Estimating the Variance of the Error Terms

- The unbiased estimator for σ_ε^2 is

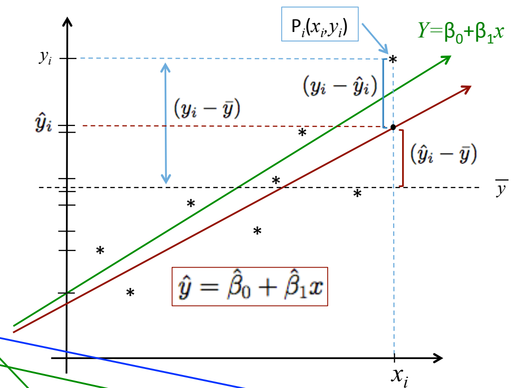
$$\hat{\sigma}_\epsilon^2 = \frac{\sum e_i^2}{n-2} = \frac{SSE}{n-2} = MSE$$

```
sse=sum(result$residuals^2)                    # 16811.16
mse=sse/(80-2)                                 # 215.5277
sigma.hat=sqrt(mse)                            # 14.68086

anova(result)
Response: Weight
          Df Sum Sq Mean Sq F value    Pr(>F)
Waist      1  79284   79284  367.86 < 2.2e-16 ***
Residuals 78  16811     216
Total     79  96095
summary(result)
lm(formula = Weight ~ Waist)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -51.7279    11.1288  -4.648 1.34e-05 ***
Waist         2.3947     0.1249  19.180  < 2e-16 ***

Residual standard error: 14.68 on 78 degrees of freedom
Multiple R-squared:  0.8251,     Adjusted R-squared:  0.8228
F-statistic:  367.9 on 1 and 78 DF,  p-value: < 2.2e-16
```

R^2 = SSR/SSTO



Note: $(y_i - \bar{y}) = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$

Surprisingly, $\sum_i (y_i - \bar{y})^2 = \sum_i (y_i - \hat{y}_i)^2 + \sum_i (\hat{y}_i - \bar{y})^2$

SSTO = SSE + SSR

Since the p-vlaue is less than 0.05, we conclude the the regression model account for a significant amount of the variability in weight.

1

Things that affect the slope estimate

- Watch the regression podcast by Dr. Will posted on our course webpage.

- Three things that affect the slope estimate:
  1. Sample size (n).
  2. Variability of the error terms (σ_ε^2).
  3. Spread of the independent variable.

Testing H_0: $\beta_1 = 0$ vs. H_1: $\beta_1 \neq 0$.

$$\frac{\hat{\beta}_1 - \beta_1}{SE_{\hat{\beta}_1}} \sim t_{(n-2)}$$

```
t.obs=(Beta1.hat-0)/SE.beta1         # 19.17971
p.value=2*(1-pt(19.18,df=78))        # virtually 0
```

```
summary(result)
lm(formula = Weight ~ Waist)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -51.7279    11.1288  -4.648 1.34e-05 ***
Waist         2.3947     0.1249  19.180  < 2e-16 ***
```

The smaller σ_ε is, the smaller the standard error of the slope estimate.

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = r\frac{s_y}{s_x}$$

$$SE_{\hat{\beta}_1} = \frac{\hat{\beta}_1\sqrt{1-r^2}}{r\sqrt{n-2}} = \frac{\hat{\sigma}_\epsilon}{\sqrt{SS_{xx}}}$$

```
MSE=anova(result)$Mean[2]           # 215.5277
SE.beta1=sqrt(MSE/SSxx)             # 0.1248554
```

```
SS=function(x,y){sum((x-mean(x))*(y-mean(y)))}
SSxy=SS(Waist,Weight)        # 33108.35
SSxx=SS(Waist,Waist)         # 13825.73
SSyy=SS(Weight,Weight)       # 96095.4 = SSTO
Beta1.hat=SSxy/SSxx          # 2.39469
```

As n increases, the standard error of the slope estimate decreases.

2

Confidence Intervals

- *The (1-α)100% C.I. for β_1:* $\hat{\beta}_1 \pm t_{(\alpha/2, df=n-2)} SE_{\hat{\beta}_1}$

Hence, the 90% C.I. for β_1 for our example is

```
Lower=Beta1.hat – qt(0.95,df=78)*SE.beta1                    # 2.186853
Upper=Beta1.hat + qt(0.95,df=78)*SE.beta1                    # 2.602528
```

```
confint(result,level=.90)
                     5 %       95 %
(Intercept) -70.253184 -33.202619
Waist         2.186853   2.602528
```

- Estimating the mean response (μ_y) at a specified value of x:

```
predict(result,newdata=data.frame(Waist=c(80,90)))
 1        2
139.8473 163.7942
```

- Confidence interval for the mean response (μ_y) at a specified value of x:

```
predict(result,newdata=data.frame(Waist=c(80,90)),interval="confidence")
       fit      lwr      upr
1 139.8473 136.0014 143.6932
2 163.7942 160.4946 167.0938
```

$$\hat{\mu}_{y|x=x^*} \pm t_{(\alpha/2, df=n-2)} \hat{\sigma}_\epsilon \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{SS_{xx}}}$$

Prediction Intervals

- Predicting the value of the response variable at a specified value of x:

```
predict(result,newdata=data.frame(Waist=c(80,90)))
 1        2
139.8473 163.7942
```

- Prediction interval for the value of new response value (y_{n+1}) at a specified value of x:

```
predict(result,newdata=data.frame(Waist=c(80,90)),interval="prediction")
       fit      lwr      upr
1 139.8473 110.3680 169.3266
2 163.7942 134.3812 193.2072
```

$$\hat{y}_{n+1} \pm t_{(\alpha/2, df=n-2)} \hat{\sigma}_\epsilon \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{SS_{xx}}}$$

```
predict(result,newdata=data.frame(Waist=c(80,90)),interval="prediction",level=.99)
       fit      lwr      upr
1 139.8473 100.7507 178.9439
2 163.7942 124.7855 202.8029
```

Note that the only difference between the prediction interval and confidence interval for the mean response is the addition of 1 inside the square root. This makes the prediction intervals wider than the confidence intervals for the mean response.

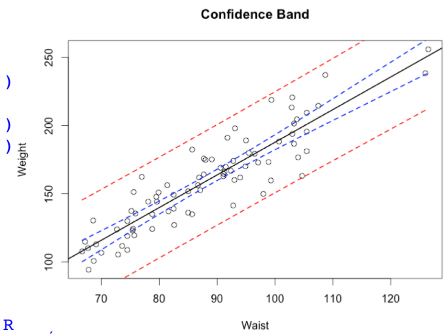
Confidence and Prediction Bands

- *Working-Hotelling (1-α)100% confidence band*: $\hat{Y}_h \pm W * se\{\hat{Y}_h\}$, $W^2 = 2F(1-\alpha, 2, n-2)$

$$se\{\hat{Y}_h\} = \hat{\sigma}_\epsilon \sqrt{\frac{1}{n} + \frac{(x_h - \bar{x})^2}{SS_{xx}}}$$

```
result=lm(Weight~Waist)
CI=predict(result,se.fit=TRUE)      # se.fit=SE(mean)
W=sqrt(2*qf(0.95,2,78))             # 2.495513
band.lower=CI$fit - W*CI$se.fit
band.upper=CI$fit + W*CI$se.fit

plot(Waist,Weight,xlab="Waist",ylab="Weight",main="Confidence Band")
abline(result)
points(sort(Waist),sort(band.lower),type="l",lwd=2,lty=2,col="Blue")
points(sort(Waist),sort(band.upper),type="l",lwd=2,lty=2,col="Blue")
```



- *The ((1-α)100% Prediction Band:*

$$se\{\hat{Y}_h\} = \hat{\sigma}_\epsilon \sqrt{1 + \frac{1}{n} + \frac{(x_h - \bar{x})^2}{SS_{xx}}}$$

```
mse=anova(result)$Mean[2]
se.pred=sqrt(CI$se.fit^2+mse)
band.lower.pred=CI$fit - W*se.pred
band.upper.pred=CI$fit + W*se.pred

points(sort(Waist),sort(band.lower.pred),type="l",lwd=2,lty=2,col="R
points(sort(Waist),sort(band.upper.pred),type="l",lwd=2,lty=2,col="Red")
```

Tests for Correlations

- *Testing H_0: $\rho = 0$ vs. H_1: $\rho \neq 0$.*

```
cor(Waist,Weight)                          # Computes the Pearson correlation coefficient, r
0.9083268
cor.test(Waist,Weight, conf.level=.99)     # Tests Ho:rho=0 and also constructs C.I. for rho
        Pearson's product-moment correlation
data:  Waist and Weight
t = 19.1797, df = 78, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
99 percent confidence interval:
 0.8409277 0.9479759
```

Note that the results are exactly the same as what we got when testing H_0: $\beta_1 = 0$ vs. H_1: $\beta_1 \neq 0$.

- *Testing H_0: $\rho = 0$ vs. H_1: $\rho \neq 0$ using the (Nonparametric) Spearman's method.*

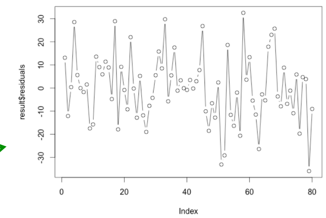
```
cor.test(Waist,Weight,method="spearman")  # Test of independence using the
        Spearman's rank correlation rho   # Spearman Rank correlation
data:  Waist and Weight
S = 8532, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
rho
0.9
```

6

Model Diagnostics

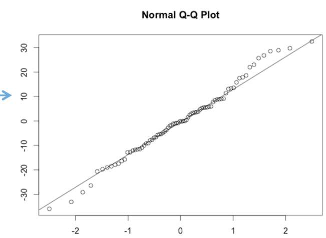
- Model: $Y_i = (\beta_0 + \beta_1 x_i) + \varepsilon_i$

  where,
  - *ε_i's are uncorrelated with a mean of 0 and constant variance σ^2_ε.*
  - *ε_i's are normally distributed.* (*This is needed in the test for the slope.*)



- Assessing uncorrelatedness of the error terms

```
plot(result$residuals,type='b')
```

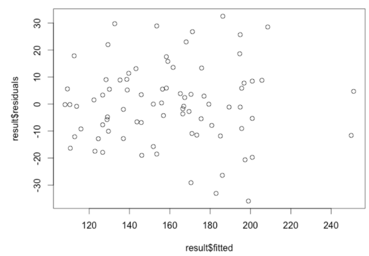


- Assessing Normality

```
qqnorm(result$residuals); qqline(result$residuals)
shapiro.test(result$residuals)
W = 0.9884, p-value = 0.6937
```

- Assessing Constant Variance

```
plot(result$fitted,result$residuals)
levene.test(result$residuals,Waist)
Test Statistic = 2.1156, p-value = 0.06764
```



7