SF2930 - Regression analysis

KTH Royal Institute of Technology, Stockholm

Lecture 2 – Simple linear regression: inference and prediction (MPV 2.3, 2.4)

February 14, 2022

Todays lecture

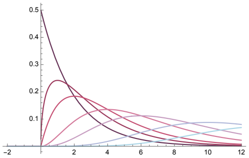
- A very short reminder on the relationship between the normal distribution, χ^2-distributions, F-distributions, and t-distributions.
- Confidence intervals and test of significance for the slope, intercept, and variance of the error term
- ANOVA (ANalysis-Of-VAriance)
- Confidence interval for mean response and prediction interval for future observations

Recall...

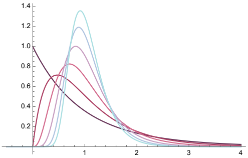
- $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$
- $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$
- $e_i = y_i - \hat{y}_i$
- n – the number of data points
- $\bar{x}$ – the mean of $x_1, x_2, \ldots, x_n$
- Least squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ minimize $f(\beta_0, \beta_1) = \sum \big(y_i - \hat{y}_i(\beta_0, \beta_1)\big)^2$

Useful distributions

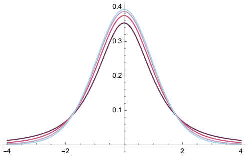
- If $X_1, X_2, \dots, X_n \sim N(0,1)$ are independent, then $\sum X_i^2 \sim \chi_n^2$ has a χ^2 distribution with n degrees of freedom.
- If $X_1 \sim \chi^2_{df_1}$ and $X_2 \sim \chi^2_{df_2}$ are independent, then $\frac{X_1/df_1}{X_2/df_2} \sim F_{d_1,d_2}$ has a F distribution with df_1 and df_2 degrees of freedom.
- If $X_1 \sim N(0,1)$ and $X_2 \sim \chi^2_{df}$ are independent, then $\frac{X_1}{\sqrt{X_2/df}} \sim t_{df}$ has a t-distribution with df degrees of freedom.



χ^2-distributions



F-distributions



t-distributions

Useful assumptions

In order to say someting about the distribution of the things we estimate from the data, we need to make additional assumptions on the error terms (ε_i). In the last lecture, we assumed that they satisifed

- $\mathbb{E}[\varepsilon_i] = 0$
- $\mathrm{Var}(\varepsilon_i) = \sigma^2$
- ε_i and ε_j are independent if $i \neq j$.

In this lecture we, in addition, assume that

$$\varepsilon_i \sim N(0, \sigma^2).$$

Note that in general, when we use the theory developed today, we should argue that this is indeed likely to hold.

The distribution of $\hat{\beta}_1$

Recall that

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum y_i(x_i - \bar{x})}{\sum (x_i - \bar{x})^2},$$

Using the assumptions on the previous slide, it follows that

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2).$$

Combining these equations, we obtain

$$\hat{\beta}_1 = \frac{\sum y_i(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} \sim \sum N\left(\frac{(\beta_0 + \beta_1 x_i)(x_i - \bar{x})}{\sum (x_i - \bar{x})^2}, \frac{\sigma^2 (x_i - \bar{x})^2}{\left(\sum (x_i - \bar{x})^2\right)^2} \right).$$

The distribution of $\hat{\beta}_1$

Since the responses y_1, y_2, ..., y_n are independent, we have

$$\hat{\beta}_1 \sim \sum N\left(\frac{(\beta_0 + \beta_1 x_i)(x_i - \bar{x})}{\sum(x_i - \bar{x})^2}, \frac{\sigma^2 (x_i - \bar{x})^2}{\left(\sum(x_i - \bar{x})^2\right)^2}\right)$$
$$= N\Bigg(\underbrace{\sum \frac{(\beta_0 + \beta_1 x_i)(x_i - \bar{x})}{\sum(x_i - \bar{x})^2}}_{\mathbb{E}[\hat{\beta}_1]}, \underbrace{\sum \frac{\sigma^2 (x_i - \bar{x})^2}{\left(\sum(x_i - \bar{x})^2\right)^2}}_{\mathrm{Var}(\hat{\beta}_1)}\Bigg).$$

Hence $\hat{\beta}_1$ has a normal distribution. Using the expressions for the mean and variance of $\hat{\beta}_1$ from the last lecture, we deduce that

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right).$$

Test of significance for the slope $\hat{\beta}_1$

Hypothesis

Assume we want to test the hypothesis $\beta_1 = \beta_{10}$, i.e.,

$$H_0\colon \beta_1 = \beta_{10}, \qquad H_1\colon \beta_1 \neq \beta_{10}.$$

Test statistic

Since $\hat{\beta}_1 \sim N\big(\beta_1, \frac{\sigma^2}{S_{xx}}\big)$, if H_0 is true, then

$$Z_0 := \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{\sigma^2 / S_{xx}}} \sim N(0,1).$$

Since σ^2 is not known, we replace σ^2 with the estimate

$$\hat{\sigma}^2 = MS_{Res} = SS_{Res}/(n-2) = \sum (y_i - \hat{y}_i)^2/(n-2).$$

However, this changes the distribution of Z_0. In detail (see MPV C.3.2),

$$(n-2)MS_{Res}/\sigma^2 \sim \chi^2_{n-2}.$$

Consequently, if H_0 is true, then

$$t_0 := \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{MS_{Res} \cdot \underbrace{(1/S_{xx})}_{\mathrm{Var}(\hat{\beta}_1)/\sigma^2}}} \sim t_{n-2}$$

$\rightarrow$ Reject H_0 with confidence level α if $|t_0| \geq t_{\alpha, n-2}$

Confidence interval for $\hat{\beta}_1$

From the previous slide, we know that if the true slope is β_1, then

$$t_0 := \frac{\hat{\beta}_1 - \beta_1}{\sqrt{MS_{Res}/S_{xx}}} \sim t_{n-2}$$

Knowing this distribution, we can calculate a confidence interval for β_1. Since

$$\begin{aligned}\alpha/2 = P(t_0 \geq t_{\alpha/2,n-2}) &= P\left(\frac{\hat{\beta}_1 - \beta_1}{\sqrt{MS_{Res}/S_{xx}}} \geq t_{\alpha/2,n-2}\right) \\ &= P\big(\hat{\beta}_1 - t_{\alpha/2,n-2}\sqrt{MS_{Res}/S_{xx}} \geq \beta_1\big)\end{aligned}$$

a $100(1-\alpha)$-percent confidence interval for $\hat{\beta}_1$ is given by

$$\hat{\beta}_1 \pm t_{\alpha/2,n-2}\sqrt{MS_{Res} \cdot \underbrace{(1/S_{xx})}_{\mathrm{Var}(\hat{\beta}_1)/\sigma^2}}$$

Test of significance for intercept $\hat{\beta}_0$

Assume we want to test the hypothesis $\beta_0 = \beta_{00}$, i.e.,

$$H_0\colon \beta_0 = \beta_{00}, \qquad H_1\colon \beta_0 \neq \beta_{00}.$$

In this case, a similar analysis shows that

$$t_0 := \frac{\hat{\beta}_0 - \beta_{00}}{\sqrt{MS_{Res}\underbrace{(1/n + \bar{x}^2/S_{xx})}_{\mathrm{Var}(\hat{\beta}_0)/\sigma^2}}} \sim t_{n-2}.$$

$\rightarrow$ Reject H_0 with confidence level α if $|t_0| \geq t_{\alpha,n-2}$

Confidence interval for $\hat{\beta}_0$

From the previous slide, we know that if the true intercept is β_0, then

$$t_0 \coloneqq \frac{\hat{\beta}_0 - \beta_0}{\sqrt{MS_{Res}(1/n + \bar{x}/S_{xx})}} \sim t_{n-2}$$

Consequently, a $100(1-\alpha)$-percent confidence interval for $\hat{\beta}_0$ is given by

$$\hat{\beta}_0 \pm t_{\alpha/2,n-2}\sqrt{MS_{Res}\underbrace{(1/n + \bar{x}^2/S_{xx})}_{\mathrm{Var}(\hat{\beta}_0)/\sigma^2}}.$$

Confidence interval for σ^2

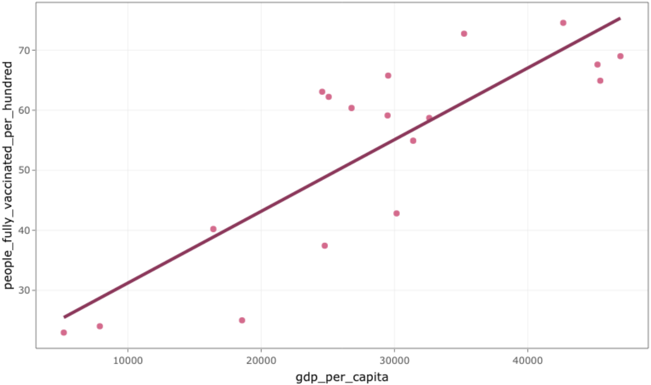
On the previous slides, we used that $SS_{Res} = (n-2)MS_{Res}/\sigma^2 \sim \chi^2_{n-2}$.
Using this observation directly, we have

$$P(\chi^2_{1-\alpha/2,n-2} \leq (n-2)MS_{Res}/\sigma^2 \leq \chi^2_{\alpha/2,n-2}) = \alpha,$$

and thus a $100(1-\alpha)$ percent CI for σ^2 is given by

$$\frac{(n-2)MS_{Res}}{\chi^2_{\alpha/2,n-2}} \leq \sigma^2 \leq \frac{(n-2)MS_{Res}}{\chi^2_{1-\alpha/2,n-2}}.$$

Example



70
60
50
40
30
people_fully_vaccinated_per_hundred
10000
20000
30000
40000
gdp_per_capita

Example

```r
df00.model <- lm(people_fully_vaccinated_per_hundred~gdp_per
    _capita, data = df00)
summary(df00.model)
```

```
Call:
lm(formula = people_fully_vaccinated_per_hundred ~ gdp_per_
    capita, data = df00)

Residuals:
    Min      1Q  Median      3Q     Max 
-16.428  -6.176  -0.675   7.997  14.445 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)   1.929e+01  6.075e+00   3.175  0.00588 ** 
gdp_per_capita 1.194e-03  1.957e-04   6.100 1.53e-05 ***
---
Signif.codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.665 on 16 degrees of freedom
Multiple R-squared:  0.6993,  Adjusted R-squared:  0.6805 
F-statistic: 37.21 on 1 and 16 DF,  p-value: 1.534e-05
```

Example

```
confint(df00.model,level=0.99)
```

```
                    0.5 %       99.5 %
(Intercept)    1.5420853852 37.028623328
gdp_per_capita 0.0006222473  0.001765602
```

Analysis of variance (ANOVA)

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

$$\begin{aligned} SS_T = \sum (y_i - \bar{y})^2 &= \sum \big((\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i) \big)^2 \\ &= \underbrace{\sum (\hat{y}_i - \bar{y})^2}_{SS_R} + \underbrace{\sum (y_i - \hat{y}_i)^2}_{SS_{Res}} + 2 \sum (\hat{y}_i - \bar{y})(y_i - \hat{y}_i) \end{aligned}$$

SS_R is called the *regression* or *model sum of squares*.

$$\sum (\hat{y}_i - \bar{y})(y_i - \hat{y}_i) = \sum (\hat{y}_i - \bar{y}) e_i = \sum \hat{y}_i e_i - \bar{y} \sum e_i = 0 - 0 = 0.$$

Fundamental analysis-of-variance identity

$$SS_T = SS_R + SS_{Res}$$

A F-test for the significance of regression

Significance of regression refers to testing whether or not the model $\beta_0 + \beta_1 x_i$ is necessary, i.e. if there is any relationship between x_i and y_i which motivates assuming that $\beta_1 \neq 0$.

General idea

By the ANOVA identity, $SS_T = SS_R + SS_{Res}$, where $SS_R = \sum(\hat{y}_i - \bar{y})^2$ and $SS_{Res} = \sum(y_i - \hat{y}_i)^2$. Note that if $\beta_1 = 0$, then $y_i = \beta_0 + \varepsilon_i$, and thus SS_R measures how much the errors vary, while SS_{Res} measure how much these would vary in an "optimal linear model" $\hat{\beta}_0 + \hat{\beta}_1 x_i$. If SS_{Res} is much smaller than SS_R, we thus expect the hypothesis $\beta_1 = 0$ to be false.

```r
SSres <- sum(df00.model$residuals^2)
SSr <- sum((df00.model$fitted.values - mean(df00$people_
    fully_vaccinated_per_hundred))^2)
SSt <- SSres + SSr

c(SSres,SSr,SSt)
```

```
[1] 1494.620 3475.847 4970.466
```

A F-test for the significance of regression

Hypothesis

$$H_0\colon \beta_1 = 0, \qquad H_1\colon \beta_1 \neq 0$$

Distribution of SS_T and SS_{Res}

- $SS_{Res} \sim \chi^2_{n-2}$ ($n-2$ degrees of freedom)
- SS_T has 1 degree of freedom
- If $\beta_0 = 0$, then $SS_R \sim \chi^2_1$
- If $\beta_0 = 0$, then SS_{Res} and SS_R are independent.

Test statistic

$$F_0 \coloneqq \frac{SS_R/df_R}{SS_{Res}/df_{Res}} = \frac{(MS_R/\sigma^2)/1}{((n-2)MS_{Res}/\sigma^2)/(n-2)} = \frac{MS_R}{MS_{Res}} \sim F_{1,n-2}$$

$\rightarrow$ We reject H_0 if $F_0 > F_{\alpha,1,n-2}$.

Example

```r
anova(df00.model)
```

```
Analysis of Variance Table

Response: people_fully_vaccinated_per_hundred
               Df Sum Sq Mean Sq F value   Pr(>F)    
gdp_per_capita  1 3476.8  3476.8   37.21 1.53e-05 ***
Residuals      16 1495.6    93.4                     
---
Signif.codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The mean response

The mean response

The function

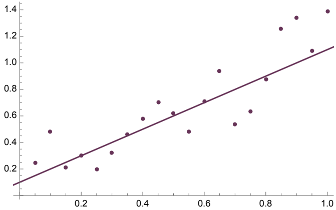
$$\mathbb{E}[y \mid x_0] = \mathbb{E}[\beta_0 + \beta_1 x_0 + \varepsilon_0] = \beta_0 + \beta_1 x_0$$

is called the *mean response* at x_0.

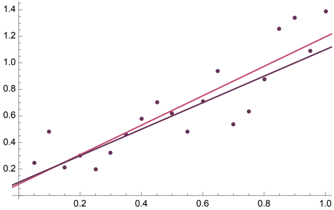
An estimate for the mean response

A point estimate for the mean response is given by

$$\widehat{\mathbb{E}[y \mid x_0]} = \hat{\beta}_0 + \hat{\beta}_1 x_0.$$



The true model and the data



The true model, the data, and the fitted line

Properties of $\widehat{\mathbb{E}[y \mid x_0]}$

Expected value

$$\mathbb{E}\Big[\widehat{\mathbb{E}[y \mid x_0]}\Big] = \mathbb{E}\big[\hat{\beta}_0 + \hat{\beta}_1 x_0\big] = \mathbb{E}\big[\hat{\beta}_0\big] + \mathbb{E}\big[\hat{\beta}_1\big] x_0 = \beta_0 + \beta_1 x_0 = \mathbb{E}[y \mid x_0].$$

Hence $\widehat{\mathbb{E}[y \mid x_0]}$ is an unbiased estimate of $\mathbb{E}[y \mid x]$.

Variance

$$\begin{aligned} \mathrm{Var}(\hat{\beta}_0 + \hat{\beta}_1 x_0) &= \mathrm{Var}\big((\bar{y} - \hat{\beta}_1 \bar{x}) + \hat{\beta}_1 x_0\big) = \mathrm{Var}\big(\bar{y} + \hat{\beta}_1 (x_0 - \bar{x})\big) \\ &\overset{\hat{\beta}_1 \perp \bar{y}}{=} \mathrm{Var}(\bar{y}) + \mathrm{Var}\big(\hat{\beta}_1 (x_0 - \bar{x})\big) = \frac{\sigma^2}{n} + \frac{\sigma^2 (x_0 - \bar{x})^2}{S_{xx}}. \end{aligned}$$

Note that the variance is increasing in $(x_0 - \bar{x})^2$.

Distribution

Since $\hat{\beta}_0$ and $\hat{\beta}_1$ are linear in $\mathbf{y}$, and $\mathbf{y}$ has a normal distribution, we know that $\hat{\beta}_0 + \hat{\beta}_1 x_0$ has a normal distribution. Using the above formulas, it follows that

$$\widehat{\mathbb{E}[y \mid x_0]} \sim N\left(\mathbb{E}[y \mid x_0], \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)\right).$$

Properties of $\widehat{\mathbb{E}[y \mid x_0]}$

Distribution

Since

$$\widehat{\mathbb{E}[y \mid x_0]} \sim N\left(\mathbb{E}[y \mid x_0], \sigma^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)\right),$$

we know that

$$\frac{\widehat{\mathbb{E}[y \mid x_0]} - \mathbb{E}[y \mid x_0]}{\sqrt{\sigma^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}} \sim N(0, 1).$$

In general however, σ^2 is unknown, and we want to replace σ^2 with its estimate MS_{Res}. Since $(n-2)MS_{Res}/\sigma^2 \sim \chi^2_{n-2}$, it follows that

$$\frac{\widehat{\mathbb{E}[y \mid x_0]} - \mathbb{E}[y \mid x_0]}{\sqrt{MS_{Res}(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}})}} \sim t_{n-2}.$$

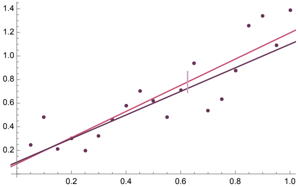
A confidence interval for the mean response

Since

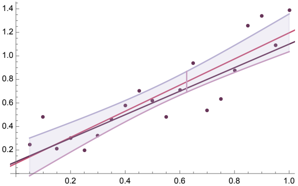
$$\frac{\widehat{\mathbb{E}[y \mid x_0]} - \mathbb{E}[y \mid x_0]}{\sqrt{MS_{Res}\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}} \sim t_{n-2}.$$

we obtain a $100(1-\alpha)$ percent confidence interval for the mean response at x_0 by

$$\mathbb{E}[y \mid x_0] = \widehat{\mathbb{E}[y \mid x_0]} \pm t_{\alpha/2, n-2}\sqrt{MS_{Res}\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}.$$



A 95%-confidence interval for the mean response at x_0



→ Note that the confidence interval is only valid for one point x_0. For simultaneous confidence intervals we must use e.g. Bonferroni.

Example

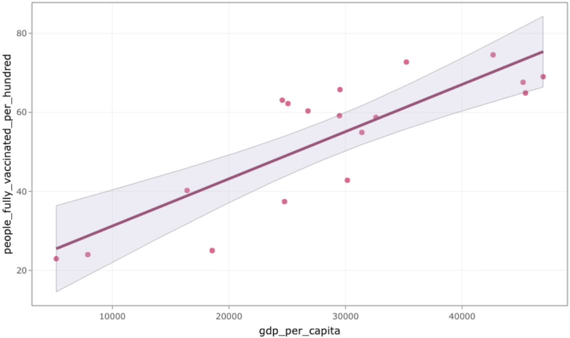
We now use R to calculate a confidence interval for the mean response at gdp_per_capita=15000.

```
newdata = data.frame(gdp_per_capita=15000)
predict(df00.model, newdata, interval="confidence")
```

```
       fit      lwr    upr
1 37.19423 29.71245 44.676
```

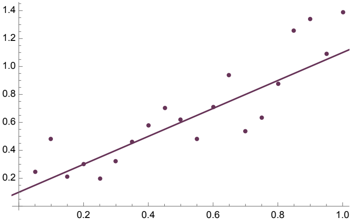
Example

```
pp <- ggplot(df00, aes(x=gdp_per_capita, y=people_fully_
     vaccinated_per_hundred)) +
  geom_point(aes(text=location),color="#D46B8D") +
  geom_smooth(method=lm, se=TRUE,color="#000000",fill="#
    B9B1D3",size=0.25,alpha=0.2) +
  geom_smooth(method=lm, se=FALSE,color="#8B375A") +
  theme_bw()
ggplotly(pp, tooltip="text")
```

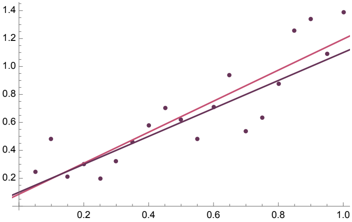


Prediction interval for future observations

Now assume that we want to say something about the distribution of a future observation y_0 at x_0. Then $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$ is a point esitmate of y_0, but the interval on the previous slide does not give a confidence interval for y_0 since $y_0 \neq \mathbb{E}[y \mid x_0]$, even if they have the same point estimates.



The true model and the data



The true model, the data, and the fitted line

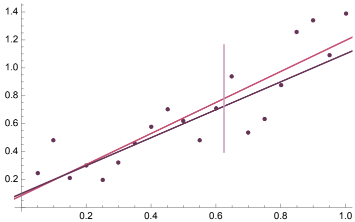
Prediction interval for future observations

To make a prediction interval for a future observation y_0 at x_0, we use that

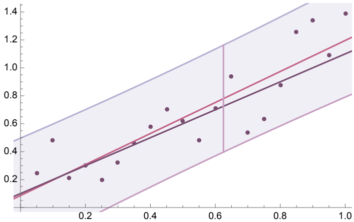
$$y_0 - \hat{y}_0 \sim N\left(0, \sigma^2\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)\right).$$

Proceeding as before by replacing σ^2 with $\hat{\sigma}^2$, we obtain a $100(1-\alpha)$ percent *prediction interval* for a future observation at x_0 by

$$\hat{y}_0 \pm t_{\alpha/2, n-2}\sqrt{MS_{Res}\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}.$$



The true model and the data



The true model, the data, and the fitted line

Example

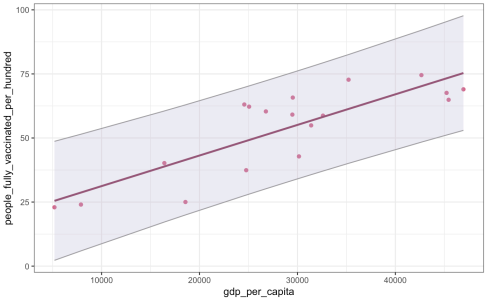
We now use R to calculate a prediction interval for a future observation at gdp_per_capita=15000.

```
newdata = data.frame(gdp_per_capita=15000)
predict(df00.model, newdata, interval="predict")
```

```
       fit      lwr      upr
1 37.19423 15.38189 59.00656
```

Example

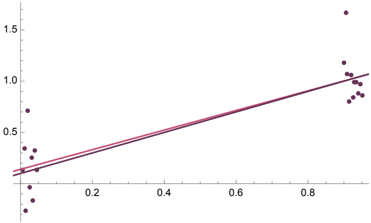
```r
predictions <- predict(df00.model, interval = "predict")
all_data <- cbind(df00, predictions)

ggplot(all_data, aes(x = gdp_per_capita, y = people_fully_vaccinated_per_hundred)) +
  geom_point(aes(text=location),color="#D46B8D")+
  geom_smooth(method=lm, se=FALSE,color="#8B375A",fill="#B9B1D3") +
  geom_line(aes(y = lwr), color = "8B375A") +
  geom_line(aes(y = upr), color = "8B375A") +
  geom_ribbon(aes(ymin=lwr,ymax=upr), fill="#B9B1D3", alpha=0.5) + theme_bw()
```



Why is it important to know the variance of $\hat{\beta}_0$ and $\hat{\beta}_1$?

1. We need to know what it is to be able to find confidence intervals
2. The width of the confidence intervals depend on the variance of our estimates. If we know the form of these, we can plan the experiment in order to make it smaller.

$$\hat{\beta}_1 \pm t_{\alpha/2,n-2}\sqrt{\frac{MS_{Res}}{S_{xx}}}$$



$$\hat{\beta}_0 \pm t_{\alpha/2,n-2}\sqrt{MS_{Res}(\frac{1}{n} + \frac{\bar{x}}{S_{xx}})}$$