DSCI561: Regression I

Lecture 5: November 29, 2017

Gabriela Cohen Freue
Department of Statistics, UBC

Review from Lect 4

- Simple linear regression: $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$
- Minimize: $S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$
- Critical points of the objective function:

$$\frac{\partial S}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\frac{\partial S}{\partial \beta_1} = -2 \sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i) = 0$$

- Least squares estimator:
  - intercept: $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$
  - slope: $\hat{\beta}_1 = \frac{r_{xy} s_y}{s_x} = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) y_i}{(n-1) s_x^2}$

(r_{xy} is the correlation between x and y; S_x and S_y are their standard deviations)

In today's lecture

- Inference:
  - distribution of estimated parameters
  - confidence intervals
  - Prediction intervals
- Extend definitions and concepts to multiple regression models

Ordinary least squares (OLS) estimators

$$\hat{\beta}_1 = \frac{r_{xy} s_y}{s_x} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{(n-1)s_x^2}$$

Estimate, based on observed data

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})Y_i}{(n-1)s_x^2} = \sum_{i=1}^{n} a_i Y_i$$

Estimator: a random variable

Note that $\hat{\beta}_1$ is a linear combination of the random variables Y_i

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i; \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2) \quad \Longrightarrow \quad Y_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2)$$

Then,

$$\hat{\beta}_1 \sim \mathcal{N}(E(\hat{\beta}_1), Var(\hat{\beta}_1))$$

? ?

Unbiased estimator

$$E[\hat{\beta}_1] = \frac{\sum_{i=1}^{n}(x_i - \bar{x})E[Y_i]}{(n-1)s_x^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})[\beta_0 + \beta_1 x_i]}{(n-1)s_x^2} = 0 + \beta_1 \frac{\sum_{i=1}^{n}(x_i - \bar{x})[x_i]}{(n-1)s_x^2} = \beta_1$$

Unknown variance unless σ is known
σ

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

Then,

$$\hat{\beta}_1 \sim \mathcal{N}\left(\beta_1, \frac{\sigma^2}{(n-1)s_x^2}\right)$$

How do we test the null hypothesis $H_0 : \beta_1 = 0$?

$$z = \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - \beta_1}{\frac{\sigma}{\sqrt{(n-1)s_x}}} \sim \mathcal{N}(0,1)$$

But σ is usually unknown!!

$$\boxed{t = \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - \beta_1}{\frac{\hat{\sigma}}{\sqrt{(n-1)s_x}}} \sim t_{n-2}}$$

where,

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}$$

Fitted values: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

standard deviation of the residuals: $r_i = y_i - \hat{y}_i$

Makes it an unbiased estimator of σ^2 : $E(\hat{\sigma}^2) = \sigma^2$

Similarly,

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x} \implies E(\hat{\beta}_0) = \beta_0 + \beta_1 \bar{x} - E(\hat{\beta}_1)\bar{x}) = \beta_0$$

$$V(\hat{\beta}_0) = V\left(\sum_{i=1}^{n} \frac{1}{n} Y_i - a_i Y_i \bar{x}\right) = V\left(\sum_{i=1}^{n} c_i Y_i\right) = \sum_{i=1}^{n} c_i^2 \sigma^2 = \left(\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2}\right)\sigma^2$$

$$\boxed{t = \frac{\hat{\beta}_0 - \beta_0}{SE(\hat{\beta}_0)} = \frac{\hat{\beta}_0 - \beta_0}{\hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2}}} \sim t_{n-2}}$$

Confidence Interval of the slope

$$\hat{\beta}_1 \pm t_{n-2,0.975} \times SE(\hat{\beta}_1)$$

where,

$$SE(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{n-1}s_x} \qquad \text{(1x1) estimate}$$

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n} \hat{e}_i^2}{n-2}} \qquad \text{Residual SD}$$

Confidence Interval of the intercept

$$\hat{\beta}_0 \pm t_{n-2,0.975} \times SE(\hat{\beta}_0)$$

where,

$$SE(\hat{\beta}_0) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2}}$$ (1x1) estimate

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n} \hat{e}_i^2}{n-2}}$$ Residual SD

```r
lm_BLDG<-lm(assessment_k~BLDG_METRE,data=dat.2)
tidy(lm_BLDG)
```

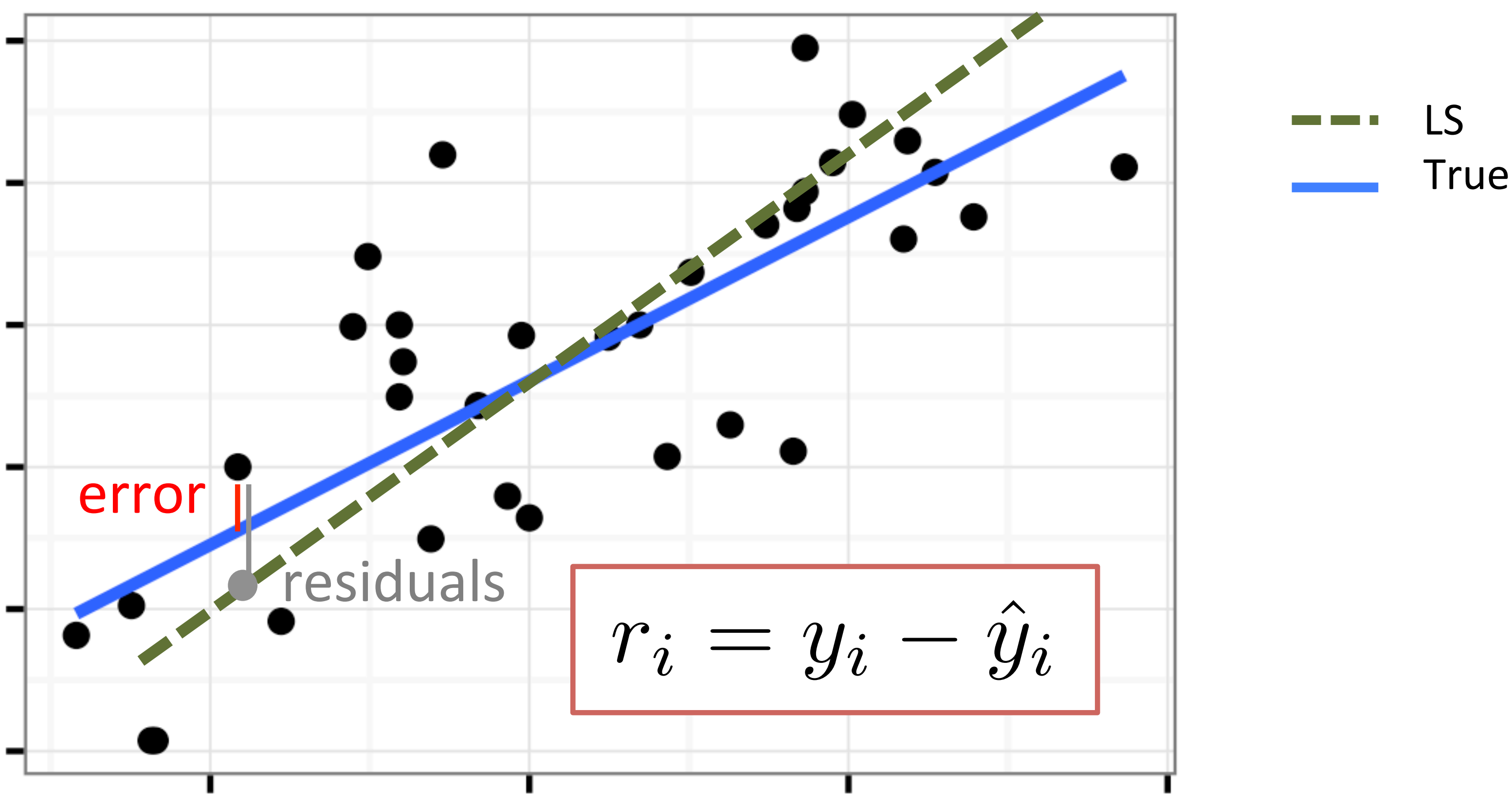
```
##          term  estimate  std.error statistic      p.value
## 1 (Intercept) 22.046047 19.4790234  1.131784 2.584662e-01
## 2  BLDG_METRE  3.079313  0.1212517 25.396038 9.282329e-83
```

$$\boxed{\begin{array}{c}\hat{\beta}_0 \\ \hat{\beta}_1\end{array}} \quad \boxed{\begin{array}{c}SE(\hat{\beta}_0) \\ SE(\hat{\beta}_1)\end{array}} \quad \boxed{\begin{array}{c}t_{\hat{\beta}_0} \\ t_{\hat{\beta}_1}\end{array}}$$

```
## Residual standard error: 137.7 on 366 degrees of freedom
```

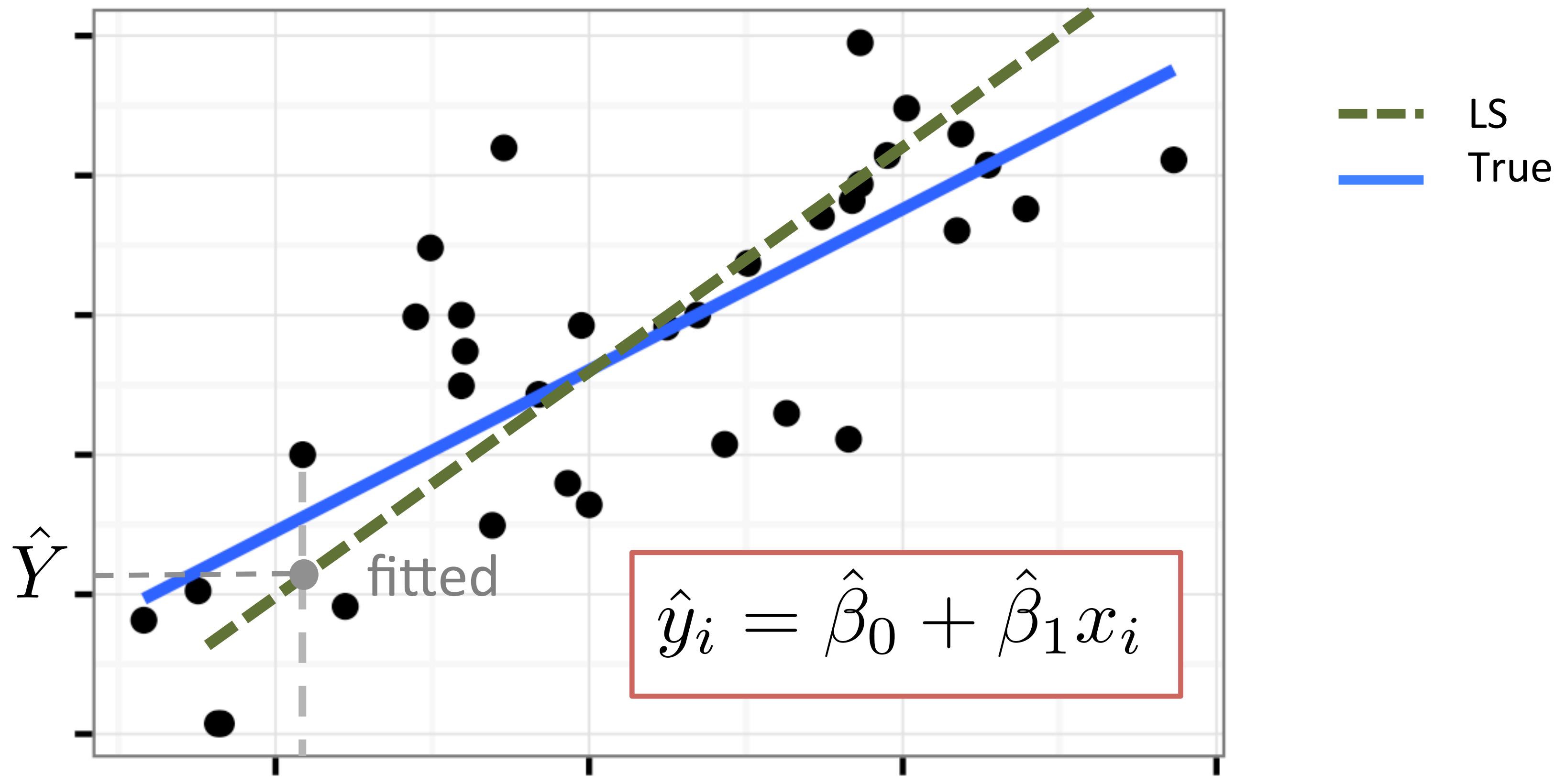
$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}} \qquad DF = n - 2 = 366$$

Residuals



The residual is the (vertical) distance between the ***estimated*** line and the real observation:

Fitted values



The fitted values are the predictions of the responses by the ***estimated*** line (points over the line)

```r
vals_n_errors <- augment(lm_BLDG) %>%
  select(assessment_k, BLDG_METRE, .fitted, .resid)
head(vals_n_errors)
```

```
##   assessment_k BLDG_METRE  .fitted      .resid
## 1          354         97 320.7394   33.26057
## 2          449        166 533.2120  -84.21204
## 3          383        108 354.6119   28.38813
## 4          536        217 690.2570 -154.25702
## 5          595        145 468.5465  126.45354
## 6          449        171 548.6086  -99.60861
```

```
## Residual standard error: 137.7 on 366 degrees of freedom
```

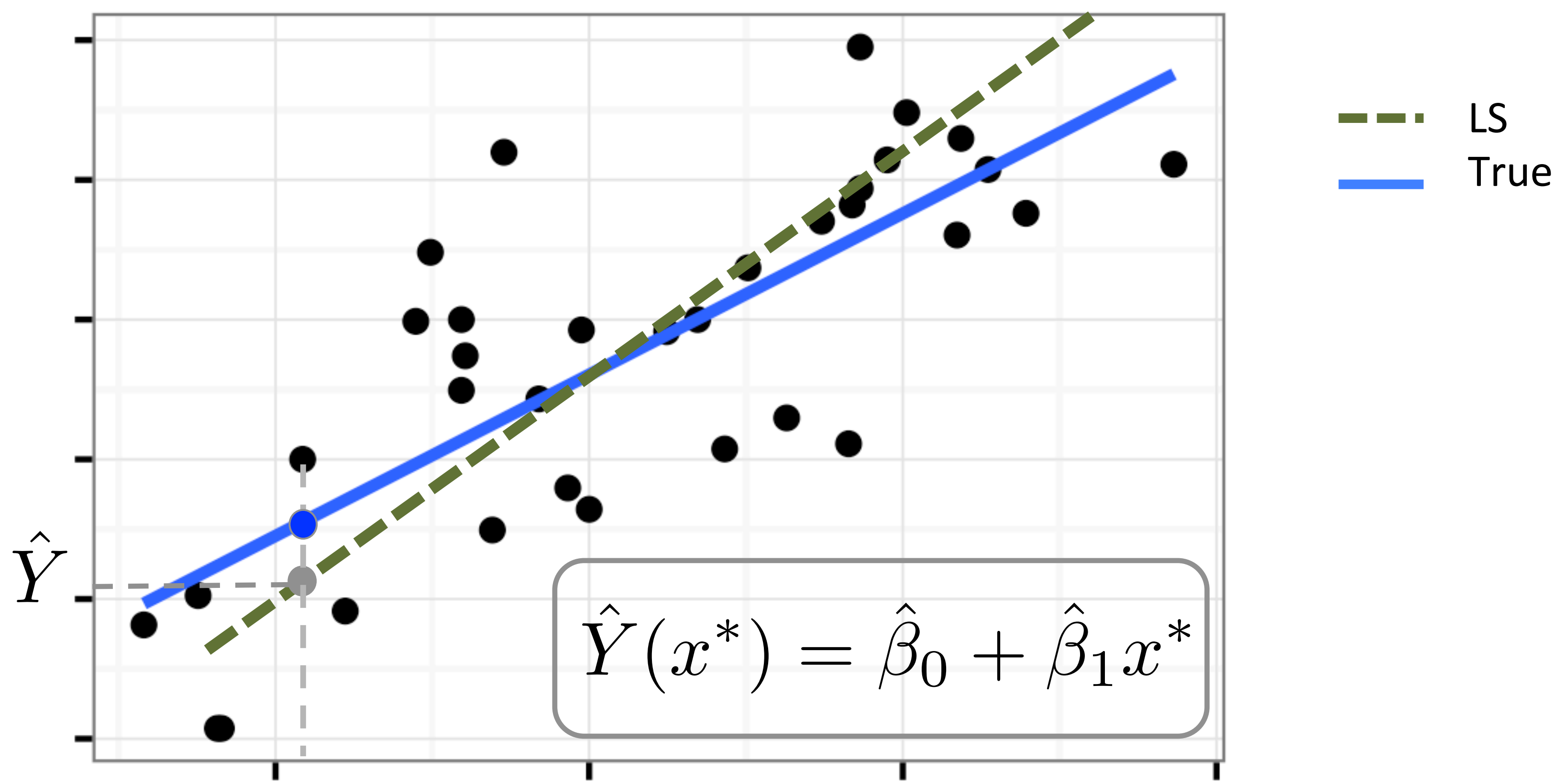
```r
sd(vals_n_errors$.resid)*sqrt((nobs-1)/(nobs-2))
```

```
## [1] 137.677
```

```r
with(dat.2,sqrt(sum((assessment_k-vals_n_errors$.fitted)^2)/(nobs-2)))
```

```
## [1] 137.677
```

Predictions



The prediction is the y-value of a point on the *estimated* line
NOTE: the grey point estimates *new* black dots *and* the blue dot

Prediction Intervals

- The grey point (fitted value, $\hat{Y}$) is used to predict new black point

- The variance of the prediction depends on the uncertainty of the estimated coefficients and that of the error that generates the data

$$\hat{Y}(x^*) \pm t_{n-2,0.975} \times SE(\hat{Y}(x^*))$$

where,

$$SE(\hat{Y}(x^*)) = \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{(n-1)s_x^2}}$$

(1x1) estimate

Confidence Interval of the prediction

- The grey point (fitted value, $\hat{Y}$) is used to estimate the blue point (i.e., the conditional expectation of Y given x*)

- The variance of the estimation depends *only* the uncertainty of the estimated coefficients

$$\hat{Y}(x^*) \pm t_{n-2,0.975} \times SE_{\hat{\mu}_{Y|x^*}}$$

$$SE_{\hat{\mu}_{Y|x^*}} = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{(n-1)s_x^2}}$$

(1x1) estimate

Note: these are not the CI of the regression coefficients

Prediction Intervals

```r
head(predict.lm(lm_BLDG, interval = "prediction"))
```

```
## Warning in predict.lm(lm_BLDG, interval = "prediction"): predictions on current data refer to _futur
```

```
##        fit       lwr      upr
## 1 320.7394  49.34739 592.1315
## 2 533.2120 262.07810 804.3460
## 3 354.6119  83.32781 625.8959
## 4 690.2570 418.67268 961.8414
## 5 468.5465 197.43962 739.6533
## 6 548.6086 277.45459 819.7626
```

Confidence Interval of the prediction

```r
predicted_fits <- data.frame(predict.lm(lm_BLDG, interval = "confidence", se.fit = TRUE)$fit)
head(predicted_fits)
```

```
##        fit      lwr      upr
## 1 320.7394 301.8987 339.5802
## 2 533.2120 518.5510 547.8731
## 3 354.6119 337.3962 371.8275
## 4 690.2570 668.8238 711.6903
## 5 468.5465 454.3953 482.6976
## 6 548.6086 533.5808 563.6364
```