Math 1710 Class 39

Regression Inference
Dr. Back

Nov. 30, 2009

Regression Inference Questions

Basic Setup:

1) Data (x_i, y_i), $1 \leq i \leq n$ leads to line of regression

$$\hat{y} = b_0 + b_1 x$$

2) Assume an ideal line

$$\hat{y} = \beta_0 + \beta_1 x$$

3) Together with an error process ϵ_i folowing an $N(0, \sigma)$ law (independent for each i.)

4) So that individual observations come from

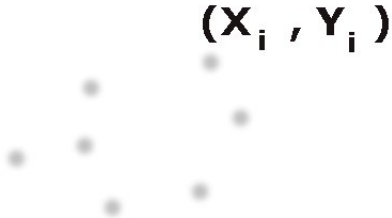
$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i.$$

Regression Inference Questions

Start with some data (x_i, y_i)

Regression Inference Questions

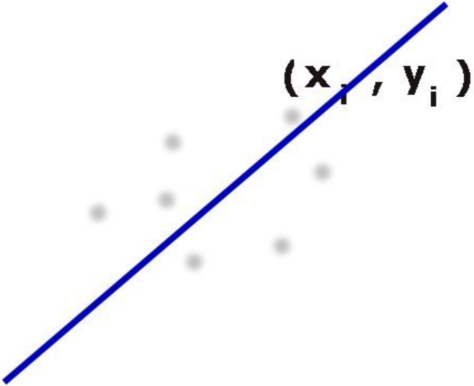
Start with some data (x_i, y_i)



Regression Inference Questions

These give the line of regression

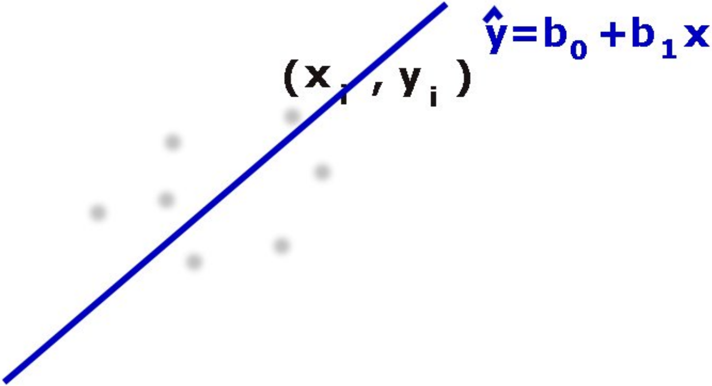
$$\hat{y} = b_0 + b_1 x_i$$



Regression Inference Questions

These give the line of regression

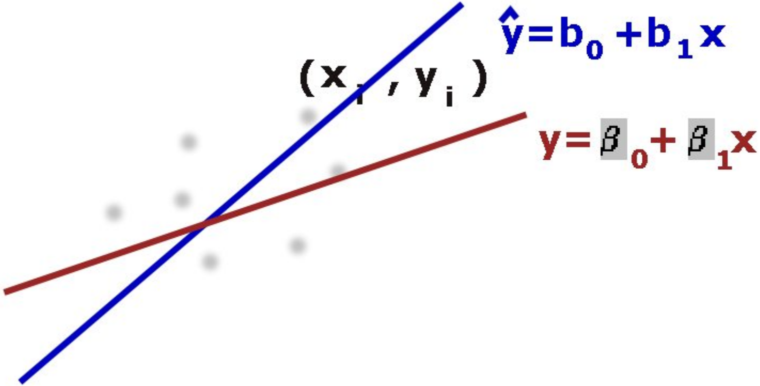
$$\hat{y} = b_0 + b_1 x_i$$



Regression Inference Questions

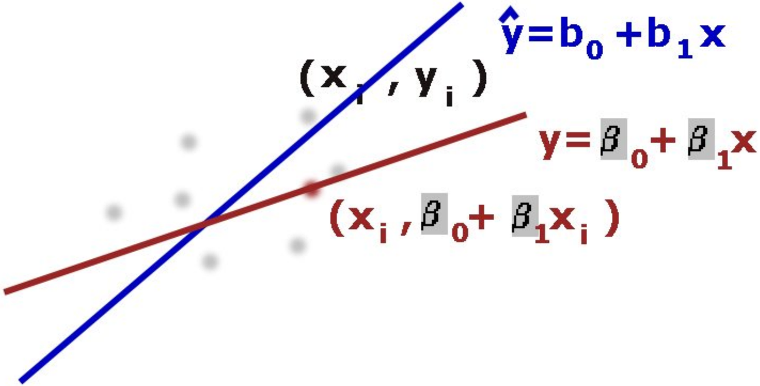
There is also an ideal line

$$\hat{y} = \beta_0 + \beta_1 x_i$$



Regression Inference Questions

We think of the data point (x_i, y_i) as arising by first plugging in x_i into the ideal line

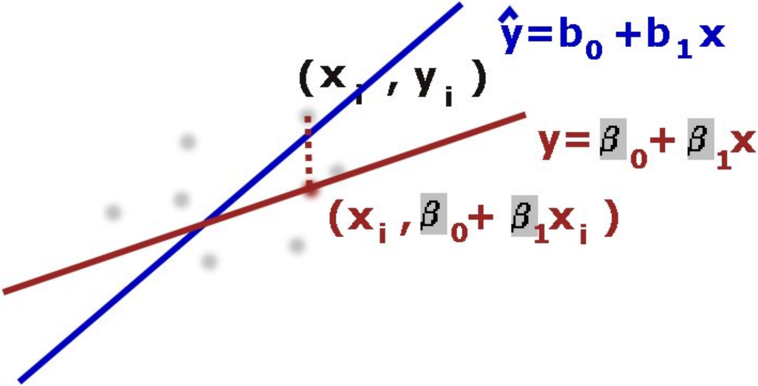


Regression Inference Questions

Then randomly choosing an offset

$$\epsilon_i \sim N(0, \sigma)$$

to see how far off the ideal line we go.

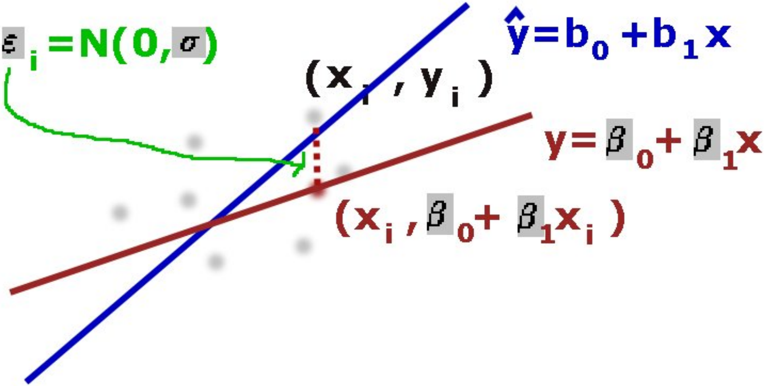


Regression Inference Questions

Then randomly choosing an offset

$$\epsilon_i \sim N(0, \sigma)$$

to see how far off the ideal line we go.



Regression Inference Questions

The errors must be independent and all have the same standard deviation σ in the ideal model.

Regression Inference Questions

Natural Questions in Regression:

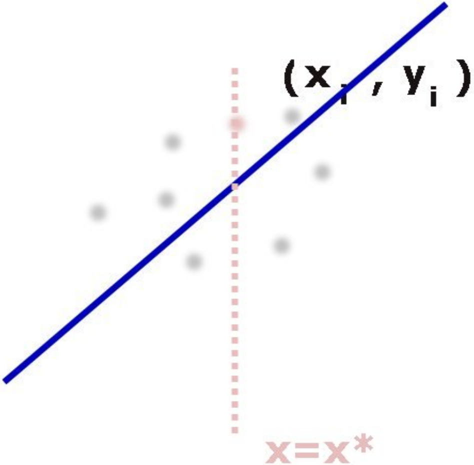
1) Estimate β_0 and β_1.
2) Estimate the accuracy of b_0 and b_1 as estimators of β_0 and β_1.
3) Estimate σ, the standard deviation of the error process.

For a given value x^* of x:

4) How accurately does the regression estimate $b_0 + b_1 x^*$ approximate an actual y observation when $x = x^*$.
5) How accurately does the regression estimate $b_0 + b_1 x^*$ approximate the average of a lot of y observations when $x = x^*$.

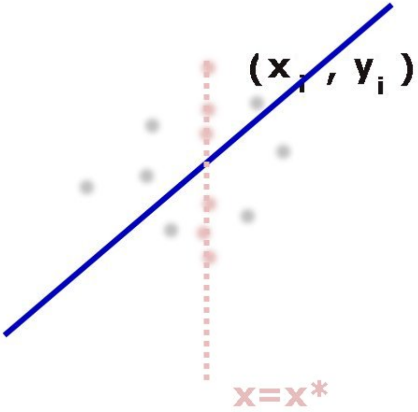
Regression Inference Questions

For $x = x^*$, we can try to predict one value of y leading to a prediction interval when $x = x^*$



Regression Inference Questions

Or we can try to predict the average of many values of y leading to a confidence interval for the mean response μ_y when $x = x^*$

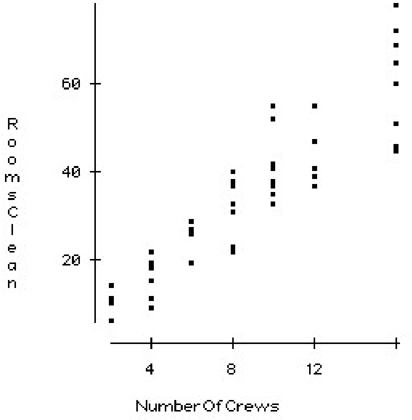


How Many Rooms Can x Crews Clean?

x crews working for a building contractor go out each night and clean y rooms.
Understand the relationship?

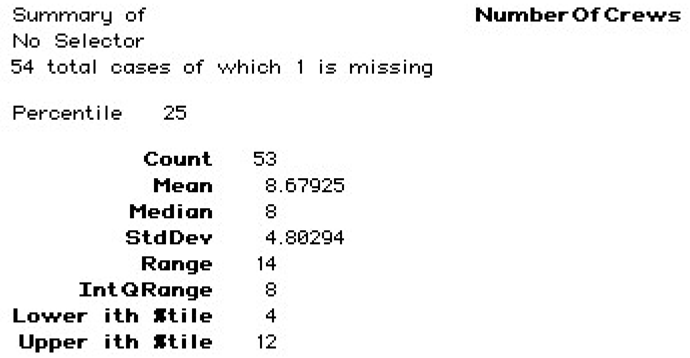
How Many Rooms Can x Crews Clean?

Scatterplot



How Many Rooms Can x Crews Clean?

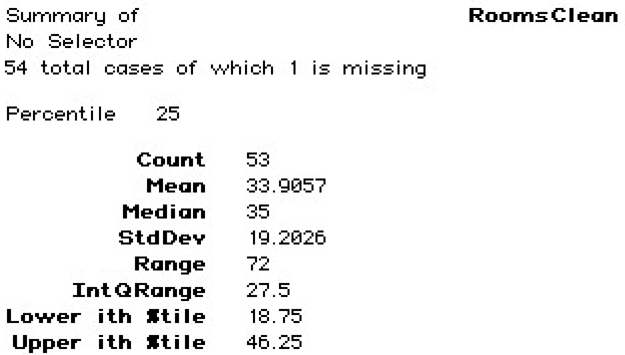
NumCrews summary



```
Summary of                                    Number Of Crews
No Selector
54 total cases of which 1 is missing

Percentile    25

           Count    53
            Mean     8.67925
          Median     8
          StdDev     4.80294
           Range    14
       IntQRange     8
Lower ith %tile      4
Upper ith %tile     12
```

How Many Rooms Can x Crews Clean?

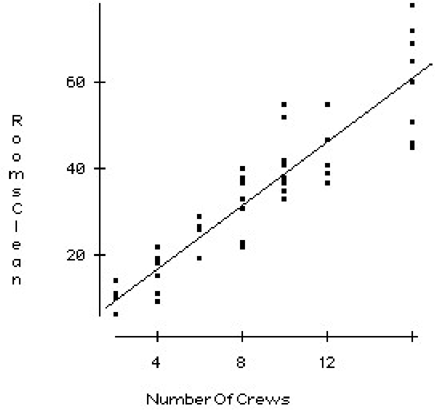
RoomsCleaned Summary



```
Summary of                                    RoomsClean
No Selector
54 total cases of which 1 is missing

Percentile    25

            Count    53
             Mean    33.9057
           Median    35
           StdDev    19.2026
            Range    72
        IntQRange    27.5
  Lower ith %tile    18.75
  Upper ith %tile    46.25
```

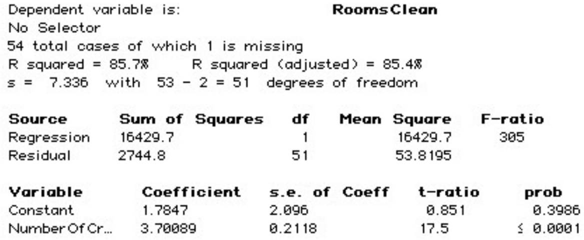
How Many Rooms Can x Crews Clean?

Scatterplot with Regression Line



How Many Rooms Can x Crews Clean?

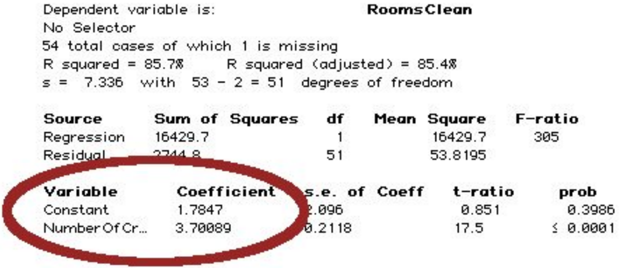
Regression Display



```
Dependent variable is:                    RoomsClean
No Selector
54 total cases of which 1 is missing
R squared = 85.7%      R squared (adjusted) = 85.4%
s =  7.336  with  53 - 2 = 51  degrees of freedom

Source       Sum of Squares    df    Mean Square    F-ratio
Regression   16429.7            1        16429.7      305
Residual     2744.8            51        53.8195

Variable       Coefficient    s.e. of Coeff    t-ratio      prob
Constant       1.7847         2.096             0.851        0.3986
NumberOfCr...  3.70089        0.2118           17.5        ≤ 0.0001
```

How Many Rooms Can x Crews Clean?

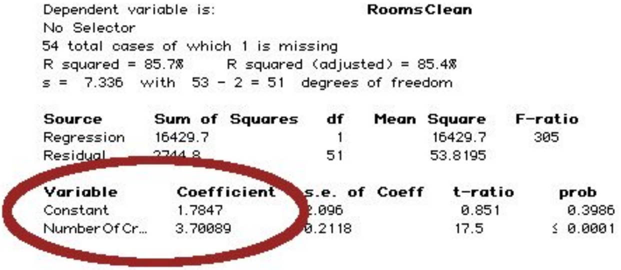
Regression Display



```
Dependent variable is:                    RoomsClean
No Selector
54 total cases of which 1 is missing
R squared = 85.7%      R squared (adjusted) = 85.4%
s =  7.336  with  53 - 2 = 51  degrees of freedom

Source        Sum of Squares     df    Mean Square    F-ratio
Regression    16429.7             1        16429.7     305
Residual      2744.8             51        53.8195

Variable          Coefficient     s.e. of Coeff    t-ratio      prob
Constant          1.7847          2.096             0.851        0.3986
NumberOfCr...     3.70089         0.2118           17.5         ≤ 0.0001
```

How Many Rooms Can x Crews Clean?

Regression Display

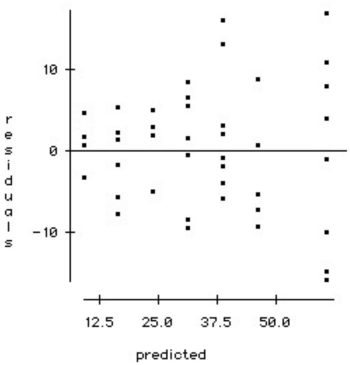


```
Dependent variable is:                    RoomsClean
No Selector
54 total cases of which 1 is missing
R squared = 85.7%      R squared (adjusted) = 85.4%
s =  7.336  with  53 - 2 = 51  degrees of freedom

Source        Sum of Squares    df    Mean Square    F-ratio
Regression    16429.7            1        16429.7       305
Residual      2744.8            51        53.8195

Variable          Coefficient    s.e. of Coeff    t-ratio      prob
Constant          1.7847         .096              0.851        0.3986
NumberOfCr...     3.70089        0.2118            17.5       ≤ 0.0001
```

$$\widehat{RoomsCleaned} = 3.70 \cdot NumCrews + 1.78$$

How Many Rooms Can x Crews Clean?

Residual Plot



How Many Rooms Can x Crews Clean?

There are important deviations from the the assumptions of an ideal linear regression model here.

Estimating the Ideal Parameters

Estimate β_0 and β_1.

Estimating the Ideal Parameters

Estimate β_0 and β_1.
This is what we studied in our original treatment of linear regression.

Estimating the Ideal Parameters

Estimate β_0 and β_1.

β_1 is estimated by b_1 and β_0 is estimated by b_0.

Estimating the Ideal Parameters

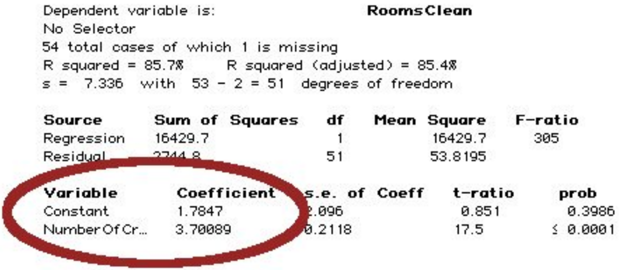
Estimate β_0 and β_1.
β_1 is estimated by b_1 and β_0 is estimated by b_0.

$$\begin{aligned} b_1 &= r\left(\frac{s_y}{s_x}\right) \\ b_0 &= \bar{y} - b_1\bar{x} \end{aligned}$$

Estimating the Ideal Parameters

Estimate β_0 and β_1.
β_1 is estimated by b_1 and β_0 is estimated by b_0.

Regression of NumRooms on NumCrews



```
Dependent variable is:                 RoomsClean
No Selector
54 total cases of which 1 is missing
R squared = 85.7%     R squared (adjusted) = 85.4%
s =  7.336  with  53 - 2 = 51  degrees of freedom

Source        Sum of Squares    df   Mean Square    F-ratio
Regression    16429.7            1       16429.7      305
Residual      2744.8            51       53.8195

Variable        Coefficient     s.e. of Coeff    t-ratio      prob
Constant        1.7847          2.096             0.851        0.3986
NumberOfCr...   3.70089         0.2118           17.5        ≤ 0.0001
```

Accuracy of the Parameters

Estimate the accuracy of b_0 and b_1 as estimators of β_0 and β_1.

Accuracy of the Parameters

Estimate the accuracy of b_0 and b_1 as estimators of β_0 and β_1.

b_1 is an unbiased estimator of β_1.

Accuracy of the Parameters

Estimate the accuracy of b_0 and b_1 as estimators of β_0 and β_1.

b_1 is an unbiased estimator of β_1.
This means that if you take many samples of size n following the same ideal model, as n gets large, the average value of b_1 will approach β_1.

Accuracy of the Parameters

Estimate the accuracy of b_0 and b_1 as estimators of β_0 and β_1.

This is the same kind of reason as that behind the (n-1) in the definition of the variance:

$$s^2 = \frac{1}{n-1} \Sigma (x_i - \bar{x})^2.$$

This makes the sample variance an unbiased estimator of σ^2.

Accuracy of the Parameters

Estimate the accuracy of b_0 and b_1 as estimators of β_0 and β_1.

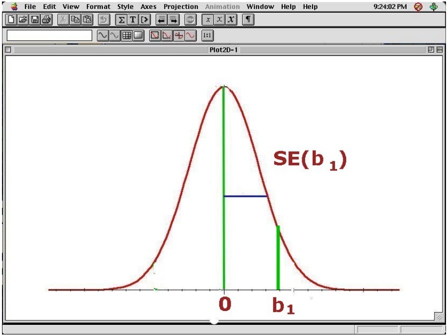
The sampling distribution of b_1 follows a rescaled t-distribution with $n-2$ df and

$$SE(b_1) = \frac{1}{\sqrt{n-1}} \frac{s}{s_x}.$$

Accuracy of the Parameters

Estimate the accuracy of b_0 and b_1 as estimators of β_0 and β_1.

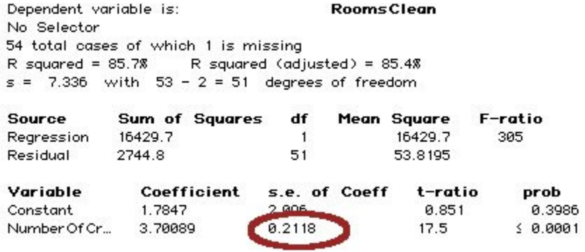
The sampling distribution of b_1 under the null hypothesis of $\beta_1 = 0$.



Accuracy of the Parameters

Estimate the accuracy of b_0 and b_1 as estimators of β_0 and β_1.

Usually one gets $SE(b_1)$ from the regression display.



```
Dependent variable is:                    RoomsClean
No Selector
54 total cases of which 1 is missing
R squared = 85.7%     R squared (adjusted) = 85.4%
s =  7.336  with  53 - 2 = 51  degrees of freedom

Source       Sum of Squares    df    Mean Square    F-ratio
Regression   16429.7            1        16429.7     305
Residual     2744.8            51        53.8195

Variable        Coefficient    s.e. of Coeff    t-ratio     prob
Constant        1.7847         2.006             0.851      0.3986
NumberOfCr...   3.70089        0.2118           17.5       ≤ 0.0001
```

Accuracy of the Parameters

Estimate the accuracy of b_0 and b_1 as estimators of β_0 and β_1.

The sampling distribution of b_1 follows a rescaled t-distribution with $n-2$ df and

$$SE(b_1) = \frac{1}{\sqrt{n-1}}\frac{s}{s_x}.$$

So a confidence interval for β_1 will be given by

$$b_1 \pm t^* SE_{b_1}$$

.

Accuracy of the Parameters

Estimate the accuracy of b_0 and b_1 as estimators of β_0 and β_1.

The sampling distribution of b_1 follows a rescaled t-distribution with $n-2$ df and

$$SE(b_1) = \frac{1}{\sqrt{n-1}}\frac{s}{s_x}.$$

Here $df = n-2 = 53-2 = 51$ and $t^* = 2.009$ for a 95% CI, so a 95% confidence interval for β_1 is:

$$b_1 \pm t^* SE(b_1) = 3.70 \pm 2.009 \cdot 0.212 = 3.70 \pm .43 = (3.27, 4.13).$$

Accuracy of the Parameters

Estimate the accuracy of b_0 and b_1 as estimators of β_0 and β_1.

The sampling distribution of b_1 follows a rescaled t-distribution with $n-2$ df and

$$SE(b_1) = \frac{1}{\sqrt{n-1}}\frac{s}{s_x}.$$

One can also do the hypothesis test

$H_0: \ \beta_1 = 0$

$H_a: \ \beta_1 \neq 0$ (or a 1 sided possibility.)

Accuracy of the Parameters

Estimate the accuracy of b_0 and b_1 as estimators of β_0 and β_1.

The sampling distribution of b_1 follows a rescaled t-distribution with $n-2$ df and

$$SE(b_1) = \frac{1}{\sqrt{n-1}}\frac{s}{s_x}.$$

One can also do the hypothesis test

H_0 : $\beta_1 = 0$

H_a : $\beta_1 \neq 0$ (or a 1 sided possibility.)

The t-statistic is

$$t = \frac{b_1}{SE(b_1)}.$$

Accuracy of the Parameters

Estimate the accuracy of b_0 and b_1 as estimators of β_0 and β_1.

The sampling distribution of b_1 follows a rescaled t-distribution with $n-2$ df and

$$SE(b_1) = \frac{1}{\sqrt{n-1}}\frac{s}{s_x}.$$

One can also do the hypothesis test

$H_0 : \ \beta_1 = 0$

$H_a : \ \beta_1 \neq 0$ (or a 1 sided possibility.)

Here the t-statistic is

$$t = \frac{b_1}{SE(b_1)} = \frac{3.70}{.212} = 17.5$$

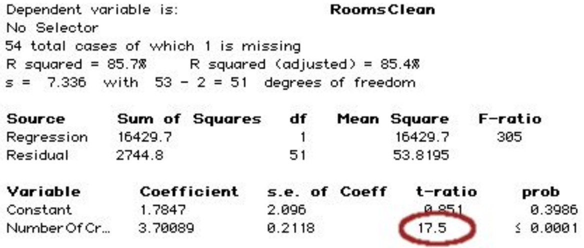
Accuracy of the Parameters

Estimate the accuracy of b_0 and b_1 as estimators of β_0 and β_1.

The sampling distribution of b_1 follows a rescaled t-distribution with $n-2$ df and

$$SE(b_1) = \frac{1}{\sqrt{n-1}}\frac{s}{s_x}.$$

One also finds this t-ratio on the regression display



```
Dependent variable is:                 RoomsClean
No Selector
54 total cases of which 1 is missing
R squared = 85.7%      R squared (adjusted) = 85.4%
s =  7.336  with  53 - 2 = 51  degrees of freedom

Source       Sum of Squares    df    Mean Square    F-ratio
Regression   16429.7            1        16429.7      305
Residual     2744.8            51        53.8195

Variable        Coefficient    s.e. of Coeff    t-ratio     prob
Constant        1.7847         2.096              0.851      0.3986
NumberOfCr...   3.70089        0.2118            17.5      ≤ 0.0001
```

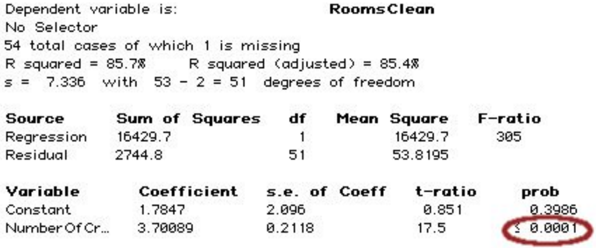
Accuracy of the Parameters

Estimate the accuracy of b_0 and b_1 as estimators of β_0 and β_1.

The sampling distribution of b_1 follows a rescaled t-distribution with $n-2$ df and

$$SE(b_1) = \frac{1}{\sqrt{n-1}} \frac{s}{s_x}.$$

with the P-value for the 2-sided test next to it:



```
Dependent variable is:                   RoomsClean
No Selector
54 total cases of which 1 is missing
R squared = 85.7%     R squared (adjusted) = 85.4%
s =  7.336  with  53 - 2 = 51  degrees of freedom

Source        Sum of Squares     df    Mean Square    F-ratio
Regression     16429.7            1       16429.7       305
Residual       2744.8            51       53.8195

Variable          Coefficient     s.e. of Coeff     t-ratio     prob
Constant           1.7847          2.096              0.851       0.3986
NumberOfCr...      3.70089         0.2118            17.5        ≤ 0.0001
```

Accuracy of the Parameters

Estimate the accuracy of b_0 and b_1 as estimators of β_0 and β_1.

The sampling distribution of b_1 follows a rescaled t-distribution with $n-2$ df and

$$SE(b_1) = \frac{1}{\sqrt{n-1}} \frac{s}{s_x}.$$

Similarly one can do CI's and HT's involving the y-intercept β_0 using

$$SE(b_0) = s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\Sigma(x_i - \bar{x})^2}}.$$

Estimating the Size of the Error Process

Estimate σ, the standard deviation of the error process.

Estimating the Size of the Error Process

Estimate σ, the standard deviation of the error process.

σ is estimated by

$$s = \sqrt{\frac{\Sigma(y_i - \bar{y})^2}{n-2}}.$$

(Almost the standard deviation of the residuals.)

Estimating the Size of the Error Process

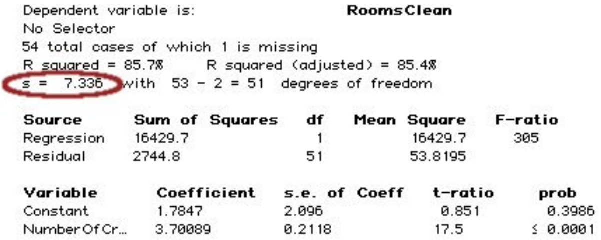
Estimate σ, the standard deviation of the error process.

σ is estimated by

$$s = \sqrt{\frac{\Sigma(y_i - \bar{y})^2}{n-2}}.$$

(Almost the standard deviation of the residuals.)

Usually we get this on the regression display



```
Dependent variable is:                     RoomsClean
No Selector
54 total cases of which 1 is missing
R squared = 85.7%       R squared (adjusted) = 85.4%
s =   7.336  with   53 - 2 = 51   degrees of freedom

Source        Sum of Squares    df    Mean Square    F-ratio
Regression    16429.7            1        16429.7      305
Residual      2744.8            51        53.8195

Variable         Coefficient    s.e. of Coeff    t-ratio      prob
Constant         1.7847         2.096             0.851       0.3986
NumberOfCr...    3.70089        0.2118           17.5       ≤ 0.0001
```

$SE(b_1)$ Formula?

First we'll look at the expectation of b_1 which will show b_1 is an unbiased estimator of b_1.

$SE(b_1)$ Formula?

$$b_1 = r\frac{s_y}{s_x} = \frac{1}{(n-1)s_x^2}\Sigma(x_i - \bar{x})(y_i - \bar{y})$$

$SE(b_1)$ Formula?

$$b_1 = r\frac{s_y}{s_x} = \frac{1}{(n-1)s_x^2}\Sigma(x_i-\bar{x})(y_i-\bar{y})$$
$$= \frac{1}{(n-1)s_x^2}\Sigma y_i(x_i-\bar{x})$$

since $\Sigma\bar{y}(x_i-\bar{x}) = 0$.

$SE(b_1)$ Formula?

Using $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ with $E(\epsilon_i) = 0$,

$$\begin{aligned} E(b_1) &= \frac{1}{(n-1)s_x^2} \Sigma E(y_i)(x_i - \bar{x}) \\ &= \frac{1}{(n-1)s_x^2} \Sigma(\beta_0 + \beta_1 x_i)(x_i - \bar{x}) \end{aligned}$$

$SE(b_1)$ Formula?

Since $\Sigma(x_i - \bar{x}) = 0$,

$$\begin{aligned} E(b_1) &= \frac{1}{(n-1)s_x^2} \Sigma E(y_i)(x_i - \bar{x}) \\ &= \frac{1}{(n-1)s_x^2} \Sigma(\beta_0 + \beta_1 x_i)(x_i - \bar{x}) \\ &= \frac{\beta_1}{(n-1)s_x^2} \Sigma x_i (x_i - \bar{x}) \end{aligned}$$

$SE(b_1)$ Formula?

Again since $\Sigma(x_i - \bar{x}) = 0$,

$$\begin{aligned} E(b_1) &= \frac{1}{(n-1)s_x^2} \Sigma E(y_i)(x_i - \bar{x}) \\ &= \frac{1}{(n-1)s_x^2} \Sigma(\beta_0 + \beta_1 x_i)(x_i - \bar{x}) \\ &= \frac{\beta_1}{(n-1)s_x^2} \Sigma x_i(x_i - \bar{x}) \\ &= \frac{\beta_1}{(n-1)s_x^2} \Sigma(x_i - \bar{x})(x_i - \bar{x}) \end{aligned}$$

$SE(b_1)$ Formula?

Using the definition of the variance of x,

$$\begin{aligned}
E(b_1) =& \frac{1}{(n-1)s_x^2} \Sigma E(y_i)(x_i - \bar{x}) \\
=& \frac{1}{(n-1)s_x^2} \Sigma(\beta_0 + \beta_1 x_i)(x_i - \bar{x}) \\
=& \frac{\beta_1}{(n-1)s_x^2} \Sigma x_i(x_i - \bar{x}) \\
=& \frac{\beta_1}{(n-1)s_x^2} \Sigma(x_i - \bar{x})(x_i - \bar{x}) \\
=& \frac{\beta_1}{(n-1)s_x^2} \Sigma(x_i - \bar{x})^2 \\
=& \beta_1 .
\end{aligned}$$

$SE(b_1)$ Formula?

The computation of the variance of b_1 (leading to $SE(b_1)$) is similar. Recall

$$b_1 = \frac{1}{(n-1)s_x^2} \Sigma y_i (x_i - \bar{x}).$$

$SE(b_1)$ Formula?

The computation of the variance of b_1 (leading to $SE(b_1)$) is similar. Recall

$$b_1 = \frac{1}{(n-1)s_x^2} \Sigma y_i (x_i - \bar{x}).$$

Using $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ with $Var(\epsilon_i) = \sigma^2$,

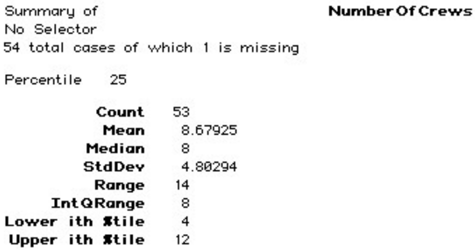
$$\begin{aligned} Var(b_1) &= \frac{1}{(n-1)s_x^2} \Sigma Var(y_i)(x_i - \bar{x})^2 \\ &= \left(\frac{1}{(n-1)s_x^2} \right)^2 \Sigma \sigma^2 (x_i - \bar{x})^2 \\ &= \frac{\sigma^2}{\Sigma (x_i - \bar{x})^2} \end{aligned}$$

$SE(b_1)$ Formula?

The computation of the variance of b_1 (leading to $SE(b_1)$) is similar. Recall

$$b_1 = \frac{1}{(n-1)s_x^2}\Sigma y_i(x_i - \bar{x}).$$

When needed, usually one finds $\Sigma(x_i - \bar{x})^2 = (n-1)s_x^2$ from a computer display



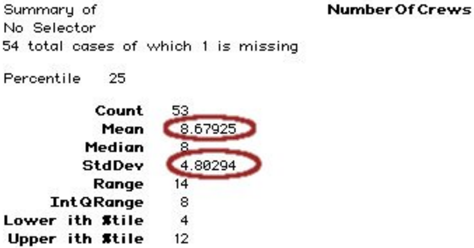
```
Summary of                                  NumberOfCrews
No Selector
54 total cases of which 1 is missing

Percentile    25

             Count    53
              Mean     8.67925
            Median     8
            StdDev     4.80294
             Range    14
         IntQRange     8
   Lower ith %tile     4
   Upper ith %tile    12
```

$SE(b_1)$ Formula?

The computation of the variance of b_1 (leading to $SE(b_1)$) is similar. Recall

$$b_1 = \frac{1}{(n-1)s_x^2} \Sigma y_i (x_i - \bar{x}).$$

When needed, usually one finds $\Sigma(x_i - \bar{x})^2 = (n-1)s_x^2$ from a computer display



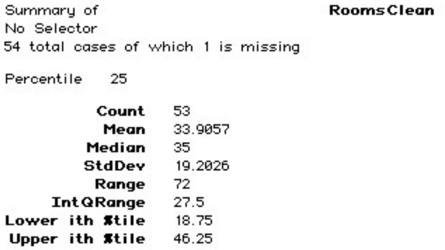
```
Summary of                              NumberOfCrews
No Selector
54 total cases of which 1 is missing

Percentile     25

          Count     53
           Mean     8.67925
         Median     8
         StdDev     4.80294
          Range     14
      IntQRange     8
Lower ith %tile     4
Upper ith %tile     12
```

$SE(b_1)$ Formula?

The computation of the variance of b_1 (leading to $SE(b_1)$) is similar. Recall

$$b_1 = \frac{1}{(n-1)s_x^2}\Sigma y_i(x_i - \bar{x}).$$

But don't confuse x and y!



```
Summary of                                   RoomsClean
No Selector
54 total cases of which 1 is missing

Percentile    25

             Count     53
              Mean     33.9057
            Median     35
            StdDev     19.2026
             Range     72
         IntQRange     27.5
 Lower ith %tile       18.75
 Upper ith %tile       46.25
```

Prediction Intervals

For a given value x^* of x:
How accurately does the regression estimate $b_0 + b_1 x^*$ approximate an actual y observation when $x = x^*$.

Prediction Intervals

For a given value x^* of x:
How accurately does the regression estimate $b_0 + b_1x^*$ approximate an actual y observation when $x = x^*$.
Since the error term ϵ_i has variance σ^2, it is natural to estimate its variance by s^2. We'll estimate the variance of the $b_0 + b_1x^*$ more carefully when we answer question 5. The resultant standard error estimate is:

Prediction Intervals

For a given value x^* of x:
How accurately does the regression estimate $b_0 + b_1 x^*$ approximate an actual y observation when $x = x^*$.
Prediction interval for y when $x = x^*$ has

$$SE_{PI} = \sqrt{s^2 + \frac{s^2}{n} + SE^2(b_1)(x^* - \bar{x})^2}.$$

The y values when $x = x^*$ also follow a t-distribution with mean $b_0 + b_1 x^*$ and $n - 2$ df. (Everything besides the first s^2 in SE_{PI} comes from the SE of the mean response formula.)

Prediction Intervals

For a given value x^* of x:

How accurately does the regression estimate $b_0 + b_1 x^*$ approximate an actual y observation when $x = x^*$.

For the cleaning example,

- $b_0 = 1.78$, $b_1 = 3.70$ where $\hat{y} = b_0 + b_1 x$.
- $n = 53$, $\bar{x} = 8.68$.
- $s = 7.336$.
- $SE(b_1) = .212$.
- The upper .025 critical value for t distribution with $df = 51$ (using 50) is $t^* = 2.009$.

Prediction Intervals

For a given value x^* of x:
How accurately does the regression estimate $b_0 + b_1 x^*$ approximate an actual y observation when $x = x^*$.
When $x^* = 10$:

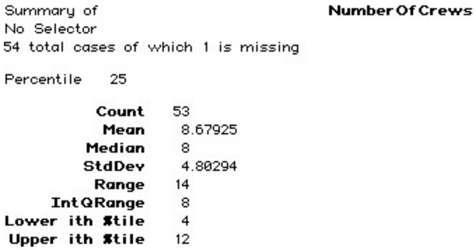
- $\hat{y} = 1.78 + 3.70 \cdot 10 = 38.78$.
- $SE_{\mu_y} = \sqrt{\frac{7.336^2}{53} + 0.212^2 \cdot (10 - 8.68)^2} = \sqrt{1.0154 + .0783} = \sqrt{1.0937} = 1.046$
- 95% CI for mean response$=38.78 \pm 2.009 \cdot 1.046 = (36.68, 40.88)$
- $SE_{pi} = \sqrt{7.336^2 + 1.0937} = 7.417$
- 95% PI for single value$=38.78 \pm 2.009 \cdot 7.417 = (23.88, 53.68)$

where we've also computed the 95% CI for the mean response using the answer to question 5 below.

Prediction Intervals

For a given value x^* of x:
How accurately does the regression estimate $b_0 + b_1 x^*$ approximate an actual y observation when $x = x^*$.

When needed, usually one finds $\Sigma(x_i - \bar{x})^2 = (n-1)s_x^2$ from a computer display

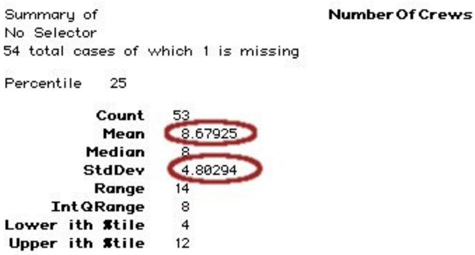


```
Summary of                                 NumberOfCrews
No Selector
54 total cases of which 1 is missing

Percentile     25

            Count     53
             Mean      8.67925
           Median      8
           StdDev      4.80294
            Range     14
        IntQRange      8
Lower  ith  %tile      4
Upper  ith  %tile     12
```

Prediction Intervals

For a given value x^* of x:
How accurately does the regression estimate $b_0 + b_1 x^*$ approximate an actual y observation when $x = x^*$.

When needed, usually one finds $\Sigma(x_i - \bar{x})^2 = (n-1)s_x^2$ from a computer display



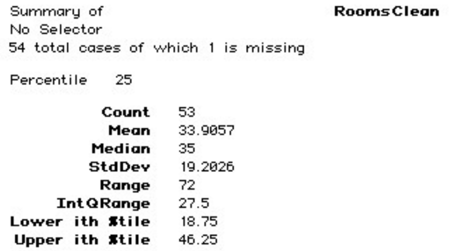
```
Summary of                                  Number Of Crews
No Selector
54 total cases of which 1 is missing

Percentile     25

            Count     53
             Mean      8.67925
           Median      8
           StdDev      4.80294
            Range     14
        IntQRange      8
   Lower ith %tile     4
   Upper ith %tile    12
```

Prediction Intervals

For a given value x^* of x:

How accurately does the regression estimate $b_0 + b_1 x^*$ approximate an actual y observation when $x = x^*$.

But don't confuse x and y!



```
Summary of                              RoomsClean
No Selector
54 total cases of which 1 is missing

Percentile     25

           Count    53
            Mean    33.9057
          Median    35
          StdDev    19.2026
           Range    72
       IntQRange    27.5
 Lower ith %tile    18.75
 Upper ith %tile    46.25
```

Confidence Intervals for Mean Response

For a given value x^* of x:
How accurately does the regression estimate $b_0 + b_1 x^*$ approximate the average μ_y of a lot of y observations when $x = x^*$.

Confidence Intervals for Mean Response

For a given value x^* of x:
How accurately does the regression estimate $b_0 + b_1 x^*$ approximate the average μ_y of a lot of y observations when $x = x^*$.
The line of regression can be written as

$$y = b_1(x - \bar{x}) + \bar{y}$$

so

$$\mu_y = b_1(x^* - \bar{x}) + \bar{y}$$

.

Confidence Intervals for Mean Response

For a given value x^* of x:

How accurately does the regression estimate $b_0 + b_1 x^*$ approximate the average μ_y of a lot of y observations when $x = x^*$.

The line of regression can be written as

$$y = b_1(x - \bar{x}) + \bar{y}$$

so

$$\mu_y = b_1(x^* - \bar{x}) + \bar{y}$$

.

We already know $Var(b_1)$ and the CLT tells us

$$Var(\bar{y}) = \frac{\sigma^2}{n}$$

which is naturally estimated by $\frac{s^2}{n}$.

Confidence Intervals for Mean Response

For a given value x^* of x:
How accurately does the regression estimate $b_0 + b_1 x^*$ approximate the average μ_y of a lot of y observations when $x = x^*$.
The line of regression can be written as

$$y = b_1(x - \bar{x}) + \bar{y}$$

so

$$\mu_y = b_1(x^* - \bar{x}) + \bar{y}$$

.

If b_1 and $\bar{y}$ were known to be independent, we'd be done and have our $SE(b_1)$ formula.
That isn't guaranteed, but a little calculation shows the covariance of b_1 and $\bar{y}$ is 0 which is good enough to establish:

Confidence Intervals for Mean Response

For a given value x^* of x:
How accurately does the regression estimate $b_0 + b_1 x^*$ approximate the average μ_y of a lot of y observations when $x = x^*$.
Confidence interval for mean response μ_y when $x = x^*$ has

$$SE_{\hat{\mu}_y} = \sqrt{\frac{s^2}{n} + SE^2(b_1)(x^* - \bar{x})^2}.$$

Confidence Intervals for Mean Response

For a given value x^* of x:
How accurately does the regression estimate $b_0 + b_1x^*$ approximate the average μ_y of a lot of y observations when $x = x^*$.
Confidence interval for mean response μ_y when $x = x^*$ has

$$SE_{\hat{\mu}_y} = \sqrt{\frac{s^2}{n} + SE^2(b_1)(x^* - \bar{x})^2}.$$

Again a rescaled t distribution with $n - 2$ df is used to form the CI for the mean response.

Regression Inference Conditions

straight enough

Regression Inference Conditons

straight enough
independent errors (no pattern in residuals)

Regression Inference Conditons

straight enough
independent errors (no pattern in residuals)
equal variance of errors

Regression Inference Condititons

straight enough
independent errors (no pattern in residuals)
equal variance of errors
normality of errors