Section 15

Multiple linear regression.

Let us consider a model

$$Y_i = \beta_1 X_{i1} + \ldots + \beta_p X_{ip} + \varepsilon_i$$

where random noise variables $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. $N(0, \sigma^2)$. We can write this in a matrix form

$$Y = X\beta + \varepsilon,$$

where Y and ε are $n \times 1$ vectors, β is $p \times 1$ vector and X is $n \times p$ matrix. We will denote the columns of matrix X by $X_1, \ldots X_p$, i.e.

$$X = (X_1, \ldots, X_p)$$

and we will assume that these columns are linearly independent. If they are not linearly independent, we can not reconstruct parameters β from X and Y even if there is no noise ε. In simple linear regression this would correspond to all Xs being equal and we can not estimate a line from observations only at one point. So from now on we will assume that $n > p$ and the rank of matrix X is equal to p. To estimate unknown parameters β and σ we will use maximum likelihood estimators.

Lemma 1. *The MLE of β and σ^2 are given by:*

$$\hat{\beta} = (X^T X)^{-1} X^T Y \quad \textit{and} \quad \hat{\sigma}^2 = \frac{1}{n}|Y - X\hat{\beta}|^2 = \frac{1}{n}|Y - X(X^T X)^{-1} X^T Y|^2.$$

Proof. The p.d.f. of Y_i is

$$f_i(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\Big(-\frac{1}{2\sigma^2}(x - \beta_1 X_{i1} - \ldots - \beta_p X_{ip})^2\Big)$$

and, therefore, the likelihood function is

$$\begin{aligned} \prod_{i=1}^n f_i(Y_i) &= \Big(\frac{1}{\sqrt{2\pi}\sigma}\Big)^n \exp\Big(-\frac{1}{2\sigma^2}\sum_{i=1}^n (Y_i - \beta_1 X_{i1} - \ldots - \beta_p X_{ip})^2\Big) \\ &= \Big(\frac{1}{\sqrt{2\pi}\sigma}\Big)^n \exp\Big(-\frac{1}{2\sigma^2}|Y - X\beta|^2\Big). \end{aligned}$$

To maximize the likelihood function, first, we need to minimize $|Y - X\beta|^2$. If we rewrite the norm squared using scalar product:

$$\begin{aligned} |Y - X\beta|^2 &= (Y - \sum_{i=1}^{p} \beta_i X_i, Y - \sum_{i=1}^{p} \beta_i X_i) \\ &= (Y,Y) - 2\sum_{i=1}^{p} \beta_i (Y, X_i) + \sum_{i,j=1}^{p} \beta_i \beta_j (X_i, X_j). \end{aligned}$$

Then setting the derivatives in each β_i equal to zero

$$-2(Y, X_i) + 2\sum_{j=1}^{p} \beta_j (X_i, X_j) = 0$$

we get

$$(Y, X_i) = \sum_{i=1}^{p} \beta_j (X_i, X_j) \quad \text{for all} \quad i \le p.$$

In matrix notations this can be written as $X^T Y = X^T X\beta$. Matrix $X^T X$ is a $p \times p$ matrix. Is is invertible since by assumption X has rank p. So we can solve for β to get the MLE

$$\hat{\beta} = (X^T X)^{-1} X^T Y.$$

It is now easy to minimize over σ to get

$$\hat{\sigma}^2 = \frac{1}{n}|Y - X\hat{\beta}|^2 = \frac{1}{n}|Y - X(X^T X)^{-1} X^T Y|^2.$$

□

To do statistical inference we need to compute the joint distribution of these estimates. We will prove the following.

Theorem. *We have*

$$\hat{\beta} \sim N\Big(\beta, \sigma^2 (X^T X)^{-1}\Big), \;\; \frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p}$$

and estimates $\hat{\beta}$ and $\hat{\sigma}^2$ are independent.

Proof. First of all, let us rewrite the estimates in terms of random noise ε using $Y = X\beta + \varepsilon$. We have

$$\begin{aligned} \hat{\beta} &= (X^T X)^{-1} X^T Y = (X^T X)^{-1} X^T (X\beta + \varepsilon) \\ &= (X^T X)^{-1}(X^T X)\beta + (X^T X)^{-1} X^T \varepsilon = \beta + (X^T X)^{-1} X^T \varepsilon \end{aligned}$$

and since

$$\begin{aligned} Y - X(X^T X)^{-1} X^T Y &= X\beta + \varepsilon - X(X^T X)^{-1} X^T (X\beta + \varepsilon) \\ &= X\beta + \varepsilon - X\beta - X(X^T X)^{-1} X^T \varepsilon = (I - X(X^T X)^{-1} X^T)\varepsilon \end{aligned}$$

we have

$$\hat{\sigma}^2 = \frac{1}{n}|(I - X(X^TX)^{-1}X^T)\varepsilon|^2.$$

Since $\hat{\beta}$ is a linear transformation of a normal vector ε it will also be normal with mean

$$\mathbb{E}\hat{\beta} = \mathbb{E}(\beta + (X^TX)^{-1}X^T\varepsilon) = \beta$$

and covariance matrix

$$\begin{aligned}
\mathbb{E}(\hat{\beta}-\beta)(\hat{\beta}-\beta)^T &= \mathbb{E}(X^TX)^{-1}X^T\varepsilon\varepsilon^TX(X^TX)^{-1} \\
&= (X^TX)^{-1}X^T\mathbb{E}\varepsilon\varepsilon^TX(X^TX)^{-1} \\
&= (X^TX)^{-1}X^T(\sigma^2 I)X(X^TX)^{-1} \\
&= \sigma^2(X^TX)^{-1}(X^TX)(X^TX)^{-1} = \sigma^2(X^TX)^{-1}.
\end{aligned}$$

This proves that $\hat{\beta} \sim N(\beta, \sigma^2(X^TX)^{-1})$. To prove that $\hat{\beta}$ and $\hat{\sigma}^2$ are independent and to find the distribution of $n\hat{\sigma}^2/\sigma^2$ we will use the following trick. This trick can also be very useful computationally since it will relate all quantities of interest expressed in terms of $n \times p$ matrix X to quantities expressed in terms of a certain $p \times p$ matrix R which can be helpful when n is very large compared to p. We would like to manipulate the columns of matrix X to make them orthogonal to each other, which can be done by Gram-Schmidt orthogonalization. In other words, we want to represent matrix X as

$$X = X_0R$$

where X_0 is $n \times p$ matrix with columns $X_0^1, \ldots, X_0^p$ that are orthogonal to each other and, moreover, form an orthonormal basis, and matrix R is $p \times p$ invertible (and upper triangular) matrix. In Matlab this can be done using economy size QR factorization

```
[X0, R]=qr(X,0).
```

The fact that columns of X_0 are orthonormal implies that

$$X_0^TX_0 = I$$

- a $p \times p$ identity matrix. Let us replace X by X_0R everywhere in the estimates. We have

$$(X^TX)^{-1}X^T = (R^TX_0^TX_0R)^{-1}R^TX_0^T = (R^TR)^{-1}R^TX_0^T = R^{-1}(R^T)^{-1}R^TR^{-1}X_0^T,$$

$$X(X^TX)^{-1}X^T = X_0R(R^TX_0^2X_0R)^{-1}R^TX_0^T = X_0RR^{-1}(R^T)^{-1}R^TX_0^T = X_0X_0^T.$$

As a result

$$\hat{\beta} - \beta = R^{-1}X_0^T\varepsilon \quad \text{and} \quad n\hat{\sigma}^2 = |(I - X_0X_0^T)\varepsilon|^2. \tag{15.0.1}$$

By construction p columns of X_0, which are also the rows of X_0^T, are orthonormal. Therefore, we can choose the last $n - p$ rows of a $n \times n$ matrix

$$A = \begin{pmatrix} X_0^T \\ \cdots \end{pmatrix}$$

to make A an orthogonal matrix, we just need to choose them to complete, together with rows of X_0^T, the orthonormal basis in $\mathbb{R}^n$. Let us define a vector

$$g = A\varepsilon, \text{ i.e. } \begin{pmatrix} g_1 \\ g_2 \\ \vdots \\ g_n \end{pmatrix} = \begin{pmatrix} X_0^T \\ \cdots \end{pmatrix} \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Since ε is a vector of i.i.d. standard normal, we proved before that its orthogonal transformation g will also be a vector of independent $N(0,\sigma^2)$ random variables $g_1,\dots,g_n$. First of all, since

$$\hat{g} := \begin{pmatrix} g_1 \\ \vdots \\ g_p \end{pmatrix} = X_0^T \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix},$$

we have

$$\hat{\beta} - \beta = R^{-1} X_0^T \varepsilon = R^{-1} \begin{pmatrix} g_1 \\ \vdots \\ g_p \end{pmatrix} = R^{-1}\hat{g}. \tag{15.0.2}$$

Next, we will prove that

$$|(I - X_0 X_0^T)\varepsilon|^2 = g_{p+1}^2 + \ldots + g_n^2. \tag{15.0.3}$$

First of all, orthogonal transformation preserves lengths, so $|g|^2 = |A\varepsilon|^2 = |\varepsilon|^2$. On the other hand, let us write $|\varepsilon|^2 = \varepsilon^T\varepsilon$ and break ε into a sum of two terms

$$\varepsilon = X_0 X_0^T \varepsilon + (I - X_0 X_0^T)\varepsilon.$$

Then we get

$$|g|^2 = |\varepsilon|^2 = \varepsilon^T\varepsilon = \Big(\varepsilon^T X_0 X_0^T + \varepsilon^T(I - X_0 X_0^T)\Big)\Big(X_0 X_0^T\varepsilon + (I - X_0 X_0^T)\varepsilon\Big).$$

When we multiply all the terms out we will use that $X_0^T X_0 = I$ since the matrix $X_0^T X_0$ consists of scalar products of columns of X_0 which are orthonormal. This also implies that

$$X_0 X_0^T (I - X_0 X_0^T) = X_0 X_0^T - X_0 I X_0^T = 0.$$

Using this we get

$$\begin{aligned} |g|^2 = |\varepsilon|^2 &= \varepsilon^T X_0 X_0^T \varepsilon + \varepsilon^T (I - X_0 X_0^T)(I - X_0 X_0^T)\varepsilon \\ &= |X_0^T\varepsilon|^2 + |(I - X_0 X_0^T)\varepsilon|^2 = |\hat{g}|^2 + |(I - X_0 X_0^T)\varepsilon|^2 \end{aligned}$$

because $\hat{g} = X_0^T\varepsilon$ so we finally proved that

$$|(I - X_0 X_0^T)\varepsilon|^2 = |g|^2 - |\hat{g}|^2 = g_1^2 + \ldots + g_n^2 - g_1^2 - \ldots - g_p^2 = g_{p+1}^2 + \ldots + g_n^2$$

which is (15.0.3). This proves that $n\hat{\sigma}^2/\sigma^2 \sim \chi^2_{n-p}$ and it is also independent of $\hat{\beta}$ which depends only on $g_1,\dots,g_p$ by (15.0.2).

□

Let us for convenience write down equation (15.0.2) as a separate result.

Lemma 2. *Given a decomposition $X = X_0 R$ with $n \times p$ matrix X_0 with orthonormal columns and invertible (upper triangular) $p \times p$ matrix R we can represent*

$$\hat{\beta} - \beta = R^{-1}\hat{g} = R^{-1} \begin{pmatrix} g_1 \\ \vdots \\ g_p \end{pmatrix}$$

for independent $N(0, \sigma^2)$ random variables $g_1, \ldots, g_p$.

Confidence intervals and t-tests for linear combination of parameters β. Let us consider a linear combination

$$c_1\beta_1 + \ldots + c_p\beta_p = c^T\beta$$

where $c = (c_1, \ldots, c_p)^T$. To construct confidence intervals and t-tests for this linear combination we need to write down a distribution of $c^T\hat{\beta}$. Clearly, it has a normal distribution with mean $\mathbb{E}c^T\hat{\beta} = c^T\beta$ and variance

$$\mathbb{E}(c^T(\hat{\beta} - \beta))^2 = \mathbb{E}c^T(\hat{\beta} - \beta)(\hat{\beta} - \beta)^T c = c^T \mathrm{Cov}(\hat{\beta})c = \sigma^2 c^T (X^T X)^{-1} c.$$

Therefore,

$$\frac{c^T(\hat{\beta} - \beta)}{\sqrt{\sigma^2 c^T (X^T X)^{-1} c}} \sim N(0, 1)$$

and using that $n\hat{\sigma}^2/\sigma^2 \sim \chi^2_{n-p}$ we get

$$\frac{c^T(\hat{\beta} - \beta)}{\sqrt{\sigma^2 c^T (X^T X)^{-1} c}} \Big/ \sqrt{\frac{1}{n-p}\frac{n\hat{\sigma}^2}{\sigma^2}} = c^T(\hat{\beta} - \beta)\sqrt{\frac{n-p}{n\hat{\sigma}^2 c^T (X^T X)^{-1} c}} \sim t_{n-p}.$$

To obtain the distribution of one parameter $\hat{\beta}_i$ we need to choose a vector c that has all zeros and 1 in the ith coordinate. Then we get

$$(\hat{\beta}_i - \beta_i)\sqrt{\frac{n-p}{n\hat{\sigma}^2((X^T X)^{-1})_{ii}}} \sim t_{n-p}.$$

Here $((X^T X)^{-1})_{ii}$ is the ith diagonal element of the matrix $(X^T X)^{-1}$. This is a good time to mention how the quality of estimation of β depends on the choice of X. For example, we mentioned before that the columns of X should be linearly independent. What happens if some of them are nearly collinear? Then some eigenvalues of $(X^T X)$ will be 'small' (in some sense) and some eigenvalues of $(X^T X)^{-1}$ will be 'large'. (Small and large here are relative terms because the size of the matrix also grows with n.) As a result, the confidence intervals for some parameters will get very large too which means that their estimates are not very accurate. To improve the quality of estimation we need to avoid using collinear predictors. We will see this in the example below.

□

Joint confidence set for β and F-test. By Lemma 2, $R(\hat{\beta}-\beta)=\hat{g}$ and, therefore,

$$g_1^2+\ldots+g_p^2=|\hat{g}|^2=\hat{g}^T\hat{g}=(\hat{\beta}-\beta)^TR^TR(\hat{\beta}-\beta)=(\hat{\beta}-\beta)^TX^TX(\hat{\beta}-\beta).$$

Since $g_i\sim N(0,\sigma^2)$ this proves that

$$\frac{(\hat{\beta}-\beta)^TX^TX(\hat{\beta}-\beta)}{\sigma^2}\sim\chi^2_p.$$

Using that $n\hat{\sigma}^2/\sigma^2\sim\chi^2_{n-p}$ gives

$$\frac{(\hat{\beta}-\beta)^TX^TX(\hat{\beta}-\beta)}{p\sigma^2}\Big/\frac{n\hat{\sigma}^2}{(n-p)\sigma^2}=\frac{(n-p)}{np\hat{\sigma}^2}(\hat{\beta}-\beta)^TX^TX(\hat{\beta}-\beta)\sim F_{p,n-p}.$$

If we take c such that $F_{p,n-p}(0,c_\alpha)=\alpha$ then

$$\frac{(n-p)}{np\hat{\sigma}^2}(\hat{\beta}-\beta)^TX^TX(\hat{\beta}-\beta)\le c_\alpha \tag{15.0.4}$$

defines a joint confidence set for all parameters β simultaneously with confidence level α.

Suppose that we want to test a hypothesis about all parameters simultaneously, for example,

$$H_0:\beta=\beta_0.$$

Then we consider a statistic

$$F=\frac{(n-p)}{np\hat{\sigma}^2}(\hat{\beta}-\beta_0)^TX^TX(\hat{\beta}-\beta_0), \tag{15.0.5}$$

which under null hypothesis has $F_{p,n-p}$ distribution, and define a decision rule by

$$\delta=\begin{cases} H_0: & F\le c \\ H_1: & F>c, \end{cases}$$

where a threshold c is determined by $F_{p,n-p}(c,\infty)=\alpha$ - a level of significance. Of course, this test is equivalent to checking if vector β_0 belongs to a confidence set (15.0.4)! (We just need to remember that confidence level = 1 - level of significance.)

□

Simultaneous confidence set and F-test for subsets of β. Let

$$s=\{i_1,\ldots,i_k\}\subseteq\{1,\ldots,p\}$$

be s subset of size $k\le p$ of indices $\{1,\ldots,p\}$ and let $\beta_s=(\beta_{i_1},\ldots,\beta_{i_k})^T$ be a vector that consists of the corresponding subset of parameters β. Suppose that we would like to test the hypothesis

$$H_0:\beta_s=\beta_s^0$$

for some given vector β_s^0, for example, $\beta_s^0=0$. Let $\hat{\beta}_s$ be a corresponding vector of estimates. Let

$$\Sigma_s=\Big((X^TX)^{-1}_{i,j}\Big)_{i,j\in s}$$

be a $k \times k$ submatrix of $(X^T X)^{-1}$ with row and column indices in the set s. By the above Theorem, the joint distribution of $\hat{\beta}_s$ is

$$\hat{\beta}_s \sim N(\beta_s, \sigma^2 \Sigma_s).$$

Let $A = \Sigma_s^{1/2}$, i.e. A is a symmetric $k \times k$ matrix such that $\Sigma_s = AA^T$. As a result, a centered vector of estimates can be represented as

$$\hat{\beta}_s - \beta_s = Ag,$$

where $g = (g_1, \ldots, g_k)^T$ are independent $N(0, \sigma^2)$. Therefore, $g = A^{-1}(\hat{\beta}_s - \beta_s)$ and the rest is similar to the above argument. Namely,

$$\begin{aligned} g_1^2 + \ldots + g_k^2 &= |g|^2 = g^T g = (\hat{\beta}_s - \beta_s)^T (A^{-1})^T A^{-1} (\hat{\beta}_s - \beta_s) \\ &= (\hat{\beta}_s - \beta_s)^T (AA^T)^{-1} (\hat{\beta}_s - \beta_s) = (\hat{\beta}_s - \beta_s)^T \Sigma_s^{-1} (\hat{\beta}_s - \beta_s) \sim \sigma^2 \chi_k^2. \end{aligned}$$

As before we get

$$F = \frac{(n-p)}{nk\hat{\sigma}^2} (\hat{\beta}_s - \beta_s)^T \Sigma_s^{-1} (\hat{\beta}_s - \beta_s) \sim F_{k,n-p}$$

and we can now construct a simultaneous confidence set and F-tests.

□

Remark. Matlab regression function 'regress' assumes that a matrix X of explanatory variables will contain a first column of ones that corresponds to an "intercept" parameter β_1. The F-statistic output by 'regress' corresponds to F-test about all other "slope" parameters:

$$H_0 : \beta_2 = \ldots = \beta_p = 0.$$

In this case $s = \{2, 3, \ldots, p\}$, $k = p - 1$ and

$$F = \frac{(n-p)}{n(p-1)\hat{\sigma}^2} \hat{\beta}_s^T \Sigma_s^{-1} \hat{\beta}_s \sim F_{p-1,n-p}.$$

□

Example. Let us take a look at the 'cigarette' dataset from previous lecture. We saw that tar, nicotine and carbon monoxide content are positively correlated and any pair is well described by a simple linear regression. Suppose that we would like to predict carbon monoxide as a linear function of both tar and nicotine content. We create a 25×3 matrix X:

```
X=[ones(25,1),tar,nic];
```

We introduce a first column of ones to allow an intercept parameter β_1 in our multiple linear regression model:

$$\mathrm{CO}_i = \beta_1 + \beta_2 \mathrm{Tar}_i + \beta_3 \mathrm{Nicotin}_i + \varepsilon_i.$$

If we perform a multiple linear regression:

```
[b,bint,r,rint,stats] = regress(carb,X);
```

We get the estimates of parameters and 95% confidence intervals for each parameter

```
b = 3.0896      bint = 1.3397   4.8395
    0.9625               0.4717   1.4533
   -2.6463             -10.5004   5.2079
```

and, in order, R^2-statistic, F-statistic from (15.0.5), p-value for this statistic

$$F_{p,n-p}(F,+\infty) = F_{3,25-3}(F,+\infty)$$

and the estimate of variance $\hat{\sigma}^2$:

```
stats = 0.9186  124.1102  0.000  1.9952.
```

First of all, we see that high R^2 means that linear model explain most of the variability in the data and small p-value means that we reject the hypothesis that all parameters are equal to zero. On the other hand, simple linear regression showed that carbon monoxide had a positive correlation with nicotine and now we got $\hat{\beta}_3 = -2.6463$. Also, notice that the confidence interval for β_3 is very poor. The reason for this is that tar and nicotine are nearly collinear. Because of this the matrix

$$(X^T X)^{-1} = \begin{pmatrix} 0.3568 & 0.0416 & -0.9408 \\ 0.0416 & 0.0281 & -0.4387 \\ -0.9408 & -0.4387 & 7.1886 \end{pmatrix}$$

has relatively large last diagonal diagonal value. We recall that Theorem gives that the variance of estimate $\hat{\beta}_3$ is $7.1886\sigma^2$ and we also see that the estimate of σ^2 is $\hat{\sigma}^2 = 1.9952$. As a result the confidence interval for β_3 is rather poor.

Of course, looking at linear combinations of tar and nicotine as new predictors does not make sense because they lose their meaning, but for the sake of illustrations let us see what would happen if our predictors were not nearly collinear but, in fact, orthonormal. Let us use economic QR decomposition

```
[X0,R]=qr(X,0)
```

a new matrix of predictor X_0 with orthonormal columns that are some linear combinations of tar and nicotine. Then regressing carbon monoxide on these new predictors

```
 [b,bint,r,rint,stats] = regress(carb,X0);
```

we would get

```
b = -62.6400     bint = -65.5694  -59.7106
    -22.2324             -25.1618  -19.3030
     0.9870              -1.9424    3.9164
```

all confidence intervals of the same relatively better size.

□

Example. The following data presents per capita income of 20 countries for 1960s. Also presented are the percentages of labor force employed in agriculture, industry and service for each country. (Data source: lib.stat.cmu.edu/DASL/Datafiles/oecdat.html)

```
COUNTRY     PCINC AGR IND SER

CANADA       1536  13  43  45
SWEEDEN      1644  14  53  33
SWITZERLAND  1361  11  56  33
LUXEMBOURG   1242  15  51  34
U. KINGDOM   1105  4   56  40
DENMARK      1049  18  45  37
W. GERMANY   1035  15  60  25
FRANCE       1013  20  44  36
BELGUIM      1005  6   52  42
NORWAY       977   20  49  32
ICELAND      839   25  47  29
NETHERLANDS  810   11  49  40
AUSTRIA      681   23  47  30
IRELAND      529   36  30  34
ITALY        504   27  46  28
JAPAN        344   33  35  32
GREECE       324   56  24  20
SPAIN        290   42  37  21
PORTUGAL     238   44  33  23
TURKEY       177   79  12  9
```

We can perform simple linear regression of income on each of the other explanatory variables or multiple linear regression on any pair of the explanatory variables. Fitting simple linear regression of income vs. percent of labor force in agriculture, industry and service:

```
polytool(agr,income,1),
```

etc., produces figure 15.1. Next, we perform statistical inference using 'regress' function. Statistical analysis of linear regression fit of income vs. percent of labor force in agriculture:

```
[b,bint,r,rint,stats]=regress(income,[ones(20,1),agr])

b = 1317.9    bint = 1094.7   1541.1
     -18.9             -26.0   -11.7

stats =   0.6315  30.8472  2.8e-005  74596.
```

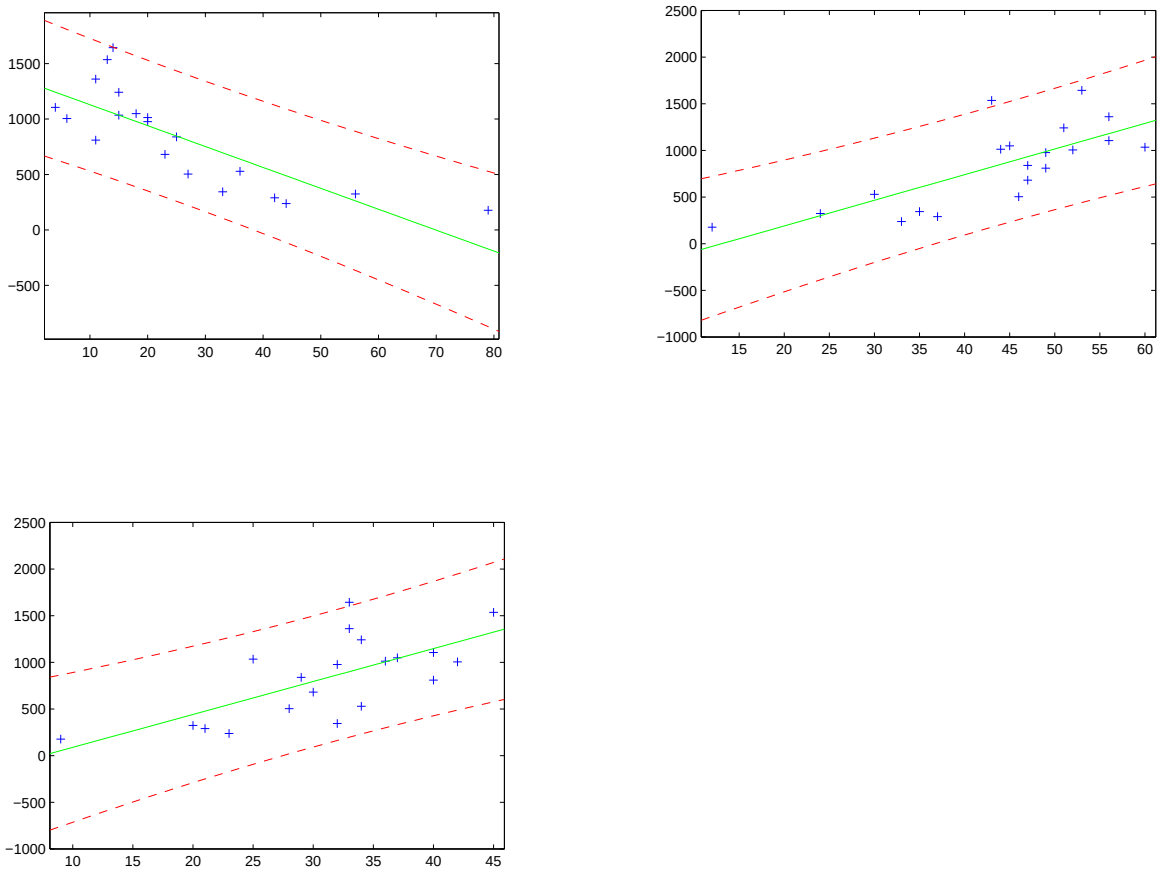

Figure 15.1: Linear regression of income on percent of labor force in agriculture, industry and service.

For income vs. percent of labor force in industry

```
[b,bint,r,rint,stats]=regress(income,[ones(20,1),ind]);

b = -359.3115    bint = -907.1807   188.5577
    27.4905                15.3058   39.6751

stats = 0.5552   22.4677   0.0002   90042
```

and for income vs. labor force in service

```
[b,bint,r,rint,stats]=regress(income,[ones(20,1),serv]);

b = -264.5199     bint =  -858.0257  328.9858
      35.3024                16.8955   53.7093
```

```
stats =   0.4742  16.2355   0.0008   106430.
```

We see that in all three cases, the hypotheses that parameters of least-squares line are both zero can be rejected at conventional level of significance $\alpha = 0.05$. Looking at the confidence intervals for the estimates of slopes we observe that the correlation of income with percent of labor force in agriculture is negative, and other two correlations are positive.

We can also perform a multiple regression on any two explanatory variables. We can not perform multiple linear regression with all three explanatory variables because they add up to 100%, i.e. they are linearly dependent. If we create a predictor matrix

```
X=[ones(20,1),agr,ind];
```

and perform multiple linear regression

```
[b,bint,r,rint,stats]=regress(income,X);
```

we get

```
b = 1272.1    bint = -632.6   3176.9
     -18.4            -39.1      2.3
       0.8            -31.4     32.9

stats =  0.6316  14.5703  0.0002  78972
```

Of course, one can find many shortcomings of this model. For example, having the entire population in agriculture results in prediction of $1272.1 - 1840 < 0$ negative income per capita.

□