PSTAT 126

Regression Analysis

Laura Baracaldo

Lecture 4
Inference

Inference and Normality assumption

1. **Estimation**: First step in Statistical Inference.

   - Point Estimates. **LS** No need of distributional assumptions, **MLE** We need distribution assumptions on the errors.
   - Interval Estimation. In order to construct Confidence Intervals we need distribution assumptions on the errors.

2. **Hypothesis Testing**: We may have a prior judgement/ believe about what values the parameters assume. We need distributional assumptions on the errors.

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, ..., n$$

We assume:

- $\epsilon_i | x_i \overset{i.i.d}{\sim} N(0, \sigma^2)$. This implies: $y_i | x_i \overset{ind}{\sim} N(\beta_0 + \beta_1 x_i, \sigma^2)$

Maximum Likelihood Estimation (MLE)

If $y_i|x_i \overset{ind}{\sim} N(\beta_0 + \beta_1 x_i, \sigma^2)$, the likelihhood function based on observations $y_1, \ldots, y_n$ can be written as:

$$\begin{aligned}
L(\beta_0, \beta_1, \sigma^2) &= \prod_{i=1}^{n} f_i(y_i|x_i) \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ \frac{-(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2} \right\} \\
&= \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\left\{ \frac{-\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2} \right\}
\end{aligned}$$

We can derive the *Maximum Likelihood Estimates (MLE)* of parameters β_0, β_1 and σ^2 by solving:

$$\underset{\beta_0, \beta_1, \sigma^2}{\arg\max}\, L(\beta_0, \beta_1, \sigma^2)$$

Maximum Likelihood Estimation (MLE)

This is equivalent to maximize the log-likelihood:

$$\arg\max_{\beta_0,\beta_1,\sigma^2} l(\beta_0, \beta_1, \sigma^2)$$

Where:

$$\begin{aligned} l(\beta_0, \beta_1, \sigma^2) &= \log\left[L(\beta_0, \beta_1, \sigma^2)\right] \\ &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2 \end{aligned}$$

MLE for β_0, β_1 and σ^2

By taking the derivatives with respect to β_0, β_1 amd σ^2 we get the equations:

$$\frac{\partial l}{\partial \beta_0} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)$$

$$\frac{\partial l}{\partial \beta_1} = \frac{1}{\sigma^2} \sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i)$$

$$\frac{\partial l}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

MLE for β_0, β_1 and σ^2

By setting the two first equations equal to zero we obtain:

$$\hat{\beta}_{1_{MLE}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad \text{and} \qquad \hat{\beta}_{0_{MLE}} = \bar{y} - \hat{\beta}_{1_{MLE}}\bar{x}$$

- Which means: **The MLE estimates of β_0 and β_1 correspond to the LS estimates!**

We get the MLE for σ^2 by setting the third equation equal to zero:

$$\begin{aligned} \hat{\sigma}^2_{MLE} &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\beta}_{0_{MLE}} - \hat{\beta}_{1_{MLE}}x_i)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{SSR}{n} \end{aligned}$$

- The MLE estimate of σ^2 is different from the LS estimate. Moreover, $\hat{\sigma}^2_{MLE}$ is *biased*.

Inference on β_0 and β_1

We can drive inference on $\hat{\beta}_0$ and $\hat{\beta}_1$ by deriving their distributions. Since $\hat{\beta}_0$ and $\hat{\beta}_1$ can be written as linear combination of normal random variables, it can be proved that:

- $\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)$
- $\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)\right)$

When σ^2 is known we can calculate confidence intervals and do hypothesis testing based of the normality of $\hat{\beta}_0$ and $\hat{\beta}_1$. But in real life problems σ^2 is unknown. What do we do in then?

Inference on β_0 and β_1

We must derive some properties on $\hat{\sigma}^2_{LS}$:

1. **Distribution:** $\frac{(n-2)\hat{\sigma}^2_{LS}}{\sigma^2} = \frac{SSR}{\sigma^2} \sim \chi^2_{(n-2)}$.
2. **Independence:** $\frac{SSR}{\sigma^2}$ is independent of $\hat{\beta}_0$ and $\hat{\beta}_1$.

Inference on β_0 and β_1

From 1 and 2 we can prove that:

- $T_0 = \frac{\hat{\beta}_0 - \beta_0}{\sqrt{MSE\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)}} \sim t_{(n-2)}$
- $T_1 = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{MSE}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}} \sim t_{(n-2)}$

With $MSE = \hat{\sigma}^2_{LS} = \frac{SSR}{n-2}$

Confidence Intervals for β_0 and β_1

We want to construct $100(1-\alpha)\%$ confidence intervals for β_0 and β_1.

- $P(-t_{1-\alpha/2;n-2} \leq T_k \leq t_{1-\alpha/2;n-2}) = 1-\alpha \qquad k = 0, 1.$

Where $t_{1-\alpha/2;n-2}$ denotes the $(1-\alpha/2)100$ percentile of the $t-$distribution with $df = n-2$.

Therefore, the $100(1-\alpha)\%$ confidence intervals for β_0 and β_1 can be constructed as:

- $100(1-\alpha)\%$ CI for β_0: $\hat{\beta}_0 \pm t_{1-\alpha/2;n-2}\sqrt{MSE\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)}$
- $100(1-\alpha)\%$ CI for β_1: $\hat{\beta}_1 \pm t_{1-\alpha/2;n-2}\sqrt{\frac{MSE}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$

Hypothesis Testing

Suppose we want to test the hypothesis:

$$H_0 : \beta_k = b_{k,0} \qquad H_1 : \beta_k \neq b_{k,0}$$

Where $b_{k,0}$ is a fixed value, $k = 0, 1$.

- The test statistic is:

$$T_k^* = \frac{\hat{\beta}_k - b_{k,0}}{\sqrt{V\hat{a}r(\hat{\beta}_k)}}$$

- Under H_0: $T_k^* \sim t_{(n-2)}$. Thus, for a significance level of $\alpha(100)\%$ we reject H_0 if $|T_k^*| > t_{1-\alpha/2;n-2}$.

Hypothesis Test for Linear association

In Linear Regression Analysis we seek to investigate the true linear association between x and y. It is possible to drive Statistical inference on this linear relationship by testing the hypothesis on β_1:

$$H_0 : \beta_1 = 0 \qquad H_1 : \beta_1 \neq 0$$

- $T_1^* = \frac{\hat{\beta}_1}{\sqrt{\hat{Var}(\hat{\beta}_1)}} = \frac{\hat{\beta}_1}{\sqrt{\frac{MSE}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}}$

Species Example - Inference on β_0, β_1

We can construct 95% Confidence Intervals for β_0, β_1:

```
data(gala, package ="faraway")
fit<- lm( Species ~ Elevation, data=gala)

CI.beta0<- c(fit$coefficients[1] - qt(0.975, df=fit$df.residual)*se.beta0,
              fit$coefficients[1] + qt(0.975, df=fit$df.residual)*se.beta0)
CI.beta0
```

```
## (Intercept) (Intercept)
##   -28.00514    50.67536
```

```
CI.beta1<- c(fit$coefficients[2] - qt(0.975, df=fit$df.residual)*se.beta1,
              fit$coefficients[2] + qt(0.975, df=fit$df.residual)*se.beta1)
CI.beta1
```

```
## Elevation Elevation
## 0.1298223 0.2717621
```

```
confint(fit)
```

```
##                   2.5 %     97.5 %
## (Intercept) -28.0051367 50.6753632
```

Species Example - Inference on β_0, β_1

We want to test whether elevation is statistically relevant when explaining the number of species:

$$H_0 : \beta_1 = 0 \qquad H_1 : \beta_1 \neq 0$$

```
data(gala, package ="faraway")
fit<- lm( Species ~ Elevation, data=gala)
T1<- fit$coefficients[2]/se.beta1;T1 # t value
```

```
## Elevation
##  5.795475
```

```
t1 <- qt(0.975, df=fit$df.residual);t1
```

```
## [1] 2.048407
```

```
if(T1>t1){print("Reject H0")
  }else{
    print("Fail to Reject H0")}
```

```
## [1] "Reject H0"
```

Species Example - Inference on β_0, β_1

```
##
## Call:
## lm(formula = Species ~ Elevation, data = gala)
##
## Residuals:
##      Min        1Q    Median        3Q       Max
## -218.319   -30.721   -14.690     4.634   259.180
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.33511   19.20529   0.590     0.56
## Elevation    0.20079    0.03465   5.795 3.18e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 78.66 on 28 degrees of freedom
## Multiple R-squared:  0.5454, Adjusted R-squared:  0.5291
## F-statistic: 33.59 on 1 and 28 DF,  p-value: 3.177e-06
```