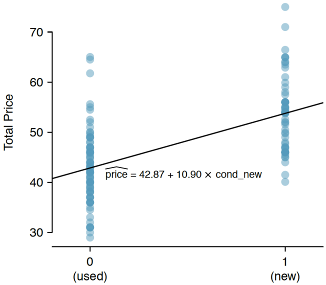
Lecture 26: Multiple Regression

Chapter 8.1

Categorical Predictor x With Two Levels



Simple Linear Regression Regression Table

eBay price of old vs new Mario Kart using $n = 141$. On page 355:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	42.87	0.81	52.67	0.0000
`cond_new`	10.90	1.26	8.66	0.0000
				$df = 139$

Confidence Interval and Hypothesis Test for β_1

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

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

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

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}

Parsimony/Occam's Razor

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.