Multiple Linear Regression (Part IV)

Lecture 7

Alexandra Chronopoulou

ILLINOIS

COLLEGE OF LIBERAL ARTS & SCIENCES

Department of Statistics
101 Illini Hall, MC-374
725 S. Wright St.
Champaign, IL 61820-5710

Learning objectives

In this lecture we will:

- construct confidence and prediction intervals for the LS coefficients.
- construct confidence and prediction intervals for μ^*.
- introduce multiple/simultaneous confidence intervals.

Confidence Intervals for β

Confidence Intervals for the β_j's

- Recall that the distribution of the LS estimators $\hat{\beta}$ is

$$\hat{\beta} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{Y} \sim \mathcal{N}_p\left(\beta, \sigma^2(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\right)$$

- An $(1-\alpha)100\%$ CI for β_j can be written as

$$\left(\hat{\beta}_j \pm T_{n-p}(\alpha/2)\, se(\hat{\beta}_j)\right) = \left(\hat{\beta}_j \pm T_{n-p}(\alpha/2)\, \hat{\sigma}\sqrt{[(\mathbf{X}^T\mathbf{X})^{-1}]_{jj}}\right)$$

where $T_{n-p}(\alpha/2)$ is the $(1-\alpha/2)$ percentile of the student T distribution with $(n-p)$ degrees of freedom.

Confidence Intervals for the β_j's in R

In R we can use the function confint(.)

```
bikeshare.mlr = lm(cnt ~ t1 +  hum + wind_speed, data=bikeshares.reg )
confint(bikeshare.mlr)
```

```
##                   2.5 %      97.5 %
## (Intercept) 2543.679114 2766.669772
## t1            41.516598   47.099480
## hum          -28.984942  -26.739780
## wind_speed    -4.941603   -1.262794
```

```
confint(bikeshare.mlr, 't1', level=0.99)
```

```
##       0.5 %   99.5 %
## t1 40.63932 47.97676
```

Confidence Region

- Just as we can use estimated standard errors and test statistics to form confidence intervals for a *single parameter*, we can also obtain a $(1-\alpha) \times 100\%$ confidence region for the *entire vector* β.
- In particular:

$$\beta - \hat{\beta} \sim \mathcal{N}\left(\mathbf{0}, \sigma^2 (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\right)$$

- Thus, the quadratic form:

$$\frac{(\beta - \hat{\beta})^T \, \mathbf{X}^\mathsf{T}\mathbf{X} \, (\beta - \hat{\beta})}{p \, \hat{\sigma}^2} \sim F_{p, n-p}$$

Confidence Region for β

- We can construct a $(1-\alpha) \times 100\%$ confidence region for β to be all the points in the following ellipsoid

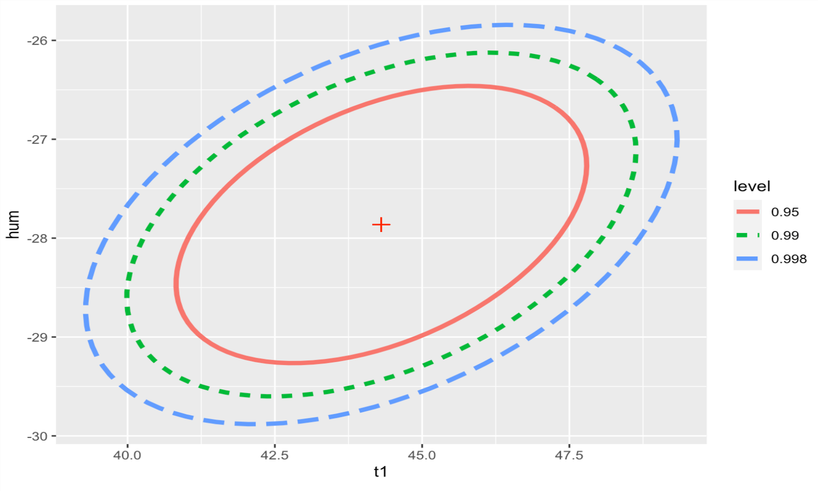
$$\frac{(\beta - \hat{\beta})^T\, \mathbf{X}^\mathsf{T}\mathbf{X}\, (\beta - \hat{\beta})}{p\, \hat{\sigma}^2} < F(\alpha; p, n-p)$$

where $F(\alpha; p, n-p)$ is defined to be the point such that:

$$\mathbb{P}\left(F_{p,n-p} > F(\alpha; p, n-p)\right) = \alpha$$

ILLINOIS

Confidence Region in Bike Shares Example



Confidence/Prediction Intervals for New Observations

- Consider $\mathbf{x}^*$ a future observation.

Goal

Similar to the simple linear regression case, we want to obtain

- an estimate
$$\mathbb{E}(Y|\mathbf{x}^*) = \mu^* = (\mathbf{x}^*)^T \beta$$
- a prediction for a future observation Y^* at $\mathbf{x}^*$.
- a confidence interval for μ^*.
- a prediction interval for y^*.

Confidence Interval for μ^*

- The Gauss-Markov theorem tells us that the BLUE (Best Linear Unbiased Estimate) of μ^* is:

$$\hat{\mu}^* = (\mathbf{x}^*)^T \hat{\beta} = (\mathbf{x}^*)^T (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y}$$

- This is just a linear transformation of $\mathbf{y}$, so we can easily derive its variance, and find its standard error.

- It can be shown that:

$$se(\hat{\mu}^*) = \hat{\sigma}\sqrt{(\mathbf{x}^*)^\mathsf{T}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{x}^*}$$

- A Confidence Interval for μ^* is given by:

$$\left(\hat{\mu}^* - T_{n-p}(\alpha/2)\, se(\hat{\mu}^*),\ \hat{\mu}^* + T_{n-p}(\alpha/2) se(\hat{\mu}^*)\right)$$

Prediction Interval for μ^*

- The best estimate for y^* at a future observation $\mathbf{x}^*$ is also

$$\hat{y}^* = (\mathbf{x}^*)^T \hat{\beta}$$

- In order to find a prediction interval (PI), we need to consider the variance due to $\hat{\beta}$ in addition to the variance associated with a new observation, which is σ^2.

- The standard error of a prediction estimate $\hat{y}^*$ is:[1]

$$se(\hat{y}^*) = \hat{\sigma}\sqrt{1 + (\mathbf{x}^*)^\mathbf{T}(\mathbf{X}^\mathbf{T}\mathbf{X})^{-1}\mathbf{x}^*}$$

- A $(1-\alpha)100\%$ PI for a new observation Y^* at $\mathbf{x}^*$ is given by:

$$(\hat{y}^* - T_{n-p}(\alpha/2)\, se(\hat{y}^*), \hat{y}^* + T_{n-p}(\alpha/2) se(\hat{y}^*))$$

[1] Note that no matter how large the sample size becomes, the width of a PI, unlike a CI, will never approach 0.

Confidence & Prediction Intervals in R

```
# create a data frame on which you would like to predict
meanvalue=apply(bikeshares.reg[,2:5],2,mean)
meanvalue
```

```
##          t1          t2         hum wind_speed
##    12.46809    11.52084    72.32495   15.91306
```

```
x=data.frame(t(meanvalue))
predict.lm(bikeshare.mlr,x,interval="confidence")
```

```
##        fit      lwr      upr
## 1 1143.102 1129.198 1157.006
```

```
predict.lm(bikeshare.mlr,x,interval="prediction")
```

```
##        fit       lwr      upr
## 1 1143.102 -691.7461 2977.949
```

ILLINOIS

Standard Errors as a function of the Mahalanobis distance

To quantify the distance between an observation vector in $\mathbb{R}^p$ and its sample mean $\bar{\mathbf{x}}$ we can use the Mahalanobis distance.

- Write $\mathbf{x}_{p\times 1} = \begin{pmatrix} 1 \\ \mathbf{z} \end{pmatrix}$ where $\mathbf{z}$ denotes the values of the $(p-1)$ predictors (without the intercept).
- We can write the sample covariance matrix of the $(p-1)$ predictor variables as:

$$\hat{\Sigma}_{(p-1)\times(p-1)} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{z}_i - \bar{\mathbf{z}})(\mathbf{z}_i - \bar{\mathbf{z}})^T$$

- The following expression can be written as:

$$(\mathbf{x}^*)^{\mathbf{T}}(\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\mathbf{x}^* = \frac{1}{n} + \frac{1}{n-1}(\mathbf{z}^* - \bar{\mathbf{z}})^T\hat{\Sigma}^{-1}(\mathbf{z}^* - \bar{\mathbf{z}})$$

The second term in the right hand side is the so-called *Mahalanobis distance from* $\mathbf{z}^*$ *to the center of the data* $\bar{\mathbf{z}}$ (the sample mean).

Confidence vs. Prediction Standard Errors

- The point estimation and prediction at a given $\mathbf{x}^*$ are the same, but their standard errors are different:

$$\begin{aligned} se(\hat{\mu}^*) &= \hat{\sigma}\sqrt{\mathbf{x}^{*\mathsf{T}}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{x}^*} \\ &= \hat{\sigma}\sqrt{\frac{1}{n} + \frac{1}{n-1}(\mathbf{z}^* - \bar{\mathbf{z}})^T\hat{\Sigma}^{-1}(\mathbf{z}^* - \bar{\mathbf{z}})} \end{aligned}$$

$$\begin{aligned} se(\hat{y}^*) &= \hat{\sigma}\sqrt{\mathbf{1} + (\mathbf{x}^*)^\mathsf{T}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{x}^*} \\ &= \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{1}{n-1}(\mathbf{z}^* - \bar{\mathbf{z}})^T\hat{\Sigma}^{-1}(\mathbf{z}^* - \bar{\mathbf{z}})} \end{aligned}$$

- Since $se(\hat{y}^*)$ has an extra 1, when the sample size n goes to infinity,
  - $se(\hat{\mu}^*) \to 0$
  - $se(\hat{y}^*) \to \sigma$

- Consider a Simple Linear Regression

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

- Given the values of x^*, the $(1-\alpha)100\%$ Confidence Interval for $\mu^* = \mathbb{E}[y|x^*] = \beta_0 + \beta_1 x^*$ is:

$$I(x^*) = (\hat{\mu}^* \pm T_{n-2}(\alpha/2)\, se(\hat{\mu}^*)) \tag{1}$$

where

$$\hat{\mu}^* = \hat{\beta}_0 + \hat{\beta}_1 x^* \text{ and } se(\hat{\mu}^*) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

- If we want confidence intervals at multiple points $(x_1^*, x_2^*, \ldots, x_m^*)$, we can use formula (1) to have confidence intervals at the m points: $I(x_1^*), I(x_2^*), \ldots, I_m(x^*)$.

Bonferroni Correction

- We know that:

$$\mathbb{P}\left(\mu_i^* \in I(x_i^*)\right) = (1 - \alpha)$$

This is the point-wise coverage probability for μ_i^* and formula (1) gives the point-wise CI.

- What about the simultaneous coverage probability? i.e.:

$$\mathbb{P}\left(\mu_i^* \in I(x_i^*), \text{ for } \ i = 1, \ldots, m\right) = ?$$

- To make sure that (for example):

$$\mathbb{P}\left(\mu_i^* \in I(x_i^*), \text{ for } \ i = 1, \ldots, m\right) = .95$$

we need to set $\alpha = 5\%/m$, which is known as the Bonferroni correction

- Let A_k denotes the event that the kth confidence interval covers μ_k^* with:

$$\mathbb{P}(\mathrm{A}_k) = (1 - \alpha)$$

- Then we can show:

$$\begin{aligned}
&\mathbb{P}(\text{All CIs cover the corresponding } \mu_k^* \text{ values}) \\
&= \mathbb{P}\left(\mathrm{A}_1 \cap \mathrm{A}_2 \ldots \cap \mathrm{A}_m\right) \\
&= 1 - \mathbb{P}(\mathrm{A}_1^c \cup \mathrm{A}_2^c \ldots \cup \mathrm{A}_m^c) \\
&\geq 1 - \mathbb{P}(\mathrm{A}_1^c) - \ldots - \mathbb{P}(\mathrm{A}_m^c) \\
&= 1 - m\alpha
\end{aligned}$$

- If we choose α/m instead of α, the simultaneous coverage probability will be $(1 - \alpha)$.

Confidence Band for Regression Line

Confidence Band

- Ideally we would like to construct a *simultaneous confidence band* (i.e., $m = \infty$) across all x^*'s. (Scheffé's Theorem - 2959). Let

$$I(x) = (\hat{r}(x) - c\hat{\sigma},\, \hat{r}(x) + c\hat{\sigma})$$

  where

$$\hat{r}(x) = \hat{\beta}_0 + \hat{\beta}_1 x,\; c\,\hat{\sigma} = \sqrt{2\,F(\alpha, 2, n-2)}\,\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

- Then,

$$\mathbb{P}\left(r(x) \in I(x) \text{ for all } x\right) \geq 1 - \alpha$$

- Can we construct a simultaneous prediction band? No!

Confidence Band vs. Pointwise Confidence Intervals

- Are confidence bands always wider than point-wise confidence intervals?

- For SLR, at a location x^*, we have

$$\begin{aligned} \text{band} \ &: \hat{\mu}^* \pm \sqrt{2F(\alpha, 2, n-2)}\, se(\hat{\mu}^*) \\ \text{interval} \ &: \hat{\mu}^* \pm T_{n-2}(\alpha/2)\, se(\hat{\mu}^*) \end{aligned}$$

- Assume $\alpha = 5\%$, you ca check which is bigger:

$$\sqrt{2F(\alpha, 2, n-2)} \text{ or } T_{n-2}(\alpha/2) = \sqrt{2F(\alpha, 1, n-2)}$$

- In fact, for any α, we have

$$T_m(\alpha/2)\sqrt{F(\alpha, 1, m)} < \sqrt{k\, F(\alpha, k, m)}$$