Confidence intervals for OLS coefficients

Stats 203, Winter 2024

Lecture 6 [Last update: January 29, 2024]

1 Review and preview

- Last week's topic was representing the error in the OLS coefficients using the standard error. On Monday, in Lecture 4, we worked out that the standard error of $\hat{\beta}_j$, conditional on the predictors $\mathbf{X}$, is

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

- However, we can't calculate the standard error from data since it depends on the unknown error variance σ^2. Last class, in Lecture 5, we discussed estimating σ^2 with the plug-in estimator

$$\hat{\sigma}^2_{\mathrm{PLUG}} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \sum_{i=1}^{p}\hat{\beta}_j X_{ij})^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \frac{1}{n}\|e\|^2.$$

  We derived, through a lengthy calculation, that the plug-in estimator is biased for the true error variance:

$$\mathbb{E}[\hat{\sigma}^2_{\mathrm{PLUG}}] = \frac{n-p-1}{n}\sigma^2,$$

  and consequently that the estimator $\hat{\sigma}^2 = \frac{1}{n-p-1}\|e\|^2$ is unbiased.

- 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 $\mathrm{Cov}[\hat{\beta}|\mathbf{X}]$, 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 $\mathrm{Cov}[\hat{\beta}|\mathbf{X}]$. So we will estimate standard errors by

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

  We now have an estimate of the error in our estimates that can be calculated using only the data at hand.

- Today, we will show how estimates of standard errors can be used to provide confidence intervals for the true β_js. In order to do this, we will need more than just the expectation and covariance of $\hat{\beta}$. Rather, we will need its entire distribution.

- Note that today's notes are in part copy/pasted from the latter half of Lecture 5 notes.

2 The linear model with normal errors

- In order to calculate the sampling distribution of $\hat{\beta}$, 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. A consequence of this is that zero-covariance implies independence for multivariate Normally distributed random vectors. Specifically, this means that for random vectors $Z \in \mathbb{R}^p$ and $Z' \in \mathbb{R}^q$ that are jointly Normal,[1] if $\text{Cov}[Z, Z'] = \mathbf{0}_{p\times q}$ then $Z \perp\!\!\!\perp Z'$.

- An important fact 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 $Y|\mathbf{X}$ is also multivariate Normal: $Y|\mathbf{X} \sim N_n(\mathbf{X}\beta, \sigma^2\mathbf{I}_n)$.

- **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 error term in our regression equation 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).

3 Sampling distribution of $\hat{\beta}, e, \hat{\sigma}^2$

- **Sampling distribution of $\hat{\beta}$.** 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 $\mathbf{X}$, $\hat{\beta}$ is also multivariate Normal:

$$\hat{\beta}|\mathbf{X} \sim N_{p+1}(\beta, \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}).$$

  It follows that each $\hat{\beta}_j$ is univariate Normal:

$$\hat{\beta}_j|\mathbf{X} \sim N(\beta_j, \sigma^2[(\mathbf{X}^\top\mathbf{X})^{-1}]_{j+1,j+1})$$

[1] Meaning that $(Z, Z') \sim N_{p+q}$.

[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.

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 completely 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?" The key is to notice that the **Z-score**

$$Z_j := \frac{\hat{\beta}_j - \beta_j}{\text{SE}[\hat{\beta}_j|\mathbf{X}]} \sim N(0,1), \quad \text{conditional on } \mathbf{X}.$$

Using this,

$$\begin{aligned}
\mathbb{P}\Big(\hat{\beta}_j \in \big[\beta_j - \text{SE}[\hat{\beta}_j|\mathbf{X}], \beta_j + \text{SE}[\hat{\beta}_j|\mathbf{X}]\big]\Big) &= \mathbb{P}\Big(\beta_j - \text{SE}[\hat{\beta}_j|\mathbf{X}] \le \hat{\beta}_j \le \beta_j + \text{SE}[\hat{\beta}_j|\mathbf{X}]\big]\Big) \\
&= \mathbb{P}\Big(- \text{SE}[\hat{\beta}_j|\mathbf{X}] \le \hat{\beta}_j - \beta_j \le \text{SE}[\hat{\beta}_j|\mathbf{X}]\big]\Big) \\
&= \mathbb{P}\Big(- 1 \le \frac{\hat{\beta}_j - \beta_j}{\text{SE}[\hat{\beta}_j|\mathbf{X}]} \le 1\Big) \\
&= \Phi(1) - \Phi(-1) \\
&\approx .683.
\end{aligned} \tag{1}$$

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.

It is obviously quite important that the probability of coverage ends up not depending on the parameters, that is, on β or σ^2. Otherwise, we would not be able to assess our level of confidence since the parameters are themselves unknown, and the whole exercise would have been of dubious value. Examining the steps in (1), we can see that the reason the probability does not depend on the parameters is that the distribution of the Z-score is $N(0,1)$, no matter what β and σ^2 are. In general, if the errors are not Normal, there is no reason to expect that rule as simple as "estimator plus and minus 1 SE" would contain the true value with a probability not depending on the parameters.

- **Sampling distribution of e.** The sampling distribution of the residuals is also Normal: noting that $e = (\mathbf{I}_n - \mathbf{H})Y = (\mathbf{I}_n - \mathbf{H})\epsilon$, we have that

$$e \sim N_p(0, \sigma^2(\mathbf{I}_n - \mathbf{H})), \quad \text{conditional on } \mathbf{X}.$$

- **Sampling distribution of $\hat{\sigma}^2$.** What follows is a somewhat fast and loose derivation of the sampling distribution of $\hat{\sigma}^2$. Throughout, all probabilistic statements are conditional on $\mathbf{X}$. We may or may not go through this in class: if we do, I will add appropriate details as relevant.

To work out the sampling distribution of $\hat{\sigma}^2$, we start by noting that

$$\begin{aligned}
\hat{\sigma}^2 &= \frac{1}{n-p-1}\epsilon^\top(\mathbf{I}_n - \mathbf{H})\epsilon \\
&= \frac{1}{n-p-1}\epsilon^\top(\mathbf{I}_n - \mathbf{H})^2\epsilon \\
&= \frac{1}{n-p-1}\|(\mathbf{I}_n - \mathbf{H})\epsilon\|^2.
\end{aligned}$$

The first equality we showed in Lecture 5, and the second equality follows since $(\mathbf{I}_n - \mathbf{H})$ is idempotent.

At this point we will need to derive another representation for $\mathbf{I}_n - \mathbf{H}$. Recall that in the Lecture 5 notes we showed that $\mathbf{I}_n - \mathbf{H}$ is a projector matrix (i.e idempotent), and that it projects onto a subspace of dimension $n-p-1$; this subspace is $\text{col}(\mathbf{X})^\perp$, the orthogonal complement of the span of the columns

of $\mathbf{X}$. Let $v_1, \ldots, v_{n-p-1}$ be a (not unique) orthonormal basis of this subspace, meaning that $\|v_k\|^2 = 1$, $v_k^\top v_\ell = 0$ for $k \neq \ell$, and $\text{span}(v_1, \ldots, v_{n-p-1}) = \text{col}(\mathbf{X})^\perp$.[3] It follows that

$$\mathbf{I}_n - \mathbf{H} = \sum_{k=1}^{n-p-1} v_k v_k^\top.$$

But this means that when applied to the vector of residuals,

$$(\mathbf{I}_n - \mathbf{H})\epsilon = \sum_{k=1}^{n-p-1} v_k v_k^\top \epsilon =: \sum_{k=1}^{n-p-1} v_k \sigma U_k,$$

where $U_k = v_k^\top \epsilon / \sigma$, and $U = (Z_1, \ldots, Z_{n-p-1}) \sim N(0, \mathbf{I}_{n-p-1})$, conditionally on $\mathbf{X}$. (A good exercise is to confirm that this is indeed the mean and covariance of U). Squaring and summing, we have that

$$\begin{aligned}\hat{\sigma}^2 &= \frac{\sigma^2}{n-p-1} \sum_{k,\ell=1}^{n-p-1} v_k^\top v_\ell U_k U_\ell \\ &= \frac{\sigma^2}{n-p-1} \sum_{k=1}^{n-p-1} U_k^2,\end{aligned}$$

with the second equality following because the v_ks are orthonormal.

We have arrived at kind of a cool result. Our unbiased estimate of error variance $\hat{\sigma}^2$, despite being an average of n squared residuals $e_1^2, \ldots, e_n^2$ which are themselves *dependent*, is equal (after appropriate rescaling) to the sum of $n-p-1$ *independent* squared standard Normal random variables $U_1^2, \ldots, U_{n-p-1}^2$.

The sum of k independent squared Normal random variables follow a distribution called the **chi-squared distribution with k degrees of freedom**, denoted χ_k^2. So we have shown that

$$(n-p-1)\hat{\sigma}^2/\sigma^2 \sim \chi^2_{n-p-1}, \quad \text{conditional on } \mathbf{X}.$$

This distribution has a density, of course, but I will not write it out as we will never use it. Likewise, the distribution of $\sqrt{n-p-1}\hat{\sigma}/\sigma$ is the square-root of a sum of squared standard Normals, which is sometimes refereed to as a **chi distribution with $n-p-1$ degrees of freedom** and denoted $\sqrt{n-p-1}\hat{\sigma}/\sigma \sim \chi_{n-p-1}$.

- **Independence of $\hat{\beta}$ and e.** So far we have separately derived the joint distribution of $\hat{\beta}$ and the marginal distribution $\hat{\sigma}^2$. It turns out that this completely determines the joint distribution of $(\hat{\beta}, \hat{\sigma}^2)$, since $\hat{\beta}$ and $\hat{\sigma}$ are independent conditional on $\mathbf{X}$.

  In fact the independence of $\hat{\beta}, \hat{\sigma}$ is a consequence of the more general independence between $\hat{\beta}$ and the residuals e. Let's compute the covariance between $\hat{\beta}$ and e, again conditioning on $\mathbf{X}$:

$$\begin{aligned}\text{Cov}[e, \hat{\beta}|\mathbf{X}] &= \text{Cov}[(\mathbf{I}_n - \mathbf{H})Y, (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top Y|\mathbf{X}] \\ &= \text{Cov}[(\mathbf{I}_n - \mathbf{H})\epsilon, (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top \epsilon|\mathbf{X}] \\ &= (\mathbf{I}_n - \mathbf{H})\text{Cov}[\epsilon, \epsilon|\mathbf{X}]\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{I}_n - \mathbf{H})\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1} \\ &= \mathbf{0}_{n\times(p+1)}.\end{aligned}$$

  The key final equality follows because $\mathbf{HX} = \mathbf{X}$, which was derived in Lecture 5 notes.

  In summary, we have shown that conditional on $\mathbf{X}$, e and $\hat{\beta}$ are multivariate Normal with covariance equal to $\mathbf{0}_{n\times(p+1)}$. Thus, they are independent. Noting that $\hat{\sigma}^2$ is a deterministic function of e, this means $\hat{\beta}$ and $\hat{\sigma}^2$ are also independent.

[3] You can find such an orthonormal basis by running Gram-Schmidt.

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 $Q_X(\alpha) := F_X^{-1}(\alpha)$. We use Φ to denote the CDF of a standard Normal random variable. Letting $z^{(\alpha/2)} := \Phi^{-1}(\alpha/2)$ denote the $\alpha/2$th quantile of a standard Normal distribution, the $(1-\alpha)$ Wald interval is

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

  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:

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

  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 standard deviation $\hat{\sigma}$ for the unknown σ. This will work well if $n-p$ is sufficiently large, but leads to intervals that are (a little) too narrow if $n-p$ is small.[4] This is because

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

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

due to the extra variability in $\hat{\sigma}$. Instead, using the results of the previous section, what we have is that

$$\begin{aligned}\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}\sqrt{[(\mathbf{X}^\top\mathbf{X})^{-1}]_{j+1,j+1}}} &= \frac{(\hat{\beta}_j - \beta_j)/\mathrm{SE}[\hat{\beta}_j|\mathbf{X}]}{\hat{\sigma}/\sigma} \\ &= \frac{Z_j}{\sqrt{\frac{1}{n-p-1}\sum_{k=1}^{n-p-1} U_k^2}}.\end{aligned}$$

Furthermore, $Z_j \sim N(0,1)$, $\sum_{k=1}^{n-p-1} U_k^2 \sim \chi^2_{n-p-1}$, and the two random variables are independent. The ratio of an independent standard Normal random variable over a chi random variable with d degrees of freedom is common enough that it gets a name: **(Student's) T-distribution with d degrees of freedom**, denoted $T \sim t_d$. So we have shown that

$$\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}[(\mathbf{X}^\top\mathbf{X})^{-1}]_{j+1,j+1}} \sim t_{n-p-1}. \tag{3}$$

As with the chi-squared distribution, the T-distribution can be defined in terms of its density function, but I won't write it down because it isn't super important.

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

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

is a $(1-\alpha)$ confidence interval for β_j. As with the Normal distribution, the T-distribution is symmetric about zero and so $t^{(\alpha/2)}_{n-p-1} = -t^{(1-\alpha/2)}_{n-p-1}$.

Importantly, the distribution in (3) does not depend on the parameters (β, σ^2), since it is the ratio of two independent random variables, both which have marginal distributions not depending on parameters.

- Now that you know how to calculate confidence intervals, you should **always** report one along with each estimated coefficient. Reporting a confidence interval requires choosing a confidence level $(1-\alpha)$. The rule of thumb corresponded to picking $\alpha = .317$, as we have seen. I will usually adhere to the (totally arbitrary) convention of choosing $\alpha = .05$, which corresponds to calculating 95% confidence intervals. The reason this became convention is first that 95% is a nice big number, and second that $z^{(.025)} = 1.96 \approx 2$, so that you can fudge and just report plus and minus two (estimated) standard errors.

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.

```
# 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
```

Interpretation of OLS coefficients.

The fully correct way to interpret the results of our regression would be to say that:

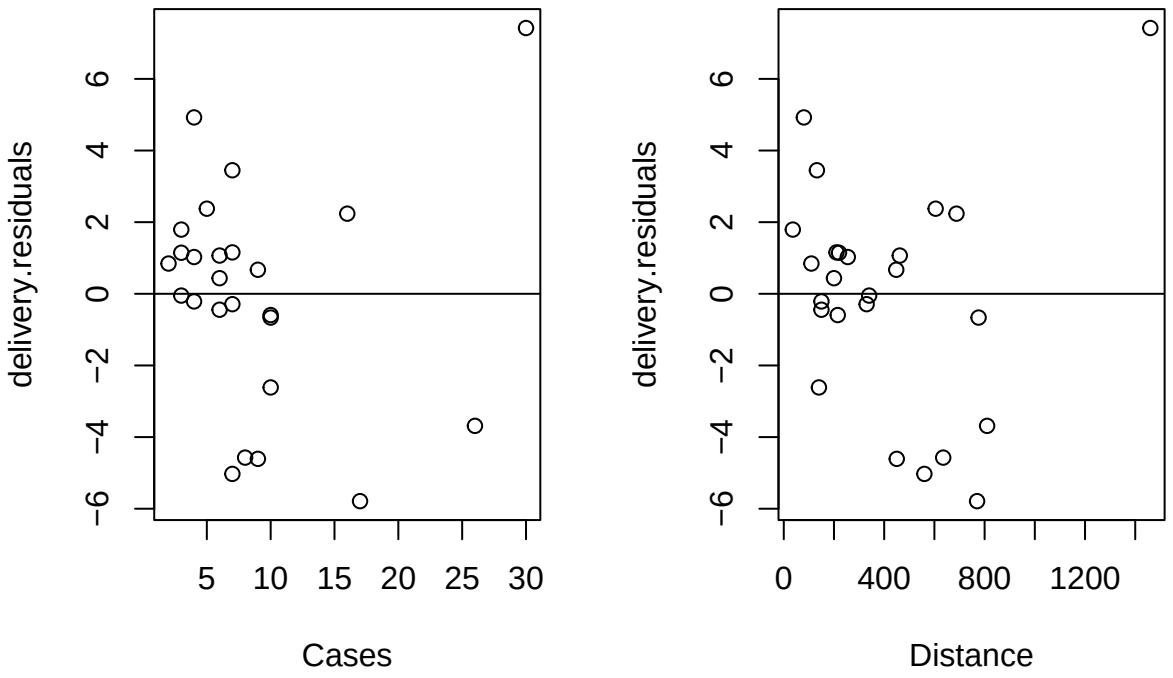
- We estimate that the difference in expected driver delivery time at two locations with the same distance walked, but different number of cases delivered, is 1.616 times the difference in number of cases delivered.
- We estimate that the difference in expected driver delivery time at two locations with the same number of cases delivered, but different distance walked, is .014 times the difference in distance walked.

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.

```
# 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.

Interpretation of confidence intervals.

The fully correct way to interpret our calculated confidence intervals would be to say that:

- We are 95% confident that the true difference in expected driver delivery time at two locations with the same distance walked, but different number of cases delivered, is between 1.26 and 1.97 times the difference in number of cases delivered.
- We are 95% confident that the true difference in expected driver delivery time at two locations with the same number of cases delivered, but different distance walked, is between .007 and .022 times the difference in distance walked.

Since this is such a mouthful, I am ok with the simpler:

- We are 95% confident that the true value of β_1 is between 1.26 and 1.97.
- We are 95% confident that the true value of β_2 is between .006 and .022.

In all places, the phrase "95% confident'' is used to refer to the long-run frequency interpretation discussed at the end of Lecture 4 notes: if the linear model with Normal errors were truly correct, and we repeatedly constructed datasets by realizing values of random variables according to this model, then the resulting confidence intervals would contain the true parameters 95% percent of the time.