11.3 Inference for Regression

In a typical setting, if we are using $(x_1, y_1), \ldots, (x_n, y_n)$ to estimate β_0, β_1, then we likely need to estimate the variance of the residuals σ^2 as well. Define a random variable S^2 by

$$S^2 = \sum (Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2$$

and let $\hat{\sigma}^2 = \frac{S^2}{n}$.

Thm: The MLE for σ^2 is $\hat{\sigma}^2$.

Proof. Consider the log-likelihood function

$$\log f(\mathbf{y}|\beta_0, \beta_1, \sigma^2, \mathbf{x}) = -\frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum (y_i - (\beta_0 + \beta_1 x_i))^2$$

Differentiating with respect to σ^2 and setting equal to 0:

$$0 = \frac{\partial}{\partial(\sigma^2)} \log f = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum (y_i - (\beta_0 + \beta_1 x_i))^2$$

and solving for σ^2 gives

$$\sigma^2 = \frac{1}{n}\sum (y_i - (\beta_0 + \beta_1 x_i))^2$$

Replacing β_0 and β_1 with their MLEs $\hat{\beta}_0$ and $\hat{\beta}_1$ gives the desired result. □

Just as the sample mean and sample variance are were independent for data $X_i \sim N(\mu, \sigma^2)$, a similar result is true for data from the linear model.

Thm: The variables $\hat{\sigma}^2$ and $\hat{\beta}_0, \hat{\beta}_1$ are independent, and $\frac{S^2}{\sigma^2}$ has χ^2 distribution with $n-2$ degrees of freedom.

Proof. We don't give a full proof here, but the idea is similar to the one used to show that sample mean and sample variance are independent. In particular, we would show that $\mathbf{Z} = (\hat{\beta}_0, \hat{\beta}_1, Y_1 - (\hat{\beta}_0 + \hat{\beta}_1 x_1), \ldots, Y_n - (\hat{\beta}_0 + \hat{\beta}_1 x_n))$ are multivariate Normal. And then show that $\hat{\beta}_0$ and $\hat{\beta}_1$ are each uncorrelated with the remaining terms. Since the vector $\mathbf{Z}$ is MVN, then this implies that $\hat{\beta}_0$ and $\hat{\beta}_1$ are actually independent of the $Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$. Finally, since S^2 is a function just of the $Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$, this implies that $\hat{\beta}_1$ and $\hat{\beta}_0$ are independent of S^2.

To see that $\hat{\sigma}^2/\sigma^2$ is $\chi^2(n-1)$, we first assume $\sigma^2 = 1$. Then we compute the MGF for the sum of squared entries of $\mathbf{Z}$ (after Normalizing $\hat{\beta}_0, \hat{\beta}_1$ to have variance 1 and mean 0), which will be the MGF of the χ^2 distribution with n degrees of freedom. Then, using independence, we factor out the MGFs for $\hat{\beta}_0^2$ and $\hat{\beta}_1^2$, which leaves the MGF of S^2, which takes the form of the MGF of a $\chi^2(n-2)$ distribution. The general result is then obtained by dividing $\mathbf{Z}$ by σ and modifying the MGF accordingly. □

Hypothesis Testing

For the remainder of this section, it will be convenient to define

$$\sigma' = \frac{S}{\sqrt{n-2}}$$

Now, the preceding theorem showed that the joint distribution of $(\hat{\beta}_1, \hat{\beta}_2)$ is bivariate Normal, and so in particular, any linear combination of these variables is Normally distributed. We will use this to derive hypothesis testing procedures for general linear combinations

$$c_0\beta_0 + c_1\beta_1$$

Specializing to $c_0 = 0, c_1 = 1$ gives a hypothesis test for β_1, while $c_0 = 1, c_1 = 0$ gives a hypothesis test for β_0. Finally, $c_0 = 1, c_1 = x$ gives a hypothesis test of $\beta_0 + \beta_1 x$ for arbitrary x.

Thm: Let c_0, c_1, c_* be specific numbers, where at least 1 of c_0, c_1 is nonzero. Consider the hypotheses

$$H_0 : c_0\beta_0 + c_1\beta_1 = c_* \quad H_0 : c_0\beta_0 + c_1\beta_1 \neq c_*$$

Define

$$U_{01} = \left[\frac{c_0^2}{n} + \frac{(c_0\bar{x} - c_1)^2}{s_x^2}\right]^{-1/2} \left(\frac{c_0\hat{\beta}_0 + c_1\hat{\beta}_1 - c_*}{\sigma'}\right)$$

For each $0 < \alpha_0 < 1$, let δ be the test which rejects H_0 when $|U_{01}| \geq F^{-1}_{t(n-2)}(1 - \alpha_0/2)$. Then δ is a level α_0 test.

Proof. The variable $c_0\hat{\beta}_0 + c_1\hat{\beta}_1$ is Normally distributed with mean $c_0\beta_0 + c_1\beta_1$ and variance

$$c_0^2\text{Var}(\hat{\beta}_0) + c_1^2\text{Var}(\hat{\beta}_1) + 2c_0c_1\text{Cov}(\hat{\beta}_0, \hat{\beta}_1) = \sigma^2\left(\frac{c_0^2}{n} + \frac{(c_0\bar{x} - c_1)^2}{s_x^2}\right)$$

Hence, under H_0, the following variable W_{01} has the standard Normal distribution:

$$W_{01} = \left[\frac{c_0^2}{n} + \frac{(c_0\bar{x} - c_1)^2}{s_x^2}\right]^{-1/2} \left(\frac{c_0\hat{\beta}_0 + c_1\hat{\beta}_1 - c_*}{\sigma}\right)$$

By the preceding theorem, the statistic S^2/σ^2 has $\chi^2(n-1)$ distribution, and is independent of $\hat{\beta}_0, \hat{\beta}_1$, hence

$$U_{01} = \frac{W_{01}}{\sqrt{\frac{1}{n-2}\frac{S^2}{\sigma^2}}}$$

has t distribution with $n-2$ degrees of freedom. □

Note: When you call on `lm()` in R to create a linear model and view the summary table, **THIS** is the hypothesis test whose P-value is reported.

Cor: Let β_1* be a specific value and consider hypotheses for slope parameter of the regression model

$$H_0 : \beta_1 = \beta_1^* \qquad H_1 : \beta_1 \neq \beta_1^*$$

Let U_1 be the test statistic

$$U_1 = s_x\frac{\hat{\beta}_1 - \beta_1^*}{\sigma'}$$

and let δ be the procedure which reject H_0 when $|U_1| \geq T^{-1}_{n-2}(1 - \alpha_0/2)$. Then δ is a level α_0 test, and the p-value for $(\mathbf{x}, \mathbf{y})$ with observed statistic u_1 is

$$P(U_1 \geq |u_1|) + P(U_1 \leq -|u_1|)$$

Cor: Suppose we wish to assess whether it is plausible that the regression line $y = \beta_0 + \beta_1 x$ passes through a particular point (x^*, y^*). This is equivalent to testing the hypotheses

$$H_0 : \beta_0 + \beta_1 x^* = y^* \qquad h_1 : \beta_0 + \beta_1 x^* \neq y^*$$

Our hypotheses are of the form $c_0 = 1, c_1 = x^*, c_* = y^*$, and the appropriate test statistic is

$$U_{01} = \left[\frac{1}{n} + \frac{(\bar{x} - x^*)^2}{s_x^2}\right]^{-1/2} \left(\frac{\hat{\beta}_0 + x^*\hat{\beta}_1 - y^*}{\sigma'}\right)$$

The test δ which rejects H_0 when $|U_{01}| \geq T_{n-2}^{-1}(1 - \alpha_0/2)$ is a level α_0 test.

Cor: Let c_0, c_1 be specific numbers, where at least 1 of c_0, c_1 is nonzero. The interval given by

$$c_0\hat{\beta}_0 + c_1\hat{\beta}_1 \pm \sigma' \left[\frac{c_0^2}{n} + \frac{(c_0\bar{x} - c_1)^2}{s_x^2}\right] F_{t(n-2)}^{-1}\left(1 - \frac{\alpha_0}{2}\right)$$

is a $1 - \alpha_0$ level confidence interval for $c_0\beta_0 + c_1\beta_1$.

Proof. Let $\omega(\mathbf{x}, \mathbf{y})$ be the set of all value c_* for which H_0 is not rejected at the α_0 level when $\mathbf{x}, \mathbf{y}$ is observed. Let $q = T_{n-2}^{-1}(1 - \alpha_0/2)$ and $SE = \sigma' \left[\frac{c_0^2}{n} + \frac{(c_0\bar{x} - c_1)^2}{s_x^2}\right]^{1/2}$. Then

$$\begin{aligned}
c_* \in \omega(\mathbf{x}, \mathbf{y}) &\iff |U_{01}| < q \\
&\iff -q < \frac{c_0\hat{\beta}_0 + c_1\hat{\beta}_1 - c_*}{SE} < q \\
&\iff -SE \cdot q < c_0\hat{\beta}_0 + c_1\hat{\beta}_1 - c_* < SE \cdot q \\
&\iff c_0\hat{\beta}_0 + c_1\hat{\beta}_1 + SE \cdot q < c_* < c_0\hat{\beta}_0 + c_1\hat{\beta}_1 - SE \cdot q
\end{aligned}$$

The result follows by noting that $\omega(\mathbf{x}, \mathbf{Y})$ is a $1 - \alpha_0$ confidence set (by a previous theorem). □

Suppose that rather creating an interval estimate for regression coefficient of the model, we wish to construct an interval (A, B) that contains a single observation Y with specified probability $1 - \alpha_0$. We can slightly modify the previous procedure to do.

Thm: Let (Y, x) be an independent observation from the model, and let $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$. Define an interval (A, B) by

$$\hat{Y} \pm T_{n-2}^{-1}(1 - \alpha_0/2)\sigma' \left[1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{s_x^2}\right]^{1/2}$$

Then $P(Y \in (A, B)) = 1 - \alpha_0$.

Proof. Since Y and $\hat{Y}$ are independent Normal and $E[Y] = E[\hat{Y}]$, then $Y - \hat{Y}$ is Normal with mean 0 and variance $\sigma + \sigma\left[\frac{1}{n} + \frac{(x-\bar{x})^2}{s_x^2}\right]^{1/2}$. Let

$$Z = \frac{Y - \hat{Y}}{\sigma\left[1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{s_x^2}\right]^{1/2}} \qquad W = \frac{S^2}{\sigma^2}$$

where Z is $N(0, 1)$ and W is $\chi^2(n - 2)$. Then $U_x = \frac{Z}{\sqrt{W/(n-2)}}$ is $t(n - 2)$, and so $P(|U_X| < T_{n-2}^{-1}(1 - \alpha_0/2)) = 1 - \alpha_0$. Inverting U_x for Y gives the desired result. □

Note that this interval is **very** sensitive to deviations from Normality, and since Y is a single observation, rather than a mean, we cannot rely on the CLT to improve our approximation as n gets larger.

Model Conditions

In practice, we will rarely know for certain whether the model conditions are satisfied. It's worth reflecting on the importance and robustness of each condition:

1. **The x values are fixed**. In some designed experiments, it is reasonable to treat the values of x as fixed and controlled. But more commonly, the values of both X and Y are random, and samples are taken from the joint distribution of X and Y.

   However, this condition was essential in deriving properties of the estimators $\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2$. It turns out that for relatively large sample sizes, the distribution of estimators does not substantially change depending on whether we treat X has fixed or random. Perhaps the larger change is to our interpretation of the model and associated parameter.

   Moreover, even if X is random, treating the values of X is fixed represents an important conceptual framework: *conditioning on observed information*. Presumably, in assessing the linear model $Y = \beta_0 + \beta_1 X + \epsilon$, we are interested in modeling Y as a function of the data X.

2. **The relationship between X and Y is linear**. This is absolutely essential. If the regression function is non-linear, the estimates for regression coefficients and predicted values will be biased, and the prediction mean squared error will be inflated.

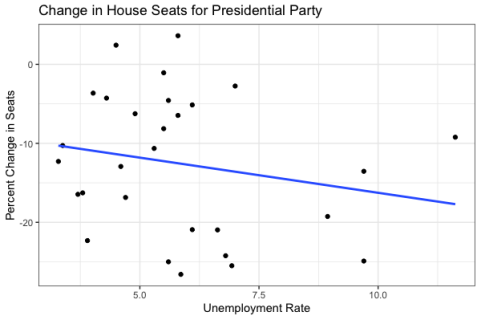
3. **The residuals are independent**. This is also essential, not only for the shape of the sampling distribution of estimators, but also their standard error. Time series data often has correlated residuals. The standard error for model coefficients are often (significantly) higher than the nominal standard error using the Linear Model.

4. **Residuals are Normally distributed**. This condition is less important. Assuming the sample size is relatively large, coefficient estimators $\hat{\beta}_0, \hat{\beta}_1$ are approximately Normally distributed (both by the CLT, as well as the asymptotic normality of MLEs), even if the residuals themselves are not Normally distributed.

R Code Demonstration

Elections for the U.S. House of Representatives occur every two years, while elections for the U.S. president occurs every 4 years. House elections in the middle of a Presidential term are called **midterm elections**. One political theory suggests that high unemployment rate corresponds to worse performance by the President's party in midterm elections. Load the following data, containing percent change in house seats for the president's party, along with unemployment rate in the year of the election. (Note that Depression Era years were excluded from this data)

```
library(openintro)
library(dplyr)
midterms_house <- midterms_house %>% filter(!year %in% c(1935, 1939))
```




```
house_mod <- lm(house_change ~ unemp, data = midterms_house)
summary(house_mod)
```

	estimate	std.error	statistic	p.value
(Intercept)	-7.364	5.155	-1.429	0.165
Slope	-0.890	0.835	-1.066	0.296