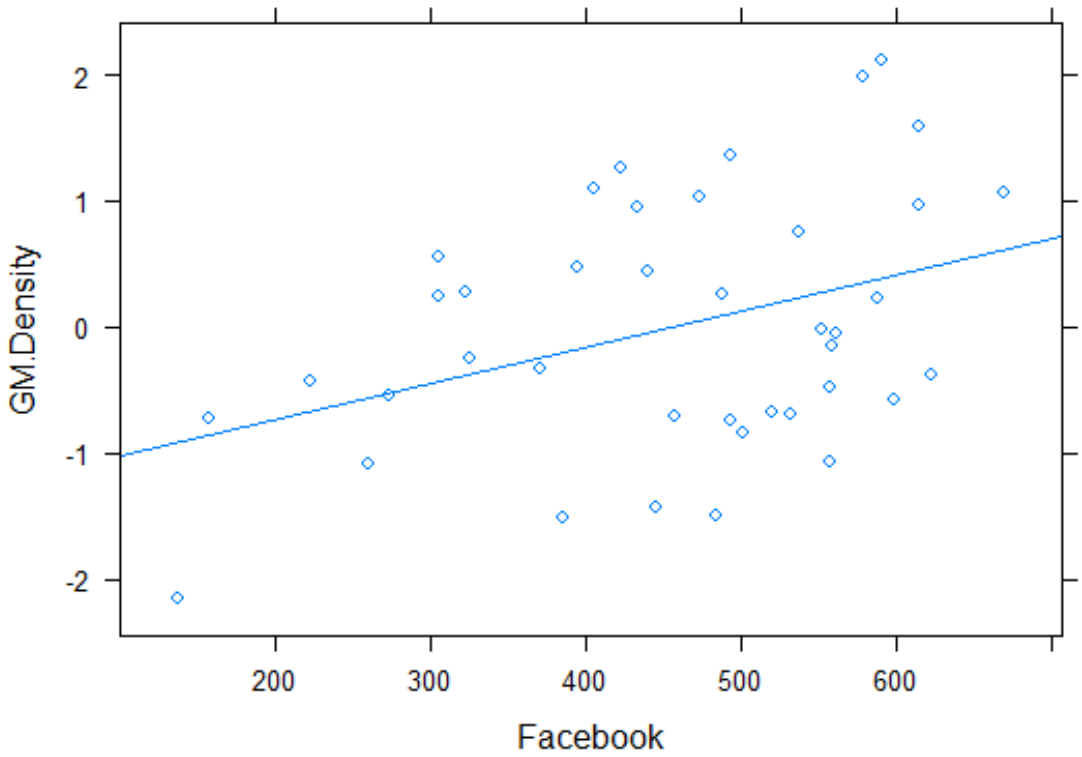
Lecture 9.3

Inference for linear regression

Grey Matter = -1.311785 + 0.002886 Friends



Is there evidence of a relationship between the number of facebook friends and grey matter density?

Inference for linear regression

We assumed $Y_i \overset{\text{indep}}{\sim} \mathcal{N}(\beta_0^\star + \beta_1^\star x_i, \sigma^2)$, i.e.,

- **Independence:** Y_i's are independent.
- **Normality:** Y_i's have normal distribution.
- **Linearity:** The mean of Y_i is a linear function of the explanatory variable x_i.
- **Common Variance:** The variance of Y_i does not depend on x or i. This is often called *homoscedasticity*. Random variables with different variances are called *heteroscedastic*.

The conclusions we make using the linear regression model are only appropriate if the assumptions are satisfied.

Inference for regression

Assuming our model assumptions hold, recall that

$$\hat{\beta}_0 \sim \mathcal{N}\left(\beta_0^\star, \sigma^2\left(\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right)\right),$$
$$\hat{\beta}_1 \sim \mathcal{N}\left(\beta_1^\star, \frac{\sigma^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right).$$

So

$$\frac{\hat{\beta}_i - \beta_i^\star}{\mathrm{sd}(\hat{\beta}_i)} \sim \mathcal{N}(0,1), \quad i = 0, 1.$$

We cannot use this statistic to construct confidence intervals or perform hypothesis testing since $\mathrm{sd}(\hat{\beta}_i)$ depends on the unknown parameter σ.

Inference for regression

Replacing σ^2 by its estimator, i.e., mean squared error (MSE),

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\Big(Y_i - [\hat{\beta}_0 + \hat{\beta}_1 x_i]\Big)^2,$$

we obtain a statistic having a t-distribution with $n-2$ degrees of freedom. That is,

$$\frac{\hat{\beta}_i - \beta_i^\star}{\mathrm{se}(\hat{\beta}_i)} \sim t_{n-2}, \quad i = 0, 1,$$

where

$$\mathrm{se}(\hat{\beta}_0) = \hat{\sigma}\sqrt{\left(\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right)},$$

$$\mathrm{se}(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}}.$$

Inference for regression

We can carry out a hypothesis test using the usual form of test statistic

$$\frac{\text{estimate} - \text{hypothesised value}}{\text{se(estimate)}}$$

We compare this to a t_{n-2} distribution to determine the p-value.

Similarly, our confidence intervals are of the form

$$\text{estimate} \pm t^* \times \text{se(estimate)},$$

where t^* is the appropriate quantile of the t_{n-2}-distribution.

Facebook friends

Q1: Is there evidence of a relationship between the number of Facebook friends and grey matter density?

$$H_0 : \beta_1 = 0, \quad \text{against} \quad H_1 : \beta_1 \neq 0$$

Q2: Give a 95% confidence interval for the slope in the linear regression model of grey matter density on the number of Facebook friends.

Predictions

Linear regression is most useful when we wish to predict how a new response variable will behave, on the basis of a new explanatory variable x.

For a given explanatory variable x, we have

$$Y \sim \mathcal{N}(\beta_0^\star + \beta_1^\star x, \sigma^2).$$

Then, $\hat{\beta}_0 + \hat{\beta}_1 x$ is a point estimator for the mean of the corresponding response, $\beta_0^\star + \beta_1^\star x$. In fact, it is an unbiased estimator since

$$\mathbb{E}\left(\hat{\beta}_0 + \hat{\beta}_1 x\right) = \mathbb{E}\left(\hat{\beta}_0\right) + \mathbb{E}\left(\hat{\beta}_1\right) x = \beta_0^\star + \beta_1^\star x.$$

How can we construct CI for this estimator?

Predictions

Let's define $\widehat{\mu}_x = \hat{\beta}_0 + \hat{\beta}_1 x$. We have

$$\begin{aligned}
\text{Var}(\widehat{\mu}_x) &= \text{Var}(\hat{\beta}_0) + x^2\text{Var}(\hat{\beta}_1) + 2x\text{Cov}(\hat{\beta}_0, \hat{\beta}_1) \\
&= \sigma^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^n (x_i - \overline{x})^2} + \frac{x^2}{\sum_{i=1}^n (x_i - \overline{x})^2} - \frac{2x\bar{x}}{\sum_{i=1}^n (x_i - \overline{x})^2} \right) \\
&= \sigma^2 \left(\frac{1}{n} + \frac{(x - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2} \right).
\end{aligned}$$

So,

$$\text{se}(\widehat{\mu}_x) = \hat{\sigma} \sqrt{\left(\frac{1}{n} + \frac{(x - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2} \right)}.$$

Predictions

Clearly, $\widehat{\mu}_x$ is normally distributed so,

$$\frac{\widehat{\mu}_x - (\beta_0^\star + \beta_1^\star x)}{\text{se}(\widehat{\mu}_x)} \sim t_{n-2}.$$

We get the $(1-\alpha)\times 100\%$ exact confidence interval for $\beta_0^\star + \beta_1^\star x$ as $\widehat{\mu}_x \pm t^\star \times \text{se}(\widehat{\mu}_x)$, i.e.,

$$\left(\hat{\beta}_0 + \hat{\beta}_1 x\right) \pm t^\star \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x-\overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}},$$

where $t^\star$ is the $(1-\alpha/2)$ quantile of the t_{n-2} distribution.

Predictions

Given an explanatory variable x, rather than calculate an interval estimate for the mean response, $\beta_0^\star + \beta_1^\star x$, we may wish to obtain an interval of the plausible response values corresponding to our model $Y \sim \mathcal{N}(\beta_0^\star + \beta_1^\star x, \sigma^2)$.

A CI refers to a parameter, or population characteristic, whose value is fixed but unknown to us. In contrast, a future value of Y is not a parameter but instead a random variable; for this reason we refer to an interval of plausible values for a future Y as a **prediction interval (PI)** rather than a confidence interval.

Here, there are two sources of uncertainty: (1) unknown mean, $\beta_0^\star + \beta_1^\star x$, and (2) unknown variance, σ^2. So, we expect the PI to be wider interval than the CI.

Predictions

Since both Y and $\hat{\beta}_0 + \hat{\beta}_1 x$ have normal distribution, are independent, and $\mathbb{E}(Y) = \beta_0^\star + \beta_1^\star x = \mathbb{E}(\hat{\beta}_0 + \hat{\beta}_1 x)$, we have

$$Y - \hat{\beta}_0 + \hat{\beta}_1 x \sim (0, \tilde{\sigma}^2).$$

We don't know $\tilde{\sigma}$ but we can estimate it. Indeed, since,

$$\begin{aligned}\tilde{\sigma}^2 &= \mathrm{Var}(Y - \hat{\beta}_0 + \hat{\beta}_1 x) = \mathrm{Var}(Y) + \mathrm{Var}(\hat{\beta}_0 + \hat{\beta}_1 x) \\ &= \sigma^2 \left(1 + \frac{1}{n} + \frac{(x - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right),\end{aligned}$$

it follows that

$$\mathrm{se}(Y - \hat{\beta}_0 + \hat{\beta}_1 x) = \hat{\sigma} \sqrt{\left(1 + \frac{1}{n} + \frac{(x - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right)}.$$

Predictions

Hence,

$$\frac{Y - (\hat{\beta}_0 + \hat{\beta}_1 x)}{\text{se}(Y - (\hat{\beta}_0 + \hat{\beta}_1 x))} \sim t_{n-2}.$$

We get the $(1-\alpha) \times 100\%$ exact **prediction interval** for Y as $\left(\hat{\beta}_0 + \hat{\beta}_1 x\right) \pm t^\star \times \text{se}\left(Y - \hat{\beta}_0 + \hat{\beta}_1 x\right)$, i.e.,

$$\left(\hat{\beta}_0 + \hat{\beta}_1 x\right) \pm t^\star \times \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x - \overline{x})^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2}},$$

where $t^\star$ is the $(1-\alpha/2)$ quantile of the t_{n-2} distribution.

Note: The "1" in the square root, makes the PI wider than the CI (though both are centered at $\hat{\beta}_0 + \hat{\beta}_1 x$). Also, as $n \to \infty$, the width of the CI goes 0, whereas the width of the PI does not.