PSTAT 126

Regression Analysis

Laura Baracaldo & Rodrigo Targino

Lecture 3
Simple Linear Regression Models Part II

Simple Linear Regression Model Assumptions

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

We assume:

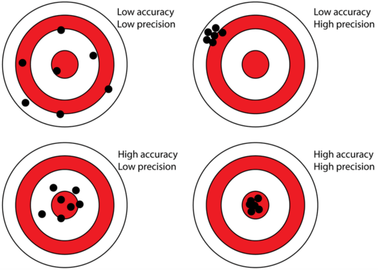
- The errors, which capture the variations unexplained by the systematic/linear component, ϵ_i, $i = 1, \ldots, n$ are unobservable i.i.d random variables.
- $E(\epsilon_i) = 0$ and $Var(\epsilon_i) = \sigma^2$.

This implies:

- $E(y_i) = \beta_0 + \beta_1 x_i$
- $Var(y_i) = \sigma^2$.
- $Cov(y_i, y_j) = 0$

Besides estimating *parameters* β_0 and β_1, we aim to estimate the random error variance σ^2, which is typically unknown.

Accuracy & Precision of the Coefficient Estimates



We evaluate the performance of our *statistics* $\hat{\beta}_0$ and $\hat{\beta}_1$ when estimating β_0 and β_1 respectively, in terms of accuracy (**Bias**) and precision (**Variance**).

Accuracy & Precision of the Coefficient Estimates

Bias: It can be proved that $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased estimators of β_0 and β_1:

- $Bias(\hat{\beta}_0) = E(\hat{\beta}_0) - \beta_0 = 0 \Rightarrow E(\hat{\beta}_0) = \beta_0$
- $Bias(\hat{\beta}_1) = E(\hat{\beta}_1) - \beta_1 = 0 \Rightarrow E(\hat{\beta}_1) = \beta_1$

Variance: It can be shown that:

- $Var(\hat{\beta}_0) = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right]$
- $Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$

Accuracy of $\hat{\beta}_1$

$$
\begin{aligned}
E(\hat{\beta}_1) &= E\left[\frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}\right] \\
&= \frac{E\left[\sum_{i=1}^n (x_i - \bar{x}) y_i - \bar{y} \sum_{i=1}^n (x_i - \bar{x})\right]}{\sum_{i=1}^n (x_i - \bar{x})^2} \\
&= \frac{\sum_{i=1}^n (x_i - \bar{x}) E\left[y_i\right]}{\sum_{i=1}^n (x_i - \bar{x})^2} \\
&= \frac{\sum_{i=1}^n (x_i - \bar{x})(\beta_0 + \beta_1 x_i)}{\sum_{i=1}^n (x_i - \bar{x})^2} \\
&= \frac{\beta_0 \sum_{i=1}^n (x_i - \bar{x}) + \beta_1 \sum_{i=1}^n (x_i - \bar{x}) x_i}{\sum_{i=1}^n (x_i - \bar{x})^2} \\
&= \beta_1 \frac{\sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x} + \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \beta_1 \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} = \beta_1
\end{aligned}
$$

Accuracy of $\hat{\beta}_0$

$$
\begin{aligned}
E(\hat{\beta}_0) &= E\left[\bar{y} - \hat{\beta}_1 \bar{x}\right] \\
&= \beta_0 + \beta_1 \bar{x} - \beta_1 \bar{x} = \beta_0
\end{aligned}
$$

Task: Fill in the details!!

Precision of $\hat{\beta}_1$

$$
\begin{aligned}
Var(\hat{\beta}_1) &= Var\left[\frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right] \\
&= \frac{Var\left[\sum_{i=1}^{n}(x_i - \bar{x})y_i - \bar{y}\sum_{i=1}^{n}(x_i - \bar{x})\right]}{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^2} \\
&= \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 Var[y_i]}{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^2} \\
&= \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}
\end{aligned}
$$

Precision of $\hat{\beta}_0$

$$
\begin{aligned}
Var(\hat{\beta}_0) &= Var\left(\bar{y} - \hat{\beta}_1 \bar{x}\right) \\
&= Var(\bar{y}) + Var(\hat{\beta}_1 \bar{x}) - 2Cov(\bar{y}, \hat{\beta}_1 \bar{x}) \\
&= \frac{\sigma^2}{n} + \frac{\bar{x}^2 \sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} - 2\bar{x}Cov(\bar{y}, \hat{\beta}_1)
\end{aligned}
$$

$$
\begin{aligned}
Cov(\bar{y}, \hat{\beta}_1) &= Cov\left(\frac{\sum_{i=1}^n y_i}{n}, \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}\right) \\
&= \frac{1}{n\sum_{i=1}^n (x_i - \bar{x})^2} Cov(\sum_{i=1}^n y_i, \sum_{i=1}^n (x_i - \bar{x}) y_i) = 0
\end{aligned}
$$

Precision of $\hat{\beta}_0$

$$\Rightarrow Var(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right)$$

Task: Fill in the details!! (Hint: Use the fact that $Cov(y_i, y_j) = 0, i \neq j$)

Gauss-Markov Theorem

The LS estimate $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1)^T$ is unbiased and has minimum variance among all unbiased linear estimators of β. In other words, $\hat{\beta}$ is said to be the best linear unbiased estimate (BLUE) of β.

It can be proved that for any other unbiased linear estimate β^*:

$$Var(\beta^*) \geq Var(\hat{\beta})$$

Estimating the variance σ^2

We assumed that $\sigma^2 = Var(\epsilon_i) = E(\epsilon_i^2) - [E(\epsilon_i)]^2 = E(\epsilon_i^2)$, (since $E(\epsilon_i) = 0$). Based on this, we could think of a natural estimate for the error variance: $\hat{\sigma}^2 = \frac{\sum_{i=1}^n \hat{\epsilon}^2}{n}$. However, it is necessary to correct by the 2 degrees of freedom (df) that were used to estimate the two parameters β_0 and β_1:

$$\hat{\sigma}^2 = MSE = \frac{\sum_{i=1}^n \hat{\epsilon}^2}{n-2}$$

Where MSE stands for *Mean Squared Error*.

It can be proved that $\hat{\sigma}^2$ is an unbiased estimator for σ^2:

$$E(\hat{\sigma}^2) = \sigma^2$$

Goodness of fit

We can measure how well the model fits the data. One way to do so is by calculating R^2, the so-called *coefficient of determination* or *percentage of variance explained*:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{SSR}{SST}$$

SSR: Residual Sum of Squares, *SST*: Total sum of squares corrected by the mean.

Its range is $0 \leq R^2 \leq 1$. Values closer to 1 indicate better fit. For simple linear regression $R^2 = r^2$, where r^2 is the correlation coefficient between x and y.
Interpretation: Proportion of the variability of y that can be explained by using x.

Species Example - Estimating σ^2

We can estimate σ^2 using the residuals from the fitted linear model:

```
data(gala, package ="faraway")
fit<- lm( Species ~ Elevation, data=gala)
sigma2.hat <- sum((fit$residuals^2))/fit$df.residual
sigma2.hat
```

```
## [1] 6187.638
```

```
sigma.hat <- sqrt(sigma2.hat) # Residual Standard Error
sigma.hat
```

```
## [1] 78.66154
```

Species Example - Estimating SE of $\hat{\beta}_0$, $\hat{\beta}_1$

```
data(gala, package ="faraway")
y<- gala$Species
x<- gala$Elevation
n<- length(y)
se.beta1<- sigma.hat/sqrt(sum((x-mean(x))^2))
se.beta1
```

```
## [1] 0.03464637
```

```
se.beta0<- sigma.hat*sqrt((1/n+mean(x)^2/sum((x-mean(x))^2)))
se.beta0
```

```
## [1] 19.20529
```

Species Example - Calculating R^2

```
data(gala, package ="faraway")
y<- gala$Species
x<- gala$Elevation
n<- length(y)
R.2 <- 1- sum((fit$residuals)^2 )/(sum((y-mean(y))^2))
R.2
```

```
## [1] 0.5453625
```

Species Example - Summary

```
##
## 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
```