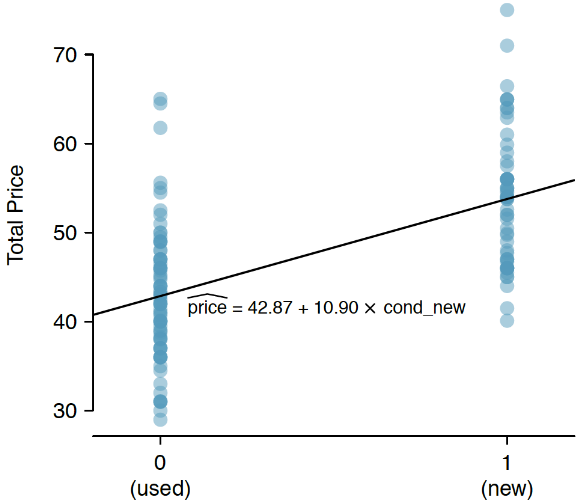
Lecture 26: Multiple Regression

Chapter 8.1

Categorical Predictor x With Two Levels



Questions for Today

Say on top of `cond_new` we are given three additional predictors:

- `stock_photo`: is there a stock photo?
- `duration`: length of the auction in days (1 to 10)
- `wheels`: number of Wii wheels included

Questions for Today

How do we simultaneously incorporate all four predictors to model the eBay auction `price`?

Multiple Regression

Use multiple regression (done via computer) where

$$\begin{aligned}\texttt{price} &= \beta_0 + \beta_1 \times \texttt{cond\_new} + \beta_2 \times \texttt{stock\_photo} + \\ &\quad \beta_3 \times \texttt{duration} + \beta_4 \times \texttt{wheels}\end{aligned}$$

or in general...

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

for k predictor variables.

Point Estimates, Fitted Values, and Residuals

Just as with SLR, we compute point estimates $b_0, b_1, \ldots, b_k$ of the parameters $\beta_0, \beta_1, \ldots, \beta_k$.

For each observation $i = 1, \ldots, n$ the i^{th} residual $e_i = y_i - \widehat{y}_i$. The point estimates minimize the least squares criterion:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2$$

Multiple Regression Results Table

On page 357:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	36.21	1.51	23.92	0.00
`cond_new`	5.13	1.05	4.88	0.00
`stock_photo`	1.08	1.06	1.02	0.31
`duration`	-0.03	0.19	-0.14	0.89
`wheels`	7.29	0.55	13.13	0.00
				$df = 136$

where $df = n - k - 1 = 141 - 4 - 1 = 136$

Interpretation of Point Estimates

The point estimates are interpreted as follows, for example:

- b_4 of `wheels`: holding all other variables constant (all other things being equal), the average difference in price associated with each additional Wii wheel is \$7.29.
- b_3 of `duration`: holding all other variables constant, every additional day of auction is associated with an average decrease of \$0.03 in price.

Comparison of Results

For simple linear regression:

	Estimate	Std. Error	t value	Pr(>|t|)
`cond_new`	10.90	1.26	8.66	0.00

For multiple regression:

	Estimate	Std. Error	t value	Pr(>|t|)
`cond_new`	5.13	1.05	4.88	0.00

Why the different point estimate?

Comparison of Result

Because `cond_new` is linearly correlated with `wheels`. We say that two predictor variables are collinear when they are correlated, and this complicates model estimation.

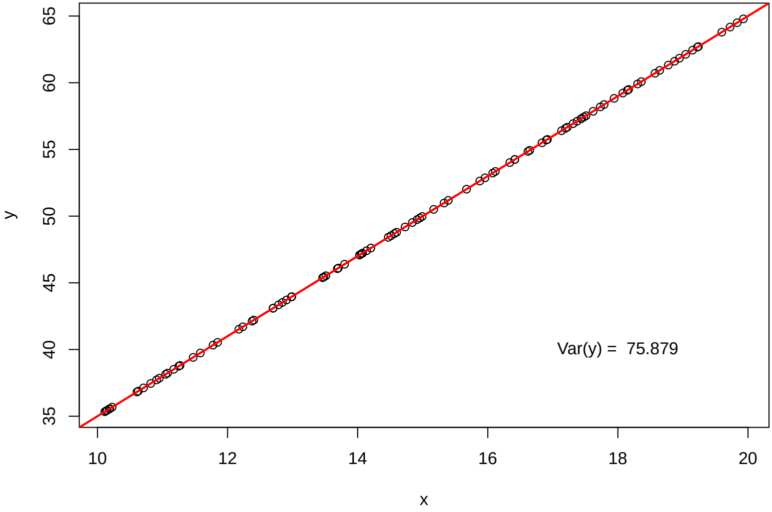
In general we must be wary of predictor variables that are collinear, because the coefficient estimates may change erratically in response to small changes in the model or the data.

R^2 to Describe the Strength of Fit

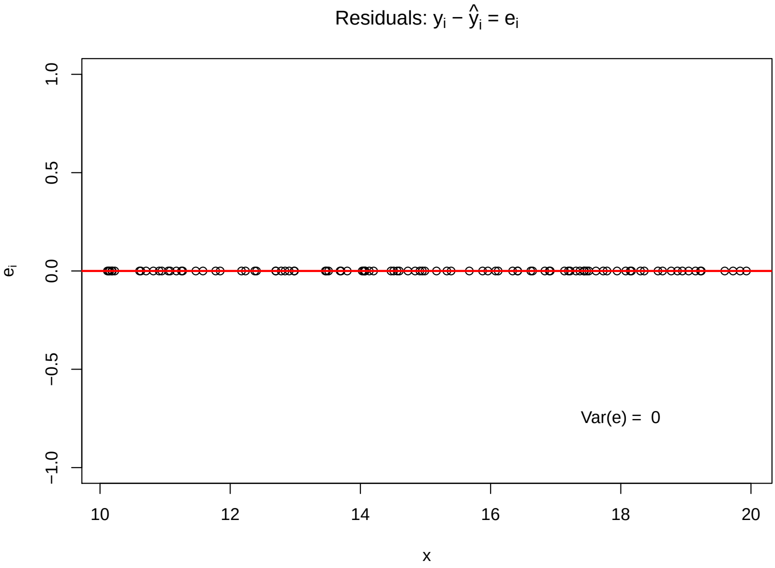
R^2 is a measure (between 0 and 1) of how well the model fits the data. It measures the proportion of the total variability in y explained by the model i.e. the least squares line:

$$
\begin{aligned}
R^2 &= 1 - \frac{\text{variability in residuals}}{\text{variability in the outcome}} \\
&= 1 - \frac{Var(e_i)}{Var(y_i)}
\end{aligned}
$$

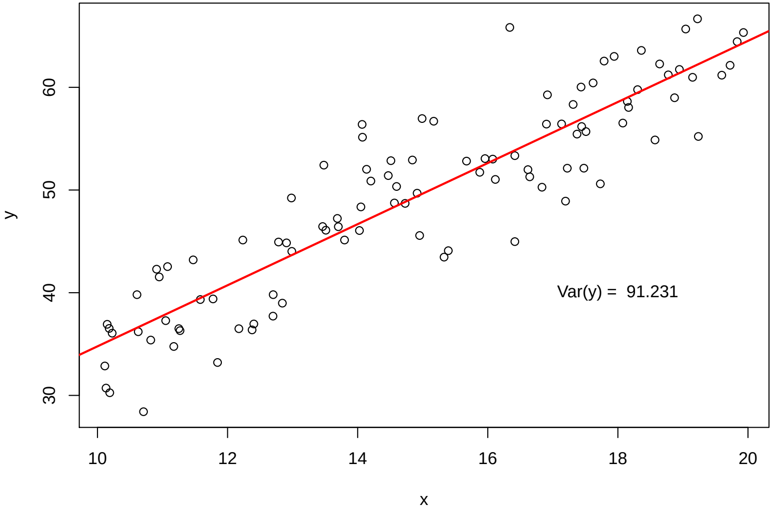
Data Set Example 1



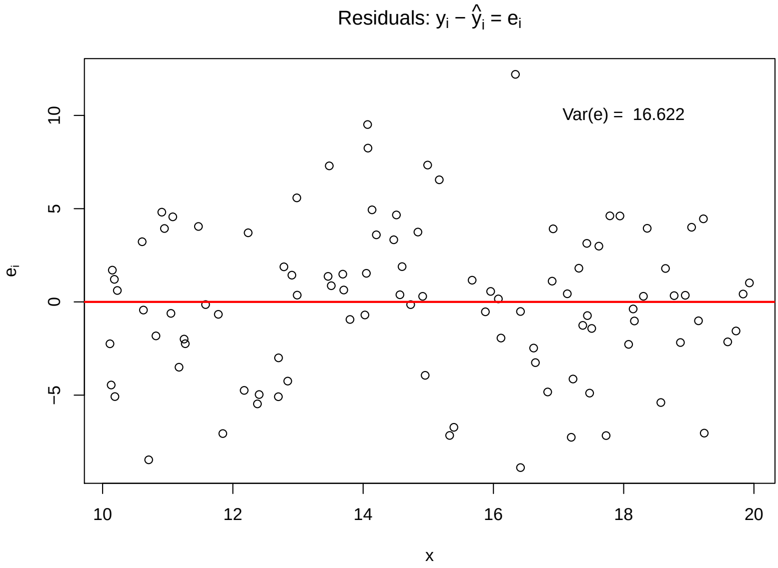
Data Set Example 1



Data Set Example 2



Data Set Example 2



R^2 vs R

Special case: Only for linear regression: R^2 is the correlation coefficient R squared. The correlation coefficient does not exist for multiple regression.

Important Concept in Model Fitting

R^2_{adj} describes the strength of fit while adhering to the following:

- Parsimony: Adoption of the simplest assumption in the formulation of a theory or in the interpretation of data.
- Occam's Razor: When you have two competing theories that make exactly the same predictions, the simpler one is the better.

Adjusted R^2_{adj}

The adjusted R^2_{adj} takes into account the number of predictor variables k you used to fit your model:

$$R^2_{adj} = 1 - \frac{\frac{Var(e_i)}{n-k-1}}{\frac{Var(y_i)}{n-1}} = 1 - \frac{Var(e_i)}{Var(y_i)} \times \frac{n-1}{n-k-1}$$

Parsimony/Occam's Razor

Say for two models applied to the same data:

- Model 1: $\widehat{y}_i = b_0 + b_1 x_{1,i}$ with residuals $e_i = y_i - \widehat{y}_i$
- Model 2: $\widehat{y}_i^* = b_0^* + b_1^* x_{1,i} + b_2^* x_{2,i}$ with residuals $e_i^* = y_i - \widehat{y}_i^*$

both models fit the data near identically, i.e. the residuals are similar

$$Var(e_i) \approx Var(e_i^*) \ldots$$

Then we should pick Model 1 over Model 2 since it is simpler: it has fewer of predictors k.

Pared Down Mario Kart Regression Output

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)       41.34       1.71   24.15  < 2e-16
condused          -5.13       1.05   -4.88 2.91e-06
stockPhotoyes      1.08       1.06    1.02    0.308
duration          -0.03       0.19   -0.14    0.888
wheels             7.30       0.55   13.13  < 2e-16
---

Residual standard error: 4.901 on 136 degrees of freedom
Multiple R-squared:  0.719, Adjusted R-squared:  0.7108
```

Duration doesn't seem to be all that informative. Why not drop it?

Pared Down Mario Kart Regression Output

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)       41.22       1.49   27.65  < 2e-16
condused          -5.18       1.00   -5.20 7.21e-07
stockPhotoyes      1.12       1.02    1.10    0.275
wheels             7.30       0.54   13.40  < 2e-16
---

Residual standard error: 4.884 on 137 degrees of freedom
Multiple R-squared:  0.719, Adjusted R-squared:  0.7128
```

Next Time

Is there a systematic way to pick which predictor variables to include?

Checking model assumptions as well.