1 Linear Regression

1.1 Simple Linear Regression Model

The linear regression model is applied if we want to model a numeric response variable and its dependency on at least one numeric factor variable. We will later see how we can add more factors, numerical as well as categorical to the model. Because of this flexibility we find regression to be a powerful tool. But again we have to pay close attention to the assumptions made for the analysis of the model to result in meaningful results.

A first order regression model (Simple Linear Regression Model)

\[ y = \beta_0 + \beta_1 x + e \]

where

\[ y = \text{response variable} \]
\[ x = \text{independent or predictor or factor variable} \]
\[ e = \text{random error, assumed to be normally distributed with mean 0 and standard deviation } \sigma \]
\[ \beta_0 = \text{intercept} \]
\[ \beta_1 = \text{slope} \]

A consequence of this model is that for a given value of \( x \) the mean of the random variable \( y \) equals the deterministic part of the model (because \( \mu_e = 0 \)).

\[ \mu_y = E(y) = \beta_0 + \beta_1 x \]

\( \beta_0 + \beta_1 x \) describes a line with y-intercept \( \beta_0 \) and slope \( \beta_1 \)
In conclusion the model implies that $y$ in average follows a line, depending on $x$. Since $y$ is assumed to also underlie a random influence, not all data is expected to fall on the line, but that the line will represent the mean of $y$ for given values of $x$. We will see how to use sample data to estimate the parameters of the model, how to check the usefulness of the model, and how to use the model for prediction, estimation and interpretation.

1.1.1 Fitting the Model

In a first step obtain a scattergram, to check visually if a linear relationship seems to be reasonable.

Example 1
Conservation Ecology study (cp. exercise 2.148 on pg. 100)

Researchers developed two fragmentation indices to score forests - one index measuring anthropogenic (human causes) fragmentation, and one index measuring fragmentation for natural causes. Assume the data below shows the indices (rounded) for 5 forests.

<table>
<thead>
<tr>
<th>noindex $x_i$</th>
<th>anthindex $y_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>34</td>
</tr>
<tr>
<td>28</td>
<td>41</td>
</tr>
<tr>
<td>28</td>
<td>45</td>
</tr>
<tr>
<td>27</td>
<td>38</td>
</tr>
<tr>
<td>23</td>
<td>25</td>
</tr>
</tbody>
</table>

The diagram shows a linear relationship between the two indices (a straight line would represent the relationship pretty well).

If we would include a line representing the relationship between the indices, we could find the differences between the measured values and the line for the same $x$ value. These differences between the data points and the values on the line are called deviations.
Example 2
A marketing manager conducted a study to determine whether there is a linear relationship between money spent on advertising and company sales.
The data are shown in the table below. Advertising expenses and company sales are both in $1000.

<table>
<thead>
<tr>
<th>ad expenses $x_i$</th>
<th>company sales $y_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.4</td>
<td>225</td>
</tr>
<tr>
<td>1.6</td>
<td>184</td>
</tr>
<tr>
<td>2.0</td>
<td>220</td>
</tr>
<tr>
<td>2.6</td>
<td>240</td>
</tr>
<tr>
<td>1.4</td>
<td>180</td>
</tr>
<tr>
<td>1.6</td>
<td>184</td>
</tr>
<tr>
<td>2.0</td>
<td>186</td>
</tr>
<tr>
<td>2.2</td>
<td>215</td>
</tr>
</tbody>
</table>

The graph shows a positive strong linear relationship between sales and advertisement expenses.

If a relationship between two variables would have no random component, the dots would all be precisely on a line and the deviations would all be 0.

\[ \text{deviations} = \text{error of prediction} = \text{difference between observed and predicted values}. \]

We will choose the line that results in the least squared deviations (for the sample data) as the estimation for the line representing the deterministic portion of the relationship in the population.

Continue example 1:
The model for the anthropogenic index ($y$) in dependency on the natural cause index ($x$) is:

\[ y = \beta_0 + \beta_1 x + e \]
After we established through the scattergram that a linear model seems to be appropriate, we now want to estimate the parameters of the model (based on sample data). We will get the estimated regression line

\[ \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \]

The hats indicate that these are estimates for the true population parameters, based on sample data.

**Continue example 2:**
The model for the company sales \( (y) \) in dependency on the advertisement expenses \( (x) \) is:

\[ y = \beta_0 + \beta_1 x + \epsilon \]

We now want to estimate the parameters of the model (based on sample data). We will get the estimated regression line

\[ \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \]

**Estimating the model parameters using the least squares line**

Once we choose \( \hat{\beta}_0 \) and \( \hat{\beta}_1 \), we find that the deviation of the \( i \)th value from its predicted value is \( y_i - \hat{y}_i = y_i - \hat{\beta}_0 + \hat{\beta}_1 x_i \).

Then the sum of squared deviations for all \( n \) data points is

\[ SSE = \sum (y_i - \hat{\beta}_0 + \hat{\beta}_1 x_i)^2 \]

We will choose \( \hat{\beta}_0 \) and \( \hat{\beta}_1 \), so that the \( SSE \) is as small as possible, and call the result the "least squares estimates" of the population parameters \( \beta_0 \) and \( \beta_1 \).

Solving this problem mathematically we get: With

\[ SS_{xy} = \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n} \quad \text{and} \quad SS_{xx} = \sum x_i^2 - \frac{(\sum x_i)^2}{n} \]

the estimate for the slope is \( \hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} \) and

the estimate for the intercept is \( \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \).

**Continue example 1:**

<table>
<thead>
<tr>
<th>( x_i )</th>
<th>( y_i )</th>
<th>( x_i^2 )</th>
<th>( y_i^2 )</th>
<th>( x_i y_i )</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>34</td>
<td>625</td>
<td>1156</td>
<td>850</td>
</tr>
<tr>
<td>28</td>
<td>41</td>
<td>784</td>
<td>1681</td>
<td>1148</td>
</tr>
<tr>
<td>28</td>
<td>45</td>
<td>784</td>
<td>2025</td>
<td>1260</td>
</tr>
<tr>
<td>27</td>
<td>38</td>
<td>729</td>
<td>1444</td>
<td>1026</td>
</tr>
<tr>
<td>23</td>
<td>25</td>
<td>529</td>
<td>625</td>
<td>575</td>
</tr>
<tr>
<td>totals</td>
<td>131</td>
<td>183</td>
<td>3451</td>
<td>6931</td>
</tr>
</tbody>
</table>
From these results we get:

\[
SS_{xy} = 4859 - \frac{(131)(183)}{5} = 64.4
\]
\[
SS_{xx} = 3451 - \frac{(131)^2}{5} = 18.8
\]
\[
\hat{\beta}_1 = \frac{64.4}{18.8} = 3.4255
\]
\[
\hat{\beta}_0 = \frac{185}{5} - 3.4255 \frac{131}{5} = -53.14
\]

So, \( \hat{y} = -53.14 + 3.4255x \) describes the line that results in the least total squared deviations (SSE) possible. Here \( SSE = 12.60 \) (found with minitab).

The intercept tells us that the anthropogenic fragmentation score would be -53 in average if the natural cause score is zero. This is probably not possible.

The slope tells us that in average the anthropogenic score increases by 3.42 point for every one point increase in the natural cause index.

![Scatterplot of anthindex vs noindex](image)

Continue example 2:
From these results we get:

\[ SS_{xy} = 3289.8 - \frac{(15.8)(1634)}{8} = 62.65 \]

\[ SS_{xx} = 32.44 - \frac{(15.8)^2}{8} = 1.235 \]

\[ \hat{\beta}_1 = \frac{62.65}{1.235} = 50.7287 \]

\[ \hat{\beta}_0 = \frac{1634}{8} - 50.7287 \frac{15.8}{8} = 104.0608 \]

So, \( \hat{y} = 104.06 + 50.73x \) describes the line that results in the least total squared deviations (SSE) possible.

Interpretation:
For every extra dollar sent on advertisement the company sales increase in average by $50.73.

When the company does not spend any money on advertisement \((x = 0)\), then the mean sales equal $104060.8.

Use

\[ SSE = SS_{yy} - \hat{\beta}_1 SS_{xy} \]

to find the \( SSE \) for this regression line.

\[ SS_{yy} = 337558 - \frac{(1634)^2}{8} = 3813.5 \]

Then \( SSE = 3813.5 - 50.7287(62.65) = 635.3469 \).

**Built in assumptions:**
When introducing the model, we stated that the error shall be normally distributed with mean zero and standard deviation \( \sigma \) independent from the value of \( x \).

This is illustrated by —————— > diagram

The first implication of the model is that the mean of \( y \) follows a straight line in \( x \), not a curve.

We write \( \mu_y = E(y) = \beta_0 + \beta_1 x \)

The model also implies that the deviations from the line are normally distributed, where the variation from the line is constant for \( x \).
The shape of the distribution of the deviations is normal.

When using the least squares estimators introduced above we also have to assume that the measurements are independent.

From this discussion we learn that another important parameter in the model (beside $\beta_0$ and $\beta_1$) is the standard deviation of the error: $\sigma$

**Estimating $\sigma$**

When estimating the standard deviation of the error in an ANOVA model we based it on the $SSE$ or $MSE$. We will do the same in regression.

For large standard deviations we should expect the deviations, that is the difference between the estimates and the measured values in the sample, to be larger than for smaller standard deviations. Because of this it is not surprising that the $SSE = \sum(\hat{y}_i - y_i)^2$ is part of the estimate.

In fact the best estimate for $\sigma^2$ is

$$\hat{\sigma}^2 = s^2 = \frac{SSE}{n-2}$$

where

$$SSE = \sum(\hat{y}_i - y_i)^2 = SS_{yy} - \hat{\beta}_1 SS_{xy}$$

and

$$SS_{yy} = \sum y_i^2 - \left(\frac{\sum y_i}{n}\right)^2$$

then the best estimate for $\sigma$ is

$$s = \sqrt{s^2} = \sqrt{\frac{SSE}{n-2}}$$

It is called the estimated standard error of the regression model.

**Continue example 1:**

Let us estimate the standard deviation $\sigma$ in the model for the fragmentation indices.

$$SS_{yy} = 6931 - \frac{(183)^2}{5} = 233.2$$

$$SSE = 233.2 - 3.4255(64.4) = 12.5978$$

then

$$s = \sqrt{\frac{12.5978}{5-2}} = 2.049$$
is the estimated standard error in the regression model.

**Continue example 2:**
The estimate for the standard deviation $\sigma$ in the model for the company sales is

$$s = \sqrt{\frac{635.3469}{8 - 2}} = 10.29$$

is the estimated standard error in the regression model.

**For interpreting $s$:**
One should expect about 95% of the observed $y$ values to fall within $2s$ of the estimated line.

### 1.1.2 Checking Model Utility – Making Inference about the slope

As the sample mean $\bar{x}$ is an estimate for the population mean $\mu$, the sample slope $\hat{\beta}_1$ is an estimate for the population slope $\beta_1$. We can now continue on the same path as we did when using sample data for drawing conclusions about a population mean: find a confidence interval, learn how to conduct a test.

In order to this we have to discuss the **distribution of $\hat{\beta}_1$.**

If the assumptions of the regression model hold then $\hat{\beta}_1$ is normally distributed with mean $\mu_{\hat{\beta}_1} = \beta_1$ and standard deviation

$$\sigma_{\hat{\beta}_1} = \frac{\sigma}{\sqrt{SS_{xx}}}$$

$\sigma_{\hat{\beta}_1}$ is estimated by

$$s_{\hat{\beta}_1} = \frac{s}{\sqrt{SS_{xx}}}$$

which is the estimated standard error of the least squares slope $\hat{\beta}_1$.

Combining all this information we find that under the assumption of the regression model

$$t = \frac{\hat{\beta}_1 - \beta_1}{s/\sqrt{SS_{xx}}}$$

is t-distributed with $df = n - 2$.

$$\text{(1 - } \alpha \text{) } \times 100\% \text{ CI for } \beta_1$$

$$\hat{\beta}_1 \pm t^* \frac{s}{\sqrt{SS_{xx}}}$$

where $t^* = (1 - \alpha/2)$ percentile of the t-distribution with $df = n - 2$.

**Continue example 1:**
A 95% CI for $\beta_1$ (df=3)

$$3.4255 \pm (3.182) \frac{2.049}{\sqrt{18.8}} \rightarrow 3.4255 \pm 1.1648$$
A 95% CI for \( \beta_1 \) is \([2.2607 ; 4.5903]\). We conclude with 95% confidence that the anthropologic index increases in average between 2.26 to 4.59 points for a one point increase in the natural cause index.

**Continue example 2:**
A 95% CI for \( \beta_1 \) (df=8-2=6)

\[
50.72 \pm (2.447) \frac{10.29}{\sqrt{1.235}} \rightarrow 50.72 \pm 22.658
\]

A 95% CI for \( \beta_1 \) is \([28.062 ; 73.288]\). We conclude that with 95% confidence the sales price increases in average between $28.1 and $73.3 for every extra dollar spent on advertisement.

Does the CI provide sufficient evidence that \( \beta_1 \neq 0 \)?
This is an important question, because if \( \beta_1 \) would be 0, then the variable \( x \) would contribute no information on predicting \( y \) when using a linear model.

Another way of answering this question is the **model utility test about** \( \beta_1 \)

1. Hypotheses: Parameter of interest \( \beta_1 \).
   
   \[
   \begin{array}{c|c}
   \text{test type} & \text{hypotheses} \\
   \hline
   \text{upper tail} & H_0 : \beta_1 \leq 0 \text{ vs. } H_a : \beta_1 > 0 \\
   \text{lower tail} & H_0 : \beta_1 \geq 0 \text{ vs. } H_a : \beta_1 < 0 \\
   \text{two tail} & H_0 : \beta_1 = 0 \text{ vs. } H_a : \beta_1 \neq 0 \\
   \end{array}
   \]
   Choose \( \alpha \).

2. Assumptions: Regression model is appropriate

3. Test statistic:
   \[
   t_0 = \frac{\hat{\beta}_1}{s/\sqrt{SS_{xx}}} \quad df = n - 2
   \]

4. P-value:
   
   \[
   \begin{array}{c|c}
   \text{test type} & \text{P-value} \\
   \hline
   \text{upper tail} & P(t > t_0) \\
   \text{lower tail} & P(t < t_0) \\
   \text{two tail} & 2P(t > abs(t_0))
   \end{array}
   \]

5. Decision: If \( P-value < \alpha \) reject \( H_0 \), otherwise do not reject \( H_0 \).

6. Put into context

**Continue example 1:**
Test if \( \beta_1 > 0 \)

1. Hypotheses: Parameter of interest \( \beta_1 \).
   \[
   H_0 : \beta_1 \leq 0 \text{ vs. } H_a : \beta_1 > 0 \text{ Choose } \alpha = 0.05.
   \]

2. Assumptions: Regression model is appropriate
3. Test statistic:
   \[ t_0 = \frac{3.4255}{2.049/\sqrt{18.8}} = 9.358, \quad df = 3 \]

4. P-value: upper tail \( P(t > t_0) < 0.005 \) (table IV).

5. Decision: \( P-value < 0.005 < 0.05 = \alpha \) reject \( H_0 \)

6. At significance level of 5% we conclude that in average the anthropogenic index increases as the natural cause index increases. The linear regression model helps explaining the anthropogenic index.

**Continue example 2:**
Test if \( \beta_1 > 0 \)

1. Hypotheses: Parameter of interest \( \beta_1 \).
   \( H_0 : \beta_1 \leq 0 \) vs. \( H_a : \beta_1 > 0 \) Choose \( \alpha = 0.05 \).

2. Assumptions: Regression model is appropriate

3. Test statistic:
   \[ t_0 = \frac{50.72 - 0}{10.29/\sqrt{1.235}} = 5.477, \quad df = 6 \]

4. P-value: upper tail \( P(t > 5.477) < 0.005 \) (table IV)

5. Decision: \( P-value < 0.05 = \alpha \) reject \( H_0 \)

6. At significance level of 5% we conclude that the mean sales increase, when the amount spend on advertisement increases.

### 1.1.3 Measuring Model Fit

#### Correlation Coefficient

The correlation coefficient
\[ r = \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}} \]
measures the strength of the linear relationship between two variables \( x \) and \( y \).

**Properties**

1. \( -1 \leq r \leq 1 \)

2. independent from units used

3. \( abs(r) = 1 \) indicates that all values fall precisely on a line

4. \( r > (<)0 \) indicates a positive (negative) relationship.
5. $r \approx 0$ indicates no linear relationship

(Diagrams)

**Continue example 1:**

\[ r = \frac{64.4}{\sqrt{18.8(233.2)}} = 0.9726 \]

indicates a very strong positive linear relationship between the two indices.

We can interpret $r$ as an estimate of the population correlation coefficient $\rho$ (Greek letter rho).

We could use a statistic based on $r$, to test if $\rho = 0$, this would be equivalent to testing if $\beta_1 = 0$, we will skip this test for that reason.

**Continue example 2:**

\[ r = \frac{62.65}{\sqrt{1.235(3813.5)}} = 0.9107 \]

indicates a very strong positive linear relationship between the company sales and the advertisement expenses.

**Coefficient of Determination**

The coefficient of determination measures the usefulness of the model, it is the correlation coefficient squared, $r^2$, and a number between 0 and 1. It’s wide use can be explained through it’s interpretation.

An alternative formula for $r^2$ is

\[ r^2 = \frac{SS_{yy} - SSE}{SS_{yy}} \]

This formula can help to understand the interpretation of $r^2$.

$SS_{yy}$ measures the spread in the $y$ values (in the sample), and $SSE$ measures the spread in $y$ we can not explain through the linear relationship with $x$. So that $SS_{yy} - SSE$ is the spread in $y$ explained through the linear relationship with $x$. By dividing this by $SS_{yy}$, we find that $r^2$ gives the proportion in the spread of $y$, that is explained through the linear relationship with $x$.

**Continue Example 1:**

The coefficient of determination in our example is

\[ r^2 = \frac{233.2 - 12.5978}{233.2} = 0.946 \]

94.6% in the variation in the anthropogenic index is explained through the natural cause index, this suggests that fragmentation of the two cause are highly related and interdependent.

**Continue Example 2:**

The coefficient of determination in the sales example is

\[ r^2 = 0.9107^2 = 0.829 \]

82.9% in the variation in the company sales can be explained through the linear relationship with the advertisement expenses.

A high coefficient of determination indicates that the model is suitable for estimation and prediction. A more detailed way of checking model fit is the Residual Analysis.
1.1.4 Residual Analysis: Checking the Assumptions

Our conclusion we draw about the population do only hold up if the assumptions we make about the population hold up.

When conducting a regression analysis we assume that the regression model is appropriate, that is:

1. the deterministic part of the model is linear in $x$.
2. the error is normally distributed, with mean 0 and standard deviation $\sigma$ (independent from $x$).

If the assumptions are reasonable we checked by analyzing a scattergram of the 2 variables of interest. We will now see how we can use further methods based on residuals to check the assumptions. The regression residuals are the differences for each measured value of $y$ with the estimate based on the estimated model. Let $i$ indicate an individual measurement, then

$$\hat{e}_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

is the residual for the $i$th measurement.

If the model is appropriate the residuals would be measurements coming from a normal distribution with mean 0 and standard deviation $\sigma$.

To check if the assumption of normality is appropriate, a QQ-plot of the residuals could be analyzed.

Continue example 1:

![QQ-Plot](image)

This QQ-plot shows no indication that the assumption of normality is violated.

If the model is appropriate then a scattergram of $x$ versus the residuals would show a random scatter of the points around zero, over the interval of observations of $x$, not showing any pattern or outliers. (Diagrams: pattern, outliers, spread depending on $x$)

Continue example 1: Too little data for this to carry sufficient information.
Continue example 2:
For this example will also use A QQ-plot to check for normality.

![Probability Plot of RESI1](RESI1.jpg)

The dots fall far off the line, but the CI included indicates no significant violation of normality.

![Scatterplot of RESI1 vs ad](RESI1 vs ad.jpg)

If the model is appropriate then a scattergram of $x$ versus the residuals would show a random scatter of the points around zero, over the interval of observations of $x$, not showing any pattern or outliers. In particular the scatter would be the same for all values of the predictor.

For this data this might be violated. There is lot of scatter around $ad=2$, but this not enough data to be conclusive.

(Diagrams: pattern, outliers, spread depending on $x$)
1.1.5 Applying the Model – Estimation and Prediction

Once we established that the model we found describes the relationship properly, we can use the model for

- estimation: estimate mean values of \( y \), \( E(y) \) for given values of \( x \). For example if the natural cause index equals 22, what value should we expect for the anthropogenic index?

- prediction: new individual value based on the knowledge of \( x \). For example, we score a different forest with a natural cause index of \( x = 22.5 \), what value do we predict for \( y = \) anthropogenic index.

Which application will we be able to answer with higher accuracy?
The key to both applications of the model is the least squares line, \( \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \), we found before, which gives us point estimators for both tasks.
The next step is to find confidence intervals for \( E(y) \) for a given \( x \), and for an individual value \( y \), for a given \( x \).

In order to this we will have to discuss the distribution of \( \hat{y} \) for the two tasks.

**Sampling distribution of \( \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \)**

1. The standard deviation of the distribution of the estimator \( \hat{y} \) of the mean value \( E(y) \) at a specific value of \( x \), say \( x_p \)

\[
\sigma_y = \sigma \sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{xx}}}
\]

This is the standard error of \( \hat{y} \).

2. The standard deviation of the distribution of the prediction error, \( y - \hat{y} \), for the predictor \( \hat{y} \) of an individual new \( y \) at a specific value of \( x \), say \( x_p \)

\[
\sigma_{(y-\hat{y})} = \sigma \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}}}
\]

This is the standard error of prediction.

**A \( (1 - \alpha)100\% \) CI for \( E(y) \) at \( x = x_p \)**

\[
\hat{y} \pm t^* s \sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{xx}}}
\]

**A \( (1 - \alpha)100\% \) Prediction Interval for \( y \) at \( x = x_p \)**

\[
\hat{y} \pm t^* s \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}}}
\]

Comment: Do a confidence interval, if interested in the mean of \( y \) for a certain value of \( x \), but if you want to get a range where a future value of \( y \) might fall for a certain value of \( x \), then do a prediction interval.
Continue example 1: The natural cause index for a forest was found to be $x_p = 27.5$, where should we expect the anthropogenic index to fall? Find a prediction interval.

For $x_p = 27.5$ find $\hat{y} = -53.14 + 3.4255(27.5) = 41.06$. With $df = 3$ $t^* = 3.182$

$$41.06 \pm 3.182(2.049) \sqrt{1 + \frac{1}{5} + \frac{(27.5 - 131/5)^2}{18.8}}, \text{ gives } 41.06 \pm 3.182(2.049)\sqrt{1.2899}$$

which is $41.06 \pm 5.736 = [35.324, 46.796]$. For a forest with natural cause index of 27.5 we expect with 95% confidence the anthropogenic index to fall between 35.3 and 46.8.

Estimate the mean of the anthropogenic index for forests with natural cause index of 27.5 at confidence level of 95%.

$$41.06 \pm 3.182(2.049) \sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}}}, \text{ gives } 41.06 \pm 3.182(2.049)\sqrt{0.2899}$$

which is $41.06 \pm 2.7194 = [38.8881, 44.3194]$. With 95% confidence the mean anthropogenic index of forests with natural cause index of 27.5 falls between 38.9 and 44.3.

Continue example 2: If the company decides to invest $2000 in advertisements, what are the sales they should expect in average? Estimate with 95% confidence.

For $x_p = 2$ find $\hat{y} = 104.06 + (50.73)2 = 205.52$. With $df = 6$ $t^* = 2.447$

$$205.52 \pm 2.447(10.29) \sqrt{\frac{1}{8} + \frac{(2 - 15.8/8)^2}{1.235}}, \text{ gives } 205.52 \pm 8.92$$

which is $[196.60, 214.44]$. If the they invest $2000$ the mean company sales fall with 95% confidence between $196600$ and $214440$.

We expect 95% of sales numbers to fall within the following interval, when the company invests $2000$ in advertisement.

$$205.52 \pm 2.447(10.29) \sqrt{\frac{1}{8} + \frac{(2 - 15.8/8)^2}{1.235}}, \text{ gives } 205.52 \pm 26.71$$

which is $[178.81, 232.23]$. 

\[15\]