1.7 Estimation and Prediction in SLR

The LS line can be used to obtain values of the response ($Y^\star$) for given values of the predictor ($X = x^\star$). There are **two variants** of this problem[8]:

- **Estimation**: Estimation of the *mean response* at $x^\star$, i.e. $\mathbb{E}(Y|X = x^\star)$. This is equivalent to estimating: $\beta_0 + \beta_1 x^\star$
- **Prediction**: Prediction of an outcome of random variable $Y^\star$ at a given value $x^\star$, where

$$Y^\star \sim \quad (\beta_0 + \beta_1 x^\star, \sigma^2)$$

The fitted value (or point estimate) for estimation and prediction is the same:

$$\hat{\beta}_0 + \hat{\beta}_1 x^\star.$$

However, the **accuracy for estimation and the one for prediction is different**. By "accuracy", we mean the expected value of the squared difference between the point estimate and the target.

1.7.1 Estimation of the Mean Response

For **estimation**, the target is $\beta_0 + \beta_1 x^\star$:

$$
\begin{aligned}
&\mathbb{E}\ \hat\beta_0 + \hat\beta_1 x^\star - \beta_0 - \beta_1 x^\star\ ^2 \\
=&Var\ \ \hat\beta_0 + \hat\beta_1 x^\star \\
=&Var\ \ \hat\beta_0\ \ + (x^\star)^2 Var\ \ \hat\beta_1\ \ + 2x^\star Cov\ \ \hat\beta_0, \hat\beta_1 \\
=&\sigma^2\ \ \frac{1}{n} + \frac{(x^\star - \bar{x})^2}{{}_i(x_i - \bar{x})^2} \\
=&\sigma^2\ \ \frac{1}{n} + \frac{(x^\star - \bar{x})^2}{S_{xx}}\ \ .
\end{aligned}
$$

Recall that all our calculations are done conditionally on $x^\star$.

A confidence interval is always reported for a *parameter.*

Confidence Interval for $\mathbb{E}(Y^\star)$

An $(1-\alpha)100\%$ Confidence Interval for the Mean Response when $x = x^\star$} is given by

$$
\hat\beta_0 + \hat\beta_1 x^\star \pm T_{n-2}(\alpha/2)\ \hat\sigma\ \ \frac{1}{n} + \frac{(x^\star - \bar{x})^2}{S_{xx}}
$$

1.7.2 Prediction

For **prediction**, the target is $Y^\star = \beta_0 + \beta_1 x^\star + e^\star$, where $e^\star \sim N(0, \sigma^2)$. This new error $e^\star$ is **independent of the previous n data points**, and as a result independent of $(\hat\beta_0, \hat\beta_1)$. Therefore,

$$\begin{aligned}
&\mathbb{E}[(\hat{\beta}_0 + \hat{\beta}_1 x^\star - Y^\star)^2] \\
&= \mathbb{E}[(\hat{\beta}_0 + \hat{\beta}_1 x^\star - \beta_0 - \beta_1 x^\star - e^\star)^2] \\
&= \mathbb{E}[(\hat{\beta}_0 + \hat{\beta}_1 x^\star - \beta_0 - \beta_1 x^\star)^2] + \mathbb{E}[(e^\star)^2] \\
&= \sigma^2 \ 1 + \frac{1}{n} + \frac{(x^\star - \bar{x})^2}{S_{XX}}
\end{aligned}$$

A prediction interval is reported for the value of a *random variable*, for example, $Y^\star$.

Prediction Interval

An $(1-\alpha)100\%$ Prediction Interval for $\hat{Y}^\star$ when $x = x^\star$ is given by

$$\hat{\beta}_0 + \hat{\beta}_1 x^\star \pm T_{n-2}(\alpha/2)\ \hat{\sigma}\ \sqrt{1 + \frac{1}{n} + \frac{(x^\star - \bar{x})^2}{S_{xx}}}$$

Remarks:

- Based upon the $Var(\hat{Y}^\star)$, the prediction interval is wider than the interval used to estimate the mean response at fixed $x = x^\star$.
- So far, we have assumed that the x-levels are known constants. So, all the previous results hold if:
  - $f(y|x)$ are independent and normally distributed with mean $\beta_0 + \beta_1 x$ and variance σ^2 conditionally on x.
  - x are independent with distribution $g(x_i)$ that does not depend on β_0, β_1, or σ^2.

University Admissions Example (Revisited)

g. What is the predicted `gpa` for a student with `entrance_score` equal to 2.1?

As we discussed, the **point estimate** is obtained by plugging in the “x” value to the fitted regression line. (You can also check this “by hand”.)

```
predict(admissions.lm, newdata=data.frame(entrance_score=2.1))
```

```
##        1
## 4.688618
```

If we want to obtain a **confidence interval** we write:

```
predict(admissions.lm, newdata=data.frame(entrance_score = 2.1), interval="co
```

```
##        fit      lwr      upr
## 1 4.688618 4.462133 4.915103
```

On the other hand, if we want to obtain a **prediction interval** we write:

```
predict(admissions.lm, newdata=data.frame(entrance_score = 2.1), interval="pr
```

```
##        fit      lwr      upr
## 1 4.688618 3.780076 5.597159
```

R **Remark:** In order for the function `predict` to work, you need to input a `data.frame` in the `newdata` attribute.