# Análise da amostra

**Tamanho da amostra:** 30 meses onde foram registrados dados sobre a fábrica.

**Variáveis**:

* Custos indiretos da empresa (CI).
* Produção da fábrica (ton).
* Produção de embalagens (emb).
* Férias (ferias).
* Dummy 2007 (Ano2007)
* Dummy 2008 (Ano2008).

**Comando:** *summary(dados\_fabrica$Ci)*

* Mínimo: 166.000
* Percentil 25: 190.300
* Mediana: 206.400
* Média: 211.100
* Percentil 75: 229.900
* Máximo: 268.200

# Análise do primeiro modelo de regressão múltipla

A regressão linear múltipla é composta pela seguinte estrutura:

* **Variável dependente:** ci (custos indiretos da empresa).
* **Variáveis independentes:** ton (produção da fábrica) e emb (produção de embalagens).

Ao analisar a dispersão entre a variável dependente e as variáveis independentes, é possível notar que produção de embalagens parece ter uma maior relação com os custos indiretos (conforme há um aumento na produção os custos indiretos também crescem). Para a variável de produção da fábrica, tal comportamento não aparece quando combinado com o custo indireto.

**Comandos:**

*custo\_indireto <- dados\_fabrica$Ci*

*producao\_fabrica <- dados\_fabrica$Ton*

*embalagens <- dados\_fabrica$Emb*

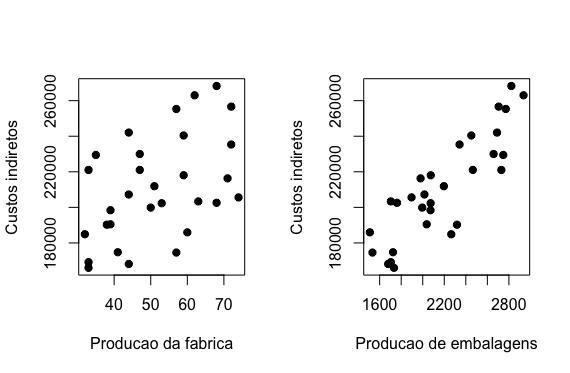
*par(mfrow=c(1,2))*

*plot(producao\_fabrica, custo\_indireto, type="n", xlab="Producao da fabrica",ylab="Custos indiretos")*

*points(producao\_fabrica, custo\_indireto, pch=19)*

*plot(embalagens, custo\_indireto, type="n", xlab="Producao de embalagens",ylab="Custos indiretos")*

*points(embalagens, custo\_indireto, pch=19)*



Ao executar a regressão, é gerada a seguinte função:

ci= 34893.14 + 986.95\*producao\_fabrica + 57.34\*embalagens

Analisando os resultados da saída da regressão é possível observar:

* **Adjusted R-squared:**  0.9475. Aproximadamente 95% da variação no custo indireto é representado pela combinação das variáveis embalagens e produção da fábrica.
* **p-value < 2.2e-16**. Como p\_value é aproximadamente 0, rejeita-se H0 ou seja há influência das variáveis independentes na variável dependente (teste do modelo).
* **Todas as variáveis demonstraram ser significativas para o modelo**.
* Para cada tonelada produzida, o custo indireto sobe 986.95.
* Para cada embalagem produzida, o custo indireto sobe 57.34.

Neste caso não será necessário aplicar a análise de multicolinearidade, pois, o modelo diz que as variáveis independentes aparentam não ter relação entre si. Ao analisa-las, as duas rejeitaram a hipótese de Bi=0.

**Comando:**

*equacao <- lm(custo\_indireto~producao\_fabrica+embalagens)*

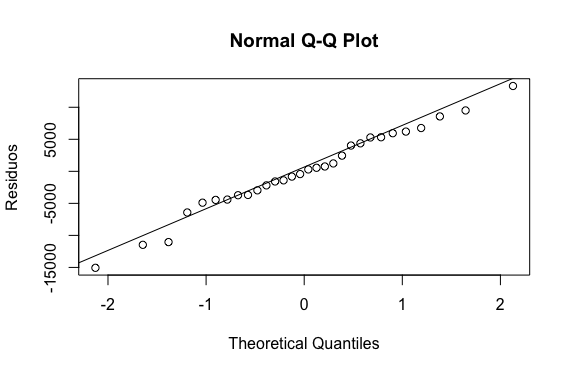
*summary(equacao)*

A partir do teste de ANOVA, foi possível provar que tanto a produção da fábrica, como a quantidade de embalagens, demonstraram ser significantes para o modelo, ou seja, exercem influência no custo indireto.

**Comando:**

*anova(equacao)*

Analisando os resíduos de forma descritiva a partir de uma reta gerada de uma distribuição normal, é possível observar que os pontos aparentam estar sobre a reta.



**Comando:**

*abline(h=0)*

*qqnorm(residuals(equacao), ylab="Residuos")*

*qqline(residuals(equacao))*

Um dos pressupostos da regressão é de que o resíduos são oriundos de uma população onde a distribuição é normal. Ao aplicar o teste de Shapiro-Wilk é possível observar que dado que p\_valor é igual a 0.9126, não se pode rejeitar a hipótese de que a amostra de resíduos venha de uma população onde a distribuição é normal.

**Comando:**

*shapiro.test(residuals(equacao))*

Outro pressuposto da regressão é de que o conjunto de resíduos referentes a cada observação X deve ter variância constante ou homogênea em toda a extensão das variáveis independentes, isto é, a dispersão de Y em relação às observações de X deve manter consistência ou ser constante em todas as dimensões da variável.

Para a realização de tal teste foram separados os resíduos a partir dos valores de mediana das variáveis independentes embalagens (2074) e producao\_fabrica (50.5) e foi realizado o teste de variância.

Foi observado que a hipótese de que os erros possuem variância constante quando separada a amostra a partir da mediana do número de embalagens e da produção, é nula, portanto, o pressuposto para a aplicação da regressão está válido.

**Comando:**

*var.test(residuals(equacao)[embalagens<=2074],residuals(equacao)*

*[embalagens>=2074])*

*var.test(residuals(equacao)[producao\_fabrica<=50.5],residuals(equacao)*

*[producao\_fabrica>=50.5])*

# Análise do segundo modelo de regressão múltipla

A regressão linear múltipla é composta pela seguinte estrutura:

* **Variável dependente:** ci (custos indiretos da empresa).
* **Variáveis independentes:** ton (produção da fábrica), emb (produção de embalagens), férias (ferias), dummy 2007 (ano2007) e dummy 2008 (ano2008).

Ao executar o modelo foi possível constatar que:

* **Adjusted R-squared:** 0.9448. A combinação entre as variáveis independentes respondem em aproximadamente 95% das variações da variável dependente.
* Nenhuma das variáveis *dummy*, além da variável férias, demonstrou representar variação no custo indireto (p\_valor maior que 5%).
* Não foi possível observar alta multicolinearidade ao realizar o teste de fator de inflação da variância nas variáveis independentes.
* Pode-se concluir que as novas variáveis incluídas no modelo não ajudam a explicar as variações no custo indireto da empresa (é possível notar que correlação de CI com as novas variáveis independentes incluídas é baixa).

**Comandos:**

*equacao\_nova<- lm(custo\_indireto~producao\_fabrica+embalagens+ferias+ano\_2007+ano\_2008)*

*summary(equacao\_nova)*

*vif(equacao\_nova)*

*sqrt(vif(equacao\_nova)) > 2*

*cor(dados\_fabrica)*

# Programa executado

*dados\_fabrica <- read.csv2("fabricachocolate.csv")*

*summary(dados\_fabrica)*

*dados\_fabrica$Ci\_residuo = (dados\_fabrica$Ci - mean(dados\_fabrica$Ci))\*(dados\_fabrica$Ci - mean(dados\_fabrica$Ci))*

*sum(dados\_fabrica$Ci\_residuo)*

*custo\_indireto <- dados\_fabrica$Ci*

*producao\_fabrica <- dados\_fabrica$Ton*

*embalagens <- dados\_fabrica$Emb*

*ferias <- factor(dados\_fabrica$ferias)*

*ano\_2008 <- factor(dados\_fabrica$Ano2008)*

*ano\_2007 <- factor(dados\_fabrica$Ano2007)*

*par(mfrow=c(1,2))*

*plot(producao\_fabrica, custo\_indireto, type="n", xlab="Producao da fabrica",ylab="Custos indiretos")*

*points(producao\_fabrica, custo\_indireto, pch=19)*

*plot(embalagens, custo\_indireto, type="n", xlab="Producao de embalagens",ylab="Custos indiretos")*

*points(embalagens, custo\_indireto, pch=19)*

*equacao <- lm(custo\_indireto~producao\_fabrica+embalagens)*

*summary(equacao)*

*confint(equacao)*

*anova(equacao)*

*summary(equacao)*

*abline(h=0)*

*qqnorm(residuals(equacao), ylab="Residuos")*

*qqline(residuals(equacao))*

*shapiro.test(residuals(equacao))*

*median(embalagens)*

*median(producao\_fabrica)*

*var.test(residuals(equacao)[embalagens<=2074],residuals(equacao)*

*[embalagens>=2074])*

*var.test(residuals(equacao)[producao\_fabrica<=50.5],residuals(equacao)*

*[producao\_fabrica>=50.5])*

*equacao\_nova <- lm(custo\_indireto~producao\_fabrica+embalagens+ferias+ano\_2007+ano\_2008)*

*summary(equacao\_nova)*

*vif(equacao\_nova)*

*sqrt(vif(equacao\_nova)) > 2*

*cor(dados\_fabrica)*