2.5 Confidence & Prediction Intervals in MLR

In this section, we extend the discussion in the Simple Linear Regression for constructing Confidence and Prediction Intervals for New Observations.

Consider $\mathbf{x}^*$ a new observation.

The **goal** is the same as before. We want to obtain:

1. an estimator for the **mean response** at x^*.
2. a prediction for a **future** observation Y^* at $\mathbf{x}^*$, i.e. $\hat{Y}^* = (\mathbf{x}^*)^T\beta$
3. a confidence interval for μ^*.
4. a prediction interval for $\hat{Y}^*$.

2.5.1 Interval Estimation of Mean Response

For given values of $X_2, \ldots, X_p$, the *mean response*, $\mu^* = \mathbb{E}(Y|\mathbf{x}^*) = (\mathbf{x}^*)^T\beta$, is estimated by

$$\hat{\mu}^* = (\mathbf{x}^*)^T\hat{\beta} = (\mathbf{x}^*)^T(\mathbf{X^T X})^{-1}\mathbf{X^T y}$$

We can easily verify that this is an **unbiased** estimator of the mean response. Indeed,

$$\mathbb{E}(\hat{\mu}^*) = \mathbb{E}((\mathbf{x}^*)^T\hat{\beta}) = (\mathbf{x}^*)^T\mathbb{E}(\hat{\beta}) = (\mathbf{x}^*)^T\beta = \mu^*$$

We also know that the variance of this estimator computes as

$$Cov(\hat{\mu}^*) = (\mathbf{x}^*)^T Cov(\hat{\beta})\mathbf{x}^* = \sigma^2(\mathbf{x}^*)^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}^*$$

In fact, according to the Gauss-Markov theorem, this is the Best Linear Unbiased Estimate of μ^*. In addition, it can be shown that its standard error is equal to

$$se(\hat{\mu}^*) = \hat{\sigma}\sqrt{(\mathbf{x}^*)^\mathbf{T}(\mathbf{X}^\mathbf{T}\mathbf{X})^{-1}\mathbf{x}^*}$$

Therefore, using the Normality assumption of the error terms we have:

$(1-\alpha)100\%$ Confidence Interval for μ^*

$$\left(\hat{\mu}^* - T_{n-p}(\alpha/2)\ se(\hat{\mu}^*),\ \hat{\mu}^* + T_{n-p}(\alpha/2)\ se(\hat{\mu}^*)\right)$$

In `R`, we can construct directly such a confidence interval, by using the `predict` function.

2.5.2 Prediction of a New Observation

The best estimate for Y^* at a future observation $\mathbf{x}^*$ is also given by

$$\hat{y}^* = (\mathbf{x}^*)^T\hat{\beta}$$

In order to find a prediction interval (PI), we need to consider the variance due to $\hat{\beta}$ *in addition to* the variance associated with a new observation, which is σ^2.

The standard error of a prediction estimate $\hat{y}^*$ is[12]:

$$se(\hat{y}^*) = \hat{\sigma}\sqrt{1 + (\mathbf{x}^*)^\mathbf{T}(\mathbf{X}^\mathbf{T}\mathbf{X})^{-1}\mathbf{x}^*}$$

Therefore,

$(1-\alpha)100\%$ Prediction Interval for a new observation Y^* at $\mathbf{x}^*$

$$\left(\hat{y}^* - T_{n-p}(\alpha/2)\ se(\hat{y}^*),\ \hat{y}^* + T_{n-p}(\alpha/2)\ se(\hat{y}^*)\right)$$

Remark: When m new observations are to be selected at the **same levels** $\mathbf{x}^*$ and their mean $\bar{Y}^*$ is the be predicted, then the $(1-\alpha)100\%$ Prediction Interval for m new observations is:

$$\left(\hat{y}^* - T_{n-p}(\alpha/2)\ se(\bar{y}^*),\ \hat{y}^* + T_{n-p}(\alpha/2)\ se(\bar{y}^*)\right),$$

where $se(\bar{y}^*) = \hat{\sigma}\sqrt{\frac{1}{m} + \mathbf{(x^*)^T(X^TX)^{-1}x^*}}$.

Birthweight Study

In practice, we have target values for which we need to estimate the mean response of values for which we want to do prediction.

Since this is not the case here, in order to illustrate both methods, we use as inputs the mean value of every predictor. This is done as follows:

```
meanvalue = apply(birthweight2[,-2], 2, mean)
meanvalue
```

```
##      Length    Headcirc   Gestation      smoker        mage      mnocig
##  51.3333333  34.5952381  39.1904762   0.5238095  25.5476190   9.4285714
##     mheight       mppwt        fage      fedyrs      fnocig     fheight
## 164.4523810  57.5000000  28.9047619  13.6666667  17.1904762 180.5000000
##      lowbwt
##   0.1428571
```

Now, we create the X `data.frame` to use as an input in the `predict()` function:

```
x=data.frame(t(meanvalue))
```

Using `R` the 95% confidence and prediction intervals are computed as follows:

```
# 95% Confidence Interval
predict.lm(birthweight.mlr1, x, interval="confidence")
```

```
##        fit     lwr      upr
## 1 3.312857 3.20642 3.419294
```

```
# 95% Confidence Interval
predict.lm(birthweight.mlr1, x, interval="prediction")
```

```
##        fit      lwr     upr
## 1 3.312857 2.614904 4.01081
```

Observe that confidence and prediction intervals become *wider* as we move *away* from the training data.

In the next plot, we illustrate the difference between confidence and prediction intervals for the full fitted model. To be able to plot the results, we keep all the predictors fixed at the mean, and we only let the variable `Gestation` vary.

First, we create the `data.frame` that contains the data to plot:

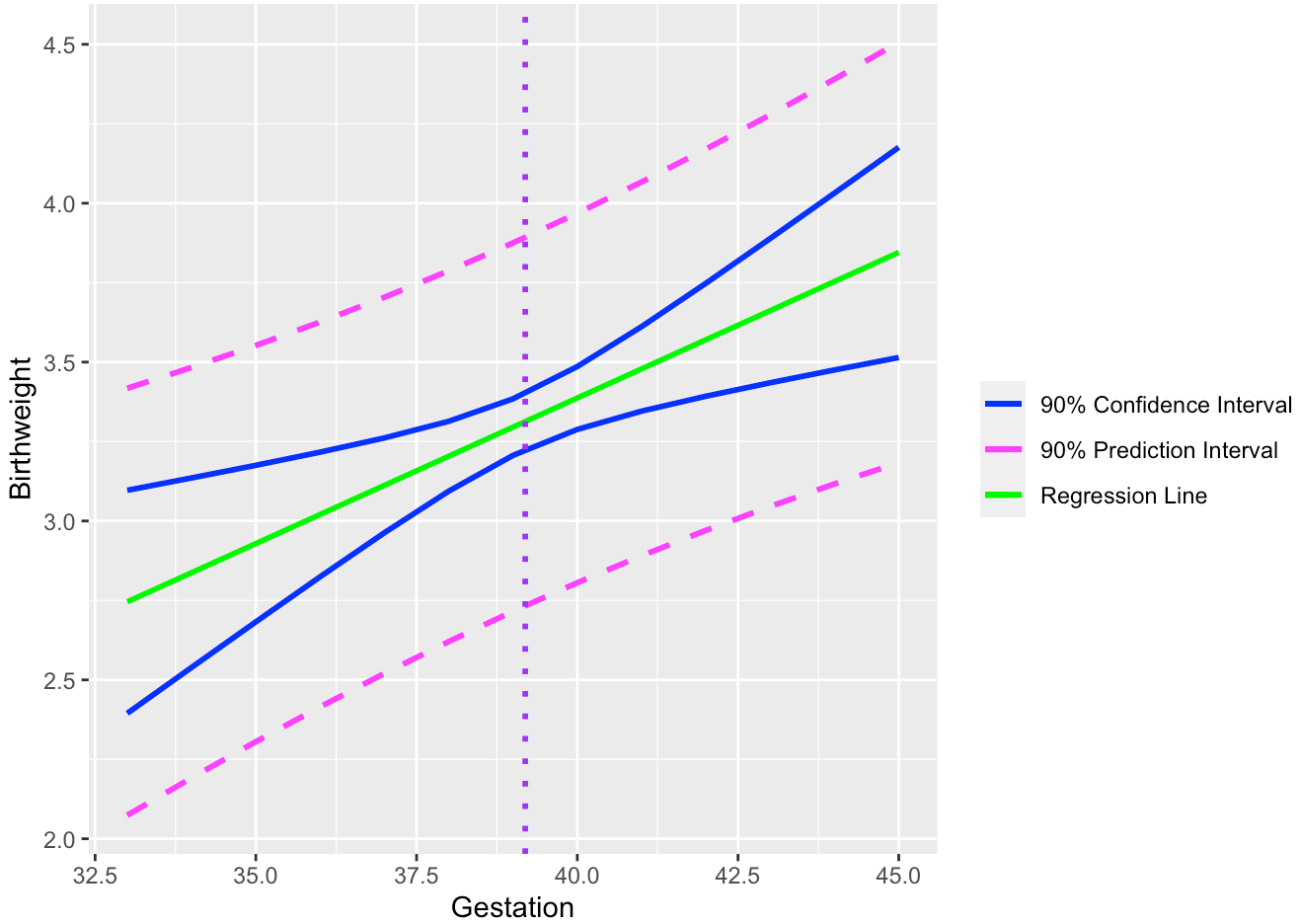
```
xnew = matrix(meanvalue[-3], 13, 12, byrow=TRUE)
xnew=data.frame(xnew)
colnames(xnew)= names(meanvalue[-3])
xnew[, 'Gestation']=33:45;
```

Then, we create the predicted objects:

```
GestCI=predict.lm(birthweight.mlr1, xnew, interval="confidence", level=0.90);
GestPI=predict.lm(birthweight.mlr1, xnew, interval="prediction", level=0.90);
```

Last, we call the `ggplot` function to create the plot:

```
ggplot(data=NULL, aes(x=33:45)) +
  geom_line(aes(y=GestCI[,1], colour="Regression Line"), size=1) +
  geom_line(aes(y=GestCI[,2], colour="90% Confidence Interval"), size=1) +
  geom_line(aes(y=GestCI[,3], colour="90% Confidence Interval"), size=1) +
  geom_line(aes(y=GestPI[,2], colour="90% Prediction Interval"), size=1, linetype
  geom_line(aes(y=GestPI[,3], colour="90% Prediction Interval"), size=1, linetype
  scale_colour_manual("", values=c("Regression Line" = "green",
                                   "90% Confidence Interval" = "blue",
                                   "90% Prediction Interval"="magenta")) +
  xlab("Gestation") +ylab("Birthweight") +
  geom_vline(xintercept = mean(birthweight2$Gestation), colour="purple", size=1,
```



Here, we did not use the default `geom_smooth` that `ggplot` has. We constructed our own point-wise confidence intervals (the blue lines). As expected the prediction intervals as much wider when compared to the confidence intervals for the same confidence level.

2.5.3 Standard Errors as a function of the Mahalanobis distance

To quantify the distance between an observation vector in $\mathbb{R}^p$ and its sample mean $\bar{\mathbf{x}}$ we can use the *Mahalanobis distance*. Let $\mathbf{x}_{p\times 1} = \begin{pmatrix} 1 \\ \mathbf{z} \end{pmatrix}$ where $\mathbf{z}$ denotes the values of the $(p-1)$ predictors (without the intercept).

We can write the sample covariance matrix of the $(p-1)$ predictor variables as:

$$\hat{\Sigma}_{(p-1)\times(p-1)} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{z}_i - \bar{\mathbf{z}})(\mathbf{z}_i - \bar{\mathbf{z}})^T$$

The following expression can be written as:

$$(\mathbf{x^*})^\mathbf{T}(\mathbf{X}^\mathbf{T}\mathbf{X})^{-\mathbf{1}}\mathbf{x^*} = \frac{1}{n} + \frac{1}{n-1}(\mathbf{z}^* - \bar{\mathbf{z}})^T\hat{\Sigma}^{-1}(\mathbf{z}^* - \bar{\mathbf{z}})$$

The term $\frac{1}{n-1}(\mathbf{z}^* - \bar{\mathbf{z}})^T\hat{\Sigma}^{-1}(\mathbf{z}^* - \bar{\mathbf{z}})$ is the so-called **Mahalanobis distance** from $\mathbf{z}^*$ to the **center** of the data $\bar{\mathbf{z}}$ (i.e. the sample mean).

Confidence vs. Prediction Standard Errors

Similar to the SimpleLinear Regression case, the point estimation and prediction at a given $\mathbf{x}^*$ are the *same*, but their standard errors are *different*. Bothe the standard errors can be exressed in terms of the Mahalanobis distance:

$$\begin{aligned} se(\hat{\mu}^*) &= \hat{\sigma}\sqrt{\mathbf{x^{*T}}(\mathbf{X}^\mathbf{T}\mathbf{X})^{-\mathbf{1}}\mathbf{x^*}} \\ &= \hat{\sigma}\sqrt{\frac{1}{n} + \frac{1}{n-1}(\mathbf{z}^* - \bar{\mathbf{z}})^T\hat{\Sigma}^{-1}(\mathbf{z}^* - \bar{\mathbf{z}})} \end{aligned}$$

$$\begin{aligned} se(\hat{y}^*) &= \hat{\sigma}\sqrt{\mathbf{1} + (\mathbf{x^*})^\mathbf{T}(\mathbf{X}^\mathbf{T}\mathbf{X})^{-\mathbf{1}}\mathbf{x^*}} \\ &= \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{1}{n-1}(\mathbf{z}^* - \bar{\mathbf{z}})^T\hat{\Sigma}^{-1}(\mathbf{z}^* - \bar{\mathbf{z}})} \end{aligned}$$

It can be shown that when the sample size n (the sample size) goes to infinity, $se(\hat{\mu}^*) \to 0$, but $se(\hat{y}^*) \to \sigma$, since $se(\hat{y}^*)$ has an extra 1.

2.5.4 Simultaneous Confidence & Prediction Intervals

In this section, we will illustrate the concepts in the Simple Linear Regression framework for simplicity in the calculations, but everything generalizes to the Multiple Linear Regression too. So, recall the SLR model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Given the values of x^*, the $(1-\alpha)100\%$ Confidence Interval for $\mu^* = \mathbb{E}(y|x^*) = \beta_0 + \beta_1 x^*$ is:

$$I(x^*) = \left(\hat{\mu}^* \pm T_{n-2}(\alpha/2)\, se(\hat{\mu}^*)\right)$$

where $\hat{\mu}^* = \hat{\beta}_0 + \hat{\beta}_1 x^*$ and

$$se(\hat{\mu}^*) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

If we want confidence intervals at **multiple points** $(x_1^*, x_2^*, \ldots, x_m^*)$, that is different x-levels, we can use this formula to construct confidence intervals at *each of the m points*, each one with significance level α:

$$I(x_1^*), I(x_2^*), \ldots, I_m(x^*)$$

This implies the point-wise coverage probability for each μ_i^* is

$$\mathbb{P}\left(\mu_i^* \in I(x_i^*)\right) = (1-\alpha)$$

and thus all the CIs above are **point-wise** confidence intervals.

But, what about the simultaneous coverage probability? i.e.:

$$\mathbb{P}\left(\mu_i^* \in I(x_i^*),\ \text{for}\ \ i = 1, \ldots, m\right) = ?$$

To make sure that (for example):

$$\mathbb{P}\left(\mu_i^* \in I(x_i^*),\ \text{for}\ \ i = 1, \ldots, m\right) = .95$$

we need to set $\alpha = 5\%/m$, which is known as the **Bonferroni Correction**.

Intuition why this works:

Let A_k denotes the event that the k-th confidence interval covers μ_k^* with:

$$\mathbb{P}(\mathrm{A}_k) = (1-\alpha).$$

Then we can show:

$$\begin{aligned}
&\mathbb{P}(\text{ All CIs cover the corresponding } \quad \mu_k^* \quad \text{values}) \\
&= \mathbb{P}\left(\mathrm{A}_1 \cap \mathrm{A}_2 \ldots \cap \mathrm{A}_m\right) \\
&= 1 - \mathbb{P}(\mathrm{A}_1^c \cup \mathrm{A}_2^c \ldots \cup \mathrm{A}_m^c) \\
&\geq 1 - \mathbb{P}(\mathrm{A}_1^c) - \ldots - \mathbb{P}(\mathrm{A}_m^c) \\
&= 1 - m\alpha
\end{aligned}$$

If we choose α/m instead of α, the simultaneous coverage probability will be $(1-\alpha)$.

■

Confidence Band for the Regression Line

Ideally we would like to construct a simultaneous confidence band (i.e., $m = \infty$) across **all** x^* 's. (Scheff'e's Theorem - 1959). Let

$$I(x) = (\hat{r}(x) - c\,\hat{\sigma},\, \hat{r}(x) + c\,\hat{\sigma})$$

where

$$\hat{r}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$$

$$c\,\hat{\sigma} = \sqrt{2\ F(\alpha, 2, n-2)}\hat{\sigma}\ \sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

Then,

$$\mathbb{P}\left(r(x) \in I(x) \text{ for all } x\right) \geq 1 - \alpha$$

Can we construct a simultaneous prediction band? No!

Confidence Band vs. Pointwise Confidence Intervals

Are confidence bands always wider than point-wise confidence intervals?

Let's answer this question in the case of a Simple Linear Regression. For SLR, at a location x^*, we have

$$\begin{aligned} \text{band} &: \hat{\mu}^* \pm \sqrt{2F(\alpha, 2, n-2)}\, se(\hat{\mu}^*) \\ \text{interval} &: \hat{\mu}^* \pm T_{n-2}(\alpha/2)\, se(\hat{\mu}^*) \end{aligned}$$

Assume $\alpha = 5\%$, then you can check that

$$\sqrt{2F(\alpha, 2, n-2)} \text{ or } T_{n-2}(\alpha/2) = \sqrt{2F(\alpha, 1, n-2)}$$

In fact, for any α, we have

$$T_m(\alpha/2)\sqrt{F(\alpha, 1, m)} < \sqrt{k\, F(\alpha, k, m)}$$

Birthweight Example

The confidence band can only be plotted in the SLR case (since multidimensional visualizations are not feasible...) So, for illustration purposes, we fit a SLR model between `Birthweight` and `Headcirc`:

```
birthweight.slr2 = lm(Birthweight~Headcirc, data=birthweight2)
summary(birthweight.slr2)
```

```
##
## Call:
## lm(formula = Birthweight ~ Headcirc, data = birthweight2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.87259 -0.28101 -0.04531  0.24732  1.33969
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -2.6472     1.0057  -2.632    0.012 *
## Headcirc      0.1723     0.0290   5.940 5.73e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4456 on 40 degrees of freedom
## Multiple R-squared:  0.4687,   Adjusted R-squared:  0.4554
## F-statistic: 35.29 on 1 and 40 DF,  p-value: 5.735e-07
```

Now, we construct all the elements of the formula above one-by-one:

- The $\hat{r}$ term is

```
r_hat = summary(birthweight.slr2)$coefficients[1, 1] + summary(birthweight.slr2)$
```

- Term c computes as:

```
n=dim(birthweight2)[1]
c = sqrt(2 * qf(.05, 2, n - 2, lower.tail = FALSE))
```

- $\hat{\sigma}$ is obtained from the formula above:

```
sigma_hat = summary(birthweight.slr2)$sigma * sqrt((1/n) +((birthweight2[,3]
```

We combine the above the estimator, lower and upper bounds in a data frame:

```
confidence.band = cbind(r_hat, lower = r_hat - c*sigma_hat, upper = r_hat + c*sig
head(confidence.band)
```

```
##         r_hat    lower    upper
## [1,] 3.210310 3.030061 3.390559
## [2,] 3.554869 3.351666 3.758073
## [3,] 4.071709 3.702870 4.440549
## [4,] 3.899429 3.593511 4.205347
## [5,] 3.727149 3.478147 3.976152
## [6,] 3.899429 3.593511 4.205347
```

From theory, we know that: For SLR, at a location x^*, we have

$$\begin{aligned} \text{band} &: \hat{\mu}^* \pm \sqrt{2F(\alpha, 2, n-2)}\, se(\hat{\mu}^*) \\ \text{interval} &: \hat{\mu}^* \pm T_{n-2}(\alpha/2)\, se(\hat{\mu}^*) \end{aligned}$$

In the following graph, we are going to plot: (a) The raw data related to the SLR model fitted above, (b) the fitted SLR regression line, (c) 95% point-wise confidence intervals, (d) the 95% confidence band calculated above.

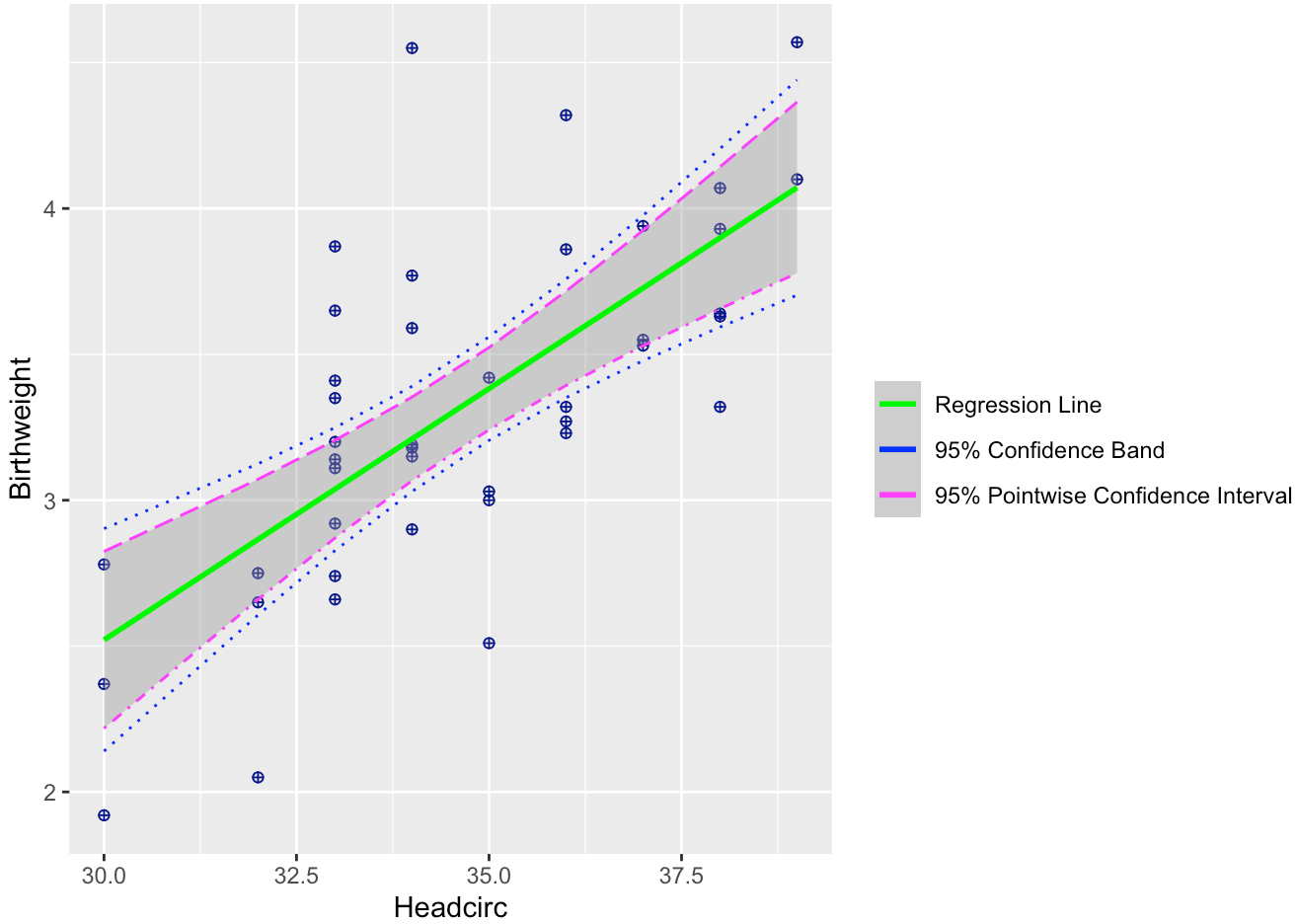
The pointwise intervals are given by:

```
t = qt(.025, n-2)
pointwise.interval = cbind(r_hat, lower = r_hat + t*sigma_hat, upper = r_hat - t*
head(pointwise.interval)
```

```
##         r_hat    lower    upper
## [1,] 3.210310 3.067017 3.353602
## [2,] 3.554869 3.393329 3.716410
## [3,] 4.071709 3.778493 4.364925
## [4,] 3.899429 3.656234 4.142625
## [5,] 3.727149 3.529200 3.925099
## [6,] 3.899429 3.656234 4.142625
```

Finally, we plot together: the raw data, regression line (green), confidence band (blue), point-wise intervals (magenta):

```
library(ggplot2)
bandplot = ggplot(data = birthweight2, aes(x=Headcirc, y = Birthweight)) +
geom_point(col = "darkblue", pch = 10) +
geom_smooth(method = "lm", formula = y~x, aes(colour="A")) +
geom_line(aes(y=confidence.band[,2], colour = "confidence.band[,2]"),  linetype=3
geom_line(aes(y=confidence.band[,3]),  colour = "blue", linetype=3) +
geom_line(aes(y=pointwise.interval[,2], colour = "pointwise.interval[,2]"),   lir
geom_line(aes(y=pointwise.interval[,3]), colour = "magenta",   linetype=5) +
scale_colour_manual(name="", values = c("green", "blue", "magenta"), labels=c("Re
plot(bandplot)
```



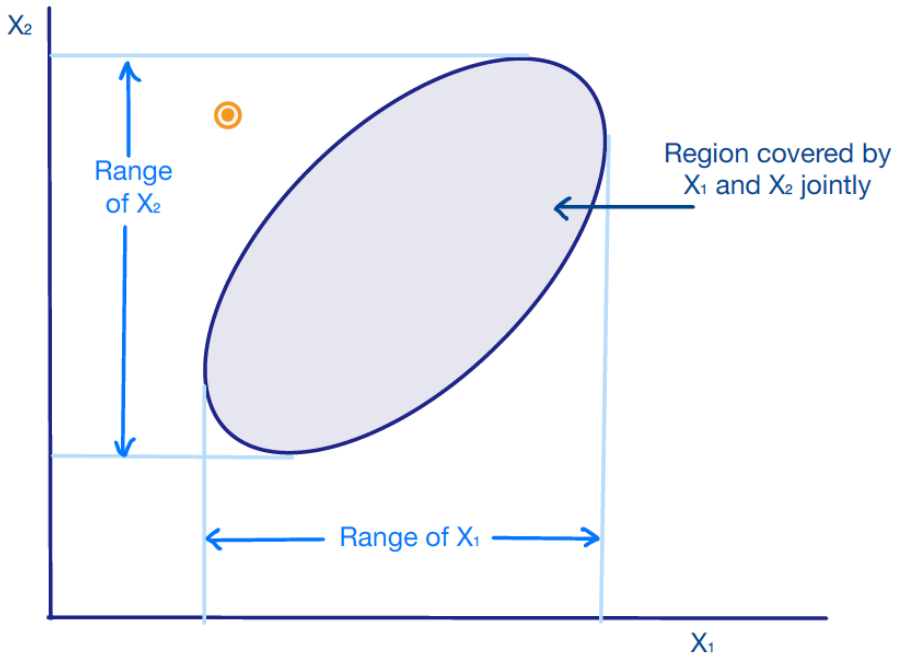
As we expected, the confidence band is slightly wider than the point-wise intervals.

2.5.5 Hidden Extrapolations

When estimating a mean response or predicting a new observation in multiple regression, one needs to be particularly careful that the estimate or prediction does not fall outside the scope of the model. The danger is that the model may not be appropriate when it is extended outside the region of the observations.

In multiple linear regression, it is particularly easy to lose track of this region since the levels of $X_1, \ldots, X_p$ **jointly** define the region, which means that one cannot merely look at the ranges of ech predictor variable.

For example, consider the following scenario:



Here, the shaded region is the region of observations for a multiple regression application with two predictor variables and the circled dot represents the values (x_1^*, x_2^*) for which a prediction is to be made. The circled dot is within the range of the predictor variables individually, yet it is well outside the joint region if observations. This is still easy to be spotted when $p = 3$, but it becomes much more difficult when the number of predictors is large. This is something that we will discuss later in the course.