Lecture 10

September 25, 2018

Model for Categorical Data

- A common model for when you have categorical predictors is
$$y_{jk} = \mu_j + \epsilon_{jk}$$
where $j = 1, \ldots, J$, the number of groups and $k = 1, \ldots, n_j$ the index of the observations within a group.

Model for Categorical Data

- A common model for when you have categorical predictors is
$$y_{jk} = \mu_j + \epsilon_{jk}$$
where $j = 1, \ldots, J$, the number of groups and $k = 1, \ldots, n_j$ the index of the observations within a group. This model says there is a different mean μ_j in each group j.

Model for Categorical Data

- A common model for when you have categorical predictors is

$$y_{jk} = \mu_j + \epsilon_{jk}$$

where $j = 1, \ldots, J$, the number of groups and $k = 1, \ldots, n_j$ the index of the observations within a group. This model says there is a different mean μ_j in each group j.

- We could rewrite this to look slightly more familiar,

$$\begin{aligned} y_i &= \mu_1 I(i \in 1st) + \ldots + \mu_J I(i \in Jth) + \epsilon_i \\ &= \mu_1 x_{1i} + \ldots + \mu_J x_{Ji} + \epsilon_i \end{aligned}$$

where x_{ji} is an indicator variable as to whether observation i is in the jth group.

Model for Categorical Data

- A common model for when you have categorical predictors is

$$y_{jk} = \mu_j + \epsilon_{jk}$$

where $j = 1, \dots, J$, the number of groups and $k = 1, \dots, n_j$ the index of the observations within a group. This model says there is a different mean μ_j in each group j.

- We could rewrite this to look slightly more familiar,

$$\begin{aligned} y_i &= \mu_1 I(i \in 1st) + \dots + \mu_J I(i \in Jth) + \epsilon_i \\ &= \mu_1 x_{1i} + \dots + \mu_J x_{Ji} + \epsilon_i \end{aligned}$$

where x_{ji} is an indicator variable as to whether observation i is in the jth group. We can call the predictor variable X_j a **dummy variable**, in that it gives 0/1 to whether it is in a group or not.

- We then get an $\mathbf{X}$ matrix

Why don't we get multiple model for each group?
If all groups come from the same population, then variance are the same for each group.
But if we have multiple models, then we assume each model has different vairance

$$\mathbf{X} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 1 & 0 & \cdots & 0 \\ 1 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 1 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ \vdots & & & \\ 0 & 0 & \cdots & 1 \\ 0 & 0 & \cdots & 1 \\ 0 & 0 & \cdots & 1 \\ \vdots & \vdots & & \vdots \end{pmatrix}$$

where within each group j, we have the same vector of predictors $\boldsymbol{x}_i^T$ repeated n_j times.

- Note that the $\hat{\mu}_j$ that solve the least-squares solution is just the mean of the observations in the group, which is intuitive.

```
Call:
lm(formula = coag ~ diet - 1, data = coagulation)

Residuals:
Min      1Q Median      3Q    Max
-5.00  -1.25   0.00   1.25   5.00

Coefficients:
       Estimate  Std. Error t value Pr(>|t|)
dietA  61.0000     1.1832   51.55   <2e-16 ***
dietB  66.0000     0.9661   68.32   <2e-16 ***
dietC  68.0000     0.9661   70.39   <2e-16 ***
dietD  61.0000     0.8367   72.91   <2e-16 ***
                              ---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.

Residual standard error: 2.366 on 20 degrees of freedom
Multiple R-squared: 0.9989,      Adjusted R-squared:
F-statistic:  4399 on 4 and 20 DF,  p-value: < 2.2e-16
```

```
 A  B  C  D
61 66 68 61

[1] 2.366432
```

- **Separate analysis** If all we are going to do is estimate the mean, we could ask why put them in a linear model. We could just take the mean per group, which as we've seen gives the same answer, and then calculate SE for them.
- ```
          A         B         C         D
  0.9128709 1.1547005 0.6831301 0.9258201
  ```
  ```

- **Separate analysis** If all we are going to do is estimate the mean, we could ask why put them in a linear model. We could just take the mean per group, which as we've seen gives the same answer, and then calculate SE for them.
-
```
        A         B         C         D
0.9128709 1.1547005 0.6831301 0.9258201
```
- Note that this is not the same estimate of SE that we got from the linear model. In particular, this allows each group to have a separate variance.

- **Separate analysis** If all we are going to do is estimate the mean, we could ask why put them in a linear model. We could just take the mean per group, which as we've seen gives the same answer, and then calculate SE for them.
-
```
        A         B         C         D
0.9128709 1.1547005 0.6831301 0.9258201
```
- Note that this is not the same estimate of SE that we got from the linear model. In particular, this allows each group to have a separate variance. Our linear model finds the same estimate of σ^2 for all observations. And then the variance of an estimate is given by $\hat{\sigma}^2(\mathbf{X}^T\mathbf{X})^{-1}$.

- X'X

```
       dietA dietB dietC dietD
dietA      4     0     0     0
dietB      0     6     0     0
dietC      0     0     6     0
dietD      0     0     0     8
```

- ```
  X'X
  ```
  Diagonal entries are numbers of samples in each group

  ```
        dietA dietB dietC dietD
  dietA     4     0     0     0
  dietB     0     6     0     0
  dietC     0     0     6     0
  dietD     0     0     0     8
  ```

- Then we see that our estimate of SE in both cases is $\hat{\sigma}_j/n_j$, but when we use the linear model, then we assume that σ_j is the same for all groups.
  ```

One way anova

- Consider the model

$$y_{ij} = \mu_i + e_{ij}, \quad \text{for } i = 1, \ldots, t, \quad \text{and} \quad j = 1, \ldots, n_i$$

where e_{ij} are i.i.d normal random variables with mean zero and variance σ^2. Let $\sum_{i=1}^{t} n_i = n$.

One way anova

- Consider the model

$$y_{ij} = \mu_i + e_{ij}, \quad \text{for } i = 1, \ldots, t, \quad \text{and} \quad j = 1, \ldots, n_i$$

  where e_{ij} are i.i.d normal random variables with mean zero and variance σ^2. Let $\sum_{i=1}^{t} n_i = n$.
- This model is used for the following kinds of situations:

One way anova

- Consider the model

$$y_{ij} = \mu_i + e_{ij}, \quad \text{for } i = 1, \ldots, t, \quad \text{and} \quad j = 1, \ldots, n_i$$

  where e_{ij} are i.i.d normal random variables with mean zero and variance σ^2. Let $\sum_{i=1}^{t} n_i = n$.
- This model is used for the following kinds of situations:
  1. There are t treatments and n subjects. Each subject is given one (and only one) of the j treatments.

One way anova

- Consider the model

$$y_{ij} = \mu_i + e_{ij}, \quad \text{for } i = 1, \ldots, t, \quad \text{and} \quad j = 1, \ldots, n_i$$

  where e_{ij} are i.i.d normal random variables with mean zero and variance σ^2. Let $\sum_{i=1}^{t} n_i = n$.
- This model is used for the following kinds of situations:
  1. There are t treatments and n subjects. Each subject is given one (and only one) of the j treatments. $y_{i1}, \ldots, y_{in_i}$ denote the scores of the subjects that received the ith treatment.

One way anova

- Consider the model

$$y_{ij} = \mu_i + e_{ij}, \quad \text{for } i = 1, \dots, t, \quad \text{and} \quad j = 1, \dots, n_i$$

  where e_{ij} are i.i.d normal random variables with mean zero and variance σ^2. Let $\sum_{i=1}^{t} n_i = n$.
- This model is used for the following kinds of situations:
  1. There are t treatments and n subjects. Each subject is given one (and only one) of the j treatments. $y_{i1}, \dots, y_{in_i}$ denote the scores of the subjects that received the ith treatment.
  2. We are looking at some performance of n subjects who can naturally be divided into t groups.

One way anova

- Consider the model

$$y_{ij} = \mu_i + e_{ij}, \quad \text{for } i = 1, \ldots, t, \quad \text{and} \quad j = 1, \ldots, n_i$$

  where e_{ij} are i.i.d normal random variables with mean zero and variance σ^2. Let $\sum_{i=1}^{t} n_i = n$.
- This model is used for the following kinds of situations:
  1. There are t treatments and n subjects. Each subject is given one (and only one) of the j treatments. $y_{i1}, \ldots, y_{in_i}$ denote the scores of the subjects that received the ith treatment.
  2. We are looking at some performance of n subjects who can naturally be divided into t groups. We would like to see if the performance difference between the subjects can be explained by the fact that there in these different groups.

One way anova

Consider different school as subjects, and students as treatment for each school. We want to know the performance for each school taking an Standard Test

- Consider the model

$$y_{ij} = \mu_i + e_{ij}, \quad \text{for } i = 1, \ldots, t, \quad \text{and} \quad j = 1, \ldots, n_i$$

  where e_{ij} are i.i.d normal random variables with mean zero and variance σ^2. Let $\sum_{i=1}^{t} n_i = n$.
- This model is used for the following kinds of situations:
  1. There are t treatments and n subjects. Each subject is given one (and only one) of the j treatments. $y_{i1}, \ldots, y_{in_i}$ denote the scores of the subjects that received the ith treatment.
  2. We are looking at some performance of n subjects who can naturally be divided into t groups. We would like to see if the performance difference between the subjects can be explained by the fact that there in these different groups. $y_{i1}, \ldots, y_{in_i}$ denote the performance of the subjects in the ith group.

- Often this model is also written as

tao`

$$, y_{ij} = \mu + \tau_i + e_{ij}, \quad \text{for } i = 1, \ldots, t, \text{ and } j = 1, \ldots, n_i \quad (1)$$

where μ is called the baseline score and τ_i is the difference between the average score for the ith treatment and the baseline score.

- Often this model is also written as

$$, y_{ij} = \mu + \tau_i + e_{ij}, \quad \text{for} \;\; i = 1, \ldots, t, \;\; \text{and} \, j = 1, \ldots, n_i \quad (1)$$

  where μ is called the baseline score and τ_i is the difference between the average score for the *i*th treatment and the baseline score.
- In this model, μ and the individual τ_is are not estimable.

- Often this model is also written as

$$, y_{ij} = \mu + \tau_i + e_{ij}, \quad \text{for } \; i = 1, \ldots, t, \;\; \text{and } j = 1, \ldots, n_i \quad (1)$$

  where μ is called the baseline score and τ_i is the difference between the average score for the ith treatment and the baseline score.
- In this model, μ and the individual τ_is are not estimable. It is easy to show that here a parameter $\lambda\mu + \sum_{i=1}^{t} \lambda_i \tau_i$ is estimable if and only if $\lambda = \sum_{i=1}^{t} \lambda_i$.

- Often this model is also written as

$$, y_{ij} = \mu + \tau_i + e_{ij}, \quad \text{for } \; i = 1, \ldots, t, \;\; \text{and } j = 1, \ldots, n_i \quad (1)$$

  where μ is called the baseline score and τ_i is the difference between the average score for the ith treatment and the baseline score.
- In this model, μ and the individual τ_is are not estimable. It is easy to show that here a parameter $\lambda\mu + \sum_{i=1}^{t} \lambda_i \tau_i$ is estimable if and only if $\lambda = \sum_{i=1}^{t} \lambda_i$.
- Because of this lack of estimability, people often impose the condition $\sum_{i=1}^{t} \tau_i = 0$.

- Often this model is also written as

$$, y_{ij} = \mu + \tau_i + e_{ij}, \quad \text{for } \ i = 1, \ldots, t, \ \text{ and } j = 1, \ldots, n_i \quad (1)$$

  where μ is called the baseline score and τ_i is the difference between the average score for the ith treatment and the baseline score.
- In this model, μ and the individual τ_is are not estimable. It is easy to show that here a parameter $\lambda\mu + \sum_{i=1}^{t} \lambda_i \tau_i$ is estimable if and only if $\lambda = \sum_{i=1}^{t} \lambda_i$.
- Because of this lack of estimability, people often impose the condition $\sum_{i=1}^{t} \tau_i = 0$. This condition ensures that all parameters μ and $\tau_1, \ldots, \tau_t$ are estimable.

- Often this model is also written as

$$, y_{ij} = \mu + \tau_i + e_{ij}, \quad \text{for } \ i = 1, \ldots, t, \ \text{ and } j = 1, \ldots, n_i \quad (1)$$

  where μ is called the baseline score and τ_i is the difference between the average score for the ith treatment and the baseline score.
- In this model, μ and the individual τ_is are not estimable. It is easy to show that here a parameter $\lambda\mu + \sum_{i=1}^{t} \lambda_i \tau_i$ is estimable if and only if $\lambda = \sum_{i=1}^{t} \lambda_i$.
- Because of this lack of estimability, people often impose the condition $\sum_{i=1}^{t} \tau_i = 0$. This condition ensures that all parameters μ and $\tau_1, \ldots, \tau_t$ are estimable.
- Moreover, it provides a nice interpretation. μ denotes the baseline response value and τ_i is the value by which the response value needs to be adjusted from the baseline μ for the group i. Because $\sum_i \tau_i = 0$, some adjustments will be positive and some negative but the overall adjustment averaged across all groups is zero.

- How does one test the hypothesis $H_0 : \mu_1 = \cdots = \mu_t$ in this model?

- How does one test the hypothesis $H_0 : \mu_1 = \cdots = \mu_t$ in this model?
- This is simply a linear model and we can therefore use the F-test. We just need to find the RSS in the full model (M) and the RSS in the reduced model (m).

- How does one test the hypothesis $H_0 : \mu_1 = \cdots = \mu_t$ in this model?
- This is simply a linear model and we can therefore use the F-test. We just need to find the RSS in the full model (M) and the RSS in the reduced model (m).
- What is the RSS in the full model? Let $\bar{y}_i = \sum_{j=1}^{n_i} y_{ij}/n_i$ and $\bar{y} = \sum_{i=1}^{t} \sum_{j=1}^{n_i} y_{ij}/n$.

- How does one test the hypothesis $H_0 : \mu_1 = \cdots = \mu_t$ in this model?
- This is simply a linear model and we can therefore use the F-test. We just need to find the RSS in the full model (M) and the RSS in the reduced model (m).
- What is the RSS in the full model? Let $\bar{y}_i = \sum_{j=1}^{n_i} y_{ij}/n_i$ and $\bar{y} = \sum_{i=1}^{t} \sum_{j=1}^{n_i} y_{ij}/n$.
- Write

- How does one test the hypothesis $H_0 : \mu_1 = \cdots = \mu_t$ in this model?
- This is simply a linear model and we can therefore use the F-test. We just need to find the RSS in the full model (M) and the RSS in the reduced model (m).
- What is the RSS in the full model? Let $\bar{y}_i = \sum_{j=1}^{n_i} y_{ij}/n_i$ and $\bar{y} = \sum_{i=1}^{t} \sum_{j=1}^{n_i} y_{ij}/n$.
- Write

$$\sum_{i=1}^{t} \sum_{j=1}^{n_i} \left(y_{ij} - \tilde{\mu}_i\right)^2 =$$

- How does one test the hypothesis $H_0 : \mu_1 = \cdots = \mu_t$ in this model?
- This is simply a linear model and we can therefore use the F-test. We just need to find the RSS in the full model (M) and the RSS in the reduced model (m).
- What is the RSS in the full model? Let $\bar{y}_i = \sum_{j=1}^{n_i} y_{ij}/n_i$ and $\bar{y} = \sum_{i=1}^{t} \sum_{j=1}^{n_i} y_{ij}/n$.
- Write

$$\sum_{i=1}^{t} \sum_{j=1}^{n_i} \left(y_{ij} - \tilde{\mu}_i\right)^2 = \sum_{i=1}^{t} \sum_{j=1}^{n_i} \left(y_{ij} - \bar{y}_i + \bar{y}_i - \tilde{\mu}_i\right)^2$$

$$=$$

- How does one test the hypothesis $H_0 : \mu_1 = \cdots = \mu_t$ in this model?
- This is simply a linear model and we can therefore use the F-test. We just need to find the RSS in the full model (M) and the RSS in the reduced model (m).
- What is the RSS in the full model? Let $\bar{y}_i = \sum_{j=1}^{n_i} y_{ij}/n_i$ and $\bar{y} = \sum_{i=1}^{t} \sum_{j=1}^{n_i} y_{ij}/n$.
- Write

$$
\begin{aligned}
\sum_{i=1}^{t} \sum_{j=1}^{n_i} \left(y_{ij} - \tilde{\mu}_i \right)^2 &= \sum_{i=1}^{t} \sum_{j=1}^{n_i} \left(y_{ij} - \bar{y}_i + \bar{y}_i - \tilde{\mu}_i \right)^2 \\
&= \sum_{i=1}^{t} \sum_{j=1}^{n_i} \left(y_{ij} - \bar{y}_i \right)^2 + 2 \sum_{i=1}^{t} (\bar{y}_i - \tilde{\mu}_i) \sum_{j=1}^{n_i} \left(y_{ij} - \bar{y}_i \right) + \sum_{i=1}^{t} n_i \left(\bar{y}_i - \tilde{\mu}_i \right)^2 \\
&=
\end{aligned}
$$

- How does one test the hypothesis $H_0 : \mu_1 = \cdots = \mu_t$ in this model?
- This is simply a linear model and we can therefore use the F-test. We just need to find the RSS in the full model (M) and the RSS in the reduced model (m).
- What is the RSS in the full model? Let $\bar{y}_i = \sum_{j=1}^{n_i} y_{ij}/n_i$ and $\bar{y} = \sum_{i=1}^{t} \sum_{j=1}^{n_i} y_{ij}/n$.
- Write

OLS, minimizing these expression

$$\begin{aligned}
\sum_{i=1}^{t}\sum_{j=1}^{n_i}\left(y_{ij}-\tilde{\mu}_i\right)^2 &= \sum_{i=1}^{t}\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}_i+\bar{y}_i-\tilde{\mu}_i\right)^2 \\
&= \sum_{i=1}^{t}\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}_i\right)^2 + 2\sum_{i=1}^{t}(\bar{y}_i-\tilde{\mu}_i)\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}_i\right) + \sum_{i=1}^{t} n_i\left(\bar{y}_i-\tilde{\mu}_i\right)^2 \\
&= \sum_{i=1}^{t}\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}_i\right)^2 + \sum_{i=1}^{t} n_i\left(\bar{y}_i-\tilde{\mu}_i\right)^2 .
\end{aligned}$$

this term is zero

- Therefore, the least squares estimate of μ_i is $\hat{\mu}_i = \bar{y}_i$.

- Therefore, the least squares estimate of μ_i is $\hat{\mu}_i = \bar{y}_i$. If we write μ_i as $\mu + \tau_i$ with $\sum_i \tau_i = 0$, then the least squares estimate of μ is $\bar{\bar{y}} =: (1/t) \sum_{i=1}^{t} \bar{y}_i$ and the least squares estimate of τ_i is $\bar{y}_i - \bar{\bar{y}}$.

- Therefore, the least squares estimate of μ_i is $\hat{\mu}_i = \bar{y}_i$. If we write μ_i as $\mu + \tau_i$ with $\sum_i \tau_i = 0$, then the least squares estimate of μ is $\bar{\bar{y}} =: (1/t) \sum_{i=1}^{t} \bar{y}_i$ and the least squares estimate of τ_i is $\bar{y}_i - \bar{\bar{y}}$.
- The RSS in the full model is

$$RSS(M) = \sum_{i=1}^{t} \sum_{j=1}^{n_i} \left(y_{ij} - \bar{y}_i \right)^2 .$$

- Therefore, the least squares estimate of μ_i is $\hat{\mu}_i = \bar{y}_i$. If we write μ_i as $\mu + \tau_i$ with $\sum_i \tau_i = 0$, then the least squares estimate of μ is $\bar{\bar{y}} =: (1/t) \sum_{i=1}^{t} \bar{y}_i$ and the least squares estimate of τ_i is $\bar{y}_i - \bar{\bar{y}}$.
- The RSS in the full model is

$$RSS(M) = \sum_{i=1}^{t} \sum_{j=1}^{n_i} \left(y_{ij} - \bar{y}_i\right)^2 .$$

- Check that the RSS in the reduced model is

$$RSS(m) = \sum_{i=1}^{t} \sum_{j=1}^{n_i} \left(y_{ij} - \bar{y}_i\right)^2 + \sum_{i=1}^{t} n_i \left(\bar{y}_i - \bar{y}\right)^2 .$$

- Therefore, the least squares estimate of μ_i is $\hat{\mu}_i = \bar{y}_i$. If we write μ_i as $\mu + \tau_i$ with $\sum_i \tau_i = 0$, then the least squares estimate of μ is $\bar{\bar{y}} =: (1/t) \sum_{i=1}^t \bar{y}_i$ and the least squares estimate of τ_i is $\bar{y}_i - \bar{\bar{y}}$.
- The RSS in the full model is

$$RSS(M) = \sum_{i=1}^{t} \sum_{j=1}^{n_i} \left(y_{ij} - \bar{y}_i\right)^2 .$$

- Check that the RSS in the reduced model is

$$RSS(m) = \sum_{i=1}^{t} \sum_{j=1}^{n_i} \left(y_{ij} - \bar{y}_i\right)^2 + \sum_{i=1}^{t} n_i \left(\bar{y}_i - \bar{y}\right)^2 .$$

only intercept

- Thus the F-statistic for testing $H_0 : \mu_1 = \cdots = \mu_t$ is

$$T = \frac{\sum_{i=1}^t n_i(\bar{y}_i - \bar{y})^2/(t-1)}{\sum_{i=1}^t \sum_{j=1}^{n_i}(y_{ij} - \bar{y})^2/(n-t)}$$

which has the F-distribution with $t-1$ and $n-t$ degrees of freedom under H_0.

Confidence Intervals for β_j

- Because $\hat{\beta} \sim N(\beta, \sigma^2(X^T X)^{-1})$, we have $\hat{\beta}_j \sim N(\beta_j, \sigma^2 v_j)$ where v_j is the corresponding diagonal entry of $(X^T X)^{-1}$.

Confidence Intervals for β_j

- Because $\hat{\beta} \sim N(\beta, \sigma^2(X^TX)^{-1})$, we have $\hat{\beta}_j \sim N(\beta_j, \sigma^2 v_j)$ where v_j is the corresponding diagonal entry of $(X^TX)^{-1}$.
- A $100(1-\alpha)$ % C.I for β_j is therefore given by

$$\hat{\beta}_j \pm z_{\alpha/2}\sigma\sqrt{v_j}.$$

Confidence Intervals for β_j

- Because $\hat{\beta} \sim N(\beta, \sigma^2 (X^T X)^{-1})$, we have $\hat{\beta}_j \sim N(\beta_j, \sigma^2 v_j)$ where v_j is the corresponding diagonal entry of $(X^T X)^{-1}$.
- A $100(1-\alpha)$ % C.I for β_j is therefore given by

$$\hat{\beta}_j \pm z_{\alpha/2} \sigma \sqrt{v_j}.$$

- But σ is not known, so we use the fact that

$$\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}\sqrt{v_j}} \sim t_{n-p-1}$$

to construct the following $100(1-\alpha)$ % C.I for β_j:

$$\hat{\beta}_j \pm t_{n-p-1}^{\alpha/2} \hat{\sigma} \sqrt{v_j}.$$

Confidence Intervals for β_j

- Because $\hat{\sigma}\sqrt{v_j}$ is the standard error for $\hat{\beta}_j$, we can write this C.I as

$$\hat{\beta}_j \pm t_{n-p-1}^{\alpha/2} s.e(\hat{\beta}_j).$$

If this interval contains the value 0, it means that the hypothesis $H_0 : \beta_j = 0$ will not be rejected at the α level.

Prediction Intervals

- Suppose we get a new subject whose explanatory variables are $x_{01}, \ldots, x_{0p}$. What would be our prediction for its response?

Prediction Intervals

- Suppose we get a new subject whose explanatory variables are $x_{01}, \ldots, x_{0p}$. What would be our prediction for its response?
- Our linear model says that the response for this new subject will be $y_0 = \beta_0 + \beta_1 x_{01} + \cdots + \beta_p x_{0p} + e_0$.

Prediction Intervals

- Suppose we get a new subject whose explanatory variables are $x_{01}, \ldots, x_{0p}$. What would be our prediction for its response?
- Our linear model says that the response for this new subject will be $y_0 = \beta_0 + \beta_1 x_{01} + \cdots + \beta_p x_{0p} + e_0$.
- Because β is estimated by $\hat{\beta}$ and e_0 is a zero mean error, our prediction for its response is simply $\hat{\beta}_0 + \hat{\beta}_1 x_{01} + \ldots \hat{\beta}_p x_{0p}$.

Prediction Intervals

- Suppose we get a new subject whose explanatory variables are $x_{01}, \ldots, x_{0p}$. What would be our prediction for its response?
- Our linear model says that the response for this new subject will be $y_0 = \beta_0 + \beta_1 x_{01} + \cdots + \beta_p x_{0p} + e_0$.
- Because β is estimated by $\hat{\beta}$ and e_0 is a zero mean error, our prediction for its response is simply $\hat{\beta}_0 + \hat{\beta}_1 x_{01} + \ldots \hat{\beta}_p x_{0p}$.
- What is the uncertainty in this prediction? This is captured by providing a prediction interval. There are usually two kinds of prediction intervals:

Prediction Intervals

- Suppose we get a new subject whose explanatory variables are $x_{01}, \ldots, x_{0p}$. What would be our prediction for its response?
- Our linear model says that the response for this new subject will be $y_0 = \beta_0 + \beta_1 x_{01} + \cdots + \beta_p x_{0p} + e_0$.
- Because β is estimated by $\hat{\beta}$ and e_0 is a zero mean error, our prediction for its response is simply $\hat{\beta}_0 + \hat{\beta}_1 x_{01} + \ldots \hat{\beta}_p x_{0p}$.
- What is the uncertainty in this prediction? This is captured by providing a prediction interval. There are usually two kinds of prediction intervals:
  1. Interval for the mean response

Prediction Intervals

- Suppose we get a new subject whose explanatory variables are $x_{01}, \ldots, x_{0p}$. What would be our prediction for its response?
- Our linear model says that the response for this new subject will be $y_0 = \beta_0 + \beta_1 x_{01} + \cdots + \beta_p x_{0p} + e_0$.
- Because β is estimated by $\hat{\beta}$ and e_0 is a zero mean error, our prediction for its response is simply $\hat{\beta}_0 + \hat{\beta}_1 x_{01} + \ldots \hat{\beta}_p x_{0p}$.
- What is the uncertainty in this prediction? This is captured by providing a prediction interval. There are usually two kinds of prediction intervals:
  1. Interval for the mean response without noise
  2. Interval for the response

Interval for the mean response

- The mean response is just $\beta_0 + \beta_1 x_{01} + \cdots + \beta_p x_{0p}$. So this interval is just a confidence interval for this parameter.

Interval for the mean response

- The mean response is just $\beta_0 + \beta_1 x_{01} + \cdots + \beta_p x_{0p}$. So this interval is just a confidence interval for this parameter.
- Write $x_0 := (1, x_{01}, \ldots, x_{0p})^T$ so that

$$\beta_0 + \beta_1 x_{01} + \cdots + \beta_p x_{0p} = x_0^T \beta.$$

Interval for the mean response

- The mean response is just $\beta_0 + \beta_1 x_{01} + \cdots + \beta_p x_{0p}$. So this interval is just a confidence interval for this parameter.
- Write $x_0 := (1, x_{01}, \ldots, x_{0p})^T$ so that

$$\beta_0 + \beta_1 x_{01} + \cdots + \beta_p x_{0p} = x_0^T \beta.$$

- How to find a $100(1-\alpha)\%$ C.I for $x_0^T \beta$? Observe that

$$\frac{x_0^T \hat{\beta} - x_0^T \beta}{\hat{\sigma}\sqrt{x_0^T (X^T X)^{-1} x_0}} \sim t_{n-p-1}.$$

Interval for the mean response

- The mean response is just $\beta_0 + \beta_1 x_{01} + \cdots + \beta_p x_{0p}$. So this interval is just a confidence interval for this parameter.
- Write $x_0 := (1, x_{01}, \ldots, x_{0p})^T$ so that

$$\beta_0 + \beta_1 x_{01} + \cdots + \beta_p x_{0p} = x_0^T \beta.$$

- How to find a $100(1-\alpha)\%$ C.I for $x_0^T\beta$? Observe that

$$\frac{x_0^T \hat{\beta} - x_0^T \beta}{\hat{\sigma}\sqrt{x_0^T (X^T X)^{-1} x_0}} \sim t_{n-p-1}.$$

Chi-square sigma hat

- Therefore

$$x_0^T \hat{\beta} \pm t_{n-p-1}^{(\alpha/2)} \hat{\sigma} \sqrt{x_0^T (X^T X)^{-1} x_0} \tag{2}$$

is a $100(1-\alpha)\%$ C.I for $x_0^T\beta$.

Interval for the response

- The response for the new subject with these explanatory variables is given by $y_0 = x_0^T \beta + e_0$.

Interval for the response

- The response for the new subject with these explanatory variables is given by $y_0 = x_0^T\beta + e_0$.
- It is therefore more sensible to obtain an interval for y_0 itself instead of $x_0^T\beta$.

Interval for the response

- The response for the new subject with these explanatory variables is given by $y_0 = x_0^T\beta + e_0$.
- It is therefore more sensible to obtain an interval for y_0 itself instead of $x_0^T\beta$.
- Because e_0 has mean zero, it is natural to center an interval for y_0 around $\hat{y}_0 = x_0^T\hat{\beta}$.

Interval for the response

- The response for the new subject with these explanatory variables is given by $y_0 = x_0^T \beta + e_0$.
- It is therefore more sensible to obtain an interval for y_0 itself instead of $x_0^T \beta$.
- Because e_0 has mean zero, it is natural to center an interval for y_0 around $\hat{y}_0 = x_0^T \hat{\beta}$.
- We want to find a such that

$$\mathbb{P}\left\{ y_0 \in [\hat{y}_0 - a, \hat{y}_0 + a] \right\} = \mathbb{P}\left\{ y_0 - \hat{y}_0 \in [-a, a] \right\} = 1 - \alpha.$$

Interval for the response

- The response for the new subject with these explanatory variables is given by $y_0 = x_0^T \beta + e_0$.
- It is therefore more sensible to obtain an interval for y_0 itself instead of $x_0^T \beta$.
- Because e_0 has mean zero, it is natural to center an interval for y_0 around $\hat{y}_0 = x_0^T \hat{\beta}$.
- We want to find a such that

$$\mathbb{P}\left\{y_0 \in [\hat{y}_0 - a, \hat{y}_0 + a]\right\} = \mathbb{P}\left\{y_0 - \hat{y}_0 \in [-a, a]\right\} = 1 - \alpha.$$

- For finding a, we need to look at the distribution of $y_0 - \hat{y}_0 = y_0 - x_0^T \hat{\beta}$.

Interval for the response

- The response for the new subject with these explanatory variables is given by $y_0 = x_0^T \beta + e_0$.
- It is therefore more sensible to obtain an interval for y_0 itself instead of $x_0^T \beta$.
- Because e_0 has mean zero, it is natural to center an interval for y_0 around $\hat{y}_0 = x_0^T \hat{\beta}$.
- We want to find a such that

$$\mathbb{P}\left\{y_0 \in [\hat{y}_0 - a, \hat{y}_0 + a]\right\} = \mathbb{P}\left\{y_0 - \hat{y}_0 \in [-a, a]\right\} = 1 - \alpha.$$

- For finding a, we need to look at the distribution of $y_0 - \hat{y}_0 = y_0 - x_0^T \hat{\beta}$.
- It is easy to see that $y_0 - \hat{y}_0$ has a normal distribution with mean zero and variance $\sigma^2 \left(1 + x_0^T (X^T X)^{-1} x_0\right)$. Therefore

$$\frac{y_0 - \hat{y}_0}{\hat{\sigma}\sqrt{1 + x_0^T (X^T X)^{-1} x_0}} \sim t_{n-p-1}$$

Interval for the response

- Is it obvious above that $\hat{\sigma}$ and $y_0 - \hat{y}_0$ are independent?

Interval for the response

- Is it obvious above that $\hat{\sigma}$ and $y_0 - \hat{y}_0$ are independent?
- Based on the above t-distribution, the interval

$$x_0^T \hat{\beta} \pm t_{n-p-1}^{(\alpha/2)} \hat{\sigma} \sqrt{1 + x_0^T (X^T X)^{-1} x_0} \qquad (3)$$

presents a $100(1-\alpha)\%$ interval for the future response y_0.

$$x_0^T \hat{\beta} \pm t_{n-p-1}^{(\alpha/2)} \hat{\sigma} \sqrt{x_0^T (X^T X)^{-1} x_0} \qquad (2)$$

Interval for the response

- Is it obvious above that $\hat{\sigma}$ and $y_0 - \hat{y}_0$ are independent?
- Based on the above t-distribution, the interval

$$x_0^T \hat{\beta} \pm t_{n-p-1}^{(\alpha/2)} \hat{\sigma} \sqrt{1 + x_0^T (X^T X)^{-1} x_0} \tag{3}$$

  presents a $100(1-\alpha)\%$ interval for the future response y_0.
- This is called the prediction interval for the response corresponding to the explanatory variable values $x_{01}, \ldots, x_{0p}$.

Interval for the response

- Is it obvious above that $\hat{\sigma}$ and $y_0 - \hat{y}_0$ are independent?
- Based on the above t-distribution, the interval

$$x_0^T \hat{\beta} \pm t_{n-p-1}^{(\alpha/2)} \hat{\sigma} \sqrt{1 + x_0^T (X^T X)^{-1} x_0} \tag{3}$$

  presents a $100(1-\alpha)\%$ interval for the future response y_0.
- This is called the prediction interval for the response corresponding to the explanatory variable values $x_{01}, \ldots, x_{0p}$.
- Note the difference between the intervals in (2) and (3). The interval in (3) also takes into account the randomness present in the error e_0 and is thus wider than the interval in (2).

Interval for the response

- Is it obvious above that $\hat{\sigma}$ and $y_0 - \hat{y}_0$ are independent?
- Based on the above t-distribution, the interval

$$x_0^T \hat{\beta} \pm t_{n-p-1}^{(\alpha/2)} \hat{\sigma} \sqrt{1 + x_0^T (X^T X)^{-1} x_0} \tag{3}$$

  presents a $100(1-\alpha)\%$ interval for the future response y_0.
- This is called the prediction interval for the response corresponding to the explanatory variable values $x_{01}, \ldots, x_{0p}$.
- Note the difference between the intervals in (2) and (3). The interval in (3) also takes into account the randomness present in the error e_0 and is thus wider than the interval in (2).
- The additional width in the prediction interval compared to the confidence interval accounts for the error in observing $x_0^T \beta$.

Interval for the response

- The additional width in the prediction interval compared to the confidence interval accounts for the error in observing $x_0^T\beta$.

Interval for the response

- The additional width in the prediction interval compared to the confidence interval accounts for the error in observing $x_0^T \beta$.
- The difference between the widths of the two intervals can be quite substantial when $x_0^T (X^T X)^{-1} x_0$ is small which typically happens when the sample size n is large.

Interval for the response

- The additional width in the prediction interval compared to the confidence interval accounts for the error in observing $x_0^T \beta$.
- The difference between the widths of the two intervals can be quite substantial when $x_0^T (X^T X)^{-1} x_0$ is small which typically happens when the sample size n is large.
- The prediction error of $x_0^T \hat{\beta}$ equals the sum of the estimation error of $x_0^T \beta$ (which is $x_0^T \hat{\beta} - x_0^T \beta$) and the deviation of the observation y_0 from its mean (which is $y_0 - x_0^T \beta$).

Interval for the response

- The additional width in the prediction interval compared to the confidence interval accounts for the error in observing $x_0^T \beta$.
- The difference between the widths of the two intervals can be quite substantial when $x_0^T (X^T X)^{-1} x_0$ is small which typically happens when the sample size n is large.
- The prediction error of $x_0^T \hat{\beta}$ equals the sum of the estimation error of $x_0^T \beta$ (which is $x_0^T \hat{\beta} - x_0^T \beta$) and the deviation of the observation y_0 from its mean (which is $y_0 - x_0^T \beta$).
- We can hope to reduce the estimation error by using a lot of data, but we still have to allow for the variablility in the observations while constructing the prediction interval. The latter component does not depend on n.