Introductory Econometrics

Regression
Functional form, Model Selection, Prediction

Monash Econometrics and Business Statistics

Semester 2, 2018

Recap

- We have studied the multiple regression model and learnt:
  1. to express it for a single observation, and, using matrix form, for n observations
  2. the OLS estimator and its derivation in matrix form
  3. the assumptions needed for the OLS estimator to be
     - 3.1 an unbiased estimator
     - 3.2 the best linear unbiased estimator
     - 3.3 normally distributed
  4. to interpret the parameters of a regression model
  5. to test a simple hypothesis about a single parameter
  6. to perform a joint test of multiple linear restrictions, and in particular testing the overall significance of a model
  7. to test a hypothesis involving a linear combination of parameters

Lecture Outline

- The world is non-linear: How useful can a linear model be? (textbook reference 2-4b, 6-2b, 6-2c)
- The interpretation of parameters in log-level, level-log and log-log models
- Transformation of persistent time series data (textbook reference 11-3, p. 395-396)
- Model selection: the adjusted R-Squared (textbook reference section 6.3) and other model selection criteria
- Confidence interval for the conditional mean and prediction interval for the target (textbook reference 6.4)

Is a linear model useful in a non-linear world?

- We all feel that the world is non-linear. Our speed in learning (or any other activity) accelerates as we grow up, gets to a peak and goes downhill eventually. How good is a linear model in this non-linear world?
- But the linear regression model only needs to be **linear in parameters**.
- y and x_1 to x_k can be non-linear transformations of variables.
- It is quite usual that y and some of x variables are logarithms of observed variables, and also some x variable can be quadratic functions of measured variables
- Example: Recall the wage example:

$$\begin{aligned} \widehat{wage} &= -128.89 + 42.06\ educ + 5.14\ IQ \\ n &= 935,\ R^2 = .134 \end{aligned}$$

Models involving logarithms: log-level

- This is not satisfactory because it predicts that regardless of what your wage currently is, an extra year of schooling will add \$42.06 to your wage.
- It is more realistic to assume that it adds a *constant percentage* to your wage, not a constant dollar amount
- How can we incorporate this in the model?
- Logarithm of y on x: $\widehat{\log(y)} = \widehat{\beta_0} + \widehat{\beta_1} x_1 + \widehat{\beta_2} x_2$
  $\widehat{\beta_1} \times 100$: the percentage change in predicted y as x_1 increases by 1 unit, keeping x_2 constant

Models involving logarithms: log-level

- In our example, we use natural **logarithm of wage as the dependent variable**

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 IQ + u$$

- Holding IQ and u fixed,

$$\Delta \log(wage) = \beta_1 \Delta educ$$

so

$$\beta_1 = \frac{\Delta \log(wage)}{\Delta educ}$$

- Useful result from calculus:

$$100 \cdot \Delta \log(wage) \approx \%\Delta wage$$

- This leads to a simple interpretation of β_1:

$$100\beta_1 \approx \%\Delta wage \text{ when } \Delta educ = 1 \text{ holding IQ constant}$$

- If we do not multiply by 100, we have the decimal version (the proportionate change).
- In this example, $100\beta_1$ is often called the *return to education* (just like an investment). This measure is free of units of measurement of wage (currency, price level).

- Let's revisit the wage equation

$$\begin{aligned} \widehat{\log(wage)} &= 5.66 + 0.039\ educ + 0.006\ IQ \\ n &= 935,\ R^2 = .130 \end{aligned}$$

- These results tell us that ...
- Warning: This R-squared is not directly comparable to the R-squared when *wage* is the dependent variable. We can only compare R-squared of two models if they have the same dependent variable. The total variation (SSTs) in $wage_i$ and $\log(wage_i)$ are completely different.

Models involving logarithms: level-log

- We can use logarithmic transformation of x as well.
- y on log of x: $\widehat{y} = \widehat{\beta_0} + \widehat{\beta_1}\log(x_1) + \widehat{\beta_2}x_2$
  $\widehat{\beta_1}/100$: the change in predicted y as x_1 increases by 1%, keeping x_2 constant
- Example: The context: determining the effect of cigarette smoking during pregnancy on health of babies. Data: birth weight in kg, family income in \$s, mother's education in years, number of cigarettes smoked per week by the mother during pregnancy

$$\widehat{bwght} = 3.22 + 0.050\log(finc) + 0.001educ - 0.013cigs$$
$$n = 1387,\ R^2 = 0.03$$

- The coefficient of $\log(finc)$: Consider newborn babies whose mothers have the same level of education and the same smoking habits. Every percentage increase in family income increases the predicted birth weight by $0.0005kg = 0.5g$.

Models involving logarithms: log-log

- log of y on log of x: $\widehat{\log(y)} = \widehat{\beta_0} + \widehat{\beta_1}\log(x_1) + \widehat{\beta_2}x_2$
  $\widehat{\beta_1}$: the percentage change in predicted y as x_1 increases by 1%, keeping x_2 constant. $\widehat{\beta_1}$ in this case is also called the estimated elasticity of y with respect to x_1 all else constant.
- Example 6.7 in the textbook: Predicting CEO salaries based on sales, market value of the firm (*mktval*) and years that CEO has been in his/her current position (*tenure*):

$$\begin{aligned} \widehat{\log(salary)} &= 4.50 + 0.16\log(sales) + 0.11\log(mktval) + 0.01 tenure \\ n &= 177,\ R^2 = .318 \end{aligned}$$

The coefficient of $\log(sales)$: In firms with the exact same market valuation with CEOs who have the same level of experience, a one percent increase in sales increases the predicted CEO salary by 0.16%

Considerations for using levels or logarithms

1. A variable must have a strictly positive range to be a candidate for logarithmic transformation.
2. Thinking about the problem: does it make sense that a unit change in x leads to a constant change in the magnitude of y or a constant % change in y?
3. Looking at the scatter plot, if there is only one x.
4. Explanatory variables that are measured in years, such as years of education, experience or age, are not logged.
5. Variables that are already in percentages (such as interest rate or tax rate) are not logged. A unit change in these variables already is a one percent change.
6. If a variable is positively skewed (like income or wealth), taking logarithms makes its distribution less skewed.

There is a good discussion in 6-2a

Other non-linear models: Quadratic terms

- We can have x^2 as well as x in a multiple regression model:

$$\hat{y} = \widehat{\beta_0} + \widehat{\beta_1} x + \widehat{\beta_2} x^2$$

- In this model

$$\frac{\partial \hat{y}}{\partial x} = \widehat{\beta_1} + 2\widehat{\beta_2} x,$$

  that is, the change in predicted y as x increases depends on x.

- Here, the coefficients of x and x^2 on their own do not have meaningful interpretations, because ...

- $\frac{\partial \hat{y}}{\partial x} = 0$ at $x = -\frac{\widehat{\beta_1}}{2\widehat{\beta_2}}$. At this level of x the predicted y is at its maximum if $\widehat{\beta_2} < 0$, and it is at its minimum if $\widehat{\beta_2} > 0$

Examples of the quadratic model

- Sleep and age: Predicting how long women sleep from their age and education level. Data: age, years of education, and minutes slept in a week recorded by women who participated in a survey

$$\widehat{sleep} = 4428.07 - 49.30age + 0.58age^2 - 13.92educ$$

$$n = 305, \ R^2 = 0.03$$

- Keeping education constant, the predicted sleep reaches its minimum at the age $\frac{49.3}{2 \times 0.58} = 42.5$
- House price and distance to the nearest train station: Data: price (000$s), area ($m^2$), number of bedrooms and distance from the train station (km) for 120 houses sold in a suburb of Melbourne in a certain month:

$$\widehat{price} = -29.08 + 1.22area + 63.76beds + 1169.71train - 687.64train^2$$

$$n = 120, \ R^2 = 0.54$$

- The ideal distance from the train station is $\frac{1169.71}{2 \times 687.64} = 0.85km$ because ...

Considerations for using a linear or a quadratic model

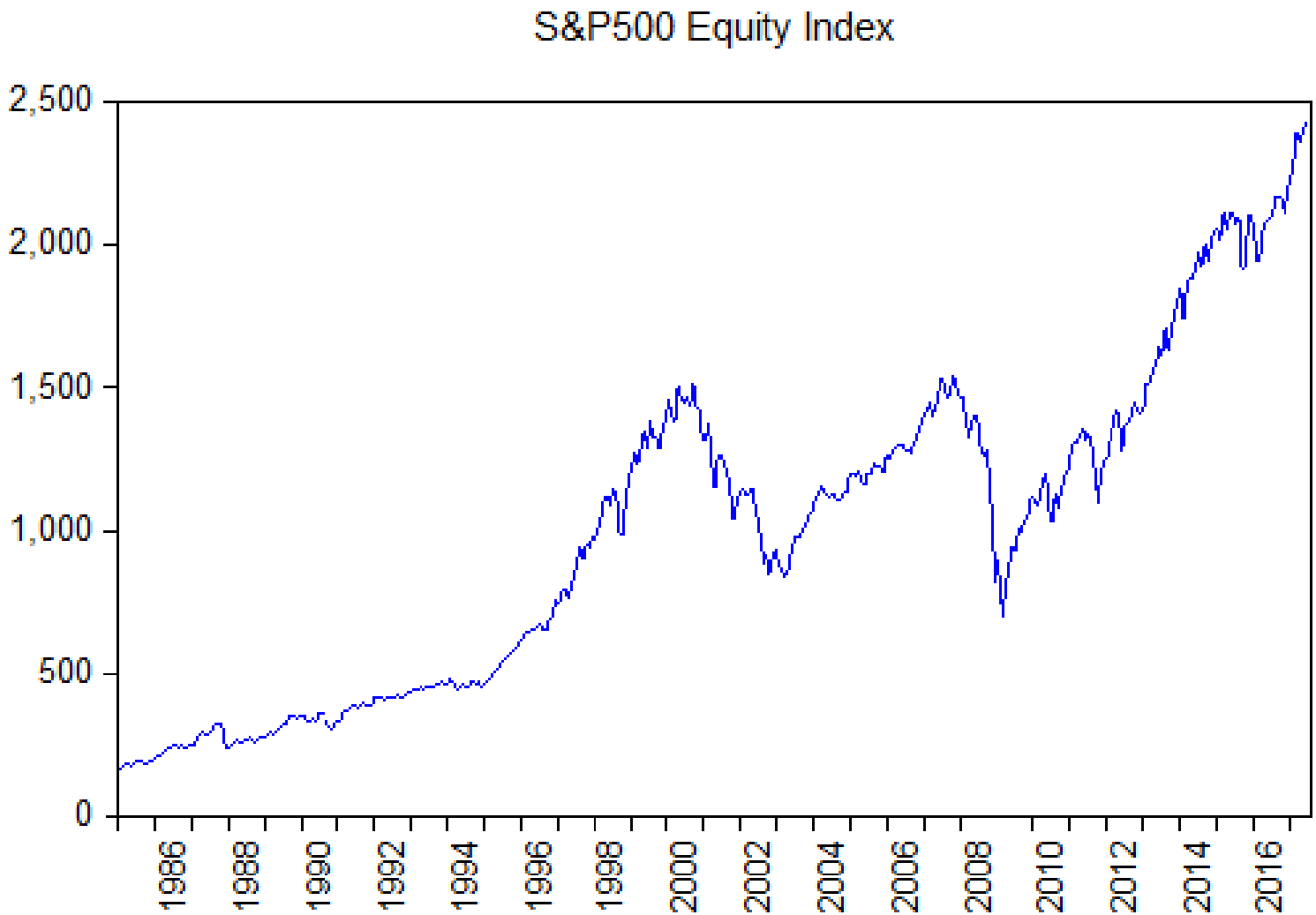
1. Thinking about the problem: is a unit increase in x likely to lead to a constant change in y for all values of x, or is it likely to lead to a change that is increasing or decreasing in x?
2. Is there an optimal or peak level of x for y? Example: *wage* and *age*, house price and distance to train station.
3. If there is only one x, looking at scatter plot can give us insights.
4. In multiple regression, there are tests that we can use to check the specification of the functional form (RESET test, to be covered later if time permits)
5. When in doubt, we can add the quadratic term and check its statistical significance, or see if it improves the adjusted R^2.

Transformation of persistent time series data

- A number of economic and financial series, such as interest rates, foreign exchange rates, price series of an asset tend to be highly persistent.
- This means that the past heavily affects the future (but not vice versa)
- A time series can be subject to different types of persistence (deterministic or stochastic).
- A common feature of persistence is lack of mean-reversion. This is evident by visual inspection of a line chart of the time series.

Empirical example

- E.g. Below is displayed the Standard and Poors Composite Price Index from January 1985 to July 2017 (monthly observations)



Transformation of persistent time series data

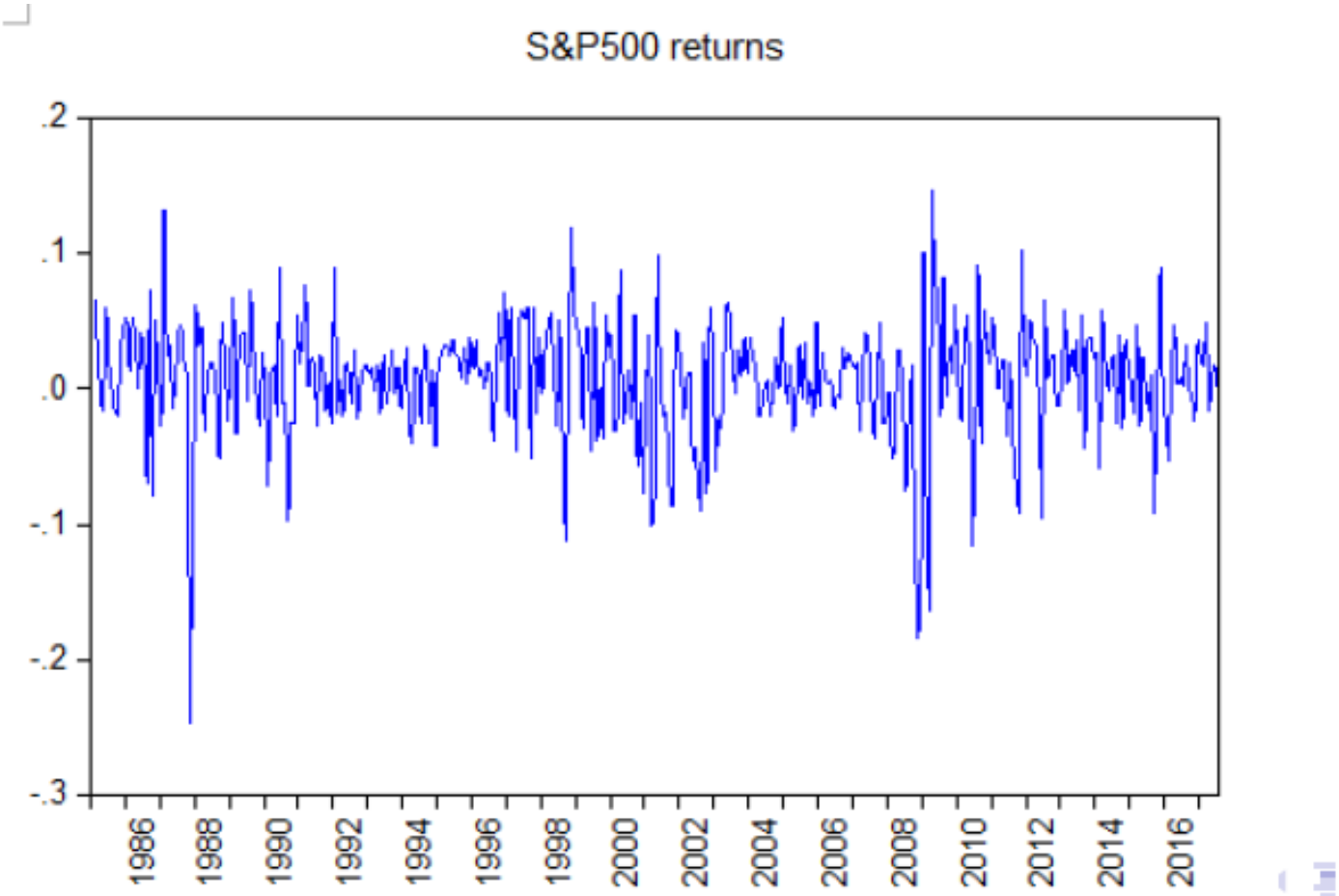
- In such cases the researcher transforms the time series by (log) differencing over the preceding period.
- The transformed series is then easier to handle and has more attractive statistical properties.
- More precisely, assume that the $S\&P$ price index at time t is denoted by P_t.
- The said log differencing is expressed as:

$$\begin{aligned}\log(P_t) - \log(P_{t-1}) &= \log(\frac{P_t}{P_{t-1}}) \\ &= \log(1 + \frac{P_t - P_{t-1}}{P_{t-1}}) = \log(1 + r_t) \approx r_t,\end{aligned}$$

where $100 \times r_t$ denotes the $\%\Delta P_t$ (for small r_t).

Transformation of persistent time series data

- By differencing the logarithmic transformation to our S&P500 price series, we obtain the S&P500 returns (sometimes called log-returns)



Model Selection Criteria

- *Parsimony* is very important in predictive analytics (which includes forecasting). You may have heard about the *KISS* principle. If not, google it!
- We want models that have predictive power, but are as parsimonious as possible
- We cannot use R^2 to select models, because R^2 always increases as we make the model bigger, even when we add irrelevant and insignificant predictors
- One can use t-stats and drop insignificant predictors, but when there are many predictors, and several of them are insignificant, the model that we end up with depends on which predictor we drop first

Model Selection Criteria

- Model selection criteria are designed to help us with selecting among competing models
- All model selection criteria balance the (lack of) fit of the model (given by its sum of squared residuals) with the size of the model (given by the number of parameters)
- These criteria can be used when modelling time series data too. Some adjustments to the formulae are needed if lagged values of the dependent variable are included as regressors in the model. Different software make the relevant alterations automatically
- There are many model selection criteria, differing on the penalty that they place on the lack of parsimony

1. Adjusted R^2 (also known as $\bar{R}^2$)

$$\bar{R}^2 = 1 - \frac{SSR/(n-k-1)}{SST/(n-1)}$$

2. Akaike Information Criteria (AIC)

$$AIC = c_1 + \ln(SSR) + 2k/n$$

3. Hannan-Quinn Criterion (HQ)

$$HQ = c_2 + \ln(SSR) + 2k\ln(\ln(n))/n$$

4. Schwarz or Bayesian Information Criterion (SIC or BIC)

$$BIC = c_3 + \ln(SSR) + k\ln(n)/n$$

c_1, c_2 and c_3 are constants that do not depend on the fit or number of parameters, so play no important role. ln is the natural logarithm. Also, all models are assumed to include an intercept

- BIC gives the largest penalty to lack of parsimony, i.e. if we use BIC to select among models, the model we end up with would be the same or smaller than the model that we end up with if we used any of the other criteria
- The order of the penalties that each criterion places on parsimony relative to fit is (for $n > 16$)

$$P(BIC) > P(HQ) > P(AIC) > P(\bar{R}^2)$$

- Remember that with BIC, HQ or AIC, we choose the model with the smallest value for the criterion, whereas with $\bar{R}^2$, we choose the model with the largest $\bar{R}^2$
- Different software may report different values for the same criterion. That is because some include c_1, c_2 and c_3 and some don't. The outcome of the model selection exercise does not depend on these constants, so regardless of the software, the final results should be the same.

- Example: Making an app to predict body fat using height (H), weight (W) and abdomen circumference (A)

Predictors	R^2	$\bar{R}^2$	*AIC*	*HQ*	*SC*
W	0.376	0.373	6.474	6.485	6.502
A	0.662	0.661	5.860	5.872	5.888
H	0.008	0.004	6.937	6.949	6.965
W, A	0.719	0.716	**5.685**	**5.702**	**5.727**
W, H	0.461	0.457	6.334	6.351	6.376
A, H	0.688	0.686	5.788	5.805	5.830
W, A, H	0.721	**0.718**	**5.685**	5.707	5.741

- For different class of models, experts use different criteria
- My favourite is HQ (because of my research on multivariate time series, and because Ted Hannan was a great Australian statistician), although not all software report HQ

Confidence Intervals for the Conditional Mean versus Prediction Intervals

- Remember that the population model

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik} + u_i, \text{ for all } i \quad (1)$$

with the CLM assumptions implying that

$$E(y_i \mid x_{i1}, \ldots, x_{ik}) = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik}, \text{ for all } i \quad (2)$$

- Our estimated regression model provides

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + ... + \hat{\beta}_k x_{ik}, \text{ for all } i \quad (3)$$

- Comparing (3) and (2), we see that $\hat{y}_i$ gives us the best estimate of the conditional expectation of y_i given $x_{i1}, \ldots, x_{ik}$
- Also, since u_i is not predictable given $x_{i1}, \ldots, x_{ik}$, $\hat{y}_i$ is also our best prediction for y_i

Confidence Intervals for the Conditional Mean versus Prediction Intervals

- As an estimator for conditional mean, the error in $\hat{y}_i$ is only due to estimation uncertainty in $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$
- We can compute $se(\hat{y}_i)$ using the estimated variance covariance matrix of estimated parameters, or we can get it with a cool trick
- The 95% confidence interval for $E(y_i \mid x_{i1}, \ldots, x_{ik})$ is

$$\hat{y}_i \pm t_{n-k-1}(0.025) * se(\hat{y}_i)$$

- The y_i itself also includes u_i, which is another source of uncertainty that our model cannot explain
- Therefore, the estimated variance of prediction error $\hat{e}_i = y_i - \hat{y}_i$ is $\hat{\sigma}^2 + [se(\hat{y}_i)]^2$, which implies

$$se(\hat{e}_i) = \sqrt{\hat{\sigma}^2 + [se(\hat{y}_i)]^2}$$

- The 95% prediction interval for y_i is

$$\hat{y}_i \pm t_{n-k-1}(0.025) * se(\hat{e}_i)$$

Are we interested in the conditional mean of y or y itself?

Examples from the real world

- Rob Hyndman's problem (Prof Hyndman is a world renowned expert in forecasting - He teaches ETC3550: Applied Forecasting. Don't miss the opportunity!): Provide an estimate of base electricity load in 2025 for every state and territory in Australia based on population growth, warming trend in temperature, growth in energy intensive industries, and other relevant factors. Here, the interest is in the mean load, for long term planning, not day to day variations in electricity load.
- EGRV's problem: Provide the probability of peak hour electricity load exceeding 2500 MW tomorrow given the temperature forecasts available now. Operations manager of power company needs this to know if they need to bring a gas powered generator on-line or not. Here, the object of interest is y itself.

Are we interested in the conditional mean of y or y itself?

Example 2

- Simon's problem (Dr Simon Angus is an economist/data scientist in the Economics Department): Simon is a religious person. He belongs to a modern church. Being a data scientist, he has counted the number of people who attended his church's meetings every Sunday for the last 351 weeks.
- Simon is asked to provide a measure of success of the church in the community in 12 weeks' time. This is a question about predicting the mean attendance in 12 weeks' time, and not getting distracted by week to week variation.
- Simon is also asked to provide the number of chairs that the church needs so that they could be 99% certain that all who attend next week's meeting would have a seat. The object of the interest is the number of attendees, not just its mean.

Computing Predictions and Prediction Intervals

Example 3: Body fat app

- We want to make a body fat app, that users enter their weight and abdomen circumference, and they get a prediction of their body app. We have data of precisely measure body fat for 252 people. After we estimate the model of body fat % (BF) conditional of weight (W) and abdomen circumference (A), we can plug in any person's W and A and get a point prediction for that person's BF%
- Example: From the estimated model

$$\widehat{BF} = -41.34812 - 0.300829W + 0.915138A$$

prediction of BF for a person with $W = 80kg$ and $A = 94cm$ is

$$-41.34812 - 0.300829 \times 80 + 0.915138 \times 94 = 20.60853$$

Computing Predictions and Prediction Intervals

A useful trick

- Or we can use our knowledge of geometry of OLS to trick the computer with a suitable reparameterisation. For example for $W = 80kg$ and $A = 94cm$:

Dependent Variable: BF
Method: Least Squares
Sample: 1 252
Included observations: 252

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-41.34812	2.412986	-17.13567	0.0000
W	-0.300829	0.042495	-7.079076	0.0000
A	0.915138	0.052535	17.41944	0.0000

R-squared	0.718726	Mean dependent var	18.93849
Adjusted R-squared	0.716467	S.D. dependent var	7.750856
S.E. of regression	4.127160	Akaike info criterion	5.684889
Sum squared resid	4241.328	Schwarz criterion	5.726906
Log likelihood	-713.2961	Hannan-Quinn criter.	5.701796
F-statistic	318.1296	Durbin-Watson stat	1.804965
Prob(F-statistic)	0.000000		

Dependent Variable: BF
Method: Least Squares
Sample: 1 252
Included observations: 252

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	20.60856	0.287066	71.79033	0.0000
W-80	-0.300829	0.042495	-7.079076	0.0000
A-94	0.915138	0.052535	17.41944	0.0000

R-squared	0.718726	Mean dependent var	18.93849
Adjusted R-squared	0.716467	S.D. dependent var	7.750856
S.E. of regression	4.127160	Akaike info criterion	5.684889
Sum squared resid	4241.328	Schwarz criterion	5.726906
Log likelihood	-713.2961	Hannan-Quinn criter.	5.701796
F-statistic	318.1296	Durbin-Watson stat	1.804965
Prob(F-statistic)	0.000000		

- The advantage of this is that we also get the standard error, which shows how well $\widehat{BF}$ estimates $E(BF \mid W, A)$. This error is due to estimation of β_0, β_1 and β_2, so we call it "estimation uncertainty"

Computing Predictions and Prediction Intervals

- There are two sources of error that make our prediction uncertain.
  1. Estimation uncertainty caused by not knowing the value of the true parameters. This uncertainty gets smaller with sample size. This is often ignored when the sample size is large.
  2. The more important source of uncertainty is u that is not predictable by our predictors, even if we knew the true values of β. This one is independent of the sample size.
- Compare 100 samples when sample size is 10 and when sample size is 100 at `https://www.geogebra.org/m/pemtjeBA`
- In the BF example: $se(\widehat{BF}) = 0.287$ and $\hat{\sigma} = 4.127$. As you see first one is much smaller than the second

Computing Predictions and Prediction Intervals

- From the output:

$$\begin{aligned}
\widehat{BF} &= 20.609 \\
se(\widehat{BF}) &= 0.287 \\
\hat{\sigma} &= 4.127 \\
\Rightarrow \widehat{Var(\hat{e})} &= 4.127^2 + 0.287^2 = 17.1145 \\
\Rightarrow se(\hat{e}) &= \sqrt{17.1145} = 4.137 \\
95\% \text{ prediction interval} &= 20.609 \pm 1.97 \times 4.137 \\
&= [12.459, 28.759]
\end{aligned}$$

- For males 14-17% is fit and 18-25% is the acceptable range
- Note how little standard error increased relative to $\hat{\sigma}$
- Also note that the critical value is very close to 2
- This justifies using $[\widehat{BF} \pm 2\hat{\sigma}]$ for the 95% prediction interval, a rule of thumb that is often used in practice

Summary

- **Modelling non-linear relationships:** The linear regression model is only linear in parameters. By using non-linear transformations (such as logarithmic or quadratic) of y or any of the x variables, we can model non-linear relationships with the regression model.
- **The adjusted R^2:** We can use $\bar{R}^2$ (and other model selection criteria) to help us choose the best model for predictive analytics.
- **Point interval and prediction interval:** We explained how to provide point prediction and a prediction interval for the target variable given specific values for explanatory variables using our estimated model
- We now understand why people use $\hat{y} \pm 2\hat{\sigma}$ as a rule of thumb for providing prediction intervals. Provided that the size of the estimation sample is large, this is a pretty good rule of thumb!