PSTAT 126

Regression Analysis

Laura Baracaldo

Lecture 5
Prediction

Estimation of Expected Response $E(y_k)$

1. **Point Estimation:** Suppose we seek to estimate the average response conditioned on the predictor: $E(y_k) = E(y_k|x_k) = \beta_0 + \beta_1 x_k$ for $k = 1, \ldots, n$. The natural solution is to plug in the estimates of β_0 and β_1:

$$\hat{E}(y_k) = \hat{y}_k = \hat{\beta}_0 + \hat{\beta}_1 x_k$$

2. **Interval Estimation:** We can provide a confidence interval for $E(y_k)$ based on the sampling distribution of $\hat{y}_k$.

Sampling Distribution of $\hat{y}_k$

- **Normality:** We have proved that $\hat{\beta}_0$ and $\hat{\beta}_1$ are normally distributed. Since $\hat{y}_k$ is a linear combination of normal random variables, it will follow a normal distribution as well.
- **Expected Value:** It can be easily shown that $\hat{y}_k$ is an *unbiased* estimator of $E(y_k)$:

$$E(\hat{y}_k) = E(\hat{\beta}_0 + \hat{\beta}_1 x_k) = \beta_0 + \beta_1 x_k = E(y_k)$$

Sampling Distribution of $\hat{y}_k$

- **Variance:** We can derive the variance of $\hat{y}_k$:

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

$(**)$ Remember: $Cov(\bar{y}, \hat{\beta}_1) = 0$

Sampling Distribution of $\hat{y}_k$

- Thus $\hat{y}_k \sim N\left(E(y_k), \sigma^2\left[\frac{1}{n} + \frac{(x_k-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right]\right)$. Or equivalently:

$$\frac{\hat{y}_k - E(y_k)}{\sqrt{\sigma^2\left[\frac{1}{n} + \frac{(x_k-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right]}} \sim N(0,1)$$

When replacing σ^2 by its estimate $\hat{\sigma}^2 = MSE$:

$$T_k = \frac{\hat{y}_k - E(y_k)}{\sqrt{MSE\left[\frac{1}{n} + \frac{(x_k-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right]}} \sim t_{n-2}$$

Confidence Interval for $E(y_k)$

- $P(-t_{1-\alpha/2;n-2} \leq T_k \leq t_{1-\alpha/2;n-2}) = 1-\alpha$

$\Rightarrow P(\hat{y}_k - t_{1-\alpha/2;n-2}\hat{SE}(\hat{y}) \leq E(y_k) \leq \hat{y} + t_{1-\alpha/2;n-2}\hat{SE}(\hat{y}_k)) = 1-\alpha$.

A $100 * (1-\alpha)\%$ CI for $E(y_k)$ is:

$$\hat{y} \pm t_{1-\alpha/2;n-2}\hat{SE}(\hat{y}_k)$$

With $\hat{SE}(\hat{y}_k) = \sqrt{MSE\left[\frac{1}{n} + \frac{(x_k-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right]}$

Prediction for a New/Future Observation

Now let's suppose we want to *predict* the individual response for a future or new observation y^* given an observed predictor x^*:

$$y^* = \beta_0 + \beta_1 x^* + \epsilon^*$$

With y^* independent of $y_1, \ldots, y_n$, $\epsilon^* \sim N(0, \sigma^2)$.

1. **Pointwise Prediction:** The natural choice for the prediction is:

$$\hat{y}^* = \hat{\beta}_0 + \hat{\beta}_1 x^*$$

Prediction for a New/Future Observation

2. **Interval Prediction:** We derive a predictive interval for y^* based on the sampling distribution of $m_k = \hat{y}^* - y^*$:

- **Normality:** m_k follows a normal distribution, since it can be written as a linear combination of normal random variables.
- **Expected value:**

$$E(m_k) = E(\hat{y}^* - y^*) = E(\hat{\beta}_0 + \hat{\beta}_1 x^* - (\beta_0 + \beta_1 x^* + \epsilon^*)) = 0$$

- **Variance:**

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

Confidence Interval for y^*

- $P(-t_{1-\alpha/2;n-2} \leq m_k \leq t_{1-\alpha/2;n-2}) = 1 - \alpha$

$\Rightarrow$ A $100 * (1 - \alpha)\%$ CI for y^* is:

$$\hat{y}^* \pm t_{1-\alpha/2;n-2}\hat{SE}pred(\hat{y}^*)$$

With $\hat{SE}pred(\hat{y}^*) = \sqrt{MSE\left[1 + \frac{1}{n} + \frac{(x^*-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right]}$

Goodness of fit - R^2 and $\bar{R}^2$

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 (Although this depends on the application).

Goodness of fit - R^2 and $\bar{R}^2$

For simple linear regression $R^2 = r^2$, where r^2 is the correlation coefficient between x and y.
Task: Try to prove this!!

Interpretation: Proportion of the variability of y that can be explained by using x.

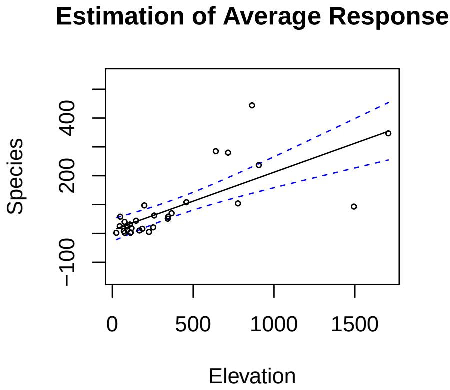
- **Adjusted** R^2: It adjusts by for the number or independent variables in a model.

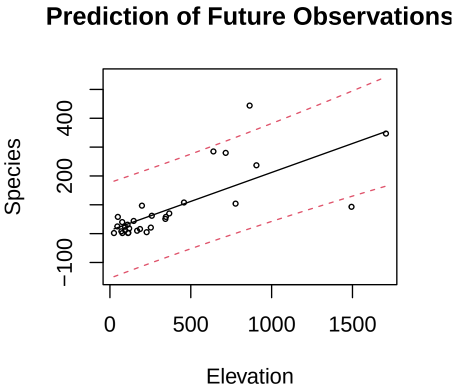
$$\bar{R}^2 = 1 - \frac{SSR/(n-2)}{SST/(n-1)}$$

Species Example - Prediction

```
data(gala, package ="faraway")
fit1<- lm( Species ~ Elevation, data=gala)
y<- gala$Species
x<- gala$Elevation
par(mfrow=c(1,2))
grid <- seq(min(x),max(x),len=100)
p1 <- predict(fit1, newdata=data.frame(Elevation=grid), se=T,
              interval="confidence")
p2 <- predict(fit1, newdata=data.frame(Elevation=grid), se=T,
              interval="prediction")
matplot(grid,p1$fit,lty=c(1,2,2),col=c(1,2,2),type="l",
        xlab="Elevation",ylab="Species",ylim=range(p1$fit,p2$fit,y))
points(x,y,cex=.5)
title("Estimation of Average Response")
matplot(grid,p2$fit,lty=c(1,2,2),col=c(1,2,2),type="l",
        xlab="Elevation",ylab="Species",ylim=range(p1$fit,p2$fit,y))
points(x,y,cex=.5)
title("Prediction of Future Observations")
```

Species Example - Prediction





Species Example - R^2 and $\bar{R}^2$

```r
R2<- cor(x,y)^2;R2
```

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

```r
R2.adjusted<- 1- (sum((fit1$residuals)^2)/fit1$df.residual)/
  (sum((y-mean(y))^2)/(n-1)); R2.adjusted
```

```
## [1] 0.5291255
```

```
##
## 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
```