STAT 425

Simple Linear Regression. Part 2

Properties of the LS estimates

- We will study the statistical properties of $(\hat{\beta}_0, \hat{\beta}_1)$ as the LS estimates of the true parameter vector (β_0, β_1)
- We will compute the mean, variance and covariance of $(\hat{\beta}_0, \hat{\beta}_1)$ and check that they are unbiased estimators.
- We can also show that they achieve the smallest mean square error (MSE) among all unbiased estimators, but we will show this result as a general result when discussing Multiple Linear Regression (MLR).
- Until this point, we only need to make the following assumptions about the 1st and 2nd moments of residuals e_i's:
  $E[e_i] = 0,\ Var(e_i) = \sigma^2,\ Cov(e_i, e_j) = 0,\ i \neq j$

- For hypothesis testing and constructing confidence/prediction intervals we need to derive the probability distribution of $(\hat{\beta}_0, \hat{\beta}_1)$
- We will assume that the e_i's have a normal distribution and are independent and identically distributed (i.i.d.). We will use the t distribution for hypothesis testing and confidence intervals. With assumptions about the 1st and 2nd order moments and for a large sample size, we can use the CLT and assume a normal distribution instead.
- Notation: Uppercase letters are normally used for Random Variables, and lowercase letters for observed values of the random variables. Uppercase letter will be also reserved for matrices. In some occasions lowercase letter will also be used for random variables.

Least-squares estimates properties

Lets assume: $y_i = \beta_0 + \beta_1 x_i + e_i$ and 1st and 2nd moments for the e_i's:

$$E[e_i] = 0, \quad \text{and} \quad Cov[e_i, e_j] = \sigma^2 \delta_{ij}$$

where $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ if $i \neq j$. The assumptions 3 about the 1st and 2nd moments of the error terms leads to the following assumption on the 1st and 2nd order moments of y conditioning on X:

$$E[y_i|x_i] = \beta_0 + \beta_1 x_i, \quad , Cov[y_i, y_j|x_i, x_j] = \sigma^2 \delta_{ij}$$

In stat425, the statistical assumption is on the conditional distribution of Y given X. So when we evaluate expectations, only y_i's are random and x_i's are treated as known, non-random constants.

The LS estimates are unbiased

$$\hat{\beta}_1 = \sum_i \frac{(x_i - \bar{x})}{S_{XX}} y_i = \sum_i c_i y_i \quad (\sum_i c_i = 0)$$

$$E[\hat{\beta}_1] = \sum_i c_i E[y_i] = \sum_i c_i(\beta_0 + \beta_1 x_i) = \beta_1 \left(\sum_i x_i c_i \right) = \beta_1$$

$$\hat{\beta}_0 = \bar{y} - \bar{x}\hat{\beta}_1$$

$$E[\hat{\beta}_0] = \frac{1}{n} \sum_i E[y_i] - E[\hat{\beta}_1]\bar{x} = \beta_0 + \bar{x}\beta_1 - \bar{x}\beta_1 = \beta_0$$

1

[1] $\sum_i x_i c_i = 1$

Mean Square Error (MSE) of the LS estimates

Since the LS estimates are unbiased, their MSE[2].

$$Var(\hat{\beta}_1) = Var(\sum_i c_i y_i) = \sigma^2 \sum_i c_i^2 = \sigma^2 \frac{1}{S_{XX}}$$

$$Var(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}} \right)$$

[3] Note that both variances depend reciprocally on S_{XX}. The larger the distance of x_i from the sample mean $\bar{x}$, the larger the value of S_{XX} and the smaller the Variance of the LS estimates.

---

[2] $MSE(\hat{\beta}) = E[\beta - \hat{\beta}]^2 = E[\hat{\beta} - E[\hat{\beta}]]^2$ is equal to the Variance of $\hat{\beta}$

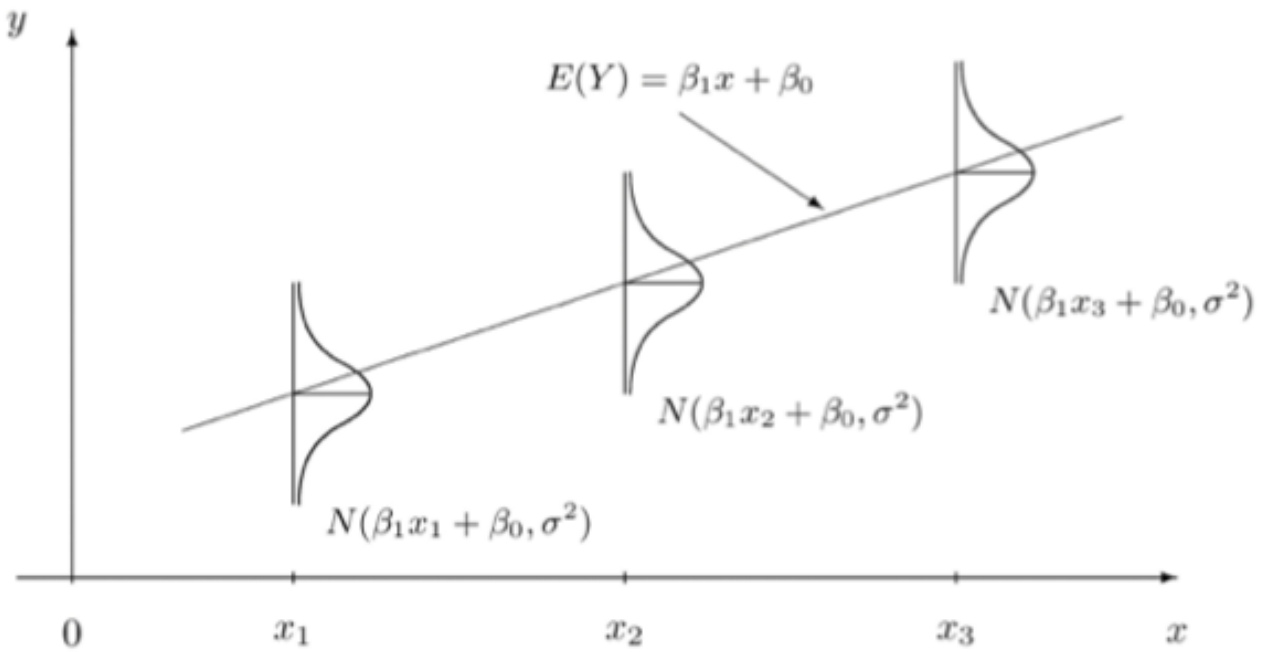
[3] We can write $\hat{\beta}_0 = \bar{y} - \sum_i c_i y_i \bar{x} = \sum_i (\frac{1}{n} - c_i \bar{x}) y_i$

Normal assumptions

Assume : $y_i = \beta_0 + \beta_1 x_i + e_i$

- The residuals e_i are assumed independent and $e_i \sim N(0, \sigma^2)$. This is equivalent to say that y_i's are independent and $y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$
- The mean function $E[y_i] = \beta_0 + \beta_1 x_i$ is a linear function of x_i
- Independence of e_i's implies Independence of y_i's
- Normality of e_i's implies Normality of y_i's
- Variance homogeneity of e_i's implies variance homogeneity of y_i's
- Since the e_i's are normal and independent, jointly they have a normal distribution. Therefore the y_i's are jointly normal, and any linear combination of the y_i's is also normal. This fact has important implications for coming inference results.

Simple Linear Regression Assumptions



Distributions of the LS estimates

- $(\hat{\beta}_0, \hat{\beta}_1)$ have a joint normal distribution with mean, variances and covariance given by:

$$E[\hat{\beta}_0] = \beta_0 \qquad Var[\hat{\beta}_0] = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}} \right)$$

$$E[\hat{\beta}_1] = \beta_1 \qquad Var[\hat{\beta}_1] = \sigma^2 \frac{1}{S_{XX}}$$

$$Cov(\hat{\beta}_0, \hat{\beta}_1) \qquad = -\sigma^2 \frac{\bar{x}}{S_{XX}}$$

- The residual sum of squares $RSS = \sum (y_i - \hat{y}_i)^2 \sim \sigma^2 \chi^2_{n-2}$.

$$E[\hat{\sigma}^2] = E[\frac{RSS}{n-2}] = \sigma^2 (n-2)/(n-2) = \sigma^2$$

- $(\hat{\beta}_0, \hat{\beta}_1)$ and RSS are independent.

Hypothesis testing

- We want to test: $H_0 : \beta_1 = c \;\; vs. \;\; H_a : \beta_1 \neq c$ (c is a known constant)
- Test statistic:

$$t = \frac{\hat{\beta}_1 - c}{se(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - c}{\hat{\sigma}/\sqrt{S_{XX}}} \sim T_{n-2}$$

- p-value$= 2\times$ the area under the T_{n2} distribution more extreme than the observed statistic t
- The p-value returned by the **R** command **lm** is for the test with $H_0 : \beta_1 = 0$

```
Call:
lm(formula = Hwt ~ Bwt, data = cats)

Residuals:
    Min      1Q  Median      3Q     Max
-3.5694 -0.9634 -0.0921  1.0426  5.1238

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.3567     0.6923  -0.515    0.607
Bwt           4.0341     0.2503  16.119   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.452 on 142 degrees of freedom
Multiple R-squared:  0.6466,        Adjusted R-squared:  0.644
F-statistic: 259.8 on 1 and 142 DF,  p-value: < 2.2e-16
```

Calculating the p-values:

```
> 2*(1-pt(abs(summary(out)$coefficients[1,3]),
+ out$df.residual))
[1] 0.6072131
> 2*(1-pt(abs(summary(out)$coefficients[2,3]),
+ out$df.residual))
[1] 0
```

Using $\alpha = 0.05$

- Reject the null hypothesis $H_0 : \beta_1 = 0$ (p-value $<< 0.05$)
- Fail to reject the null hypothesis: $H_0 : \beta_0 = 0$ (p-value $>> 0.05$)

F-test and ANOVA

An alternative way to test $\beta_1 = 0$ is based on the F-test. Recall the following decomposition of the variance: TSS = FSS + RSS.

Sum of Squares	Expression	df
TSS	$\sum_i (y_i - \bar{y})^2$	$n-1$
FSS	$\sum_i (\hat{y}_i - \bar{y})^2$	1
RSS	$\sum_i (y_i - \hat{y}_i)^2$	$n-2$

If $\beta_1 \neq 0$, we would expect a large amount of variation in Y is explained by the regression model, i.e., FSS is large. But how *large* is large? For the `cats` data, if we measure `Hwt` by kg, FSS will be much smaller, but whether `Bwt` is a good predictor for `Hwt` shouldn't be affected by the scale of `Hwt`.

Source	df	SS	MS	F
Regression	1	FSS	FSS$/1$	MS(reg)/MS(err)
Error	$n-2$	RSS	RSS$/(n-2)$	
Total	$n-1$	TSS		

Under $H_0 : \beta_1 = 0$, the F-test statistic (scale-invariant)

$$F = \frac{\text{MS(reg)}}{\text{MS(err)}} = \frac{\text{FSS}}{\text{RSS}/(n-2)} \sim F_{1,n-2}.$$

It can be shown that the F-test statistic is equal to the square of the t-test statistic (for testing $\beta_1 = 0$) and their p-values (for testing $\beta_1 = 0$) are the same. So they are essentially the same test; in other words, you can ignore the F-test in the R output for SLR.

ANOVA table for the *cats* example

Use function anova:

```
> anova(out)
Analysis of Variance Table

Response: Hwt
           Df Sum Sq Mean Sq F value    Pr(>F)
Bwt         1 548.09  548.09  259.83 < 2.2e-16 ***
Residuals 142 299.53    2.11
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
```

F-test is equivalent to the square of the t-test for SLR:

```
> summary(out)$fstatistic
   value    numdf    dendf
259.8348   1.0000 142.0000
> summary(out)$coefficients[2,3]^2
[1] 259.8348
```

Estimation/Prediction at a New Case

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[4]:

1. Estimation: We want to estimate the mean response at $x*$. This is equivalent to estimate: $\beta_0 + \beta_1 x^\star$
2. Prediction of an outcome of random variable $Y^\star$ at a given value $x^\star$, where

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

The fitted value (or point estimate) for estimation and prediction are the same: $\hat{\beta}_0 + \hat{\beta}_1 x^\star$. However the accuracy for estimation and the one for prediction are different.

[4]Estimation looks to get information from the data about a fixed but parameter, while prediction looks to get information about a random variable

Measure of accuracy: expected value of the squared difference between the point estimate and the target.

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

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

- 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, ie. is independent of $(\hat{\beta}_0, \hat{\beta}_1)$

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

- Error for estimation: $\sigma^2 \left(\frac{1}{n} + \frac{(x^\star - \bar{x})}{S_{XX}} \right)$
- Error for prediction: $\sigma^2 \left(1 + \frac{1}{n} + \frac{(x^\star - \bar{x})^2}{S_{XX}} \right)$
- Errors are not the same for all values of $x^\star$. It is smaller when $x^\star$ is close to $\bar{x}$.
- Prediction error $>$ estimation error.
- There are two sources of uncertainty when doing prediction at $x^\star$:
  1. the first is from the n sample points $(x_i, y_i) i = 1, \ldots, n$ which are used to estimate the LS line.
  2. the second is from the random error $e^\star \sim N(0, \sigma^2)$, which is the error we cannot avoid even if we knew (β_0, β_1). This is why, even when the sample size n goes to infinity, we can have the estimation error go to 0 but not the prediction error.

Confidence interval and prediction interval

- A confidence interval is always reported for a parameter, for example, $E[Y^\star|x^\star] = \beta_0 + \beta_1 x^\star$. The $(1-\alpha)100\%$ confidence interval for $\beta_0 + \beta_1 x^\star$ is:

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

- A prediction interval is reported for the value of a random variable, for example, $y^\star$. The $(1-\alpha)100\%$ prediction interval for $y^\star$ is:

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

Confidence and prediction interval of Hwt (in grs) for Bwt=2.15 kg

```
> predict(out, newdata=data.frame(Bwt = 2.15), se=TRUE,
+ interval="confidence")$fit
      fit      lwr     upr
1 8.316572 7.945395 8.68775
> predict(out, newdata=data.frame(Bwt = 2.15), se=TRUE,
+ interval="prediction")$fit
      fit      lwr      upr
1 8.316572 5.421611 11.21153
```

Association/Correlation vs. Causation

- The statement "X causes Y " means that changing the value of X will change the distribution of Y. When X causes Y , X and Y will be associated but the reverse is not, in general, true. Association does not necessarily imply causation.
- If the data are from a randomized study, then the causal interpretation is correct
- If the data are from a observational study, then the association interpretation is correct.