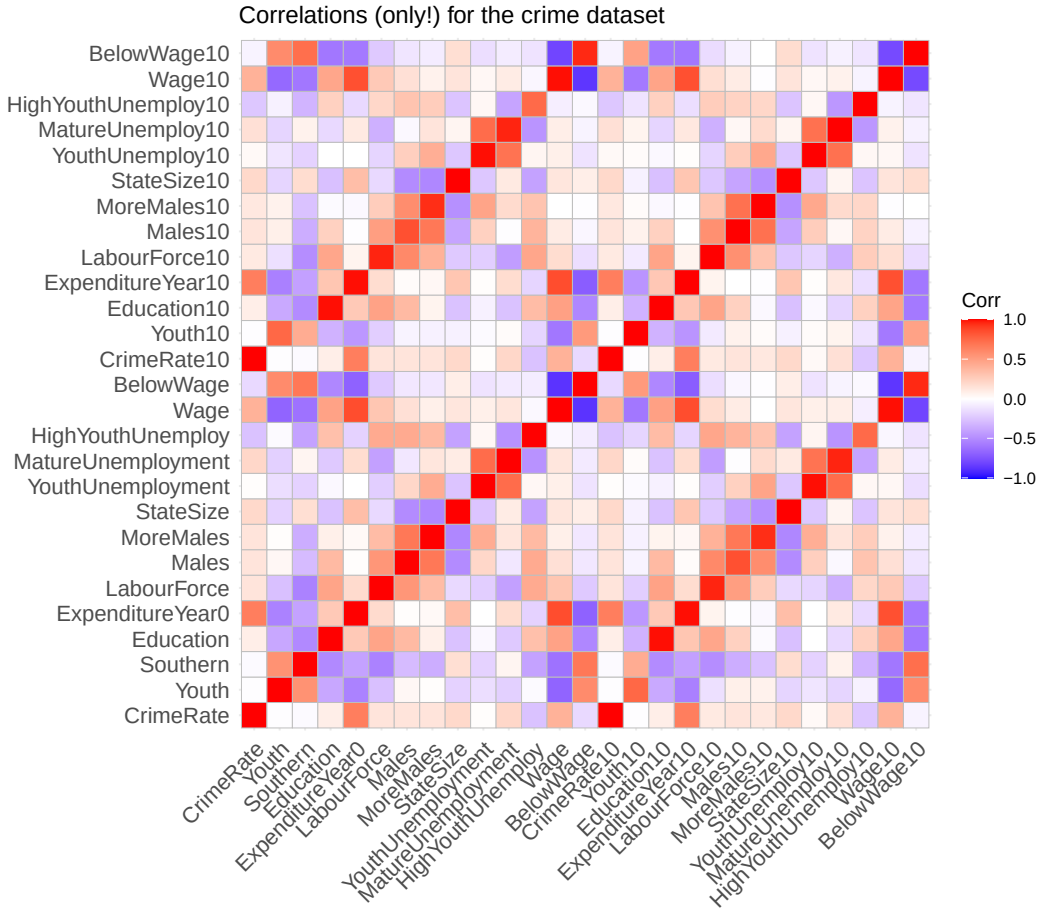
Linear Models

Nikodem Lewandowski

Predictions vs. Correlations

```
#these are registered violent incidents per 100k citizens
cors <- cor(cbs, method = 'spearman')
ggcorrplot(cors, method="square")+
  ggtitle("Correlations (only!) for the crime dataset")
```



University of Gdańsk

Correlation

Correlation is a statistical measure that indicates the extent to which two variables are related. In other words, it shows how strong the relationship is between two variables.

Spearman's rank correlation coefficient

d_i is the difference in paired ranks and n is number of cases.

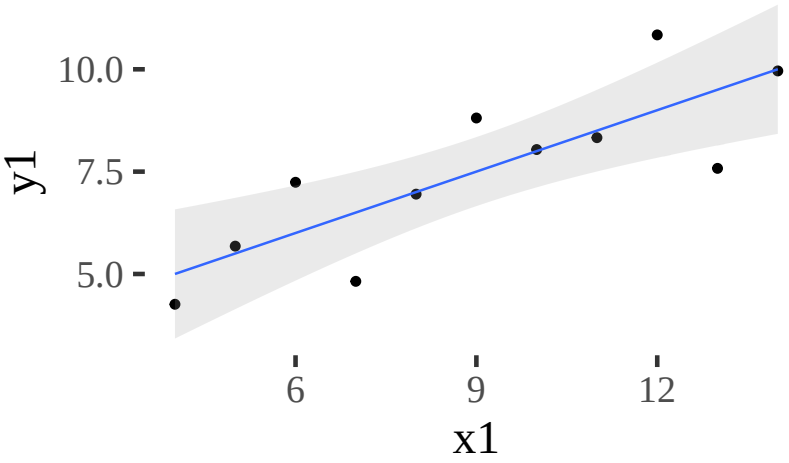
$$r_s = 1 - \frac{6 \sum d^2}{n(n^2 - 1)}$$

Pearson's correlation coefficient:

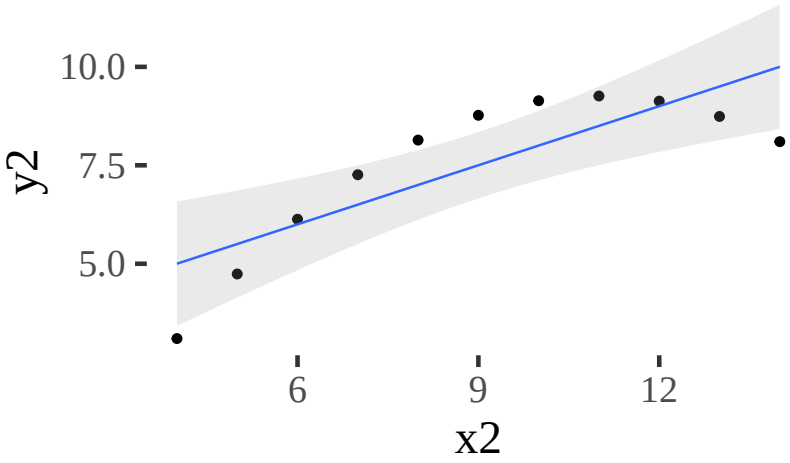
$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2} \sqrt{(y_i - \bar{y})^2}}$$

Anscombe's quartet (Pearson's correlation)

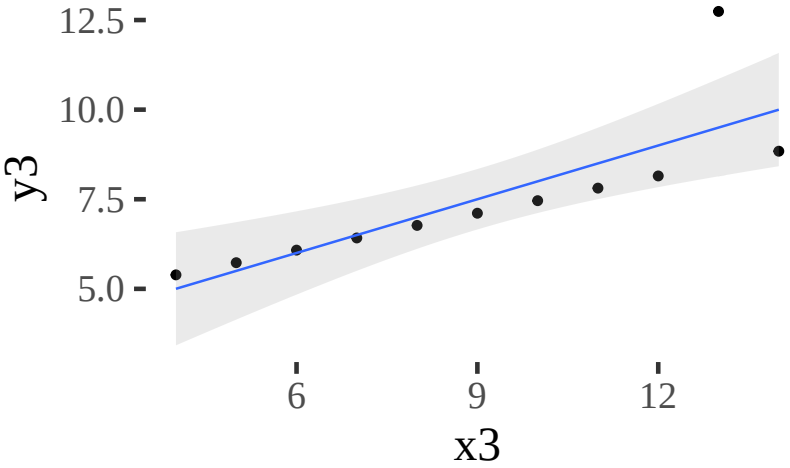
Correlation coefficient = 0.82



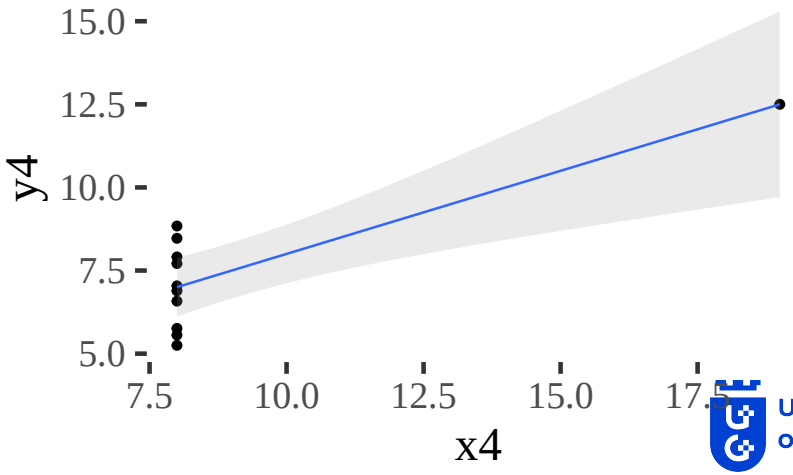
Correlation coefficient = 0.82



Correlation coefficient = 0.82



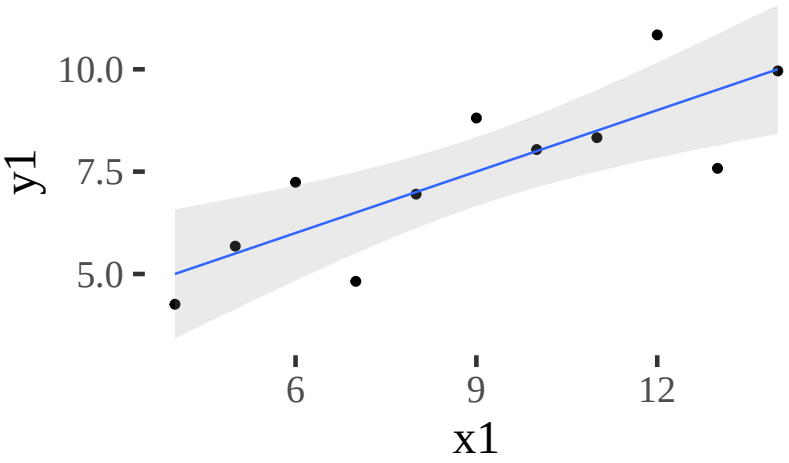
Correlation coefficient = 0.82



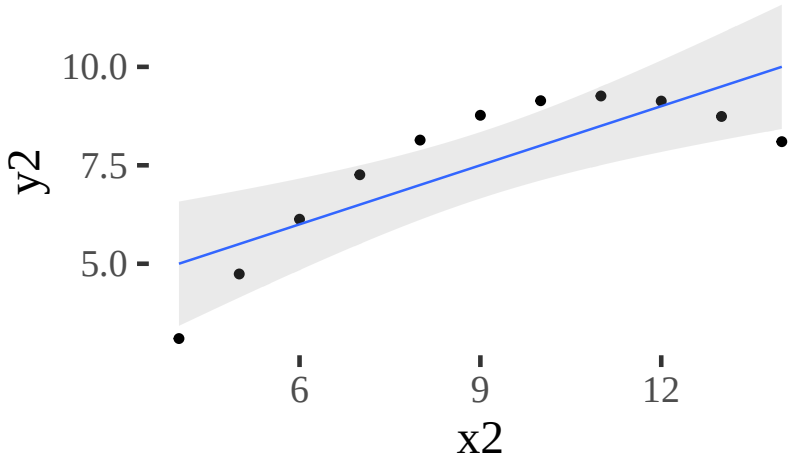
University of Gdańsk

Anscombe's quartet (Spearman's correlation)

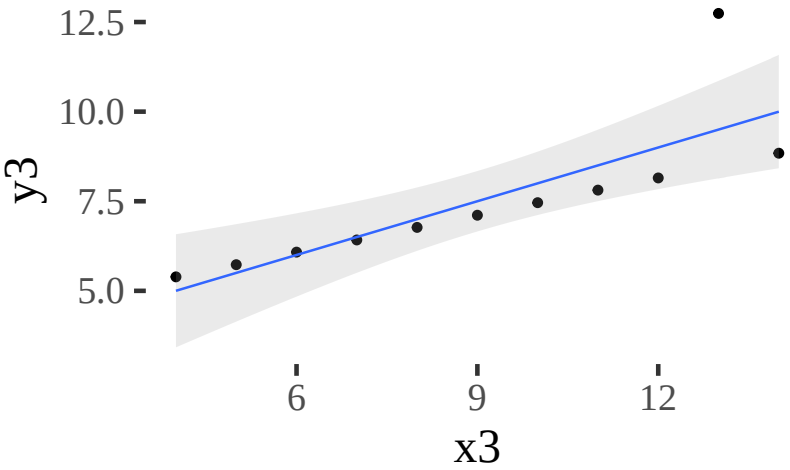
Correlation coefficient = 0.82



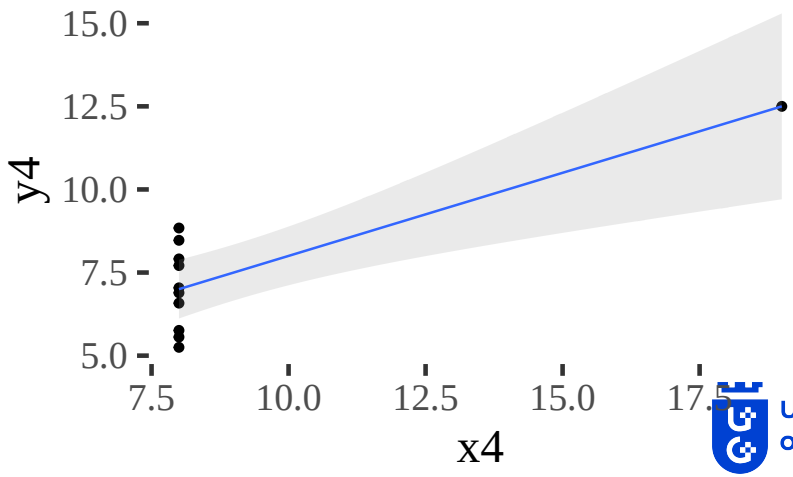
Correlation coefficient = 0.69



Correlation coefficient = 0.99

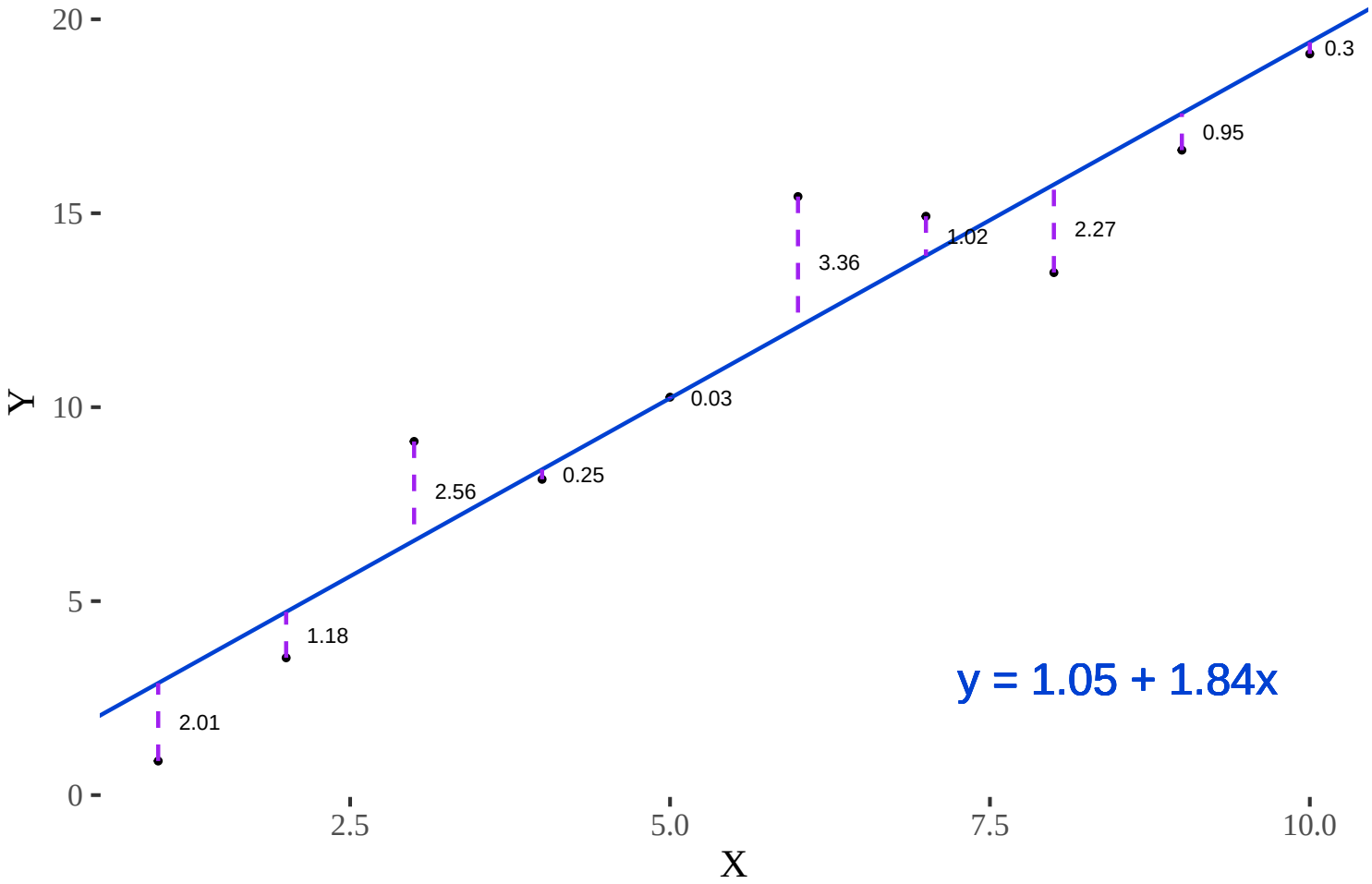


Correlation coefficient = 0.5



University of Gdańsk

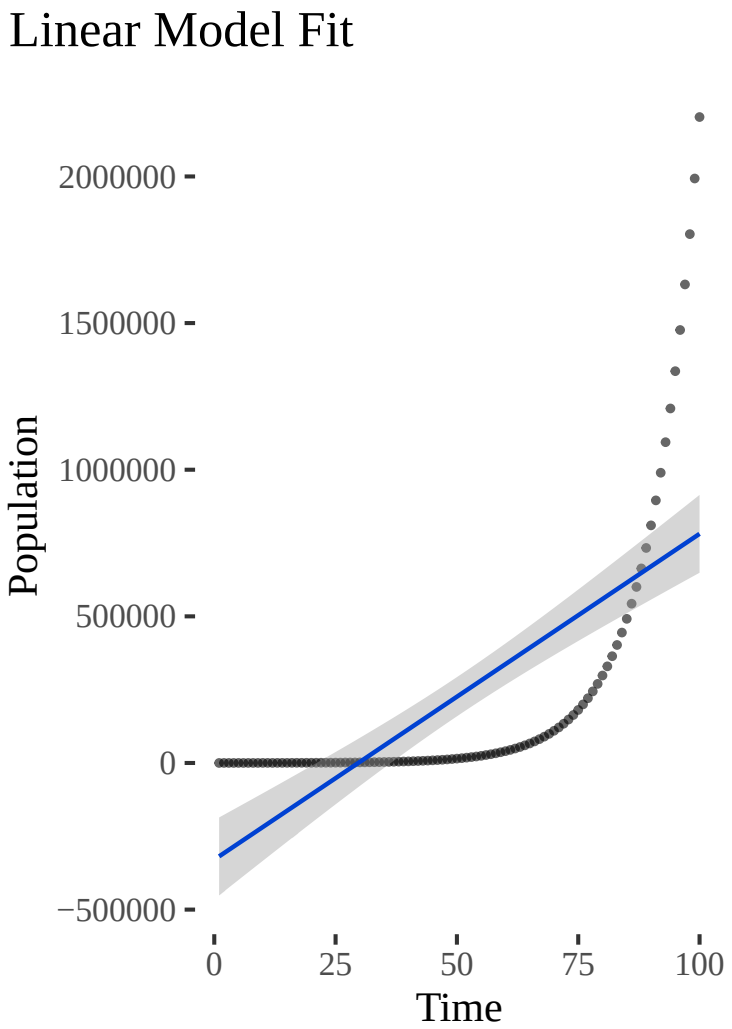
Simple linear regression

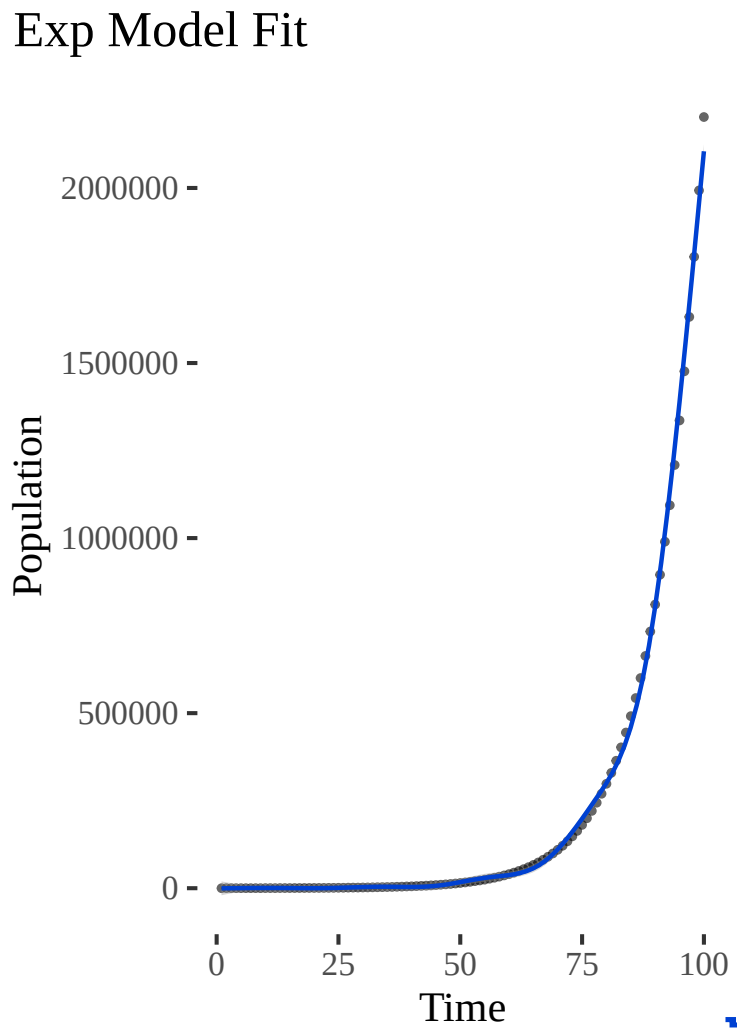


$$y = a + bx$$

University of Gdańsk

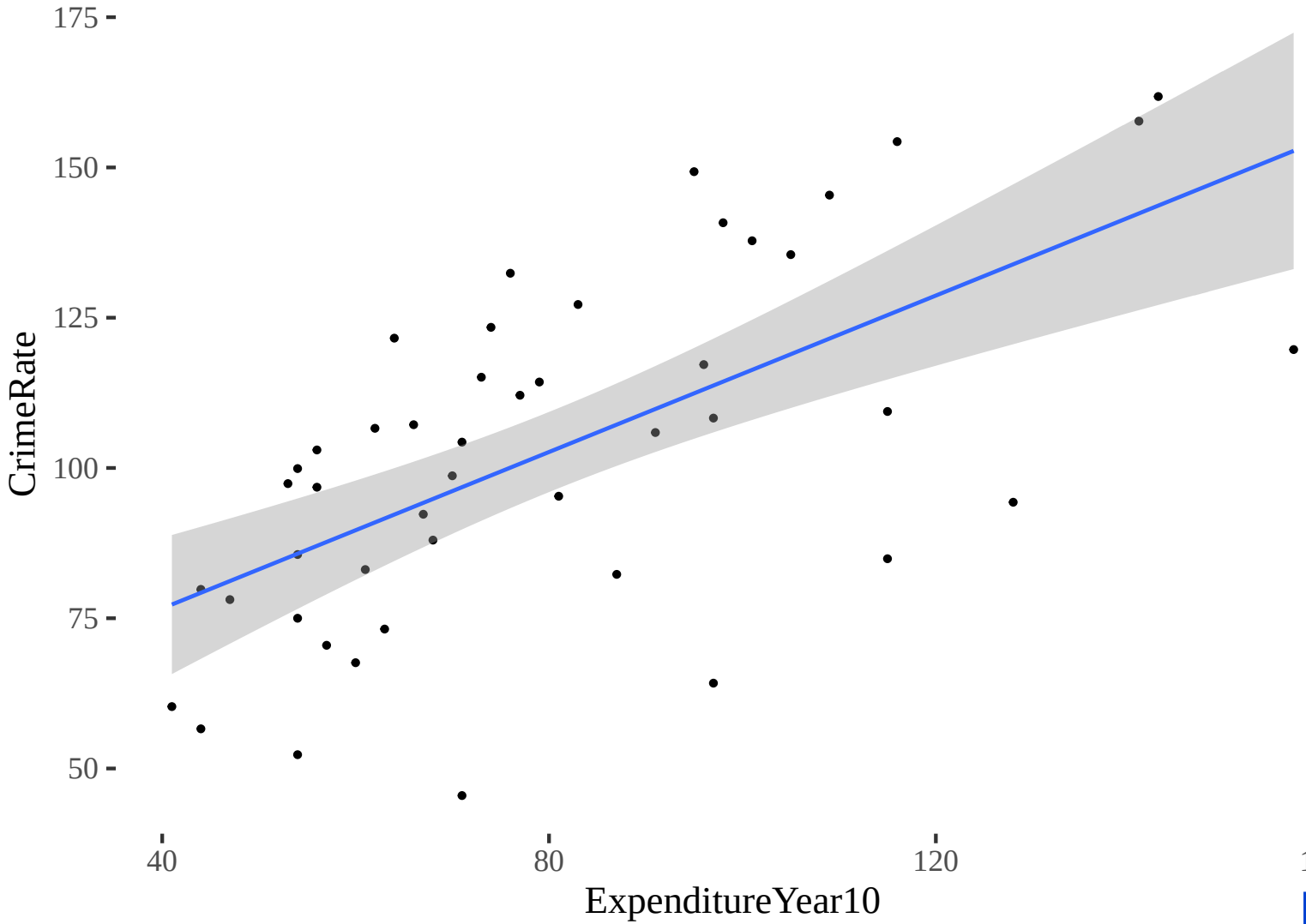
Linear Model is NOT Universal





University of Gdańsk

Linear model



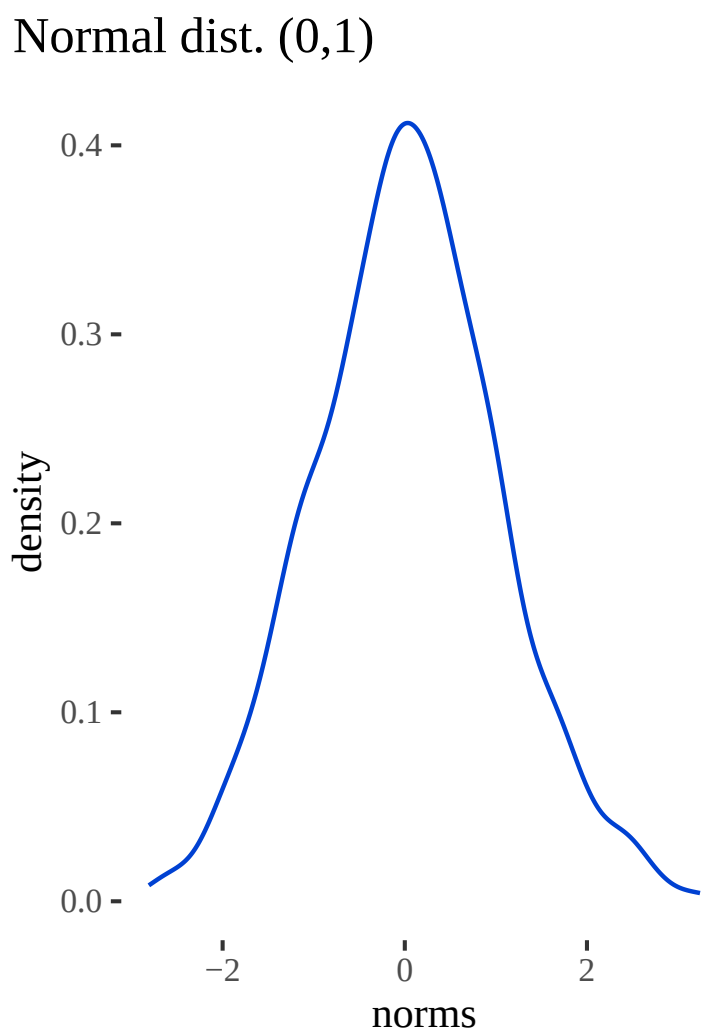
University of Gdańsk

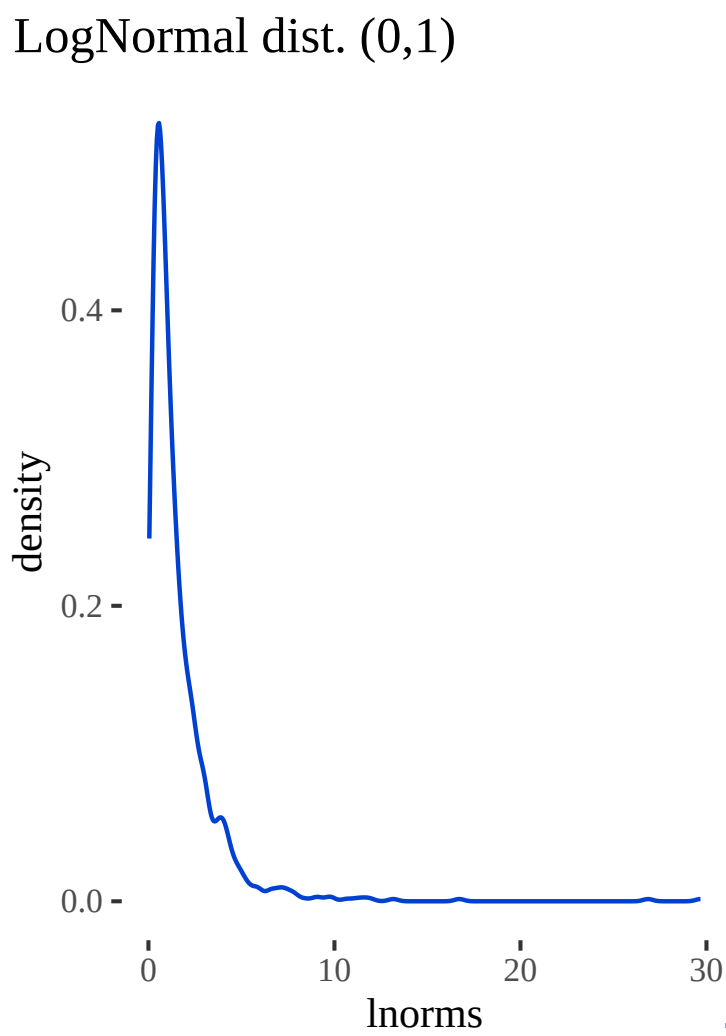
Model of Expenditure vs Crime Rate

$$
\begin{aligned}
\text{rate} &\sim \text{Normal}(\mu, \sigma) \\
\mu &\sim a + b * exp10 \\
a &\sim \text{Normal}(50, 20) \\
b &\sim \text{LogNormal}(0, 1) \\
\sigma &\sim \text{Uniform}(0, 30)
\end{aligned}
$$

University of Gdańsk

Normal vs LogNormal





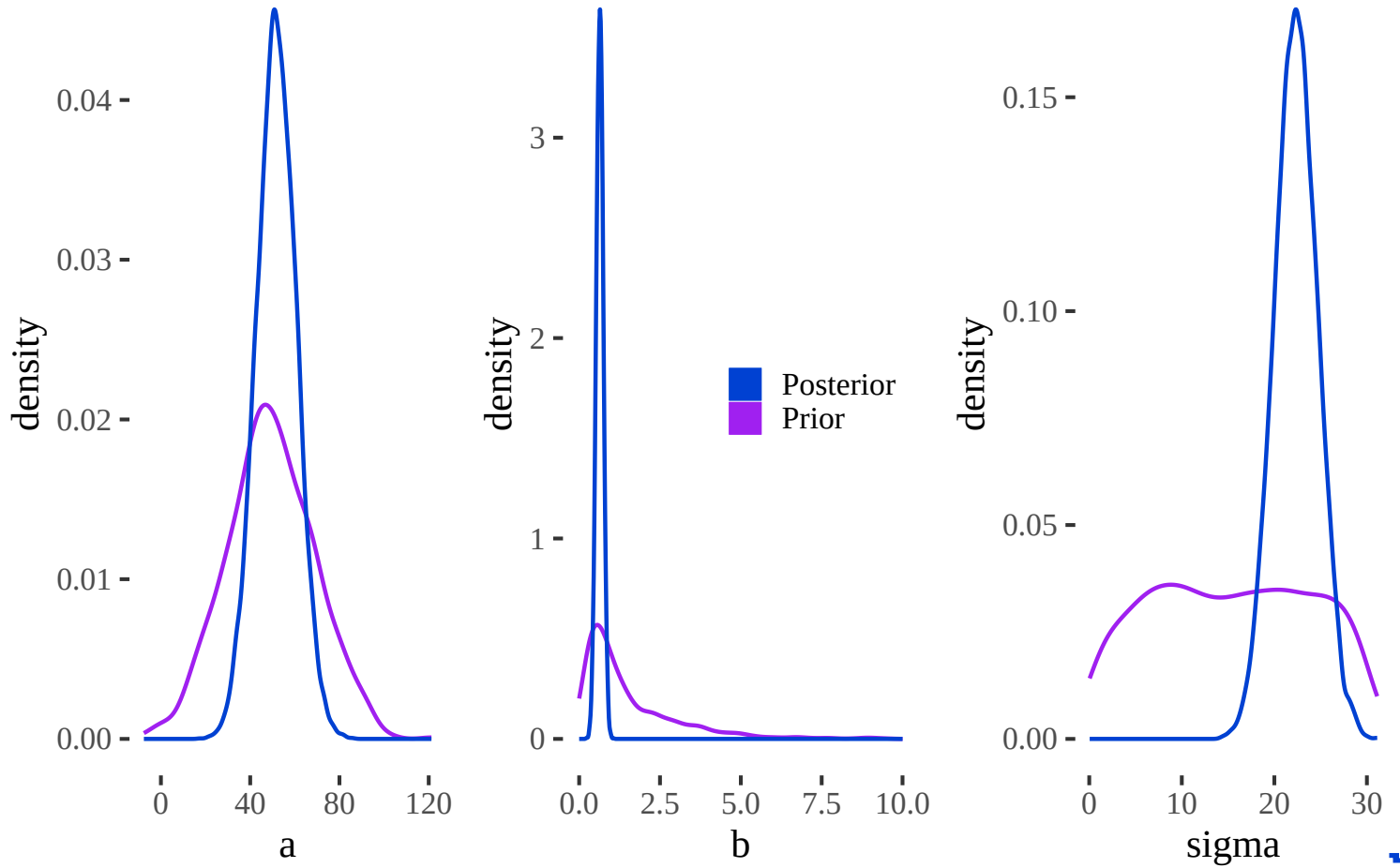
University of Gdańsk

Running the Model

```
dat <- list(
  rate = cbs$CrimeRate,
  exp10 = cbs$ExpenditureYear10)

# We could have standardize the data!
# For simplicity we won't do it now

expenditureModel<- quap(
  alist(
    rate ~ dnorm( mu , sigma ) ,
    mu <- a + b * exp10,
    a ~ dnorm(50, 20) ,
    b ~ dlnorm( 0 , 1 ) ,
    sigma ~ dunif(0, 30)
  ), data = dat
)
```

University of Gdańsk

Priors and Their Posteriors



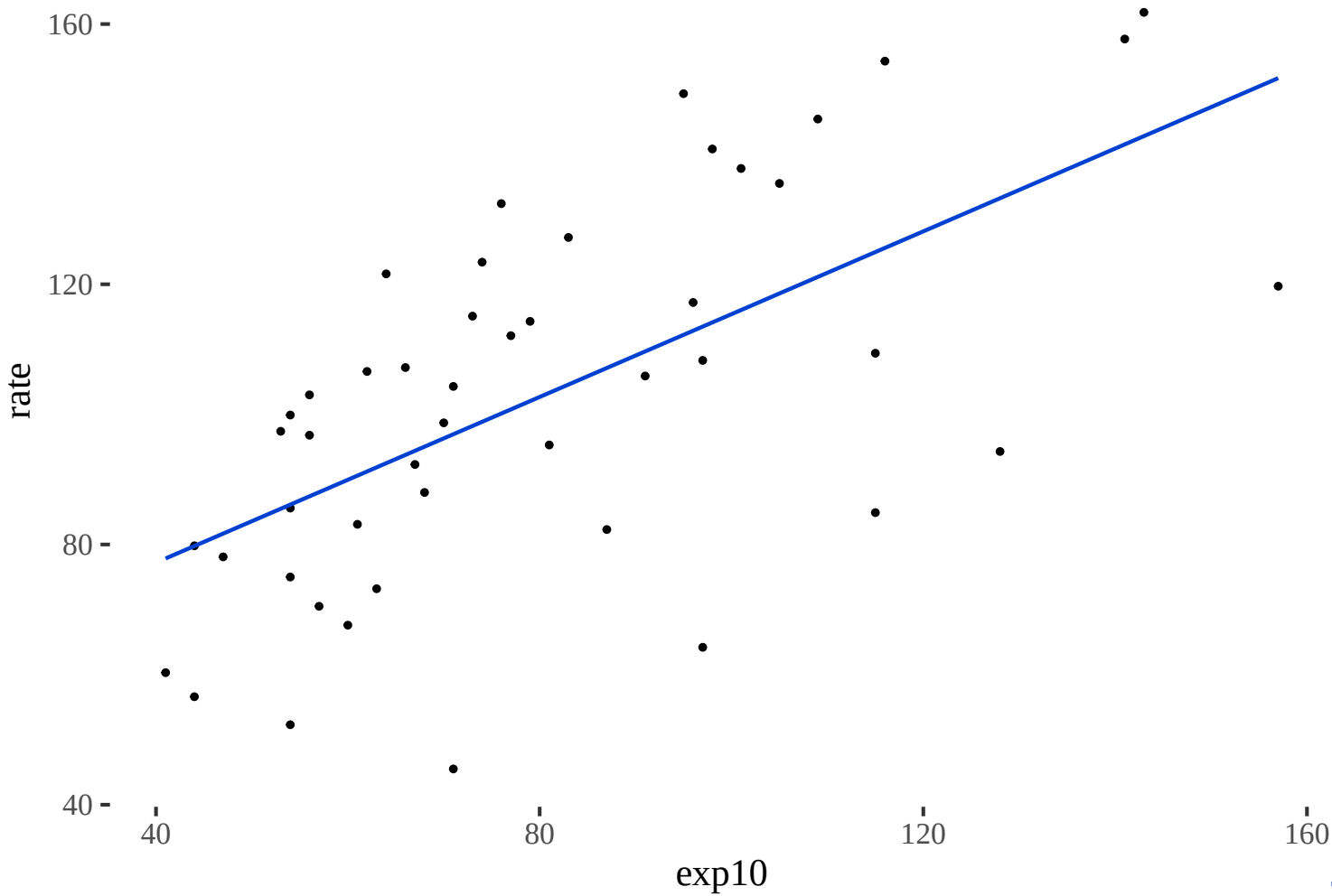
University of Gdańsk

The mean line

```
post <- extract.samples(expenditureModel)

a_map <- mean(post$a)
b_map <- mean(post$b)
x = dat$exp10

y_vals <- a_map + b_map * x

ggplot()+ geom_point( aes(x= dat$exp10, y = dat$rate)) + th +
geom_line(aes(x= x, y = y_vals), col= UGblue, linewidth = 1)+
  labs(x= 'exp10', y= 'rate')
```

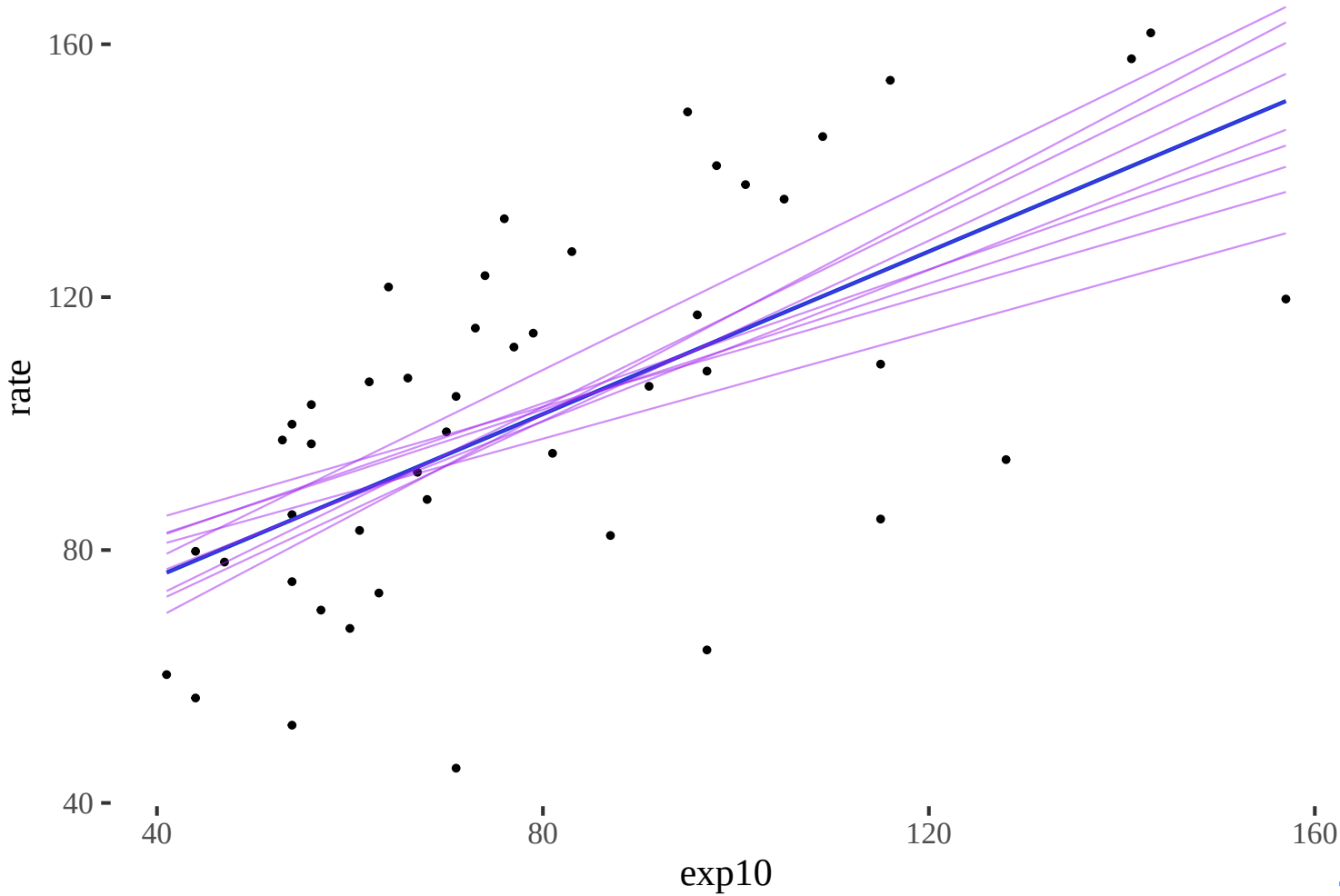
University of Gdańsk

The mean line



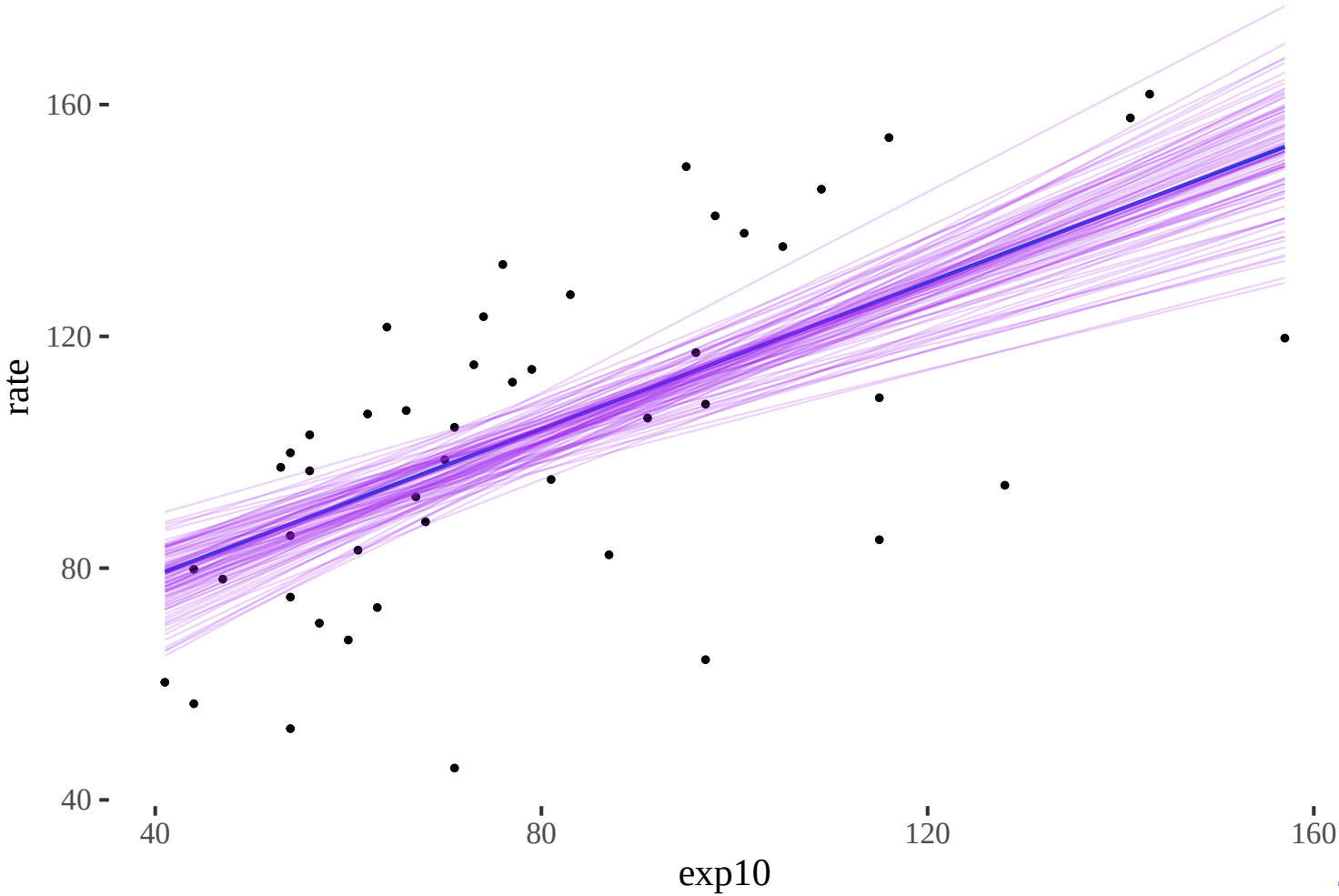
University of Gdańsk

Mean line and 10 sampled lines



University of Gdańsk

Mean line and 100 sampled lines



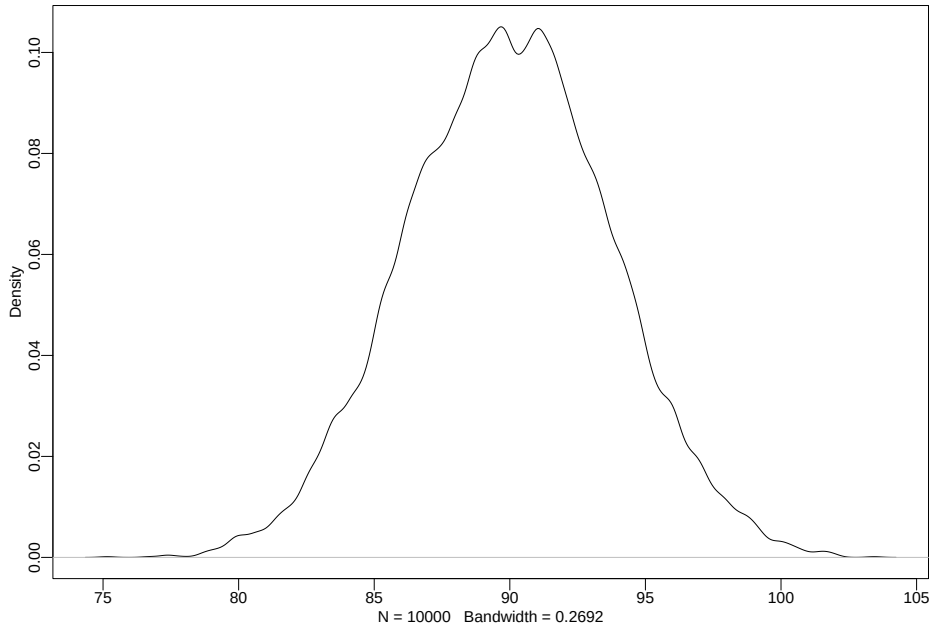
University of Gdańsk

Predictive Power!

```
samples <- extract.samples(expenditureModel)
new_exp10 <- 60
predicted_rate <- samples$a + samples$b * new_exp10

precis(predicted_rate)[1:4]
```

```
                  mean       sd     5.5%    94.5%
predicted_rate 89.95242 3.774019 83.82167 95.99207
```

```
dens(predicted_rate)
```



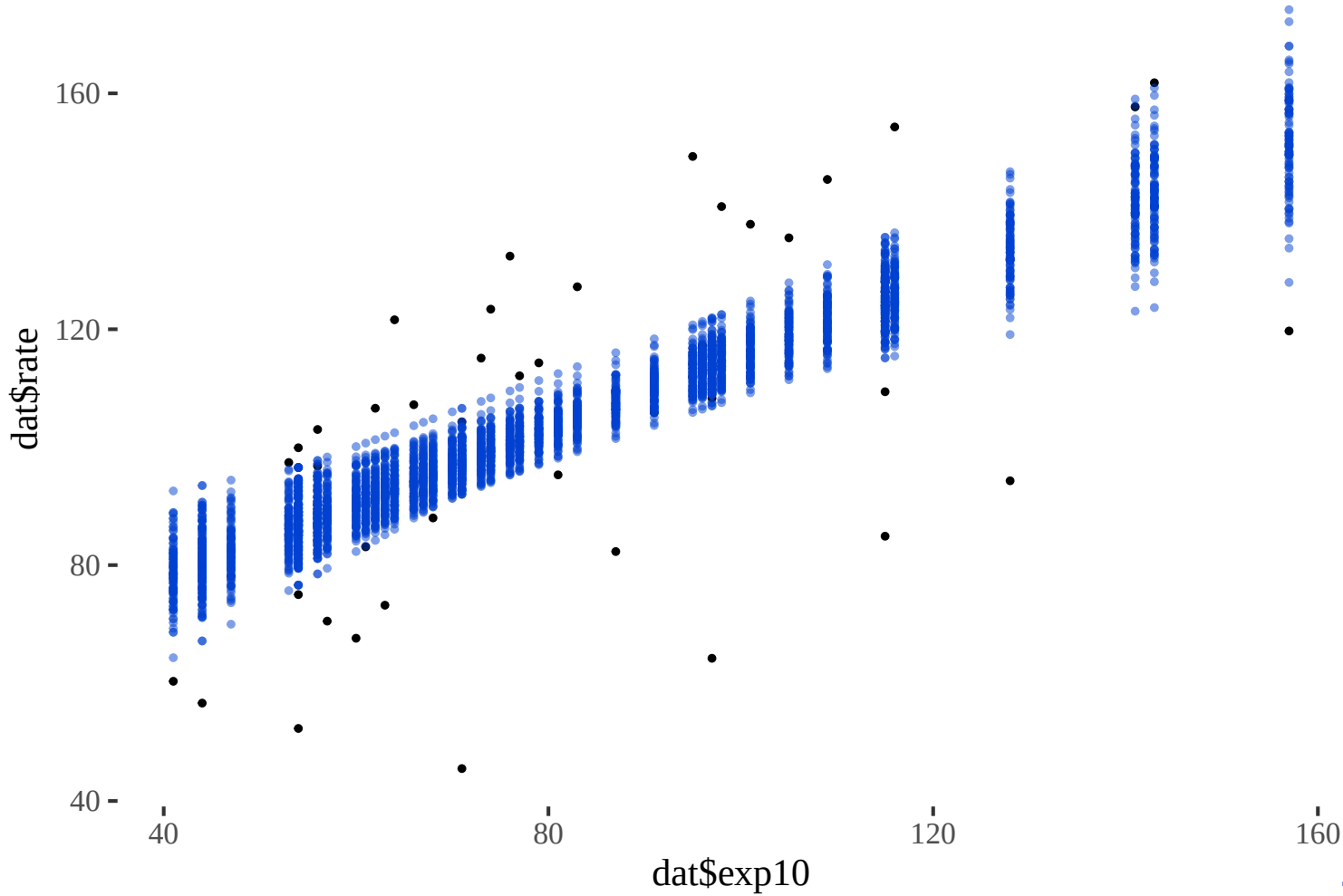
University of Gdańsk

Predictions Evaluation

```
mu <- link(expenditureModel, data = data.frame(exp10 = dat$exp10), 100)

str(mu)

predictions <- data.frame(
  exp10 = rep(dat$exp10, each = 100),
  rate = as.vector(mu),
  sample = rep(1:100, times = 47)
)

ggplot() +
  geom_point( aes(x= dat$exp10, y = dat$rate)) + th +
  geom_point(data = predictions, aes(x = exp10, y = rate),
             col = UGblue, alpha = 0.5)
```

University of Gdańsk

Predictions Evaluation



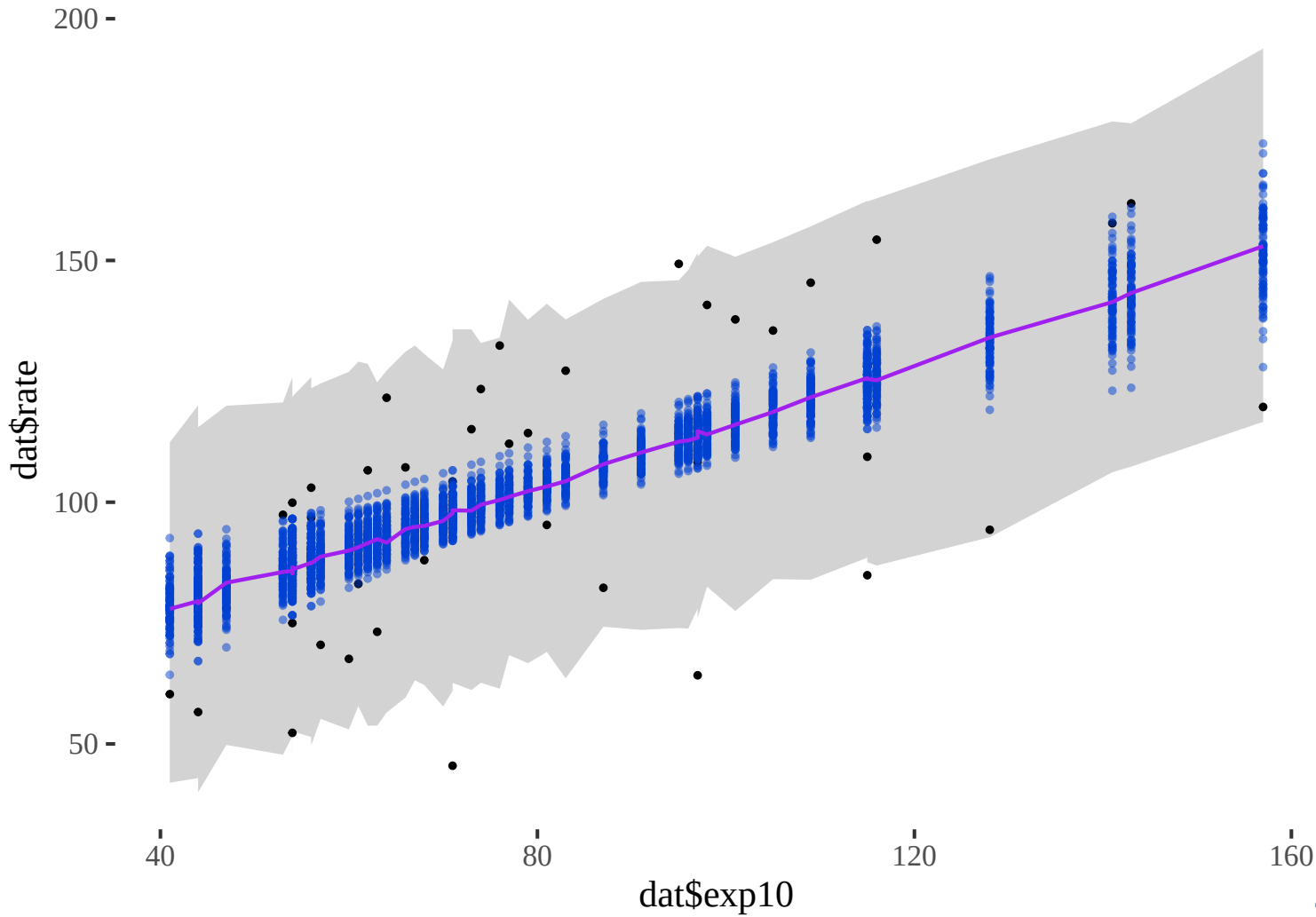
University of Gdańsk

Posterior Predictive Check

```
sim_rate <- sim(expenditureModel,
                data = data.frame(exp10 = dat$exp10))

mu_sim <- apply(sim_rate, 2, mean)
hpdi_sim <- data.frame(t(apply(sim_rate, 2, HPDI)))
names(hpdi_sim) <- c("low", "high")

ggplot() +
  geom_ribbon(aes(x = dat$exp10, ymin = hpdi_sim$low, ymax = hpdi_sim$high),
              fill= 'lightgray')+
  geom_point( aes(x= dat$exp10, y = dat$rate)) + th +
  geom_point(data = predictions, aes(x = exp10, y = rate),
             col = UGblue, alpha = 0.5)+
  geom_line(aes(x= dat$exp10, y= mu_sim), col = 'purple', linewidth= 1)
```

University of Gdańsk

Posterior Predictive Check



University of Gdańsk

Polynomial Model - Overfitting Example

```
dat$exp10_2 <- dat$exp10^2
dat$exp10_3 <- dat$exp10^3


expenditureModelPoly <- quap(
  alist(
    rate ~ dnorm( mu , sigma ) ,
    mu <- a + b1 * exp10 + b2 * exp10_2 +
      b3 * exp10_3,
    a ~ dnorm(50, 20) ,
    c(b1, b2, b3) ~ dnorm( 1 , 3 ) ,
    sigma ~ dunif(0, 30)
  ), data = dat
)
```

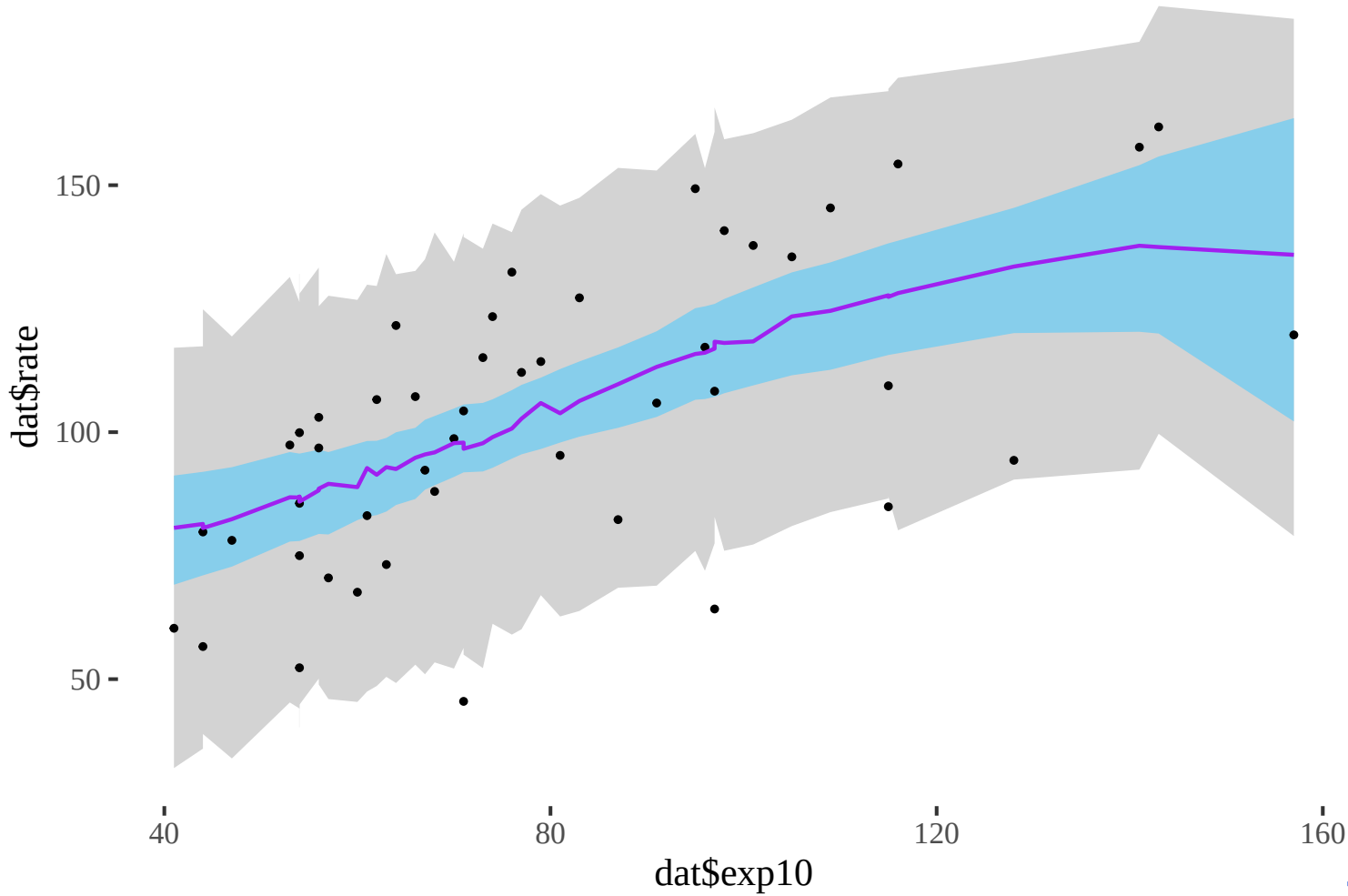
University of Gdańsk

Polynomial Model - Overfitting Example

```
pred_df <- list(exp10 = dat$exp10,
                exp10_2 = dat$exp10^2,
                exp10_3 = dat$exp10^3
)
mu_poly_mean <- link(expenditureModelPoly, data = pred_df)

hpdi_poly_mean <- data.frame(t(apply(mu_poly_mean, 2, HPDI)))
names(hpdi_poly_mean) <- c("low", "high")

mu_poly <- sim(expenditureModelPoly, data = pred_df)
mean_poly <- apply( mu_poly, 2, mean)
hpdi_poly <- apply( mu_poly, 2, HPDI, prob = .89)
hpdi_poly <- data.frame(t(apply(mu_poly, 2, HPDI)))
names(hpdi_poly) <- c("low", "high")


ggplot() +
  geom_ribbon(aes(x = dat$exp10, ymin = hpdi_poly$low,
                  ymax = hpdi_poly$high), fill= 'lightgray')+
  geom_ribbon(aes(x = dat$exp10, ymin = hpdi_poly_mean$low,
                  ymax = hpdi_poly_mean$high), fill= 'skyblue')+
  geom_point( aes(x= dat$exp10, y = dat$rate)) + th+
  geom_line(aes(x= dat$exp10, y= mean_poly), col = 'purple', linewidth= 1)
```

University of Gdańsk

Overfitted Example



University of Gdańsk