HIGH DIMENSIONAL STATISTICS

LECTURE 1: LINEAR MODEL

Msc programme in Mathematics and Statistics

University of Helsinki

Matti Pirinen

BASICS OF STATISTICS

- Linear model summary material is available on the course home page

- If you need some more material about the fundamental ideas of statistics and/or R language, you can read more material here
  - https://www.mv.helsinki.fi/home/mjxpirin/medstat_course/
  - What are parameter estimates, standard errors, confidence intervals, P-values, Normal distribution?
  - How to do basic things with R language?

REGRESSION

- What is the expected value of outcome *Y* for a sample whose predictor values are *X* = *x*? How does E(*Y* | *X* = *x*) depend on *x*?
- Linear regression example: Expected sales depends on the amount of advertisement

$$\texttt{sales} = \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times \texttt{newspaper} + \epsilon.$$

- Logistic regression example: Risk of default (p) depends on credit card balance
  - $\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \times \text{balance}$
- Regression function is a simple function of coefficients (β_k)
- We find estimates $\left(\hat{\beta}_k\right)$ for coefficients
  - Inference: What do we know about the values of coefficients?
  - Prediction: What is predicted outcome value for a new sample *X* = *x*' ?

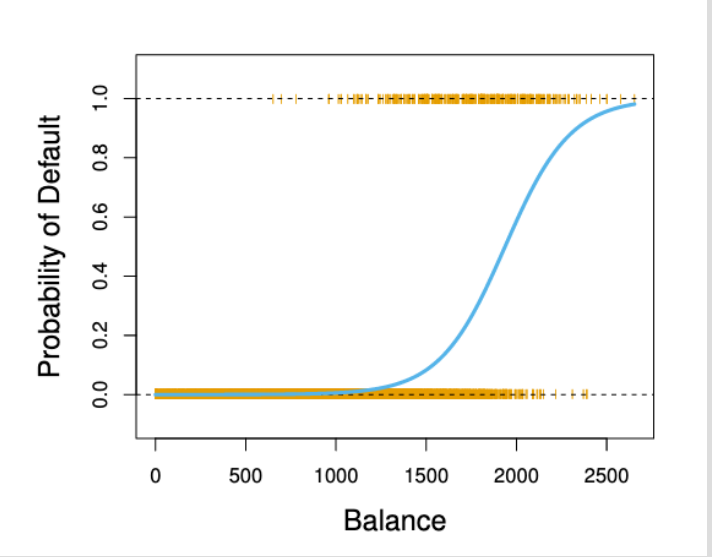
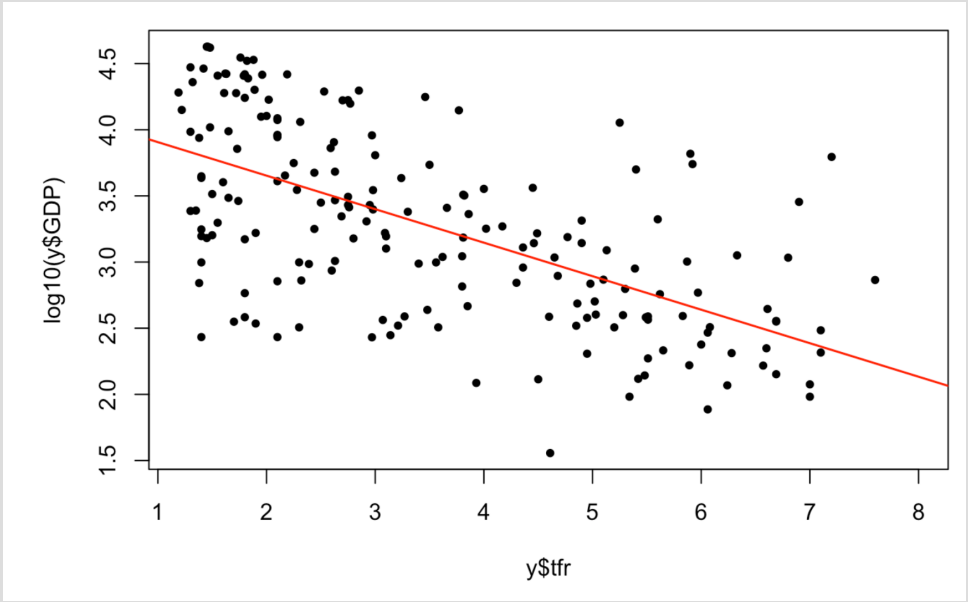


Fig. from ISL

FITTED MODEL

```
## Call:
## lm(formula = log10(GDP) ~ tfr, data = y)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.4358 -0.3703 -0.0134  0.4197  1.4586
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.15995    0.09179   45.32   <2e-16 ***
## tfr         -0.25333    0.02343  -10.81   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5647 on 188 degrees of freedom
##   (17 observations deleted due to missingness)
## Multiple R-squared:  0.3834, Adjusted R-squared:  0.3801
## F-statistic: 116.9 on 1 and 188 DF,  p-value: < 2.2e-16
```



Each unit of total fertility rate associates with a decrease of -0.25 in $\log_{10}$(GDP) or a multiplication of GDP by 10^(-0.25)=0.56.

What is the uncertainty related to these estimates?

GDP = gross domestic product
tfr = total fertility rate, avg. no. of children per a woman

STANDARD ERRORS AND CONFIDENCE INTERVALS

```
## Coefficients:
##             Estimate Std. Error
## (Intercept)  4.15995    0.09179
## tfr         -0.25333    0.02343
```

95% confidence interval from confint( ) function

```
              2.5 %     97.5 %
 3.9788741  4.3410321
-0.2995475 -0.2071029
```

95% confidence interval (95%CI):
If we were to repeat the model fitting to many data sets of this kind, then, on average, in 95% of our data sets, the 95%CI would cover the true value of the parameter. This is a **frequentist** concept, i.e., defined through what would happen on average across many repeated data sets.

In many cases, we can interpret the 95%CI in a **Bayesian** way as an approximation to a 95% **credible interval**:
Interval that covers the true value with 95% probability. This is a **subjective** concept, i.e., combines analyst's prior distribution with data.

(For this Bayesian interpretation of 95%CI, we need to assume a uniform prior distribution and an approximately Gaussian likelihood.)

Standard error = standard deviation of the sampling distribution of parameter estimates. This is a frequentist concept that describes how much the estimates were expected to vary.

For example, if we were to estimate tfr's coefficient from many similarly structure data sets, then the SD of the estimates would be ~ 0.023.

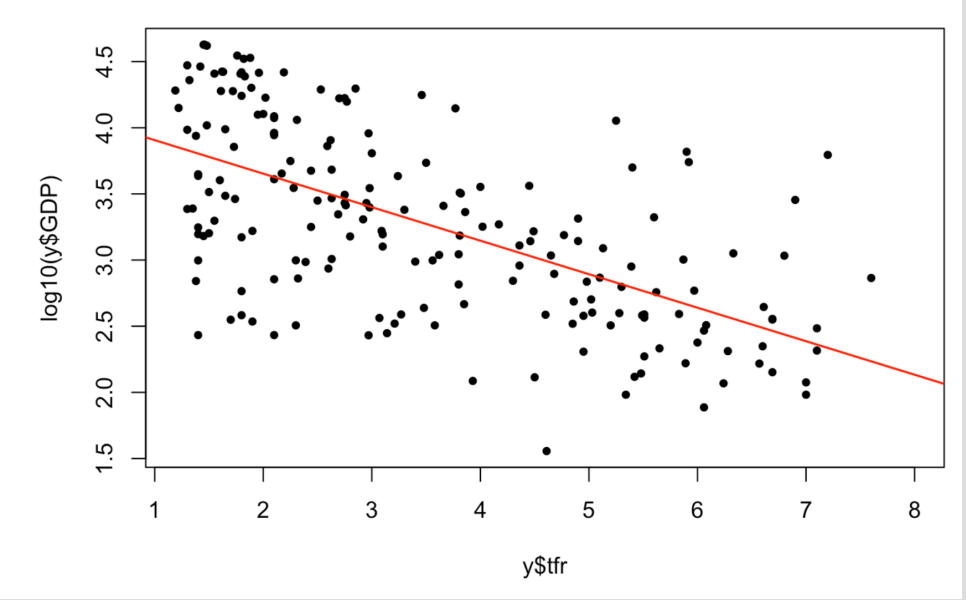
One can approximate 95%CI by adding 1.96 x SE on both sides of the estimate:
$-0.253 \pm 1.96*0.0234 = (-0.299, -0.207)$

This is very close to the values given above by R's *confint( )* function.

PREDICTION

lm.tfr model coefficients:

```
## Coefficients:
##             Estimate Std. Error
## (Intercept)  4.15995    0.09179
## tfr         -0.25333    0.02343
```



Prediction for a country with tfr = 4 is 4.15995 – 0.25333*4 = 3.14663.

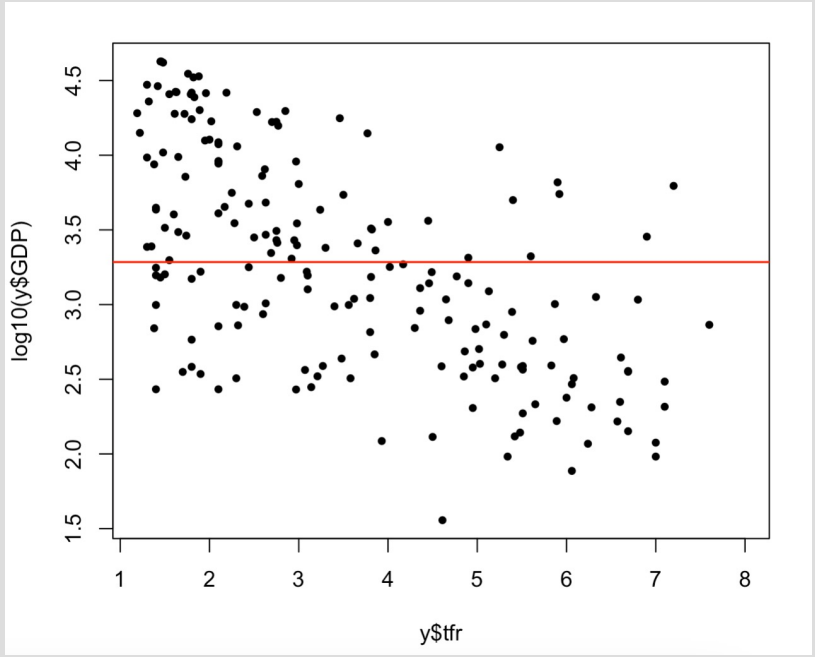
Uncertainty of the prediction can be described by a 95% prediction interval. Frequentist interpretation is that in 95% of the data sets, the prediction interval covers the true value. Here 95% pred. int. is (2.03, 4.26).

```
> predict(lm.tfr,  newdata = data.frame(tfr = 4), interval = "pred")
       fit      lwr     upr
1 3.146652 2.029485 4.26382
```

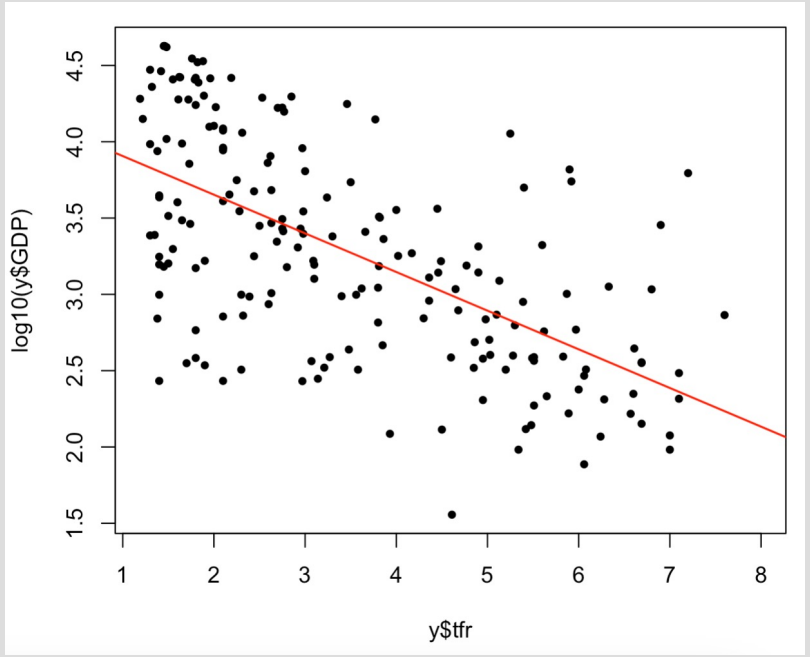
VARIANCE EXPLAINED

```
## Multiple R-squared:  0.3834, Adjusted R-squared:  0.3801
```

Linear model with only intercept (y = a)



Linear model with also slope (y = a + b x)



Residual: deviation of outcome from predicted value.

Variance of outcome variable is the same as variance of residuals of the left-hand model, var(y - a) = 0.5145.

Variance of residuals on the right-hand model is var(y - a - b x) = 0.3189.

Variance explained by the model is 0.5145 – 0.3189 = 0.1956.

R^2 is the proportion of variance explained: 0.1956/0.5145 = 0.38.

ASSUMPTIONS OF LINEAR MODEL

Let's list the assumptions behind the standard linear model and properties of its least squares estimates (LSE).

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i$$

1. Additivity and linearity. We assume that each predictor acts **additively** (there is + between terms that involve different predictors) and that the effect of each predictor on outcome is **linear** (predictor is simply multiplied by a β coefficient). (How to extend linear model outside these assumptions?)
2. Error terms are independent of each other and of predictors. If this does not hold, then the amount of information in data does not correspond to the number of observations, and statistical inference based on theoretical distributions will be invalid. Additionally, LSE is not an optimal unbiased point estimate but a ganeralized least squares estimation, that takes into account the correlation between errors, gives more precise estimates.
3. Errors have same variance (homoscedasticity). If this does not hold, then a weighted linear regression would give more precise estimates.
4. Errors are Gaussian (i.e. have a normal distribution). Under this assumptions LSE coincides with the maximum likelihood estimate and hence has many optimality properties. However, LSE has several optimality properties even without this assumptions. For example, Gauss-Markov theorem says that LSE $\widehat{\beta}$ has the smallest sampling variance among all linear and unbiased estimators of β as long as errors are homoscedastic and uncorrelated, no matter what is their distribution. Gaussian errors is in practice the least important assumption out of the ones listed here.