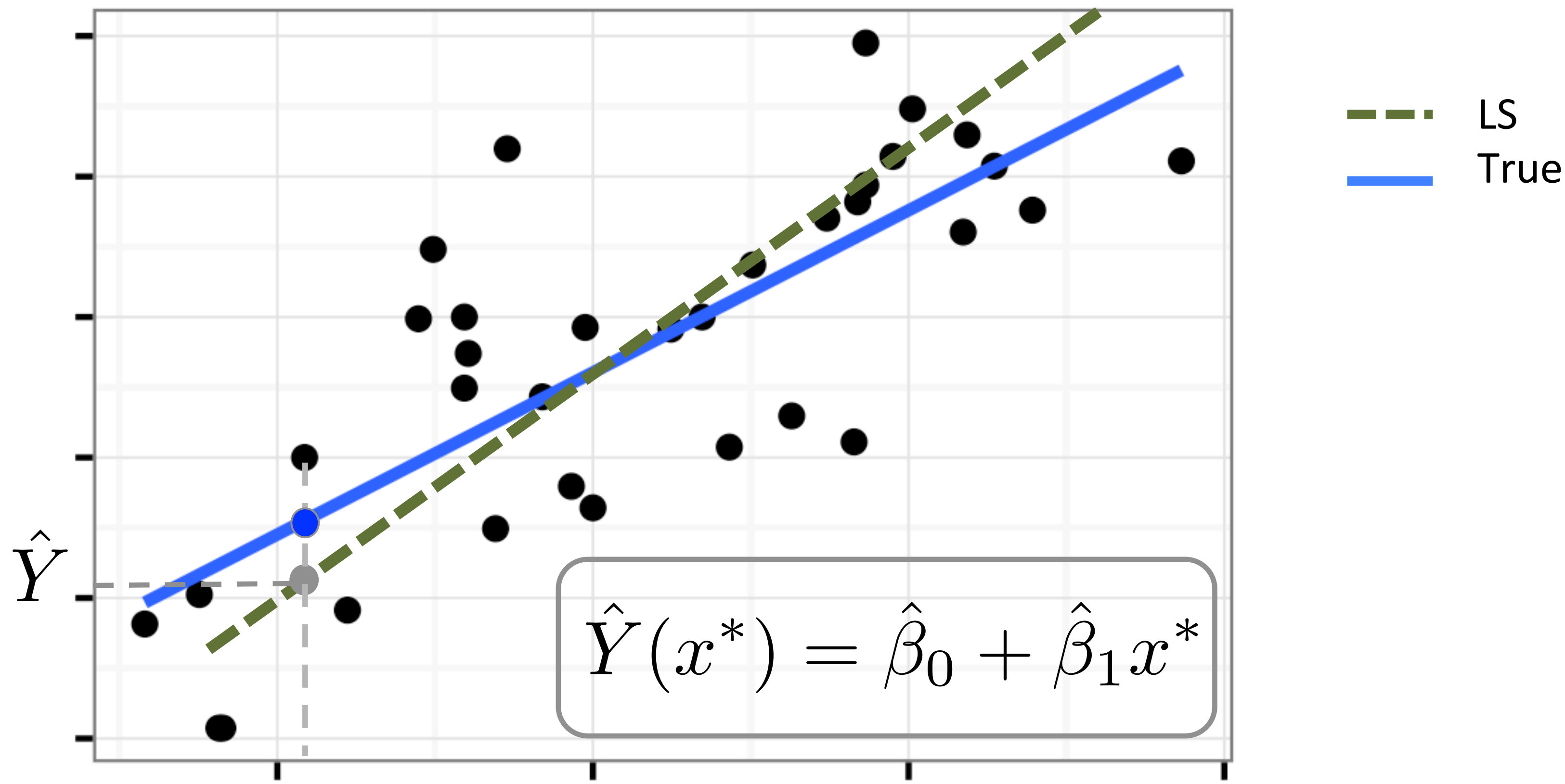
DSCI561: Regression I

Lecture 6: December 4, 2017

Gabriela Cohen Freue
Department of Statistics, UBC

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

Review from Lect 5

- **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

$$\hat{Y}(x^*) \pm t_{n-2,0.975} \times SE(\hat{Y}(x^*)) \quad SE(\hat{Y}(x^*)) = \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{(n-1)s_x^2}}$$

- **Confidence Interval for 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

$$\hat{Y}(x^*) \pm t_{n-2,0.975} \times SE_{\hat{\mu}_{Y|x^*}};\ SE_{\hat{\mu}_{Y|x^*}} = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{(n-1)s_x^2}}$$

Prediction Intervals

```
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

```
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
```

In today's lecture

- Inference: using bootstrapping
- Extend definitions and concepts to multiple regression models

Statistical Inference

Sample

Population

$$\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1)$$

$$\boldsymbol{\beta} = (\beta_0, \beta_1)$$

$$\boxed{\hat{\beta}_i \sim ?}$$

$$\beta_i \sim \mathcal{G}$$

What assumptions are you willing to make?

Ordinary least squares (OLS) estimators

Last class:

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

Then,

$$\implies Y_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2) \text{ and } \hat{\beta}_1 \sim \mathcal{N}\left(\beta_1, \frac{\sigma^2}{(n-1)s_x^2}\right)$$

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

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

Bootstrapping

- But…, if the assumptions are not good??

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

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

- Which distribution can we use to assess the significance of t??

Bootstrapping

Sample

Population

$$\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1)$$

$$\boldsymbol{\beta} = (\beta_0, \beta_1)$$

$$\beta_i \sim \mathcal{G}$$

$$\left.\begin{aligned}
\hat{\boldsymbol{\beta}}^1 &= (\hat{\beta}_0, \hat{\beta}_1)^1 \\
\hat{\boldsymbol{\beta}}^2 &= (\hat{\beta}_0, \hat{\beta}_1)^2 \\
&\cdots \\
\hat{\boldsymbol{\beta}}^B &= (\hat{\beta}_0, \hat{\beta}_1)^B
\end{aligned}\right\} \quad \hat{\beta}_i \sim ?$$

We can get the sampling distribution of the estimators

Bootstrapping (cont.)

- Draw a new sample of size *n* from the observed data, with replacement.
- With replacement: some observations will re-appear (some once, some twice, etc) and some observations may not appear at all in a given bootstrap sample.
- Compute the test-statistic from the bootstrap sample. That is the bootstrap statistic.
- We do that B times (B should be *large*). We look at the distribution of our bootstrap statistics.

Bootstrapping in Regression

- X random: we sample rows of the dataframe

$$z_i = (y_i, x_{1i}, \ldots, x_{pi})$$

- X fixed: $y_i = \hat{y}_i + \hat{e}_i$
  - we sample from the residuals to generate bootstrap samples, X is fixed!

Bootstrapping

Sample	Population

$$\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1)$$

$$\boldsymbol{\beta} = (\beta_0, \beta_1)$$

$$\beta_i \sim \mathcal{G}$$

$$\left.\begin{aligned}
\hat{\boldsymbol{\beta}}^1 &= (\hat{\beta}_0, \hat{\beta}_1)^1 \\
\hat{\boldsymbol{\beta}}^2 &= (\hat{\beta}_0, \hat{\beta}_1)^2 \\
&\cdots \\
\hat{\boldsymbol{\beta}}^B &= (\hat{\beta}_0, \hat{\beta}_1)^B
\end{aligned}\right\} \frac{1}{B}\sum_{b=1}^{B} \hat{\beta}_i^b$$

The population is to the sample
as
The sample is to the bootstrap sample

Continues in Lecture 7 and lectures_code.pdf