mean given by the average of all but the first and last observation in the sample: $\tilde{\beta} = \frac{\Sigma_{i=2}^{n-1} Y_i}{n-2}$ of β. Show that it is linear and unbiased, compute its variance and show that it is bigger than for OLS.

5.1.5. The **distribution** of the estimators.

In the classical regression model (the errors are independent $N(0, \sigma^2)$, the $X's$ non-stochastic) the OLS estimators are normally distributed (this follows from the facts that the estimators are linear in Y_i; Y_i are normally distributed if ε_i are; a linear combination of normal variables is normal).

$\hat{\beta}_0$ is $N(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{X}^2}{\Sigma(X_i - \bar{X})^2}\right)$; $\hat{\beta}_1$ is normal $N(\beta_1, \sigma^2 \frac{1}{\Sigma(X_i - \bar{X})^2})$.

In the more general regression model under suitable assumptions, the distribution of OLS estimators can be approximated by a normal distribution (asymptotic distribution).

5.2. **Confidence intervals and hypotheses tests in the simple linear regression.**

5.2.1. Confidence intervals and hypotheses tests for the coefficients and significance of the model.

$\hat{\beta}_0$ is $N(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{X}^2}{\Sigma(X_i - \bar{X})^2}\right))$; denote the standard deviation of the distribution by $\mathrm{SD}\left(\hat{\beta}_0\right)$; standardize:

$$z = \frac{\hat{\beta}_0 - \beta_0}{SD\left(\hat{\beta}_0\right)} \sim N(0,1).$$

Similarly, $\hat{\beta}_1$ is $N(\beta_1, \sigma^2 \frac{1}{\Sigma(X_i - \bar{X})^2})$; denote the standard deviation of the distribution by $\mathrm{SD}\left(\hat{\beta}_1\right)$; standardize:

$$z = \frac{\hat{\beta}_1 - \beta_1}{SD\left(\hat{\beta}_1\right)} \sim N(0,1).$$

For confidence intervals and hypotheses tests **when σ^2 is known** we can use **the normal distribution**.

When σ^2 is not known we use the estimator $\hat{\sigma}^2 = \frac{1}{n-2}\Sigma e_i^2 = \frac{1}{n-2}\Sigma(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$.

For $\hat{\beta}_0$ substituting estimated $\hat{\sigma}^2$ gives as estimated variance $\hat{\sigma}^2\left(\frac{1}{n}+\frac{\bar{X}^2}{\Sigma(X_i-\bar{X})^2}\right)$ and **standard error** of the estimator is

$$s(\beta_0)=\hat{\sigma}\sqrt{\frac{1}{n}+\frac{\bar{X}^2}{\Sigma(X_i-\bar{X})^2}}.$$

Similarly, for $\hat{\beta}_1$

$$s\left(\hat{\beta}_1\right)=\hat{\sigma}\sqrt{\frac{1}{\Sigma(X_i-\bar{X})^2}}.$$

Then the ratio has a t **distribution with d.f.=n-2**.

$$\begin{aligned}\frac{\hat{\beta}_0-\beta_0}{s\left(\hat{\beta}_0\right)} &\sim t_{n-2};\\ \frac{\hat{\beta}_1-\beta_1}{s\left(\hat{\beta}_1\right)} &\sim t_{n-2}.\end{aligned}$$

Example. At a public utility with 312 employees the linear model to explain the effect of age of employee on wage was estimated.

Model $y_i=\beta_0+\beta_1 x_i+\varepsilon_i$, where y_i represents wage of employee i, x_i- age. Assume classical regression model. Results of estimation.

coefficients	value	st.error	$t-ratio$	p-value
constant	46950.64	4551.36	10.32	0.0000
age	309.27	95.10	3.25	.0013

For p-value the $t-ratio$ is evaluated using the t_{310}, since d.f.=n-2=312-2.

Interpretation of coefficients.

Significance testing.

$H_0:\beta_1=0$, vs $H_1:\beta_1\neq 0$.

Rejection of H_0 means that the explanatory varible does play a role and helps (at the approprite level of the test) explain variation in the dependent vriable of interest.

Here the p-value indicates that the age slope is significant.

95% confidence interval for the slope coefficient provides ($t_{300,.025}=1.96$):

$$309.27-1.96\cdot 95.10<\beta_1<309.27+1.96\cdot 95.10.$$

In that example despite the significance of the regression coefficients the fit was poor: $R^2=.033$.

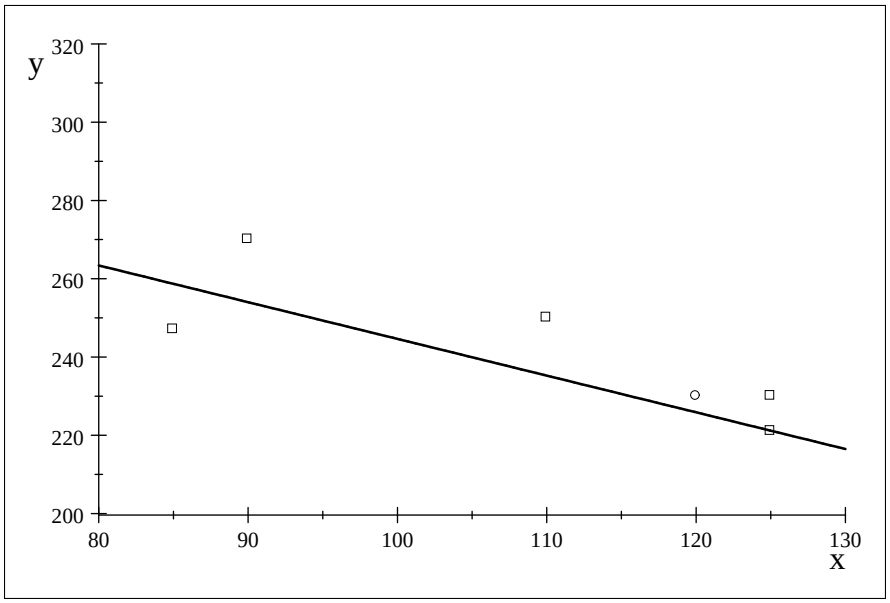
It turns out that an important variable was not considered: gender of the employee. Once that was included the R^2 increased to .298. (we shall examine this in Multiple regression part of the course).

Example. Demand for long-distance bus travel route depending on fare. $n=11$.

The estimated regression $\hat{y}_i = 338.43 - .938x_i$, where x_i is a real fare index (varies the ticket fare from the average real fare with index $\bar{x}$=100). The output from the OLS

coefficients	value	st.error	$t-ratio$
constant	338.43	19.72	11.16
fare index	$-.938$	.1909	-4.91

The coefficients are significant.



Estimated $\hat{\sigma} = 5.974$; $R^2 = .728$.

Although R^2 indicates goodness of fit it is not a statistic with a known distribution.

Test of goodness of fit.

In order to perform a **test of goodness of fit** of the regression model we can apply **ANOVA**.

The null is that the regression does not add to explaining the dependent variable. In simple regression this means $H_0 : \beta_1 = 0$. (later in multiple regression we'll see the separate importance of this test).

We also say that we are testing one **restriction**: $\beta_1 = 0$.

To calculate the coefficients various sums of square were computed: TSS=S_{yy}; SSresiduals=Σe_i^2=$\Sigma \left(y_i - \hat{y}_i\right)^2 = \Sigma \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2$, then SSexplained=$\Sigma(\hat{y}_i - \bar{y})^2 = TSS - SSresid$. Enter them into the ANOVA table:

source	SS	d.f.	MSS
regression (explained)	860.81	1	860.81
error(residuals)	321.19	9	35.69
total	1182.00	10	

D.f. in regression = # of restrictions=1.

The F ratio computed for the example is $\frac{860.81}{35.69} = 24.119$. The F ratio is distributed as $F_{1,9}$; for $\alpha = .05$ the critical value is 5.12. The null is rejected and the regression is significant.

Note. For simple regression $F-$ratio is equal to $t^2_{\hat{\beta}_1}$.

Algebraic proof (optional). $t^2_{\hat{\beta}_1} = \frac{\hat{\beta}^2_1}{\hat{s}^2_{\hat{\beta}_1}}$, substitute (derived earlier) $s^2_{\hat{\beta}_1} = \frac{SSresid/(n-2)}{S_{XX}}$ and (recall) the SSexplained=$\hat{\beta}^2_1 S_{XX}$.

In this example $4.91^2 = 24.108$, the difference is due to rounding error in computation.

Forecasting with the simple regression model.

Suppose that we wish to estimate the demand for bus travel at $x = 125$.

There are two different questions:

(a) one is what is the **expected** demand $E(y$ at $x = 125)$?

(b) the other is fare index will be 125 next week, what will be the **actual** demand y at $x = 125$?

For (a) we want $E(y|x = 125) = \beta_0 + \beta_1 \cdot 125$.

For (b) we would like to have $y(x = 125) = \beta_0 + \beta_1 \cdot 125 + \varepsilon$ (including the value of ε that will occur for $x = 125$).

Start with question (a). **Forecast of the expected value of** y.

We do not have β_0 and β_1, but we have estimators and can estimate $E(y|x = 125)$ by $\hat{y}(x = 125) = 338.43 - .938 \cdot 125 = 221.18$.

This is an **unbiased estimator of expectation**.

Indeed,

$E(\hat{y}(x)) = E(\hat{\beta}_0) + E(\hat{\beta}_1) \cdot x = \beta_0 + \beta_1 \cdot 125 = E(y|x = 125)$.

How accurate is our forecast? We evaluate the **mean squared error (MSE) of the forecast**.

The **forecast error** (=difference between the estimated forecast and the true value we are trying to forecast) is $e^f = \hat{y}(x) - E(y|x) = \hat{\beta}_0 + \hat{\beta}_1 \cdot x - \beta_0 - \beta_1 \cdot x$.

The mean square error of the forecast for the expected value (equals the variance of forecast error since $Ee^f = 0$) :

$$MSE^f = E\left(e^f\right)^2 = E\left(\hat{\beta}_0 - \beta_0\right)^2 + 2E\left(\hat{\beta}_0 - \beta_0\right)(\hat{\beta}_1 - \beta_1) \cdot x + x^2 E(\hat{\beta}_1 - \beta_1)^2.$$

Derivation of the expression (optional).

First, the term $E\left(\hat{\beta}_0-\beta_0\right)(\hat{\beta}_1-\beta_1)=$

$$\begin{aligned}
& E(\bar{Y}-\hat{\beta}_1\bar{X}-\beta_0)(\hat{\beta}_1-\beta_1) \\
= \; & E\left(\beta_0+\beta_1\bar{X}+\frac{1}{n}\Sigma\varepsilon_i-\hat{\beta}_1\bar{X}-\beta_0\right)(\hat{\beta}_1-\beta_1) \\
\text{using } \hat{\beta}_0 \; = \; & \bar{Y}-\hat{\beta}\bar{X} \text{ and } \bar{Y}=\beta_0+\beta_1\bar{X}+\frac{1}{n}\Sigma\varepsilon_i, \text{ then} \\
= \; & E\left(\beta_1\bar{X}-\hat{\beta}_1\bar{X}+\frac{1}{n}\Sigma\varepsilon_i\right)(\hat{\beta}_1-\beta_1) \\
= \; & -\bar{X}E\left(\hat{\beta}_1-\beta_1\right)^2+\frac{1}{n}\Sigma E\varepsilon_i\hat{\beta}_1.
\end{aligned}$$

Show that $\frac{1}{n}\Sigma E\varepsilon_i\hat{\beta}_1=0$.Indeed, substitute $\hat{\beta}_1=\Sigma w_jY_j=\beta_0\Sigma w_j+\beta_1\Sigma w_jX_j+\Sigma w_j\varepsilon_j=\beta_1+\Sigma w_j\varepsilon_j$ (recall $\Sigma w_i=0;\Sigma w_jX_j=1$),

$$\begin{aligned}
\text{since } E\varepsilon_i \; = \; & 0 \text{ it follows that } \frac{1}{n}\Sigma E\varepsilon_i\beta_1=0, \\
\text{then } \frac{1}{n}\Sigma E\varepsilon_i\hat{\beta}_1 \; = \; & \frac{1}{n}\Sigma E\varepsilon_i\left(\Sigma_{j=1}^n w_j\varepsilon_j\right) \\
\text{and (recall } E\varepsilon_i \; = \; & 0; E\varepsilon_i\varepsilon_j=\sigma^2, \text{ if } i=j, 0 \text{ otherwise}): \\
\text{Thus } \Sigma_{i=1}^n E\varepsilon_i\Sigma_{j=1}^n w_j\varepsilon_j \; = \; & \Sigma w_j\sigma^2=0.
\end{aligned}$$

So substituting $-\bar{X}E\left(\hat{\beta}_1-\beta_1\right)^2=-\sigma^2\frac{\bar{X}}{\Sigma(X_i-\bar{X})^2}$ we get the result:

$$\begin{aligned}
MSE^f \; = \; & \sigma^2(\frac{1}{n}+\frac{\bar{X}^2}{\Sigma(X_i-\bar{X})^2}-2\frac{\bar{X}x}{\Sigma(X_i-\bar{X})^2}+\frac{x^2}{\Sigma(X_i-\bar{X})^2}) \\
= \; & \sigma^2\left(\frac{1}{n}+\frac{(x-\bar{X})^2}{\Sigma(X_i-\bar{X})^2}\right).
\end{aligned}$$

Confidence interval for the forecast of the expected value.

Thus a $1-\alpha$ confidence interval for the forcast of the expected value is

$$[\hat{y}-t_{n-2,\alpha/2}\hat{\sigma}\sqrt{\left(\frac{1}{n}+\frac{(x-\bar{X})^2}{\Sigma(X_i-\bar{X})^2}\right)},\hat{y}-t_{n-2,\alpha/2}\hat{\sigma}\sqrt{\left(\frac{1}{n}+\frac{(x-\bar{X})^2}{\Sigma(X_i-\bar{X})^2}\right)}]$$

This interval depends in particular on how far from the mean of the sample is the point at which we wish to forecast.

Forecast of the actual value of y.

For (b) add unknown ε. Thus for the forecast of the actual value $y\,(x)=\beta_0+\beta_1x+\varepsilon$ at some x the estimate of the forecast is still the same, $\hat{\beta}_0+\hat{\beta}_1x$.

This is an unbiased forecast. (Show).

The forecast error:

$$e^f=\hat{y}(x)-y(x)=\hat{\beta}_0+\hat{\beta}_1\cdot x-\beta_0-\beta_1\cdot x-\varepsilon.$$

The MSE of the forecast:

$$MSE^f = \sigma^2 \left(1 + \frac{1}{n} + \frac{(x - \bar{X})^2}{\Sigma(X_i - \bar{X})^2}\right).$$

It is more difficult to predict a particular outcome than the expected value due to the fact that the actual value of y will include the random error that cannot be predicted but is taken account of in the forecast error and increases (relative to the MSE for forecast of the expected value) the MSE of the forecast.

Example. Consumption function.

Y denotes consumption, X - income.

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

The estimated model is for a sample of size $n = 25$, $\left(\text{where } \bar{X} = 61.92, \Sigma X_i^2 = 99862\right)$

$$\hat{Y} = 8.83 + .76X$$

with $\hat{\sigma}^2 = .6548$; $R^2 = .993$. The standard errors were $\left(SE\left(\beta_0\right)\right)^2 = .669$; $\left(SE\left(\beta_1\right)\right)^2 = .000167$.

Significance test: for $\hat{\beta}_0$ the t-statistic $\frac{8.83}{\sqrt{.669}} = 10.796$; for $\hat{\beta}_1$ we get $\frac{.76}{\sqrt{.000167}} = 58.811$. With the asymptotic distrbution t_{23} these are highly significant ($H_0 : \beta_. = 0$ is rejected).

Prediction.

Suppose that we wish to predict consumption level for income at 73.36.

$\hat{Y} = 8.83 + .76 \cdot 73.36 = 64.584$.

Construct the 90% CI for the prediction.

We need to compute $\Sigma X_i - \bar{X})^2 = \Sigma X_i^2 - n\bar{X}^2 = 99862 - 25 \cdot 61.92^2 = 4009.8$. Also $\hat{\sigma} = \sqrt{.6548} = 0.809\,20$.

Then

$$[64.584 - t_{23,.05} 0.809 \sqrt{\left(1 + \frac{1}{25} + \frac{(73.36 - 61.92)^2}{4009.8}\right)}, 64.584 + t_{23,.05} 0.809 \sqrt{\left(1 + \frac{1}{25} + \frac{(73.36 - 61.92)^2}{4009.8}\right)}]$$

Substitute $t_{23,.05} = 1.71$ and compute $1.71 \cdot 0.809 \sqrt{\left(1 + \frac{1}{25} + \frac{(73.36-61.92)^2}{4009.8}\right)} = 1.4328$

$$\begin{aligned} & [64.584 - 1.4328,\ 64.584 + 1.4328] \\ = \ & [63.151,\ 66.017]. \end{aligned}$$

Suppose we were predicting the expected value of consumption at $x = 73.36$.The predicted value is the same, 64.584. But now the MSE is smaller:

$.6548 \left(\frac{1}{25} + \frac{(73.36-61.92)^2}{4009.8}\right) = .047564$ as compared to the MSE of the forecast of the actual value,

$.6548\left(1+\frac{1}{25}+\frac{(73.36-61.92)^2}{4009.8}\right)=0.702\,36.$

Forecast accuracy (summary).

Forecast accuracy of the expected value is better (confidence intervals narrower) than for the actual value.

Forecast accuracy depends on the distance of the point X at which we wish to forecast from the mean of the $x's$, $\bar{X}$; the farther from $\bar{X}$ the wider the confidence interval.

A few comments on OLS regression.

(a) Changing **units of measurement**.

Changing measurement of y (say $\tilde{y}=100y$) but not x inflates the $\beta's$ accordingly (by 100).

So e.g. regress wages on years of schooling and use wages in \$ ves 1000\$.

Changing unit of measurement of x (say, $\tilde{x}=100x$) reduces the slope, β_1 by that (divide by 100) but does not change the intercept.

In the example with O-rings measure the temperature in C rather than F. Note that here not just scale change but also shift in X : (°F - 32) x 5/9 = °C

This will affect both the intercept and slope: $y=\beta_0+\beta_1(b\tilde{X}+a)+\varepsilon=\tilde{\beta}_0+\tilde{\beta}_1\tilde{X}+\varepsilon$, where $\tilde{\beta}_0=\beta_0+a\beta_1;\tilde{\beta}_1=b\beta_1$.

R^2 does not depend on units of measurement.

(b) Incorporating **nonlinearities** in regression.

model	dep.var.	indep.var.	effect of change in x	interpretation of β_1
level/level	y	x	$\Delta y=\beta_1\Delta x$	change in y per unit change in x
level/ log	y	$\log x$	$\Delta y=(\beta_1/100)\%\Delta x$	change in y per 1% change in x
log/level	$\log y$	x	$\%\Delta y=(100\beta_1)\Delta x$	% change in y per unit change in x
log/log	$\log y$	$\log x$	$\%\Delta y=\beta_1\%\Delta x$	% change in y per 1% change in x

Examples.

1. Model $\log(wage)=\beta_0+\beta_1 educ+u$; then in expectation $\%\Delta wage=(100\beta_1)\Delta educ$ we get percent change in wage for each additional year of education, so the the change in wage is assumed to increase here as education increases providing wage as an expenential function $\exp(\beta_0+\beta_1 educ+u)$.

2. Constant elasticity model: $\log(salary)=\beta_0+\beta_1\log(sales)+u$. Here β_1 is the elasticity of salary with respect to sales.

(c) Problem of **outliers.**

Outliers in the y direction.

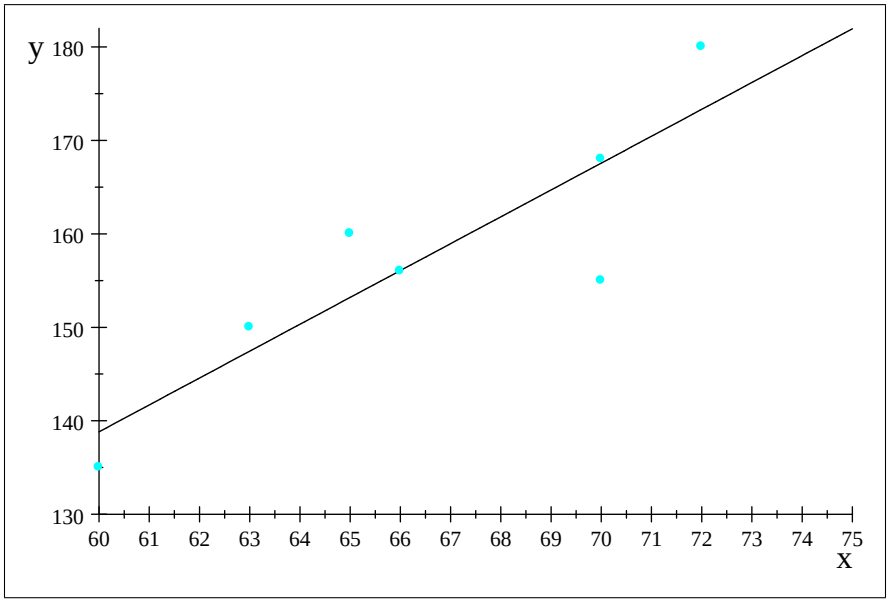
An unusually large +ve or -ve ε_i was drawn. This results on a value of y that is far from the expected value. The OLS regression is affected; because of squaring of the residuals the regression line is drawn in the direction of the outlier.

Example.

Consider the data

X	Y
70	155
63	150
72	180
60	135
66	156
70	168
65	160

$\hat{Y} = 2.874\,7X - 33.657$



Suppose one of the observations were $(66, 186)$ instead of $(66, 156)$, then $\hat{y} = 2.721\,2x - 19.156$.

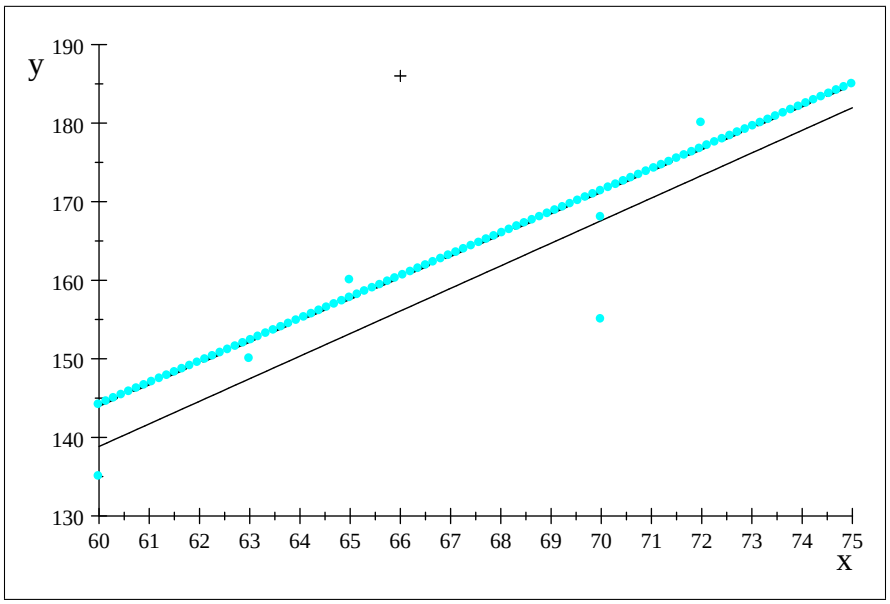
This result is very different. Is this due to the fact that the replacement observation is actually an outlier?

We proceed as follows to answer the question whether an observation is an outlier.

Dropping this observation altogether we get

$\hat{y} = 2.874\,3x - 33.617$.

Standard error of the regression is $\hat{\sigma} = \sqrt{270.24/4} = 8.219$.



(On the graph the fat regression line for the sample that includes the "suspect observation" shows that it moves towards the outlier and possibly misrepresents the majority of observations. Removing this observation produces a regression line that is fairly close to the other observations.) **The residual** from the dropped observation to the regression line is $e = 186 - (2.8743 \cdot 66 - 33.617) = 29.913$ if we **standardize the residual** by the standard error, $\hat{\sigma}$, $\frac{29.913}{8.219} = 3.6395$. If this comes from the regression model the ratio should be distributed as t_4. A **high value** or correspondingly low prob-value indicates (with some degree of confidence) that this is an **outlier**. $\Pr(t_4 > 3.64) < .02$.

A common way of dealing with outliers (once they were identified by this procedure) is to remove them. The advantage is that the standard errors are smaller, the disadvantage is that maybe the true variance is large and what we thought was an outlier was simply an observation that resulted from the true large variance.

Another way of dealing with outliers is to replace OLS estimation by a method of estimation that is robust to outliers in the y direction. For example, LAD (least absolute deviations estimator). This estimator minimizes the sum of absolute deviations (not squared):

$$(\beta_0, \beta_1) = \arg\min \Sigma |y_i - a - bx_i| .$$

This is similar to median as opposed to sample average. The average is very sensitive to even one outlying observation, but the median is not.

Outliers in the x direction.

Otliers in the x direction are more difficult to detect because generally OLS benefits from the large spread of the X's.

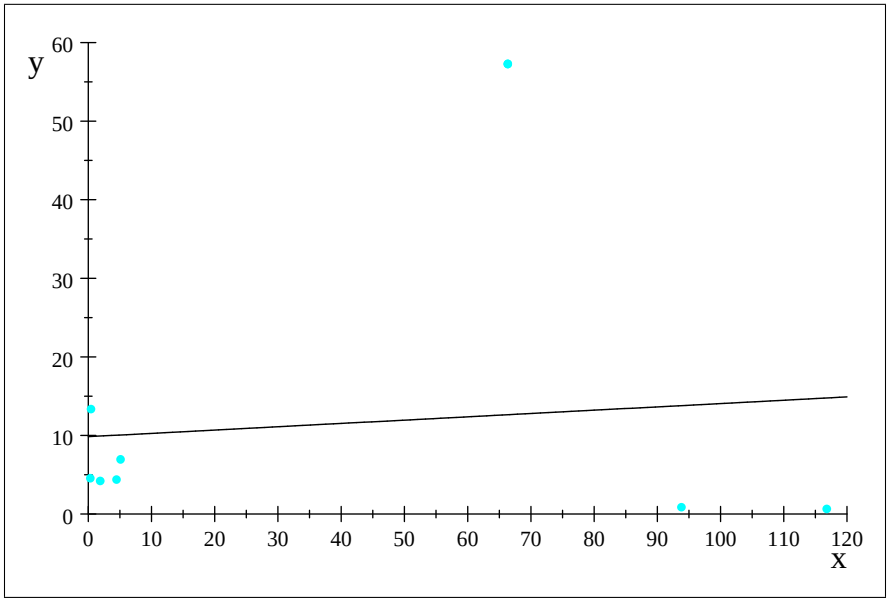
Example. How is brain weight in different species related to body weight?

Animal	Body weight, kg	Brain weight, gr
Mountain beaver	1.35	465
Cow	465	423
Grey wolf	36.33	119.5
Goat	27.66	115
Dipliodocus	11700	50
Asian elephant	2547	4603
Donkey	187.1	419
Horse	521	665
Guinea pig	1.04	5.5
Potar monkey	10	115
Cat	3.3	25.6
Giraffe	529	680
Gorilla	207	406
Human	62	1320
African elephant	6654	5712
Triceratops	9400	70
Rhesus monkey	6.8	179
Kangaroo	35	56
Chimpanzee	52.16	440
Brachiosaurus	87000	154.5
Pig	192	180

Regression for a subsample (including some dinosaurs - very high body weight):

$x = .01 \cdot body$	$y = .01brain$
4.65	4.23
117	.50
5.29	6.80
2.07	4.06
.62	13.20
94	.70
.5	4.40
66.54	57.12

$\hat{y} = 4.212\,3 \times 10^{-2} x + 9.845\,7$



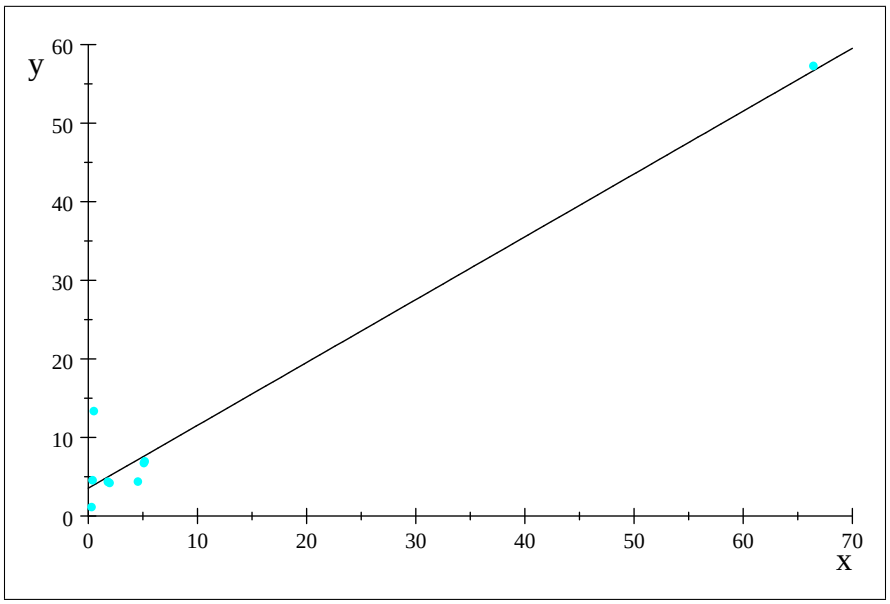
Because of the outliers (around X=100 and beyond which combine with very low y's - they correspond to the dinosaurs) the regression line does not pass close to any observations and is a poor representation of the majority of the sample.

Do dinosaurs not fit the general pattern of the other animal species?

Regression for a subsample (no dinosaurs - observations replaced).

4.65	4.23
5.29	6.80
2.07	4.06
.62	13.20
.5	4.40
5.2	6.6
.4	1
1.9	4.2
66.54	57.12

New regression line $\hat{y} = 0.799\,69x + 3.\,544\,6$



The regression line is close to the majority of observations. The outlier now possibly is humans (outlier in the y direction) with high brain weight relative to body. Note that the elephant with high body weight fits on quite well.

An alternative to removing outliers is to apply estimation methods that are robust to outliers. LAD is not robust to outliers in the X direction, but there are other methods.

Analysis of the model and extensions

We relax some of the assumptions of the model.

Here we consider the properties of the estimators conditionally on X.

Generally to have appropriate properties we need the X's to satisfy some assumptions, for example we may assume that $E\left(\frac{1}{\Sigma(X_i-\bar{X})^2}\right)$ exists for any n. It is the behavior of the random $\frac{1}{\Sigma(X_i-\bar{X})^2}$ that is important.

1. Random regressors.

$$y_i = \beta_0 + \beta_1 X_i + \varepsilon_i.$$

Consider models where the random (X_i, ε_i) are all distributed independently and identically for $i = 1, ..., n$.

Assumptions and results.

(a) the joint distribution $\{(X_i, \varepsilon_i)\}$ is such that

$$E(\varepsilon_i|X) = 0, E(\varepsilon_i\varepsilon_j|X) = \sigma^2 \text{ if } i = j, 0 \text{ otherwise.}$$

Consider $\hat{\beta}_1 = \frac{\Sigma(X_i-\bar{X})(y_i-\bar{y})}{\Sigma(X_i-\bar{X})^2} = \beta_1 + \frac{\Sigma(X_i-\bar{X})\varepsilon_i}{\Sigma(X_i-\bar{X})^2}$.

Then $E\left(\hat{\beta}_1\right) = \beta_1$, since $E\left(\frac{\Sigma\left(X_i-\bar{X}\right)\varepsilon_i}{\Sigma\left(X_i-\bar{X}\right)^2}\right) = E\left(\frac{\Sigma\left(X_i-\bar{X}\right)E(\varepsilon_i|X)}{\Sigma\left(X_i-\bar{X}\right)^2}\right) = 0$ (assuming the $X's$ are such that expectation exists).

So the estimator is still **unbiased**.

Variance.

$var\hat{\beta}_1 = E\left(\hat{\beta}_1 - \beta_1\right)^2 = E\left(\frac{\Sigma\left(X_i-\bar{X}\right)\varepsilon_i}{\Sigma\left(X_i-\bar{X}\right)^2}\right)^2 = E\left(\frac{\Sigma\left(X_i-\bar{X}\right)^2 E(\varepsilon_i^2|X)}{\left(\Sigma\left(X_i-\bar{X}\right)^2\right)^2}\right) = \sigma^2 E\left(\frac{1}{\Sigma\left(X_i-\bar{X}\right)^2}\right)$ (assuming expectation exists).

Conditionally on X the variance $var(\hat{\beta}_1|X)$ is the same as before:$\sigma^2\frac{1}{\Sigma\left(X_i-\bar{X}\right)^2}$

(b) Conditional distribution $\varepsilon_i|X$ is $N\left(0,\sigma^2\right)$.

Conditionally on X the estimator $\hat{\beta}_1$ is normally distributed since it is than a linear combination of normal variables, ε_i, i=1,...,n:

$N\left(\beta_1, \frac{\sigma^2}{\Sigma\left(X_i-\bar{X}\right)^2}\right)$.

All inference on the coefficients (confidence intervals, hypotheses tests) can then proceed as before.

3. **Weaker still: relaxing normality of errors:** the joint distribution $\{(X_i, \varepsilon_i)\}$ is such that

$$E(\varepsilon_i|X) = 0, E(\varepsilon_i\varepsilon_j|X) = \sigma^2 \text{ if } i = j, 0 \text{ otherwise.}$$

The unbiasedness result and the computation of variance for the coefficients does not change.

However, even conditionally on X the distribution is no longer normal.

It can be shown (with appropriate assumptions for X that typically hold, so we do not focus on them now), that the OLS estimator ia consistent and asymptotically normal, in other words for large enough sample size, the OLS estimator can be well approximated by the normal distribution, so that confidence test statistics have distributions that are well approximated by the usual distributions and confidence intervals can be constructed and hypotheses testing can proceed as before.

To show this two important types of theorem are needed, **Laws of Large Numbers** and **Central Limit Theorem**.

These theorems are about **sample averages**. For a sample $\{y_i\}_{i=1}^n$ from some distribution with mean μ, variance σ^2 consider the sample average, $\bar{y} = \frac{1}{n}\Sigma y_i$.

Law of large numbers will state that under some conditions:

$$\bar{y} - \mu \to_p 0.$$

This means that for any ε the value $\Pr\left(|\bar{y} - \mu| > \varepsilon\right) \to 0$ as $n \to \infty$.

So $\bar{y}$ converges to μ with high probability.

What is **the rate** of this convergence?

For example, $\frac{1}{n} \to 0$ and $\frac{1}{n^4} \to 0$, but $\frac{1}{n} >> \frac{1}{n^4}$ so $\frac{1}{n^4}$ converges to zero at a faster rate.

If you multiply $\frac{1}{n}$ by n it will not converge to zero any more (but $\frac{1}{n^4}$, even multiplied by n still converges to zero, need to multiply by n^4).

It turns out that $\bar{y}-\mu$ converges to zero at the rate $\frac{1}{\sqrt{n}}$. So multiplied by $\sqrt{n}$ it no longer converges to zero, but has some distribution; **central limit theorem** states that it converges to a normal distribution.

Central limit theorem will state that under some assumptions the distribution of $\sqrt{n}\left(\frac{\bar{y}-\mu}{\sigma}\right)$ is such that it converges to $N(0,1)$ - standard normal. The σ in the denominator can be replaced by the estimated value, s, without affecting convergence to normal.

Optional proofs.

1. Law of large numbers (LLN).

Denote $y_i-\mu$ by z_i.

Then z_i, $i=1,...,n$ is a random sample of **independent, identically distributed random variables from a distribution with mean 0 and variance** σ^2. Under the condition on z_i we prove LLN:

$$\Pr(|\bar{z}|>\varepsilon)\to 0 \text{ for any } \varepsilon \text{ as } n\to\infty.$$

By Chebyshev's theorem

$$\begin{aligned}\Pr(|\bar{z}|>\varepsilon) &\leq \frac{E\left(\frac{1}{n}\Sigma z_i\right)^2}{\varepsilon^2}=\frac{\frac{1}{n^2}\left[\Sigma_{i=1}^n E\left(z_i^2\right)+\Sigma_{i\neq j}E(z_iz_j)\right]}{\varepsilon^2}\\ &= \frac{\frac{1}{n}\sigma^2}{\varepsilon^2}\to 0 \text{ as } n\to\infty.\end{aligned}$$

■

2. Central limit theorem.

Define the Moment Generating Function for a random variable y: $M_y(t)=E(\exp(ty))$. There is a one-to-one relation between the distribution and the moment generating function.

An important theorem states: If for a sequence of random variables $\bar{y}_n$ and a random variable y the sequence of moment generating functions, $M_{\bar{y}_n}(t)$ converges to $M_y(t)$ as $n\to\infty$, then the distribution functions $F_{\bar{y}_n}(x)\to F_y(x)$.

We shall show that if y_i is i.i.d. and $E|y|^3$ exists, then the sequence of averages $\bar{y}_n=\frac{1}{n}\Sigma_{i=1}^n y_i$ is such that the distribution of $\sqrt{n}\frac{\bar{y}_n-E(y)}{(var(y))^{1/2}}$ converges to a standard normal distribution.

Let $\frac{y_i-E(y)}{(var(y))^{1/2}}=z_i$; $\bar{z}_n=\frac{1}{n}\Sigma z_i$. Note that $Ez_i=0; var(z_i)=1$.

We want to prove that the distribution of $\sqrt{n}\bar{z}_n$ converges to $N(0,1)$. We shall use the theorem about the moment generating function. In the first step we shall derive the limit of $M_{\sqrt{n}\bar{z}_n}(t)$ as $n\to\infty$. in the second step we shall derive the moment generating function for the standard normal variable and show that the limit forund in step 1 equals this function.

Step 1.

$M_{\sqrt{n}\bar{z}_n}(t) = E\left(\exp(t\frac{1}{n}\Sigma\sqrt{n}z_i)\right) = E\left(\Pi\exp\left(\frac{tz_i}{\sqrt{n}}\right)\right)$; expectation of a product of independent variables is the product of expectations:

$$E\left(\Pi\exp\left(\frac{tz_i}{\sqrt{n}}\right)\right) = \Pi E\left(\exp\left(\frac{tz_i}{\sqrt{n}}\right)\right) = \left(E\left(\exp\left(\frac{tz_i}{\sqrt{n}}\right)\right)\right)^n, \quad (4)$$

where the last equality uses the fact that all the variables z_i have identical distributions.

Next, write an expansion for the exponent

$$E\left(\exp\left(\frac{tz_i}{\sqrt{n}}\right)\right) = E\left(1 + \frac{tz_i}{\sqrt{n}} + \frac{1}{2!}\left(\frac{tz_i}{\sqrt{n}}\right)^2 + \frac{1}{n\sqrt{n}}R_n\right),$$

where by the remainder formula $|R_n| < a\left|tz_i\right|^3$ for some constant $a > 0$.

We get by substituting the expectations of z_i

$$\begin{aligned} E\left(1 + \frac{tz_i}{\sqrt{n}} + \frac{1}{2!}\left(\frac{tz_i}{\sqrt{n}}\right)^2 + R_n\right) &= \left(1 + \frac{t^2}{2}\frac{1}{n} + \frac{1}{n\sqrt{n}}ER_n\right) \\ &= 1 + \frac{1}{n}(\frac{t^2}{2} + \frac{1}{\sqrt{n}}ER_n). \end{aligned}$$

We need to take this to the power n for (4)

$$\begin{aligned} &\left(1 + \frac{1}{n}(\frac{t^2}{2} + \frac{1}{\sqrt{n}}ER_n)\right)^n \\ = &\left(\left(1 + \frac{1}{n}(\frac{t^2}{2} + \frac{1}{\sqrt{n}}ER_n)\right)^{n(\frac{t^2}{2} + \frac{1}{\sqrt{n}}ER_n)^{-1}}\right)^{(\frac{t^2}{2} + \frac{1}{\sqrt{n}}ER_n)}. \end{aligned}$$

Now to take limits: as $n \to \infty$ the expression $w = \frac{1}{n}(\frac{t^2}{2} + \frac{1}{\sqrt{n}}ER_n)$ converges to zero, $w^{-1} \to \infty$, then by the definition of the base of natural logarithms,

$$\left(1 + \frac{1}{n}(\frac{t^2}{2} + \frac{1}{\sqrt{n}}ER_n)\right)^{n(\frac{t^2}{2} + \frac{1}{\sqrt{n}}ER_n)^{-1}} = (1 + w)^{\frac{1}{w}} \to e.$$

At the same time $\frac{t^2}{2} + \frac{1}{\sqrt{n}}ER_n \to \frac{t^2}{2}$ since ER_n is bounded and $\frac{1}{\sqrt{n}} \to 0$.

We get then that $M_{\sqrt{n}\bar{z}_n}(t) \to \exp\left(\frac{t^2}{2}\right)$.

Step 2.

Now consider the moment generating function for the standard normal variable, x :

$$\begin{aligned} E\left(\exp(tx)\right) &= \int \frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}\left(-2tx + x^2\right))dx \\ &= \frac{1}{\sqrt{2\pi}}\int \exp(-\frac{1}{2}\left(-2tx + x^2 + t^2\right))\exp\left(\frac{1}{2}t^2\right)dx \\ &= \exp\left(\frac{1}{2}t^2\right)\frac{1}{\sqrt{2\pi}}\int \exp(-\frac{1}{2}w^2)dw = \exp\left(\frac{1}{2}t^2\right), \end{aligned}$$

by change of variable $w = (x - t)$ and since $\frac{1}{\sqrt{2\pi}} \int \exp(-\frac{1}{2}w)dw = 1$.

So for $N(0,1)$ the moment generating function is $M_{N(0,1)}(t) = \exp\left(\frac{t^2}{2}\right)$.■

Why are the theorems about **sample averages** useful?

Take the OLS estimator, $\hat{\beta}_1 = \frac{\Sigma(X_i - \bar{X})(y_i - \bar{y})}{\Sigma(X_i - \bar{X})^2} = \frac{\frac{1}{n}\Sigma(X_i - \bar{X})(y_i - \bar{y})}{\frac{1}{n}\Sigma(X_i - \bar{X})^2}$. This is a ratio of two averages.

$$\hat{\beta}_1 = \beta_1 + \frac{\frac{1}{n}\Sigma(X_i - \bar{X})\varepsilon_i}{\frac{1}{n}\Sigma(X_i - \bar{X})^2}.$$

We know that $\{X_i, \varepsilon_i\}$ are i.i.d.,$E(X_i - \bar{X})\varepsilon_i = 0$. Then by Law of Large Numbers the average $\frac{1}{n}\Sigma(X_i - \bar{X})\varepsilon_i$ converges in probability to zero.

The denominator $\frac{1}{n}\Sigma(X_i - \bar{X})^2$ by Law of Large Numbers converges to variance of X (assuming it exists), that is $varX > 0$.

Then since the ratio $\frac{\frac{1}{n}\Sigma(X_i - \bar{X})\varepsilon_i}{\frac{1}{n}\Sigma(X_i - \bar{X})^2}$ converges on probability to zero, $\hat{\beta}_1$ converges in probability to β_1, it is a **consistent** estimator.

Suppose that the Central Limit Theorem applies to $z_i = \frac{(X_i - \bar{X})\varepsilon_i}{var\left((X_i - \bar{X})\varepsilon_i\right)}$. Then it can be shown that the distribution of$\sqrt{n}\left(\hat{\beta} - \beta\right)$ converges to a normal distribution $N\left(0, p\lim \frac{\sigma^2}{\frac{1}{n}\Sigma\left(X_i - \bar{X}\right)^2}\right)$. Here the asymptotic variance (variance of the asymptotic normal distribution, $p\lim \frac{\sigma^2}{\frac{1}{n}\Sigma\left(X_i - \bar{X}\right)^2}$, as before can be estimated by $\frac{s^2}{\frac{1}{n}\Sigma\left(X_i - \bar{X}\right)^2}$.

In many other problems estimators and statistics are functions of some averages and similar analysis applies.

Checking that assumptions hold.

Assumptions required that the variances of the errors exist and be the same, σ^2, and the all the error covariances be zero.

One could examine the plots of the residuals to see whether these assumptions are plausible. (There are also formal tests).

Normality of a distribution and boundedness of moments.

Why is normality important? If the data is generated by a normal, then the distributions of the estimators, test statistics are known exactly in many cases (e.g. the sample mean has a normal distribution, the t-ratio is exactly Student's t etc.); there is no need to worry about how good the asymptotic approximation is.

Some economic models postulate normality for example latent utility model: $y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, with ε_i normally distributed.

Violations of normality are common for financial variables; incorrectly assuming normality may lead to very misleading results.

Tests of normality can be applied to regression residuals.