Chapter 7- Simple Linear Regression

March 7, 2018

7.2 SLR Model

Simple Linear Regression (SLR) is a model that relates the *mean* of a single **response variable** (Y) to a single **explanatory variable** (X) through a linear relationship (think $Y = mX + b$ from algebra).

We write the SLR model as

$$\mu\{Y|X\} = \beta_0 + \beta_1 X$$

where β_0 is the *intercept* and β_1 is the *slope*.

Model Assumptions:

1. There is a normally distributed subpopulation of responses for each value of the explanatory variable (see Figure 1). (**Normality**)
2. The means of the subpopulation fall on a straight line function of the explanatory variable. (**Linearity**)
3. The subpopulation **standard deviations** are all **equal** to σ. (**Equal Variance**)
4. The selection of an observation from any of the subpopulations is independent of the selection of any other observation. (**Independence**)

There are no assumptions about the distribution of X.

This model is usefull in a large number of scenarios, and it can allow us to *interpolate* the relationship between observed values. However, it is typically not good to *extrapolate* outside of the observed data, as there is no way for us to check the assumptions of the model outside our observed data. It also might not be a good idea to interpolate through data when there is poor coverage (i.e. large gaps of data in the explanatory variable).

The parameters of this model are β_0, β_1, and σ. In this chapter, we estimate these parameters through the *method of least squares*, and then denote these *estimated* parameters as $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\sigma}$. Be sure to distinguish whether you are referring to the parameter or the estimated parameter by using the proper "hat" notation. For example, β_0 is an unknown parameter whereas $\hat{\beta}_0$ is a number calculated from data.

7.3 Least squares regression

We can write the *fitted* model or the *regression equation* as

$$\hat{\mu}\{Y|X\} = \hat{\beta}_0 + \hat{\beta}_1 X$$

Further, if we denote the ith response as Y_i, then the *fitted* value for Y_i is

$$fit_i = \hat{\mu}\{Y_i|X_i\} = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

and the residuals (the difference between the *observed*) are

$$res_i = Y_i - fit_i$$

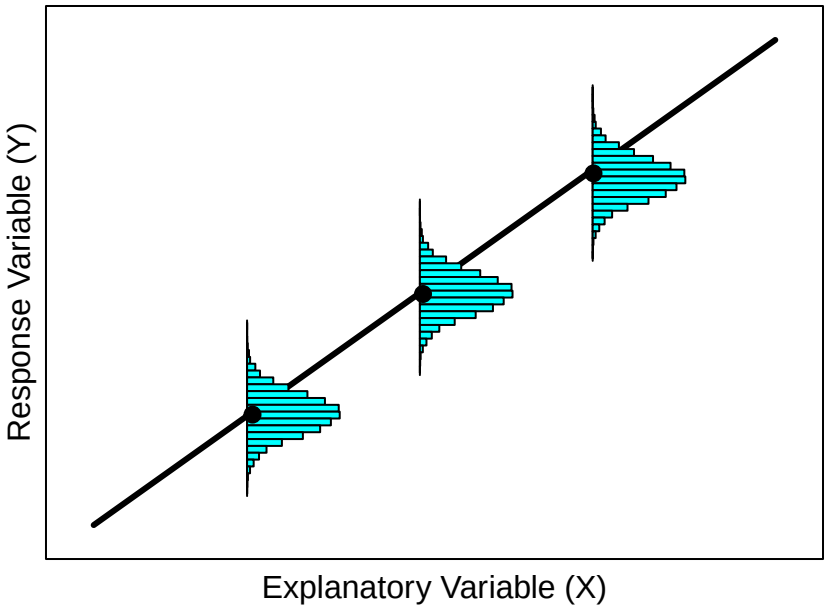


Figure 1: Normally distributed subpopulations along the regression line

An example with synthetic data

Let's generate some synthetic data to play around with using the SLR model on page 1. The fake data will look like:

$$\mu\{Y|X\} = 20 + 1.1 \times x.$$

Here's the R code that generates this synthetic data:

```
set.seed(0)
x = runif(100, 0, 30)
beta_0 <- 21
beta_1 <- 1.1
sigma <- 20
y = beta_0 + beta_1*x + rnorm(100, 0, sigma)

# Fit the SLR model using lm
lm1 <- lm(y~x)
summary(lm1)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -42.380 -15.187   0.625  11.414  50.507
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  19.8606     3.8845   5.113 1.57e-06 ***
## x             1.0516     0.2209   4.760 6.69e-06 ***
```

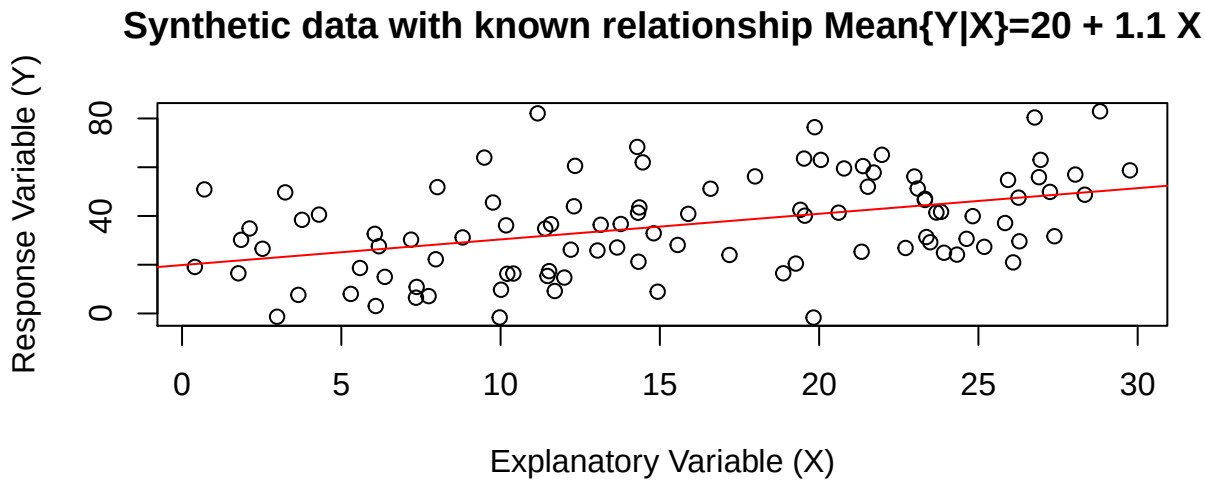
```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.82 on 98 degrees of freedom
## Multiple R-squared:  0.1878, Adjusted R-squared:  0.1795
## F-statistic: 22.65 on 1 and 98 DF,  p-value: 6.694e-06
```

The regression output shows that the estimated y-intercept $\hat{\beta}_0 = 19.9$, the estimated slope $\hat{\beta}_1 = 1.05$ and the associated standard errors. So the regression equation that predicts y for any value of x in the range from 0 to 30 is

$$\hat{\mu}\{Y|X\} = 19.9 + 1.05x.$$

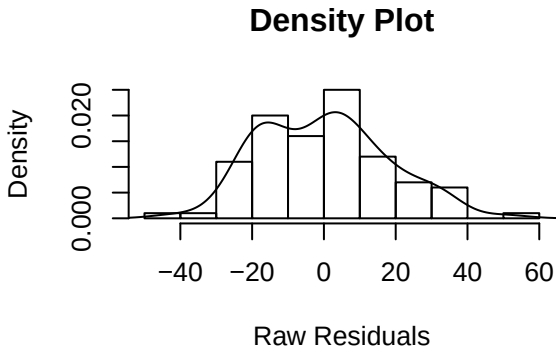
Let's plot the data using a scatterplot. The red line is the "best fit line."

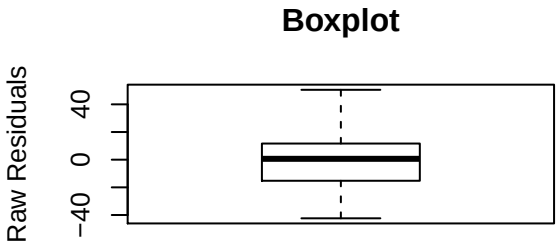
```
plot(x, y, xlab = "Explanatory Variable (X)",
     ylab = "Response Variable (Y)",main="Synthetic data with known relationship Mean{Y|X}=20 + 1.1 X")
abline(a = coef(lm1)[1], b = coef(lm1)[2], col = "red")
```

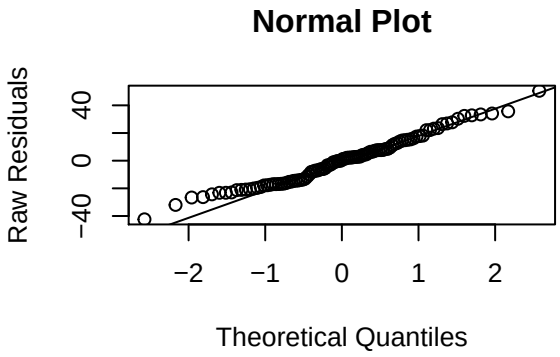


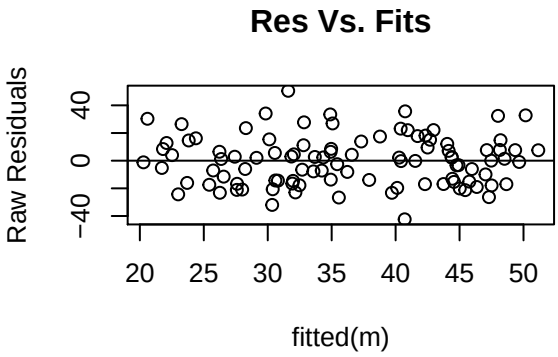
The vertical distance between the fitted line and the observed value is the residual. The assumptions of the model can be checked using the same residual plots that we learned about for ANOVA. For these synthetic data, of course they satisfy the SLE assumptions because that's how we generated it!

```
source("http://www.math.montana.edu/parker/courses/STAT411/diagANOVA.r")
diagANOVA(lm1)
```









An example with real data

The proportion of male births in the United States from 1970 to 1990, from Exercise 24 on p. 201, is shown in the next plot. The SLR model is

$$\mu\{Births|Year\} = \beta_0 + \beta_1 \times Year.$$

The least squares regression line is also shown in the plot. R code:

```
library(Sleuth3)
d = ex0724
names(d)
```

```
## [1] "Year"        "Denmark"     "Netherlands" "Canada"      "USA"
```

```
summary(d[,c(1,5)])
```

```
##       Year           USA
##  Min.   :1950   Min.   :0.5120
##  1st Qu.:1961   1st Qu.:0.5122
##  Median :1972   Median :0.5126
##  Mean   :1972   Mean   :0.5126
##  3rd Qu.:1983   3rd Qu.:0.5128
##  Max.   :1994   Max.   :0.5134
##                 NA's   :24
```

```r
dim(d)
```

```
## [1] 45  5
```

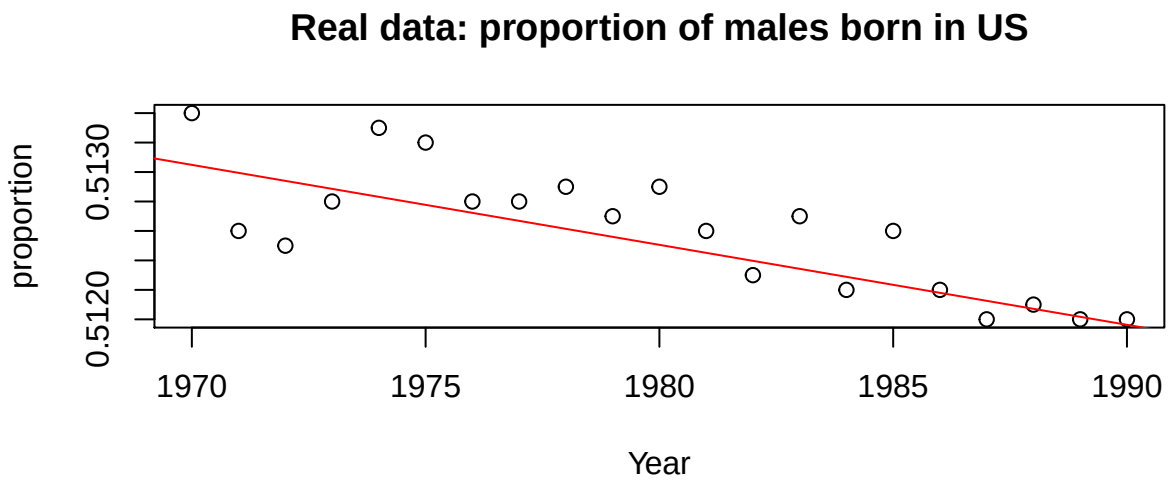
```r
dn=na.omit(d)
#summary(dn[,c(1,5)])
dim(dn)
```

```
## [1] 21  5
```

```r
plot(dn$Year,dn$USA,xlab="Year",ylab="proportion",main="Real data: proportion of males born in US")
lm2 = lm(USA ~ Year,data=dn)
summary(lm2)
```

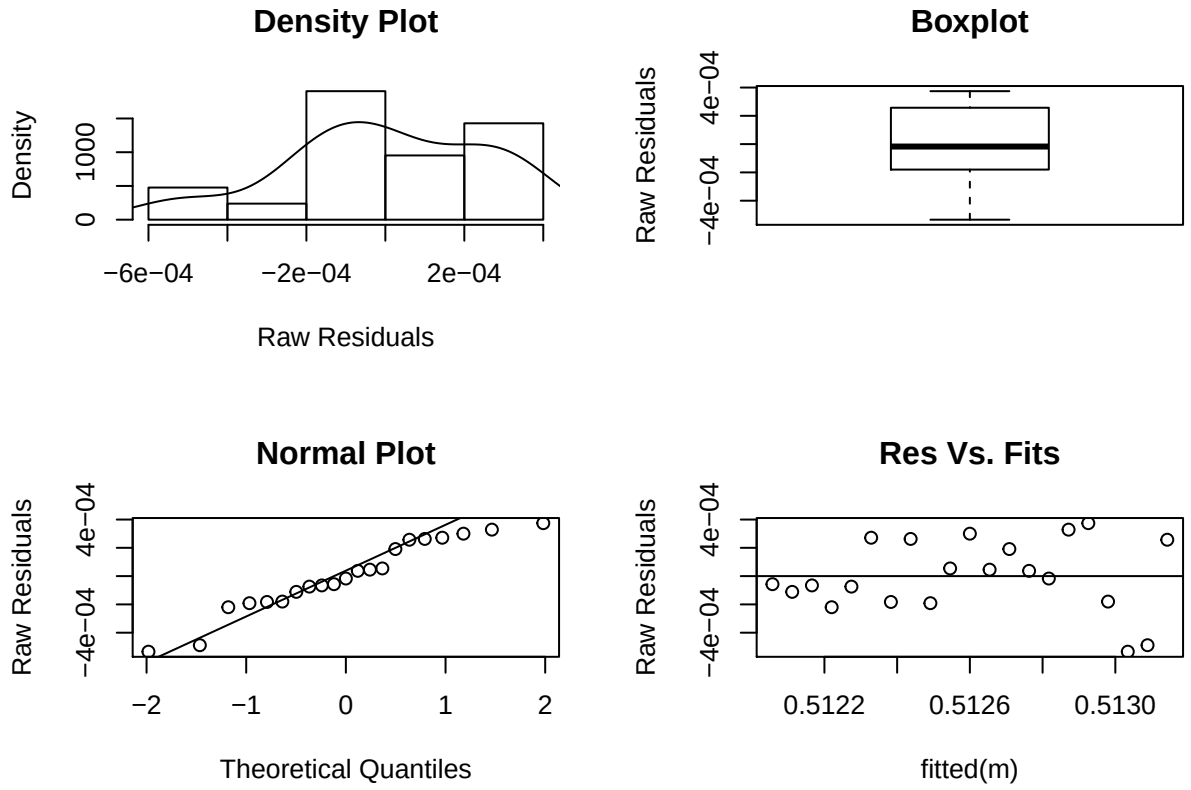
```
##
## Call:
## lm(formula = USA ~ Year, data = dn)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -5.343e-04 -1.800e-04 -1.714e-05  2.571e-04  3.743e-04
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.201e-01  1.860e-02  33.340  < 2e-16 ***
## Year        -5.429e-05  9.393e-06  -5.779 1.44e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0002607 on 19 degrees of freedom
## Multiple R-squared:  0.6374, Adjusted R-squared:  0.6183
## F-statistic:  33.4 on 1 and 19 DF,  p-value: 1.439e-05
```

```r
abline(0.62,-5.429e-05,col="red")
```



Now check the assumptions.

```
diagANOVA(lm2)
```



Results from SLR

Through some calculus and linear algebra, it can be shown that the "best fit" line (i.e., the values of β_0 and β_1 that minimize the sum of the squared residuals) is

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \\ \hat{\beta}_0 &= \bar{Y} - \hat{\beta}_1 \bar{X}. \end{aligned}$$

It can also be shown that our estimate of σ is

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n} res_i^2}{\text{Degrees of freedom}}}$$

where the degrees of freedom for the SLR model is $n - 2$. It can also be shown that in the sampling distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$, the standard deviation of of $\hat{\beta}_0$ and $\hat{\beta}_1$ are

$$
\begin{aligned}
SD(\hat{\beta}_1) &= \sigma\sqrt{\frac{1}{(n-1)s_X^2}} \\
SD(\hat{\beta}_0) &= \sigma\sqrt{\frac{1}{n} + \frac{\bar{X}^2}{(n-1)S_X^2}}
\end{aligned}
$$

We estimate these quantities by plugging our value of $\hat{\sigma}$ into these equations:

$$
\begin{aligned}
SE(\hat{\beta}_0) &= \hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{X}^2}{(n-1)S_X^2}} \\
SE(\hat{\beta}_1) &= \hat{\sigma}\sqrt{\frac{1}{(n-1)s_X^2}}
\end{aligned}
$$

7.4 Inferential tools: tests and CIs

We now have all of the tools to make inference on β_0 and β_1. Remember, it might not always make sense to conduct these tests, as it would depend on the research questions of interest. However, if we were interested, we can test these coefficients with a test statistic that has a t distribution with $n-2$ degrees of freedom:

$$
t = \frac{\hat{\beta}_i - \beta_{0i}}{SE(\hat{\beta}_i)}
$$

Where β_{0i} is the hypothesized value for β_i. Your book doesn't use this notation, but rather I just made it up so that it makes sense to talk about hypothesis testing for hypothesized values other than zero.

Summary output of the linear model will give us t-statistics for testing two hypothesis tests. The first tests whether the y-intercept is zero or not:

$$
H_0 : \beta_0 = 0 \text{ vs } H_a : \beta_0 \neq 0.
$$

These SLR hypotheses can be re-phrased in terms of the mean of Y at $X = 0$:

$$
H_0 : \mu\{Y|X = 0\} = 0 \text{ vs } H_a : \mu\{Y|X = 0\} \neq 0.
$$

The second tests whether the slope is zero or not:

$$
H_0 : \beta_1 = 0 \text{ vs } H_a : \beta_1 \neq 0.
$$

These SLR hypotheses can be re-phrased in terms of correlation:

$$
H_0 : \text{there is no correlation between } x \text{ and } y \quad \text{vs} \quad H_a : \text{there is a correlation between } x \text{ and } y.
$$

The two-sided p-values for the tests of β_i are calculated by comparing the t test statistic to a t-distribution with $n-2$ degrees of freedom.

```
#Two-sided p-values in the regression output above for the real data
2*(1-pt(33.34, 19))
```

```
## [1] 0
```

```r
2*(1-pt(abs(-5.779),19))
```

```
## [1] 1.439765e-05
```

Two-sided $100\times(1-\alpha)\%$ CIs are calculated by the following equation:

$$\hat{\beta}_i \pm t_{1-\alpha/2,df=n-2} SE(\hat{\beta}_i)$$

```r
# 95% CIs by hand
0.6201 + c(-1,1)*qt(.975,19)*0.0186
```

```
## [1] 0.5811698 0.6590302
```

```r
-5.429e-05   + c(-1,1)*qt(.975,19)*9.393e-06
```

```
## [1] -7.394977e-05 -3.463023e-05
```

```r
# 95% CIs by R's confint function
confint(lm2)
```

```
##                     2.5 %        97.5 %
## (Intercept)  5.811580e-01  6.590134e-01
## Year        -7.394606e-05 -3.462537e-05
```

Centering trick

Suppose you do not really care whether the mean of Y (i.e., the line) is different than 0 at $X = 0$, i.e. you do not care about the hypotheses $H_0 : \beta_0 = \mu\{Y|X = 0\} = 0$ vs $H_a : \beta_0 = \mu\{Y|X = 0\} \neq 0$.

For the male birth data, we certainly do not want to extrapolate from the Year 1970AD to attempt to estimate Births in the Year 0AD. Instead, it may be more interesting to estimate the proportion of male births in 1990; for (a silly) example:

$$H_0 : \beta_0 = \mu\{Births|Year = 1990\} = 0 \quad \text{vs} \quad H_a : \beta_0 = \mu\{Births|X = 1990\} \neq 0.$$

```r
# Test and CI for mean Births  at Year 1990
lm3 = lm(USA ~ I(Year-1990),data=dn)
summary(lm3)
```

```
##
## Call:
## lm(formula = USA ~ I(Year - 1990), data = dn)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -5.343e-04 -1.800e-04 -1.714e-05  2.571e-04  3.743e-04
##
## Coefficients:
##                  Estimate Std. Error  t value Pr(>|t|)
## (Intercept)     5.121e-01  1.098e-04 4663.051  < 2e-16 ***
## I(Year - 1990) -5.429e-05  9.393e-06   -5.779 1.44e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0002607 on 19 degrees of freedom
## Multiple R-squared:  0.6374, Adjusted R-squared:  0.6183
## F-statistic:  33.4 on 1 and 19 DF,  p-value: 1.439e-05
```

```
confint(lm3)
```

```
##                          2.5 %        97.5 %
## (Intercept)     5.118273e-01  5.122870e-01
## I(Year - 1990) -7.394606e-05 -3.462537e-05
```

7.4.2 Estimatng the mean response (Y) at a particular value of $X = X_0$

At some specified value (X_0) of the explanatory variable, the response variable Y has a normal distribution.

The mean of the distribution is

$$\mu\{Y|X_0\} = \beta_0 + \beta_1 X_0.$$

We will estimate the mean response using

$$\widehat{\mu}\{Y|X_0\} = \widehat{\beta}_0 + \widehat{\beta}_1 X_0$$

The Standard Error has $n-2$ degrees of freedom:

$$SE[\widehat{\mu}\{Y|X_0\}] = \widehat{\sigma}\sqrt{\frac{1}{n} + \frac{(X_0 - \overline{X})^2}{(n-1)s_x^2}}.$$

This equation shows that the uncertainty in the mean of Y gets larger for values of X_0 farther from the mean $\overline{X}$!

An individual $100(1-\alpha)\%$ t-CI for the mean value of Y at $X = X_0$ for each of a few values of X_0 is

$$\widehat{\mu}\{Y|X_0\} \pm t_{1-\alpha/2, df=n-2} SE[\widehat{\mu}\{Y|X_0\}]$$

A family of *Workman-Hotelling* CIs maintain a family-wise confidence level of $100(1-\alpha)\%$ for the mean value of Y at $X = X_0$ over as many values of X_0 as you would like

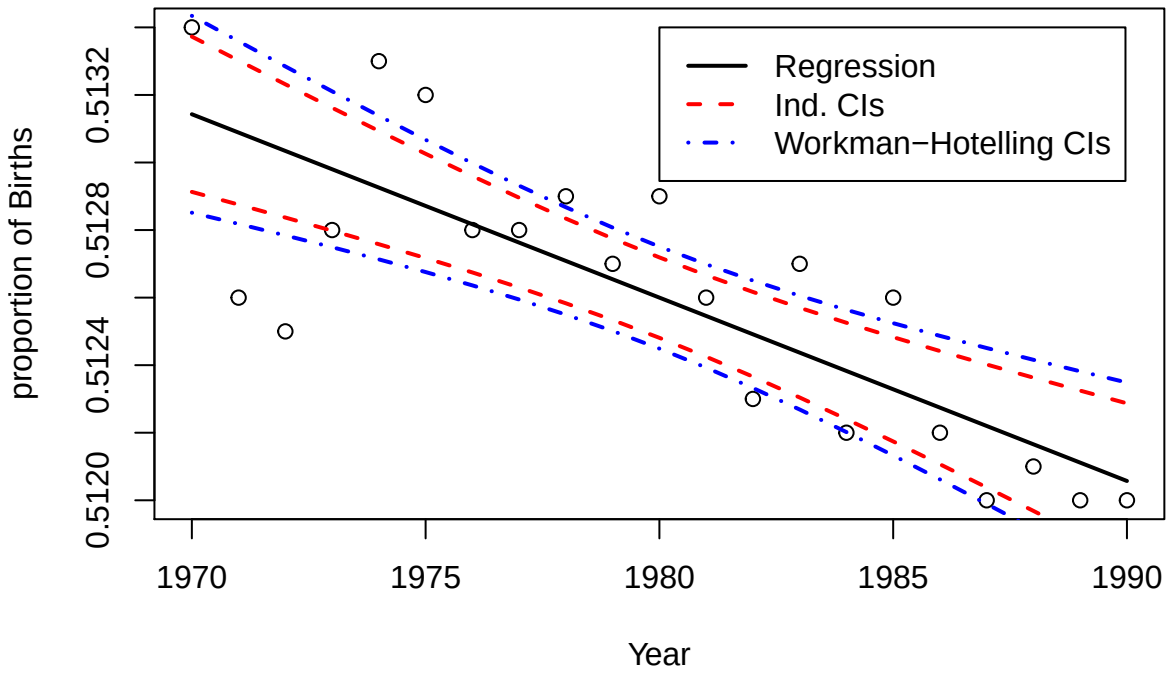
$$\widehat{\mu}\{Y|X_0\} \pm \sqrt{2F_{2,n-2}(1-\alpha)} \times SE[\widehat{\mu}\{Y|X_0\}].$$

```
# Set up a new data.frame of X0 values that we want to construct CIs for
new <- data.frame(Year = seq(1970, 1990,1))  #21 values between 0 and 30

# Get the SEs and individual CIs for the mean Y at each X0
est.mean.ses <- predict(lm2, newdata=new, se.fit=TRUE,interval="confidence")
head(data.frame(new,est.mean.ses$fit))
```

```
##   Year       fit       lwr       upr
## 1 1970 0.5131429 0.5129130 0.5133727
## 2 1971 0.5130886 0.5128753 0.5133018
## 3 1972 0.5130343 0.5128370 0.5132315
## 4 1973 0.5129800 0.5127980 0.5131620
## 5 1974 0.5129257 0.5127581 0.5130933
## 6 1975 0.5128714 0.5127170 0.5130258
```

```
conf.BAND.WH.low <- est.mean.ses$fit[,1] - sqrt(2*qf(.95,2,19))*est.mean.ses$se.fit
conf.BAND.WH.hi <- est.mean.ses$fit[,1] + sqrt(2*qf(.95,2,19))*est.mean.ses$se.fit

plot(dn$Year, dn$USA, ylab="proportion of Births", xlab="Year")    # scatterplot of data
lines(new$Year, est.mean.ses$fit[,1], lty=1, lwd=2)        # add the fitted line
lines(new$Year, est.mean.ses$fit[,2], lty=2, lwd=2, col=2) # lower ind. CL
```

```
lines(new$Year, est.mean.ses$fit[,3], lty=2, lwd=2, col=2) # upper ind. CL
lines(new$Year, conf.BAND.WH.low, lty=4, lwd=2, col=4)     # lower WH CL
lines(new$Year, conf.BAND.WH.hi, lty=4, lwd=2, col=4)      # upper WH  CL
legend(x=1980,y=.5134,legend=c("Regression","Ind. CIs","Workman-Hotelling CIs"),col=c(1,2,4),
       lty=c(1,2,4),lwd=2)
```



The red bands were drawn using individual (or pointwise) CIs. At any given value of Year (for $Year = 1970, 1971, ..., 1990$) the CI is interpreted by saying that we are 95% confident the true mean proportion of male births lies in the corresponding CI. But we are not controlling for a family of 21 CIs. Each CI individually has a confidence level of 95% but the combined level for all of them is much less than 95%.

The blue bands were drawn using Workman Hotelling CIs. These control the family-wise confidence level at 95% and hence we can be 95% confident that all 21 mean $Births$ (corresponding to $Year = 1970, 1971, ..., 1990$) lie inside these CIs.

7.4.3 Prediction of a future response

You may not want to estimate the true mean from a group of individuals. Instead, you may want to estimate a future response for a single individual. Because estimating the mean is different than estimating a future response, most statisticians say that we are *predicting* the future response as opposed to estimating it.

Before, wanted to estimate the true mean response over a group of indiviudals at a value of $X = X_0$. Now we want to predict the future response for a single individual at a value of $X = X_0$.

At some specified value (X_0) of the explanatory variable, the response variable Y has a normal distribution.

The mean of the distribution is $\mu\{Y|X_0\} = \beta_0 + \beta_1 X_0$. We will predict the response using

$$\text{Pred}\{Y|X_0\} = \widehat{\beta}_0 + \widehat{\beta}_1 X_0$$

To derive the Standard Error of $\text{Pred}\{Y|X_0\}$, note that

$$Y - \text{Pred}\{Y|X_0\} = [Y - \mu\{Y|X_0\}] - [\widehat{\mu}\{Y|X_0\} - \mu\{Y|X_0\}].$$

This equation is interpreted as

$$\text{PREDICTION ERROR} = \text{RANDOM SAMPLING ERROR} + \text{ESTIMATION ERROR}.$$

Taking the variance of both side we get that the Standard Error has $n-2$ degrees of freedom:

$$SE[\text{Pred}\{Y|X_0\}] = \widehat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(X_0 - \overline{X})^2}{(n-1)s_x^2}}.$$

This equation shows that the uncertainty in the prediction of Y gets larger for values of X_0 farther from the mean $\overline{X}$!

An individual $100(1-\alpha)\%$ t *prediction interval* (PI) for the futurevalue of Y at $X = X_0$ for each of a few values of X_0 is

$$\text{Pred}\{Y|X_0\} \pm t_{1-\alpha/2, df=n-2} SE[\text{Pred}\{Y|X_0\}]$$

A family of *Workman-Hotelling* PIs maintain a family-wise confidence level of $100(1-\alpha)\%$ for the future values of Y at $X = X_0$ over k values of X_0

$$\text{Pred}\{Y|X_0\} \pm \sqrt{kF_{k,n-2}(1-\alpha)} \times SE[\text{Pred}\{Y|X_0\}].$$

```
# Remember there are 21 Year (X0) values in the new data.frame
dim(new)
```

```
## [1] 21  1
```

```
k=21

# Get the individual PIs for the individual predicted responses Y=Births at each X0 Year
pred.ses <- predict(lm2, newdata=new, se.fit=TRUE,interval="prediction")
head(data.frame(new,pred.ses$fit))
```

```
##   Year       fit       lwr       upr
## 1 1970 0.5131429 0.5125509 0.5137348
## 2 1971 0.5130886 0.5125028 0.5136743
## 3 1972 0.5130343 0.5124542 0.5136144
## 4 1973 0.5129800 0.5124049 0.5135551
## 5 1974 0.5129257 0.5123550 0.5134964
## 6 1975 0.5128714 0.5123045 0.5134384
```

```
# Unfortunately, pred.ses$se.fit = SE[hat mu{Y|X}]
#   ie pred.ses$se.fit=est.mean.ses$se.fit
# SE[Pred{Y|X}] = sqrt(pred.ses$residual.scale^2 + pred.ses$se.fit^2)
pred.BAND.WH.low <- pred.ses$fit[,1] - sqrt(k*qf(.95,k,19)*(pred.ses$residual.scale^2
                                                            + pred.ses$se.fit^2))
pred.BAND.WH.hi <- pred.ses$fit[,1] + sqrt(k*qf(.95,k,19)*(pred.ses$residual.scale^2
                                                           + pred.ses$se.fit^2))

plot(dn$Year, dn$USA, ylab="Births", xlab="Year",ylim=c(.51,.517))    # scatterplot of data
```

```
lines(new$Year, pred.ses$fit[,1], lty=1, lwd=2)        # add the fitted line
lines(new$Year, pred.ses$fit[,2], lty=2, lwd=2, col=3) # lower ind. PL
lines(new$Year, pred.ses$fit[,3], lty=2, lwd=2, col=3) # upper ind. PL
lines(new$Year, pred.BAND.WH.low, lty=4, lwd=2, col=6) # lower WH PL
lines(new$Year, pred.BAND.WH.hi, lty=4, lwd=2, col=6)  # upper WH PL
legend(x=1980,y=.517,legend=c("Regression","Ind. PIs","Workman-Hotelling PIs"),col=c(1,3,6),
       lty=c(1,2,4),lwd=2)
```

The green bands were drawn using individual (or pointwise) PIs. At any given value of Year (for $Year = 1970, 1971, ..., 1990$) the PI is interpreted by saying that we are 95% confident the true mean proportion of Births for a single Year lies in the corresponding PI. But we are not controlling for a family of 21 PIs. Each PI individually has a confidence level of 95% but the combined level for all of them is much less than 95%.

The pink bands were drawn using Scheffe (a generalization of Workman Hotelling) PIs. These control the family-wise confidence level at 95% and hence we can be 95% confident that all 21 future values of $Births$ (corresponding to $Year = 1970, 1971, ..., 1990$) lie inside these PIs.

Comparing these PIs to the CIs generated earlier, it is clear that it is much easier to estimate the mean response $\mu\{Births|Year\}$ as opposed to predicting an individual response $Births|Year$ for a single individual.

7.5 Correlation

- The **sample correlation coefficient** describes the strength of linear relationship between any two quantitative random variables X and Y.

$$r_{XY} = \frac{\frac{1}{(n-1)} \sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{s_X s_Y}$$

- **Relationship between correlation and regression:** The above equation can be algebraically manipulated to show that
$$\hat{\beta}_1 = r_{XY} \frac{s_y}{s_x}$$
- Correlation does not depend on distinguishing between the response and explanatory variables.
- Correlation is unitless.
- $-1 \leq r_{XY} \leq 1$. $|r_{XY}|$ close to 1 suggests a strong linear relationship with the sample data falling close to the line. r_{XY} close to 0 suggests a weak linear relationship with the sample data not falling close to the line.
- For SLR, $R^2 = r_{XY}^2$. Recall that for ANOVA, $R^2 = 1 - SSF/SSR$.
- Correlation only measures the degree of linear association between two quantitative variables.
  - It is possible for there to be a non-linear relationship between X and Y and have a correlation of ZERO.
  - Hair color and gender cannot be "correlated" (although they can have an associtation)