1 Inferential Methods for Correlation and Regression Analysis

In the last section Correlation and Regression Analysis was studied as a method for describing bivariate continuous data. The sample Pearson Correlation Coefficient and the sample Regression Line were obtained for describing and measuring the quality and strength of the linear relationship between two continuous variables.

In this section a model for the underlying population will be developed. Then this model will be used to obtain inferential methods for this kind of data.

1.1 The Simple Linear Regression Model

Definition:
The simple linear regression model assumes that there is a line with intercept \( \beta_0 \) and slope \( \beta_1 \), called the true population regression line, that describes the relationship between the variable \( x \) and \( y \).

When a value of the independent variable \( x \) is fixed and an observation on the dependent variable \( y \) can be written as,

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

Basic Assumptions of the Simple Linear Regression Model

1. The distribution of the random deviation \( e \) has population mean \( \mu_e=0 \).
2. The standard deviation of \( e \) is the same for any particular value of \( x \). It is denoted by \( \sigma \).
3. The distribution of \( e \) is normal.

Randomness in \( e \) implies that the outcome of \( y \) for a given \( x \) is also random, so that \( y \) is a random variable.

The assumptions of the model imply:

That for every \( x \), \( y \) is a normal distributed random variable with mean \( \alpha + \beta x \) and standard deviation \( \sigma \).

The slope \( \beta_1 \) of the population regression line is the average change in \( y \) associated with a 1-unit increase in \( x \). The intercept \( \beta_0 \) is the height if the regression line when \( x = 0 \). The value of \( \sigma \) describes the extent to which \((x, y)\) observations deviate vertically from the regression line. Figure 15.3 in the textbook is a good illustration of the model.
1.2 Estimating the Population Regression Line

If the linear regression model seems to be appropriate for two population variables (random variables). The parameters $\beta_0$ and $\beta_1$ are usually unknown to the investigator. They can be estimated by using sample data $((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n))$.

Assume that the relationship of $x$ and $y$ can be described by a Simple Linear Regression Model and that the observations were drawn independently.

Under these assumptions point estimates of $\beta_0$ and $\beta_1$ are the intercept and the slope of the least squares line, respectively. That is

point estimator of $\beta_1 = b_1 = \frac{SS_{xy}}{SS_{xx}}$

point estimator of $\beta_0 = b_0 = \bar{y} - b_1 \bar{x}$

The estimated regression line is identical to the least squares line seen in the last chapter

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

Is $x^*$ a specific value of $x$, then $b_0 + b_1 x^*$ can be interpreted as

1. a point estimate of the mean of $y$, when $x = x^*$, and

2. a point estimate of an individual observation for $y$, when $x = x^*$.

In addition to estimating $\beta_0$ and $\beta_1$ there is also interest in estimating $\sigma$ the spread of the random deviation of the regression line. This estimate will be based on the Sum of Squares for Error:

$$SSE = \sum (y - \hat{y})^2$$

where $\hat{y}_1 = b_0 + b_1 x_1, \hat{y}_2 = b_0 + b_1 x_2, \ldots, \hat{y}_n = b_0 + b_1 x_n$ are the fitted or predicted $y$ values and the residuals are

$y_1 - \hat{y}_1, y_2 - \hat{y}_2, \ldots, y_n - \hat{y}_n$

The residuals give the vertical distance of the observations to the regression line. This makes $SSE$ a measure of the extent to which the the sample data spreads out about the estimated regression line.

**Estimation of $\sigma$**

The statistic for estimating the variance $\sigma^2$ is

$$s_e^2 = \frac{SSE}{n - 2}$$

where

$$SSE = \sum (y - \hat{y})^2 = \sum y^2 - b_0 \sum y - b_1 \sum xy$$

The estimate of the standard deviation $\sigma$ is

$$s_e = \sqrt{s_e^2}$$

It is also called the standard error of the estimate
σ measures the usual vertical deviation of a point \((x, y)\) in the population from the population regression line. Similarly, \(s_e\) measures the typical deviation of a sample point \((x, y)\) of the sample regression line.

**Example:**
Is cardiovascular fitness related to an athlete’s performance in a 20km ski race?

\[ x = \text{time to exhaustion running on a treadmill (in minutes)} \]
\[ y = \text{20km ski time (in minutes)} \]

<table>
<thead>
<tr>
<th></th>
<th>7.7</th>
<th>8.4</th>
<th>8.7</th>
<th>9.0</th>
<th>9.6</th>
<th>10.0</th>
<th>10.2</th>
<th>10.4</th>
<th>11.0</th>
<th>11.7</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>71.0</td>
<td>71.4</td>
<td>65.0</td>
<td>68.7</td>
<td>64.4</td>
<td>63.0</td>
<td>64.6</td>
<td>66.9</td>
<td>62.6</td>
<td>61.7</td>
</tr>
</tbody>
</table>

A scatterplot of this data shows a linear decrease in ski–time by increasing treadmill time.

We use a SLRM to relate the two variables

\[ y = \beta_0 + \beta_1 x + e, \quad e \sim \mathcal{N}(0, \sigma) \]

Straightforward calculation gives:

\[
\begin{align*}
 n &= 11 \\
 \sum x &= 106.3 \\
 \sum y &= 728.70 \\
 \sum x^2 &= 1040.95 \\
 \sum xy &= 7009.91 \\
 \sum y^2 &= 48,390.79
\end{align*}
\]

From these calculate

\[
SS_{xy} = \sum xy - \frac{(\sum x)(\sum y)}{n} = 7009.91 - \frac{(106.3)(728.70)}{11} = -31.9818
\]

and

\[
SS_{xx} = \sum x^2 - \frac{(\sum x)^2}{n} = 1040.95 - \frac{(106.3)^2}{11} = 13.705.
\]

\[
SS_{yy} = \sum y^2 - \frac{(\sum y)^2}{n} = 48,390.79 - \frac{(728.70)^2}{11} = 117.727.
\]
so that:
\[ r = \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}} = -0.796 \]

Indicating a strong negative linear relationship between the two variables.

To estimate the regression line:
\[ b_1 = \frac{SS_{xy}}{SS_{xx}} = -\frac{31.9818}{13.705} = -2.3335 \quad \text{and} \quad b_0 = \bar{y} - b \cdot \bar{x} = 66.2455 + 2.3335 \cdot 9.6636 = 88.7956 \]

The point estimate for the decrease in the mean ski–time for every minute increase in treadmill time is 2.33 minutes.

To estimate \( \sigma \)
\[
SSE = \sum y^2 - b_0 \sum y - b_1 \sum xy = 43.097
\]
\[
s^2_e = \frac{SSE}{n-2} = 4.789
\]
\[
s_e = \sqrt{s^2_e} = 2.188
\]

For athletes with the same treadmill time the ski race time has an estimated standard deviation of 2.19 minutes.

### 1.3 Residual Analysis

We will be using methods from inferential statistics to estimate \( \beta_1 \) using a confidence interval and to conduct tests concerning the slope. These methods are only appropriate, if the Simple Linear Regression Model is an appropriate description of the population, otherwise the results are misleading and might result in wrong conclusions.

Therefore it is necessary to at least have a look at the sample data to check if there are any indications that the model is violated.

**The model assumptions:**

- The error is normally distributed
- The standard deviation of the error is the same for all values of the predictor variable.

In order to check these assumptions using the sample data, we will use residuals, which are observed values for the error.

**Definition:**
A residual is the difference between an observed value of the response variable and the value predicted by the regression line.

\[ e_i = y_i - \hat{y}_i \]
\[ = y_i - (b_0 + b_1 x_i) \]

**Example:** The residuals for the example are
The mean of the residuals is always zero.

- In order to check the first assumption we will do a Normal Probability Plot, or a histogram. Here I chose to do a histogram for the residuals:

This histogram does not indicate a strong deviation from being bellshaped, it is symmetric and does not show outliers. We would find that there is no strong evidence against the assumption that the error is normal.

- To check if the standard deviation is the same for all values of $x$ we do a residual plot.
A residual plot is a scatterplot of the residuals against the explanatory variable.
The residual plot should show an even band of data points scattered around zero. As it does in this example.

Problems with the second assumption would be indicated if the residual plot either shows a pattern in the residual (see below), or if for example the standard deviation in the residuals changes with increasing $x$ (see below).

1.4 Inference for the Slope of the Population Regression Line

The slope $\beta_1$ is the average change in $y$, when $x$ is increased by 1 unit. This interpretation makes $\beta_1$ an interesting parameter. In addition we find, that if $\beta_1 = 0$ the Population Regression Line would be parallel to the $x$-axis, thus a change in $x$ would not have an impact on $y$, which would mean that $x$ and $y$ are independent.

These two properties lead us to ask for a confidence interval for $\beta_1$, in order to be able to estimate it more properly and for a statistical test concerning $\beta_1$.

Before we can obtain these methods we have to study the sampling distribution of the statistic to be used for estimating $\beta_1$, which is $b_1$.

Properties of the Sampling Distribution of $b_1$
When the basic assumptions of the simple linear regression model are met, the following claims hold:

1. The population mean of $b$ is $\beta$: $\mu_b = \beta_1$. Thus, $b_1$ is an unbiased estimate of $\beta_1$.

2. The population standard deviation of $b_1$ is

$$\sigma_{b_1} = \frac{\sigma}{\sqrt{SS_{xx}}}$$

3. The statistic $b_1$ has a normal distribution (which is a consequence of $e$ being normally distributed).

If $\sigma_{b_1}$ is large the estimate $b_1$ does not have to be close to the true $\beta_1$. In order for $\sigma_{b_1}$ to be small

- $\sigma$ should be small, that means little variation about the population regression line, and
- $(x - \bar{x})^2$ should be large, that means far spread out $x$ values are in favor for a smaller variation in the estimate $b_1$ for $\beta_1$.

**Standardizing $b_1$**

Since $\sigma_{b_1}$ depends on the unknown $\sigma$ it cannot be used for a standardizing $b_1$, in order to develop a statistic that can be used for developing a confidence interval or test. Instead use

$$s_{b_1} = \frac{s_e}{\sqrt{SS_{xx}}}$$

as estimate of $\sigma_{b_1}$ the standard deviation of $b_1$.

This leads to:

$$t = \frac{b_1 - \beta_1}{s_{b_1}}$$

which is t-distributed with $df = n - 2$ (for a Simple Linear Regression Model). This t-statistic does not depend on any unknown parameters besides $\beta_1$, so it can be used for developing a confidence interval and a test concerning $\beta_1$.

**Confidence Interval for $\beta_1$**

A $(1 - \alpha)\%$ Confidence Interval for the slope $\beta_1$ from a Simple Linear Regression Model is given by:

$$b_1 \pm (t \text{ critical value})^{n-2}_{1-\alpha} \cdot s_{b_1}$$

Where the t critical value is based on $df = n - 2$ and $1 - \alpha$ (Table IV).

**Continue example:** The calculation of a 95% confidence interval for $\beta_1$ requires the t critical value with $n-2=9$ degrees of freedom and $1 - \alpha = 0.95$: 2.26. The resulting confidence interval is then

$$b_1 \pm t \text{ critical value} \cdot s_{b_1} = -2.3335 \pm (2.26) \cdot (0.591) = -2.3335 \pm 1.336 = (-3.671; -0.999)$$

Based on the sample, we are 95% confident that the true average decrease in ski-time associated with a one minute increase in treadmill time is between 1 and 3.7 minutes.
Hypothesis Test Concerning $\beta_1$

- **Test type**

<table>
<thead>
<tr>
<th>Test type</th>
<th>$H_0$</th>
<th>$H_a$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Upper tail</td>
<td>$\beta \leq \beta_0$</td>
<td>$\beta &gt; \beta_0$</td>
</tr>
<tr>
<td>Lower tail</td>
<td>$\beta \geq \beta_0$</td>
<td>$\beta &lt; \beta_0$</td>
</tr>
<tr>
<td>Two tail</td>
<td>$\beta = \beta_0$</td>
<td>$\beta \neq \beta_0$</td>
</tr>
</tbody>
</table>

Choose $\alpha$.

- **Assumptions:** The Simple Linear Regression Model applies to $x$ and $y$.

- **Test statistic:**
  
  $$t_0 = \frac{b_1 - \beta_0}{s_{b_1}}$$
  
  with $df = n - 2$

- **P-value:**

<table>
<thead>
<tr>
<th>Test type</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Upper tail</td>
<td>$P(t &gt; t_0)$</td>
</tr>
<tr>
<td>Lower tail</td>
<td>$P(t &lt; t_0)$</td>
</tr>
<tr>
<td>Two tail</td>
<td>$2 \cdot P(t &gt; abs(t_0))$</td>
</tr>
</tbody>
</table>

- **Decision:**

- **Context:**

  The two tail test of $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 \neq 0$ is a test to assure that there is in fact a linear connection between $x$ and $y$, it is also called “Model Utility Test”.

**Continue Example:**

Let’s do the Model Utility Test for the example above at a significance level of $\alpha = 0.05$.

1. **Hypotheses:**

   $$H_0 : \beta_1 = 0 \text{ versus } H_a : \beta_1 \neq 0$$

   and $\alpha = 0.05$. Do Model Utility Test.

2. **Assumptions:** Assume the relationship of $x$ and $y$ can be described by a Simple Linear Regression Model:

3. **Test statistic:**

   $$t_0 = \frac{b_1 - 0}{s_{b_1}} = \frac{-2.3335}{0.591} = -3.948 \text{ with } df = 9$$
4. **P-value**: This is a two tail test, thus (according to table IV) 
\[ p-value = 2 \cdot P(t > abs(t_0)) < 2 \cdot 0.005 = 0.01 \]

5. **Decision**: Since the p-value \( \leq \alpha \) the null hypothesis is rejected.

6. **Conclusion**: We conclude that \( \beta_1 \neq 0 \) and there is a linear relationship between ski-time and treadmill time.

The investigators are asking if there is statistical evidence to say that every additional minute on the treadmill will reduce the ski time by at least 2 minutes.

1. **Hypotheses**: 
   \( H_0 : \beta_1 \geq -2 \) versus \( H_a : \beta_1 < -2 \), at significance level \( \alpha = 0.05 \). Use Hypothesis Test Concerning \( \beta \).

2. **Assumptions**: Assume treadmill time and ski time are properly described by a simple linear regression model (Model Utility Test was significant).

3. **Test statistic**:
   \[ t_0 = \frac{b_1 - \beta_0}{s_{b_1}} = \frac{-2.3335 - (-2)}{0.591} = -0.564 \quad \text{with} \quad df = 9 \]

4. **P-value**: This is a lower tail test test, thus
   \[ p-value = P(t < t_0) = P(t > -t_0) = P(t > 0.564) > 0.1 \]

5. **Decision**: Since the p-value > \( \alpha \) the null hypothesis can not be rejected.

6. **Conclusion**: Even though the average decrease in ski time is greater than 2 \( (b = -2.3335) \) in the sample, the sample does not provide enough evidence that \( \beta_1 \) ((population) mean decrease in ski time for every extra minute on the treadmill) is less than \(-2\) at significance level of 5%.