2.6 Inference about the regression parameters

Example 2.5. Overheads.
A company builds custom electronic instruments and computer components. All jobs are manufactured to customer specifications. The firm wants to be able to estimate its overhead cost. As part of a preliminary investigation, the firm decides to focus on a particular department and investigates the relationship between total departmental overhead cost (Y) and total direct labor hours (X). The data for the most recent 16 months are plotted in Figure 2.14.

Two objectives of this investigation are

1. to summarize for management the relationship between total departmental overhead and total direct labor hours.

2. to estimate the expected and to predict the actual total departmental overhead from the total direct labor hours.

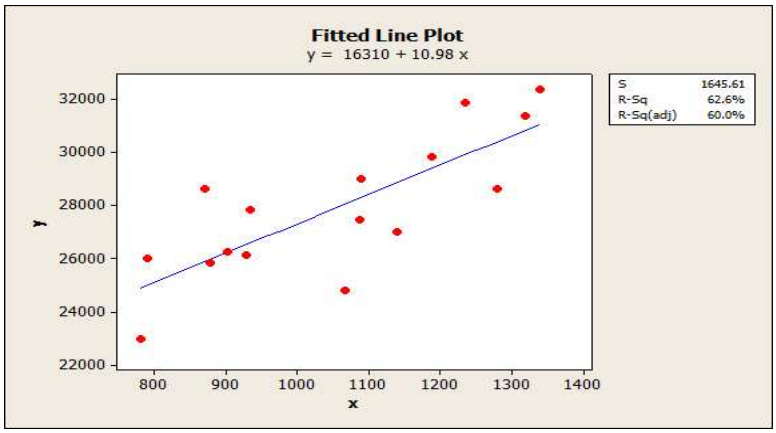

Figure 2.14: Plot of overheads data

```
The regression equation is
Ovhd = 16310 + 11.0 Labor

Predictor     Coef    SE Coef      T      P
Constant     16310   2421       6.74   0.000
Labor        10.982  2.268      4.84   0.000

S = 1645.61   R-Sq = 62.6%   R-Sq(adj) = 60.0%
```

```
Analysis of Variance
Source          DF       SS          MS      F      P
Regression       1     63517077   63517077  23.46  0.000
Residual Error  14     37912232    2708017
Total           15    101429309

Unusual Observations
Obs   Labor   Ovhd    Fit    SE Fit  Residual  St Resid
  6   1067    24817  28028    413     -3211     -2.02R

R denotes an observation with a large standardized residual.
```

Comments:

- The model fit is $\widehat{y}_i = 16310 + 11x_i$. There is a significant relationship between the overheads and the labor hours ($p < 0.001$ in ANOVA).
- The increase of labor hours by 1 will increase the mean overheads by about £11 ($\widehat{\beta}_1 = 11.0$).
- There is rather large variability in the data; the percentage of total variation explained by the model is rather small ($R^2 = 62.6$).

The model allows us to estimate the total overhead cost as a function of labour hours, but as we noticed, there is large variability in the data. In such a case, the point estimates may not be very reliable. Anyway, point estimates should always be accompanied by their standard errors. Then we can also find confidence intervals (CI) for the unknown model parameters, or test their non-significance.□

Note that for the simple linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \text{where } \varepsilon_i \underset{iid}{\sim} N(0, \sigma^2), \tag{2.13}$$

we obtained the following LSE of the parameters β_0 and β_1:

$$\widehat{\beta}_0 = \bar{Y} - \widehat{\beta}_1 \bar{x}$$
$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

We now derive results which allow us to make inference about the regression parameters and predictions.

2.6.1 Inference about β_1

We proved the following result in Section 2.3.

Theorem 2.3. *In the full simple linear regression model (SLRM) the distribution of the LSE of β_1, $\widehat{\beta}_1$, is normal with the expectation $\mathrm{E}(\widehat{\beta}_1) = \beta_1$ and the variance $\mathrm{var}(\widehat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$, that is*

$$\widehat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right). \tag{2.14}$$

□

Remark 2.3. For large samples, where there is no assumption of normality of Y_i, the sampling distribution of $\widehat{\beta}_1$ is approximately normal. □

Theorem 2.3 allows us to derive a confidence interval (CI) for β_1 and a test of non-significance for β_1. After standarisation of $\widehat{\beta}_1$ we obtain

$$\frac{\widehat{\beta}_1 - \beta_1}{\sigma/\sqrt{S_{xx}}} \sim \mathcal{N}(0,1).$$

However, the error variance is usually not known and it is replaced by its estimator. Then the normal distribution changes to a Student t-distribution. The explanation is following.

Lemma 2.1. *If $Z \sim N(0,1)$ and $U \sim \chi^2_\nu$, and Z and U are independent, then*

$$\frac{Z}{\sqrt{U/\nu}} \sim t_\nu.$$

□

Here we have,

$$Z = \frac{\widehat{\beta}_1 - \beta_1}{\sigma/\sqrt{S_{xx}}} \sim \mathcal{N}(0,1).$$

We will see later that

$$U = \frac{(n-2)S^2}{\sigma^2} \sim \chi^2_{n-2}$$

and S^2 and $\widehat{\beta}_1$ are independent. It follows that

$$T = \frac{\frac{\widehat{\beta}_1 - \beta_1}{\sigma/\sqrt{S_{xx}}}}{\sqrt{\frac{(n-2)S^2}{\sigma^2(n-2)}}} = \frac{\widehat{\beta}_1 - \beta_1}{S/\sqrt{S_{xx}}} \sim t_{n-2}. \tag{2.15}$$

Confidence interval for β_1

To find a CI for an unknown parameter θ means to find values of the boundaries A and B which satisfy

$$P(A < \theta < B) = 1 - \alpha$$

for some small α, that is for a high confidence level $(1-\alpha)100\%$. From (2.15) we have

$$P\left(-t_{\frac{\alpha}{2},n-2} < \frac{\widehat{\beta}_1 - \beta_1}{S/\sqrt{S_{xx}}} < t_{\frac{\alpha}{2},n-2}\right) = 1-\alpha, \tag{2.16}$$

where $t_{\frac{\alpha}{2},n-2}$ is such that $P(|T| < t_{\frac{\alpha}{2},n-2}) = 1-\alpha$.

Rearranging the expression in brackets of (2.16) gives

$$P\left(\widehat{\beta}_1 - t_{\frac{\alpha}{2},n-2}\frac{S}{\sqrt{S_{xx}}} < \beta_1 < \widehat{\beta}_1 + t_{\frac{\alpha}{2},n-2}\frac{S}{\sqrt{S_{xx}}}\right) = 1-\alpha. \tag{2.17}$$

That is the CI for β_1 is

$$[A, B] = \left[\widehat{\beta}_1 - t_{\frac{\alpha}{2},n-2}\frac{S}{\sqrt{S_{xx}}},\ \widehat{\beta}_1 + t_{\frac{\alpha}{2},n-2}\frac{S}{\sqrt{S_{xx}}}\right]. \tag{2.18}$$

The calculated values of $\widehat{\beta}_1$, S and S_{xx} for the overhead costs (Example 2.5) are the following

$$\widehat{\beta}_1 = 10.982,\ \ S = 1645.61,\ \ S_{xx} = 526656.9.$$

Also $t_{0.025,14} = 2.14479$. Hence, the 95% CI for β_1 is

$$\begin{aligned}[a,\ b] &= \left[10.982 - 2.14479\frac{1645.61}{\sqrt{526656.9}},\ 10.982 + 2.14479\frac{1645.61}{\sqrt{526656.9}}\right] \\ &= [6.11851,\ 15.8455]\end{aligned}$$

We would expect (with 95% confidence) that one hour increase in labour will increase the cost between £6.12 and £15.82.

Test of $H_0 : \beta_1 = 0$ **versus** $H_0 : \beta_1 \neq 0$

The null hypothesis $H_0 : \beta_1 = 0$ means that the slope is zero and a better model is a constant model

$$Y_i = \beta_0 + \varepsilon_i, \qquad \varepsilon_i \underset{iid}{\sim} \mathcal{N}(0, \sigma^2)$$

showing no relationship between Y and X. From (2.15) we see that if H_0 is true, then

$$T = \frac{\widehat{\beta}_1}{\frac{S}{\sqrt{S_{xx}}}} \underset{H_0}{\sim} t_{n-2}. \tag{2.19}$$

This statistic can be used as a test function for the null hypothesis.

We reject H_0 at a significance level α when the calculated, for a given data set, value of the test function, T_{cal}, is in the rejection region, that is

$$|T_{cal}| > t_{\frac{\alpha}{2},n-2}.$$

Many statistical software give the p-value when testing a hypothesis. When the p-value is smaller then α then we may reject the null hypothesis on a significance level $\leq \alpha$.

Remark 2.4. Square root of the variance $\mathrm{var}(\widehat{\beta}_1)$ is called the standard error of $\widehat{\beta}_1$ and it is denoted by $se(\widehat{\beta}_1)$, that is

$$se(\widehat{\beta}_1) = \sqrt{\frac{\sigma^2}{S_{xx}}}.$$

Its estimator is

$$\widehat{se(\widehat{\beta}_1)} = \sqrt{\frac{S^2}{S_{xx}}}.$$

Often this estimated standard error is called the standard error. You should be aware of the difference between the two. □

Remark 2.5. Note that the $(1-\alpha)100\%$ CI for β_1 can be written as

$$\left[\widehat{\beta}_1 - t_{\frac{\alpha}{2},n-2}\widehat{se(\widehat{\beta}_1)},\ \ \widehat{\beta}_1 + t_{\frac{\alpha}{2},n-2}\widehat{se(\widehat{\beta}_1)}\right]$$

and the test statistic for $H_0 : \beta_1 = 0$ as

$$T = \frac{\widehat{\beta}_1}{\widehat{se(\widehat{\beta}_1)}}.$$

□

As we have noted before we can also test the hypothesis $H_0 : \beta_1 = 0$ using the Analysis of Variance table and the F test. In this case the two tests are equivalent since if the random variable $W \sim t_\nu$ then $W^2 \sim F_{1,\nu}$.

2.6.2 Inference about β_0

Since we are studying the relationship between X and Y, we are most interested in β_1. However, we can also carry out inference about β_0. The LSE of β_0 is

$$\widehat{\beta}_0 = \bar{Y} - \widehat{\beta}_1 \bar{x}.$$

We have already seen the following result.

Theorem 2.4. *In the full SLM the distribution of the LSE of β_0, $\widehat{\beta}_0$, is normal with the expectation $\mathrm{E}(\widehat{\beta}_0) = \beta_0$ and the variance $\mathrm{var}(\widehat{\beta}_0) = \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\sigma^2$, that is*

$$\widehat{\beta}_0 \sim \mathcal{N}\left(\beta_0, \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]\right). \tag{2.20}$$

□

Corollary 2.1. *Assuming the full simple linear regression model, we obtain*

CI for β_0:

$$\left[\widehat{\beta}_0 - t_{\frac{\alpha}{2},n-2}\widehat{se(\widehat{\beta}_0)},\ \ \widehat{\beta}_0 + t_{\frac{\alpha}{2},n-2}\widehat{se(\widehat{\beta}_0)}\right]$$

Test of the hypothesis $H_0 : \beta_0 = \beta_0^\star$:

$$T = \frac{\widehat{\beta}_0 - \beta_0^\star}{\widehat{se(\widehat{\beta}_0)}} \underset{H_0}{\sim} t_{n-2},$$

where

$$\widehat{se(\widehat{\beta}_0)} = \sqrt{S^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}.$$

□

The calculated values for the overhead costs (Example 2.5) are following

$$\widehat{\beta}_0 = 16310,\ \ \widehat{se(\widehat{\beta}_0)} = 2421$$

Hence, the 95% CI for β_0 is

$$\begin{aligned}[a,\, b] &= [16310 - 2.14479 \times 2421,\ 16310 + 2.14479 \times 2421] \\ &= [11117.5,\ \ 21502.5]\end{aligned}$$

We would expect (with 95% confidence) that even if there is zero hours of labor, the overhead cost is between £11117.5 and £21502.5.

2.6.3 Inference about $\mathrm{E}(Y|X = x_i)$

In the simple linear regression model, we have

$$\mu_i = \mathrm{E}(Y|X = x_i) = \beta_0 + \beta_1 x_i$$

and its LSE is

$$\widehat{\mu}_i = \widehat{\mathrm{E}(Y|X = x_i)} = \widehat{\beta}_0 + \widehat{\beta}_1 x_i.$$

We may estimate the mean response at any value of X which is within the range of the data, say x_0. Then,

$$\widehat{\mu}_0 = \widehat{\mathrm{E}(Y|X = x_0)} = \widehat{\beta}_0 + \widehat{\beta}_1 x_0.$$

Similarly as for the LSE of β_0 and for β_1 we have the following Theorem.

Theorem 2.5. *In the full SLRM the distribution of the LSE of μ_0, $\widehat{\mu}_0$, is normal with the expectation $\mathrm{E}(\widehat{\mu}_0) = \mu_0$ and the variance $\mathrm{var}(\widehat{\mu}_0) = \sigma^2 \left(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right)$, that is*

$$\widehat{\mu}_0 \sim N\left(\mu_0, \sigma^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]\right). \tag{2.21}$$

□

Corollary 2.2. *In the full simple linear regression model, we have*

CI for μ_0:

$$\left[\widehat{\mu}_0 - t_{\frac{\alpha}{2},n-2}\widehat{se(\widehat{\mu}_0)},\ \ \widehat{\mu}_0 + t_{\frac{\alpha}{2},n-2}\widehat{se(\widehat{\mu}_0)}\right]$$

Test of the hypothesis $H_0 : \mu_0 = \mu^*$:

$$T = \frac{\widehat{\mu}_0 - \mu^*}{\widehat{se(\widehat{\mu}_0)}} \underset{H_0}{\sim} t_{n-2},$$

where

$$\widehat{se(\widehat{\mu}_0)} = \sqrt{S^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}.$$

□

Remark 2.6. Care is needed when estimating the mean at x_0. It should only be done if x_0 is within the data range. Extrapolation beyond the range of the given x-values is not reliable, as there is no evidence that a linear relationship is appropriate there.

□

2.6.4 Prediction Interval for a new observation

Apart from making inference on the mean response we may also try to do it for a new response itself, that is for an unknown (not observed) response at some x_0. For example, we might want to predict an overhead cost for another department of the same structure whose total labor hours are x_0 (Example 2.5). In this section we derive a *Prediction Interval (PI)* for a response

$$Y_0 = \beta_0 + \beta_1 x_0 + \varepsilon_0 = \mu_0 + \varepsilon_0, \quad \varepsilon_0 \sim \mathcal{N}(0, \sigma^2)$$

for which the *point prediction* is $\widehat{\mu}_0 = \widehat{\beta}_0 + \widehat{\beta}_1 x_0$.

By Theorem 2.5 we have

$$\widehat{\mu}_0 \sim \mathcal{N}(\mu_0, a\sigma^2),$$

where $a = \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)$.

To obtain a prediction interval (PI) for the unknown observation we may use the point predictor and its distribution as follows. First, we will find the distribution of $\widehat{\mu}_0 - Y_0$. Note that for

$$\widehat{\mu}_0 - Y_0 = \widehat{\mu}_0 - (\mu_0 + \varepsilon_0),$$

we have $\mathrm{E}(\widehat{\mu}_0 - Y_0) = 0$ and

$$\mathrm{var}(\widehat{\mu}_0 - Y_0) = \mathrm{var}(\widehat{\mu}_0) + \mathrm{var}(\mu_0 + \varepsilon_0) = a\sigma^2 + \sigma^2 = \sigma^2(1 + a).$$

This is because $\widehat{\mu}_0$ is the estimator based on the random sample $Y_1, \ldots, Y_n$ and not on Y_0, i.e., it is independent of Y_0. We get,

$$\widehat{\mu}_0 - Y_0 \sim \mathcal{N}(0, \sigma^2(1 + a)).$$

Standardizing $\widehat{\mu}_0 - Y_0$ and replacing σ^2 by its estimator S^2 gives

$$\frac{\widehat{\mu}_0 - Y_0}{\sqrt{S^2\,[1 + a]}} \sim t_{n-2}.$$

Hence, a $(1 - \alpha)100\%$ PI for Y_0 is

$$\widehat{\mu}_0 \pm t_{\frac{\alpha}{2}, n-2} \sqrt{S^2 \left\{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right\}}.$$

This interval is wider than the CI for the mean response μ_0. This is because to predict a new observation rather than a mean, we need to add variability of the

additional random error ε_0. Again, we should only make predictions for values of x_0 within the range of the data.

A part of MINTAB output for the example on the overhead cost (Example 2.5) gives the confidence and prediction intervals (here they are for $x_0 = 1000$ hours.)

```
Predicted Values for New Observations
New
Obs    Fit    SE Fit       95% CI           95% PI
  1   27292     428    (26374, 28210)  (23645, 30939)

Values of Predictors for New Observations
New Obs     x
  1       1000
```

MINITAB

Stat → Regression → Fitted Line Plot...
Options
Prediction intervals for new observations
1000

We may say, with 95% confidence, that when the total direct labour hours are equal to 1000, then the expected total departmental cost would be between $£26374$ and $£28210$, however if we were to observe the total cost for a 1000 hours of labour it might be anything between $£23645$ and $£30939$.

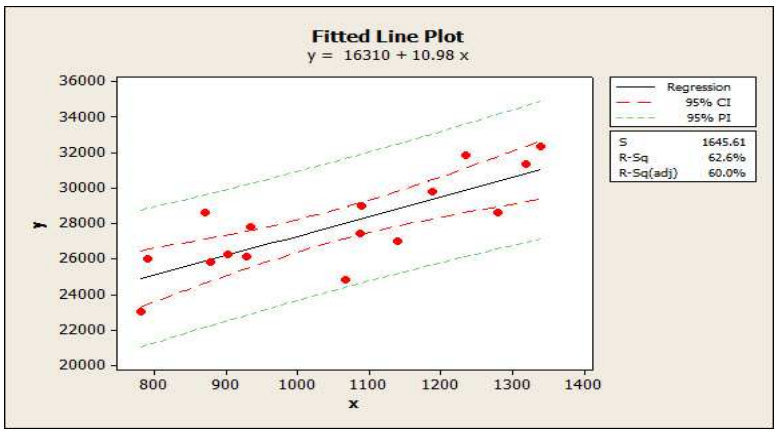

Figure 2.15: Data, fitted line plot, CI for the mean and PI for a new observation at any x_0.

To obtain such plot in MINITAB:

Stat $\rightarrow$ Regression $\rightarrow$ Fitted Line Plot...
Options
Display Options
✓ Display confidence interval
✓ Display prediction interval
Confidence level: 95.0