Confidence intervals for OLS coefficients

Stats 203, Winter 2024

Lecture 5 [Last update: February 10, 2024]

1 Review and preview

- Last class we covered the meaning (and not meaning) of parameters in the linear model, and derived the bias, covariance, and standard error of OLS estimators when the model is correct.
- The "rule of thumb" we have discussed in class to this point says that "β_j is roughly equal to $\hat{\beta}_j$, give or take a standard error." Today we will replace this rule of thumb with something more precise: **confidence intervals**. At the end of the Lecture 4 notes, there is a little demonstration of how standard errors and the rule of thumb work on some simulated data. We didn't get to that last time, so we'll start today's class with it.
- Mathematically, we could write the range of values given by the rule of thumb as

$$\left[\hat{\beta}_j - \sigma\sqrt{[(\mathbf{X}^\top\mathbf{X})^{-1}]_{(j+1),(j+1)}}, \hat{\beta}_j + \sigma\sqrt{[(\mathbf{X}^\top\mathbf{X})^{-1}]_{(j+1),(j+1)}}\right] \tag{1}$$

  There are two issues with this (aside from the formula being ugly). The first is that σ itself is an unknown parameter, and so I cannot actually compute (1) from the data. Second, why did I go plus and minus one SE, as opposed to two SEs, or .5 SEs, etc? In order to compute confidence intervals, we will need to address both issues.
- Today's class has a lot of material. At the end of the notes, there is a bit of data analysis showing how to compute confidence intervals in R, which I suspect we may not get to in class. If we don't, we will start with that next Monday.

2 Estimating the error variance, and standard error

- To estimate the error variance we will use start by using **plug-in method**. The plug-in method was mentioned at the bottom of Lecture 1 notes, although we didn't cover it in class. The plug-in method is a little bit less formally prescribed than the method of least squares: basically, we try to estimate a parameter by writing it in terms of simple quantities (e.g means, variances, and covariances) that we *do* know how to estimate, and then plugging in these estimates.
- **Plug-in estimate for error variance.** To apply the plug-in method, notice that if (X, Y) follow the linear model then

$$\sigma^2 = \mathbb{E}[(Y - \beta_0 - \sum_{j=1}^{p} \beta_j X_j)^2].$$

  Since we do not know the expectation on the right hand side, we replace it by a sample mean, leading to

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2.$$

Of course this is still not an estimate as it depends on the unknown β. So we apply the plug-in method again, this time plugging in the OLS estimates $\hat{\beta}$ for β:

$$\hat{\sigma}^2_{\text{PLUG}} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \sum_{j=1}^{p}\hat{\beta}_j x_{ij})^2.$$

Now this is an estimate! If we plug in random variables $(X_1, Y_1), \ldots, (X_n, Y_n)$ following the linear model (say), we get back the estimator

$$\hat{\sigma}^2_{\text{PLUG}} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \sum_{j=1}^{p}\hat{\beta}_j X_{ij})^2.$$

As usual, the notation is the same whether we are talking about estimates (functions of data) or estimators (functions of random variables) but the meaning is completely different.

As an alternative derivation, notice that $\hat{\sigma}^2_{\text{PLUG}}$, can be written in terms of the OLS residuals, which we talked about in Lecture 3. Recall that fitted values and residuals are given by,

$$\hat{y} = \mathbf{X}\hat{\beta}, \quad e = y - \hat{y},$$

respectively. So $\hat{\sigma}^2_{\text{PLUG}} = \frac{1}{n}\sum_{i=1}^{n} e_i^2 = \frac{1}{n}\|e\|^2$ is just the mean of the squared residuals, or equivalently, the squared norm of the residual vector, divided by n.

- As with any other estimator, $\hat{\sigma}^2_{\text{PLUG}}$ itself has an expectation and a variance. Computing its expectation is (in my opinion) the first truly complex derivation of this class, in that if I sat you down for an hour and asked you to figure it out, I wouldn't be surprised if you failed. The reason is that the derivation relies on several geometric properties of OLS that are not obvious until after I tell you about them (or unless you're C.F. Gauss.)

- **The hat matrix, and a little more geometry.** A bit of algebra shows that the fitted values and residuals can be written as

$$\hat{y} = \mathbf{H}y, \quad e = (\mathbf{I}_n - \mathbf{H})y, \quad \text{where} \quad \mathbf{H} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top.$$

This matrix $\mathbf{H}$ is important enough that it gets a name: the **hat matrix**. The hat matrix turns out to satisfy a number of nice algebraic properties. In increasingly interesting order:

  - It is symmetric $\mathbf{H}_{ij} = \mathbf{H}_{ji}$.
  - It is positive semi-definite: $v^\top\mathbf{H}v \geq 0$ for all $v \in \mathbb{R}^n$.
  - It is **idempotent**:

$$\mathbf{H}^2 = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top = \mathbf{H}.$$

  - It leaves anything in the span of the columns of $\mathbf{X}$ unchanged. Recall that $v \in \mathbb{R}^n$ is in the span of the columns of $\mathbf{X}$ if $v = \mathbf{X}b$ for some $b \in \mathbb{R}^{p+1}$. But then,

$$\mathbf{H}v = \mathbf{H}\mathbf{X}b = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}b = \mathbf{X}b = v.$$

  - Assuming $\mathbf{X}$ is full rank, $\mathbf{H}$ has trace equal to $p+1$. Remember that the **trace** of a matrix is the sum of its diagonal elements: $\text{tr}(\mathbf{A}) = \sum_{i=1}^{n}\mathbf{A}_{ii}$. Then

$$\begin{aligned}\text{tr}(\mathbf{H}) &= \text{tr}(\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top) \\ &= \text{tr}(\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}) \\ &= \text{tr}(\mathbf{I}_{p+1}) \\ &= p+1,\end{aligned}$$

where the second equality uses the "trace trick" $\text{tr}(\mathbf{AB}) = \text{tr}(\mathbf{BA})$.

These set of linear algebraic properties have a corresponding geometric interpretation: the hat matrix is a projection. Specifically, it is the linear transformation which projects a vector $y \in \mathbb{R}^n$ onto the subspace spanned by the columns of $\mathbf{X}$. The trace represents the dimension of the subspace it projects onto, which if $\mathbf{X}$ is full rank is $p + 1$.

Notice that $\mathbf{I} - \mathbf{H}$ is also symmetric, positive semi-definite, idempotent.

$$(\mathbf{I} - \mathbf{H})^2 = \mathbf{I} + \mathbf{H}^2 - 2\mathbf{I}\mathbf{H} = \mathbf{I} - \mathbf{H}.$$

The corresponding geometric interpretation: $\mathbf{I} - \mathbf{H}$ is also a projection matrix, which projects a vector $y \in \mathbb{R}^n$ into the subspace *orthogonal to* the span of the columns of $\mathbf{X}$.

To see the use of this kind of geometry, consider the residual vector $e = Y - \mathbf{X}\hat{\beta} = (\mathbf{I} - \mathbf{H})Y$.[1] Now e is a random vector, and its (conditional-on-$X_1, \ldots, X_n$) covariance can be simply computed:

$$\begin{aligned}
\text{Cov}[e|X_1, \ldots, X_n] &= \text{Cov}[(\mathbf{I} - \mathbf{H})Y|X_1, \ldots, X_n] \\
&= (\mathbf{I} - \mathbf{H})\text{Cov}[Y|X_1, \ldots, X_n](\mathbf{I} - \mathbf{H})^\top \\
&= \sigma^2(\mathbf{I} - \mathbf{H})(\mathbf{I} - \mathbf{H})^\top \\
&= \sigma^2(\mathbf{I} - \mathbf{H}).
\end{aligned}$$

Notice that, unlike the errors, the residuals are *not* independent.

- **Bias of plug-in estimate of error variance.** We start by using the idempotency of $\mathbf{I} - \mathbf{H}$,

$$\begin{aligned}
\frac{1}{n}\|e\|^2 &= \frac{1}{n}\|(\mathbf{I}_n - \mathbf{H})Y\|^2 \\
&= \frac{1}{n}Y^\top(\mathbf{I}_n - \mathbf{H})^\top(\mathbf{I} - \mathbf{H})Y \\
&= \frac{1}{n}Y^\top(\mathbf{I}_n - \mathbf{H})Y \\
&= \frac{1}{n}(\mathbf{X}\beta + \epsilon)^\top(\mathbf{I}_n - \mathbf{H})(\mathbf{X}\beta + \epsilon) \\
&= \frac{1}{n}\epsilon^\top(\mathbf{I}_n - \mathbf{H})\epsilon,
\end{aligned}$$

where the first equality uses the definition of residuals e, the second equality is just the definition of squared-norm, the third equality uses the idempotency of $\mathbf{I} - \mathbf{H}$, the fourth equality uses the linear model equation $Y = \mathbf{X}\beta + \epsilon$, and the last equality follows since $\mathbf{H}\mathbf{X}\beta = \mathbf{X}\beta$ and so $(\mathbf{I}_n - \mathbf{H})\mathbf{X}\beta = 0$.

Now, to compute the bias, we take a conditional expectation, starting from the final expression in the

[1] Notice that the exact same notation is also used for residuals of OLS estimates, and OLS estimators.

previous chain of equalities:

$$\begin{aligned}\mathbb{E}\Big[\frac{1}{n}\epsilon^\top(\mathbf{I}_n-\mathbf{H})\epsilon|X_1,\dots,X_n\Big] &= \frac{1}{n}\mathbb{E}\Big[\sum_{i,j=1}^n \epsilon_i\epsilon_j(\mathbf{I}_n-\mathbf{H})_{ij}|X_1,\dots,X_n\Big]\\ &= \frac{1}{n}\sum_{i,j=1}^n(\mathbf{I}_n-\mathbf{H})_{ij}\mathbb{E}\Big[\epsilon_i\epsilon_j|X_1,\dots,X_n\Big]\\ &= \frac{1}{n}\sum_{i,j=1}^n(\mathbf{I}_n-\mathbf{H})_{ij}\mathbb{E}\Big[\epsilon_i\epsilon_j\Big]\\ &= \frac{1}{n}\sum_{i=1}^n(\mathbf{I}_n-\mathbf{H})_{ii}\mathbb{E}\Big[\epsilon_i^2\Big]\\ &= \frac{\sigma^2}{n}\sum_{i=1}^n(\mathbf{I}_n-\mathbf{H})_{ii}\\ &= \frac{\sigma^2}{n}\sum_{i=1}^n\mathrm{tr}(\mathbf{I}_n-\mathbf{H})\\ &= \frac{\sigma^2(n-p-1)}{n}.\end{aligned}$$

The third equality uses that $\epsilon \perp\!\!\!\perp X$, the fourth equality that ϵ_i, ϵ_j are independent and mean-zero, the fifth equality that $\mathrm{Var}[\epsilon_i] = \sigma^2$ for all i.

- The conclusion of this extended algebraic foray is that the plug-in estimate of variance is biased: $\mathbb{E}[\hat{\sigma}^2_{\text{PLUG}}] \neq \sigma^2$. But the nature of the bias is simple enough – and more importantly, doesn't depend on any unknown parameters besides σ^2 – to be easily fixed. Letting

$$\hat{\sigma}^2 = \frac{n}{(n-p-1)}\hat{\sigma}^2_{\text{PLUG}} = \frac{1}{n-p-1}\|e\|^2,$$

  it follows from the previous derivations that $\mathbb{E}[\hat{\sigma}^2] = \sigma^2$.

- In some rare problems, we might care about estimating σ^2 itself. For example, if the errors represented measurement error of some instrument, than estimating σ^2 would tell us how inaccurate the instrument is, which could be useful if we want to know when it needs replacing.

  More commonly, we will be more interested in estimating the covariance matrix $\mathrm{Cov}[\hat{\beta}|X_1,\dots,X_n]$, and in particular, the diagonal elements of this covariance. The previous work tells us that $\hat{\sigma}^2(\mathbf{X}^\top\mathbf{X})^{-1}$ is an unbiased estimator for $\mathrm{Cov}[\hat{\beta}|X_1,\dots,X_n]$. So we will estimate standard errors by

$$\widehat{SE}[\hat{\beta}_j|X_1,\dots,X_n] = \hat{\sigma}\sqrt{[(\mathbf{X}^\top\mathbf{X})^{-1}]_{(j+1,j+1)}}$$

3 Normal errors, and the sampling distribution of $\hat{\beta}$

- Having answered our first question – how to estimate the error variance, and standard error? – we turn to the second – what is the right number of standard errors to add and subtract from $\hat{\beta}$? It turns out that in order to properly answer this question, we need to make (yet) another assumption, on top of all those made by linear model: that the errors are **multivariate Normal**.

- **Linear model with Normal errors.** A random vector $Z \in \mathbb{R}^n$ is distributed multivariate Normal with mean μ and covariance $\mathbf{\Sigma}$, $Z \sim N_n(\mu, \mathbf{\Sigma})$, if it has density

$$p_Z(z) = \frac{1}{\sqrt{(2\pi)^n\det(\mathbf{\Sigma})}}\exp(-\frac{1}{2}(z-\mu)^\top\mathbf{\Sigma}^{-1}(z-\mu)).$$

Here $\det(\mathbf{\Sigma})$ is the determinant of $\mathbf{\Sigma}$. I won't define it as we won't have much use for it.

Notice that multivariate Normal distributions are completely determined by their mean and covariance, just like univariate Normal distributions are determined by their mean and variance.

- An important factor about multivariate Normals is that they are closed under linear transformations. That is, if $Z \sim N_n(\mu, \mathbf{\Sigma})$, and $\mathbf{A} \in \mathbb{R}^{p \times n}$ is a fixed matrix, then $\mathbf{A}Z \sim N_p(\mathbf{A}\mu, \mathbf{A}\mathbf{\Sigma}\mathbf{A}^\top)$. In fact, a useful criterion for checking multivariate Normality is that Z is multivariate Normal if and only if $v^\top Z$ is univariate Normal for every $v \in \mathbb{R}^n$.

- The linear model with Normal errors asserts the following: $(X_1, Y_1), \ldots, (X_n, Y_n)$ are independently distributed, with
$$Y = \mathbf{X}\beta + \epsilon,$$
where $\epsilon \perp\!\!\!\perp \mathbf{X}$, and $\epsilon \sim N_n(0, \sigma^2 \mathbf{I}_n)$. This is equivalent to saying that the conditional distribution of $Y | X_1, \ldots, X_n$ is also multivariate Normal: $Y | X_1, \ldots, X_n \sim N_n(\mathbf{X}\beta, \sigma^2 \mathbf{I}_n)$.

- We have already calculated the mean and variance of the OLS estimators in the linear model, without assuming Normal errors. Assuming Normal errors, however, we can figure out their entire **sampling distribution**.[2] Remember that
$$\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top Y.$$
Thus, conditional on $X_1, \ldots, X_n$, $\hat{\beta}$ is also multivariate Normal:
$$\hat{\beta} | X_1, \ldots, X_n \sim N_{p+1}(\beta, \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}).$$
It follows that each $\hat{\beta}_j$ is univariate Normal:
$$\hat{\beta}_j | X_1, \ldots, X_n \sim N(\beta_j, \sigma^2 [(\mathbf{X}^\top \mathbf{X})^{-1}]_{j+1, j+1})$$
You might have wondered why, in Lectures 2 and 4, we only computed bias and (co)variance of OLS, instead of (say) its entire moment generating function. One reason is that, when the errors are Normal, the distribution of $\hat{\beta}$ is complete determined by these two quantities.

- The sampling distribution of $\hat{\beta}_j$ can be used to answer questions such as "what is the probability that $\hat{\beta}_j$ is within 1 standard error of β_j?" Noting that
$$\frac{\hat{\beta}_j - \beta_j}{\mathrm{SE}[\hat{\beta}_j | X_1, \ldots, X_n]} \sim N(0, 1),$$
we have
$$\begin{aligned}
\mathbb{P}\Big(\hat{\beta}_j \in \big[\beta_j - \mathrm{SE}[\hat{\beta}_j | X_1, \ldots, X_n], \beta_j + \mathrm{SE}[\hat{\beta}_j | X_1, \ldots, X_n]\big]\Big) &= \mathbb{P}\Big(\beta_j - \mathrm{SE}[\hat{\beta}_j | X_1, \ldots, X_n] \leq \hat{\beta}_j \leq \mathrm{SE}[\hat{\beta}_j | X_1, \ldots, X_n]]\Big) \\
&= \mathbb{P}\Big(- \mathrm{SE}[\hat{\beta}_j | X_1, \ldots, X_n] \leq \hat{\beta}_j - \beta_j \leq \mathrm{SE}[\hat{\beta}_j | X_1, \ldots, X_n]] \\
&= \mathbb{P}\Big(- 1 \leq \frac{\hat{\beta}_j - \beta_j}{\mathrm{SE}[\hat{\beta}_j | X_1, \ldots, X_n]} \leq 1\Big) \\
&= \Phi(1) - \Phi(-1) \\
&\approx .683.
\end{aligned}$$
So, under the linear model with Normal errors, there is about a 68.3% chance that $\hat{\beta}_j$ is within one SE of the true value β_j. Using the frequency-based interpretation of probability, this gives us a sense of how often the rule of thumb delivers a range that actually contains the true parameter.

[2] Remember from Lecture 2 that the distribution of an estimator is called a sampling distribution, with the modifier "sampling" used to distinguish it from the distribution of the random variables in the model.

- **Is real data Normal?** We know from the Central Limit Theorem that Normality (approximately) holds when the random variable in question is the sum or average of many independent random variables. So if we believe the total "error" in our regression can itself be further decomposed into the sum of many different independent random variables, then perhaps Normality is plausible. This is the case in e.g communication networks, where errors are due to the cumulative effect of many different interferences in the network.

  In many applications, however, there's no better justification for Normality than any of the other assumptions in the linear model. Making strong assumptions like Normal errors was important in the pre-computer age, when it led to a simple and tractable method for computing confidence intervals. The Normal theory is still taught today, partly due to inertia, partly because it gives a baseline against which to compare other, more sophisticated methods, which we will learn about later in this course. And when we do use the Normal theory, it is important to check that the Normal assumption is plausible through appropriate diagnostics (which, again, will come later).

4 Confidence intervals

- At last, we are prepared to give confidence intervals for parameters in the linear models.
- Since confidence intervals are a source of much confusion, let's start with a crisp mathematical definition of what we're after. For a given $\alpha \in (0,1)$, a $(1-\alpha)$ **confidence interval for** β_j is an interval $C(x_1, y_1, \ldots, x_n, y_n) = [L(x_1, y_1, \ldots, x_n, y_n), U(x_1, y_1, \ldots, x_n, y_n)]$ computable from the data, such that if $(X_1, Y_1), \ldots, (X_n, Y_n)$ come from the linear model:

$$\mathbb{P}\Big(\beta_j \in C_j(X_1, Y_1, \ldots, X_n, Y_n)\Big) \geq 1 - \alpha. \tag{2}$$

  Notice what is random in (2) – the interval C_j – and what is fixed – the unknown parameter β_j.

- Just like with estimates, the standard notation is to abbreviate both $C_j(x_1, y_1, \ldots, x_n, y_n)$ *and* $C_j(X_1, Y_1, \ldots, X_n, Y_n)$ by C_j. In this case, whether C_j is a function of data (and thus a fixed interval) or a function of random variables (and thus a random interval) depends on the context.

Wald intervals with known variance. There are different ways to compute confidence intervals, but for the most part we will stick with intervals constructed by taking the estimate, and adding and subtracting a certain number of standard errors. The resulting intervals are called the **Wald intervals**.

- When the error variance is known, the $(1-\alpha)$ **Wald interval** for β_j is based on quantiles of the Normal distribution. Recall that the **quantile** of a distribution is the inverse CDF; that is, if X has CDF F_X, then the quantile $Q_X(\alpha) := F_X^{-1}(\alpha)$. Letting $z^{\alpha/2} := \Phi^{-1}(\alpha/2)$ denote the $\alpha/2$th quantile of a standard Normal distribution, the Wald interval

$$C_j = \Big[\hat{\beta}_j + z^{(\alpha/2)}\sigma\sqrt{[(\mathbf{X}^\top \mathbf{X})^{-1}]_{j+1,j+1}}, \hat{\beta}_j + z^{(1-\alpha/2)}\sigma\sqrt{[(\mathbf{X}^\top \mathbf{X})^{-1}]_{j+1,j+1}}\Big]$$

  Notice that the Wald interval is symmetric as $z^{(\alpha/2)} = -z^{(1-\alpha/2)}$.

Let's verify that this is truly a $(1-\alpha)$ confidence interval. As usual, we condition on the predictors:

$$\begin{aligned}
\mathbb{P}\big(\beta_j \in C_j|\mathbf{X}\big) &= \mathbb{P}\Big(\hat{\beta} z^{(\alpha/2)}\sigma\sqrt{[(\mathbf{X}^\top\mathbf{X})^{-1}]_{j+1,j+1}} \leq \beta_j \leq \hat{\beta} + z^{(1-\alpha/2)}\sigma\sqrt{[(\mathbf{X}^\top\mathbf{X})^{-1}]_{j+1,j+1}}|\mathbf{X}\Big) \\
&= \mathbb{P}\Big(z^{(\alpha/2)}\sigma\sqrt{[(\mathbf{X}^\top\mathbf{X})^{-1}]_{j+1,j+1}} \leq \beta_j - \hat{\beta}_j \leq z^{(1-\alpha/2)}\sigma\sqrt{[(\mathbf{X}^\top\mathbf{X})^{-1}]_{j+1,j+1}}|\mathbf{X}\Big) \\
&= \mathbb{P}\Big(z^{(\alpha/2)} \leq \frac{\beta_j - \hat{\beta}_j}{\sigma\sqrt{[(\mathbf{X}^\top\mathbf{X})^{-1}]_{j+1,j+1}}} \leq z^{(1-\alpha/2)}|\mathbf{X}\Big) \\
&= \Phi\Big(z^{(1-\alpha/2)}\Big) - \Phi\Big(z^{(\alpha/2)}\Big) \\
&= 1 - \alpha/2 - \alpha/2 \\
&= 1 - \alpha.
\end{aligned}$$

By the law of total expectation, $\mathbb{P}(\beta_j \in C_j) = 1 - \alpha$, so C_j is a $(1-\alpha)$ confidence interval, both marginal and conditionally on $\mathbf{X}$.

- **Wald intervals with unknown error variance.** What about when the error variance is unknown? The obvious guess is to simply "plug in" the estimated error variance $\hat{\sigma}$ for the unknown σ. This will work well if $n-p$ is sufficiently large, but leads to intervals that are (ever so slightly) too narrow if $n-p$ is small.[3] This is because

  $$\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}\sqrt{[(\mathbf{X}^\top\mathbf{X})^{-1}]_{j+1,j+1}}} \not\sim N(0,1),$$

  due to the extra variability in $\hat{\sigma}$. Instead it follows a distribution called **(Student's) T-distribution with $n-p-1$ degrees of freedom**:

  $$\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}\sqrt{[(\mathbf{X}^\top\mathbf{X})^{-1}]_{j+1,j+1}}} \sim t_{n-p-1}.$$

  Letting $t_{n-p-1}^{(\alpha)}$ be the αth quantile of the T-distribution, we have that the Wald interval

  $$C_j = \Big[\hat{\beta}_j + t_{n-p-1}^{(\alpha/2)}\sigma\sqrt{[(\mathbf{X}^\top\mathbf{X})^{-1}]_{j+1,j+1}}, \hat{\beta}_j + t_{n-p-1}^{(1-\alpha/2)}\sigma\sqrt{[(\mathbf{X}^\top\mathbf{X})^{-1}]_{j+1,j+1}}\Big]$$

  is a $(1-\alpha)$ confidence interval for β_j.

[3] A typically cited if totally arbitrary cutoff is $n-p > 20$.

Delivery example redux: computing SEs

Recall the delivery example from Lecture 3, in which we attempted to predict driver delivery time using number of cases delivered and distance walked. Our model was that the data were observed values of independent samples of $(Time_i, Cases_i, Distances_i)$, each distributed according to

$$\text{Time}_i = \beta_0 + \beta_1 \text{cases}_i + \beta_2 \text{distance}_i + \epsilon_i,$$

where ϵ_i are independent of the predictors, $\mathbb{E}[\epsilon_i] = 0$ and $\text{Var}[\epsilon_i] = \sigma^2$. This was deemed plausible on the basis of some (hasty) EDA. So we fit a linear model using OLS.

```r
# read in the data
delivery = read.table("delivery.txt")

# fit the linear model
delivery.lm = lm(Time ~ Cases + Distance,data = delivery)

# report the coefficients
delivery.coefficients = coefficients(delivery.lm) # the same as delivery.lm$coefficients
print(delivery.coefficients,digits = 2)
```

```
## (Intercept)       Cases    Distance
##       2.341       1.616       0.014
```

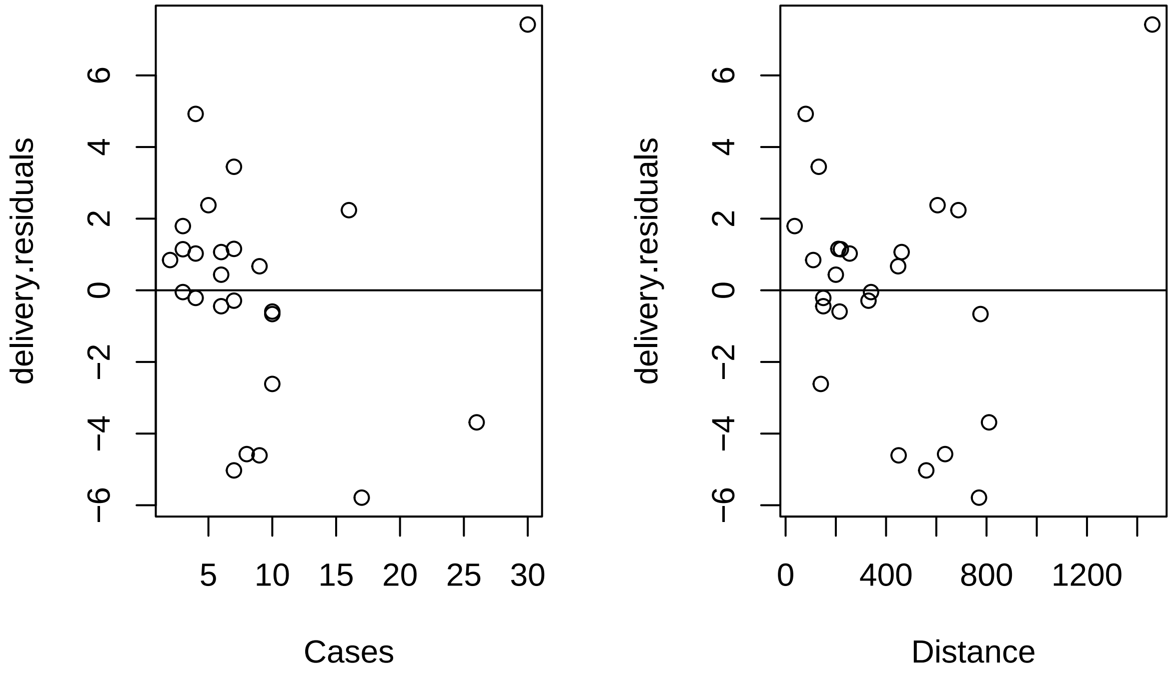
Our interpretation of the coefficients is…

A taste of diagnostic plotting. Before proceeding to inference, we would like to check the assumptions of the model. Usually this boils down to making a bunch of plots. For example, to check whether the errors are independent of the predictors, we could plot the errors against each predictor…except we don’t know the errors. (This is why diagnostics are hard.)

Instead, we plot the residuals.

```r
# residuals
delivery.residuals = delivery.lm$residuals

# residual plots
par(mfrow = c(1,2)) # format the output
plot(x = delivery$Cases,y = delivery.residuals,xlab = "Cases"); abline(h = 0)
plot(x = delivery$Distance,y = delivery.residuals,xlab = "Distance"); abline(h = 0)
```



In these **residual plots** we are hoping to see an *absence* of pattern. That is, if the errors come from the linear model, we should expect to see that the residuals are

- centered at the horizontal line (checking that $\mathbb{E}[\epsilon] = 0$),
- roughly equal width (checking that $\text{Var}[\epsilon] = \sigma^2$),
- no discernible trend as a function of the predictors (ϵ independent of the predictors).

This is fairly debatable for the observed residuals but we proceed anyway.

Estimating SEs

We could compute estimates of the SEs using the formula.

```
X = cbind(1,delivery$Cases,delivery$Distance)
n = nrow(X); p = ncol(X) - 1
P = solve(t(X) %*% X)
sigmahat = sqrt(sum(delivery.residuals^2)/(n - p - 1))
sehat = sqrt(diag(P)) * sigmahat
print(sehat,digits = 3)
```

```
## [1] 1.09673 0.17073 0.00361
```

As usual, there is a simpler way.

```
summary(delivery.lm)$coefficients[,"Std. Error"]
```

```
## (Intercept)       Cases    Distance
## 1.096730168 0.170734918 0.003613086
```

The two agree.

Confidence intervals

Finally, we compute 95% Wald confidence intervals using the formula.

```
alpha = .05
qnalpha = qnorm(1 - alpha/2)
cbind(delivery.coefficients - qnalpha * sehat, delivery.coefficients + qnalpha * sehat)
```

```
##                    [,1]       [,2]
## (Intercept) 0.191679515 4.49078278
## Cases       1.281272920 1.95054150
## Distance    0.007303308 0.02146634
```

Or, the simpler way.

```
confint(delivery.lm,level = .95)
```

```
##                   2.5 %     97.5 %
## (Intercept) 0.066751987 4.61571030
## Cases       1.261824662 1.96998976
## Distance    0.006891745 0.02187791
```

Notice that the latter intervals are slightly wider, since they use the quantile of the T-distribution with $n - p - 1 = 22$ degrees of freedom, rather than the quantile of the standard Normal distribution.